N°2 | STEPWISE

DESSINE MOI UN DATA ENGINEER

Quantmetry


dessine moi
un DATA
ENGINEER

S O M M A I R E

PRÉFACE

Imaginer, concevoir et mettre en oeuvre des solutions Big Data performantes, travailler avec des Data Scientists pour traduire des algorithmes en code exécutable à grande échelle adapté aux architectures,... le champ des compétences résumé par le terme de Data Engineer est large et peut varier en fonction des sociétés, il évolue au rythme effréné des technologies, dans un marché très demandeur, alors que de nouvelles formes de travail se développent.

Dans ce contexte, certes passionnant et très stimulant pour de jeunes ingénieurs très sollicités, il est parfois difficile de s'y retrouver et de faire des choix éclairés. Quelles compétences dois-je développer et par quel biais ? Vers quel type de sociétés dois-je me tourner (start-ups, grands groupes ou sociétés spécialisées) ? Quels projets dois-je privilégier ?

Par cet ouvrage, nous souhaitons apporter ici quelques éléments de réponse qui aideront chacun à faire les choix les plus adaptés à leurs ambitions, et notamment quelques témoignages de ces ingénieurs qui croient dans le progrès et construisent à grande vitesse le monde de demain, non sans se poser quelques questions.

Guillaume Bodiou, Directeur chez Quantmetry

INTRODUCTION

Après le succès planétaire de Dessine moi un Data Scientist et l'apparition de nouveaux challenges, de nouveaux profils voient le jour dont celui des Data Engineers. Des profils parfois difficiles à débusquer, qui sont pourtant bel et bien indispensables pour relever les défis des années à venir. L'objectif de cet ouvrage est de vous expliquer ce qu'est un Data Engineer en l'an 2018 de notre ère. Quelles sont leurs missions ? Où se cachent-ils ? Et finalement naît-on Data Engineer ou le devient-on ?

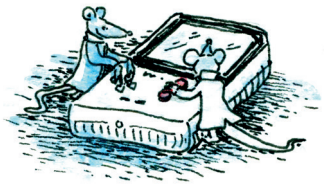
Une chose devenue assez évidente au cours de ces dernières années est que l'Intelligence Artificielle est toujours le sujet qui fait couler des litres et des litres d'encre. Vertueuse ou perfide ? Chacun a son positionnement. Quoi qu'il en soit, ce sujet tend à prendre de plus en plus d'importance. Corollaires directs de ce phénomène, les métiers de la Data sont toujours plébiscités par les entreprises et leurs côtes de popularité n'ont de cesse de monter en flèche.

À l'image de la fonction de Data Scientist qui connaît un succès total auprès des étudiants et professionnels, celle de Data Engineer a déjà commencé à devenir une nouvelle fonction tendance dans l'univers de la Data. En effet, les entreprises, étant maintenant assez matures pour évoluer sur des projets d'industrialisation du traitement de la donnée, elles cherchent à recruter des Data Engineers pour mettre en place ces plans d'envergure.

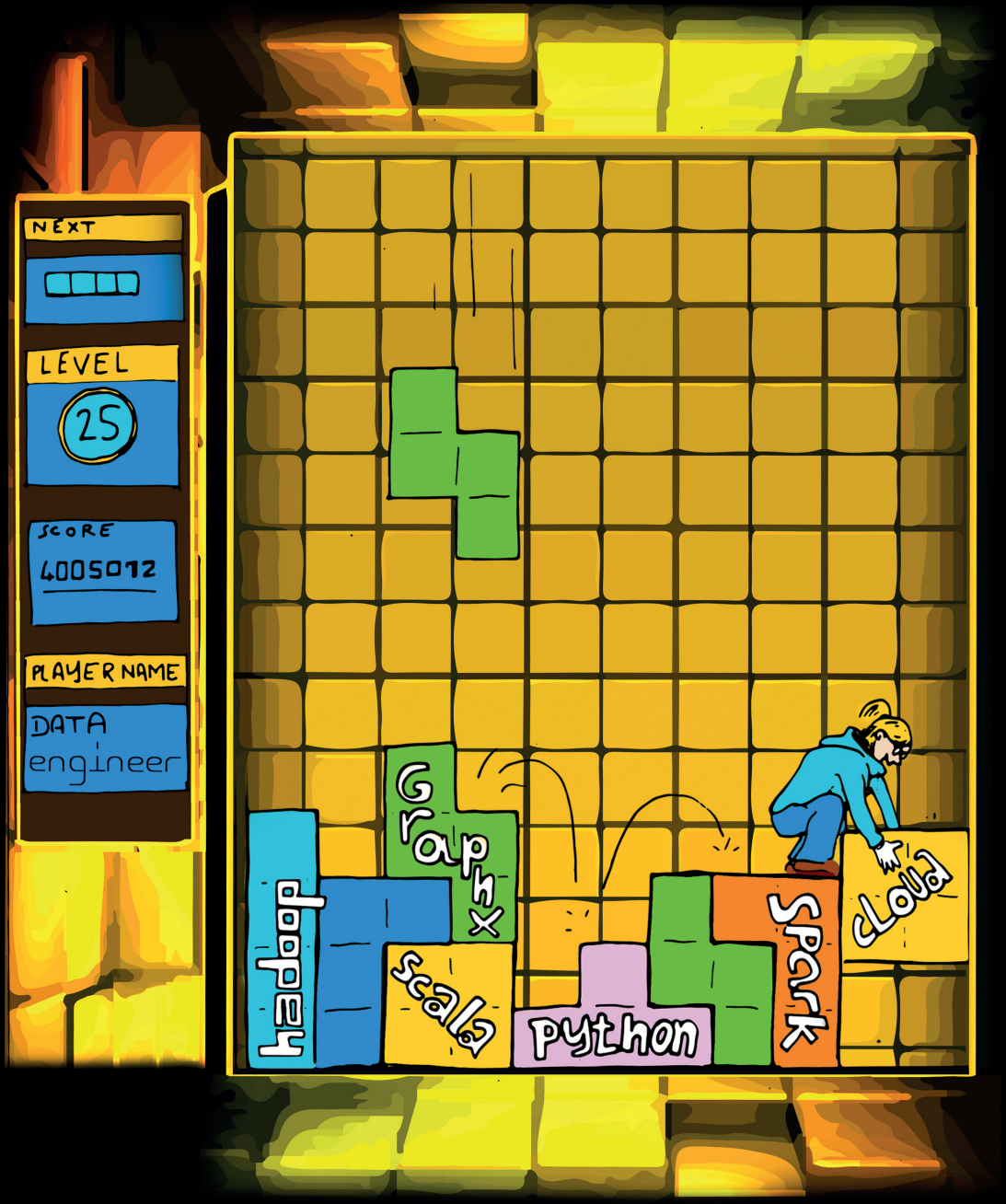
Malgré cet engouement, les entreprises peinent à recruter ces profils techniques. Le temps de l'entreprise en position dominante par rapport à l'employé est révolu depuis bien longtemps. Aujourd'hui, c'est le candidat qui est en position de force (et il le sait) en particulier dans l'univers de la Data et plus spécifiquement pour les Data Engineers. Ces derniers sont très sollicités à travers tous les canaux (réseaux sociaux, salons, sites web spécialisés en programmation type Stack Overflow – GitHub, meetup…). Le marché du recrutement des métiers du Data Engineering est encore plus tendu que celui de la Data Science.

Mais comment expliquer ce décalage entre demande et offre ? Comme nous l'avons dit précédemment , les entreprises ont compris qu'un POC (Proof Of Concept) n'était pas la finalité d'un cas d'usage de la data mais bien une étape initiale dans un processus débouchant sur la création de valeur consécutive à ces projets innovants (pour plus de détails sur l'industrialisation des projets data, n'hésitez pas à (re)consulter notre livre blanc "Y a t-il une vie après les POCs"). Ce gain de maturité a permis de démocratiser la fonction de Data Engineer, encore peu connue auparavant.

En parallèle de cela, les écoles ont eu un temps de latence important (qui tend à être rattrapé mais avec encore des difficultés) pour se rendre compte du besoin naissant et fulgurant de profils techniques spécialisés en Big Data. Malgré une structuration des cursus, il existe encore trop peu de parcours type pour devenir Data Engineer permettant aux étudiants de travailler sur des gros volumes de données ou avec des cas d'usage nécessitant du temps réel. C'est un sujet qui sera abordé plus loin dans cet ouvrage.



Le Tetris du Data Engineer



1. UN PROFIL TECHNIQUE... MAIS PAS QUE !

Beaucoup d'écoles se proposent de former à la Data, en abordant la Data Science (mathématiques, statistiques) et/ou la Big Data (Programmation, développement) à travers des enseignements théoriques et pratiques. Ce type de cursus permet à l'apprenti Data Engineer de "s'armer" pour ses premières expériences professionnelles. Mais il devra néanmoins s'attacher à développer d'autres compétences pour pouvoir faire face à l'ensemble des challenges qui l'attendent.

En effet, la fonction de Data Engineer nécessite des compétences hautement pointues, autant d'un point de vue de la **technique**, ce que l'on appellera les hard skills, représentant les compétences « dures » : **le savoir-faire**, que du point de vue de la **personnalité**, ce que l'on appellera les soft skills, représentant les compétences « douces » : **le savoir-être**.

A. LES *HARD SKILLS*

Elles représentent le cœur de la fonction de Data Engineer. Ces compétences sont primordiales pour pouvoir assumer un tel poste. Voici une liste, non exhaustive, des compétences primaires à détenir pour comprendre et travailler sur des enjeux Big Data. Cette liste fait l'objet d'une présentation plus visuelle à travers « la boussole du Data Engineer » plus loin dans ce livre blanc :

- Comprendre le fonctionnement et maîtriser le développement d'application dans des langages de programmation tels que Python, Scala et Java.
- Maîtriser les bases de données (SQL, NoSQL, SGBDR)
- Comprendre le fonctionnement des technologies Big Data et des systèmes distribués (M/R, Hadoop, Event messaging…)
- Connaître et savoir appliquer les méthodes agiles
- Avoir des connaissances en systèmes et réseaux (Unix, TCP/IP, ...)
- Avoir des notions d'infrastructure IT on-premise ou sur le cloud
- Savoir utiliser les outils du DevOps (versionning, CI/CD, …)

En ayant ces atouts en main, le Data Engineer peut d'ores et déjà commencer à appréhender relativement sereinement les enjeux et problématiques liés aux projets Data.

B. LES *SOFT-SKILLS*

Oui, un bagage de connaissances techniques solide est impératif pour embrasser la fonction de Data Engineer. Néanmoins, le milieu de la Data étant en perpétuelle évolution, nous assistons à des changements, des remises en question, des évolutions qui font que les us et pratiques d'aujourd'hui seront différentes de celles de demain. Se concentrer uniquement sur les compétences techniques, qui sont vouées à être remises en cause avec le temps par d'autres avancées, ne suffisent pas aujourd'hui pour être efficace et compétent dans ce domaine.

C'est ainsi que certaines qualités sont nécessaires pour pouvoir évoluer dans cette fonction. Nous avons pu en dégager trois principales qui permettent d'envisager sereinement un job de Data Engineer :

- La curiosité : une des qualités les plus importantes pour un Data Engineer. En effet, l'univers de la Data évolue à une vitesse vertigineuse, ce qui a pour conséquence de changer, faire évoluer les pratiques d'hier, obligeant ainsi les Data Engineers à se mettre dans une **veille permanente** tant pour assurer les missions et autres projets (R&D par exemple) que pour rester attractif sur le marché de l'emploi.

- La communication : l'activité du Data Engineer est complexe et ce dernier sera amené à travailler avec d'autres interlocuteurs : Data Engineers, Data Scientists, des opérationnels… Il est nécessaire d'adapter son discours en fonction de la situation, car ses interlocuteurs n'ont pas tous les connaissances et/ou le background technique suffisant pour comprendre les enjeux. Il est du devoir du Data Engineer d'expliquer, de mettre son discours à la portée de son audience, pour que ces derniers puissent en assimiler les enjeux et messages clés.

- L'esprit d'équipe : Comme dit précédemment, les Data Engineers sont amenés à travailler en collaboration avec d'autres acteurs (techniques ou non). Il est primordial qu'ils puissent organiser leurs tâches en fonction du but commun et ne pas jouer la carte du "loup solitaire", ce qui aurait pour conséquence d'être contre productif et donc de ralentir le projet.

En conclusion, être un Data Engineer ne se résume pas à une stack de compétences techniques accumulées. En effet, il est essentiel de pouvoir travailler sur ses hard-skills tout autant que sur ses soft-skills.

C. LA BOUSSOLE DU DATA ENGINEER

(voir feuille détachable)

2. DES EXPÉRIENCES VARIÉES TU AURAS

Une fois les différents pré-requis validés, plusieurs parcours sont possibles et permettent de s'orienter vers différents types d'entreprises et différents types de métiers.

A. LEVEL ONE : DATA ENGINEER JUNIOR

Côté Junior, il n'existe pas de formations qui préparent au métier de Data Engineer. De manière général, les futurs Data Engineers sont issus d'école d'ingénieurs généralistes. Ces école qui offrent des spécialisations à mi-chemin entre le développement informatique classique et une spécialisation dans les technologies Big Data, permettent d'armer les jeunes diplômés à prétendre à un premier poste en tant que Data Engineer junior, dans tous types de sociétés qui présentent des problématiques de traitement de données massives.

Pour démarrer en tant que Data Engineer, des parcours plus atypiques sont également possibles car plusieurs passerelles peuvent y mener. Youness, nous explique son parcours et comment ses études l'ont mené à un poste de data engineer dans le monde du conseil.

Raconte-nous ton parcours :

Après l'obtention de mon Bac S j'ai fait une licence en maths informatique puis j'ai intégré une école d'ingénieur, l'école Saint Etienne qui forme des ingénieurs généralistes. J'ai pu découvrir les nouvelles technologies d'information et de communication. Ensuite, je me suis spécialisé en informatique, j'ai fait un échange académique avec l'université de Québec et c'est là où je me suis vraiment spécialisé en développement java. A l'issue de mes études, je me suis installé à Paris en tant que développeur web avec la technologie Java/JEE puis j'ai continué en java avec plusieurs clients dans le secteur de la finance. Ensuite, grâce à toutes mes missions j'avais suffisamment de recul pour voir le processus informatique dans une grande entreprise et j'ai décidé de m'orienter dans la data

Qu'est ce qui t'a poussé à devenir Data Engineer ?

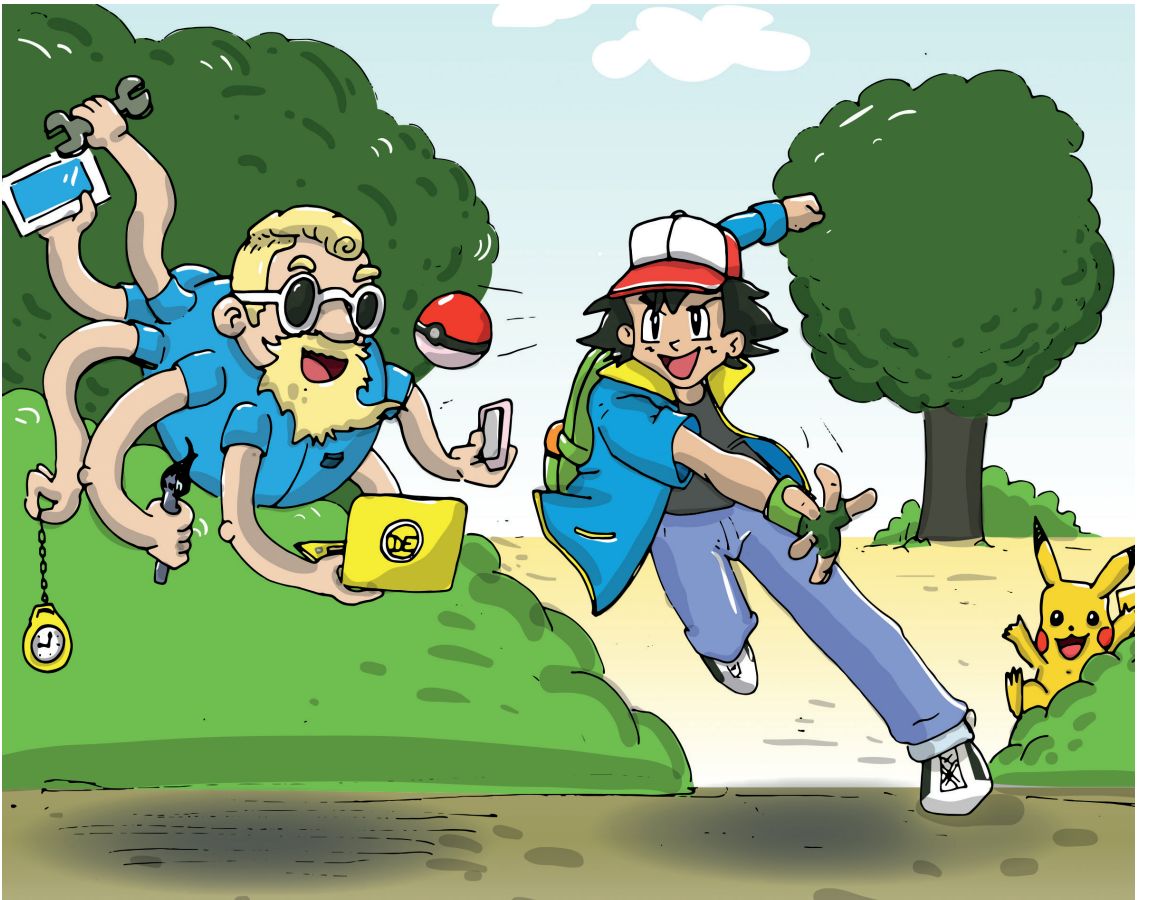
J'ai toujours été passionné par la quantité de data qui est traitée chaque jour avec les réseaux sociaux et les flux sur internet, cela prend de plus en plus d'ampleur et je trouvais que c'était une spécialité d'avenir puisque toutes les entreprises, et même la société se transforme avec toutes ces données. Je me suis dit que j'avais le background technique nécessaire pour ce qui est du développement et de l'architecture pour intégrer ce domaine. Aussi, j'ai fait une formation en Big Data au CNAM où j'ai découvert le machine learning, spark, hadoop et c'est cela qui m'a permis de devenir Data Engineer.

Comment es-tu monté en compétence ?

Le domaine d'activité de Data Engineer, c'est très vaste : les infrastructures, le développement, il faut connaître aussi les techniques de machine learning parce qu'il est difficile d'industrialiser des modèles qu'on ne connaît pas. Il faut avoir des connaissances et des best practices en développement pour produire un code qui est industrialisable, automatisé et robuste. Pour ce qui est des infrastructures, je travaille principalement sur du cloud et je prépare actuellement un certificat AWS donc je me forme grâce à des MOOC.

Tes projets pour l'avenir ?

Actuellement, je vise l'expertise technique donc j'aimerais devenir architect data pour participer à toute la chaîne de la conception, la mise en place et la maintenance dans la Big Data. En parallèle, je souhaiterais aussi avoir des responsabilités managériales et gérer des équipes mais plutôt sur le volet technique.



B. LEVEL UP : DATA ENGINEER SENIOR

Pour les seniors, les profils peuvent être variés mais les compétences qui sont demandées ont toutes un point commun : des connaissances sur des technologies Big Data. Issus d'un parcours BI ou encore anciennement développeurs, la curiosité et l'auto-formation permettent d'arriver à des postes plus "expérimentés". Johann, nous explique son parcours et les choix qui l'ont mené à devenir Senior Data Engineer.

Raconte-nous ton parcours :

Après un DEUG MIAS (mathématiques et informatique), j'ai décidé de faire un IUP en informatique, avec des cours en systèmes et réseaux ainsi qu'en Bases de données. En parallèle de mes études, j'ai travaillé en tant que freelance pour accompagner des entreprises sur la création de leur site web.

A l'issue de ma formation, j'ai décidé d'effectuer un stage dans le web, sur un projet d'agrégateur RSS, où je suis ensuite resté en CDI. Après trois ans, j'ai décidé de partir dans une start-up où j'ai construit un réseau social d'entreprise. C'était ma première expérience avec des bases de données NoSQL. J'ai travaillé sur l'implémentation de ce réseau social avec Elastic search, ce qui m'a permis de découvrir d'autres aspects du métier.

J'ai suis ensuite changé d'entreprise dans laquelle j'ai continué à monter en compétences sur l'optimisation de la segmentation clients (en analysant notamment les mouvements de souris sur les pages visitées). Suite à cela, beaucoup de questions m'ont assailli et j'ai eu la conviction que la Data allait devenir un sujet de première importance à l'avenir.

Qu'est ce qui t'a poussé à devenir Data Engineer ?

Depuis ma première expérience, je me suis intéressé à l'utilisation de la donnée et des études de cas qui pouvait s'en dégager, c'est comme ça que j'ai commencé à vouloir monter en compétences sur des sujets Big Data.

Commençant à sentir que mes missions devenaient répétitives sur la partie web, j'ai décidé de continuer à évoluer sur les sujets de Big Data.

Comment es-tu monté en compétences ?

J'ai essentiellement appris en pratiquant sur les différentes missions sur lesquelles j'ai travaillé.

Mais pour les personnes qui souhaitent se former, je recommande les MOOC. Un bon Data Engineer doit être à la fois curieux et capable d'apprendre vite.

C'est un métier qui évolue très rapidement à la fois sur l'aspect connaissances techniques,

mais aussi sur les enjeux côté utilisateur.

Par exemple, dans le secteur du marketing, on accorde de plus en plus d'importance à la publicité ciblée, on parle d'anticiper les publicités personnalisées à la seconde près !

Tes projets pour l'avenir ?

Continuer sur de nouveaux projets et continuer à renforcer mes compétences sur les nouvelles technologies.

C. CO-OP : DATA SCIENTIST

Loin du clivage qui peut sembler exister, les deux métiers ne sont pas si éloignés et ont tendance à se recouvrir sur plusieurs compétences (à quand une fusion ?). Beaucoup de Data Scientists sont intéressés par les problématiques de Data Engineering, de DevOps et de développement produit. Nicolas et Ysé, Data Scientists ont passé le cap et nous racontent la différence parfois ténue entre ces deux métiers.

Nicolas

Raconte-nous ton parcours

Après une formation à Central SupElec, j'ai décidé de me spécialiser en Mathématiques & statistiques appliquées à la Data Science.

J'ai effectué mon stage de fin d'études dans un laboratoire de recherches médicales à Singapour, où j'ai travaillé sur du traitement d'images afin de classifier les images présentes dans la base de données du laboratoire. J'ai pu travailler sur des méthodologies de Deep Learning, appliquées à de l'analyse d'image.

J'ai donc voulu continuer dans ce domaine, mais sans forcément choisir un secteur d'activité de prédilection. C'est ainsi que j'ai atterri chez Quantmetry en tant que Data Scientist Junior. Mes premières missions étaient essentiellement sur des POCs (proof of concept), dans des domaines variés (banques, assurances, industries, santé…). Au début de 2015, le marché a évolué et on a senti de plus en plus un besoin de nos clients de s'orienter vers des formats d'industrialisation et j'ai donc commencé à en faire. ,Qui dit industrialisation, dit également des compétences à la fois sur des outils Big Data (Hadoop, Spark, Scala…) mais également la mise en place de l'architecture chez nos clients. J'ai petit à petit vu mon poste évoluer passant de la pure Data Science, à un poste hybride entre Data Engineer et Data Scientist.

Comment es-tu arrivé à devenir Data Engineer ?

Dans un premier temps, je me suis rendu compte que c'est difficile de travailler sur des missions longues d'industrialisation uniquement avec des compétences en Data Science. De plus, l'état du marché change, et d'un point de vue commercial, il y a de moins en moins de missions où l'on fait seulement des POCs. J'ai donc commencé à travailler davantage sur des missions allant de l'élaboration de la stratégie data jusqu'à son industrialisation, j'ai donc évolué à un poste de Data Scientist très opérationnel, avec une très grande composante de développement de produits et d'applications poussées en machine Learning.

C'est ce qui m'a permis de concrétiser des projets, et plus je travaillais sur des missions d'industrialisation, plus je me rendais compte que, finalement, la Data Science et le Data Engineering, ne sont pas de fonctions isolées, mais plutôt complémentaires. Dans un premier temps, le Data Engineer va préparer la donnée et les flux ce qui va permettre au Data Scientist d'avoir une base de travail « clean ».

Comment es-tu monté en compétences ?

J'ai beaucoup appris par la pratique sur mes premières missions, je suis montée en compétences sur les sujets au fur et à mesure qu'ils arrivaient. J'ai surtout bénéficié de la politique de formation interne, j'ai profité des différents BBL (Brown Bag Lunch) et du transfert de compétences que j'ai pu trouver chez Quantmetry (notamment à travers la capitalisation). C'était également beaucoup de travail personnel, j'ai lu beaucoup d'articles sur le sujet. Mais, finalement, tout Data Scientist chez Quantmetry, va se trouver confronté(e) à des sujets Big Data, il faut donc être agile d'esprit et curieux.

Tes projets pour l'avenir ?

Après mon expérience sur le projet AIDA (assistant virtuel type voicebot), j'aimerais continuer à monter en compétences sur le développement de produits data et à solidifier mes compétences globales sur le Machine Learning et son application en Intelligence Artificielle. De plus, j'aimerais traiter de plus en plus de projets de développement « produit » de A à Z, ce qui me permettra de mettre en application à la fois mes compétences en Data Science et en Data Engineering.

Ysé

Raconte-nous ton parcours

Après une formation à Dauphine en mathématiques appliquées et statistiques, suivi d'un master en Big Data à Télécom Paris Tech, je me suis orientée vers le conseil afin de découvrir des cas d'usage d'application de Data Science, dans des domaines d'activités différents. J'ai donc rejoint Quantmetry pour mon stage de fin d'études en commençant

par des missions de marketing dans le secteur de la banque et de l'assurance sur des campagnes de Test&Learn. J'ai ensuite poursuivi sur plusieurs autres projets, tels que de la visualisation de données d'un réseau de transport pour schématiser les déplacements des voyageurs. J'ai commencé, un peu comme tout le monde au début de Quantmetry, sur des missions de POCS, et je me suis rendue compte que finalement nous ne montrions que la faisabilité des projets, sans avoir le côté challengeant de prouver que le projet pouvait fonctionner.

Qu'est-ce qui t'a poussée à devenir Data Engineer ?

J'ai travaillé sur plein de projets différents, mais toujours avec cette frustration de ne pas savoir si finalement le client allait utiliser ce que j'avais conçu.

Grâce à ma première mission « produit » (la création d'un Dashboard permettant d'analyser les données et de détecter des fraudes), j'ai eu une première vision de comment construire une API et développer un site web (en front et back end). Malgré le fait que la partie Data Science et Machine Learning était infime, j'ai enfin vu qu'il y avait des utilisateurs qui se connectaient au Dashboard. J'ai commencé à voir le côté opérationnel, de voir les « end-users », ce projet m'a également permis d'adapter mon Dashboard en fonction des besoins des utilisateurs et donc d'avoir des challenges qui évoluaient constamment.

Comment est-ce que tu es montée en compétences ?

Au fur et à mesure des missions en travaillant sur des « longs projets », j'ai commencé à vraiment développer, à faire des tests unitaires, à avoir des réflexions sur l'optimisation de code pour les phases d'industrialisation. De plus, j'ai développé une vision produit, ce qui m'a forcé à prendre davantage le temps de construire un code destiné à évoluer.

En plus des articles sur Medium, j'ai beaucoup appris des podcasts que j'ai pu écouter (Software engineering Daily, Big Data Hebdo, Google Cloud Platform). Il faut également avoir une très bonne connaissance des bases de données, une rigueur sur le code et chercher des méthodologies de travail qui permettent de tout automatiser.

Quels sont tes projets pour l'avenir ?

Après mon expérience sur le projet AIDA et le développement de l'assistant vocal, j'ai envie de continuer sur le développement de logiciels et d'applications, essentiellement destinés à des personnes comme vous et moi, qui ont un fort potentiel et une vraie valeur ajoutée (citymapper, waze…).

D. LEVEL DESIGN : DATA ARCHITECT

Il n'existe pas un parcours type pour devenir Data Architect : c'est un mélange de compétences à la fois variés et complémentaires. Malgré les formations qui existent déjà, il est impératif de s'intéresser à l'évolution des différentes technologies qui se développent sur le marché, une technologie qui est utilisée aujourd'hui peut vite devenir obsolète. Fouad, nous raconte son parcours et nous explique les différentes expériences qui l'ont conduit à devenir Data Architect chez Quantmetry.

Quel a été ton parcours pour devenir Data Architect ?

J'ai passé un bac, un DUT GEA (Gestion des Entreprises et des administrations) puis un BTS informatique.

En 20 ans d'expérience, j'ai eu l'occasion de passer par des secteurs d'activités différents (éditeur de logiciel, télécom, industrie…) et des types de structures variées : la start-up , le grand groupe, ma propre SSII et j'ai aussi été freelance. J'ai donc pu rencontrer beaucoup de systèmes d'informations différents.

Pour ce qui est des missions qui m'ont permis de devenir Data Architect, cela a, bien entendu, été progressif. Au départ, j'ai rédigé des documentations et des préconisations de déploiement pour des opérateurs télécom, j'ai pu gérer une plateforme technique (migration, design, évolutions des architectures) puis j'ai fait de la qualité, de la base de données, de l'infrastructure ou encore du support N3. J'ai aussi eu l'opportunité de piloter une refonte complète de système d'information et de développer des stratégies Cloud.

Cela m'a permis de me familiariser avec plein de technologies différentes en étant confronté à des contraintes d'architecture unique à chaque fois. A chaque nouvelle expérience, j'ai ajouté de nouvelles composantes techniques ou fonctionnelles à mon profil.

A la base, je suis donc architecte mais mes deux dernières expériences m'ont permis de découvrir la Data Science avec le Machine Learning, Scala, Spark et Hadoop et ainsi de devenir Data Architect.

Qu'est ce qui t'a poussé à devenir Data Architect ?

Je pars toujours du principe que l'informatique c'est un tout. Ce que j'aime dans le métier d'architecte c'est qu'il casse les silos des différents métiers (réseaux, systèmes, industrialisation, pilotage, monitoring, performance ou sécurité). Le Data Engineer doit rassembler toutes ses compétences car il est l'interface entre la machine et ce que le client a pu imaginer ou ce que le Data Scientist a pu manipuler et surtout, c'est lui qui doit faire l'industrialisation. Ce métier est un pur produit des méthodes agiles et il peut intervenir sur tout ou partie des projets et c'est vraiment intéressant. Personnellement, c'était une suite logique de mon parcours même si je n'avais pas fait beaucoup de développement, mais c'est comme tout, cela s'apprend.

Comment es-tu monté en compétence ?

Je l'ai fait au fur et à mesure de mon parcours. De plus, en étant en start-up, j'ai vite compris que j'étais un produit avec une date limite de consommation et que je pouvais très vite être dépassé par les nouvelles technologies. Donc, c'est une veille de technologies permanente et une formation personnelle avant tout : lire des livres ou faire des MOOC, être curieux. C'est encore plus vrai de nos jours : aujourd'hui, et essentiellement en Big Data, nous ne sommes plus confrontés à des technologies génériques. Chaque base de données est spécifique, chaque problématique nécessite un outil qui lui est propre. Le Data Engineer et le Data Architect doivent développer des stratégies sur-mesure. Pour conclure, pour monter en compétence il faut prendre des risques, c'est-à-dire se remettre en question et sortir de sa zone de confort, être à l'affût de la nouveauté pour toujours aller plus loin et progresser.

Tes projets pour demain ?

Je voudrais développer mes connaissances en NLP et autour du traitement de l'image pour continuer à progresser sur des langages ou des frameworks. Aussi, augmenter mes compétences en Data Science pour échanger de manière plus fluide avec les Data Scientists et tout type de corps de métiers pour être capable de communiquer avec les personnes d'une DSI ou un client (que ce soit technique ou non). Et puis, pourquoi pas continuer cette aventure IT à l'étranger ?

3. TES MISSIONS SI TU L'ACCEPTES...

Le coeur de métier du Data Engineer consiste principalement à **développer des applications informatiques** en utilisant des techniques variées allant de la programmation système à l'implémentation de technologies Big Data. En particulier, il **travaille souvent de pair avec des Data Scientists** ou des Data Analysts dans le cadre des projets. En effet, avant de pouvoir raffiner la donnée, il faut la collecter depuis les entrepôts bruts, la nettoyer, la transformer, l'organiser et la restituer pour leur permettre de travailler une matière exploitable et valorisable. Rien de bien nouveau diront certains, ce type de traitements existe, en effet, en informatique décisionnelle depuis 15 ans et les outils (ETL, Stockage RGBD, Reporting) sont matures et efficaces.

Dans ce chapitre, nous vous proposons des exemples de missions réalisées pour différents clients, qui illustrent bien le rôle du Data Engineer dans la réalisation des projets Data.

Mais alors, **quelles sont les spécificités des missions du Data Engineer qui en font un profil si recherché** aujourd'hui ?

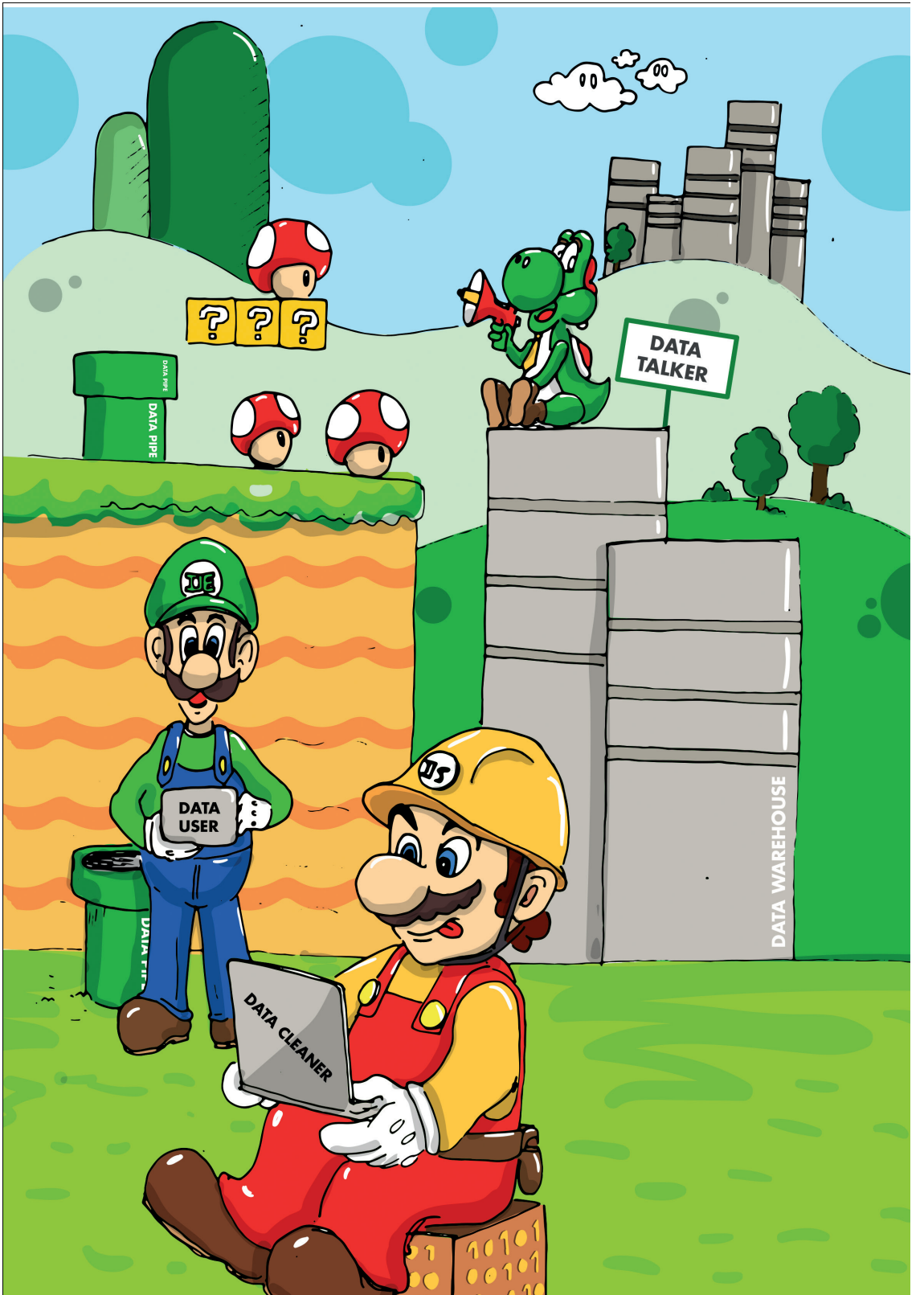
A. MISSION #1: TRAITEMENT DE DONNÉES TEMPS RÉEL

Pour le compte d'un Client dans le domaine de l'énergie, nous sommes intervenus dans la mise en place d'une plateforme d'acquisition de données (IoT) issues de capteurs provenant de systèmes de production d'énergie (soit plusieurs milliards de données reçues et à traiter chaque jour).

Le challenge est de pouvoir traiter et stocker ces séries temporelles en moins de 2 minutes (entre la production de la valeur et son stockage dans un format exploitable par les Métiers).

La chaîne de traitement de la donnée est relativement complexe puisqu'elle fait appel à :

- des référentiels pour :
  - valider les données (exclure les données inconnues, les valeurs hors normes etc.)
  - normaliser les paramètres : associer les paramètres propres à chaque constructeur en des paramètres standards
  - enrichir les données en méta-données (constructeur, type de capteur, numéro de version etc.)
- des transformations pour :
  - convertir les dates dans un référentiel commun (UTC)
  - convertir les valeurs en unités du Système International


DATA PIPE
DATA TALKER
DATA USER
DATA CLEANER
DATA WAREHOUSE

- des systèmes de stockage différents afin de répondre aux différents besoins Métiers :
    - Datawarehouse pour des besoins analytiques et d'analyse en mode "batch", pour des populations d'analystes
    - RDMBS pour des besoins d'exposition de la donnée vers des systèmes nécessitant une forte réactivité (ex : site Web), pour des utilisateurs finaux
    - Fichiers plats pour des besoins d'analyse par les Data Scientists

Une fois les données stockées dans le système, l'équipe de Data Scientists est intervenue afin d'élaborer des algorithmes de maintenance prévisionnelle, à partir de données parfaitement nettoyées et validées.

Cette mission a nécessité des compétences en architecture (type Lambda), du PySpark pour le traitement des données volumineuses, l'utilisation de middleware messages (MOM) pour l'ingestion et le stockage temporaire des valeurs de capteurs, l'utilisation de bases de données SQL et NoSQL pour le stockage des données exposées aux utilisateurs.

B. MISSION #2 - TRAITEMENT DE DONNÉES BATCH

Pour le compte d'un client dans le domaine des télécoms, nous sommes intervenus afin de réconcilier différentes sources de données, provenant d'environ 200 systèmes différents (base clients, données du Service Clients, les usages, la navigation Web, la consommation des différents produits etc.).

L'objectif est de fournir une vision à 360° de chaque client.

Les données sont traitées "par lot" (batch), du fait des contraintes des différents systèmes en amont (les données sont extraites en masse, pendant la nuit).

Le challenge est de réconcilier en l'espace de quelques heures l'ensemble de ces données hétérogènes pour n'avoir au final (entre autres), qu'une unique table contenant la liste des quelques millions de clients et des attributs les caractérisant (soit environ 200 informations).

Cela nécessite un réel travail en équipe parmi des équipes Métiers et IT car chaque système porte sa propre logique Métier (par exemple, un client pour la Direction Financière est une personne physique qui paie une facture ; un client pour un vendeur est une personne physique qui s'est abonnée à un service ; un client pour le réseau est une personne physique qui a un accès provisionné…) et son propre identifiant de client (un numéro de téléphone, une adresse mail, un identifiant de connexion, un numéro de contrat etc.).

Ce travail a également permis aux Data Scientists de travailler sur des données bien préparées avec des règles métier connues et partagées, afin qu'ils se concentrent uniquement sur leur compétence propre : l'Intelligence Artificielle.

Outre les compétences techniques essentielles (gros volumes, contraintes temporelles), des compétences fonctionnelles sont nécessaires pour comprendre les concepts manipulés.

Cette mission a nécessité des compétences en PySpark sur environnement HADOOP Cloudera.

La **maîtrise des concepts théoriques du Big Data et des techniques associées** est un premier différenciateur incontournable dans la palette du Data Engineer, il se doit de connaître aussi bien les technologies de stockage que celles de processing des données volumineuses en batch ou en temps réel.

Sa connaissance des principes et des **pratiques DevOps** est également typique. En tant que constructeur de traitements ayant pour vocation à s'exécuter en production et à délivrer un niveau de service fiable à ses utilisateurs, le Data Engineer doit maîtriser les techniques de tests automatisés, de déploiement continu et d'ordonnancement des traitements. Ces compétences sont particulièrement adaptées aux projets de Data Science qui nécessitent des processus de livraison courts et agiles ainsi que des problématiques de suivi de la dérive et de réentrainement périodique des modèles.

Le **cloud et les services associés** sont aujourd'hui un incontournable levier d'efficacité dans la réalisation des projets Data, le Data Engineer se doit de savoir tirer profit de ces infrastructures. Sans forcément devoir être certifié sur l'ensemble des "cloud-providers" du marché, une connaissance critique de la grille des solutions proposées ainsi qu'une expérience dans leur utilisation sont de vraies atouts.

C. MISSION #3 - ENVIRONNEMENTS DE TYPES "ON-PREMISE", "CLOUD" OU "HYBRIDE"

Selon les clients, nous sommes amenés à intervenir sur différents types d'environnements :

- "On-premise" ; le client possède son propre data center et ses infrastructures informatiques.
- "Cloud" ; le client "loue" des ressources informatiques à des fournisseurs spécialisés (OVH, Amazon Web Services, Azure, Google Cloud Platform)
- "Hybride" ; le client possède son infrastructure informatique mais pour certains besoins (ou dans un but de migration), le client "loue" des ressources sur le "Cloud".

Opérer sur ces différentes infrastructures ne nécessite pas les mêmes compétences.

Sur les technologies de type "Cloud", il est nécessaire de bien appréhender les technologies mises en oeuvre par ces "Cloud providers" et de développer dans la philosophie du produit, afin que le bénéfice en retour soit le plus important possible ; cela impose une très bonne connaissance de l'écosystème proposé par le Cloud provider et de bien respecter les best-practices (version du langage de programmation et uses-cases supportés notamment). Chez Amazon Web Services, il existe notamment plus de

100 services, dont 10 rien que pour le stockage des données. Les formations proposées par ces "Cloud Provider" sont un excellent point de départ pour appréhender ce type d'environnement.

Sur les technologies de type "On-premise", le choix des services (catalogue de services) est généralement plus restreint que pour le "Cloud" car la DSI préfère se concentrer sur quelques technologies clés et ne pas multiplier les outils pour des usages proches (pour des questions de coût et de compétences internes).

Il est donc important de bien comprendre l'architecture en place et les outils disponibles avant de proposer une solution. Généralement, les grandes briques logicielles installées sont des outils bénéficiant d'un support éditeur et soumis à licence (bien que des offres "open-source" émergent en complément des offres traditionnelles).

Sur les technologies "Hybrides", le "coeur" du SI et la production de données restent généralement "On-Premise" et certains traitements (réalisation de Proof Of Concept, lourds traitements Big Data etc.) peuvent être réalisés sur le "Cloud", afin de bénéficier d'une puissance "élastique" de calcul, ou d'une infrastructure temporaire. Ce type d'architecture permet aux entreprises, généralement des grands comptes, de maîtriser certains aspects, notamment réglementaires, tout en bénéficiant de la souplesse du Cloud. Toutefois, des contraintes surviennent, comme la latence et le débit du réseau entre les 2 environnements ; à prendre en compte dans la solution retenue !

L'utilisation de multiples **API de collecte de données** voire le **scrapping** (extraction automatique de contenus présents sur le Web) d'informations utiles sur le net ne font pas peur à notre Data Engineer, bricoleur et débrouillard, qui peut trouver ici un moyen d'aller capter ces données exogènes (produites à l'extérieur de l'entreprise) tant convoitées par ses collègues Data Scientists.

Une autre mission du Data Engineer est de participer au **design de la solution** sur laquelle il travaille, en autonomie ou en collaboration avec un Data Architecte. Ce type de tâche nécessite une prise de recul pour appréhender la solution de bout en bout, identifier les points durs et proposer une solution technologique en adéquation avec les besoins du client (compétences, performances, coût, …)

Le Data Engineer joue également un **rôle de formateur ou de lead technique** auprès de ses collègues, qu'ils soient Data Scientists, Analysts ou même chef de projet. Il est important de savoir partager, expliquer et justifier des choix techniques, des méthodes de développement ou des difficultés techniques pour assurer la réussite du projet.

Enfin, il est également un **acteur clé de la Data Gouvernance**. En étant au plus proche de la donnée, il est le plus à même d'évaluer la qualité et l'accessibilité de la donnée et doit être force de proposition pour gagner en maturité. Il peut également accomplir des tâches de Data Steward pour contribuer au management de l'information.


meet-up
sea of data
keyboard cemetery
IT volcano
unix river
data science
data forest
scalatown
mooc
brain lands
python mountains
programmation
hadoop bay
data earth
golf of pocs
java wood
devops hills
big data
développement
sql island

4. UN PARCOURS DONT TU ES LE HÉROS

Il n'existe pas encore de voie royale pour accéder au poste de Data Engineer. Plusieurs parcours sont néanmoins possibles pour arriver à ce poste tant convoité par les entreprises.

Pour les étudiants voulant se former au Big Data, ces derniers peuvent choisir un parcours orienté informatique, leur permettant d'acquérir des compétences dans diverses technologies (C#, C++, Python, Java/J22E, SQL, NoSQL, Hadoop, MongoDB…).

Les profils seniors développent leurs compétences « sur le tas » par de l'autoformation (MOOC, cours du soir), lors de projets professionnels ou lors de formation en entreprise.

A. TOOLBOX

La boîte à outlis reste un élément essentiel à tout Data Engineer. Celle-ci est différente pour chacun, en fonction des appétences, des sujets favoris ou encore des méthodes.

La constitution d'une tool-box se fait dès la phase de formation, puis tout au long de la carrière du Data Engineer. Pour être à même de l'alimenter de façon continuelle, plusieurs réflexes doivent être adoptés. Pour se perfectionner, il est nécessaire d'avoir une activité de veille importante abordant les technologies utilisées en Data Engineering. Cela peut se faire à travers des blogs d'entreprises référentes dans le monde de la Big Data, des publications de professionnels spécialisés en Big Data ou des réseaux sociaux (Twitter, LinkedIn, Medium....)

D'autres façons de se tenir au courant des avancés en Big Data existent, telles que la participation aux meetups (Data Engineer Paris, Paris NLP, Paris Data Geeks) et conférences (Devoxx). En effet, aujourd'hui, ces derniers sont légion et peuvent aborder des sujets très généralistes, comme des sujets très précis et pointus, permettant de monter en compétences. Au-delà du meetup/conférence, cela permet de se forger un réseau en rencontrant d'autres individus passionnés ayant les mêmes centres d'intérêt : la Big Data.

Pour finir, un site web de référence que tout bon Data Engineer se doit de connaître : Stack Overflow. Ce dernier s'adresse à l'ensemble des développeurs cherchant une réponse à leurs problématiques techniques. Indirectement, il permet une montée en compétences dans le sens où d'autres développeurs se conseillent en proposant des pistes, ce qui permet ainsi d'améliorer son projet et donc indirectement ses méthodes.

B. LES M.O.O.C.

Enfin, les M.O.O.C. (Massive Online Open Courses) sont devenus de véritables outils de référence en termes d'apprentissage. Ces derniers se sont imposés dans l'univers de la Data. Plusieurs plateformes (Coursera, OpenClassrooms) permettent de s'auto-former ou

de renforcer ses connaissances sur un langage précis, une technologie Cloud…

Voici quelques exemples de M.O.O.C essentiels en Data Engineering :

- " Big Data Analysis with Scala and Spark "
- " Functional Programming Principles in Scala "
- " Spécialisation Data Engineering on Google Cloud Platform "

Bien entendu, d'autres cours sont disponibles et plus spécifiques en fonction du besoin ou de la tâche que le Data Engineer devra effectuer.

En résumé, travailler sur des sujets Big Data nécessite la mobilisation de nombreuses compétences, tant en développement, qu'en gestion de projet ou bien encore en DevOps. Il est nécessaire pour mener à bien une carrière de Data Engineer de vouloir se former en permanence.

TOUTE CHOSE COMMENCE PAR UN CHOIX...

Comme dirait Morpheus dans Matrix Reloaded, "everything begins with choice..." Et c'est en effet le cas. Au delà de choisir une formation ou un MOOC, il y'a la question du choix pour exercer son métier de Data Engineer. Tout le monde sera d'accord sur le fait, qu'aujourd'hui, la Data représente une fabuleuse opportunité pour les entreprises, permettant ainsi à une grande partie des profils Data Engineer, entre autre, de pouvoir exercer leur profession dans de multiples secteurs (santé, transports, banque, médias, divertissement, énergie….), différents types d'entreprises (grand groupe, conseil, PME, start-up, ENS) et différents projets (sur un produit, services à un client).

Le Data Engineer est au coeur de la transformation digitale de l'entreprise et ses activités sont de plus en plus portées par l'industrialisation des modèles d'Intelligence Artificielle.

Son recrutement est donc essentiel pour toute entreprise se réclamant "Data-Driven" !

Le rôle du Data Engineer est, dans son domaine d'intervention, très large et sur tout le cycle de vie du projet :

- En amont, afin de comprendre l'environnement du client (Cloud, On-premise), l'urbanisation du SI et les différentes briques le composant
- Pendant le projet, pour comprendre les besoins et exigences du Métier, afin de proposer une solution appropriée
- Après le projet, afin de valider l'exploitabilité de sa solution

Enfin, d'un point de vue technologique, le Data Engineer doit réaliser une veille constante sur les technologies émergentes, car dans ce domaine, les initiatives sont nombreuses mais certaines sont abandonnées. Il est donc important de bien comprendre "le marché" pour réaliser une solution à partir de briques logicielles pérennes et performantes.

Le Data Engineer et le Data Scientist interviennent sur des étapes distinctes :

1. Le Data Engineer intervient sur l'acquisition, manipulation/traitement de la données en vue de son stockage, dans un format et une modélisation appropriés pour le Métier.
2. Le Data Scientist intervient à partir des données préalablement mises à disposition par le Data Engineer, en vue d'élaborer des modèles répondant aux demandes du Métier.
3. Le Data Engineer se charge alors d'industrialiser les modèles réalisés, afin qu'ils répondent aux exigences de la DSI.

Selon les profils et les compétences des collaborateurs, il n'est pas rare qu'un Data Engineer puisse réaliser les activités d'un Data Scientist (et inversement) ; tout est une question de goût et compétences !

En termes de perspectives, un Data Engineer a les bases nécessaires pour évoluer vers un poste de Data Architect, afin d'être responsable de la conception, du déploiement et de l'administration de plateformes de calculs et de stockage distribués, afin de répondre aux besoins fonctionnels émis par le Métier.

En conclusion, le métier de Data Engineer est en train de transformer et de façonner de façon significative le monde de la Tech. C'est une profession en devenir qui a un énorme potentiel d'évolution, d'importance et d'impact sur les entreprises et qui se trouve au carrefour de l'innovation.

QUANTMETRY - BUILDING AI WITH PIONEERS

Qu'en est-il pour les Data Engineers au sein de Quantmetry ? Et bien, c'est une vraie aventure apprenante qui s'ouvre à ceux qui tentent l'expérience, en rejoignant l'équipe des pionniers de l'IA.

Durant nos missions, il n'est pas rare de voir nos Data Architects et Data Scientists venir épauler nos Data Engineers lors de sprints agiles pour préparer les données et les mettre à disposition en co-construisant des flux de données ou des "Data Pipeline". Une partie sera dédiée à la mise en valeur des restitutions avec des outils de DataViz ou des développements custom (Dash, Bokeh, ...).

Les Data Engineers de Quantmetry, ont l'opportunité de travailler sur la construction de l'architecture Big Data qui va héberger les traitements (Infrastructure as code sur le cloud, configuration d'une distribution Hadoop sur un cluster de QBox…).

Travailler à Quantmetry signifie que vous prenez part à l'ensemble de la mission. C'est-à-dire de l'identification des sources hétérogènes en phase de cadrage, à la compréhension et l'implémentation des règles Métiers pour finir à la participation des ateliers de restitutions, vous permettant ainsi d'être en interaction avec le client, le Métier et la DSI de l'entreprise.

Nos Data Engineers ont également **une âme de chercheur (#testfailandlearn)** : à travers un binôme de choc avec un Data Scientist, ils participent aux **sujets de R&D alignés avec nos offres** (Image, NLP, …) pour imaginer **les solutions d'IA at scale de demain**.

Pour alimenter leur Tool-Box et rester au fait des nouvelles technologies, nos Data Engineers participent ou donnent des cours à travers, entre autres, nos formations ou nous BBL, des moments en petits comités pour monter en compétences tout en mangeant des burgers, salades et pizzas…

LA BOUSSOLE DU **DATA ENGINEER** CHEZ QM

PROFILS DE DATA ENGINEER

DESSINE MOI UN DATA ENGINEER

Quantmetry
Building AI with pioneers

GRAPHISME ET ILLUSTRATIONS

Aurélien Gomez
Aureliengomez@gmail.com
neografix-creation.com

IMPRESSION & BROCHAGE

SCRIPT LASER
29, boulevard Malesherbes
75008 Paris

Merci à nos contributeurs :

Maya Azouri

Guillaume Bodiou

Victor Berne

Olivier Denti

Justine Deshais

Etienne Fongue

Martin Le Loc