MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

ÉCOLE NATIONALE SUPÉRIEURE DE MANANGEMENT
ENSM. Pôle Universitaire de KOLÉA

MEMOIRE DE FIN D'ETUDE

Master Professionnel en Management Stratégique et Système d'Information

LES FACTEURS CLÉS DE SUCCÈS DU MÉTIER DE DATA SCIENTIST
CAS - BIG MAMA TECHNOLOGY-

Elaborée par : GRINE Hiba Miled **Encadrée par** : GHIDOUCHE AÏT YAHIA Kamila

ANNEE **2016/2017**

RESUME

Suite à l'explosion du volume de données, les systèmes d'information ont connu une extrême sophistication et une très haute technicité, sollicitant par cela des compétences variés, et interdisciplinaire. Notre recherche s'intéresse à ces compétences, et a comme but de comprendre qu'est-ce qu'un data scientist ? Quelles sont ses compétences et rôles liés à l'interaction de l'environnement technologique complexe des BIG DATA ? Sans souci d'exhaustivité, c'est ceux à quoi nous tentons de répondre. Notre recherche présente une liste des facteurs clés de succès constituant le métier d'un data scientist et des perspectives axés sur la généralisation et la contextualisation des facteurs.

Mots clés : Data Scientist, Big data, Machine Learning, FCS

ABSTRACT

Due to the explosion of the volume of data, information systems have been extremely sophisticated and highly technical, requiring a variety of interdisciplinary skills. Our research focuses on these skills, and aims to understand what a data scientist is? What are its competencies and roles related to the interaction of the complex technological environment of the BIG DATA. Without looking for exhaustiveness, this is what we try to answer, our research presents a list of the key success factors constituting the profession of a data scientist. Our perspective focuses on the generalization and contextualization of factors.

Key words: Data scientist, Big data, Machine Learning, FCS

ملخّص

بعد انفجار في حجم المعطيات، شهدت نظم المعلومات تطور ذات تقنية عالية، يتطلب من خلالها مهارات متنوعة ومتعددة التخصصات. يركز بحثنا على هذه المهارات، ويهدف إلى فهم ما هو عالم المعطيات؟ ما هي المهارات والأدوار المرتبطة بتفاعله مع البيئة التكنولوجية المعقدة للبينات الضخمة ؟ تلك هي الاسئلة التي نحاول من خلالها الإجابة على اطروحتنا، يقدم بحثنا قائمة من العوامل الرئيسية التي تشكل نجاح عالم البيانات، و نوجه عملنا الى وجهات النظر التي تركز على تعميم وسياقة هاته العوامل.

كلمات مفتاحيّة: عالم المعطايات ، المعطيات الضخمة، تعلم الالي، عوامل النج

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES TERMES ANGLO-SAXONS

BIG DATA : Données massives

BUSINESS INTELIGENCE : L'informatique décisionnelle

Buzz Word : Expression à la mode

Cloud Computing : L'informatique en nuage ou nuagique ou l'infonuagique

Clustering : classification non supervisé

Data Analayse : L'analyse de données

Data manipulation : manipulation de données

Data Miner : Expert de classification et de segmentation d'un très grand volume de données

Data science : Science de donnée

Data science : science des données

Data scientist : Scientifique de la donnée, manipulateur de de la donnée

Data : Les données

Dataset : Un jeu de données, ensemble de valeurs (ou données).

Framework : cadre d'applications

INPUT : Intrant, tous les produits nécessaires au fonctionnement d'un ensemble

Machine Learning : Apprentissage automatique

Mastering : maîtriser

Need to have : les facteurs clés de succès nécessaires et indispensable au métier,

Nice to have : Les facteurs clés de succès optionnels

Open Source : code source ouvert

Skills : compétences

Userfriendly : convivial, facile à utiliser par ses usagers.

REMERCIMENTS :

Pour commencer, nous remercions ALLAH le tout puissant de nous avoir donné la force et la patience d'accomplir ce modeste travail.

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui nous tenons témoigné toute notre reconnaissance. Mes remerciements vont à :

Nos très chers parents pour nous avoir encouragés, supportés, soutenus, motivés, épaulés. Sans eux, nous n'en serons pas là.

À nos chères sœurs Zineb, Khaoula, Linda pour leurs aides précieuses, surtout dans cette période éprouvante qu'est la dernière ligne droite.

À Mme Kamila Ghidouche Ait Yahia, notre promotrice de cette recherche, pour avoir accepté de nous soutenir dans l'aventure de ce mémoire, pour sa patience, son temps, ses orientations et ses conseils précieux.

À Mr Ryad Zenine, notre tuteur, pour nous avoir permis d'apprendre sur le métier de data scientist, de nous avoir consacré son temps, pour ses explications, formations et aide précieuse.

Aux data scientists, pour leur temps, attention, orientation et accueil sympathique lors des jours précédant de stage.

À notre meilleure amie Ikram Loubna Nour, pour nous avoir épaulé moralement tous les jours dans la construction de ce mémoire.

A notre cher ami et collègue Billel pour ses conseils et le soutien qui nous accorde quotidiennement.

Nous remercions également les marqueteurs Malek pour son aide, puis Khaled pour sa gentillesse, sa bonne humeur, sa drôlerie et son encouragement tout au long de ce travail.

Nous exprimons également toute notre reconnaissance et gratitude à l'administration et à l'ensemble du corps enseignant de l'ENSM, pour leurs efforts à nous garantir la continuité et l'aboutissement de ce programme de Master,

Aux personnes qui nous ont apporté leur aide et qui ont contribué à l'élaboration de ce travail ainsi qu'à la réussite de cette formidable année académique.

Dans l'impossibilité de citer tous les noms, nos sincères remerciements vont à les toute personnes ayant permis à la réalisation de ce mémoire par leurs conseils, encouragements, sourires, mots tendres.

INTRODUCTION

1. Introduction :

L'importance d'un système d'information dans une entreprise n'est plus à démontrer. La pérennité de l'entreprise dépend de son bon fonctionnement. Celui-ci facilite et optimise les processus métier de l'entreprise et toutes les fonctions en sont concernées (Brasseur, 2013, p. 11).

On oublie parfois que la matière première du système d'information, est la donnée, selon toutes les confrontations que cela implique. Il est clair que la donnée est devenue un facteur-clé de succès pour les entreprises, néanmoins il faut savoir en faire bon usage pour gagner en compétitivité (Brasseur, 2013, p. 12).

Depuis quelques années, on assiste à une prise de conscience de plus en plus importante de l'avantage compétitif que l'on peut tirer des données. Les initiatives des entreprises en ce sens annoncent une tout autre tournure, une tournure vers une nouvelle ère.

Une nouvelle ère, portée par un nouveau phénomène vient révolutionner l'usage quotidien de la *data* et bouleverser l'équilibre des entreprises, il s'agit de l'explosion du volume de données appelé plus communément " *BIG DATA*".

Le développement spectaculaire d'internet et en particulier la transformation numérique des entreprises en conséquence provoquent une avalanche de données (Brasseur, 2013). De l'histoire de navigation aux localisations GPS, jusqu'au rythme cardiaque, à la météo et au solde des comptes courants, un milliard de données est récolté par les mobiles, applications et autres objets connectés.

Ces grandes données sont ainsi considérées comme fournissant à l'organisation le grand potentiel pour générer de nouvelles idées ou pour créer de nouvelles formes de valeur. Il est clair qu'un grand volume de données nécessite de faire les choses à grande échelle, ce qui indique que pour réaliser, capturer et tirer pleinement parti des avantages potentiels des grandes données, de grandes compétences et un grand savoir-faire en la matière doivent être promus (Brasseur, 2013, p. 136).

En réponse à ce besoin et à l'accélération des grands environnements de données, une nécessité pour un "data scientist" est apparue et a fortement augmenté dans le marché. La demande provient de petites et moyennes entreprises basées sur les données ainsi que de grandes entreprises sous la pression de la concurrence mondiale pour l'innovation et la qualité du service et des offres de produits.

En effet, avec l'arrivée de cette technologie *BIG DATA*, les attentes sont extrêmement fortes, tant d'un point de vue business, que technique. Le référentiel métier de la branche du numérique, de l'ingénierie, des études et du conseil et de l'événement (OPIIEC[1]), définit le Data Scientist comme "un expert de la gestion et de l'analyse pointue de données massives ("big data"), qui détermine à partir de sources de données multiples et dispersées des indicateurs permettant la mise en place d'une stratégie répondant à une problématique.

[1] L'Observatoire Paritaire de l'Informatique, de l'Ingénierie, des Etudes et du Conseil

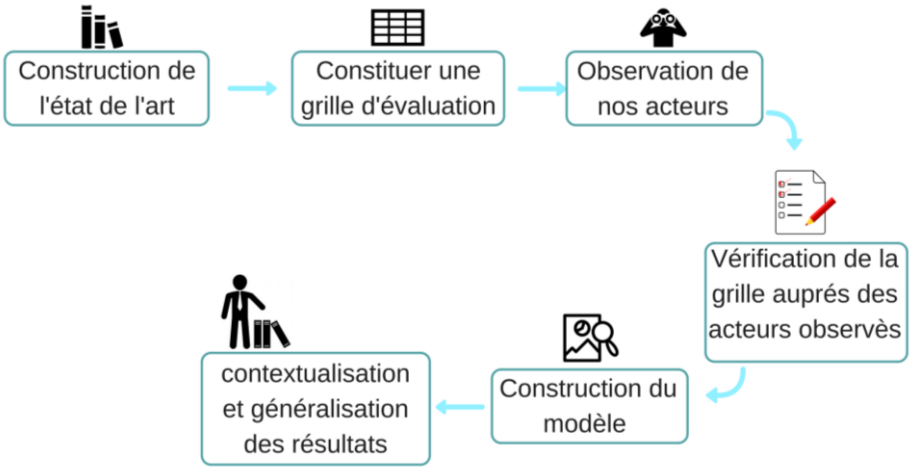
Il est donc spécialisé en statistiques, mathématiques, informatique et connaît parfaitement le secteur ou la fonction d'application des données analysées'' (OPIIEC, 2016). Cette définition nous donne déjà une petite idée du métier de data scientist, cependant il n'existe pas de définition claire et rigoureuse d'un data scientist, de sa formation, des compétences requises et son domaine d'expertise.

Pour tenter d'élucidation ce phénomène, notre travail de recherche vise à identifier les facteurs clés de succès de la réussite du métier de data scientist. Aussi afin de mener à bien cette étude, des bases théoriques tirées des plus rigoureuses bases de données existantes ont été choisies, et une pluralité de démarches méthodologiques a été adoptée.

Pour commencer, on s'inspire des travaux de recherche d'Aline Scouarnec "L'observation des métiers : définitions, méthodologie et "actionnabilité" en GRH[1]", qui met en évidence les différentes méthodologies et outils utilisés dans l'observation des compétences et métiers. Ces travaux s'inscrivent dans notre démarche, qui est le pluralisme méthodologique, et la méthode adoptée nous permet de se pourvoir des moyens conceptuels nécessaires afin d'analyser et de trianguler les observations et informations recueillies sur les acteurs et de les comparer à notre matériau théorique.

Notre présentation théorique s'organise en fonction de ces démarches :

Figure 01 : Objet de notre recherche



Source : Schéma réalisé par nos soins en utilisant l'outil Canva

Une enquête exploratoire a été menée auprès de nos acteurs afin de mettre en évidence ce métier de data scientist et les conséquences en termes de compétences, d'évolutivité et de formation. Notre paradigme étant constructiviste, les théories citées nous servent de cadre d'interprétation.

Notre travail se divise en quatre parties. Ainsi la 1ére partie est consacrée à la formulation de la problématique et à la pertinence de recherche. La 2éme partie est un voyage temporel dans la revue de littérature et une définition du cadre conceptuel de l'étude. Le 3éme chapitre détaille le cadre méthodologique entrepris pour mener à bien ce travail. Enfin le 4éme partie expose et commente les résultats obtenus à partir d'une méthode inductive. Pour conclure, nous proposons un modèle généraliste du data scientist dans le contexte algérien.

[1] Gestion des Ressources Humaines

CHAPITRE I : PROBLÉMATIQUE

1. Contexte et Objectifs de la recherche :

Le cadre de la recherche est défini dans ce chapitre. Cette démarche est importante, car elle permet au lecteur de connaître les règles posées par le chercheur pour effectuer son travail. Pour bien cadrer notre recherche, nous avons adopté la méthodologie de la recherche compréhensive de "Hervé Dumez", qui semble être bien structurée et pertinente pour notre recherche.

1.1. Contexte de la recherche :

Nous assistons à un bouleversement mondial, à une 3eme révolution industrielle qui est l'explosion du volume de la donnée (Rifkin, 2011), nous vivons actuellement une digitalisation ou *datafication* massive du quotidien, photo, message, documents, navigation web, vidéo, localisation GPS… Tout se numérise, s'archive et devient accessible à l'analyse (Laurent, 2016, p. 19).

Les grandes entreprises ont saisi l'enjeu et l'importance de l'aspect de la *data analyse* (Brasseur, 2013, p. 20). Dans ce sens, des vagues technologiques se sont succédées au cours de ces trente dernières années, le rythme des évolutions technologiques liées au traitement de l'information a triplé, voire quadruplé et a commandé beaucoup d'adaptabilité. Des systèmes centralisés d'IBM en 1960, aux ERP[1] 1990, la montée du *cloud computing* en 2000, à l'avènement du web 2.0 en 2003, qui ont facilité le partage et la diffusion de l'information, suivie par l'explosion du commerce électronique et tous les progrès réalisés dans la capacité de stockage et de calcul des données, ont nécessité un effort important en termes de compétences et de savoir-faire technologique. L'environnement dans lequel croit cette donnée est de plus en plus complexe et ce dernier nécessite une pluridisciplinarité focalisée sur la méthodologie d'analyser l'exploitation et la valorisation de ces données.

Un data scientist est donc, au centre d'un tétraèdre de compétences interconnectées : mathématiques, informatique, statistiques et connaissances métier, ces compétences s'accompagnent de maîtrise de l'environnement technologique du *BIG DATA* (Philippe Besse & Béatrice Laurent, 2016, p. 86), tout en ayant une sensibilisation aux aspects juridiques et éthiques lié au traitement de la donnée.

Dans ce contexte, notre champ d'étude appréhendé dans le cadre de notre travail est celui de l'identification des facteurs clés de succès de ce nouveau métier de la *data science*, ou plus précisément quels sont les compétences, le savoir-faire technologique et stratégique nécessaire dans le cadre d'un projet *BIG DATA*.

Cette problématique est abordée à travers deux catégories de facteurs clés (*need to have, nice to have*). Notre intérêt se porte particulièrement sur l'observation de ce nouveau métier complexe, en émergence.
Pour mieux comprendre ce métier, une période de trois mois passée au sein de l'entreprise BIG MAMA Technology nous a permis d'être au plus près de nos acteurs.

L'objectif général de cette étude est de contribuer à une meilleure définition et identification des facteurs clés de succès du métier de *data scientist*. La recherche menée est définie comme étant une recherche descriptive, car nous mettons l'accent sur la description et la classification des compétences d'un data scientist. Puis narrative, car nous narrons les

[1] Entreprise Ressource Planning (progiciel de gestion intégré)

discours tenus par les acteurs, les connaissances dont ils disposent, les situations qu'ils traversent, leurs pratiques, routine et évolution. Les résultats de recherches seront par la suite l'objet de rapprochement et de comparaison de notre matériau théorique.
Elle est également observationnelle Grâce à l'observation participante, nous essayons de comprendre le métier de data scientist, les compétences nécessaires liées à l'environnement complexe des *BIG DATA*, l'évolution de l'équipe data scientists, ainsi que leurs interactions avec différents projets.

4

2. Pertinence de la recherche :

Cette recherche a permis de dresser et d'évaluer un modèle centré sur le métier, les rôles et les compétences d'un *data scientist*. Ainsi, les fondements théoriques et méthodologiques de ce domaine de recherche ont été approfondis. Cette réflexion s'appuie sur différentes sciences, à savoir les ressources humaines, la sociologie et le management. Mais, dans le but de comprendre le métier en lui-même, l'environnement dans lequel il opère, l'étude s'appuie principalement sur la recherche et la compréhension de théories et de concepts en mathématiques, statistiques et informatique. Car il est désormais essentiel pour les organisations de comprendre les enjeux de l'explosion des données, d'une part, et de l'urgence de se munir des compétences pour la valorisation de ces données d'autre part.
Ainsi, les recherches sur les compétences et les rôles constituant ce métier de *data scientist* ont des clés déterminantes pour le début de notre quête de recherche. L'apport fondamental des recherches portant sur les compétences des data scientist réside dans la mise en exergue des facteurs déterminants ce métier, d'un point de vue fonctionnel et managérial.

L'ensemble des apports développés ci-après confirme cette assertion.

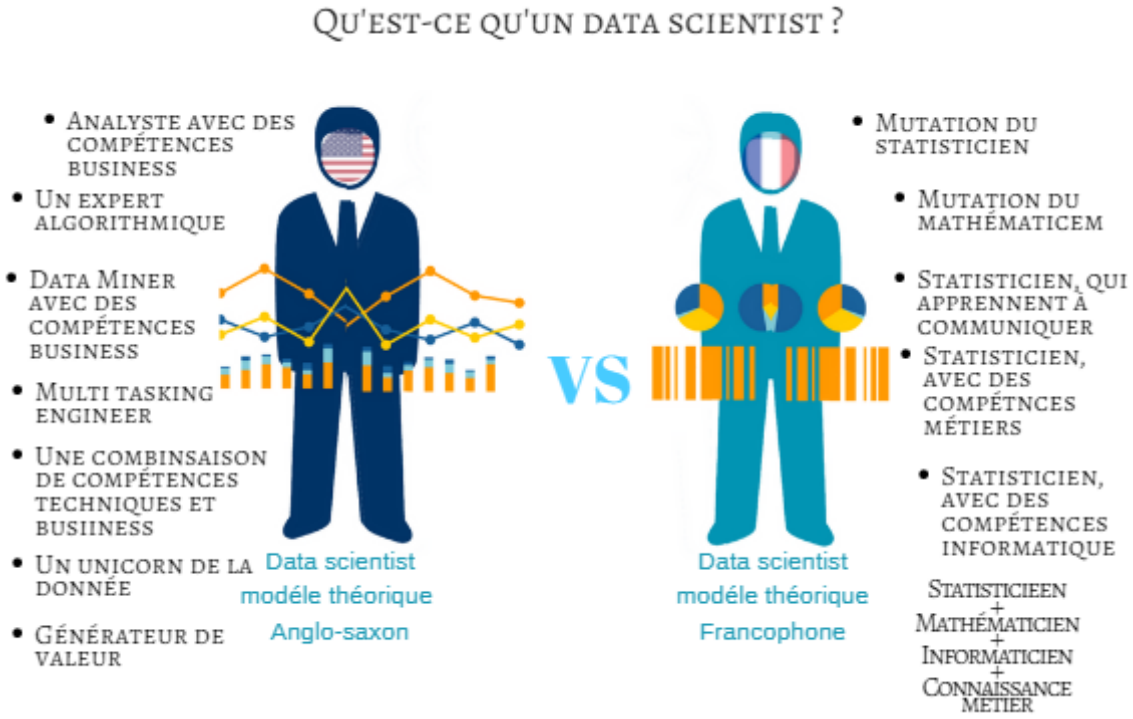
2.1. Apport théorique :

Les apports théoriques sont conduits selon deux grands modèles théoriques : le data scientist selon le modèle théorique francophone et le data scientist selon le modèle théorique anglo-Saxon.

2.1.1. Apport selon les modèles francophones et anglo-saxons :

La littérature parcourue traitant le sujet de data science selon le modèle français et anglo-saxon avait deux visions quelque peu nuancées du métier, notre approche normative nous a amenées à recueillir 171 articles différents traitant le métier de data scientist (les détails de notre approche et de la méthodologie de recueil des articles sont cités dans le 2e et 3e chapitre), dans les 171 articles recueillis, 12 d'entre eux étaient de la littérature issue de pays francophones.
La figue ci-dessous illustre la nuance existante entre un data scientist décrit dans les théories francophones et celui décrit dans les théories anglo-saxons.

Figure 2 : Qu'est-ce qu'un data scientist ? Modèle Francophone VS modèle Anglo-saxon



Source : Schéma réalisé par nos soins, en utilisant l'outil graphique Canva

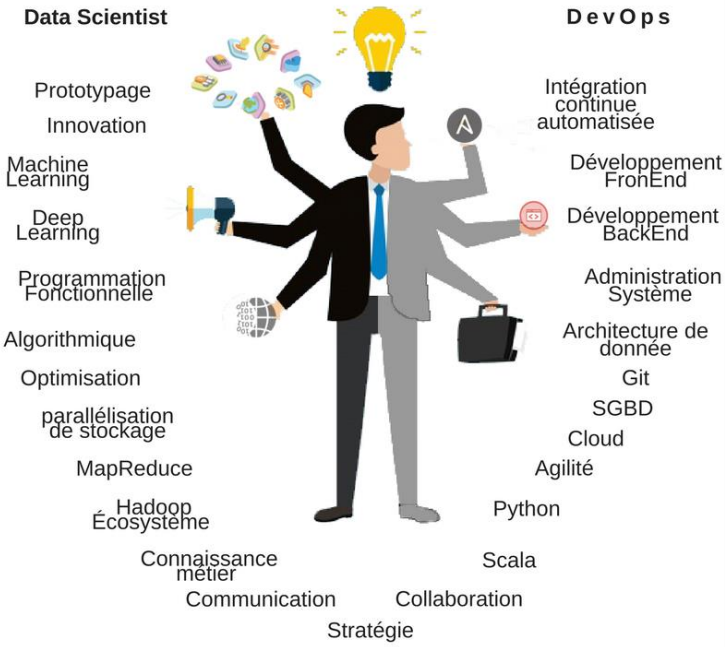
Comme le montre l'illustration, les deux modèles sont bien nuancés, les francophones affirment que la data science est une mutation du statisticien/mathématicien confronté à une technologie nouvelle et à un nouvel environnement de données. Dans leurs papiers Big Data Analytics- Retour vers le Futur- les auteurs affirment que : ''Le statisticien, prend des formes successives d'un data scientist au gré des développements brutaux des outils informatiques (Philippe Besse, 2014, p. 2)'' d'autres chercheurs francophones rejoignent l'idée, tels que Philippe Besse[1], Béatrice Laurent[2] ''le data scientist devient un enrobage publicitaire (packaging) pour masquer des approches statistiques limitées, ou élaborées, exécutées dans des environnements technologiques sophistiqués'' (Philippe Besse, 2014, p. 84). En comparaison avec le modèle Anglo-saxon qui affirme que le data scientist est une révolution d'un nouveau métier, à la croisée de plusieurs compétences, analytiques, mathématiques, informatique, business et communication , né d'un besoin imminent en la compréhension des besoins métiers, pour les transcrire et les modéliser en solutions technologiques. Dans ce sens, Mariano Martínez, affirme dans son papier « How to Think Like a Data Scientist » ''*Data Scientist is the term for a new type of analysts possessing both the business knowledge to understand the problems companies are faced with as well as the technical skills to generate decision-relevant information from large-scale datasets* (G.Illescas et al., 2016, p. 154)''

Dans notre contexte, les data scientists observés, se rapprochent des deux modèles suivants, Cependant, ils sont issus de formations différentes et ont tous été modelés pour rentrer dans le moule de data scientist, répondant aux besoins et aux spécificités de l'entreprise et au

[1] Chercheur Université de Toulouse, INSA : Institut de mathématiques de Toulouse
[2] Chercheur Université de Toulouse, INSA : Institut de mathématiques de Toulouse

contexte dans lequel ils opèrent et évoluent, qui est le contexte algérien. Le data scientist que nous traitons, fera l'objet d'un nouveau modèle. C'est ainsi, que nous contribuerons par notre recherche, à enrichir la littérature. La figure ci-dessous illustre le data scientist selon le contexte algérien.

Figure 3 : data scientist contexte algérien[1]



Source : Schéma réalisé par nos soins en utilisant l'outil Canva[2]

2.1.2. Apport selon la complexité et la nouveauté du sujet :

L'intérêt pour le métier de data scientist a pris de l'ampleur entre les années 2010 et 2012, aux États Unis, bien plus tardivement en France et plus récemment en Algérie. Compte tenu de la nouveauté du métier, plusieurs entités organisationnelles ignorent l'existence d'un tel métier, mais ils ont en tout de même, subi les conséquences.

Aussi, traiter et étudier un métier issu d'un environnement technologique et informatique, qui nous y méconnu est quelque peu complexe. Celui-ci nécessite une compréhension approfondie du contexte technologique de mise en œuvre et des paradigmes qui en découlent, notamment les choix et méthodologie d'algorithme, les langages de programmation, les modèles mathématiques et statistiques ainsi que l'architecture de la parallélisassion des stockages et calculs et finalement, les rôles, les compétences et les facteurs clés du métier.

Un coup d'œil des travaux antérieurs nous a permis de mesurer la pertinence et l'apport de notre recherche à la littérature traitant les rôles et les compétences des métiers du *BIG DATA*. A notre connaissance, peu d'auteurs s'étaient penchés sur le rôle du data scientist, la petite enquête qui a été menée au niveau des bibliothèques des écoles du pôle universitaire (ESC[3],

[1] La discussion du modèle est détaillée dans le chapitre IV
[2] Les technologies citées sur cette illustration sont définies dans le glossaire ajouté en annexe
[3] Ecole Supérieur de Commerce

EHEC[1], ENSSEA[2]) a permis de révéler qu' aucun des mémoires ne portait sur le thème de data science ou de data scientist ou portait un intérêt aux métiers analytiques des *BIG DATA*. Une autre recherche menée auprès des bibliothèques de L'ESI[3] et de l'ÉNPA[4], a révélé aussi l'existence de projets de fin d'études traitant dans les BIG DATA, utilisant des modèles d'apprentissage statistique ou automatique, mais sans expliciter et mettre en avant les rôles et les compétences d'un data scientist, ou d'essayer de comprendre et d'identifier les composantes du métier.

Notre mémoire sera donc le premier à aborder ce sujet et à contribuer à la construction d'un modèle pour définir le data scientist dans le contexte algérien, face à des besoins spécifiques.

Notre travail permettra de mettre en évidence plus en détail le métier de data scientist et espérant ainsi enrichir la littérature y afférente. En outre, la démarche méthodologique mise en œuvre pourra être reproduite dans le cadre d'autres recherches afin de mettre à l'épreuve nos propositions.

2.2. Apport managérial :

Aujourd'hui, la nouveauté du métier fait qu'il est difficile de savoir qu'est-ce qu'un data scientist, que font-ils, que délivrent-ils, Dj Patil et Thomas H.Devenport publient dans le *Harvard Business Review ''Juste en parcourant les offres d'emploi pour les data scientists, on comprend que les employeurs ne savent pas vraiment ce qu'ils cherchent exactement, et c'est probablement pourquoi tout le monde se plaignait de la pénurie de data scientist sur le marché du travail de nos jours''* (Davenport T. H., 2012, p. 75)[5]. Suite à cette ambiguïté, et indéfinité du métier de *data scientist*, notre recherche contribuera à une meilleure définition et compréhension du métier, tant d'un point de vue macro que micro.

D'un point de vue macro, notre recherche tentera de satisfaire la curiosité de ceux intéressés par ce domaine de répondre aux interrogations de nombreuses personnes voulant se lancer dans des projets big data et se demandant quels seraient les profils appropriés, leurs besoins, leurs formations ainsi que les enjeux liés à ce métier. Cela les aiderait également à cerner et cibler leurs besoins en ressources humaines.

D'un point de vue micro, en donnant une vision claire et précise du rôle du data scientist, nous contribuerons au sein de l'entreprise BIG MAMA à la réalisation d'un plan de recrutement et d'une stratégie RH[6].

2.2.1. Contribution au plan de recrutement et stratégie RH :

Comment trouver la personne qui sera amenée à être *data scientist* ?

Dans le contexte actuel (Algérie), peu de formations proposent un contenu orienté *data science*[7] et big data, pour être directement opérationnel à la sortie de l'université. Cependant,

[1] Ecole des Hautes Etudes Commerciales
[2] Ecole Nationale Supérieure de Statistique et d'Economie Appliquée
[3] Ecole Supérieur d'Informatique
[4] Ecole Nationale Polytechnique d'Alger
[5] Traduite de l'anglais par nos soins
[6] Ressource Humaine
[7] Science de donnée, à vérifier définition dans le glossaire en annexe.

la personne qui sera amenée à être data scientist, doit pouvoir se munir des compétences clés (*need to have*), pour pouvoir intégrer une équipe déjà formée et opérationnelle.
Par notre analyse, nous mettons au point un plan de recrutement qui définit clairement les facteurs clés (*need to have*) à avoir pour intégrer l'équipe de *data scientist*, accompagnés d'une grille d'évolution permettant de déterminer les facteurs clé de succès (*nice to have*), pour évoluer rapidement et se hisser au niveau exigé par l'entreprise.

De ce fait, l'apport est double, il consiste d'abord à identifier ce que cherche l'entreprise comme profil de manière assez claire et le verbaliser. Ainsi, le processus devient plus clair en termes de recrutement. Par la suite, le processus permettrait d'identifier les cibles à atteindre, pour avoir un data scientist productif et opérationnel. Si le point de départ est ainsi déterminé, le point d'arrivée serait clair, ce qui permettra, par la suite, d'enclencher la réflexion sur la partie formation.

3. Question de la recherche :

Dans cette phase, il s'agit de définir une problématique de recherche précise, c'est-à-dire un aspect particulier du champ d'étude. Notre recherche identifie les facteurs clés de succès du métier de data scientist et recueille les compétences nécessaires à ce métier et son évolution au fil du temps et des projets, à partir d'une analyse exploratoire. Nous repérons les interactions vécues en lien avec les projets et l'environnement de travail. L'analyse de l'évolution des projets et compétences des data scientists au cours de la période étudiée, nous permet d'établir des grilles et des modèles que nous pourrons comparer avec notre matériau théorique. Pour cela, nous désirions répondre à la problématique suivante.

- **Quels sont les facteurs clés de succès déterminant le métier de data scientist ?**

Pour répondre à cette problématique, nous posons un ensemble de sous-questions conductrices :

- Qu'est-ce qu'un data scientist ?
- Que faut-il pour réussir dans le métier ?
- Quel est son environnement technologique de travail ?
- Quelles sont les difficultés et les problèmes adressés par ce métier ?

4. Contexte organisationnel :

Big Mama Technology est une jeune start-up créée en 2014, sous la forme juridique SARL[1], créée par deux associés. Constituée de 09 employés. Le statut lui procure de nombreux avantages quant à la limitation des responsabilités.

4.1. Big mama Technology : domaines d'expertise, valeur ajoutée et différenciation :

Dans cette partie, nous présenterons brièvement l'entreprise, à travers son nom, son domaine d'expertise et ses spécificités

[1] Société A Responsabilité Limitée.

Big Mama, histoire d'un nom :
La société Big Mama tire son nom et ses valeurs de 4 sources :

- L'objectif d'une marque est de "marquer" les esprits. Les mots "Big Mama" interpellent. Ils utilisent l'humour pour entrer dans l'esprit des gens (Big mama Technology, 2016, p. 2).
- Le choix de cette marque permet de créer un lien (en utilisant la référence à la maternité) Mais aussi pour montrer leur capacité à prendre de la perspective et faire preuve d'humilité (Big mama Technology, 2016, p. 2).
- La ressemblance phonétique entre Big Mama et Big Data (secteur dans lequel la société évolue).
- L'allusion - par contradiction sémantique - à Big Brother et au livre d'Orwell "1984", dans lequel la surveillance et le voyeurisme sont dénoncés.

Effectivement, nous constatons que la connotation du nom reste quelque peu humoristique, tel qu'il ressort d'une enquête menée auprès d'une catégorie de personnes (amis et familles) pour évaluer que peut évoquer la marque « BIG mama ». En effet, très peu associent Big mama au secteur des Big data. Le nom « Big mama » est associé plutôt à un film de comédie américaine, où à de la cuisine italienne,
Il est vrai que ce choix de nom est risqué, mais il est aussi osé et suscite l'intérêt, il est même risqué vis-à-vis des clients qui seraient susceptibles de ne pas les prendre au sérieux du premier coup et aussi vis-à-vis du secteur qui est un secteur très sensible. Le nom est très intéressant, quand bien même il demeure peu commun, il est rare que des entreprises portent un nom semblable. En effet, ce nom attire l'attention et marque les esprits et cela pousse à s'intéresser à l'identité de l'entreprise, à son activité et à son histoire.
D'un point de vue marketing, la marque est le reflet de l'entreprise, tel un miroir. Pour cela, une bonne marque est censée être facilement mémorisable, euphonique, facilement prononçable, significative et évocatrice (Béatrice Durand-Mégret, 2012, p. 62),
Nous avons mené une enquête à l'aide d'un questionnaire mis en ligne, les réponses et les mots qui revenaient le plus souvent sont présentés dans le tableau ci-dessous :

Tableau 1 : Qualité commerciale de la marque Big mama technology.

Les qualités commerciales de la marque ''Big mama Technology''	
Facilement mémorisable	Des syllabes faciles à retenir
Euphonique	La marque Big mama technology est agréable à entendre, c'est une succession harmonieuse de consonnes et voyelles, et cette appellation anglo-saxonne lui donne un côté branché
Facilement prononçable	En français, comme en Anglais, il n'y a aucune difficulté de prononciation.
Significative	Big Mama Technology : évoque la maternité, le soin, les big data, la technologie,
Evocatrice	Big mama Technology : Informatique, protection, sécurité, film comédie américaine, gastronomie, restauration.

Les réponses recueillies des internautes, confirment que le nom de la marque Big mama Technology a une forte liaison avec le film de comédie américaine portant le même nom et renvoi également à la restauration et à la gastronomie. Cependant, en prenant un peu plus de temps en lisant le nom complet « Big mama Technology » la signification renvoie à la sécurité, au soin, à la protection informatique et technologique.

4.2. Big mama Technology, responsabilité éthique :

"Créer de la valeur sans perdre ses valeurs" (Kofman, 2009) Big mama s'inspire de cette ligne de conduite pour ses projets, à l'heure où les nouvelles technologies, en particulier celles qui touchent à la sécurité des individus et aux données des personnes (morales ou physiques), menacent de plus en plus la vie privée des gens.
A terme, à travers sa stratégies, BIG mama, véhicule une perspective qui permet de protéger les personnes et de proposer une vision nouvelle des data.

''Dans un monde où nous serons capables de prédire (au sens d'anticiper), quelle personne sera susceptible d'acheter tel et tel produit, quel employé quittera son poste demain, qui va se marier avec qui, etc… Quand les machines ne seront plus seulement capables de battre les Hommes aux échecs ou rédiger des articles de journaux, ou répondre à des mails, mais dans tous les aspects de la vie qui auront une dimension stratégique, il faudra des gens qui soient suffisamment armés en matière d'éthique pour ne pas basculer vers un chaos technologique'' (Big mama Technology, 2016, p. 7) à travers ce passage cité dans le business plan, l'on comprend clairement que l'entreprise Big Mama a été imaginée dans le souci de démontrer que l'Intelligence Artificielle pouvait être un outil au profit de l'amélioration humaine, et non l'inverse.
Dans ce sens ;

- L'entreprise refuse des projets jugés non conformes à ses perceptions personnelles (et probablement subjective) de l'éthique.
- L'entreprise a conçu un logiciel algorithmique de formation à distance à la Data Science[1].

4.3. Big mama technology, structure organisationnelle et management :

La structure organisationnelle de BIG mama, est une structure plate, le management de l'entreprise est piloté dans cette perspective de structure aplatie. Ce choix de structure est certes, une conséquence du nombre du personnel, mais aussi un choix, dans une vision de création de valeur et de construction d'environnement favorable à l'interaction et à la collaboration des employés. Ce modèle d'interaction a été créé dans le but de faire évoluer l'équipe et que chacun puisse apprendre sur les technologies de l'autre, se développer, et être le plus polyvalent possible. La vraie richesse est le capital humain pour permettre aux employés d'exploiter et de révéler au mieux leurs richesses humaines. Pour ce faire, l'information doit absolument circuler de manière optimale. La mécanique de

[1] L'objectif du logiciel n'est pas de former des Data Scientist, mais de donner des bases solides à des étudiants, et une opportunité de pouvoir se développer par la suite sur ces métiers.et leurs inculqué une culture d'éthique et de respect à la vie privée confronté à la manipulation et traitement de la donnée.

fonctionnement globale, la synergie, la confiance, la liberté et la motivation doivent favoriser les conditions d'éclosion de cette valeur.

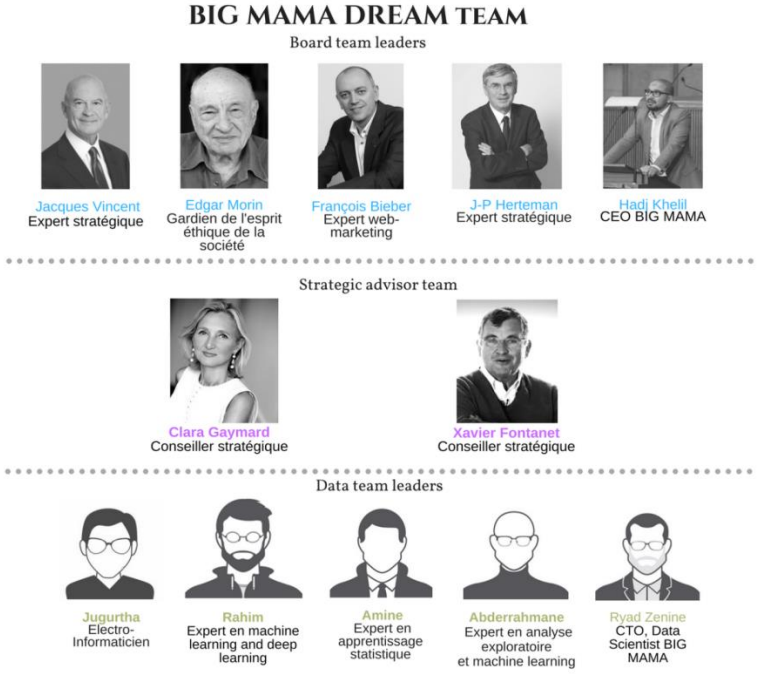
Cependant, en adoptant cette structure, l'entreprise se retrouve à piloter en micro management. C'est-à-dire contrôler et observer étroitement le travail des employés. Cela amène à donner une grande importance aux détails, et avoir une grande exigence dans la qualité de travail.

4.4. Equipe Big mama :

L'équipe Big Mama est composée de deux hémisphères complémentaires :

- **L'équipe dirigeante :** qui définit le cap stratégique (pour BIG MAMA mais parfois aussi pour leurs clients), et qui gère la phase cruciale de traduction algorithmique de la stratégie de création de valeur.
- **Les conseillers stratégiques :** qui discutent des décisions stratégiques, et la faisabilité des projets.
- **Les Data-Scientists** : chargés de créer les algorithmes et les logiciels pour la résolution des problématiques client, ayant plusieurs compétences dans leur domaine de prédilection (algorithmes, informatique, intelligence artificielle).

Figure 04 : L'équipe BIG MAMA



Source : Étude réalisée par nos soins, en utilisant l'outil Canva

La figure ci-dessus illustre la structure, et l'organisation des équipes au sein de l'entreprise, comme nous l'avons expliqué plus haut, la structure est organisée en équipes : (équipe dirigeante, équipe stratégiste, équipe data scientist) pilotées par un micro management et chapoté et orienté par les stratégistes et conseillés.

4.5. Modèle Economique :

BIG MAMA est développeur de solutions logicielles algorithmiques exclusivement en B2B, utilisant tout le savoir-faire autour de la machine learning, du deep learning et de l'apprentissage statistique. Les outils qui utilisent ingèrent, traitent les données et restituent des informations pour les clients.

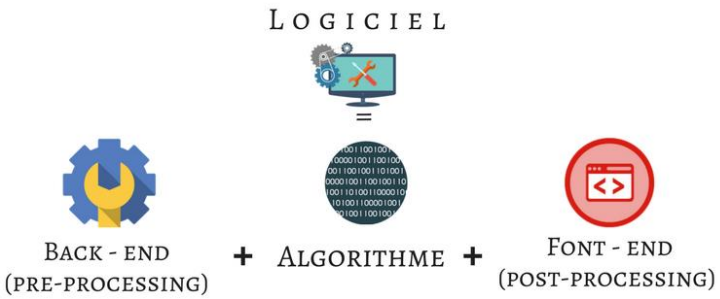
Le choix du modèle B2B est justifié comme étant un positionnement imposé par le secteur, et justifié par les moyens. En d'autres termes en rentrant dans un domaine, il faut faire le choix d'être un artisan ou un industriel, quand l'entrepreneur n'a pas les moyens d'être industriel, il devient forcément artisan, ce choix est argumenté par le fondateur de l'entreprise comme suit : ''*Big mama est un tailleur talentueux qui fabrique des costumes sur mesure pour les clients''* (Khelil, 2017)

Cependant, ce modèle de prestation sur mesure, présente des limites. Dans de telles perspectives, il est difficile de stabiliser le modèle économique, ce dernier ayant une croissance organique, poussé par les attentes du client et influencée par le terrain et l'expérience.

Historiquement, l'entreprise BIG mama est un laboratoire de développement d'algorithmes. Malheureusement, les algorithmes, aussi performants soient-ils, ne sont que très rarement monétisables en l'état. Ce qui intéresse le marché, ce sont des outils qui soient les plus clairs possibles, ergonomiques et opérationnellement utilisables. Fort de ce constat, durant ses dernières années d'activité, BIG MAMA s'est construite une compétence interne en développement de logiciels (web et mobile).

Cette nouvelle compétence interne, leur permet aujourd'hui de construire un catalogue de logiciels "customisables" et "prêts à la vente" (voir descriptif ci-dessous). Pour finir, BIG MAMA possède une approche singulière en ce qui concerne son processus de "Recherche et Développement" interne. En effet, elle produit seule des logiciels lorsqu'une opportunité est identifiée. Sinon, le logiciel est conçu avec un client/partenaire qui contribue par ses données et son *input* métier dans les phases de développement et qui prend ensuite en charge une partie du développement commercial.
L'illustration ci-dessous décrit la composition de la solution produite par BIG MAMA.

Figure 05 : Cœur de métier BIG MAMA



Source : Étude réalisée par nos soins, en utilisant l'outil graphique Canva

4.6. Les spécificités de Big mama Technology:

Une recherche et développement pilotées par la stratégie:
Une fois que la société a réalisé sa capacité à développer des logiciels, il a fallu comprendre lesquels. Surtout, il a fallu mettre au point un outil d'identification systématique. Une méthodologie leur a permis d'identifier quels logiciels doivent-ils développer.

BIG MAMA a fait le choix de faire piloter ses capacités de R & D par le département stratégie.

03 raisons justifient ce choix :

- La stratégie est une activité rentable en soi qui permet de fabriquer de la valeur pour BM du fait de la facturation des prestations de conseil.
- La stratégie lui permet d'avoir en interne un outil très précis de cartographie du marché planétaire du big data. Cette connaissance (qui est un enjeu majeur), permet d'anticiper les synergies les plus rémunératrices et d'en profiter pour développer des outils ciblés.
- L'articulation [Développements algorithmiques - Stratégie] est une combinaison différenciante, qui consiste à faire précéder tout développement par une analyse stratégique qui viendra formaliser le besoin et traduire dans une problématique algorithmique un sujet qui serait jugé pertinent.

Cette tâche de traduction (d'un sujet business en une problématique algorithmique) est précisément l'une des valeurs fondamentales sur laquelle va s'appuyer Big Mama pour générer la plus-value chez le partenaire ou l'utilisateur de ses logiciels.
Elle permettra d'aller plus loin dans l'approche et de tirer la quintessence des chiffres et des données de l'entreprise. En effet, un mauvais diagnostic ou réalisé de manière incomplète, peut baisser le rendement de l'utilisation des *BIG DATA*.

4.7. Business Map (carte d'affaire) :

La société Big Mama Technology est composée de deux entités indépendantes, mais liées d'un point de vue stratégique. Il existe une relation basée sur le partage de valeurs en fonction de l'origine du projet .Cela dit, une relative distinction existe au niveau des métiers de chacun

Tableau 2: business map Big mama

Provenance du projet / Partage de valeur	Big mama France	Big mama Algérie
Projet émane de Paris	66%	33%
le projet émane d'Alger	33%	66%
Projet Conjoitement apporté	50%	50%

Source : Réalisé par nos soins

Big Mama Algérie est :
L'interlocuteur pour les clients MENA et Asie, fournisseur de ressources humaines d'appoint pour Big Mama France
Big Mama France : est dédiée à la stratégie, l'interlocuteur pour les clients du reste du monde

Ci-dessous l'illustration venant résumer le mode organisationnel de la structure.

Figure 04: Business Map BIG MAMA

Source: Business plan BIG MAMA

La réunion de ces deux entités parfaitement indépendantes permet une synergie qui permet à chacune de régler des problématiques business fondamentales :

Pour Big Mama France, l'embauche durable de data-scientiste est problématique, alors que Big Mama Algérie peut fournir la ressource dont elle dispose pour l'instant. En revanche, Big Mama Algérie ne dispose pas de profils pouvant apporter la compétence stratégique nécessaire à son développement.

Une fois ce double besoin satisfait, chaque entité s'occupe de la partie du globe qui lui convient le mieux (exemple : les clients algériens exigent que leurs données soient traitées en Algérie et ainsi de suite pour les autres clients).

CHAPITRE II : REVUE DE LITTÉRATURE ET CADRE CONCEPTUEL

1. Revue de littérature et cadre conceptuel :

Dans cette partie, nous allons construire les points de repères théoriques nécessaires à la description du domaine du BIG DATA, d'où découle la *data science* et qui fait le métier de *data scientist.*

Dans une première partie, nous présentons l'apport théorique sur les *BIG DATA* concernant l'historique de l'émergence de ce dernier. Après une introduction générale du concept, nous exposerons les différents travaux et recherches scientifiques orientés vers ce thème. Dans la seconde partie, nous avons regroupé différents modèles théoriques et travaux scientifiques servant à mettre en évidence les changements qui ont amenés à une nouvelle science nommée *data science* (science des données).
Pour finir, et en suivant la même méthodologie de collectes d'informations, nous mettons le point sur les travaux, articles, publications identifiant et caractérisant un *data scientist.*

1.1. La Revolution *BIG DATA:*

"Ultimately, big data is more about attitude than tools; data-driven organization look at big data as a solution, not a problem. (Roger Magoulas and Ben Lorica, 2009)"

Nous sommes tous, aujourd'hui conscient de la place qu'a pris internet, et de la façon dont cela a transformé le fonctionnement des entreprises, mais ce que la plupart ignore, c'est qu'il y a une nouvelle tendance technologique, moins visible et tout aussi transformatrice, et dont l'impact est aussi énorme que son nom, les *BIG DATA.*
D'où vient ce phénomène ? Que recouvre ce *BuzzWord*[1] *?*
Afin de répondre au mieux à ces questions, une recherche dans les revues théoriques abordant les *BIG DATA* a été menée.

1.2. Envol historique BIG DATA :

"Big Data: when the size and performance requirements for data management become significant design and decision factors for implementing a data management and analysis system. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration" (O'Reilly Media, 2009).

*"Le Big Data (*la traduction française préfère les termes de) « Données massives », ou « Méga données » est l'explosion de cette production des données numériques hétérogènes (multi structures, multimodales, multi-versions) en volumes tellement importants qu'il devient impossible de traiter avec les outils traditionnels de gestion de bases de données" (Lebbah, *2016)*.

[1] expression à la mode

Nous l'aurons compris à travers ces deux définitions que les BIG DATA, est l'émergence de données volumineuses, ayant des caractéristiques spécifiques, nécessitant un traitement et une technologie dédié. Cependant, la littérature parcourue ne présente pas une définition claire et officielle du terme *BIG DATA*, et il est difficile de le décrire en quelques mots, pour bien comprendre les *BIG DATA*, nous avons parcourus l'histoire de l'émergence de ce phénomène.

Le terme *BIG DATA* fut utilisé pour la première fois en 1997 par Michael Cox et David Ellsworth dans l'article « *Application controlled demand paging for out-of-core visualization*» publié lors de la 8ème conférence de l'IEEE[1] sur la visualisation, (Press, 2013). Néanmoins, les pratiques du BIG DATA sont apparues bien avant, de nombreux secteurs comme l'astronomie, la recherche sur le génome et d'autres faisaient du *BIG DATA* sans le savoir depuis bien longtemps.
Le tableau ci-dessous illustre un bref aperçu de la longue histoire de la pensée et des innovations qui nous ont amenés à l'aube de la décennie des *BIG DATA*.

Tableau 3 : Historique de l'émergence des big data

Année	Histoire
1865	**1ere utilisation du terme *Business Intelligence*** Le terme « business intelligence » est utilisé par Richard Millar Devens dans son encyclopédie des anecdotes business et commerciales, décrivant comment le banquier *Henry Furnese* a obtenu un avantage sur les concurrents en recueillant et en analysant de façon structurée les informations pertinentes de ses activités commerciales. Ceci est censé être la première étude d'une entreprise mettant l'analyse des données à des fins commerciales (Marr, 2015).
1881	**La 1ère machine de calcul** Le Bureau du recensement des États-Unis rencontre le plus grand problème de l'histoire, il estime qu'il faudrait près de 10 ans pour analyser et traiter toutes les données recueillies dans le recensement de 1880. En 1881, un jeune ingénieur employé par le bureau - Herman Hollerith - produit ce qui deviendra connu sous le nom de *Hollerith Tabulating Machine*. À l'aide de cartes perforées, il réduit 10 ans de travail à trois mois et atteint sa place dans l'histoire en tant que père du calcul automatisé moderne. Le bureau est connu aujourd'hui sous le nom d'IBM (Mines, 2011).
1928	**L'apparition de stockage de données moderne** Fritz Pfleumer, un ingénieur germano-autrichien, invente une méthode de stockage d'informations sur bande magnétique. Les principes qu'il développe sont toujours utilisés aujourd'hui. La grande majorité des données numériques sont stockées magnétiquement sur des disques durs informatiques (Wikipedia, 2016).
1944	Fremont Rider, bibliothécaire de l'Université Wesleyen, aux États-Unis, a publié un article intitulé *The Scholar and the Future of the Research Library*. Dans l'une des premières tentatives de quantification d'information produite, il observe que pour stocker toutes les œuvres académiques et populaires de valeur produites, les bibliothèques américaines devraient doubler leurs capacités tous les

[1] Institute of Electrical and Electronics Engineers

16 ans. Cela l'a amené à prédire que la bibliothèque de Yale, d'ici 2040, contiendra 200 millions de livres répartis sur (6 000) six mille étagères (Rider, 1944).

1962 Les premières étapes sont prises en matière de reconnaissance de la parole, lorsque l'ingénieur IBM William C.Dersch présente la Shoebox Machine à l'Exposition universelle de 1962. Il peut interpréter des chiffres et seize mots parlés en anglais dans des informations numériques (IBM, 2009).

1970 Le mathématicien d'IBM Edgar F Codd présente son *framework* « base de données relationnelles ». Le modèle fournit le *framework* que de nombreux services de données modernes utilisent aujourd'hui, pour stocker des informations sous un format hiérarchique, qui peut être consulté et manipulé par quiconque. Avant, pour avoir accès aux données des banques de mémoire d'un ordinateur on faisait appel à un expert (Bruchez, 2015).

En outre, depuis la fin des années cinquante la « business intelligence » était déjà un concept populaire qui voit une popularité croissante avec les nouveaux logiciels et systèmes émergents pour analyser les performances commerciales et opérationnelles.

1991 **L'émergence d'internet**

L'informaticien Tim Berners-Lee a annoncé la naissance de ce qui deviendrait Internet tel que nous le connaissons aujourd'hui. Dans une publication dans le groupe Usenet alt.hypertext (un groupe destiné aux amateurs d'hypertextes), il définit les spécifications d'un réseau de données mondialement interconnecté accessible n'importe où, par n'importe qui (Wikipedia, The birth of the World Wide Web, 1991).

1997 Michael Lesk publie son article « Combien d'information existe-t-elle dans le monde » ? Théorisant que l'existence de 12 000 petabytes n'est peut-être pas une supposition déraisonnable. Il souligne également que même à ce stade précoce de son développement, le web augmente de 10 fois par an. Une grande partie de ces données, souligne-t-il, ne seront jamais vues par personne et ne donneront donc aucune idée (Lesk, 1997).

Google Search débute également cette année - et pour les 20 prochaines années (au moins), son nom deviendra un raccourci pour la recherche de données sur Internet (Marr, 2015).

2000 *In how much information* ? Peter Lyman et Hal Varian (maintenant économistes en chef chez Google) ont tenté de mesurer pour la première fois la quantité d'informations numériques dans le monde et son taux de croissance. Ils ont conclu que : "La production annuelle totale des imprimés, de contenu cinématographique, optique et magnétique nécessiterai environ 1,5 milliard de gigabytes de stockage. C'est l'équivalent de 250 mégaoctets par personne pour chaque homme, femme et enfant sur Terre " (Peter Layman, Hal Varian, 2000).

2001 Doug Laney, analyste chez Gartner, définit, pour la premiére fois, dans son article, « *3D Data Management : Controlling Data Volume, Velocity and*

Variety », trois paramètres qui deviendront les caractéristiques de Big Data : volume de données, vitesse et variété (Laney, 2001)

Cette année, on voit également la première utilisation du terme « Saas[1] » dans l'article « Strategic Backgrounder: Software as a Service » publié par *the Software and Information Industry Association* en collaboration avec *Tripletree association*, qui est un concept fondamental pour de nombreuses applications basées sur le cloud (Kevin Green, 2001).

2005 **L'apparition du Web 2.0 (l'explosion des volumes de données)**
le Web 2.0 est un web généré par les utilisateurs où la majorité du contenu sera fournie par les utilisateurs des services plutôt que les fournisseurs de services eux-mêmes. Tout ceci grâce à l'intégration de pages Web traditionnelles de style HTML avec de vastes bases de données back-end en SQL[2]. Une année après son apparition Facebook compte déjà 5,5 millions d'utilisateurs (Marr, 2015),

2006 Cette année, on voit également la création de *Hadoop* par Doug Cutting- le *framework* open source créé spécifiquement pour le stockage et l'analyse des ensembles *Big Data*. Sa flexibilité et facilité pour la création d'applications distribuées le rend particulièrement utile pour gérer les données non structurées (voix, vidéo, texte brut, etc.) que nous générons et collectons de plus en plus (Bonaci, 2015).

2008 **Le *BIG DATA au quotidien***
Wired le fameux site web publie le concept de *Big Data* aux masses avec leur article « *The End of Theor y: The Data Deluge Makes the Scientific Model Obsolete* » La fin de la théorie: le déluge de données rend le modèle scientifique obsolète, l'article publié met en avant l'importance qu'a pris le *BIG DATA*, et les modèles mathématiques, et comment ses derniers ont rendu les sciences obsolètes (ANDERSON, 2008).

2010 Selon le rapport « *How Much Information* » publié par « Enterprise Server Information » Les serveurs du monde entier traitent 9,57 zettabytes (9,57 trillion de gigaoctets) d'informations, équivalent à 12 gigaoctets d'informations par personne, par jour), on estime que 14,7 exabytes de nouvelles informations sont produites cette année-là (James E. Short, Roger E. Bohn, Chaitanya Baru, 2010).

2011 L'entreprise américaine moyenne avec plus de 1 000 employés enregistre plus de 200 téraoctets de données selon le rapport Big Data : The Next Frontier for innovation, competition and productivity par McKinsey Global Institute (James Manyika and al., 2011)

2014 Cette année est marqué par la montée des utilisateurs mobiles. D'après un rapport de « THE RADICATI GROUP, Inc » le nombre a atteint plus de 5,6 milliards, de plus en plus de personnes utilisent des appareils mobiles pour accéder aux données numériques, à l'instar des ordinateurs de bureau ou domestiques (Radicati, 2014).

[1] Software as a service
[2] Structured Query Language

2016	**Une révolution portée par l'Internet des objets** Un rapport du Business Insider, compte 6.6 milliard d'objets connectés, la combinaison de la sécurité, du mobile, du cloud, des *BIG DATA*, ainsi que les développements technologiques portant sur l'analyse avancée des données, sont aujourd'hui les principaux moteurs de l'IOT (Newman, 2017)[1].

Source : réalisé par nos soins

Ce voyage dans l'histoire, nous familiarise d'avantage avec le phénomène *BIG DATA*, et nous constatons ainsi que les BIG DATA sont la conséquence des avancées technologiques et de l'explosion des données, mais pas que, comme le mentionne Davenport, les BIG DATA sont définies non seulement par la taille mais aussi par leur manque de structure: "Les BIG DATA se réfèrent à des données trop importantes pour s'adapter à un seul serveur , trop déstructuré pour s'intégrer dans une base de données de lignes et colonnes, trop circulaire et instable pour entrer dans un entrepôt de données statiques. Bien que sa taille reçoive toute l'attention, l'aspect le plus difficile des grandes données implique vraiment son manque de *structure* (Davenport, 2014)". Ainsi notre recherche, porte sur l'émergence des compétences et du métier *data scientist* opérant dans cet environnement des *BIG DATA*.

1.3. Envol Historique data science

Si le mot *BIG DATA* est le nouveau BuzzWord du 21ème siécle, le mot *data science* (science des données) est apparu bien avant, la science des données est aussi vielle que les statistiques et les mathématiques. Ce sujet a déjà fait l'objet de nombreuses études, et enquête universitaire.

Cela remonte au 17ème siècle, Graunt le père des statistiques dans son ouvrage « *Natural and Political observations, made upon the Bills of Mortality* », a été le 1er à mettre au point un système de prévision sur des bases statistiques en 1663 (Wikipedia, John Graunt, 2007)

Cependant, les avancées technologiques, et la montée du volume d'informations (*BIG DATA*), ont donnés un nouveau concept, une nouvelle utilisation et de nouvelles conditions à la *data science.*

- Nouveau concept par apport à la structure de donnée : quel est sa source, quel est sa granularité, son type, et sa nature (donnée statique, donnée et temps réel…etc) ;
- Nouveau contexte : comment la donnée a été générée ;
- Nouvelle condition : état de la donnée, (peut-on utiliser la donnée tout de suite, ou doit-on d'abord la nettoyer… etc).

Tous ces changements, et mutations qu'a connu la data science ont fait l'objet de préoccupations et recherche scientifique, nous les présentons dans le tableau ci-dessous.

[1] Internet Of Things

Tableau 4 : Envol historique data science

Année	Article traitant le sujet data science
1962	John W. Tukey écrit "The Future of Data Analysis", où il met en avant l'évolution des modèles statistiques et mathématiques vers l'analytique et de la place qu'a prise l'analyse des données dans les statistiques, lui donnant un caractère scientifique.
1977	L'Association internationale pour l'informatique statistique (IASC[1]), constituée en tant que section de l'ISI[2] prend conscience de l'importance de la création de valeurs via les modèles statistiques et mathématique et souligne l'urgence d'intégrer la data science au cœur de l'institut. "*C'est la mission de l'IASC de lier la méthodologie statistique traditionnelle, la technologie informatique moderne et la connaissance des experts du domaine afin de convertir les données en informations et connaissances*".
1989	Gregory Piatetsky-Shapiro organise et préside le premier atelier sur la découverte des connaissances dans les bases de données (KDD[3]). En 1995, cet atelier fait l'objet d'une conférence annuelle ACM[4] SIGKDD[5] sur la découverte de connaissance et l'exploration de données (KDD) (KDnuggets, 2017).
1996	Pour la première fois, l'importance de la «*data science* » comme domaine d'études a été officiellement reconnue et incluse dans le titre de la conférence académique «Science des données, classification et méthodes connexes» (Fédération internationale des sociétés de classification).
1997	1997 Dans sa conférence inaugurale H. C. Carver en statistique de l'Université du Michigan, le professeur C. F. Jeff Wu (actuellement au Georgia Institute of Technology), demande que les statistiques soient renommées les sciences des données et les statisticiens pour être renommé *data scientist* (Strickland, 2014).
2002	Une nouvelle revue, Data Science Journal, a été lancée sous le nom (Committee on Data for Science and Technology 2002)
2009	Troy Sadkowsky créa pour la première fois le groupe dédié à la data science, et à la communauté *data scientist* sur LinkedIn, qui sera suivie par le site web officiel datascientist.net
2010	Les travaux d'Hilary Mason et Chris Wiggins dans l'articl "A Taxonomy of Data Science" démontrent que la data science est le fruit de l'interaction d'expertise approfondie en piratage, en statistiques et apprentissage automatique et une expertise en mathématiques et analytique, tout en prenant des décisions créatives et en ayant une ouverture d'esprit dans un contexte scientifique ".

[1] International Association for Statistical Computing
[2] International Statistical Institute
[3] Knowledge Discovery in Databases
[4] Association for Computer Machinery
[5] The community for data mining, data science and analytics

2011	Suivi des travaux de David Smith en 2011, la data science a été interprétée comme une combinaison de résolution de problèmes, d'analyse de données et de piratage informatique. Il est à noter que les connaissances et les compétences techniques de haut niveau ont été communément identifiées (par exemple, les compétences de piratage informatique) ainsi que les compétences analytiques de haut niveau requises pour les chercheurs universitaires et les scientifiques.

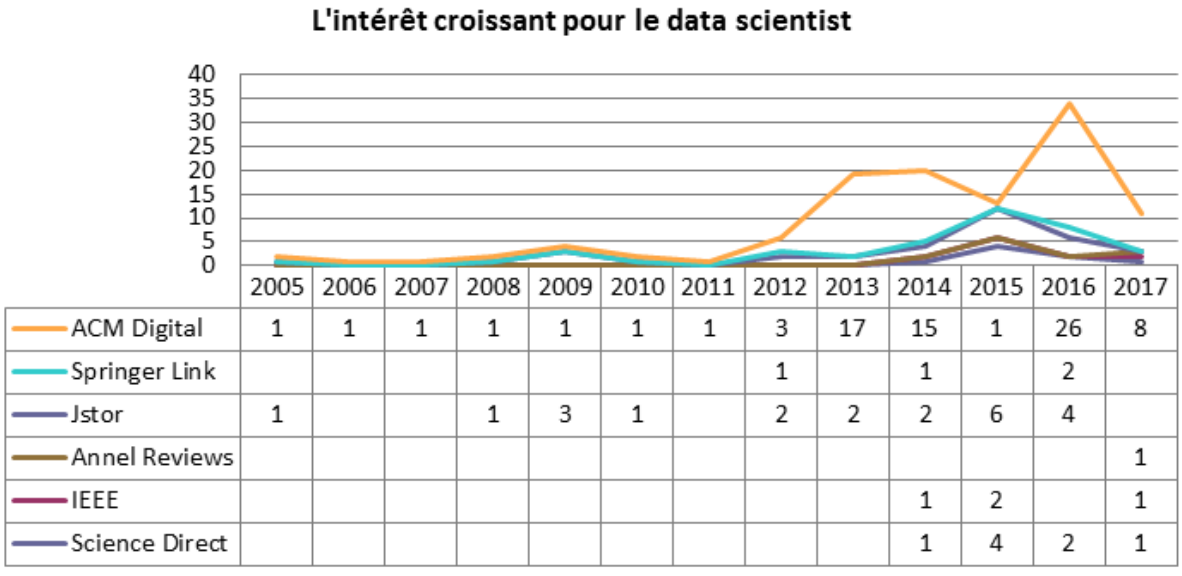
Source : Réalisé par nos soins

En résumé, il semble que l'évolution des concepts de la *data science* correspondent aux innovations dans les technologies informatiques et les systèmes d'information. Il semble également qu'il ait attiré l'attention des universitaires et des écoles d'ingénieurs, ces derniers sont en train d'intégrer et d'adapter le contenu pédagogique autour de la data science.

Après s'être familiarisé avec les BIG DATA et la data science, et avons compris les changements que cela implique, il est maintenant important de se poser la question, quelle compétence faudrait-il pour tirer parti des *big data* ? Qu'est-ce qu'un data scientist ?

Pour répondre au mieux à cette question, nous avons procédé par approche normative, pour la collecte de notre matériau, les étapes de la collecte et la méthodologie adoptées sont détaillées dans notre troisième chapitre dédié à la méthodologie. Seuls les résultats sont détaillés dans cette partie dédiée à la revue de littérature.

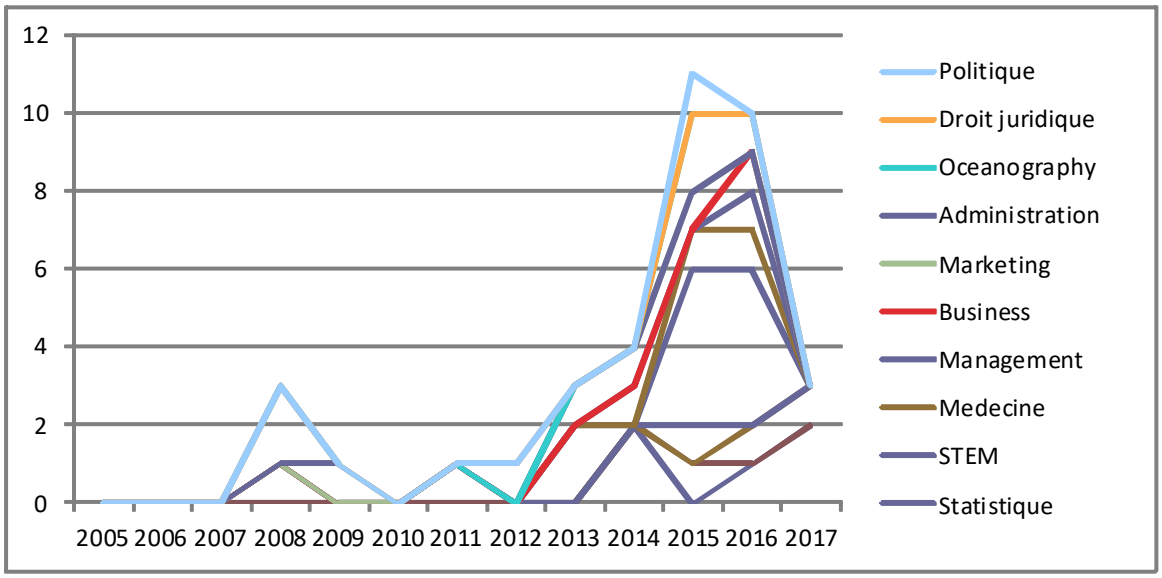
Figure 07 : L'intérêt croissant pour le data scientist :



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
ACM Digital	1	1	1	1	1	1	1	3	17	15	1	26	8
Springer Link								1		1		2	
Jstor	1			1	3	1		2	2	2	6	4	
Annel Reviews													1
IEEE										1	2		1
Science Direct										1	4	2	1

Source : Réalisé par mes soins, en utilisant l'outil graphique de Word

Figure 08 : L'intérêt croissant pour les data scientists dans différents domaines :



Source : Réalisé par mes soins, en utilisant l'outil graphique de Word

Les deux figures, montrent des séries temporelles pour les dynamiques des changements au cours de la même période dans les six bases de données que nous avons parcourus. Les deux graphique montrent clairement un accroissement de l'intérêt pour un data scientist à partir de l'année 2012. Les six bases de données publient plus d'études ces dernières années (2010-2017) vis-à-vis des années précédentes (2005- 2010). **Le graphique 08** montre les différents domaines et secteurs portant un intérêt pour la data science au cours de la période allant de (2005 à 2017), la data science ne se contente pas d'un seule domaine, et depuis l'année 2007, de nombreux domaines, exprime déjà le besoin d'un data scientist.

1.4. A la découverte du métier de scientiste :

Dans la littérature parcourue, nous constatons que plusieurs chercheurs ont étudié les problèmes et les contraintes liés au travail *d'un data scientist*, en axant particulièrement, sur les contraintes technologiques et techniques liées aux mathématiques, algorithmiques et statistiques. Cependant, très peu de recherches ont étudié les rôles et les compétences d'un *data scientist,* d'un point fonctionnel et managérial. Comme le souligne Derrick Harris ''*Qu'est-ce qu'un data scientist, et quelles sont les compétences nécessaires pour être un data scientist reste un sujet de débat intense* (Harris, 2013)*'',* Steven Miller le rejoint en ajoutant *''Il n'existe pas encore de normes officielles, pour définir un data scientist et par conséquent, tout le monde peut s'attribuer lui-même ou elle-même le pseudonyme data scientist''* (Miller, 2014).

Dans la section suivante, nous présentons nos résultats de recherche concernant les différentes définitions d'un *data scientist*, ses rôles et les compétences requises, nos résultats de recherches sont issus de différentes sources (bases de données scientifiques, site web, témoignages…etc).

1.4.1. Résultats :

Compte tenu de notre approche normative nous avons diversifié les sources de collecte de données pour déterminer le matériau à examiner. Cela nous amène à collecter 171 articles publiés dans les six bases de données académiques (pour une explication détaillée, voir le chapitre méthodologie), l'insuffisance de définitions dans ces articles, nous pousse par la suite, à élargir notre recherche pour inclure des sources non académiques provenant de sites web industriels et d'informations.

Le tableau 5 énumère les premières définitions académiques tirées des articles et travaux scientifiques des 06 bases de données, ensuite les définitions tirées des sites web des grands acteurs de l'industrie. Le contenu du tableau 1 est trié par l'année de publication du document par ordre croissant.

Tableau 5 : Qu'est-ce qu'un data scientist ?

Publication Scientifique/Année	**Page**	**Définition : *Un data scientist* est …**
2005	27	Les data scientists sont les informaticiens, les ingénieurs et les programmeurs, les experts en matière de bases de données et de logiciels, les experts disciplinaires, les conservateurs et les experts, les bibliothécaires, les archivistes et d'autres, qui sont essentiels à la gestion réussie d'une collecte de données numériques (NSF, 2005)[1]
2008	09	Un data scientist est un meneur d'enquête et d'analyse créatives; Il améliorer par la consultation, la collaboration et la coordination la capacité des autres à mener des recherches à l'aide de collections de données numériques; il est à l'avant-garde dans le développement de concepts novateurs dans la technologie des bases de données et les sciences de l'information, y compris les méthodes de visualisation des données et de découverte de l'information, et les applique dans les domaines de la science et de l'éducation pertinents pour la collecte; il met en œuvre les meilleures pratiques et la technologie; pour servir de mentor pour le début ou la transition des enquêteurs, des étudiants et d'autres personnes intéressées à la recherche de données scientifiques; il conçoit et met en œuvre des programmes d'éducation et de sensibilisation qui permettent aux chercheurs, aux éducateurs, aux étudiants et au grand public de bénéficier des avantages de la collecte de données et de la science de l'information numérique (Brown, 2008)
2012	70	Les *data scientist* sont les gens qui comprennent comment pêcher des réponses aux importantes questions business dans le tsunami d'information non structurée d'aujourd'hui (Davenport T. H., 2012).

[1] NFS : National Science Foundation

2012	12	Un *data scientist* est une personne qui combine la connaissance des statistiques, des technologies de l'information et de la conception de l'information avec de solides compétences en communication et en collaboration. Le lieu de travail d'un *data scientist* doit procurer l'autofourniture d'informations, pour lui permettre d'être curieux et créatif (A.Cooper, 2012).
2013	11	''Un data scientist est un métier à l'intersection de trois domaines d'expertise : Informatique, Statistique et Mathématiques, et Connaissances métier (Abiteboul S. & s., 2014)."
2013	15	Un *data scientist* est une personne qui prend des matières premières (dans ce cas, des données) et utilise les compétences, les connaissances et la vision nécessaire pour l'élaborer en une valeur unique (Mohanty et al, 2013).
2016	154	*Data Scientist* est le terme pour un nouveau type d'analystes possédant à la fois les connaissances business pour comprendre les problèmes rencontrés par les entreprises ainsi que les compétences techniques pour générer des informations pertinentes pour la prise de décision à partir d'ensembles de données à grande échelle (G.Illescas et al., 2016).
2017	66	Les data scientists aident les organisations à extraire le signal du bruit, ils analysent des montagnes d'information virtuelles, qui ont des idées sur comment le business se fait, et qui améliorent les produits et les services (Tunkelang, 2017).

Publications Industrielles/ Année	**Source**	**Définition : *Un data scientist* est...**
2010	Fondateur Fast Forward Labs	''Un *data scientist* est quelqu'un qui combine, mathématiques, algorithmes et une compréhension du comportement humain avec la possibilité de pirater des ensemble de systèmes pour obtenir des réponses à des questions humaines intéressantes à partir de données (Mason, 2010) ''
2011	Forbes	''Les data scientists sont avec un esprit analytique, statistiquement et mathématiquement sophistiqués, qui peuvent inférer des connaissances sur les entreprises et d'autres systèmes complexes à partir de grandes quantités de données (Hillion, 2011)''.
2012	IBM	''Un data Scientist est un profil interdisciplinaire qui combine l'apprentissage automatique, les statistiques, l'analyse avancée

		et la programmation. C'est une nouvelle forme d'art qui élabore des idées cachées et met les données au travail à l'ère cognitive. Dans ce sillage de l'innovation, les data scientist peuvent détruire les barrières de données et développer des idées qui changent le monde (IBM, Data science, 2010)''
2012	Fondateur DataKind	''Un *data scientist* est un hybride rare, un informaticien possédant des capacités de programmation pour construire des logiciels, afin de combiner et gérer des données provenant de diverses sources, et un statisticien qui sait comment tirer des idées de l'information. Il combine les compétences pour créer de nouveaux prototypes avec la créativité et la rigueur de poser et de répondre aux questions les plus profondes sur les données et les secrets qu'elle détient (Porway, 2012)''.
2011	ASA	''Nous devons dire aux gens que les data scientists sont ceux qui ont un sens du déluge de données sur les sciences, l'ingénierie et la médecine ; ils fournissent des méthodes d'analyse de données dans tous les domaines, de l'histoire de l'art à la zoologie; il n y a pas plus excitant que d'être un *data scientist* au 21e siècle en raison des nombreux défis provoqués par l'explosion de données dans tous ces domaines Présidente'' ASA (Geller, 2011).
2011	O'reilly	Un scientifique de données est un mélange unique de compétences qui peuvent à la fois débloquer les idées de données et raconter une histoire fantastique via les données (Patil, 2011) "[1]
2010	O'reilly	Les data scientists sont ces personnes impliquées dans la collecte de données, en lui redonnant une forme traitable, en lui créant son histoire et en présentant cette histoire aux autres (Loukides, 2010)
2015	BBVA	''Le data scientist va bien au-delà du Power Point auquel nous sommes habitués à voir dans le monde de l'innovation : sa responsabilité commence par la conception d'un prototype avec les technologies les mieux adaptées au problème en main (Hadoop, MongoDB, Spark, Python, R), Et se termine par la supervision de sa mise en œuvre'' (Teleña, 2015)[2].
2012	The Gruardian	Un data scientist est un ingénieur ayant de la créativité, la ténacité, la curiosité et des compétences techniques, pour la collecte de données, la visualisation, l'apprentissage

[1] Dr. DJ Patil un Data Scientist à Residence at Greylock Partners et anciennement data scientist à LinkedIn.

[2] Teleña Sergio Álvarez Directeur de la stratégie global et data Science au BBVA.

	automatique et la programmation informatique. Ils font parler la donnée, pour une meilleure prise de décision (Howard, 2012)[1]

Source : réalisé par nos soins

Les 17 définitions listées dans le tableau 1 montrent différentes conceptions de qu'est-ce qu'un data scientist, et quels sont ses rôles et missions. La définition et l'identification des perspectives du métier se sont affiner au fil du temps, et se sont développées en parallèle avec le développement de l'environnement du travail *BIG DATA.*

Ces résultats nous ont permis de recueillir assez d'informations pour constituer une grille regroupant l'ensemble de compétences et de rôles attribués à un data scientist. Le tableau ci-dessous comporte l'ensemble de ces compétences communes retrouvées dans les différents travaux scientifiques, ou des discours de grands acteurs de l'industrie BIG DATA.

Tableau 6 : Grille de compétences, « Skills Grid »

Attributs	US National Science	Sheridan Brown &	A.Cooper 2012	S.Abiteboul, 2014	G.Illescas 2016	Hilary Manson 2010	Steve Hilion 2011	IBM 2012	Jake Porway 2012	SAS 2011	BBVA 2015	The Gruardian	Total
Algorithmique						X						X	2
Communication			X										1
Compétences business					X	X	X		X	X		X	6
Créativité		X	X						X			X	4
Curiosité			X				X					X	3
Data Manipulation	X	X				X		X	X	X		X	7
Data science technologies (Machine learning/deep learning/prototypage								X	X		X	X	4
Esprit d'équipe/collaboration			X										1
Expertise logicielle	X		X	X	X				X			X	6
Innovation		X									X		2
Leadership		X											1
Mathématiques (Linear, Algèbre, Analyse, optimisation)				X	X	X	X						4
Programmation Back-end/Front-End											X	X	2
SGBD	X	X						X	X				4
Statistique (Classique/Temporel/Spatial/Bayésienne			X	X	X		X	X	X				6
Technologies BIG DATA					X						X		2
Visualisation		X										X	2

[1] Jeremy Howard, président et Chef data scientist, Kaggle

Source : réalisé par nos soins, en recueillant les définitions de l'état de l'art

Presque toutes les définitions semblent d'écrire qu'est-ce qu'un data scientist, ce qu'un data scientist produit et livre (produit) et comment (processus de production les résultats souhaités). Notre sujet cherche à connaître quels sont les facteurs clés de succès du métier de data scientist. Pour y répondre, nous avons regroupées l'ensemble des compétences à partir des 17 définitions listées. Ces résultats vont nous permettre de mener une enquête comparative avec nos acteurs à partir du tableau ci-dessus les discussions sont détaillées dans notre quatrième chapitre.

2. Cadre conceptuel

Nous allons introduire dans cette partie une brève définition des concepts cités et abordés, tout au long de ce mémoire, afin de mettre le lecteur dans le cadre de notre sujet de recherche.

Dans le but de comprendre qu'est-ce qu'un data scientist ? Quels sont ses rôles et compétences, il est important de situer le data scientist dans son contexte, et de comprendre et de connaître son champ d'application.

2.1. Caractéristiques des big data

Comment nous l'avons vu, dans les précédentes définitions, les big data sont la conséquence de la digitalisation du quotidien, tout ce qui digital est de la data. Le matériel informatique ainsi que les logiciels d'aujourd'hui ne peuvent cependant pas traiter un volume aussi important de données. Celui-ci est ainsi si important et si complexe et dynamique, qu'il ne peut être traité, stocké, analysé, et gérer avec les outils traditionnels de traitement de données (RIJMENAM, 2014, p. 5).

Heureusement, grâce aux avancés technologiques, il existe aujourd'hui des outils, algorithmes, et logiciels, qui transforment la donnée en valeur, en information utile, les renseignements délivrés par cette information, permettent à l'entreprise de prendre les bonnes décisions, d'améliorer son efficacité, de diminuer ses coûts, et accroître son revenu, les implications de la révolution big data sont vastes, elles impactent tout type d'entreprise. (RIJMENAM, 2014, p. 5).

2.2. Mesurer l'impact des big data :

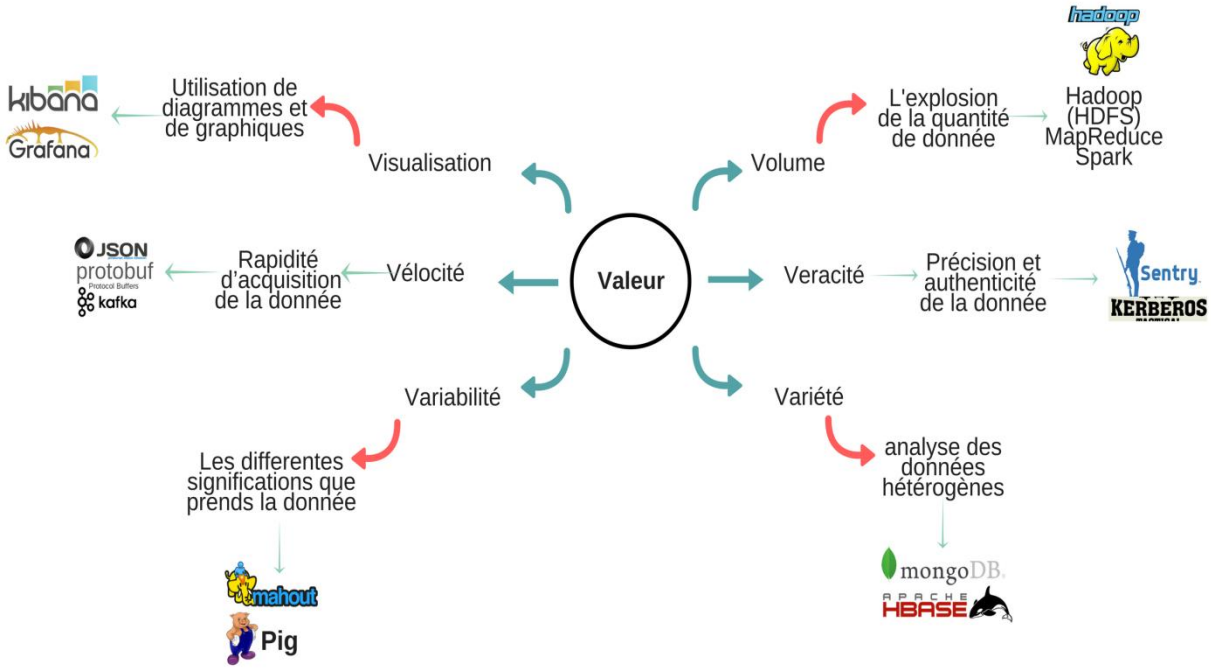
Afin de mesurer l'impact des big data, il est important de connaître ses caractéristiques essentielles, connus plus communément sous les ‘‘V’’ du big data :

De nombreux travaux présentent les big data sous les « 03 V ». La première définition de Big Data a été développée par Meta Group (font partie de Gartner aujourd'hui) en décrivant les trois caractéristiques appelées «3V» : Volume, Vélocité et Variété. IBM a ajouté ensuite, un quatrième V appelé : Véracité, pour décrire la qualité de la donnée, Par la suite, Oracle a ajouté un cinquième V appelé : Valeur, mettant en évidence la valeur ajoutée des Big Data (Abdelkader Baaziz, 2013, p. 20). Pour finir, nous complétons ces caractéristiques de deux autres V, Variabilité et Visualisation ajouté par l'auteur Mark VAN RIJMENAM, pour expliquer l'impact et l'implication d'une stratégie Big data. En combinant les caractéristiques,

nous obtenons une définition plus exhaustive appelée "7V": Volume, Vélocité, Variété, Valeur, Véracité, Variabilité et Visualisation.

La figure ci-dessous illustre le changement de paradigmes poussés par les Big data :

Figure 9 : les problèmes posés par les Big data et les solutions technologiques correspondantes



Source : Schéma réalisé par nos soins en utilisant l'outil graphique Canevas[1]

La figure ci-dessus, illustre les 7V des big data, et les 7 problèmes posés, ainsi que les solutions technologiques développées aujourd'hui, pour y capturer toute la valeur. Le but ultime des big data est la création de valeur. Pour cela les big data regroupent une famille d'outils qui répondent aux 6 problématiques : un Volume de données importantes à traiter, une grande variété d'informations (en provenance de plusieurs sources, structurées, semi-structurées, non-structurées), soulevant chacune une question technologique et méthodologique spécifique à son analyse et traitement, un certain niveau de Vélocité à atteindre - c'est-à-dire de fréquence de création, collecte, traitement/analyse et partage de ces données, une vérification de la véracité de la donnée, quant à sa conformité, et sa précision, une variabilité de la donnée à prendre en considération, en effet celle-ci prend plusieurs significations, et une importance colossale quant à la Visualisation de cette dernière (RIJMENAM, 2014, pp. 7,12).

Tous ces outils propulsent le data scientist vers de nouveaux horizons de recherche mais aussi dans un écosystème, une jungle aux imbrications matérielles, logicielles, économiques... excessivement complexes, et la maîtrise de ces logicielle est l'un des facteurs clés de succès d'un data scientist (Philippe Besse, 2014, p. 2).

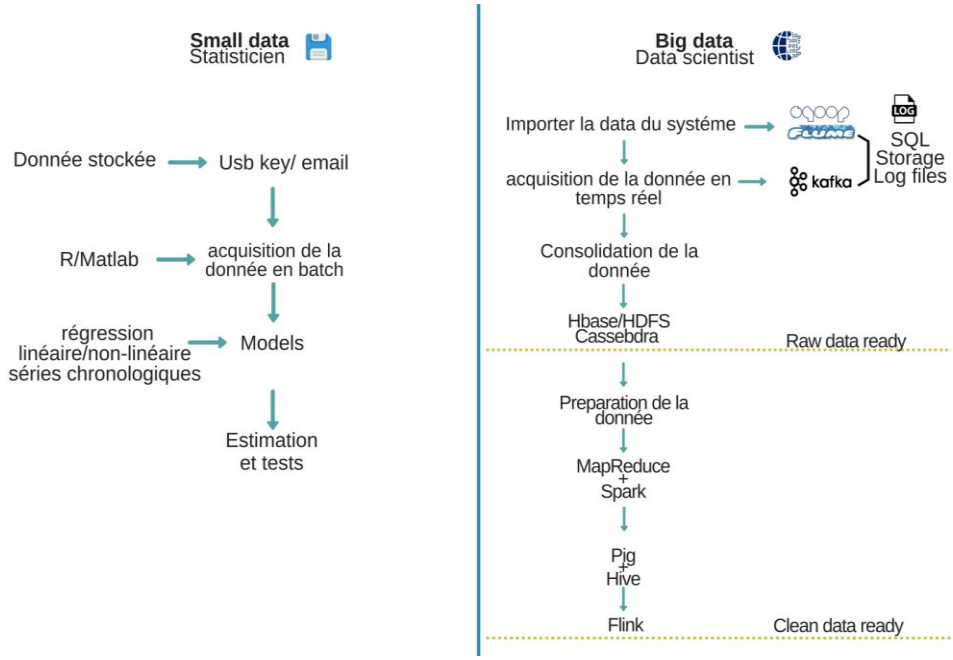
[1] Cette illustration a été créée, grâce au livre "Field guide to Hadoop : An Introduction to Hadoop, its ecosystem, and Aligned Technologies", Kevin Sitto and Marshall Presser, O'reilly

Effectivement, le statisticien, traite un volume de donnée en Méga-Octets (MO), cela ne nécessitait pas une méthodologie et technologique particulière, les technologies utilisées sont dotées d'une interface graphique ergonomique *userfriendly*[1], avec un assemblage de logiciels dédiés intégrant un langage de requête SQL[2] (de base), explorations statistiques (réduction de dimension) classification non supervisée (clustering) modélisation stratégique classique[3], et algorithmique (arbre de décision, réseaux de neurones) (Philippe Besse, 2014, p. 7).

Aussi, beaucoup de ces méthodes mathématiques et statistiques utilisées sont en commun dans la data science, mais le contexte technologique de mise en œuvre est unique. Le data scientist est confronté à de nouvelles techniques statistiques, mathématique et algorithmique plus complexe liée à la multiplicité de la donnée, tels que SVM[4] (algorithme de séparation permet la comparaison dans la distance entre les objets traitées), les modèles d'apprentissage statistiques, automatique et profond (Machine Learning, Statistical Learning and Deep Learning), Lasso (Philippe Besse, 2014, p. 10), la parallélisation de stockage, et la nouvelle structure des données réparties en nœuds, syntax de la programmation fonctionnelle facilement distribuable en (MapReduce) pour l'utilisation des langages tels que scala, clojure, ainsi leurs confrontation à des problèmes de perpétuelle optimisation, quand à l'équilibre biaie/variance (crossvalidation[5]) pour optimiser sans cesse les performances du modèle (Philippe Besse & Béatrice Laurent, 2016, p. 80).
Afin d'illustrer, la jungle dans lequel le data scientist opére, nous prenons deux cas différents de préparation de la donnée, dans deux environnements différents, face à deux volumes différents de donnée. La figure ci-dessus illustre les différentes étapes de la procédure de préparation de la donnée, dans deux cas différents big data/ small data :

[1] Un système est dit « user friendly » quand il est convivial, facile à utiliser par ses usagers.
[2] Structured Query Language
[3] Méthodes et modèle statistique, d'exploration et de modélisation de donnée
[4] Support Vector Machine (machine à vecteur de support)
[5] Validation croisée Définition disponible dans le glossaire ajouté en annexe

Figure 10 : procédure de préparation de la donnée big data/ small data

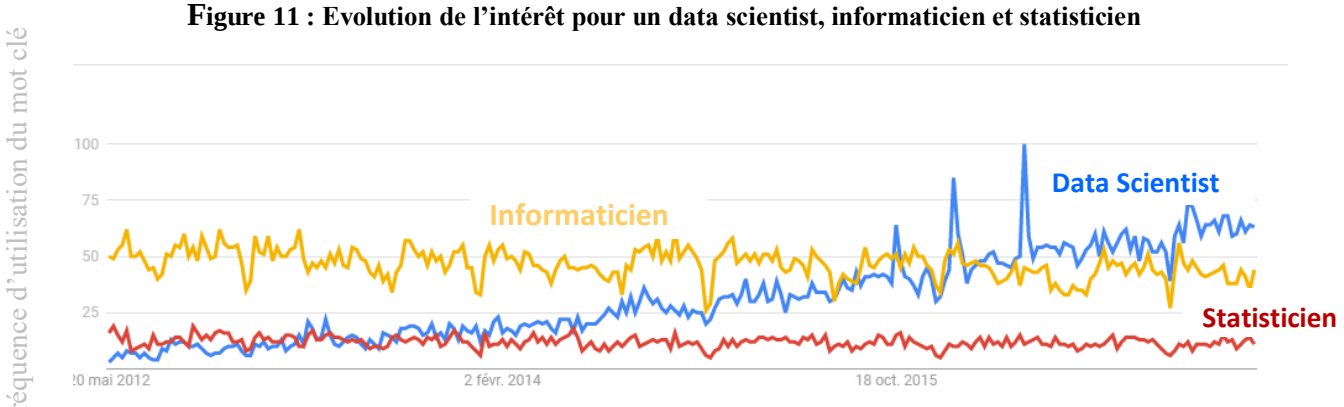


Source : Schéma réalisé par nos soins en utilisant l'outil graphique Canevas[1]

Le schéma démontre, effectivement, que le challenge pour un data scientist est de taille, il n'est pas seulement question de maîtriser les outils technologiques, mais de les implémenter, les administrer et faire tout un travail en amont de mathématiques et statistiques pour la préparation des modèles (algorithmes prédictifs), une réflexion et une analyse business et stratégique en aval, pour délivrer la solution au final.

Dans le contexte d'aujourd'hui, conduit par la digitalisation et la *dataficatio* des entreprises, la demande pour les compétences en big data et en science des données est en accroissement, au détriment des autres domaines. En effet, en effectuant une recherche sur Google Trends, pour voir à quelle fréquence la demande et l'intérêt d'un data scientist a évolué par apport des autres métiers, recouvrant plus au moins des compétences communes, tel que l'informaticien et le statisticien.

Figure 11 : Evolution de l'intérêt pour un data scientist, informaticien et statisticien



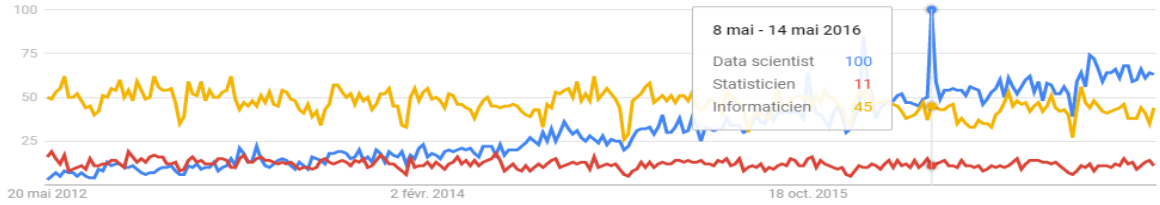
Source : Google Trends, (Google tendance des recherches)

[1] Toutes les technologies, et les outils sitès sont définis dans le glossaire ajouté en annexe

Les résultats reflètent la proportion de recherches portant sur les mots clés (Data scientist, Informaticien, Statisticien) dans toutes les régions du monde et pour une période allant du 20 Mai 2012 au 15 Mai 2017, par rapport au taux d'utilisation de ce mot clé sur le moteur de recherche Google, la valeur de 100 indique une valeur deux fois plus supérieure aux autres tendances. Ainsi, une valeur de 50 signifie que le mot clé a été utilisé moitié moins souvent, et une valeur de 0 correspond à une région ayant enregistré moins de 1 % de correspondances, par rapport à celle qui a obtenu 100.

Le graphique montre un pique atteignant les 100 pour un data scientist pour la période allant de janvier à mai 2016, contre un décroissement de l'intérêt pour un informaticien et un statisticien.

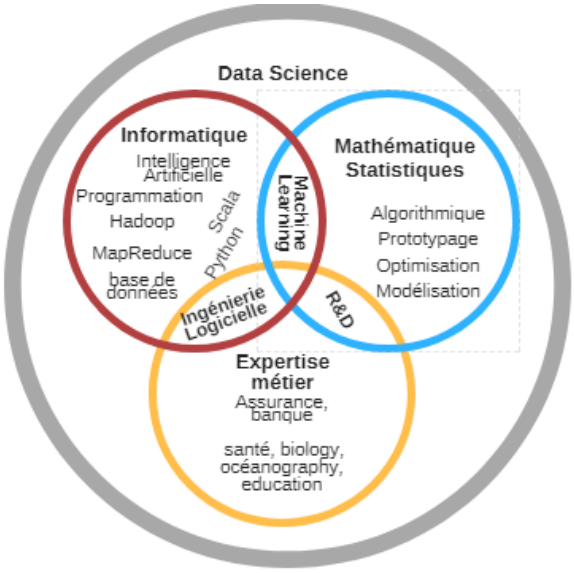
Figure 12 : Evolution de l'intérêt pour un data scientist



Source : Google Trends

Les courbes démontrent clairement que durant les quatre dernières années, l'intérêt pour le data scientist l'emporte sur le métier de statisticien et d'informaticien. Il est vrai qu'aujourd'hui la data science se retrouve à la croisée de plusieurs disciplines (voir figure), ce qui explique clairement l'intérêt croissant, et la hausse de la demande du métier de data scientist.

Figure 13 : Ensemble de desciplines de base de la data science



Source : Schéma réalisé par nos soins en utilisant l'outil graphique Canva

Le data scientist recouvre plusieurs métiers en un seul, l'informaticien, le mathématicien et statisticien et l'expert métier (Corea, 2016, p. 25), l'arrivé de ce profil a bouleversé le marché de travail, et ceux pour les raisons suivantes :

- Le data scientist est entrain de substituer les métiers de statisticien, informaticien, expert data…
- La chaîne de valeur est concentrée vers un seul métier et n'est pas dispersée sur plusieurs fonctions,
- Un seul individu peut probablement être plus productif, et faire des tâches plus rapidement que cinq personnes différentes travaillant sur le même problème dans même échéance, et ceux en raison de sa spécialisation et de ses connaissances et sa flexibilité de haut niveau,
- L'embauche d'un data scientist devrait coûter moins cher qu'un total de cinq semi spécialistes (Corea, 2016, p. 26).

CHAPITRE III : CADRE MÉTHODOLOGIQUE

1. Choix méthodologiques et épistémologiques

Nous avons choisi d'insérer notre recherche dans un cadre épistémologique de type constructiviste. Dans le but d'effectuer un travail cohérent, nous allons expliciter les raisons pour lesquelles cette perspective a été choisie et en quoi elle oriente les questions de recherche.

Comme le relève Charlier : ''*une manière de voir ou d'étudier un problème étant aussi une manière de ne pas voir…*'' (Charlier , 1998) Nous tenons à préciser ici qu'il s'agit d'une option prise et que nous sommes conscients que d'autres approches sont possibles.

4.3.5 1.1. L'approche constructiviste

Pour Guba, l'approche constructiviste se définit par la cohérence entre trois caractéristiques : la nature de la réalité, la relation établie entre le chercheur et l'objet à connaître, la méthode à mettre en œuvre pour construire la connaissance (Guba & Lincoln, 1985)

4.3.5.1 1.1.1. La nature de la réalité :

4.3.5.2 Dans l'approche constructiviste la réalité n'existe pas en soi mais est une création qui s'effectue dans la relation entre l'individu et son environnement (Guba & Lincoln, 1985).

Comme de nombreuses autres recherches, le but de notre étude est d'explorer en profondeur le métier de data scientist et l'environnement dans lequel il opère. Nos données se fondent sur les réalités forcément subjectives et contextualisées de nos sujets.

4.3.5.3 1.1.2. La relation établie entre le chercheur et l'objet à connaître

A ce propos, Charlier affirme : '' l'épistémologie est subjectiviste. Les découvertes sont littéralement une création issue d'une interaction entre le chercheur et l'objet à connaître'' (Charlier, 1998). Dans cette optique, nous menons notre enquête auprès de nos acteurs pendant une période qui s'étend du mois de Janvier au mois de Mai, pour pouvoir cerner au mieux les comportements et les interactions de nos acteurs dans leur environnement, en excluant toute forme de subjectivité.

Ce point de départ, nous pousse à adopter une approche qualitative, outillée de méthode normative, du paradigme pragmatiste.

2. Méthode et instrument de mesure :

2.1. Une recherche qualitative :

En voulant comprendre qu'est-ce qu'un data scientist, et quels ont les facteurs clés de succès constituant le métier, nous abordons un sujet complexe. La relation entre les compétences et les rôles, liées à l'environnement technologique dans lequel ce dernier opère d'un point de vue fonctionnel et managérial ont été peu mis en relation. Afin de découvrir au mieux, ces facteurs clés de succès nous avons choisi d'effectuer une recherche qualitative.

2.2. Une recherche qualitative outillée de pratiques normatives :

‘‘*Sans une théorie de l’objectivité toute institution sociale devient inexplicable’’* (Frega, 2015, p. 5)

Pour étudier nos acteurs en excluant toute forme de subjectivité. Nous outillons notre recherche qualitative, par des pratiques normatives, qui sont la base de la sociologie, pour étudier toute forme et toute dimension de comportements et d’interactions humaines.

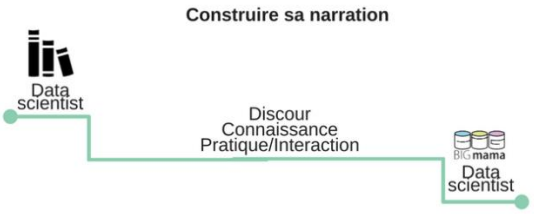
De plus la normativité explique comment, par le biais de quelles actions et de quels discours, les agents mobilisent et modifient les ordres normatifs qui régissent leur vie en commun (Frega, 2015, p. 9)

4.3.6 En extrapolant ceci sur notre recherche, nous tenons à découvrir comment le data scientist étudié dans notre cas, se distingue du moule de data scientist cité dans les théories, et ce, en se basant sur une construction de théorie interdisciplinaire, en science sociale et managériale, qui nous permettra de comprendre les conditions de validité de notre démarche, qui demande à son tour qu’on mobilise une pluralité de perspectives et de démarches.

2.3. Une recherche descriptive

Dans une telle recherche une partie descriptive est nécessaire. Comme le dit Charlier : ‘‘une première démarche est de décrire avec précision et rigueur le phénomène étudié en le situant dans le contexte de la situation particulière appréhendée par le chercheur.’’ (B.Charlier, 1998, p. 123) Dans notre cas, la recherche descriptive, nous permet de détailler le contexte dans lequel se situe notre champ d’étude d’un point de vue théorique, en toute objectivité, elle nous permettra de décrire l’environnement dans lequel notre data scientist opère, les rôles et les compétences selon l’état de l’art et le cadre conceptuel de notre champ d’étude. Aussi cela nous permettra d’orienter et de guider la construction de notre narration.

Figure 14 : construction de la narration



Source : réalisé par mes soins en utilisant l’outil Canva

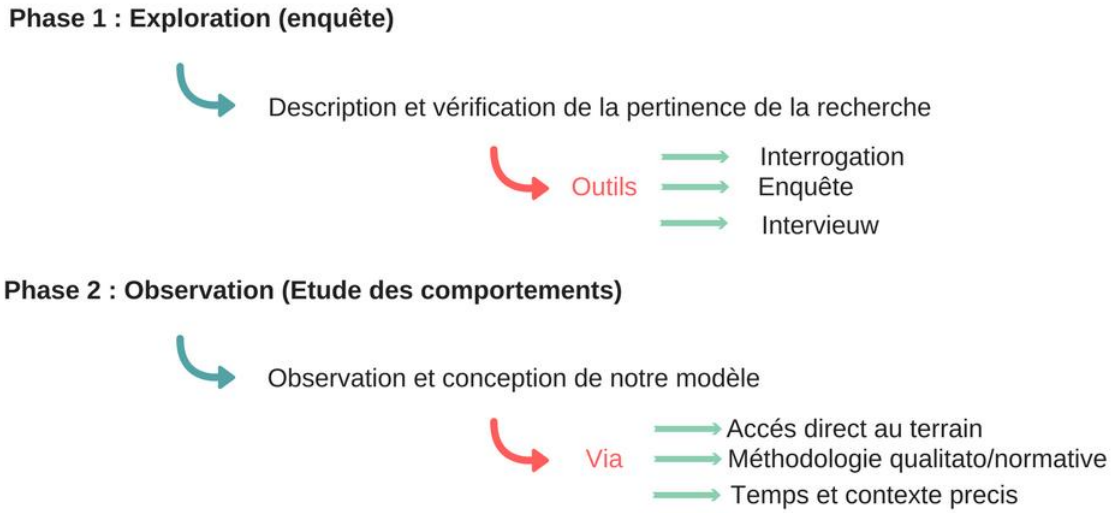
2.4. Une recherche narrative :

‘‘La narration a pour but d’expliquer une dynamique sociale de l’interaction des individus, et l’on est dans une vision d’exploitation de l’information, pour la production de la connaissance’’ (Dumez, 2016, p. 129) Nous construisons notre narration sur la base de notre recherche descriptive, mais en excluant toute théorie. Nous présentons les acteurs et leurs actions et interactions, leurs discours et interprétations, qui font l’objet de rapprochement et

de comparaison des discours tenus par nos data scientists, aux connaissances techniques dont ils disposent, sur les différents projets et situations de travail qu'ils traversent, à leurs pratiques, routine et évolution.

Dans la partie qui va suivre, nous allons illustrer par quelle méthode, nous avons procédé, au recueil de notre matériau.

Figure 15 : Méthodologie d'observation du métier de data scientist



Source : réalisé par nos soins, en utilisant l'outil Canva

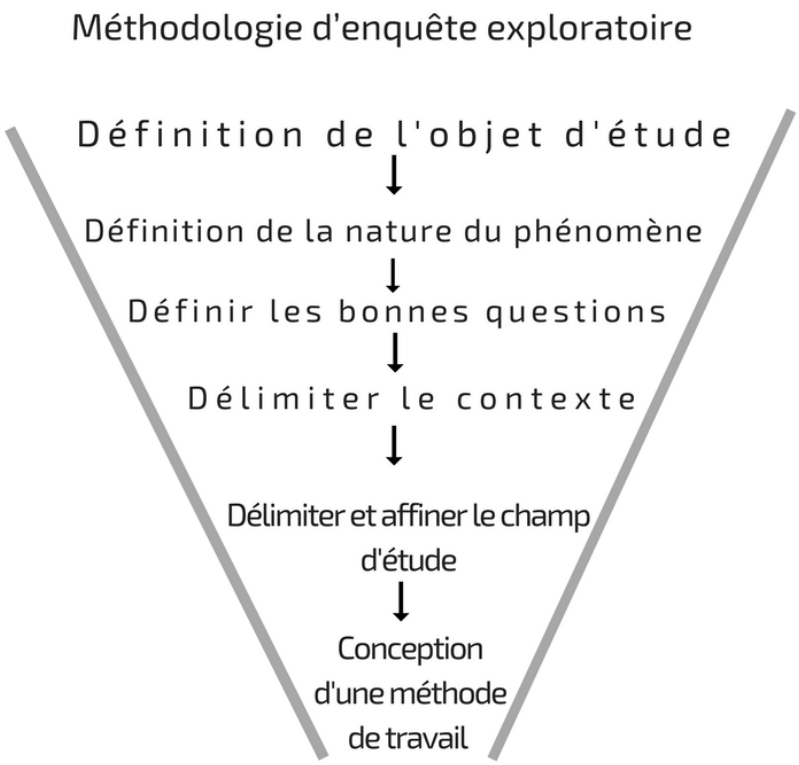
2.5. Une recherche exploratoire :

Dans un deuxième temps, nous effectuons un travail exploratoire. Nous mettons en relation les données analysées à des moments différents afin de mettre en évidence des changements éventuels de tensions ou de perception du dispositif. « Ces comparaisons diachroniques fondent la démarche de recherche exploratoire. » (Charlier, 1998, p. 123). Par la confrontation des cas étudiés, nous pensons faire émerger de nouvelles hypothèses.

Nous précisons que, dans un 1er temps, nous visons à comprendre, anticiper les évolutions de nos data scientists, face aux différents projets auxquels ils sont confrontés. Par la suite, nous repérons et identifions les rôles et compétences de chacun, et les qualifications correspondantes à chaque phase du projet. Enfin nous identifions les évolutions des data scientists au fur et à mesure des projets et à leurs complications. Ces observations feront l'objet d'une d'approche inductive, qui part du discours des interviewés, et s'enrichit d'un cadre théorique afin de proposer un modèle type de notre data scientist.

Le schéma ci-dessous nous permet d'illustrer notre méthodologie par enquête d'exploration

Figure 16 : Méthodologie d'enquête exploratoire



Source : Réalisé par nos soins en utilisant l'outil graphique Canva

3. Procédure de collecte de données :

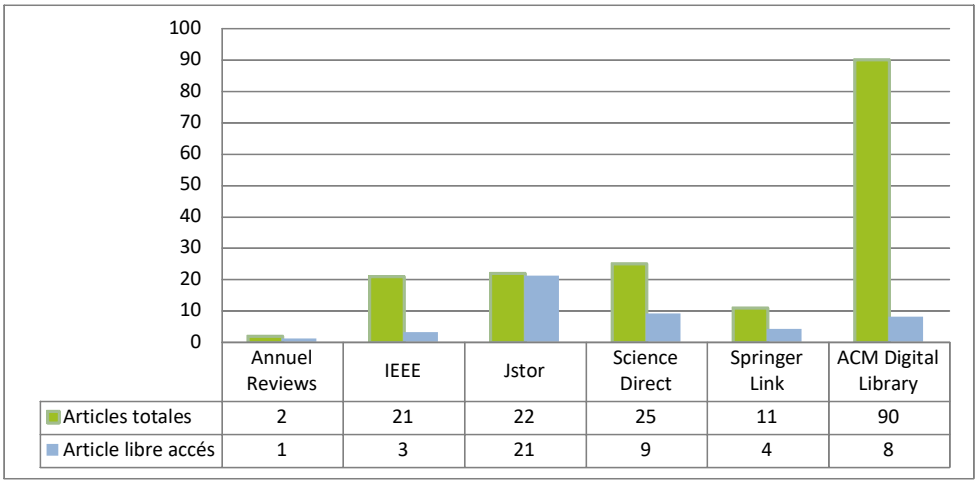
Afin de répondre à ces questions, nous avons d'abord procédé à une revue systématique de la littérature existante. Une recherche systématique accumule un recensement relativement complet de la littérature pertinente.

3.1. Le recueil de données par approche normative

Nous nous inspirant des travaux de (Chatfield et al, 2014) sur le receil du matériau source, et de l'article (Frega, Les pratiques normatives, 2015,) qui comportent trois étapes structurées pour mener une démarche normative. Cette approche, viendra compléter notre cadre de recherche. Nous procédons dans cette partie comme suit :

Tout d'abord, nous avons utilisé "data scientist" comme mot-clé principal pour notre stratégie de recherche. Deuxièmement, nous identifions six bases de données académiques majeures à travers lesquelles nous recherchons des articles de revues hautement cités : Annuel Reviews, IEEE, Jstor, Science Direct, Springer Link, comptenu de la différence et variances des paramètres des moteurs utilisés sur les bases de données, nous avons utilisé la stratégie de requête générique suivante: (Titre OU Résumé) CONTIENT ("data scientist") ET (Année de publication) = (2005-2017) ET (Type de publication) = (Article de journal), en excluant les articles rédigés dans d'autres langues que l'anglais et le français .Cette stratégie de recherche a révélé 171 articles publiés de différentes catégories, articles scientifiques, revues spécialisées, revues généralistes, handbooks, communications…etc.

Figure : 17 : Tendance globale dans les principales bases de donnée académique 2005-2017



Source : Réalisé par mes soins, en utilisant l'outil graphique de Word

Cette démarche de recherche à cependant montré une documentation insuffisante sur les rôles et les compétences d'un data scientist et sur les exigences professionnelles correspondantes.

Pour élargir notre recherche, dans la deuxième phase, nous avons utilisé Google Scholar avec une requête générique de mots clés sur "data scientist". Une première recherche avec quelques filtres appliqués (années 2005-2017) a généré 7 380 résultats avec des documents de tout genre (livre, article de journal, publication magazine…etc.), les résultats donnent des articles redondants, pour mieux ciblés et filtrer la recherche, nous procédons à la recherche par combinaisons de mots clés, le tableau ci-dessus illustre les mots clés utilisés, et les résultats obtenus dans les deux langues (français et anglais).

Tableau 7 : résultats de recherche Google Scholar

Mots clés	Résultats
«data scientist »	7380 résultats (anglais + français)
Data scientist + skills	2 770
Data scientist + competencies	505 résultats
Data scientist + formation	1 130
Data scientist + profil	101
Data scientist + experience	3 170
Data scientist + metier	7 résultats

Source : réalisé par nos soins, à partir des résultats de recherche Google Scholar

La troisième phase consiste à des recherches sur des sites web qui comprenaient des acteurs connus de l'industrie des BIG DATA et leurs publications sur les data scientist tels que : IBM, SAS, Harvard Business Revieuw, Cloudera, BBVA, LinkedIn, O'reilly, Forbes, ASA, The Guardian, Microsoft, et autres... Ces résultats de recherche nous ont permis de recueillir 16 définitions différentes d'un data scientist. Le recueil de données auprès de nos acteurs :

Pour analyser les compétences et les rôles de nos acteurs, différents outils de recueil d'information ont été combinés, ce pluralisme méthodologique s'inscrit dans le cadre de notre approche normative. Le recueil des données s'est effectué par plusieurs étapes, tout au long des trois mois passé au saint de l'entreprise d'accueil, les étapes de collecte se sont effectuées comme suit :

3.1.1. Méthodologie de recueil de données au prés de nos acteurs :

- Etat de l'art
- Construction de la grille d'analyse
- Entretiens semi-directifs,
- Groupe de travail, participation au projet,
- Enregistrements vidéo d'entretiens professionnels

3.2. Sélection et recrutement des interviewés :

4.3.6.1 Nous avons passé une période de trois (03) mois auprès de nos acteurs, nous les avons vus interagir, évoluer, nous avons aussi, été, à un moment, acteurs durant les projets dans lesquelles nous les avons observés. Cependant, afin d'arriver à notre but de recherche, il a été nécessaire, de structurer notre recueil de données de manière officielle, pour cela nous menons une enquête par entretiens, où nous interrogeons nos acteurs durant différentes phases de leurs projets, durant des périodes séparées dans le temps, car l'on ne recherche plus autant à observer les comportements, que d'analyser les attitudes. Le choix du recueil de données par entretien, est poussé par l'approche adoptée pour traiter notre projet de recherche, qui est l'approche qualitative

3.2.1. Entretien semi-directif :

Pour répondre aux questions, nous menons notre enquête par des entretiens semi-dirigés. Les questions sont ciblées, mais restent ouvertes dans le sens où l'interviewé à une grande liberté de réponse. Cette forme de questionnement favorise l'approfondissement de la réflexion et évite ainsi de rester dans un discours de surface,

3.2.2. Elaboration du guide d'entretien :

Les résultats de l'état de l'art nous ont permis de construire une grille de compétences, que nous avons appelés ''skills grid[1]''. Ensuite, nous avons organisé un ensemble d'entretiens concentré sur trois thématiques : compétences, rôle et enfin, formation scolaire et évolution. Deux ensembles de questions d'enquête se sont prêtés à une analyse de regroupement (entretiens compétences, et entretiens formation). Les thématiques sont présentées dans le tableau ci-dessous :

[1] La grille de compétences est ajoutée en annexe

Tableau 10 : Thématique des entretiens

N°	Thématique	Description
1	Compétences	Des questions portant sur les compétences et les savoirs acquis par les data scientist, en comparaison avec la grille des compétences et savoir, que nous avons constitué grâce à l'état de l'art.
2	Rôles durant les projets	Plusieurs questions, réparties sur plusieurs phases, pendant une période de trois mois des différents projets menés par les data scientist, le but était de tirer les compétences nécessaires pour l'atteinte des objectifs des projets
3	Formation et évolution	Une série de questions portant sur la formation scolaire et les expériences précédentes, ainsi que l'évolution au cours des projets, le but était de connaître le point de départ des data scientist, et faire une triangulation des informations recueillies grâce aux deux thématiques précédentes, pour connaître le point d'arrivé, afin de mesurer l'évolution des data scientist.

Source : réalisé par nos soins

L'ordre des questions a été choisi afin de permettre aux interviewés d'entrer petit à petit dans l'entretien.

Déroulement des entretiens :

Chaque interviewé a été interrogé à trois reprises, tout au long des projets durant les trois mois de notre stage. La répartition s'est effectuée ainsi :

Tableau 11 : déroulement des entretiens

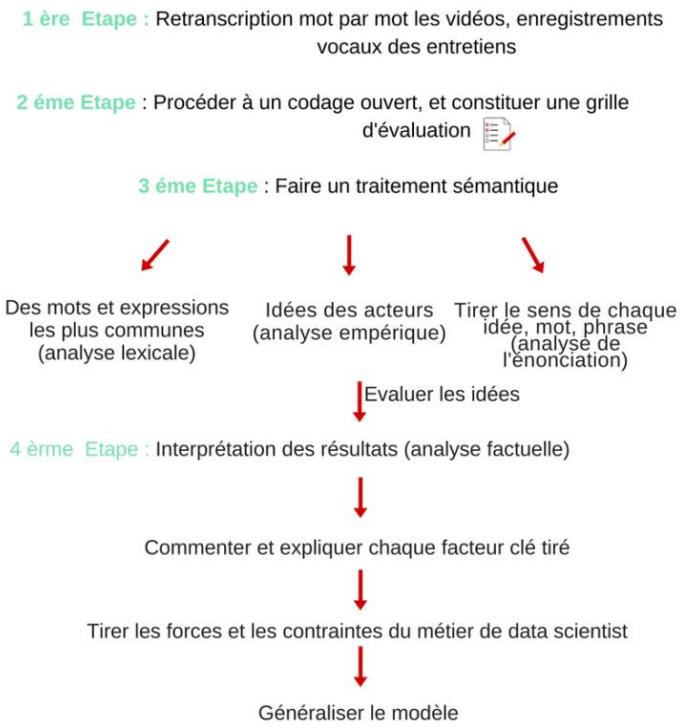
Date	Déroulement	Durée
26 Février 2017	1er enregistrements vidéo et audio d'entretiens avec les data scientists, pour connaître les projets sur lesquels ils travaillaient, les data scientists avaient chacun un projet, ou travaillaient mutuellement sur des projets différents.	3h30 à 4h
24 Avril 2017	2e série d'entretiens avec les data scientists, pour rester à jour avec les projets, certains allaient être, de nouveaux venaient d'arriver.	2h à 2h 30
14 Mai 2017	Dernière série d'entretiens, celle-ci consistait à mesurer l'évolution des compétences des data scientist, au cours des projets précédents.	1h 50 minutes à 2h

Source : Réalisé par nos soins

4. L'analyse qualitative de contenu

Comme le dit Bardin : « Le but de l'analyse de contenu est l'inférence de connaissances relatives aux conditions de production à l'aide d'indicateurs. » (Bardin, 1983, p. 39). Nous allons définir ces indicateurs dans le but de faire ressortir les représentations des interviewés sur la thématique nous intéressant, les détails de l'analyse sont illustrés dans le schéma suivant :

Figure 18 : analyse qualitative du contenu



Source : réalisé par nos soins, en utilisant l'outil graphique Canva

Le schéma montre clairement la procédure de traitement et d'interprétation de nos résultats, qui est l'étape la plus importante dans un travail de recherche, dans cette phase de travail, la méthode d'analyse doit nous aider à nous protéger de notre subjectivité. A ce sujet, Albarello relève : « *Rien n'est plus facile que de faire dire aux acteurs sociaux interrogés ce que l'on souhaite qu'ils aient dit*. » (Albarello, 1999, p. 57). Comme chercheurs, nous allons nous efforcer d'être rigoureux à chaque étape d'analyse de contenu. C'est-à-dire au moment du traitement des données en unités de sens, lors de leur catégorisation et finalement lors de leur analyse.

5. Considération éthique :

Etant donné que dans toute recherche, des règles éthiques et déontologiques sont à respecter et annoncer aux enquêtés. Pour le bon déroulement du stage, nous signons un NDA[1]avec l'entreprise (un accord de non-divulgation et non communication de secret ou de confidentialité).

Tout au long des entretiens, nous avons mis nos acteurs au courant des objectifs de l'étude, et nous avons pris la permission auprès des responsables pour interroger et interviewer les employés et les collaborateurs, nous ferons l'engagement éventuel de communiquer les résultats de l'enquête, et que nous participant par notre étude à la réalisation d'une stratégie RH et d'un plan de recrutement, qui seront opérationnels à la fin de notre travail. Tout au long

[1] Non-Disclosure Agreement

CHAPITRE IV : RESULTATS ET DISCUSSIONS

1. Présentation des résultats :

Ce chapitre sera consacré à la présentation des résultats de notre étude et à la réponse à notre problématique. Ainsi la 1re partie du chapitre traite les résultats des entretiens et présente une retranscription de notre observation participante durant les trois mois passés au sein de l'entreprise Big mama Technology, la seconde partie sera consacré aux commentaires et discussion autour des résultats. Et enfin nous clôturerons ce chapitre par les problèmes et difficultés que rencontrent adressées par le métier.

1.1. Description des interviewés :

Notre question de recherche porte sur les facteurs clés de succès, constituant du métier de *data scientist.*

La liste des interviewés est constitué de quatre (04) data scientists, dont un *team leader* (CTO[1]). Nous les présentons dans le tableau ci-dessous :

Tableau 11 : Présentation des interviewés

N°	Présentation du profil	Durée de l'entretien
1	CTO, et co-fondateur de l'entreprise, titulaire d'un master II en probabilités et modèle aléatoire LPMA[2] faculté de Paris, data scientist depuis 2014, sa formation en mathématique, et sa formation parallèle en informatique lui ont permises dès son master I de se spécialiser en data science.	30 mn
2	Data scientist chez Big mama technology depuis le 03 avril 2016, expert en analyse exploratoire et Machine Learning. Titulaire d'un master en Intelligence Artificielle de l'USTHB, son projet de fin d'étude, et ses deux mois de phase d'intégration chez Big mama lui ont permis d'être data scientist niveau intermédiaire.	1 heure
3	Data scientist chez Big mama technology depuis le 21 Mai 2016, Expert en apprentissage statistique et technology Hadoop. Ingénieur en système informatique issus de l'Ecole Nationale d'Informatique, son projet de fin d'étude, et sa période d'intégration lui ont permis d'être data scientist depuis le mois d'Août 2016	35 mn
4	Data scientist junior chez Big mama technology depuis le 11 décembre 2016, expert en Deep Learning. Il est titulaire d'un master en Intelligence Artificielle de l'USTHB, ses qualités en programmation et son projet de fin d'étude lui ont permis d'être data scientist junior.	1 heure 30 minutes

Source : Étude Réalisée par nos soins

2. Traitement et analyse des données :

Pour le traitement de notre matériau, nous avons procédé par une analyse qualitative du contenu[3],

2.1. Compétences

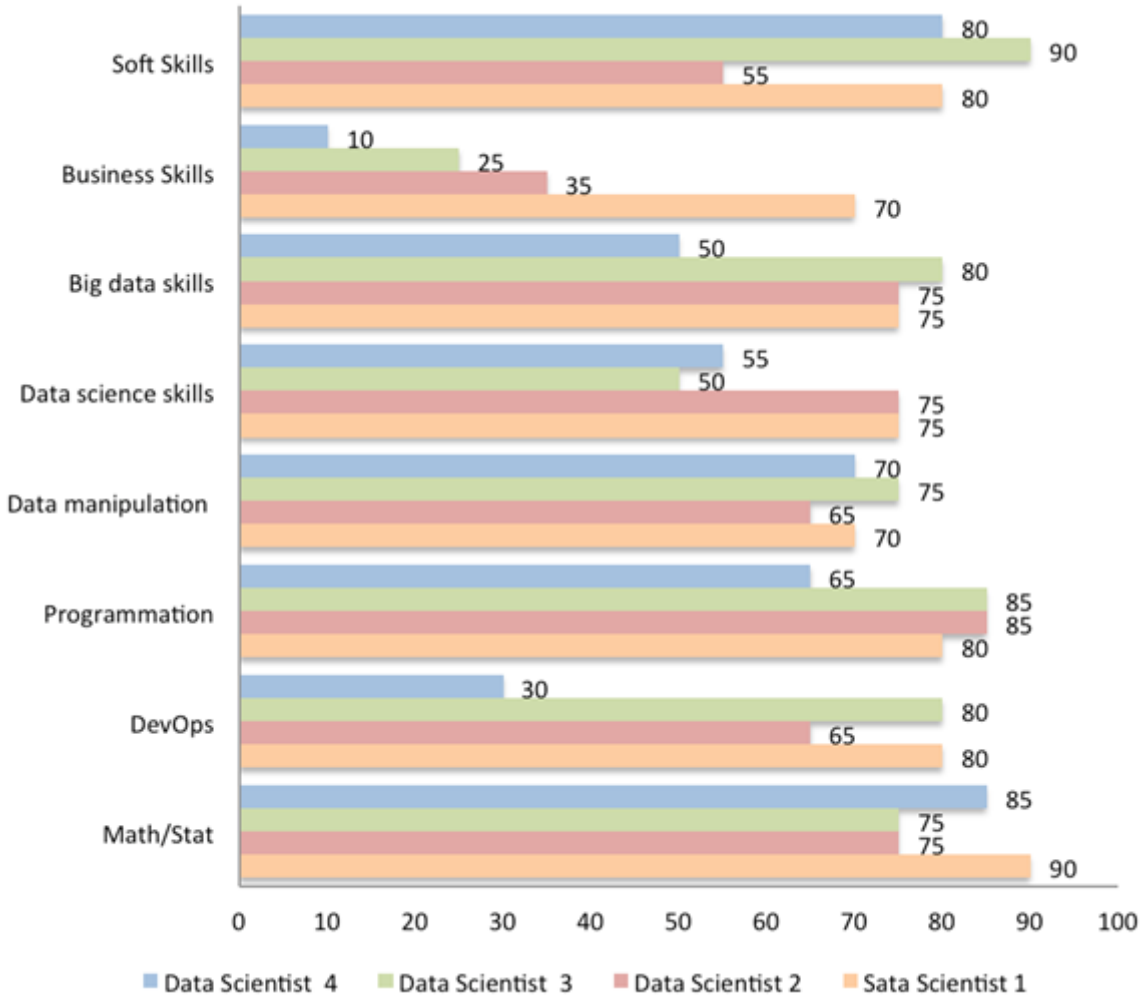
Pour commencer, la 1ére série d'entretiens, portaient sur la thématique des compétences. Nous avons demandé aux data scientists de classer un ensemble diversifié de compétences de la '' skills grid[4]'', qui comporte un ensemble de 37 compétences génériques qui selon

[1] Cheif Technology Officer (directeur technique)
[2] Laboratoire de probabilités et modèles aléatoires
[3] Les détails de cette démarche sont explicités dans le chapitre méthodologie
[4] Skills grid, est une grille constituant l'ensemble de compétences d'un data scientist, tiré de l'état de l'art

nous, couvrait l'éventail des compétences d'un data scientist. Pour quantifier les compétences, nous leur avons demandé de classer sur une échelle de 1 à 5, à quel point ils étaient forts en Mathématiques, Statistiques, Algorithmique… Ensuite, pour analyser les réponses, nous avons regroupé les compétences en 8 catégories, chaque catégorisation s'est faite, par apport aux compétences fortement liées,

Les résultats de cette série d'entretiens sont représentés dans le graphique ci-dessous :

Figure 19 : Résultats de la classification des compétences de data scientists



Source : Étude réalisée par nos soins grâce à la skills grid

Les histogrammes, ci-dessus représentent la classification donnée par chaque data scientist sur ses compétences. Dans le but d'illustrer et de comparer les résultats de chacun, la classification sur cinq a été convertie en pourcentage. Comme le montre les barres, le niveau des data scientist, n'est pas le même, il y a quand même, une grande variabilité dans le classement de leurs compétences.

Pour les « business skills » ou les compétences business et stratégie, le data scientist 1, qui est le data scientist team leader réalise un score de 70%. Dans un contexte big data, les décisions sont pilotées par la stratégie et lorsqu'on est data scientist et team leader, les

décisions qui sont prises quant à la construction des modèles, oùchaque tâche doit impérativement répondre à la meilleure solution business, ou plus concrètement, savoir trouver la solution, la plus profitable. Il réalise également mais aussi un total de score de 90% en mathématiques, statistiques, probabilités, et algorithmique, dû à sa formation universitaire en mathématique.

Le data scientist 3, réalise un score de 80% en compétences de technologie big data, quant au data scientist 2, se classe au même niveau ou presque avec le data scientist 3, et ce à plusieurs compétences, (mathématiques, data science, big data), ceci s'explique par les profils des deux data scientist, qui ont déjà une année d'expérience au sein de l'entreprise, et ont un niveau de data scientist intermédiaire. Le data scientist 4 quant à lui, est classé dernier, celui-ci n'a que quelques mois d'expérience au sein de l'entreprise, il est data scientist junior depuis trois mois.

Cependant, les résultats de cette classification ne sont pas suffisants, pour répondre à notre problématique, les data scientists se sont classé eux même, les réponses sont ainsi plus subjectives. Donc, pour donner plus de poids à notre enquête et pour expliquer les variabilités, dans les niveaux de chaque data scientist, nous menons une 2^{éme} série d'entretiens portées sur la formation et l'évolution des data scientist au sein de l'entreprise, cela va nous permettre d'expliquer les résultats obtenus plus haut[1].

2.2. Formation et évolution :

Dans le but de mesurer l'évolution des data scientists et de comprendre davantage la classification de leurs compétences obtenues suite à notre première enquête, nous menons une deuxième série d'entretiens sur leurs parcours scolaire et professionnel, d'où viennent-ils[2] ? Quels sont les compétences et les savoirs, qui leurs ont permis de s'orienter vers la data science et surtout comment ont-ils évolué durant leurs parcours professionnels, toutes les réponses à ces questions sont apportées plus bas.

La 1ère série d'entretiens nous a déjà donnée un petit aperçu sur les points forts et les points faibles de chacun, croisés avec leurs parcours scolaires, nous mettons au point une fiche d'identité de chaque data scientist :

[1] L'ensemble de questions d'enquête porté sur les compétences et la formation des data scientist, se sont prêtés à une analyse de regroupement.

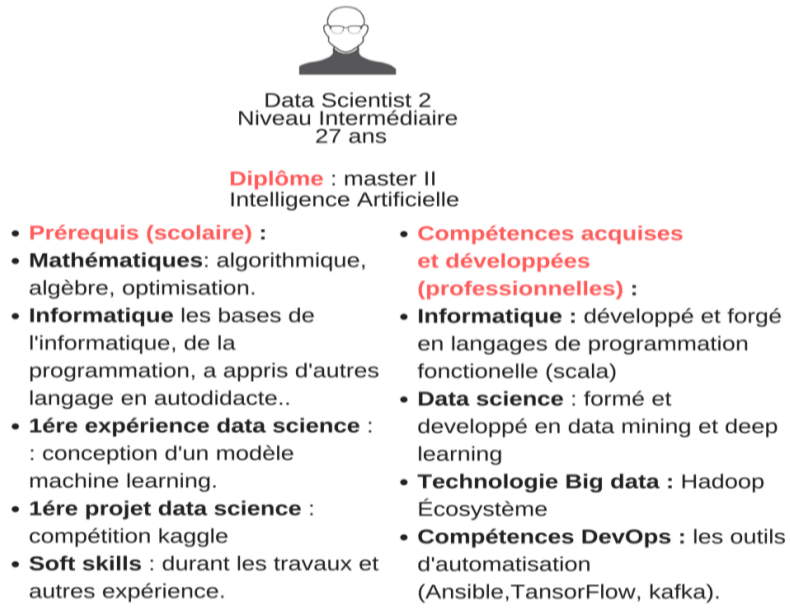
[2] Le guide d'entretien pour cette deuxième série, est ajouté en annexe

Figure 20 : Fiche d'identité data scientist 1 ''Skills ID''

Source : Fiche réalisée par nos soins, en utilisant l'outil graphique Canva

Comme le montre l'illustration, le data scientist 1 est de formation mathématique appliquée, ce cursus offre des connaissances en mathématiques très proches et liés à la data science tels que les méthodes de l'analyse statistique et algorithmique (datamining, classification, ..) notamment celles spécifiques au big data (Machine Learning, les réseaux de neurones), ainsi que les méthodes d'optimisation, modélisation aléatoire etc... Toutes ces connaissances mathématiques sont indispensables à la data science. De plus, pour être un data scientist efficient, une formation en informatique (programmation, administration système), ainsi qu'en business et stratégie est requise.

Figure 21 : Fiche identité data scientist 2 ''Skills ID''



Source : Fiche réalisée par nos soins, en utilisant l'outil Canva

Le data scientist 2 est titulaire d'un master en Système Informatique Intelligent (Intelligence Artificielle). La data science est l'une des branches de l'intelligence artificielle, la formation offre un socle de connaissances et de compétences en data mining, en résolution de problèmes combinatoires en apprentissage automatique, réseau de neurones ainsi que des bases en informatique, mathématiques et probabilités. Grâce à sa formation et sa 1re expérience en machine learning durant son projet de fin d'étude, le data scientist 2, est aujourd'hui un data scientist de niveau intermédiaire.

Figure 22 : Fiche d'identité data scientist 3 ''Skills ID''

Data Scientist 3
Niveau Intermédiaire
26 ans

Diplôme : ingénieur d'état en informatique

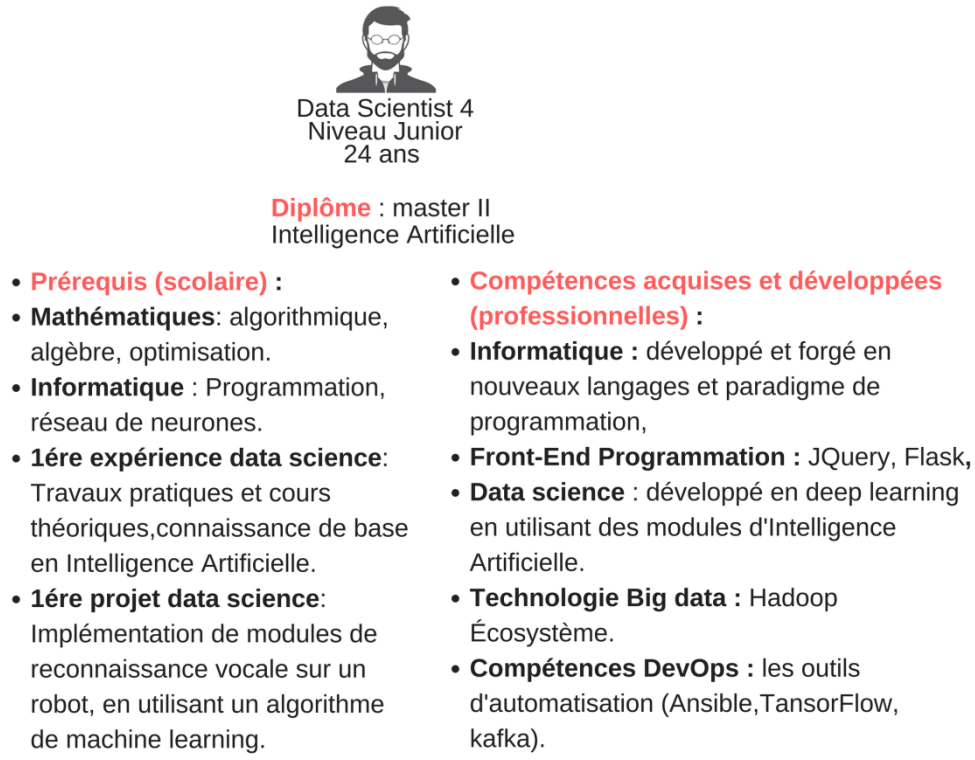
- **Prérequis (scolaire) :**
- **Mathématiques**: algorithmique, algèbre, optimisation.
- **Informatique** : connaissance de base de l'informatique, les systèmes, réseau, analyse de donnée.
- **1ére expérience data science**: Projet de fin d'études, en utilisant des algorithme de machine learning.
- **1ére projet data science**: Conception d'un algorithme de classification.

- **Compétences acquises et développées (professionnelles) :**
- **Informatique :** développé et forgé en nouveaux langages et paradigme de programmation (scala, AKA),
- **Data science** : développé apprentissage statistique et machine.
- **Technologie Big data :** Hadoop Écosystème.
- **Compétences DevOps :** les outils d'automatisation (Ansible,TansorFlow, kafka).
- **Soft skills** : gestion et optimisation du temps.

Source : Fiche réalisée par nos soins en utilisant l'outil graphique Canva

Le data scientist 3, quant à lui, est issu d'un parcours d'ingénieur, titulaire d'un diplôme d'ingénieur d'état en informatique option systèmes informatiques, son cycle préparatoire, lui a permis d'apprendre toutes les bases des sciences d'ingénieurs (mathématiques, statistiques, probabilités…), s'est ensuite spécialisé en système informatique (système d'exploitation, système distribué, analyse de données, réseau..). Il a eu sa 1erexpérience data science lors de son projet de fin d'étude, en développant un algorithme de détection d'attaques informatiques. Son parcours scolaire, et son expérience en data science lui conferent un profil de data scientist assez polyvalent.

Figure 23 : Fiche d'identité data scientist 4 ''Skills ID''



Source : Fiche réalisée par nos soins en utilisant l'outil Canva

Le data scientist 4 est titulaire d'un master en système d'Informatique Intelligent. Sa formation lui a permis d'acquérir des connaissances en matière de machine learning et des réseaux de neurones. Suite à sa première expérience, en créant un modèle de reconnaissance vocale, grâce à un algorithme de machine learning et de ses compétences en programmation, le data scientist 4 a pu facilement rejoindre et s'intégrer dans une équipe de data scientist déjà formée.

Ces fiches d'identités nous donnent déjà une idée sur les éventuelles formations, qui convergent vers la data science, mais surtout sur les différentes compétences que l'on retrouve chez un data scientist. Comme nous l'avons déjà cité dans les chapitres précédents, les profils de data scientists peuvent varier, car jusqu'à présent, il n'existe pas de définition claire des compétences et rôles d'un data scientist, ni de formation académique spécialisée ou orientée data science, ni de modèles d'apprentissage comme l'affirme Lei Liu and al. ''Un consensus commun sur le titre, la définition ou la carrière d'un scientifique de données est absente. "Il n'existe pas de structure de carrière définie pour les scientifiques de données, et c'est un problème majeur qui doit être résolu..." (Swan, 2008, p. 19) La data science traverse de nombreuses disciplines et fait face au défi d'exiger des talents interdisciplinaires, des spécialistes data et des experts dans différents domaines, doivent se transformer facilement en data scientist tout en restant fidèles à leur domaine d'études, physicien, chimiste, informaticien, etc. (Lei Liu1 et al., 2009, p. 206)[1].

[1] Traduit de l'anglais par nos soins, à l'aide de Google Translator

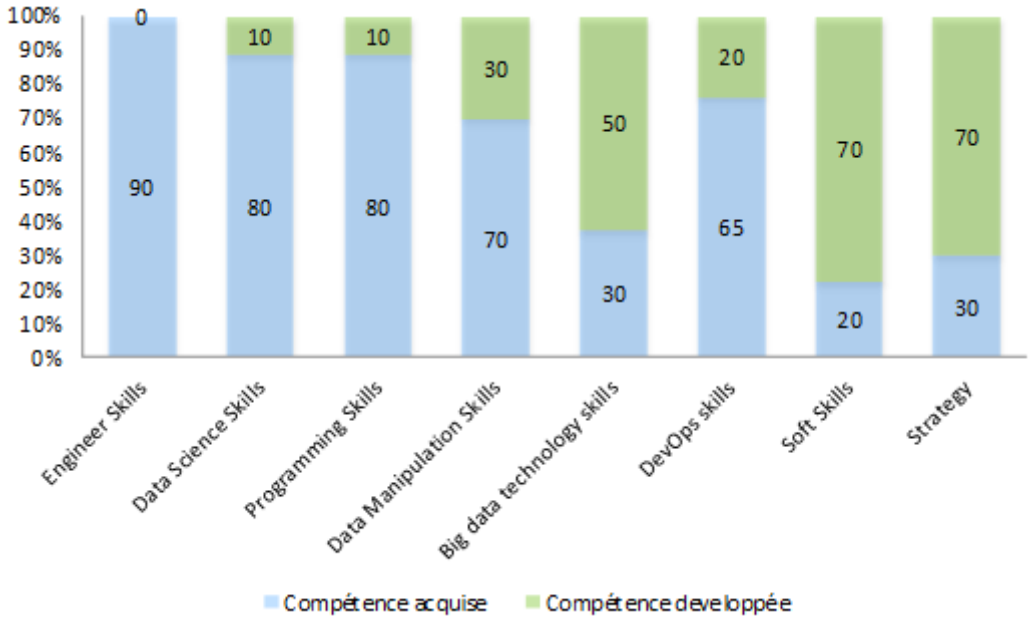
4.3.7

2.2.1. Evolution des compétences des data scientists

Après avoir eu un aperçu clair des compétences de nos data scientists, nous allons à présent, présenter les résultats de la série d'entretiens d'évolution de leurs compétences.

Les résultats de la ''skills grid'' nous ont permis de croiser les compétences actuelles des data scientist (les compétences acquises jusqu'à aujourd'hui), avec leurs prérequis, les compétences et savoirs acquis durant leurs parcours scolaires, ou expériences précédentes, et mesurer l'évolution des profils au long des projets, à leurs confrontations à l'environnement des big data et à leurs interactions et collaborations. Il est à noter que suite à leurs recrutements, les data scientists passent par une période d'intégration allant de un (01) à six (06) mois (la période est limité, selon l'urgence des projets en cours, et du niveau du recrues). Durant cette période d'intégration, les data scientists, passent des concours de type Kaggle[1], où améliorent des scores de prédiction d'algorithmes des anciens projets.

Figure 24 : Grille d'évolution data scientist 1



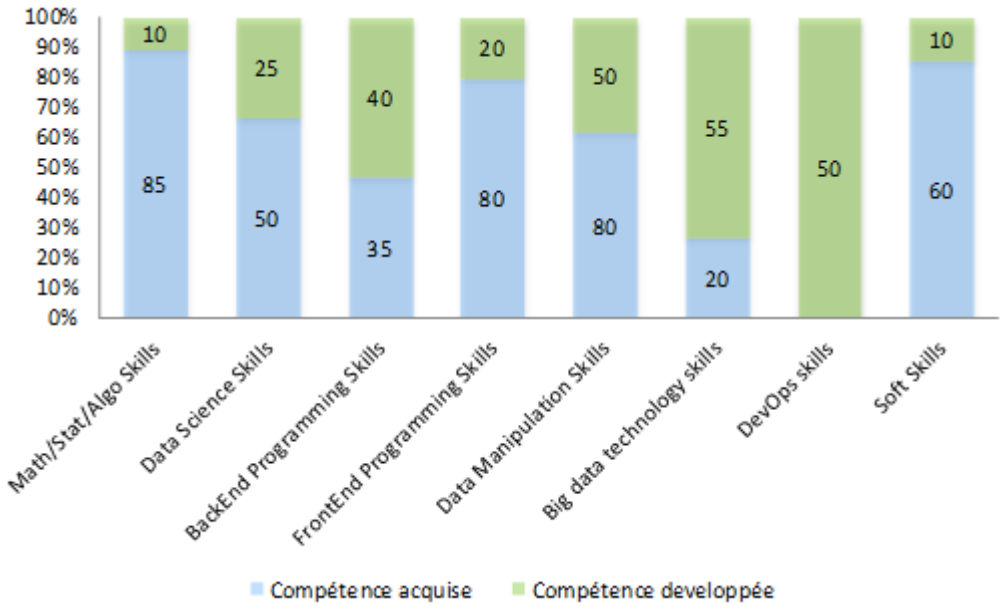
Source : Etude réalisée par nos soins à l'aide des résultats des entretiens

Comme nous l'avons mentionné sur les fiches d'identité «Skills ID» la formation du data scientist 1, lui a permis d'avoir très tôt de solides connaissances en mathématique et data science, les barres affichent des score de 90% en mathématique et 80% en data science. Aussi étant CTO et team leader, le data scientist 1 au-delà de ses compétences techniques, a d'autres missions et rôles à accomplir, liées à la compréhension des enjeux business et marketing de chaque solution algorithmique délivrée, ainsi que la communication au sein de

[1] Plateformes regroupant des challenges en data science proposée par des entreprises et instituts de recherche.

l'équipe pour coordonnées et chapoter les travaux, également, des missions de communication et présentations des solutions auprès des clients, des partenaires et investisseurs, ces missions lui ont permis de se développer en soft skills et stratégie.

Grille d'évolution data scientist 2

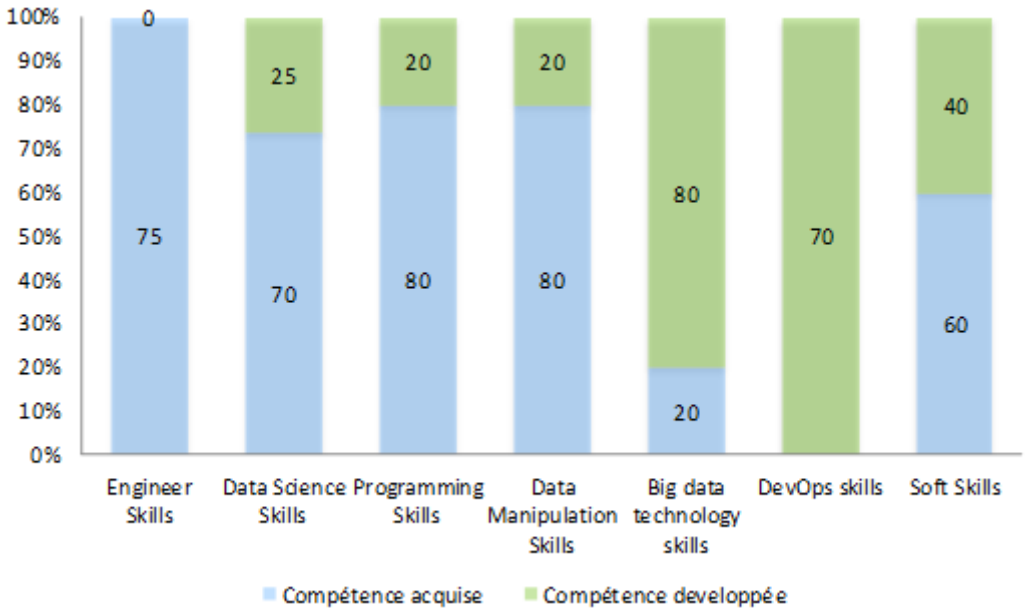


Source : Etude réalisé par nos soins, en utilisant l'outil graphique Word

Sur ce graphique, les évolutions les plus importantes que l'on constate sont au niveau des technologies big data et compétences *DevOps*[1], le data scientist en question avait eu deux mois de période d'intégration, où, il a développé D'avantage ses compétences en data science et en stratégie. Une fois la période d'intégration achevée, le data scientist était face à son premier projet et les premières étapes étaient la préparation de l'environnement, déployer les outils et les technologies nécessaire, pour crée un environnement favorable, au traitement de données massives. N'ayant eu aucune expérience précédente dans l'administration, le déploiement et l'automatisation de ces outils, mais ayant le socle nécessaire pour mener ce genre d'opérations (qui est le langage de programmation python) le data scientist s'est très vite familiarisé avec ce genre de pratiques et continue à se perfectionner au fil des projets.

[1] Ensemble de compétences, d'outils et de pratiques visant à automatisé et aligner l'ensemble des outils et des infrastructures (exploitation, administration système, réseau, bases de données), exemple : Ansible, Vagrant

Grille d'évolution data scientist 3

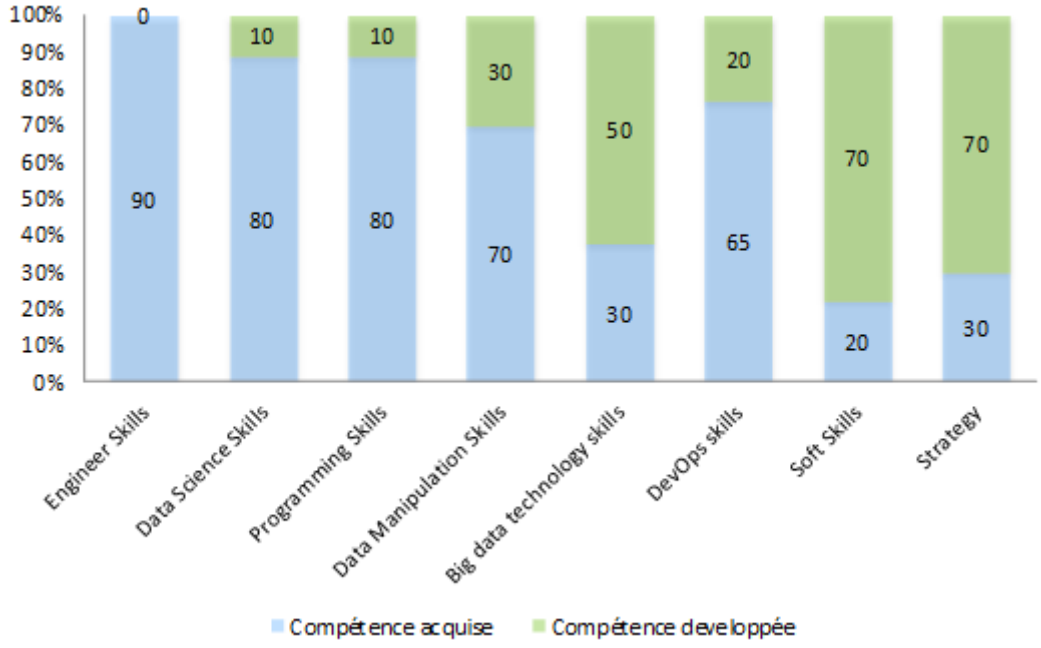


Source : Etude réalisé par nos soins, en utilisant Word

Le data scientist 3, est issu d'un parcours d'ingénieur en informatique, sa formation en système informatique lui à permis de développer des compétences solides en programmation et en opérations d'administration systèmes réseaux, de supervision, de sécurisation, mais aussi de mise en place des déploiements matériels et logiciels. Toutes ces connaissances lui ont permis de se familiariser et de se développer rapidement dans le déploiement et l'administration des outils de l'écosystème Hadoop, de l'automatisation et de gestion multi-nœuds[1], ''*j'ai un background d'ingénieurs, cela m'as permis de connaître les bases de la programmation. Je me documentais sur les fonctions que sur lesquelles je bloquais. C'était aussi la même chose pr les technology, le fais d'avoir un minimum de connaissances sur l'architecture d'une techno, j'apprenais les modules petit à petit au fil des projets* (Data scientist 3, 2017)''.

[1] Mode qui permet la répartition le stockage et le traitement des données volumineuses sur plusieurs machines.

Grille d'évolution data scientist 4



Source : Etude réalisée par nos soins en utilisant Word

Le data scientist 4 n'a que quatre mois d'expérience chez l'entreprise big mama technology, son profil est pour l'instant junior. Le 1er mois du data scientist 4 au sein de l'entreprise, était consacré à sa formation. Le data scientist, étant d'une formation en intelligence artificielle il disposait déjà des compétences nécessaires, pour mener un projet de data science, mais celui-ci n'avais aucune expérience, en big data, il n'avait jamais manipulé ou administrer des outils de calcul sur les big data. Sa formation et son évolution étaient axés spécialement sur ces technologies là (Hadoop écosystème, stockage et traitement parallélisé (MapReduce), Spark, et autres outils de manipulation big data[1])''les difficultés que j'ai eu étaient liées principalement à mes lacunes en technologie big data, lors de mon premier projet, en essayant de résoudre un problème, j'ai failli faire couler le serveur, car je n'avais jamais eu d'expérience, avec un aussi grand volume de données et il fallait vraiment faire attention aux performances de l'algorithme. J'ai résolu ça en me documentant et en faisant plusieurs essaies (Data scientist 4, 2017)''

2.3. Projets :

Cette troisième série d'entretiens, est dédiée principalement aux projets, Le but est de les examiner de façon approfondie à partir de l'analyse des entretiens. Les deux dernières séries d'entretiens portaient sur les compétences des data scientists, nous nous intéressons dans cette partie-là, à la mise en pratique de ces compétences-là, de leurs rôles et des facteurs clés de succès à l'aboutissement des projets Big data :

[1] Les outils et technologies cité, sont expliqués dans le glossaire, ajouté en annexe

Une étude de terrain à partir des trois derniers projets au sein de l'entreprise Big mama, fait ressortir un certain nombre de facteurs importants dans le succès du métier. Les facteurs que nous avons tirés sont regroupés en deux catégories à savoir :

- Need to have
- Nice to have

Need to have : ce sont les facteurs clés de succès, jugé nécessaires et indispensable au métier,

Nice to have : ce sont les facteurs clés de succès, jugés optionnels, mais qui sont cependant importants.

2.3.1. Les facteurs clés de succès (Need to have)

1. *Connaissance approfondie en mathématique, statistiques, probabilité et algorithmique:*

Comprendre le monde à travers les données, impliquer le tout "de la formulation du problème à ses conclusions " (Wild CJ, 1999, p. 1).

Les data scientists doivent maîtriser les théories statistiques de base : collecte de données, modélisation. Ils doivent formuler le problème, l'analyser et trouver les limites de leurs modèles, mais aussi recueillir puis analyser les données pour fournir des informations, trouver des corrélations, etc. Mais également, la perspective algorithmique de l'apprentissage et la perspective prédictive de la pensée statistique. L'accent est mis sur les méthodes communes d'apprentissage du machine/ stastical learning et deep learning (apprentissage machine et statistique et profond). Toutes ses compétences statistiques s'appuyant sur des fondements mathématiques et informatiques.

Les mathématiques : les data scientist doivent comprendre la structure sous-jacente des modèles communs utilisés en statistique et en machine learning et résoudre des problèmes d'optimisation en mettant en pratique des outils qui incluent le calcul, l'algèbre linéaire, les probabilités et l'algorithmique (méthode de gradient, les algorithmes d'arbre de décision), etc...

L'algorithmique : Les data scientists utilisent des compétences algorithmiques pour la résolution de problèmes. L'un des facteurs clés de succès d'un data scientist est de traduire algorithmique du problème à résoudre, Cela inclut la définition d'exigences claires à un problème, la décomposition du problème, et enfin solutionner le problème par programmation dans un langage appropriée.

Tout ceci montre à quel point la data science est intrinsèquement interdisciplinaire et savoir combiner l'ensemble des compétences, pour solutionner un problème de data science est un facteur clés de succès primordial.

2. *Mastering machine learning (apprentissage automatique) :*

Le machine learning est l'un des principaux outils d'un data scientist, son but est d'apprendre à une machine à 'apprendre' à résoudre un problème en elle-même, elle apprend à identifier des caractéristiques sur la donnée, pour lui permettre d'établir des prédictions. Tout ceci se fait en recevant un ensemble de données du client, puis utiliser ces données pour répondre à sa

problématique spécifique. Cette pratique se trouve, donc, au cœur du métier du data scientist, car, d'une part, elle combine des approches statistiques et mathématiques se basant sur des approches d'arbres de décision, régression linéaire, réseaux neuronaux, réseaux bayésiens, machines à vecteurs… etc. et d'autres part un algorithme de machine learning est caractérisé par sa performance (capacité de prédire). L'un des facteurs clés de succès d'un data scientist, est de savoir évaluer tout algorithme d'apprentissage sur son jeu de données en équilibrant au mieux le biais/variances du sur/sous-apprentissage; le data scientist doit pouvoir trouver l'équilibre parfait pour que l'algorithme puisse atteindre le meilleur score en performance, en d'autres termes, afin d'être capable de bien prédire.

3. *Compétences poussées en informatique (DeVops, Big data) :*

Travailler avec des données nécessite des compétences poussées en informatique, les data scientists doivent se doter d'outils et de compétences pour faire face à la jungle de Hadoop et de son écosystème qui nécessite la maîtrise des technologies de pointe, en stockage et distribution parallélisée, (utilisation de systèmes distribués et à la mise en œuvre de méthode de calcul haute performance). L'utilisation de ces technologies nécessite des compétences de devOps, le data scientist travaille sur des machines virtuelles où il doit déployer et administrer tout l'écosystème Hadoop sous MapReduce[1], ensuite configurer et automatiser les flux entre les différents nœuds. Ils doivent également s'armer de compétences en logiciels et algorithmes associés avec une compréhension des principes de programmation pour pouvoir programmer dans plusieurs langages et dans différentes technologie

4. *Connaissance métier et vision business :*

Les données n'existent pas dans le vide, mais proviennent d'un contexte particulier. La connaissance de ce contexte est nécessaire pour analyser les données, les data scientists, reçoivent différents clients issus de différents domaines d'activités : assurance, banque, télécommunication, santé, éducation etc… Trouver la bonne solution au client, c'est poser la bonne question business, rattaché au contexte de provenance de la donnée. Avant de procéder à la résolution technique du problème, le data scientist doit répondre au besoin primordial du client, qui est génère du profit (Data Scientist 1;, 2017), Pour cela le data scientist doit connaître le domaine et le secteur du client sur le bout des doigts et prendre en considération tous les enjeux business du domaine pour pouvoir apporter la meilleur solution algorithmique. Cela nécessite une recherche approfondie sur le secteur, sur les éventuelle parties prenantes, les intérêts et les tendances de ce dernier, les chiffres clés et la concurrence direct et indirect dans le secteur.

5. *Collaboration et travail d'équipe :*

La data science est un travail en équipe, l'équipe de data scientist, est formée de profils différents, mais ayant tous le même but, délivrer une solution qui puisse répondre aux attentes du client. L'équipe de data scientists chez Big mama Technology est récente de formation, ses membres ont évolué et progressé au long des projets, et ce suite à la collaboration et le travail d'équipe. Quand le data scientist 2 avait un bug au moment de déployer Hadoop ou Spark, le data scientist 3 intervenait, pour lui débugger son code. Quand l'équipe est confrontée à une

[1] Définit dans le glossaire ajouté en annexe

nouvelle technologie le data scientist 1 (team leader) apporte la documentation et trouve les cours nécessaires pour son équipe. De plus, plusieurs formations en interne avaient eu lieu, chacun intervenait dans chaque phase du projet de l'autre, l'espace et la structure organisationnelle de l'entreprise favorisait cet environnement. Mis à part l'environnement favorable à la collaboration et les nouvelles technologies de collaboration telles que Slack et Git nous jugeons que savoir collaborer est l'un des facteurs clés de succès prédominant de la réussite dans le métier.

6. *Innovation :*

Nous jugeons que lorsque l'on est dans un domaine de technologie de pointe, comme les big data, l'innovation est un facteur clés de succès décisif, que ça soit l'innovation dans les modèles ou dans les méthodes d'analyse. Les data scientists doivent, d'une part être à la pointe de tous les progrès et innovations dans le secteur des big data et, d'autres part, faire eux même preuve d'innovation intellectuelle, durant les trois projets auxquels les data scientist étaient confrontés. L'innovation dans les méthodes d'analyser, d'interpréter, d'améliorer les modèles, fut l'un de leur meilleur allié. Dans un métier aussi complexe, les data scientist sont en perpétuelle confrontations à des problématiques d'ordre technique, qu'il s'agisse des contraintes de calculs haute performance, ou de manque de donnée. Pour avoir le meilleur score, les data scientist font preuve d'innovation pour s'extirper de la meilleure manière possible.

Durant le projet B, les data scientists étaient confrontés à un problème de performance, auquel ils devaient faire face. Pour se mettre dans le contexte, les data scientists traitent tout type de données, (text, image, vidéos, bande sonores) chaque variété de données, nécessitant des méthodologies et des technologies spécifiques quant à son traitement, dans le cas du projet B, les data scientists étaient confrontés à des bandes sonores qui étaient très gourmandes en ressources, les data scientists devaient trouver la meilleurs solution pour traiter et extraire des informations pertinentes de ses bandes sonores sans affecter les performances du modèle. Pour ce faire, ils avaient converti les bandes sonores en spectrogrammes et utilisé des techniques de traitement d'images pour extraire les caractéristiques des bandes sonores. *On avait un compromis entre les qualités des informations qu'on peut extraire d'une bonde sonore et les performances de calcul, pour extraire cette information et permettre à l'algorithme de faire des prédiction en temps réel. On a cherché la meilleur façon de ça, en construisant plusieurs modèles, qui pouvaient nous permettre d'avoir les deux (extraire les meilleurs informations, ne pas consommer beaucoup de ressources)* (Data Scientist 4,, 2017).

7. *Ouvert et prêt à la formation et au développement*

Les data scientists sont confrontés, dans leurs quotidiens à plusieurs difficultés et contraintes techniques posé par le traitement et le calcul à haute performance. Pour surmonter ces difficultés, les data scientists, passent une grande partie de leur temps à se documenter, et à apprendre, en testant de nouvelles librairies, en implémentant de nouveaux modules, apprendre de nouvelles technologies, un nouveau langage, un nouveau paradigme de programmation etc....Lva vitesses à laquelle les technologies évoluait, les obligeais à être à jour. Le but était d'apprendre vite et de manière efficace. Les data scientists de l'entreprise Big mama Technology n'avaient jamais travaillé auparavant avec des technologies big data,

mais leur capacité à apprendre et à se développer, était l'un de leurs facteurs clés de succès ''*La première phase du projet A était de déployer une plateforme Hadoop, pour distribuer la donnée et lancer les tâches de data science, c'était la première fois que je faisais ça, je ne maitrisais pas ses outils auparavant, mais je me formais au fur et à mesure, à chaque étape du projet*.'' (Data scientist 3, 2017), '' *Durant mon travail sur le projet B, il y avait déjà un prototype, qui était fait avec un algorithme de foret aléatoire. Durant mon travaille, j'ai Adopté une approche deep learning, en utilisant des algorithmes de réseaux de neurones, je me suis formé et j'attendais de passer sur un problème concret* (Data Scientist 4,, 2017)*''*

8. *Battre tous les scores :*

Le travail d'un data scientist est de développer des modèles prédictifs. Cependant pour être performant un modèle doit avoir le meilleur score en prédiction, Plus concrètement plus le score est élevé, plus le modèle fait de bonnes prédictions, et plus il sera susceptible de répondre aux attentes du client et le satisfaire. Les data scientists doivent faire un travail en amont pour chercher et voir l'état de l'art d'un modèle et des efforts en aval, pour améliorer les performances d'un modèle, en atteignant un score supérieur ou égal à celui de l'état de l'art. Des protypes de modèles prédictifs, sont disponibles en open source, avec un score de base. Pour s'entrainer les data scientists, travaillaient sur l'amélioration de ces scores, car délivrer un modèle, qui puisse avoir le meilleur score en prédiction est un facteur clé de succès colossale pour un data scientist. Les enjeux sont avant tout économiques et la survie de l'entreprise en dépend. La finalité d'un modèle prédictif est commerciale, et la satisfaction client est l'une des valeurs primordiales de l'entreprise Big mama Technology.

Nous jugeons que ces sept facteurs cités sont des facteurs clés de succès indispensables à la réussite dans le métier de data scientist. Nous passons à présent des facteurs optionnels (préférables) à la réussite et la pérennité du métier.

9. *Polyvalence :*

Généralement dans un département data, où les projets big data sont pilotés par les DSI, les données sont stockées et traitées en interne. Chapeauté par le Data Office Manager et entourés par des compétences techniques, le data scientist exécute de simples tâches de manipulation de données. Le travail dans ces départements est réparti selon la structure matricielle de la DSI, sur l'ensemble des profils : l'architecte big data gère l'installation de Hadoop sur les différentes machines, le *devOps* big data s'appuiera ensuite, sur ce nouvel environnement mis à sa disposition par l'architecte pour commencer à configurer, administrer, et déployer les outils et les automatiser (Pig, Hive, Json, Ansible, LogStach). Une fois tout cela est mis en place, le du data scientist, intervient pour accomplir des tâches de data science sur l'environnement mis en place par l'architecte et le DevOps, ensuite le *data visualizer*, met en place des interfaces pour la visualisation et la communication des résultats de son équipe.
Dans d'autres cas où l'entreprise est un laboratoire de big data externalisé, ou tout simplement une entreprise se lançant dans un projet big data, la première contrainte à laquelle l'entreprise fait face est le manque de compétences, ou de spécialistes et le coût associé au recrutement de ces derniers. Pour y faire face, les profils sélectionnés doivent impérativement faire preuve de polyvalence, ou bien construire des équipes polyvalentes, en les formant et les initiant à plusieurs technologies et pratiques techniques. Les data scientists sont souvent amenés à

occuper plusieurs fonctions. La finalité de constituer des équipes aussi polyvalentes est l'optimisation du temps, et d'erreurs. Les équipes sont ainsi plus efficaces. Imaginons que le data scientist doit attendre que le devOps big data déploie les outils nécessaires, le devOps doit attendre le data architecte pour lui préparer les machines…etc, cela ferait perdre énormément de temps et d'efficacité à l'entreprise. La polyvalence est un facteur clés de succès crucial pour un data scientist. Il est ainsi plus productif et beaucoup plus accessible en terme de budget (un data scientist polyvalent coute forcément moins cher que plusieurs spécialistes). Par ce dernier facteur nous clôturons notre liste des facteurs clés de succès nice to have, optionnel, mais tout de même important pour la réussite dans le métier de data scientist.

2.3.2. Nice to have :

10. Des perspectives orientées recherche et développement :

Un data scientist est censé apporter des solutions aux problématiques posées par les clients. Souvent, face à l'immaturité du secteur des big data, les clients ont des problématiques qui ne s'avent pas posé, ou des données et qui ne s'avent quoi en faire, ni comment valoriser ces donnée, dans ces cas-là, l'équipe des data scientists, avec des connaissances métier, propose des axes de recherche, à partir du contexte de la provenance de la donnée, associé au secteur et domaine d'activité du client. Après l'acquisition de la donnée, les data scientists se transforment en vrais chercheurs d'or, ils regardent la donnée, et par intuition, proposent les meilleures perspectives pour le client, en lui proposant des solutions adaptées au marché, et répondant à ses besoins. ''*On doit définir et chercher les variables, il faut observer et transformer la donnée, et voir en quoi cette variable pourrait intéresser le client, cela se fait par intuition, et selon la pertinence de la variable (Data scientist 2,, 2017)*''

Egalement en interne, l'entreprise fait de la veille sur des axes technologiques, observe les tendances, teste la faisabilité des solutions et à partir de là, propose ces solutions aux partenaires et investisseurs.

Qu'il s'agisse de faire de la recherche et développement en interne ou pour les clients, nous considérons que cela fait partie des facteurs clé de succès dans le métier de data scientist.

11. Etre un alchimiste de la donnée :

La data science est la science de transformation de la donnée, nous avons appelé ce facteur clé de succès ainsi, car nous jugeons que le travail d'un data scientist est tel un alchimiste, qui transforme les matériaux en or. La donnée que les data scientists reçoivent des dataset clients contient ce que les data scientists appellent la donnée sale, les valeurs sont aberrante et ne sont pas cohérentes, le contenu est mal formé, il y a des colonnes inutiles… etc. Le rôle du data scientist est la transformation de cette donnée en donnée exploitable, manipulable, pour y extraire les informations utiles et susceptibles de répondre à la a problématique business du client. Ce travail demande beaucoup de patience et de concentration, les données sont analysées par une phase appelé analyse exploratoire, durant laquelle, le data scientist se familiarise avec la donnée, pour passer à l'étape suivante, qui est la transformation de cette donnée, elle est ensuite triée, et organisé dans des tables, pour commencer à extraire les

variables intéressantes et pertinentes, qui pourraient répondre au besoin des clients. ‘‘ *Dans un modèle de machine learning théorique, on a un data set, et les données viennent directement groupées sous forme de tableau. Le tableau est constitué de deux Cologne les X et les Y, les X ce sont les attributs dépendant du contexte, et les Y se sont les variables à prédire. Dans notre cas, c’est à nous de constituée le tableau, à partir des X extrait de l’analyse exploratoire, et à partir de là nous construisons notre modèle pour prédire nos variables Y* ’’ (Data scientist 2,, 2017)

12. Gérer et optimiser son temps :

Au cours des différents projets, les data scientists font face à de nombreuses problématiques liées à la qualité de la donnée, à la complexité de l’environnement big data, aux performances du modèle… etc. Cependant, la plus grande contrainte reste le temps. Les data scientists, étaient conditionné par les dates limite des projets, cette contrainte les poussé à chercher les outils et méthodes qui leur épargnaient de faire certaines tâches de data science manuellement, ‘‘il faut automatiser tout ce qui est automatisable’’ (Data scientist 1;, 2017). Aussi, il a fallu faire des efforts personnels pour apprendre et se familiariser avec une nouvelle technologie avec diligence. Savoir optimiser et gérer son temps de la manière la plus efficiente est un facteur clé de succès crucial à la réussite dans le métier de data scientist. ‘‘*Quand c’était des technologies que je n’avais jamais utilisé, je me documentais avec des livres, j’allais droit au but, et je me familiarisais très vite avec les outils sur lesquelles je travaillais*’’ (Data scientist 3, *2017)*

13. Sortir de sa zone de confort, et se confronter au risque :

Les data scientists travaillent avec une collection diversifiée d’outils. Ils sont en permanence confrontés à apprendre de nouveaux outils et dans certains cas, contribuer au développement de nouveaux outils eux-mêmes. Les environnements informatiques, à la fois logiciels et matériels, changent rapidement et fréquemment et ces changements ont des conséquences sur la structure des données, le stockage des données et le calcul. Toutes ces cogitations procurait une instabilité dans la routine des data scientists, pour cela ils doivent s’armer de patience et être proactifs, toujours ouverts à faire face aux changements. ‘‘*On programmait tous nos projets avec le langage Python, et vu le volume de données qu’on traitait, on commençait à avoir de sérieux problème de performances, le traitement prenais énormément de temps, et on avait une date limite pour la livraison de la solution chez le client, j’ai donné à mon équipe deux jours pour apprendre un nouveau langage qu’ils n’avaient jamais utilisé auparavant, dans les deux jours qui ont suivi, tous nos projets étaient codés dans ce nouveau langage’’* (Data scientist 1; 2017).

Par ce dernier facteur nous clôturons, notre liste des facteurs clés de succès nice to have, optionnel, mais tout de même important à la réussite dans le métier de data scientist.

Nous ne prétendons pas en dresser une liste exhaustive puisque les recherches et expérimentations dans ce domaine sont en perpétuelle évolution. Cependant, nous tenons à rendre compte des éléments essentielles constituant et identifiant le travail d’un data scientist, *dans un environnement de big data.*

2.3.3. **Les contraintes liés au métier de data scientist (contexte big mama):**

Après avoir vu les facteurs clés de succès du métier de data scientist, nous allons dans cette partie exposer brièvement les contraintes liées au métier.

1. *Les big data sont gourmand en ressources :*

L'une des premières contraintes auxquelles les data scientists sont confrontées est le volume des données traités. Les données brutes sur lesquelles s'appuie le big data nécessitent un traitement parfois lourd. Les algorithmes et les croisements, nécessitait une large bande passante[1], plus il y a de la donnée plus cela nécessite une grande capacité de calcul, et cela implique donc une infrastructure et des outils dédiés, autrement dit un engagement coûteux et gourmand en ressources.

2. *L'immaturité du secteur big data en Algérie :*

La stratégie big data est considérée comme le moteur de l'innovation, de la satisfaction client et de la réalisation de plus grandes marges de profits (Barton, 2010). Le secteur des big data fait ses premiers pas en Algérie, la plupart des entreprises restent immatures sur le sujet, et ont une prise de conscience insuffisante sur les défis du big data, et surtout sur les connaissances et pratiques liée au domaine et aux solutions possibles et envisageables dans une conduite de projet big data. Les entreprises ne sont pas aujourd'hui prêtes et équipées des ressources humaines et matérielles suffisantes pour gérer le Big Data et pour penser à des éventuelles perspectives vers ce domaine. Cela influe énormément sur les jeunes entreprises qui se lancent dans ce secteur, telles que Big mama technology qui, pour l'instant opère plus avec des clients étrangers que locaux, De cette immaturité se déclinent deux autres conséquences majeures qui sont détaillé plus bas.

3. *L'incompétence en BI et ses conséquences :*

Comme nous l'avons cité plus haut, chaque projet big data nécessite que l'on repense à l'infrastructure technologique pour que des solutions dédiées à des analyses de données gigantesques et sophistiqués puissent venir s'y greffer (Myriam Karoui, 2012), cette réflexion s'articule sur une stratégie business intelligence au sein de l'entreprise, qui puisse avoir les ressources humaines et matérielles pour collecter, stocker, et organiser la donnée de façon cohérente, et lui donnant une forme plus au moins exploitable. Cependant, face à cette incompétence, les entreprises stockent de la donnée d'une manière très anarchique, ou pire encore, elles ne stockaient tout simplement pas de données. Lorsque ces mêmes entreprises rencontrent des problématiques nécessitant des solutions en Intelligence Artificielle, ou expriment un besoin ou une conscience de l'exploitation de sa donnée, l'entreprise big mama technology fait face, à de nombreuses problématiques, telles que la collecte de la donné, relative au secteur du client, qui souvent ne se rapproche pas de la donnée du monde réel, car cette dernière est collecté à partir de dataset open source trouvé sur le net, ou souvent la donnée reçu est une donnée sale.

[1] Le nombre de Mo ou de kilo Bit pouvant arriver sur une machine (ou en partir).

4. *Le contexte de la culture de l'information :*

L'une des autres problématiques qui découle des deux précédentes, est la culture Algérienne liée à l'information et la crainte de l'effet Big Brother, c'est-à-dire l'utilisation des données personnelles des clients à des fins pas très éthiques, ceci est considéré comme un réel frein à toute perspectives big data.

5. *Le manque de documentation :*

La nouveauté du secteur de la data science, exprime un manque en pratique et documentation, plusieurs problèmes qui sont posés durant les projets, ont du mal à trouver une réponse, que ce soit sur les livres ou sur internet. Très peu de sources traitent les problèmes liés à l'utilisation des nouvelles technologies big data, telles que l'utilisation de l'écosystème Hadoop. Parfois les data scientists se retrouvaient face à un modèle, qui n'as jamais été entrepris avant, il n'y a donc pas l'état de l'art, pas de score à améliorer, pas de librairie à exploiter.

Par ce dernier point, nous clôturons la liste des contraintes liées au métier d'un data scientist, et nous arrivons à la fin des résultats de notre étude.

2.3.4. Synthèse inductive1 :

Sur la base de nos résultats, nous nous proposons de reprendre les modèles présentés dans notre état de l'art de qu'est-ce qu'un data scientist, et quels sont les facteurs constituant le métier, et de les adapter en tenant compte de l'apport d'autres théories et de notre recherche. Par cela nous enrichissons la théorie, en y apportant notre vision.

Nous commençons par la proposition d'une fiche de poste comportant l'ensemble d'éléments nécessaires déterminant les rôles et responsabilités d'un data scientist dans une entreprise.

2.3.5. Fiche métier : data scientist

Objectif d'un data scientist :

- Transformer la data brute en données exploitables pour l'entreprise (avec une vision business à la clé)
- Répondre à la problématique métier de son client, et les traduire en solution et modèle algorithmique
- Concevoir des algorithmes pour savoir quelles données récolter, les méthodes de récolte, leur utilité…Etc

Missions et responsabilités

- État de l'art : Analyse et veille des algorithmes existants (documentation)
- Analyse exploratoire de données : explorer, trier, purger les données clients et y sortir la pertinence

[1] Regarder schéma ajouté en annexe

- Protyper les algorithmes de recherches et de traitement d'information
- Etude de faisabilité : faire des tests à partir d'un échantillon de données client ou de donnée virtuel pour mesurer la performance de l'algorithme
- R&D : chercher et développer les meilleurs solutions et modèles qui répondent aux besoins et au contexte de la donnée du client
- Conception et implémentation des algorithmes
- Test et amélioration continue des performances de l'algorithme
- Suivi technique et maintenance
- Rédaction de documentation de l'algorithme (et éventuellement dépôt de brevet)

Technologies employées :

- Bases de données NoSQL : MongoDB, Hadoop, MapReduce
- Langages : Python, Skala, Go
- Technologies: Back-end Full stack, Front-end, Administration systéme

Environnements et structures d'accueil :

- Startups et grands groupes ayant une data driven stratégie ou data driven objectif
- Service Data Science ou Big Data, rattaché au CTO ...

Études et parcours :

- Bac +5/8, Parcours, statistiques (data maning), mathématiques (modèles aléatoire) et informatique, spécialité Intelligence artificielle (Écoles ou Universités), toutes spécialité informatique ayant un intérêt pour le machine learning

Compétences demandées :

- Mathématiques et algorithmie
- Programmation de base C, java
- Langage de programmation fonctionnelle Python, Scala
- Anglais courant, lu écrit et parlé

Savoir-être requis :

- Passionné par le traitement de l'information et les problématiques Big Data
- Curieux et ouvert d'esprit : Veilleur dans l'âme, prêt à découvrir et envisager des choses inédites
- Art de remise en question

Ainsi, nous espérons que cette fichier métier fasse l'objet d'une sérieuse auprès des entreprises.
Aussi nous ésperons que les écoles de mathématiques, informatiques et statistiques puissent intégrer et orienté les objectives pédagogique pour répondre à ce besoin d'un data scientist, en ajoutant des modules tels que machine learning, analyse exploratoire des données etc..

CONCLUSION

Conclusion :

Le but de notre recherche a été de contribuer à enrichir la théorie par l'identification et la définition du métier de data scientist, ses rôles et ses compétences dans un environnement big data. Les résultats de notre recherche nous ont permis d'obtenir un ensemble de facteurs clés de succès du métier de data scientist, ce qui définissait le cœur de notre problématique. Les résultats obtenus révèlent que l'ensemble des facteurs clés identifiés constituent des conditions nécessaires pour optimiser le succès d'un data scientist au sein des entreprises.

Cependant, on peut tirer de nombreux enseignements de cette réflexion. Les big data demeurent un facteur important à l'innovation et pour faire face à la concurrence. Selon une étude de *Markets and Markets*, le marché du Big Data progresse annuellement de 26% et atteindra 46 milliards de dollars en 2018. Nous soulignons ainsi l'urgence pour les entreprises et les institutions publiques ou privées de prendre conscience de ce phénomène. En Algérie, la concurrence se fait de plus en plus rude et les entreprises grandes comme petites ne, sont pas épargnées et le passage vers une stratégie big data devient presque obligatoire pour assurer la survie et la pérennité ces entreprises.

En effet, les big data demandent des big compétences et de big solutions technologiques. Nous jugeons donc, que le métier de data scientist est aujourd'hui d'une importance capitale. Face à l'immaturité du secteur big data en Algérie, les entreprises ont beaucoup de retard à rattraper. Ainsi, elles doivent préparer leur environnement interne à une conduite vers une stratégie big data, en préparant les compétences nécessaires qui se doivent d'être sensibles à la dimension technologique mais aussi managériale des données afin d'être créatives dans le traitement et l'analyse de cette masse de données.

Avant de conclure, nous voudrions relever les limites de notre démarche. Ce travail de recherche a été pour nous comme une série d'aventures. Le choix de la thématique a été la première étape. Etant passionné par l'informatique, et cherchant toujours à relever les défis de nous améliorer dans le domaine, partir à la quête du métier de data scientist était pour nous la meilleur occasion d'élargir nos champs de compétences. Mais, sur le plan théorique, tout était à découvrir. Nous nous sommes lancées dans l'aventure, pour décoder ce que nos acteurs disaient et faisaient au quotidien. La démarche qualitative par entretien a été notre deuxième défi, le but a été d'effectuer un travail de qualité dans un souci de rigueur scientifique. En tant que novice en la matière, nous avons pu constater que rien n'est jamais aussi simple que dans les livres ! L'analyse de nos données a été notre dernière aventure. C'était sans aucun doute la plus longue et difficile, mais également la plus intéressante et enrichissante.

Perspectives :

La validation des FCS identifiés pourrait se faire auprès d'une plus large population de data scientists et dans un contexte hors de l'entreprise en question qui a, certes, ses propres spécificités.

Ces FCS pourraient faire l'objet d'expérimentation en Algérie et servir de feuille de route dans le cadre d'une éventuelle intégration d'un projet big data au sein d'une entreprise.

REFERENCES BIBLIOGRAPHIQUES

Bibliographie

A.Cooper. (2012). Analytics and big data—reflections from the Teradata Universe conference. Dans T. Universe (Éd.), *Analytics and big data—reflections from the Teradata Universe conference*, (p. 12).

Abdelkader Baaziz, L. Q. (2013, Decembre 06). How to use big data technologies to optimize operations in upstream petroleum industry. *International Journal of Innovation*, pp. 30-42.

Abiteboul S., B. F., & s. (2014). L'émergence d'une nouvelle filière de formation : data scientist. *Rapport technique*, p. 11.

ANDERSON, A. C. (2008, Juin 28). *Science*. Consulté le Avril 20, 2017, sur WIRED: https://www.wired.com/2008/06/pb-theory/

B.Charlier. (1998). *Apprendre et changer sa pratique d'enseignement. Expériences d'enseignants.* Bruxelles: De Boeck.

Barton, D. e. (2010). Making Advanced Analytics Work for You. *Harvard Business Revieuw, 90*(N°10), pp. 78-83.

Béatrice Durand-Mégret, N. V. (2012). *La boîte à outils du responsable marketing* (Vol. 192 pages). (Dunod, & É. :. édition, Éds.) Paris, France: Dunod.

Big mama Technology. (2016, Janvier 20). Big Mama Business Plan. Alger, Algerie.

Bonaci, M. (2015, Avril 11). *The history of Hadoop.* Consulté le Avril 20, 2017, sur medium: https://medium.com/@markobonaci/the-history-of-hadoop-68984a11704

Brasseur, C. (2013). *Enjeux et usages du BIG DATA* (Vol. 207). Paris: Lavoisier.

Brown, A. S. (2008, Juillet). The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs. *Key Perspectives Ltd*, p. 09.

Bruchez, R. (2015). *Les bases de données NoSQL et le BIG DATA* (éd. Blanche). (Eyrolles, Éd.) Paris, France.

Charlier , B. (1998). Apprendre et changer sa pratique d'enseignement. Expériences d'enseignants. (Persée, Éd.) *Recherche & Formation, 186-190*, p. 187.

Chatfield et al. (2014). Data Scientists as Game Changers. *25th Australasian Conference on Information Systems* (pp. 1-11). Dec 2014, Auckland, New Zealand: Faculty of Engineering and Information Sciences.

Corea, F. (2016, Janvier). A Chimera Called Data Scientist: Why They Don't Exist (But They Will in the Future). (S. Links, Éd.) *Big Data Analytics: A Management Perspective*, pp. 25-30.

Data Scientist 1;. (2017, 4 24). Compétences data scientist. (G. H. Miled, Intervieweur)

Data scientist 1;. (2017, Mai 11). Intervieux projet C. (G. H. Miled, Intervieweur)

Data scientist 2,. (2017, Avril 24). Intervieuw projet A. (G. H. Miled, Intervieweur)

Data scientist 3. (2017, Mai 11). Intervieuw compétences et évolution. (G. H. Miled, Intervieweur)

Data scientist 3,. (2017, Mai 11). Intervieuw projet A. (G. H. Miled, Intervieweur)

Data scientist 4. (2017, Mai 11). Entrevieuw compétences et évolution. (G. H. Miled, Intervieweur)

Data Scientist 4,. (2017, Mai 11). Intreview projet B. (G. H. Miled, Intervieweur)

Davenport, T. (2014, Mars 3). Big Data at Work : DISPELLING THE MYTHS, UNCOVERING THE OPPORTUNITIES. (H. B. Review, Éd.) *Harvard Business Revieuw*, 70-76.

Davenport, T. H. (2012, Octobre). Data scientist: The sexiest Job of the 21st century. *Harvard Business Review*, pp. 70-76.

Dumez, H. (2016). *Méthodologie de la recherche qualitative.* Paris: Vuibert.

Frega, R. (2015, Avril 23). *Les pratiques normatives.* (R. SociologieS, Éd.) Consulté le Avril 24, 2017, sur Open Edition: http://sociologies.revues.org/4969

Frega, R. (2015,, Février 23). *Les pratiques normatives.* Consulté le Avril 11, 2017, sur SociologieS: http://sociologies.revues.org/4969

G.Illescas et al. (2016). How to Think Like a Data Scientist: Application of a Variable Order Markov Model to Indicators Management. (S. International, Éd.) *Springer International Publishing*, pp. 153-163.

Geller, N. (2011). "Don't shun the 'S' word", in which she defends statistics. (https://www.safaribooksonline.com/library/view/doing-data-science/9781449363871/ch01.html, Éd.) *Amstat new article*, p. 76.

Guba & Lincoln. (1985). Évaluation et paradigme constructiviste.

Harris, D. (2013, Juin 11). *GIGAOM*. Consulté le Avril 22, 2017, sur Kaggle now has 100K data scientists, but what's a data scientist?: https://gigaom.com/2013/07/11/kaggle-now-has-100k-data-scientists-but-whats-a-data-scientist/

Hillion, S. (2011, Octobre 11). *Product and engineering leadership in analytics*. Consulté le Avril 11, 2017, sur Forbes: http://www.forbes.com/sites/danwoods/2011/10/11/emc-greenplums-steven-hillion-on-what-is-a-data-scientist/

Howard, J. (2012, Mars 02). *What is a data scientist*. Consulté le Avril 17, 2017, sur The Gruadian: https://www.theguardian.com/news/datablog/2012/mar/02/data-scientist

IBM. (2009). *IBM Shoebox history*. Consulté le Avril 19, 2017, sur IBM: https://www-03.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html

IBM. (2010). *Data science*. Consulté le Avril 13, 2017, sur IBM: https://www.ibm.com/analytics/us/en/technology/data-science/

James E. Short, Roger E. Bohn, Chaitanya Baru. (2010, Janvier 02). *ESI report.* (C. f.-S. Research, Éd.) Consulté le Avril 20, 2017, sur Centre for Large-Scale Data Systems Research: http://clds.sdsc.edu/sites/clds.sdsc.edu/files/pubs/ESI-Report-Jan2011.pdf

James Manyika and al. (2011, May). *Reports.* Consulté le Avril 20, 2017, sur McKinsey Global Institute: http://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx

KDnuggets. (2017, Mars 11). *Gregory Piatetsky-Shapiro*. Consulté le Avril 21, 2017, sur KDnuggets : http://www.kdnuggets.com/gps.html

Kenneth Cukier, V. M.-S. (2014). *L'augmentation des données importantes* . Robert Laffont.

Kevin Ashton. (2009). *Articles.* Consulté le Avril 20, 2017, sur RFID Journal: http://www.rfidjournal.com/articles/pdf?4986

Kevin Green, B. K. (2001). *Software as a service.* Consulté le Avril 20, 2017, sur Software and Information Industry Association SIIA: http://www.crmlandmark.com/library/saas_aug04_white_paper_offer.pdf

Khelil, H. (2017, Janvier 20). Business plan BIG mama Technology. Alger: Algérie.

Kofman, F. (2009). *L'entreprise consciente : Comment créer de la valeur sans oublier les valeurs* (Vol. 441). Paris, France: Des îlots de résistance .

Laney, D. (2001, Février 6). *3D Data Management: Controlling Data Volume, Velocity and Variety.* (M. Groupe, Éd.) Consulté le Avril 20, 2017, sur Gartner: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Laurent, P. B. (2016). From Statistician to Data Scientist. *Pedagogical Developments at INSA Toulouse* (pp. 5-6). Toulouse: Soci´et´e Franc¸aise de Statistique (SFdS),.

Lebbah, M. (2016). Journée Scientifique « Big Data & Data Science ». *Journée Scientifique « Big Data & Data Science »* (p. 1). Tunis: Département des mathématiques et sciences naturelles, Académie Tunisienne des Sciences, des Lettres et des Arts, Bait al-Hikma.

Lei Liu1 et al. (2009, Octobre 24). BUILDING A COMMUNITY OF DATA SCIENTISTS: AN EXPLORATIVE. (C. N. Center, Éd.) *Data Science Journal, 8*, pp. 201-208.

Lesk, M. (1997). *How Much Information Is There In the World?* Consulté le Avril 20, 2017, sur lesk.com: https://courses.cs.washington.edu/courses/cse590s/03au/lesk.pdf

Loukides, m. k. (2010, Juin 02). *What is data science?* Consulté le Avril 03, 2017, sur O'reilly: https://www.oreilly.com/ideas/what-is-data-science

Luhn, H. (1958, Octobre). A Business Intelligence System. (IBM, Éd.) *IBM Journal, 6*, 1-6.

Marr, B. (2015, Fevrier 25). *A brief history of big data everyone should read*. Consulté le 19 04, 2017, sur World Economic Forum: https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/

Mason, H. (2010, Septembre). *a taxonomy of data science*. Consulté le Avril 13, 2017, sur Dataist.com: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

Miller, S. (2014, Mars 26). Collaborative Approaches Need to Close the Big Data Skills Gap. (IBM, Éd.) *Journal of Organization Design*, pp. 1-5.

Mines, C. U. (2011, October 28). *Computing history*. Consulté le Avril 20, 2017, sur Columbia University: http://www.columbia.edu/cu/computinghistory/hollerith.html

Mohanty et al. (2013). *Big Data Imperatives.* India: Apress.

Mucchielli, A. (1991). *Les méthodes qualitatives* (Vol. 126). (P. u. France, Éd.) Paris, France: Presses universitaires de France.

Myriam Karoui, G. D. (2012). *Big Data : Mise en perspective et enjeux pour les entreprises.* Consulté le Mai 26, 2017, sur ResearchGate: https://www.researchgate.net/profile/Aurelie_Dudezert/publication/274062564_Karoui_M_Devauchelle_G_Dudezert_A_2014_Big_Data_Mise_en_perspective_et_enjeux_pour_les_entreprises_N_Special_Big_Data_Revue_Ingenierie_des_Systemes_d%27Information_19_3_73-92_Herm

Newman, P. (2017, Janvier 12). *THE INTERNET OF THINGS REPORT.* Consulté le Avril 20, 2017, sur Business Insider: http://www.businessinsider.fr/us/the-internet-of-things-2017-report-2017-1/

NSF. (2005). Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. (N. S. Board, Éd.) *National Science Board*(NSB-05-40).

O'Reilly Media, I. (2009). Release 2.0. *O'reilly Radar*, 03.

OPIIEC. (2016). *Data scientist*. Consulté le Mars 22, 2017, sur referentiels metiers opiiec: http://referentiels-metiers.opiiec.fr/fiche-metier/113-data-scientist

Patil, D. (2011, Septembre 09). *Building data science team*. Consulté le Avril 03, 2017, sur O'reilly Radar: http://radar.oreilly.com/2011/09/building-data-science-teams.html

Peter Layman, Hal Varian. (2000). *In how much information.* California: School of Information, University of California, Berkeley.

Philippe Besse & Béatrice Laurent. (2016, Juin). De statiscticient à Data scientist développements pédagogiques à L'INSA de Toulouse. (S. f. statistiques, Éd.) *Statistique et enseignement*, pp. 76-93.

Philippe Besse, A. G.-M. (2014, May 24). Big Data Analytics - Retour Vers le Futur - De statisticien à Data Scientist. *Math.ST*, pp. 1-14.

Porway, J. (2012, Mars 02). *What is a data scientist?* Consulté le Avril 11, 2017, sur The Guardian: https://www.theguardian.com/news/datablog/2012/mar/02/data-scientist

Press, G. (2013, 05 09). *a short story about BIG DATA*. Consulté le 04 19, 2017, sur Forbes: https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#2cbd535a65a1

Radicati, S. (2014, Fevrier). *Reports.* Consulté le Avril 20, 2017, sur The Radicati Group, Inc.: http://www.radicati.com/wp/wp-content/uploads/2014/01/Mobile-Statistics-Report-2014-2018-Executive-Summary.pdf

Rider, F. (1944). The Scholar and the Future of the Research Library: A Problem and Its Solution. *Hadham press*.

Rifkin, J. (2011). *the third indutriel revolution.* Washington, D.C: Broché.

RIJMENAM, M. v. (2014). *Think Bigger : developing a successful big data strategy for your business* (éd. Business/ Information Management, Vol. 278). (AMACOM, Éd.) New york, Broadway, Etats Unis: American Management Association.

Roger Magoulas and Ben Lorica, B. D. Big Data: Technologies and techniques for large scales data. *Release 2.0.* NEW YORK.

Scouarnec, A. (2004, Mars). L'observation des métiers : définition, méthodologie et "actionnabilité" en GRH. *Management & Avenir*(N°1), pp. 23-42.

Strickland, J. S. (2014). *Data Science and Analytics for Ordinary People.* (I. Lulu, Éd.) Etats Unis: Simulation Educators.

Swan, A. &. (2008, Juillet). The skills, role and career structure of data scientists and curators: An Assessment of Current Practice and Future Needs. *Key Perspectives Ltd*, pp. 1-32.

Teleña, S. Á. (2015, Mars 11). *Data Scientists, the "unicorn" of data*. Consulté le Avril 13, 2017, sur BBVA: https://bbvaopen4u.com/en/actualidad/data-scientists-unicorn-data-what-are-they-what-do-they-do-and-how-will-they-change-world

Truskowski, R. M. (2003, 02). The evolution of storage systems. *IBM Systems Journal*, p. 42.

Tunkelang, D. (2017, Fevrier 20). Career of the Month. *National Science Teachers Association, 79*(N°6), p. 66.

Wikipedia. (1991, Aout 06). *The birth of the World Wide Web*. Consulté le Avril 21, 2017, sur Wikipedia: https://fr.wikipedia.org/wiki/Tim_Berners-Lee

Wikipedia. (2007, Juin 16). *John Graunt*. Consulté le Avril 21, 2017, sur Wikipedia: https://fr.wikipedia.org/wiki/John_Graunt

Wikipedia. (2016, Octobre 11). *Fritz Pfleumer*. Consulté le Avril 20, 2017, sur Wikipedia: https://en.wikipedia.org/wiki/Fritz_Pfleumer

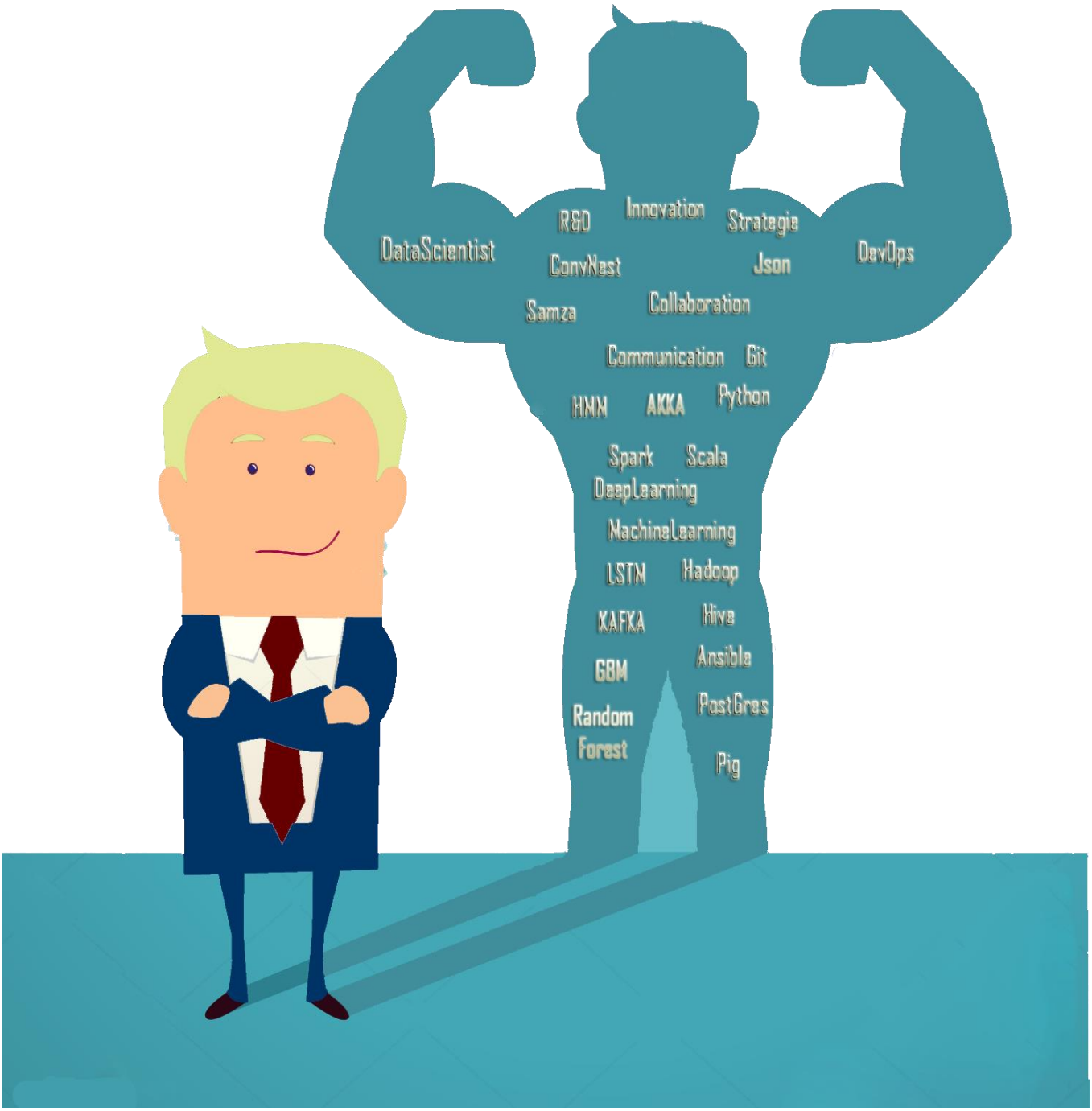
Wild CJ, P. M. (1999). Statistical thinking in empirical enquiry. *Int. Stat. Rev.*(67(3):223–65), pp. 1-68.

ANNEXE A

Figure synthèse inductive

Modèle data scientist – contexte Algérie-

Figure : Modèle data scienitist contexte Algérie

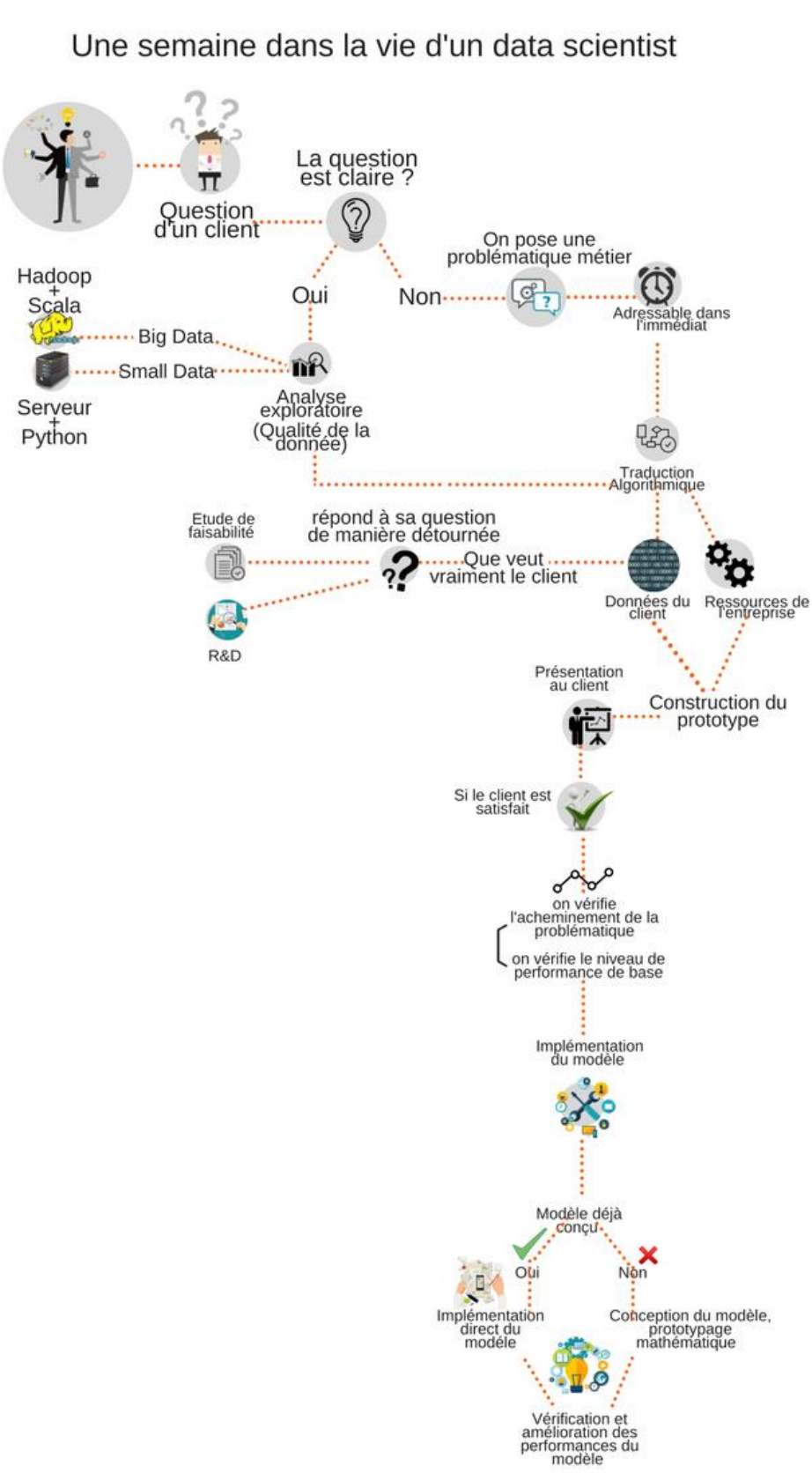


5 Source : Réalisé par nos soins en utilisant Adobe Photoshop

ANNEXE B

Schéma récapitulatif

Une semaine dans la vie d’un data scientist



Source : Réalisé par nos soins en utilisant l'outil Canva

ANNEXE C

GUIDE D'ENTRETIEN

CHEIF TECHNOLOGY OFFICER

Guide d'entretien du Cheif Technology Officer

Bonjour, je m'appelle **GRINE Hiba Miled** je suis stagiaire, je mène présentement une recherche scientifique dans le cadre du projet de fin d'étude pour l'obtention d'un *Master Management Stratégique et Systéme d'information*, portant sur **les facteurs clés de succès du métier de data scientist.** Au cours de l'entretien j'aimerais vous poser quelques questions, sur trois thématiques. La première va porter sur vos compétences en tant que data scientist, la seconde sur votre formation et cursus scolaire, et enfin la dernière sur les projets menés dans l'entreprise. L'objectif est de connaitre le métier de data scientist, mieux comprendre l'environnement dans lequel vous évolués, ainsi que les facteurs déterminant ce métier, finalement, très indéfini, pour commencer j'aimerai vous posez quelques questions sur l'entreprise.

Informations de l'interviewé

Nom……**Prénom**……..…….
Poste occupé……...…**Depuis**…….…….
Poste précédent……..……..

Informations sur l'entreprise

1. Quels sont les principaux objectifs de l'entreprise ?

……
……
……

2. Quel est le nombre d'employés au sein de votre entreprise ?

……

*Pouvez-vous me justifiez le choix de la structure organisationnelle, et celle du statut juridique, est-ce un choix ou une conséquence du nombre ?

……
……

3. Parlez-moi des activités de l'entreprise, que fait Big mama exactement ?

……

*Permettez-moi de détailler encore davantage, comment justifiez-vous le choix de la prestation sur mesure et du disant « positionnement haute gamme »

……
……

……

4. Quel le modèle économique de l'entreprise ?

……
……
……

*Quels vos motivations et ambitions prochaines concernant le modèle économique ?

Equipe de l'entreprise

5. Comment avez-vous choisis vos data scientists, quels étaient les critères de recrutements ?

……

……

6. Comment définissez-vous un data scientist ?

……
……

……

*Pouvez-vous m'en dire davantage sur les éventuelles ambitions d'agrandir l'équipe ?

……
……
……

Clôture

1. Avez-vous autre chose à ajouter concernant les perspectives et ambitions de l'entreprise ?

……
……
……

Je vous remercie pour votre amabilité et le précieux temps que vous m'avez consacré et je vous souhaite une très bonne journée.

ANNEXE D

GUIDE D'ENTRETIEN DATA SCIENTIST

COMPETENCES ET PREREQUIS

Guide d'entretien des fonctions

Bonjour, je m'appelle **GRINE Hiba Miled** je suis stagiaire, je mène présentement une recherche scientifique dans le cadre du projet de fin d'étude pour l'obtention d'un *Master Management Stratégique et Système d'information*, portant **sur les facteurs clés de succès du métier de data scientist.** Au cours de l'entretien, j'aimerais vous poser quelques questions sur deux thématiques, dont la première sur les savoirs et les connaissances acquises durant votre parcours scolaire, et la seconde le développement de ses connaissance, et des nouvelles compétences développé de produits au sein de votre entreprise. Cela me permettra de connaitre l'entreprise, de mieux comprendre le processus de développement actuel, ainsi que le niveau d'implication et la contribution des différentes fonctions. En outre, cela m'aidera à mettre en évidence les motivations et freins pour une éventuelle amélioration.

Informations de l'interviewé

Nom……**Prénom**……..…….
Poste occupé……...…**Depuis**…….…….
Poste précèdent……..……..

Connaissance et savoir développé durant le cursus scolaire, ou expérience précédente

1. Qu'avez-vous fait comme étude ?

……
……

2. Quelles sont les savoirs et les connaissances développées au cours de votre formation?

……
……

3. Quelles compétences/ connaissance vous ont permis, d'être directement opérationnel en tant que data scientist ?

……
……

4. Avez-vous eu une expérience précédentes en data science/ machine learning/big data ?

……
……

Intégration à l'entreprise

5. Avez-vous eu des problèmes d'intégrations les premiers jours ?

……
……

*combien à durée votre période d'intégration ?

……
……

*Qu'avez-vous appris durant votre période d'intégration ?

……
……

*Qu'elles étaient les formations que vous-avez eu avant votre premier projet ?

……
……

6. Avez-vous rencontrez des difficultés d'ordre technique ?

……

*Pouvez-vous me dire qu'elles étaient ses contraintes ?

……
……

*Comment les aviez-vous surmontées ?

……

……

7. Concrètement, pouvez-vous m'en dire un peu plus sur votre rôle au tant que data scientist?
……
……

8. Afin d'évaluer votre évolution au tant que data scientist, nous mettons à votre disposition deux grilles, merci de bien vouloir le remplir. En y introduisant les dates clés depuis votre intégration à big mama, et toutes celles qui se sont marqués par des événements quelconque.

Clôture

9. Avez-vous autre chose à ajouter concernant le processus de développement de produits dans votre firme ?

……
……
……

Je vous remercie pour votre amabilité et le précieux temps que vous m'avez consacré et je vous souhaite une très bonne journée.

ANNEXE E

Grille de compétences ‘‘Skills grid’’ et d’évolution

Compétences des data scientists

Skills	Oui	%
Mathématiques		
Statistiques /classique/ temporelSpatial/bayésienne		
Probabilités		
Algorithmique		
Data science skills	**Oui**	**%**
Machine Learning		
Deep Learning		
Prototyping		
Back-End Programming		
Python		
Scala		
Java		
Autres		
Front-End Prog		

Data Manipulation	**Oui**	**%**
SGBdD		
Data mining		
Data visualisation		
Data Analyse		
Data architecture		
BIG DATA technologies	**Oui**	**%**
Hadoop		
Map/reduce		
Spark		
Mango Db		
Aurtres		
Soft Skills	**Oui**	**%**
Communication		
Leadership		
Creativity		
Innovation		
Collaboration		
Curiosity		

Compétences des data scientists

Business Skills	Oui	
Strategy		
Marketing		
Business analytics		
Project management		
Administration systems	Oui	
Cloud		
Drupal		
Autres		

Date clés	Compétences développé a Big mama jusqu'à aujourd'hui	Connaissance/ Compétences acqu durant votre cursus/expérience

Optimisation	**Oui**	
Convex		
Linear		
Global		

ANNEXE F
GLOSSAIRE

GLOSSAIRE :

Mot	Signification
Accuracy rate	Nombre de documents pertinents existant dans un fonds documentaire, à partir duquel sont calculés les taux de rappel et de précision
ACP	Réduction dimension
Apache Flink	Apache Flink est un framework de gestion des flux open source.
Apache Flume	Flume est un logiciel destiné à la collecte et à l'analyse de fichiers de log. L'outil est conçu pour fonctionner au sein d'une architecture informatique distribuée et ainsi supporter les pics de charge.
Apache Hbase	HBase est un système de gestion de base de données non-relationnel distribué, écrit en Java, disposant d'un stockage structuré pour les grandes tables.
Apache Hive	Apache Hive est une infrastructure d'entrepôt de donnée intégrée sur Hadoop permettant l'analyse, le requêtage via un langage proche syntaxiquement de SQL ainsi que la synthèse de données
Apache Kerberos	Kerberos est un outil de sécurité et d'authentification réseau, il fournit une authentification forte basée sur la fourniture de tickets cryptés sécurisés entre les clients demandant l'accès aux serveurs et les fournissant d'accès.
Apache Pig	Pig est une plateforme haut niveau pour la création de programme MapReduce utilisé avec Hadoop.
Apache Sentry	Sentry est un service d'authentification dans Hadoop, qui fournir des services d'authentification aux composants de l'écosystème Hadoop.
Apache Spark	Spark est un framework open source de calcul distribué. Il s'agit d'un ensemble d'outils et de composants logiciels structurés selon une architecture définie. Ce produit est un cadre applicatif de traitements big data pour effectuer des analyses complexes à grande échelle.
Apache Sqoop	Sqoop est une interface en ligne de commande de l'application pour transférer des données entre des bases de données relationnelles et Hadoop
Apache Storm	Apache Storm est un framework de calcul et de traitement de flux distribué écrit principalement dans le langage de programmation Clojure.
Biais	erreur d'approximation, Le biais est l'erreur provenant d'hypothèses erronées dans l'algorithme apprentissage. Un biais élevé peut être lié à un algorithme qui manque de relations pertinentes entre les données en entrée et les sorties prévues (sous-apprentissage)
Classification automatique	La classification automatique - clustering - est une étape importante du processus d'extraction de connaissances à partir de données. Elle vise à découvrir la structure intrinsèque d'un ensemble d'objets en formant des regroupements - clusters - qui partagent des caractéristiques similaires.
Clojure	Un langage de programmation fonctionnel compilé, multi-plateforme et destiné à la

	création de programmes sûrs et facilement distribuables
ClouderaML	collection de librairie Java et de lignes de commande sous licence Apache pour aider les data scientist, à effectuer des tâches courantes d'élaboration de données et d'évaluation de modèles. généralement destiné aux étudiants qui souhaitent comprendre les techniques de construction de modèles d'apprentissage machine robustes et évolutifs sur Hadoop.
Cross-validation	La validation croisée(« cross-validation ») est une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage
Data science	La science des données est l'extraction de connaissance d'ensembles de données. Elle emploie des techniques et des théories tirées de plusieurs autres domaines plus larges des mathématiques, la statistique, la théorie de l'information et la technologie de l'information, notamment le traitement de signal, des modèles probabilistes, l'apprentissage automatique et statistique, la programmation informatique, l'ingénierie de données, la reconnaissance de formes, la visualisation, l'analytique prophétique, la modélisation d'incertitude, le stockage de données, la compression de données et le calcul à haute performance.
Deep Learning	L'apprentissage profond (en anglais deep learning, deep structured learning, hierarchical learning) est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures articulées de différentes transformations non linéaires.
Git	Git est un logiciel de gestion de versions décentralisé, utilisé pour la gestion de projets informatique, qui permet de versionner le code source.
Grafana	Grafana est le plus souvent utilisé pour visualiser des données de séries chronologiques pour l'infrastructure internet et l'analyse d'applications.
Hadoop	Framework open source écrit en Java destiné à faciliter la création d'applications distribuées (au niveau du stockage des données et de leur traitement) et échelonnables (scalables) permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données. Chaque nœud est constitué de machines standard regroupées en grappe. Tous les modules de Hadoop sont conçus dans l'idée fondamentale que les pannes matérielles sont fréquentes et qu'en conséquence elles doivent être gérées automatiquement par le framework.
HDFS	Le HDFS est un système de fichiers distribué, extensible et portable développé par Hadoop à partir du GoogleFS. Écrit en Java, il a été conçu pour stocker de très gros volumes de données sur un grand nombre de machines équipées de disques durs banalisés. Il permet l'abstraction de l'architecture physique de stockage, afin de manipuler un système de fichiers distribué comme s'il s'agissait d'un disque dur unique.
Intégration continue automatisée	L'intégration continue est un ensemble de pratiques utilisées en génie logiciel consistant à vérifier à chaque modification de code source que le résultat des modifications ne produit pas de régression dans l'application développée, et permet d'automatiser l'exécution des suites de tests et de voir l'évolution du développement du logiciel.
JSON	JavaScript Object Notation est un format de données textuelles, générique. Il permet de représenter de l'information structurée.
KAFKA	Kafka est un système de messagerie distribué de stockage et d'échange de données

	émises en temps réel. Originellement développé chez LinkedIn, et maintenu au sein de la fondation Apache depuis 2012.
Kibana	Kibana est un greffon open source de visualisation de données pour Elasticsearch. Il fournit des fonctions de visualisation sur du contenu indexé dans une grappe Elasticsearch. Les utilisateurs peuvent créer des diagrammes en barre, en ligne, des nuages de points, des camemberts et des cartes de grands volumes de données
Kmeans	Algorithme d'apprentissage non supervisé, à l'origine du traitement du signal, qui est populaire pour l'analyse par grappes dans l'exploration de données.
Lasso	En statistiques, le lasso est une méthode de contraction des coefficients de la régression utilisé pour traiter les données quand $n < p$
Logfile	Log file désigne le fichier contenant des enregistrements datés et classés par ordre chronologique, ces derniers permettent d'analyser pas à pas l'activité interne du processus et ses interactions avec son environnement.
Machine Learning	Le machine learning est une des branches de l'intelligence artificielle. C'est une discipline qui est apparue avec les premiers travaux de recherche sur les statistiques et les réseaux neuronaux, il sert à développer des processus d'apprentissage permettant à une machine d'évoluer, sans que ses algorithmes ne soient modifiés, avec pour objectif la construction d'un modèle prédictif. Selon le contexte et les données disponibles, une machine peut apprendre de plusieurs façons (supervisée, non supervisée, renforcée, par transfert…) et selon plusieurs approches (arbres de décision, régression linéaire, réseaux neuronaux, réseaux bayésiens, machines à vecteurs…).
Mahout	Apache Mahout ™ est une application qui permet de créer un environnement permettant de développer rapidement des applications d'apprentissage machine scalable et évolutives.
MangoDB	MongoDB est un système de gestion de base de données orientée documents, répartissable sur un nombre quelconque d'ordinateurs
MapReduce	MapReduce est un patron d'architecture de développement informatique, dans lequel sont effectués des calculs parallèles, et souvent distribués, de données potentiellement très volumineuses. Les termes « map » et « reduce », et les concepts sous-jacents, sont empruntés aux langages de programmation fonctionnelle. 1- 'map' pour découper la base ou chaînes de données en sous-segments. 2- 'reduce' pour recoller les morceaux après le traitement distribué sur plusieurs noeuds (ou process 'shuffle' entre processeurs)
Matlab	MATLAB est un environnement de programmation facile et performant pour les ingénieurs et les statisticiens, ayant un langage de programme sous le même nom qui permet de manipuler des matrices, d'afficher des courbes et des données, de mettre en œuvre des algorithmes.
Multitasking	dans l'ingénierie informatique, le multitasking désignant un type de système d'exploitation multitâche, capable de traiter en même temps plusieurs programmes informatiques. Il a ensuite été décliné pour s'appliquer à l'humain, désignant le fait de pratiquer plusieurs activités en même temps en utilisant plusieurs moyens de communication de manière simultanée.
Need to have	Les nécessités à avoir. (facteurs clés de succès primaire).

Nice to have	Les préférences à avoir. (facteur clés de succès secondaire).
Omique	les sciences omiques, incorporent diverses disciplines, tels que la génomique, protéomique, métabolomique, celle ci analysent un grand volume de données, qui s'interprètent par le biaie du bioinformaticien.
Overfitting	sur apprentissage
Protocol Buffers	Protocol Buffers est un format de sérialisation avec un langage de description d'interface développé par Google.
Réseau de neurones	Un réseau de neurones artificiels, est un ensemble d'algorithmes dont la conception est à l'origine très schématiquement inspirée du fonctionnement des neurones biologiques, et qui par la suite s'est rapproché des méthodes statistiques.
Scala	Scala est un langage de programmation multi-paradigme conçu pour exprimer les modèles de programmation courants dans une forme concise et élégante. Son nom vient de l'anglais Scalable language qui signifie à peu près « langage adaptable » ou « langage qui peut être mis à l'échelle ». Il peut en effet être vu comme un métalangage.
Scalability	la capacité d'un produit à s'adapter à un changement d'ordre de grandeur de la demande (montée en charge), en particulier sa capacité à maintenir ses fonctionnalités et ses performances en cas de forte demande
SGBD	Acronyme de système de gestion de bases de données.
Sparse :	modèle parcimonieux (KO)
SQL	Sigle de Structured Query Language, en français langage de requête structurée) est un langage informatique normalisé servant à exploiter des bases de données relationnelles.
SVM	Support vector machine
Variance	erreur d'estimation, Une variance élevée peut entraîner un sur-apprentissage, c'est-à-dire modéliser le bruit aléatoire des données d'apprentissage plutôt que les sorties prévues
Vélocité	Grande rapidité dans le mouvement
Voracité	avidité de satisfaire un besoin