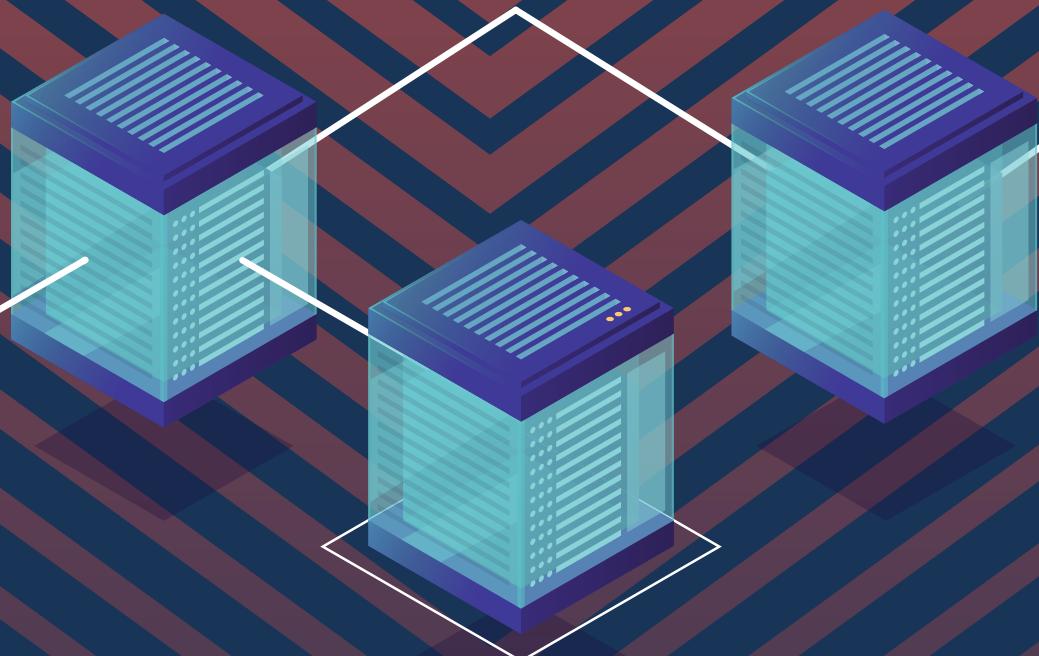




Saagie

Data Fabric



LA DATA FABRIC POUR INDUSTRIALISER LES DATA LABS

Comment passer les initiatives
Big data / IA de l'expérimentation
à la production ?

Livre Blanc

Partie 1

Avant tout, les bases !

Comment le définir ?	5
Quels objectifs ?	6
Quelle équipe constituer ?	7
Pour aller plus loin - Comment l'outiller ?	11

L'expérimentation, comment ça se passe ?

L'importance d'échanger avec les métiers et d'identifier des "quick wins" simples à transformer	13
Le déroulé des POCs (Proof Of Concepts)	15

Partie 2

Amener les travaux du Data Lab en production ?

Ce qui empêche les projets du Data Lab de passer en production	21
Le juge de paix - la Data Fabric	24
Les pratiques DevOps dans la Data Science	32

AVANT PROPOS

85%

des projets Big Data ou intégrant de l'Intelligence Artificielle n'arrivent jamais en production¹

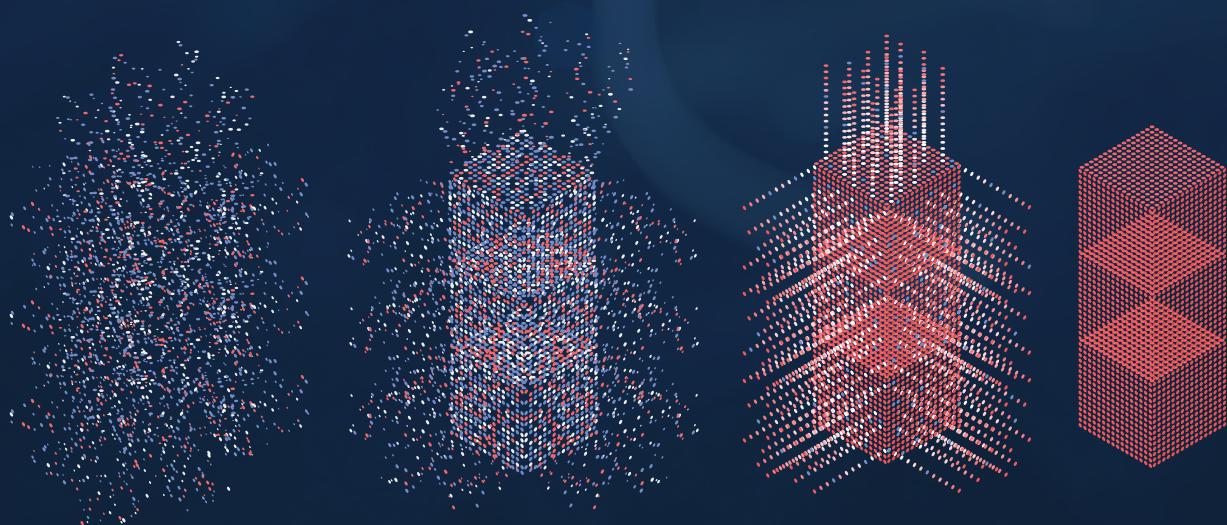
Le Big Data et l'Intelligence Artificielle se démocratisent de plus en plus dans les entreprises, mais tardent à démontrer leur valeur et à tenir leurs promesses. En effet, de nombreuses entreprises ont lancé des initiatives data/IA, mais très peu ont été capables de les amener en production. Et d'après le dernier sondage de Gartner effectué auprès des DSI (CIO Survey 2018), seul 4% des projets / initiatives IA ont été déployés. Cela s'explique par de nombreuses raisons :

Une approche artisanale dans la manière d'appréhender les projets

Des technologies en constante évolution, difficiles à appréhender et à intégrer

Un manque d'alignement entre les différentes équipes concernées par le projet data de l'entreprise

Des usages limités à une problématique spécifique, au lieu de capitaliser sur ce qui a été réalisé et l'étendre à d'autres départements et géographies



L'objectif de ce livre blanc est de vous donner les clés pour faire face à tous ces challenges, et que vous puissiez complètement intégrer l'Intelligence Artificielle au coeur de votre activité !

¹ <https://www.gartner.com/newsroom/id/3466117>

PARTIE 1

AVANT TOUT, LES BASES !

Data Lab

Selon Gartner, indépendamment de la taille de l'entreprise et de son niveau de maturité, les projets de Data Science et IA n'arrivent souvent pas au bout à cause d'un manque d'alignement, de vision stratégique et d'appui de la direction.



En quoi cela consiste ?

Il s'agit d'une entité pluridisciplinaire dont la vocation est d'optimiser l'utilisation des données de l'entreprise et de leur donner une valeur métier, dédiée à l'expérimentation et à la qualification « fonctionnelle » de ces données. Le Data Lab est très utile car il permet de faire évoluer l'architecture existante sans tout casser au nom de la révolution Big Data.

La création d'un Data Lab est une étape essentielle dans votre projet Big Data

Il ne s'agit plus simplement de lire des tableaux de bord mais d'aller plus loin dans l'exploitation des données de l'entreprise en cherchant les signaux faibles, les tendances, les prévisions, les fraudes et autres cas d'usages dont nous parlerons plus loin dans ce livre blanc.

AVANT TOUT, LES BASES !

Pour construire votre Data Lab les étapes importantes sont :

- ◆ La définition des objectifs
- ◆ La constitution d'une équipe
- ◆ La mise à disposition d'outils communs pour réunir votre équipe



Il s'agit d'un changement culturel important pour l'entreprise qui est de comprendre le caractère expérimental de l'approche Data Science, avec une forte approche mathématique et métier au sein du Data Lab. C'est l'un des points les plus complexes à maîtriser dans la construction de votre Data Lab.

AVANT TOUT, LES BASES !

Quels objectifs ?

Pour que la mise en place d'un Data Lab soit un succès, la première condition est de bien définir les objectifs, c'est-à-dire définir le cap !

Un de ces objectifs peut être d'adresser certains cas d'usage métiers : la segmentation clients, la prévision de ventes, l'automatisation des processus internes (par exemple le traitement de facture), ou plus simplement de fournir des indicateurs de performance consolidés pour l'ensemble des entités de l'entreprise.

Si ces objectifs ne sont pas clairement exprimés / fixés, le risque est d'avoir un Data Lab qui tourne "à vide", n'étant pas capable de délivrer de la valeur pour l'entreprise, entraînant un effet déceptif vis-à-vis du Big Data et de l'Intelligence Artificielle.



Quelle équipe à constituer ?

Une fois que les objectifs ont été définis, il convient maintenant de constituer l'équipe du Data Lab. Là aussi, cette étape est critique.

En effet, même si la mise en place d'un Data Lab peut s'avérer complexe d'un point de vue technique (ce sujet sera développé ultérieurement) c'est surtout une question d'organisation et de compétences : le fait de ne pas disposer des bonnes personnes et de ne pas avoir mis en place la gouvernance nécessaire peut entraîner l'échec du Data Lab. Il est donc indispensable de recruter de bons éléments, avec des compétences variées.

AVANT TOUT, LES BASES !

Les profils à rechercher

Les "Data people" ont le vent en poupe et il est de plus en difficile de trouver les bons candidats, mais comme nous le verrons, mieux vaut construire une équipe aux profils variés, chacun ayant sa spécialité, plutôt que chercher à trouver la perle rare maîtrisant l'ensemble des concepts, algorithmes et langages...

Les profils à rechercher sont des candidats ayant :

Des connaissances approfondies des données et du métier permettant d'identifier les différentes sources de données, de déterminer les hypothèses à valider, puis de communiquer sur les résultats.



Une maîtrise technique des outils de traitement de données dédiés à la mise en oeuvre de l'architecture technique, au développement d'interfaces de programmation (API) ou la conception de tableaux de bord / rapports (profil que l'on trouve dans les équipes décisionnelles ou de Business Intelligence). Si vous avez de grosses volumétries et du temps réel à gérer, préférez des profils qui maîtriseront des technologies telles que Spark.

Une expertise mathématique et statistique afin de choisir les meilleures méthodes pour construire un modèle, puis interpréter les résultats d'une analyse. Il s'agira plutôt des profils de Data Scientists

AVANT TOUT, LES BASES !

Les profils à rechercher

L'Administrateur Système, qui permettra la mise en place de l'infrastructure IT.

Le Data Architect qui met en place les choix d'architecture du Data Lab.

Le Data Scientist qui va être en charge de la conception des algorithmes et modèles analytiques et prédictifs.

Le Data Engineer qui assure la mise en place des différentes séquences de traitement de la donnée (on parlera de pipeline de données) jusqu'à leur mise en production.

Les Experts Métiers qui auront un oeil avisé sur les données dont ils auront besoin avec une connaissance des problématiques business, ils doivent être impliqués dès le début du projet et être des sponsors de ce dernier et de votre Data Lab !

Le Product Owner qui aidera les Experts Métier à formaliser leurs besoins et supervisera le développement du produit (ou de la solution) intégrant les modèles analytiques.

Le Business / Data Analyst qui permettra de mettre sous forme de tableaux de bord et graphiques les résultats concluant des POCs et des projets en production.



AVANT TOUT, LES BASES !

Comment outiller le Data Lab ?

Référencer vos sources de données : d'où viennent-elles ? Comment les regrouper ? Quelles sont les difficultés liées à leur utilisation ? Pour y répondre, il est nécessaire de disposer d'une quantité et d'une qualité suffisante de données disponibles, de référencer ces données qui seront ensuite explorées, tout en restant vigilant sur **leur conformité vis-à-vis des réglementations en vigueur (comme le RGPD)**.

Structurées ou non, toutes les données à disposition de l'entreprise peuvent être utiles pour optimiser la réalisation d'un use case. **Les différentes données de votre SI constituent la principale (et la plus évidente) source à exploiter.** Les logs que les utilisateurs laissent lors de l'utilisation de vos applications, de vos logiciels ou lorsqu'ils naviguent sur votre site internet sont également une importante source de données utilisable pour optimiser votre produit, vos ventes...

Certaines entreprises spécialisées dans la revente de données

peuvent aussi constituer une source de données intéressantes avec des informations telles que : données géographiques, socio-professionnelles, habitudes de consommation... Ces entreprises ou organismes publics ou privés diffusent des données peu sensibles, utilisables gratuitement et représentent un moyen supplémentaire d'enrichir votre Data Lake (voir ci-dessous).

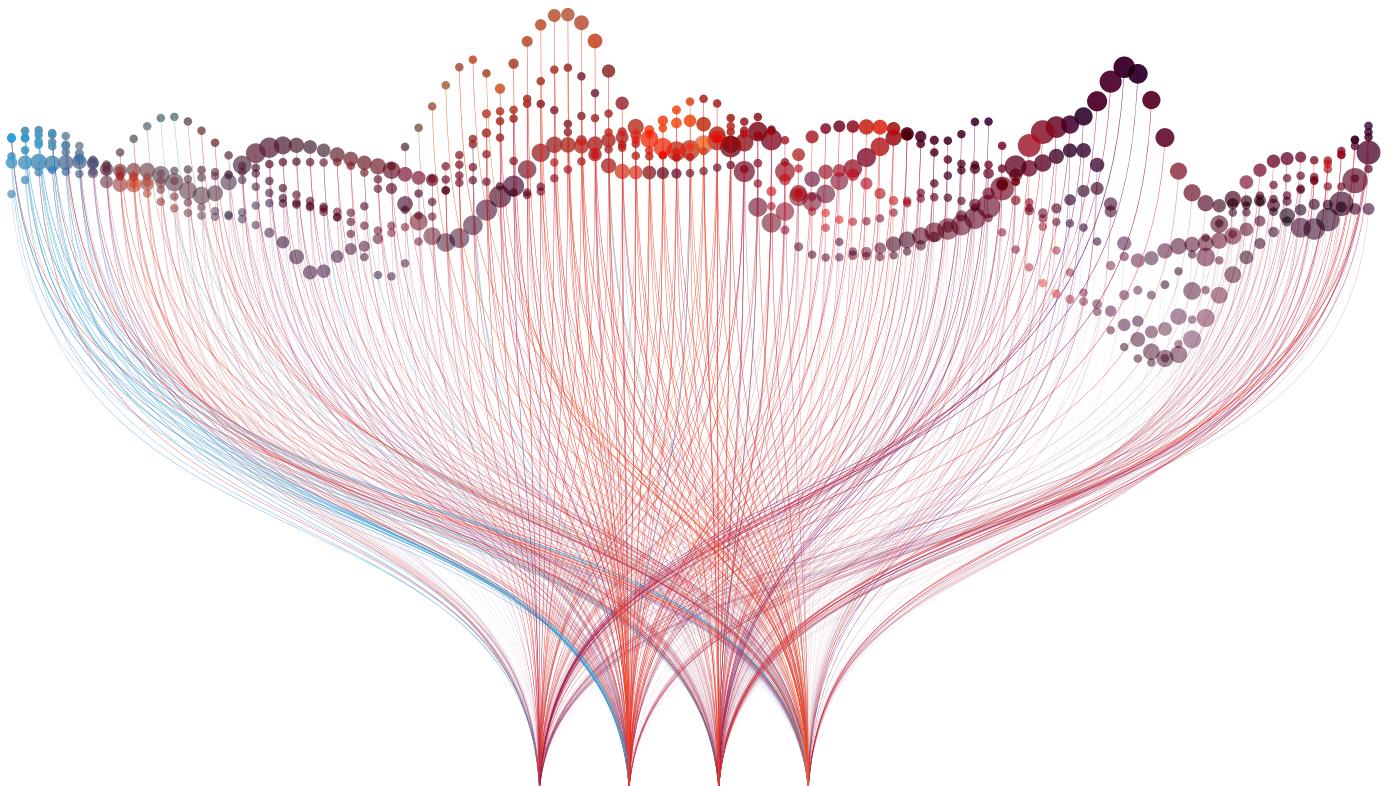
A savoir

Un autre des pré-requis nécessaires est la possibilité pour les Data Scientists d'accéder rapidement à la donnée.



AVANT TOUT, LES BASES !

Mettre en place un Data Lake



Le principe du Data Lake (ou lac de données) est assez simple, **toutes vos données sont regroupées et accessibles en un seul et même endroit**. L'avantage considérable d'un lac de données, comparativement à un Data Warehouse (ou entrepôt de données), réside d'une part dans sa capacité à croiser et enrichir des données beaucoup plus simplement, tout au long du projet, et à en accroître leur disponibilité (même en cas de panne) grâce à un mécanisme de réPLICATION.

AVANT TOUT, LES BASES !

Pour aller plus loin...

L'exploitation d'un Data Lake par le Data Lab implique l'utilisation d'un certain nombre de technologies :



ETL (Extract - Transform - Load) par batch (lot) ou streaming (continu, temps réel)

- SQuOOP
- Spark
- Kafka Stream
- Talend



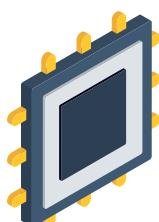
Statistiques et modélisation

- Spark
- Java
- Python
- R



Format de stockage des fichiers

- Avro
- Parquet
- ORC



Technologies liées au stockage de données

- Stockage distribué :
 - Hadoop et son système de fichier HDFS
 - HBase
- Stockage rapide d'accès (SQL, NoSQL) :
 - Cassandra
 - Mongo DB
 - Elasticsearch
 - PostgreSQL

AVANT TOUT, LES BASES !

Pour aller plus loin...

Les technologies de Data Science

Langages de programmation

- Python
- R
- Scala (Spark)

Deep Learning

- Tensor Flow
- Keras
- Pytorch

Interface de programmation (Notebook)

- Jupyter
- Zeppelin

Environnement de développement (IDE)

- R Studio
- Jupyter Lab



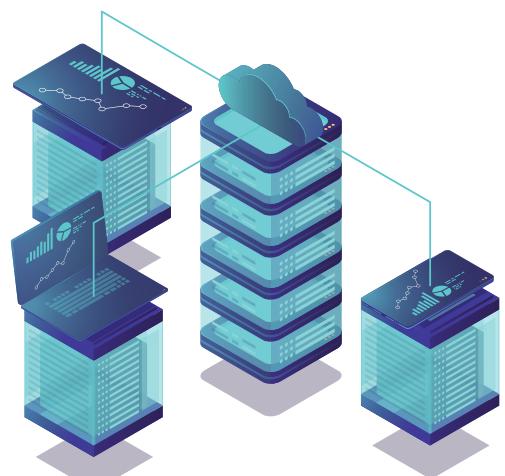
Comme vous pouvez le constater, la réalisation d'un projet Big Data / IA va impliquer grand nombre technologies disparates et peu intégrées.

Leur utilisation va varier en fonction de la composition de votre Data Lab, ce qui va complexifier la chose.

En se déployant sur le Data Lake, la Data Fabric de Saagie va pouvoir outiller l'intégralité du Data Lab.

Comment ?

En proposant un environnement unique où ces différentes technologies vont être pré-assemblées et orchestrées les unes avec les autres tout en étant maintenues à jour, ce qui va grandement simplifier leur mise à disposition pour l'ensemble du Data Lab.

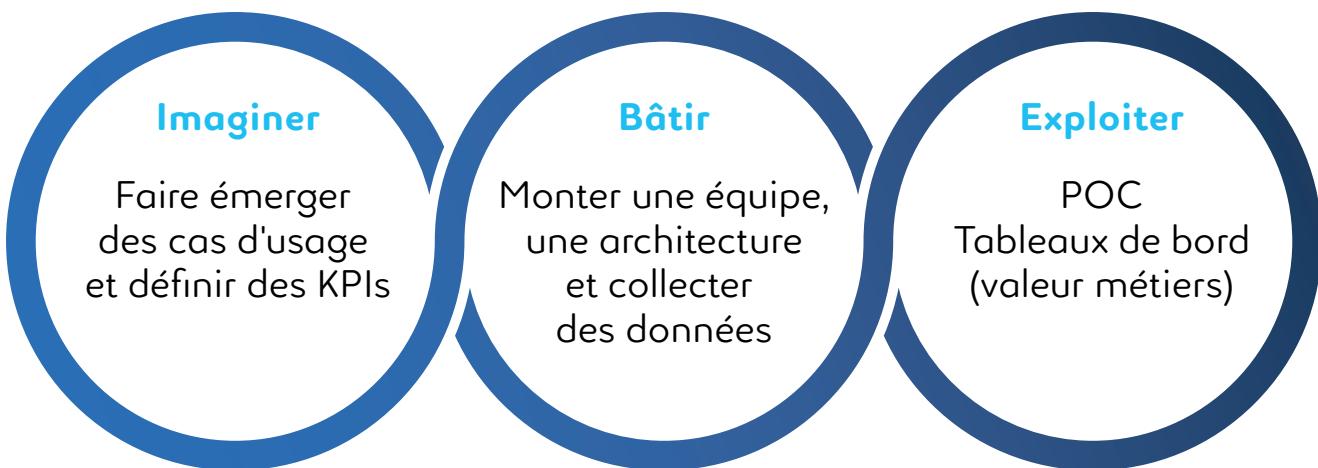


L'EXPÉRIMENTATION, COMMENT ÇA SE PASSE ?

L'importance d'échanger avec les métiers et d'identifier des "quick wins" simples à transformer

Les initiatives autour de l'IA et du Big Data doivent être guidées par un besoin métier et répondre à une problématique de l'entreprise, sans quoi les projets risquent de ne pas de produire l'effet escompté.

Paradigme



La nécessité d'amener une vision Business dans le Data Lab

Il est important de créer une relation forte entre les métiers et le Data Lab car la finalité du projet sera d'autant plus pertinente et apportera plus rapidement des résultats.

En résumé, un projet Big data / IA, c'est :

90% de design thinking* pour faire émerger vos cas d'usage

*Le Design Thinking une approche consistant à appliquer les méthodes et la philosophie utilisées par les designers pour résoudre certains problèmes. Il s'appuie en grande partie sur un processus de co-créativité impliquant des retours de l'utilisateur final.

L'EXPÉRIMENTATION, COMMENT ÇA SE PASSE ?

La sélection des cas d'usage va se dérouler de la manière suivante :

Divergence

- ◆ Faire émerger le plus grand nombre de cas d'usage

Convergence

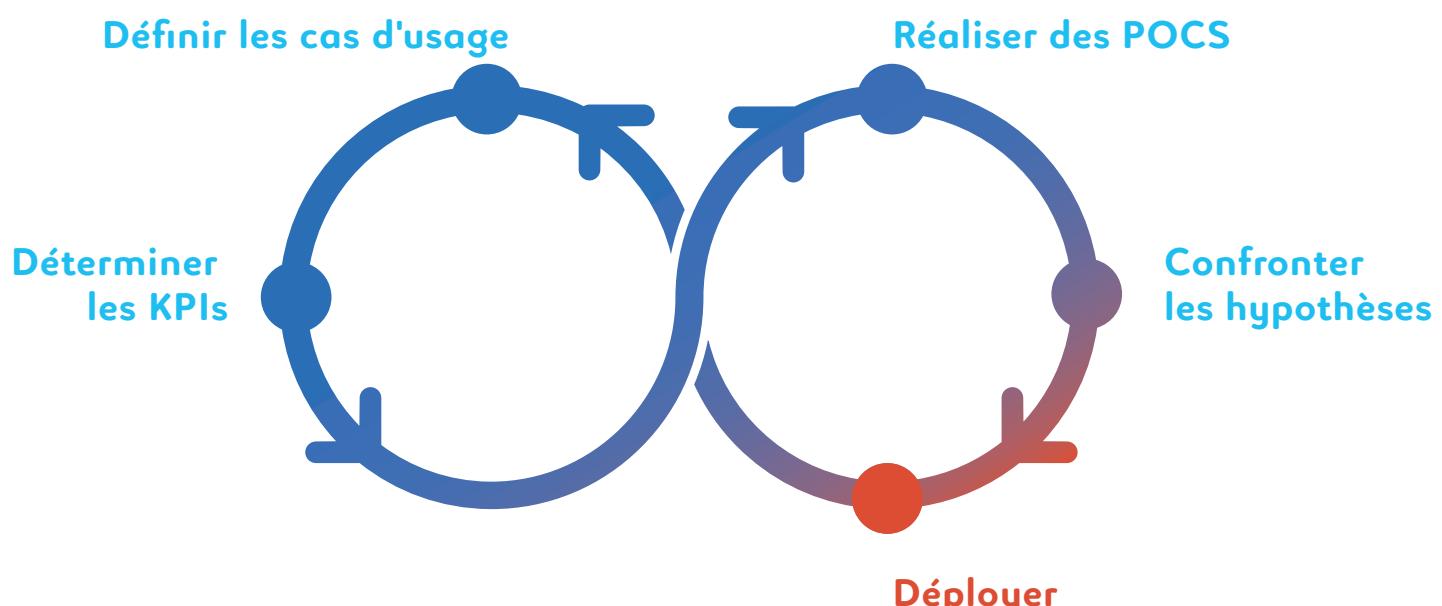
- ◆ Identifier les cas d'usage pertinents
- ◆ Identifier les cas d'usage réalisables

Quelques questions sont à considérer pour sélectionner les cas d'usage :

Le cas d'usage produit-il une valeur ajoutée métier ? Des KPIs peuvent-ils être définis pour mesurer la valeur métier créée ?

Le POC (Proof Of Concept) que je vais réaliser pourra-t-il intégrer des processus métiers opérationnels ?

Autre point à considérer : plutôt que d'adresser une multitude de cas d'usage en même temps, il est préférable de se focaliser sur un ou deux cas d'usage, afin de démontrer rapidement de la valeur et se familiariser avec les nouvelles approches qui seront mises en place.



L'EXPÉRIMENTATION, COMMENT ÇA SE PASSE ?

Le déroulement des POCs

Produire de la valeur répondant à vos use cases métiers à partir de votre Data Lake n'est pas si simple pour le Data Scientist. Lors du POC une collaboration étroite entre le Data Scientist et le Data Engineer est essentielle et celle-ci se déroule en deux phases :

- ◆ **Le pré-traitement (ou preprocessing)**, qui consiste dans le nettoyage, l'exploration et surtout la compréhension des données de l'entreprise. Cette partie est souvent assez longue mais essentielle pour que les Data Scientists puissent travailler la donnée.
- ◆ **L'algorithme qui passe par la construction de modèles descriptifs ou prédictifs visant à apporter de la valeur ajoutée.**



L'EXPÉRIMENTATION, COMMENT ÇA SE PASSE ?

Pour aller plus loin



Compréhension

Il est indispensable de travailler avec les Experts Métiers pour comprendre les données vraiment utiles, et déterminer si l'on a besoin d'utiliser des données externes à l'entreprise (par exemple l'Open Data)



Exploration

Cette étape va permettre de déterminer si les données sont intéressantes et fiables. Elle va mettre en évidence les données manquantes (anomalies, champs vides) et décrire les variables (avec des indicateurs tels que la moyenne, variance, quartile, classe, saisonnalité) nécessaires à l'élaboration de KPIs.

L'EXPÉRIMENTATION, COMMENT ÇA SE PASSE ?

Pour aller plus loin

La préparation de données (Data Preparation) et les étapes de pre-processing et feature engineering

Preprocessing

Cette étape consiste à supprimer des variables trop faibles ou avec trop de champs manquants, mais aussi des valeurs aberrantes (ex : 0 quand la data n'est pas disponible). C'est une tâche qui peut s'avérer très longue et laborieuse.

C'est pourquoi Saagie, en tant que solution ouverte sur son écosystème, intègre des applications partenaires pour simplifier cette opération.



Feature engineering

Pour compléter la préparation des données on applique très souvent une étape de "feature engineering". Cette étape consiste à créer de nouveaux jeux de donnée et à calculer de nouvelles caractéristiques utiles aux futurs modèles* sur la base de données actuelles. Il est parfois utile de calculer des données en amont pour optimiser les données utiles. Durant cette période, il y a bien entendu de nombreux échanges avec les métiers pour affiner la compréhension des données.

En général, 80% du code de développement va être utilisé pour faire du "feature engineering".

Cette étape de préparation est bien souvent la plus longue dans un projet Big Data / IA / Data Science. On estime qu'elle prend au minimum 75% du temps du projet.

*Le modèle est le résultat d'un algorithme de Data Science appliqué à un jeu de données particuliers

L'EXPÉRIMENTATION, COMMENT ÇA SE PASSE ?

Pour aller plus loin

Valorisation des données (création de modèles et analyse de résultats)

Les algorithmes de Machine Learning (apprentissage automatique) tels que les méthodes régression, de regroupement ou encore de classification vont permettre d'adresser différents cas métiers et générer ainsi de la valeur;



La prédiction (ou régression), à partir des historiques il est possible de prédire les futurs ventes ou stocks, et d'optimiser la gestion de l'offre et de la demande. D'autres données externes et libres d'usage (les données Open Data notamment) peuvent aussi être utiles comme par exemple la météo qui peut avoir un impact dans certains secteurs (comme l'agriculture), ou par exemple le trafic routier, les périodes de vacances, etc.



Les anomalies, il s'agit ici de détecter les données qui ne sont pas cohérentes avec l'ensemble d'un jeu de données. Il s'agit ensuite de déterminer si l'anomalie est avérée ou aberrante. La recherche d'anomalies est particulièrement utilisée dans les cas de détection de fraude.



La segmentation (ou clustering), ici on va séparer et regrouper des données parcellaires mais avec des caractéristiques identiques. Par exemple identifier pour des régions spécifiques des clients ayant un salaire inférieur à 30 000€ par an pour promouvoir certaines offres.

Il est souvent nécessaire de transformer ces données vers des formats optimisés pour la création d'algorithmes, tel que le framework Spark qui utilise la mémoire vive des clusters Hadoop ; il est aussi possible d'utiliser un moteur de requête tel que Impala ou Drill. D'autres frameworks, comme Spark Streaming, vous permettent une gestion complètement dynamique afin de préparer et traiter les données en temps réel dès leur arrivée dans le Data Lake.



L'EXPÉRIMENTATION, COMMENT ÇA SE PASSE ?



Conseil : adopter la méthode agile !

"Fail Fast, Try Again"

Pour le bon déroulement du projet, une démarche agile (par exemple Scrum ou Kanban), consistant en échanges réguliers et confrontation des résultats avec les équipes métier (expert métier & Product Owner) est conseillée. Cela créera une dynamique itérative permettant de faire rapidement progresser le projet (amélioration continue).

PARTIE 2

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION

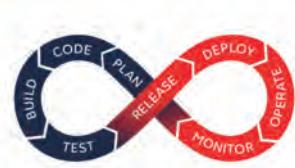


AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Ce qui empêche les projets du Data Lab de passer en production

Concevoir des algorithmes est d'ordre académique, les déployer est d'ordre économique

Data Science Lab



Data Science Production



Promote

Fail Fast, Try Again

Toute la difficulté réside dans le fait de passer d'un environnement où l'innovation, l'exploration et la pro-activité règnent en maître vers un environnement dit de production, répondant aux critères de stabilité, de sécurité et de gouvernance de l'entreprise, où les différents algorithmes vont pouvoir être déployés, supervisés et où les différentes applications métiers vont pouvoir être développées.

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Ce qui empêche les projets du Data Lab de passer en production

L'approche Shadow IT

De nombreuses initiatives Big Data sont mises en place sans que les équipes IT soient informées ou impliquées. C'est ce que l'on appelle l'approche "shadow IT". Malheureusement, lorsqu'il faut basculer les travaux en production, cela est très difficile car les solutions choisies bien souvent ne répondent pas aux critères du département IT qui refuse alors de déployer ces travaux.

Vouloir concevoir sa propre plateforme Big Data pour gérer de bout en bout ses projets

Cette approche rencontre de nombreuses difficultés techniques : elle implique d'assembler et d'intégrer de nombreuses technologies disparates, et d'être en capacité de les maintenir à jour de manière très régulière. Cela complexifie grandement la conception de cette plateforme, et bien souvent le ROI n'apparaît que très tardivement.

A titre d'exemple, Uber a mis environ 18 mois pour développer sa plateforme Big data. Cela vous donne une idée de l'ampleur du chantier, sachant que Uber fait partie des entreprises les plus innovantes de la Silicon Valley*.



*<https://eng.uber.com/michelangelo/>

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Des approches artisanales

Il y a une grosse différence d'outillage entre les technologies dites de Data Science et les technologies plus fréquemment utilisées dans le développement et pour la mise en production. Par exemple, en langage Python, il existe des librairies de modélisation avancées (Scikit-Learn) que l'on ne retrouve pas dans une technologie comme Java.

Ce qui rend difficile

- le déploiement des travaux à une plus grande échelle
- la reproductibilité de ces travaux pour le reste de l'entreprise

Et cela peut amener dans certains cas des développeurs à réécrire complètement le code des data scientists entraînant une énorme perte de temps et une diminution de la productivité.

La bunkerisation du Data Lake

A vouloir trop sécuriser l'accès du Data Lake, on peut se retrouver dans une situation assez ubuesque où

- aucune donnée n'entre
- aucune donnée ne sort
- aucun cas d'usage n'est adressé !



AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Le juge de paix - la Data Fabric

Big Data, Data Science, intelligence artificielle... Si ces termes gagnent chaque jour en popularité, peu d'initiatives voient en réalité le jour. De nombreuses entreprises ont des projets et cas d'usage à adresser, mais encore faut-il pouvoir les mettre en production. La Data Fabric apparaît alors comme une solution prometteuse. C'est un concept qui a émergé outre-atlantique ces derniers mois dans la presse spécialisée, avec plusieurs définitions (Forbes, NetworkWorld...). Nous allons vous expliquer dans quelle mesure elle peut vous aider à concrétiser vos projets Big Data et IA.

Qu'est-ce qu'une Data Fabric ?



Par manque d'expertise, de temps, de technologie ou de moyens, rares sont aujourd'hui les entreprises qui peuvent gérer leurs données seules. Elles sont néanmoins nombreuses à avoir réalisé l'enjeu que peut représenter leur traitement. C'est là qu'intervient la Data Fabric. Elle vous permet de gouverner, d'exploiter et de sécuriser vos données en temps réel, mais surtout de développer des applications métiers afin de répondre à vos problématiques.

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Une Data Fabric est une solution logicielle de gestion de données.

Disponible dans le cloud et "on-premise" (sur site), elle accélère la transformation numérique des entreprises en facilitant la mise en production des projets Big Data / IA.



A l'intersection entre les plateformes de Data Management, de Data Science et le Data Lake, elle représente un ensemble cohérent de solutions logicielles et applicatives, indépendantes des choix de l'architecture de l'IT. Elle offre une solution plus complète en permettant de gérer de bout en bout le cycle de vie de vos données : collecte, stockage, traitement, modélisation, déploiement, supervision, gouvernance. Ainsi, de nombreuses problématiques métiers vont pouvoir être adressées grâce à l'ensemble de technologies disponibles au sein d'une Data Fabric.

L'autre avantage d'une Data Fabric, c'est sa capacité à proposer une vue différente sur les données de l'entreprise, vue qui va pouvoir ensuite être partagée à l'ensemble des équipes. Des profils moins experts pourront y avoir accès, et apporter aussi une vision métier aux données de l'entreprise.

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

N'est pas Data Fabric qui veut

Selon Dan Kuznetzky, directeur de la recherche chez IDC Group, une Data Fabric doit répondre à ces différents critères :

- ◆ Combiner des données de divers systèmes, quelle que soit leur taille, et les rendre disponibles aux applications tout en garantissant vitesse et fiabilité.
- ◆ Offrir un accès aux données aux systèmes, quelle que soit leur localisation : dans le Data Center, sur des environnements cloud, ou à la périphérie du réseau.
- ◆ Offrir un environnement unifié : les documents doivent y être facilement accessibles, la sécurité doit y être garantie et la capacité de stockage doit être suffisante.



La Data Fabric n'est pas une plateforme Data Science

On pourrait s'y méprendre, mais Data Fabric et plateforme de Data Science sont bien deux outils distincts. Dans une vision simpliste, une plateforme de Data Science sert à développer des algorithmes afin de concrétiser des projets d'Intelligence Artificielle et plus particulièrement de Machine Learning ou de Deep Learning. Elle n'est pas toujours adaptée aux profils métiers pour qui les algorithmes doivent au préalable être intégrés dans une application pour être interprétés. En revanche, la Data Fabric est un véritable écosystème qui permet la gestion des données, de leur extraction jusqu'à leur consommation, en passant par leur traitement. Contrairement à la plateforme de Data Science, son but premier est la mise en production des projets Big Data / IA de l'entreprise.

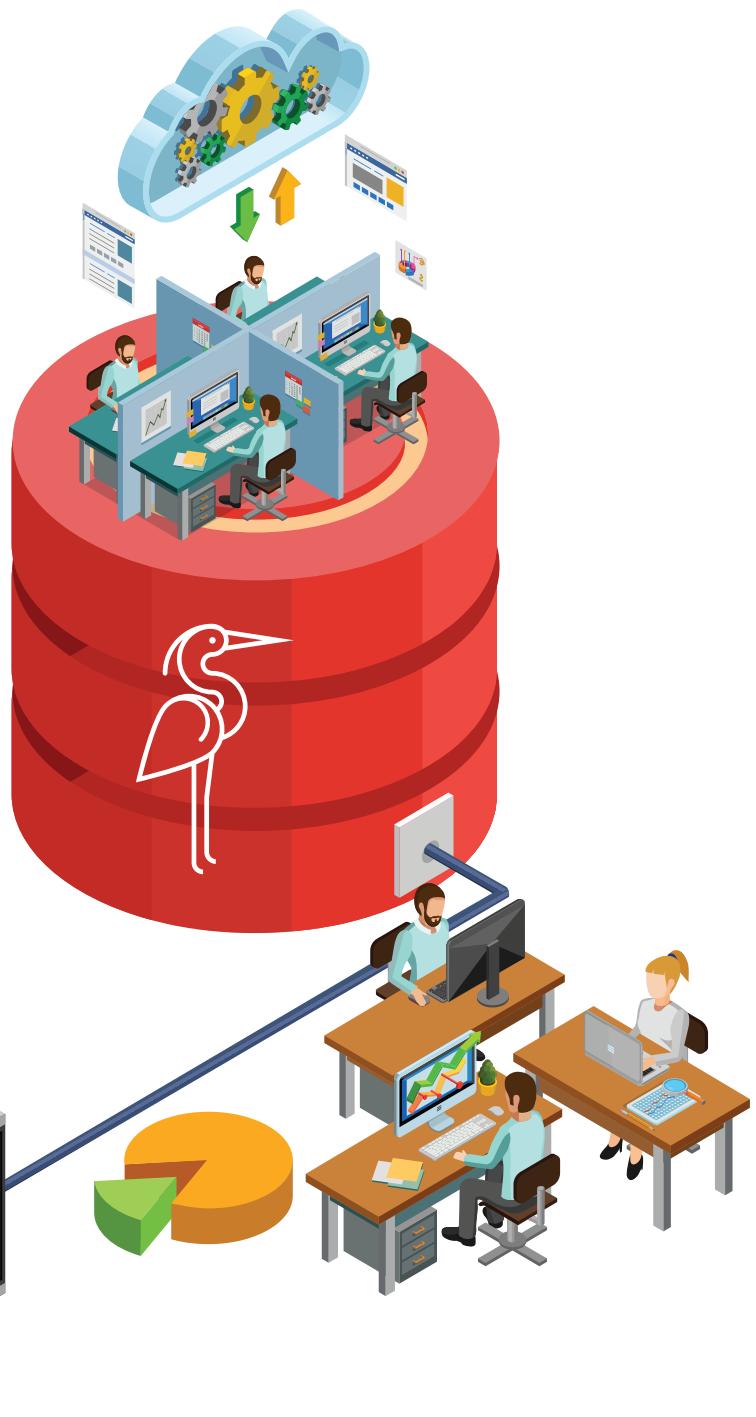
Toutes les technologies data y sont assemblées, les profils métiers peuvent facilement accéder aux données, les profils plus techniques profitent de son ouverture qui leur permet de travailler sur n'importe quel langage (R, Python...). Pour faire simple, la Data Fabric, plus complète, pourrait englober une plateforme Data Science.

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Pourquoi choisir la Data Fabric de Saagie ?

Pour son ouverture :

à titre d'exemple, la Data Fabric de Saagie supporte de multiples technologies comme HDFS, Impala, Hive, Drill, Spark, Sqoop, Elasticsearch, PostgreSQL, Talend, Java, Scala, R, Python, Jupyter, Docker, Zeppelin, Mongo DB et MySQL. Elle offre aussi une compatibilité complète avec les dernières versions de ces technologies, mais aussi avec des versions moins récentes. La Data Fabric se charge ensuite d'assurer la cohésion entre ces différents outils et technologies.



AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Pour sa flexibilité : les traitements sur les données y sont rendus possibles, peu importe l'endroit où ces données sont hébergées (sur site ou dans le cloud public - Amazon Web Services, Microsoft Azure et Google Cloud). La Data Fabric apparaît donc comme une alternative viable face aux plateformes de Data Management et va pouvoir s'adapter à de nombreux cas d'usage.

Pour ses capacités de gouvernance : en facilitant la documentation des données, en historisant les différents traitements réalisés et en contrôlant les accès à ces données, la Data Fabric de Saagie va simplifier la mise en conformité avec le RGPD. En harmonisant les différents processus liés au traitement et à l'utilisation de la donnée, elle permettra aux métiers de disposer de données de confiance, et donc apportera de la valeur pour ces derniers.



AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Enfin, la Data Fabric fédère vos équipes. Elle facilite la collaboration entre les membres de l'équipe du Data Lab (Data Engineers, Data Scientists, Business Analysts, Data Stewards, responsables IT/Ops) et leur fournit les outils leur permettant de mener à bien leurs projets :



Pour les Data Engineers :

la possibilité de créer des pipelines de données pour collecter, nettoyer, traiter la donnée et alimenter les différents modèles préparés par les Data Scientists.

Pour les Data Scientists :

l'accès à une quantité plus large de données et aux dernières versions des langages de programmation; des fonctionnalités pour déployer leurs développements à plus grande échelle.

Pour les Data Analysts :

l'accès simplifié à des données de confiance pour travailler des vues métier.

Pour les Data Stewards :

des outils pour documenter la donnée.

Pour les responsables IT/Ops :

un environnement sécurisé pour amener en production les initiatives data et gérer les accès à la donnée.

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Et pour quelle finalité ?

Les cas d'usage adressés avec la Data Fabric de Saagie sont multiples.

Vous êtes ainsi en capacité de :

- réduire votre taux d'attrition,
- segmenter vos clients,
- optimiser la supply chain,
- améliorer la chaîne de production, entre autres...

Transformation digitale, passage à l'ère numérique... quelle que soit l'expression utilisée, les entreprises sont en train de changer, et le temps presse. Afin d'exploiter leurs données, elles ont besoin d'une solution simple mais complète. En mettant du Devops dans la Data Science, la Data Fabric permet de tirer profit de ses données et rend possible la prise rapide de décisions ciblées par l'exploitation, le tri et l'analyse des données en fonction des métiers.



[CONCRÉTISER VOS PROJETS BIG DATA ET IA >](#)

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

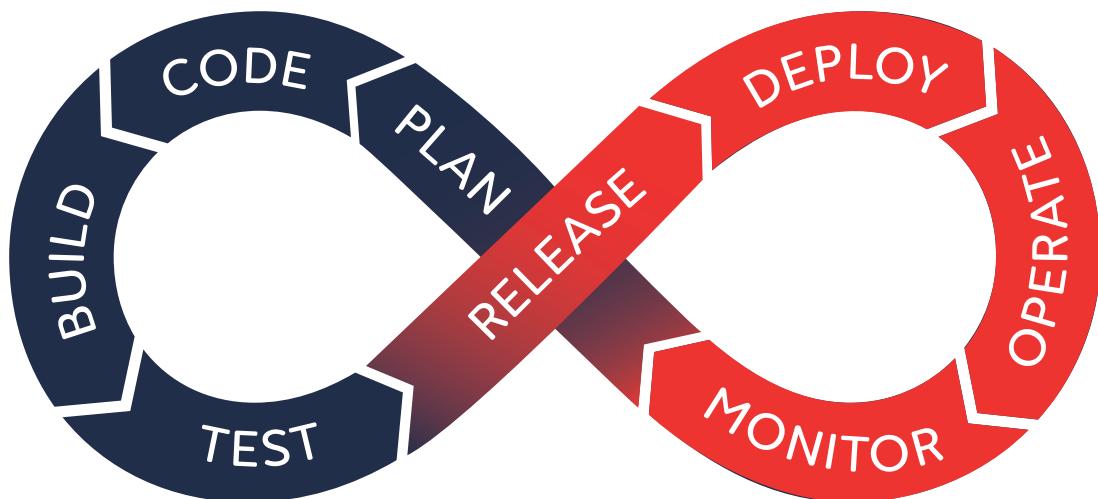
Amener les pratiques DevOps dans la Data Science

Dans cette partie, nous allons vous expliquer comment les pratiques DevOps peuvent faciliter la mise en production des initiatives Big Data et IA et apporter de la valeur à l'entreprise.

Qu'est-ce que l'approche DevOps

Basée sur les principes Lean et Agile, l'approche DevOps rassemble responsables opérationnels et développeurs. On parle de "dev" pour tout ce qui se rattache au développement d'un logiciel, d'"ops" pour l'exploitation et l'administration de son infrastructure. Les pratiques DevOps visent à unifier l'ensemble. En pratique, il s'agit de l'automatisation et du suivi de chacune des étapes de la création d'un logiciel, de son développement à son déploiement, mais aussi de son exploitation dans la durée.

On associe couramment DevOps à agilité qui favorise les cycles courts, l'itération ou encore des déploiements plus fréquents. L'objectif de cette démarche est de délivrer un logiciel en continu et donc modifiable, qui permet à la fois de prendre en compte les retours clients, mais aussi de saisir plus d'opportunités commerciales. Les principaux avantages de ces pratiques sont aussi la collaboration de différentes équipes qui amène à un déploiement accéléré et donc à des coûts réduits.



AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Le modèle CRISP est-il encore d'actualité en 2018

Le modèle CRISP (Cross Industry Standard Process for Data Mining) est communément utilisée par les experts en Data Mining pour résoudre les problèmes qui se posent à eux. Il se découpe en six phases principales :

- **Connaissance du métier**
- **Connaissance des données**
- **Préparation des données**
- **Modélisation des données**
- **Evaluation**
- **Déploiement**

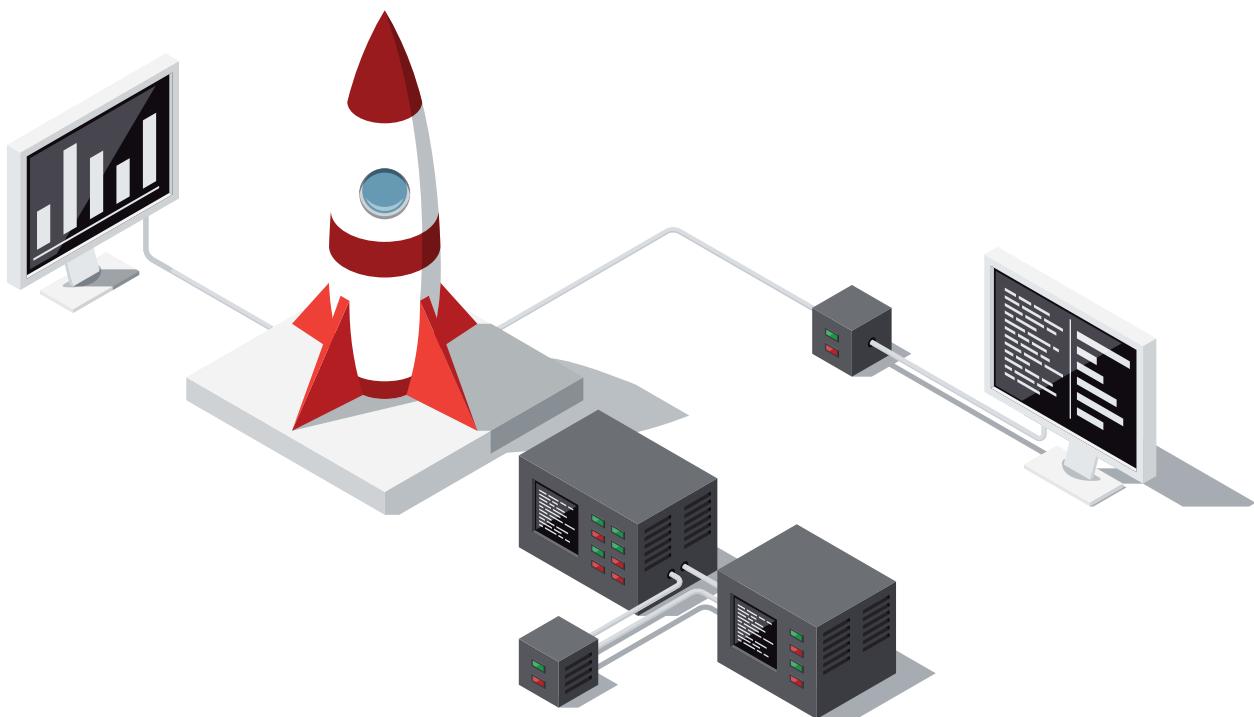


Ce modèle était jusqu'à présent celui qui a prédominé au niveau des équipes Data Science. Cependant, avec l'arrivée des modèles de déploiement continu (Continuous Delivery) et d'innovation continue (Lean Startup), est-ce que cette manière de travailler entre les équipes de Data Science et les équipes IT doit perdurer ? En effet, les désavantages liés à ce modèle sont principalement une certaine longueur au niveau des feedbacks, entraînant un manque de réactivité quant à l'amélioration des travaux et réduisant même leur efficacité.

¹https://fr.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Étapes Release & Deploy



Ces étapes étant complémentaires, nous les présentons ensemble. Pour faire simple, la “release” est la sortie d’une première version stable d’un “package” (ensemble de fichiers informatiques nécessaires à l’exécution d’un logiciel, intégrant par exemple du code et des configurations). Le “deploy” concerne son déploiement dans un environnement spécifique (développement, recette, pré-production, production).

La Data Fabric de Saagie va pouvoir standardiser la manière de déployer un processus traitant de la donnée (traitement)

1. en automatisant le déploiement d’un package
2. avec des environnements préconfigurés pour exécuter les applications
3. en historisant les différentes versions des packages déployés et des exécutions des applications

Cela permet aux équipes Data Lab de pouvoir développer, avec des critères de production, leurs packages et cas d’usage, et les répéter quel que soit l’environnement.

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Étape Operate

L'étape “Operate” de la boucle DevOps va consister à “opérer” les développements précédemment déployés (les traitements). Et cette étape va se décomposer de la manière suivante :



1. Ordonnancer/Orchestrer les traitements (dans le domaine de la data, il y a beaucoup de traitements par "lots", notamment dans le cas d'un apprentissage de modèle de Data Science).

La Data Fabric de Saagie dispose d'un orchestrateur intégré permettant au Data Lab de planifier les différents traitements issus de ses explorations. Cet ordonnanceur va pouvoir combiner les différentes technologies utilisées pour développer les cas d'usage de manière totalement transparente.

2. Superviser l'état de l'ensemble des traitements

La Data Fabric va pouvoir remonter les statuts de tous les traitements et les intégrer au reste du SI.

3. Diagnostiquer les problèmes de production (erreur de traitement, lenteur de traitement). Avec la Data Fabric, l'ensemble des informations liées aux traitement est historisée (sous forme de logs notamment), offrant une traçabilité très avancée pouvant par exemple enrichir un registre de traitement (RGPD).

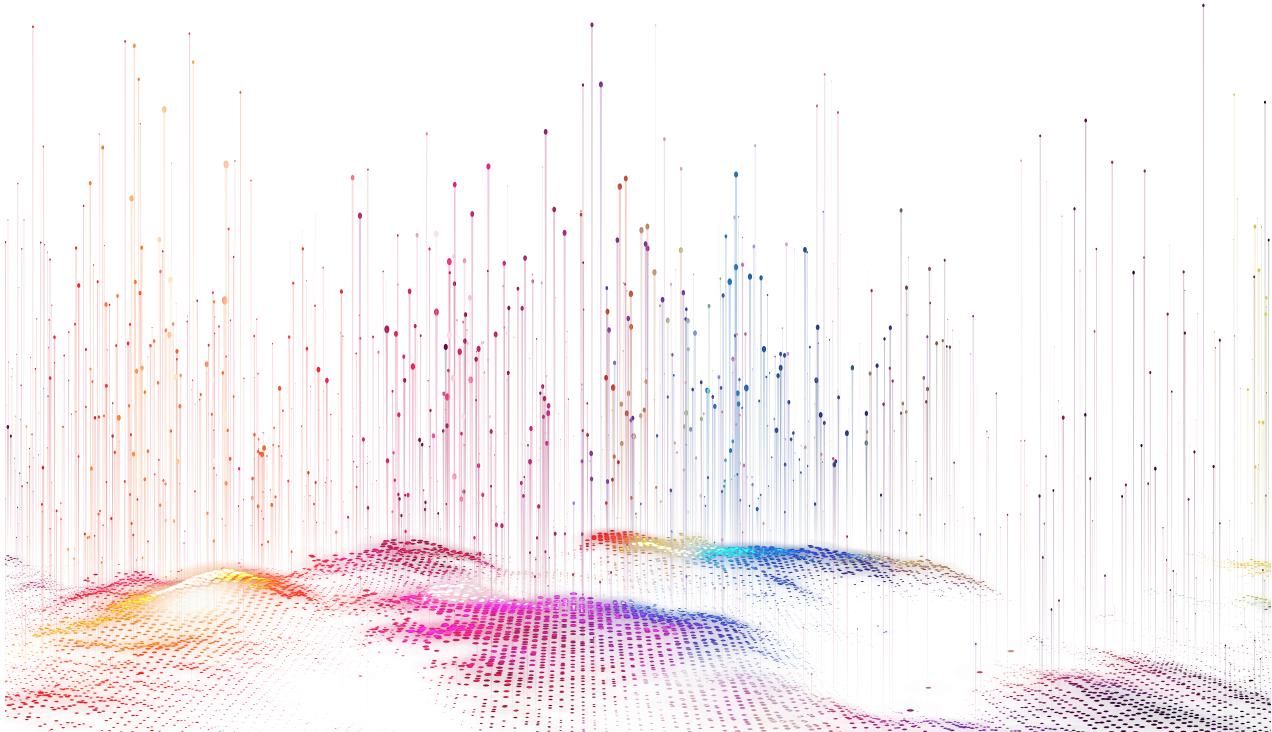
4. Contrôler l'état et les versions des différents frameworks/technologies utilisés et des traitements déployés.

Avec la Data Fabric, il est possible d'exécuter des traitements dans plusieurs versions d'une même technologie. Cela signifie par exemple qu'il est possible de lancer des anciens batchs Spark en version 1.5 pour avoir de la stabilité, et de l'autre côté tester les nouveautés de Spark en version 2.3. Le Data Lab a ainsi la possibilité d'innover de manière transparente et contrôlée.

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Étape Monitor

Il s'agit d'un travail de surveillance et de veille continue. Dans le monde de la donnée, cela consiste à mesurer les effets du traitement et de ses versions ultérieures sur le cas d'usage. Pour résumer, est-ce que les améliorations apportées sur mon cas d'usage / traitement ont eu un impact sur mon business ? (par exemple est-ce que suite à une mise à jour de mon modèle de churn, j'ai eu une augmentation de mes faux positifs ?)



Avec la Data Fabric de Saagie, les responsables de Data Lab (par exemple les CDO) vont avoir une vue consolidée de son utilisation, et de l'activité de l'ensemble des équipes (technologies utilisées, type de traitements implémentés, complexité des workflows, etc). Cela va permettre d'identifier des synergies, promouvoir le partage entre les équipes, réduire les doublons en termes de travaux pour plus d'efficacité.

AMENER LES TRAVAUX DU DATA LAB EN PRODUCTION ?

Le processus d'itération

Les méthodes classiques de développement (type cycle en V) ne sont pas optimisées pour intégrer efficacement une boucle de retour (feedback loop) de la part des métiers. Itérer implique de proposer une nouvelle version de son projet data en minimisant le délai entre 2 versions (passer de mois en semaines, semaines en jours, jours en heures). C'est là qu'interviennent les méthodes agile qui amènent un état d'esprit visant à incrémenter de la valeur au fur et mesure des itérations et tester leur impact.

Au-delà de l'état d'esprit, il n'est pas possible d'y arriver sans un bon outillage permettant de s'affranchir des contraintes liées à l'infrastructure, à la complexité d'orchestrer et d'intégrer des technologies data peu matures, et au respect des critères de production.

La Data Fabric de Saagie a été conçue pour s'affranchir de toutes ces contraintes, et pour que :

- Les Data Labs aient la liberté d'innover
- La DSI puisse avoir un contrôle et une traçabilité fine de ces technologies utilisées
- Ces 2 mondes puissent itérer sur des cas d'usage de la manière la plus fine possible (ex: promouvoir des versions de traitements d'un environnement à un autre de manière facile et automatisée).
- La "Feedback Loop" avec les métiers soit la plus courte possible
- La valeur créé par le Data Lab arrive en production
- Les initiatives Big Data et IA aient un véritable impact sur l'activité de l'entreprise

AGIR AVEC SAAGIE

Faites de vos projets Big Data et IA une réalité



Voir la vidéo

Réconciliez vos équipes Data, IT et Métier pour qu'elles délivrent de la valeur pour votre entreprise.

Affranchissez-vous de la complexité technologique pour adresser plus rapidement vos cas d'usage

Saagie Data Fabric

Une solution prête à l'emploi orchestrant le meilleur des technologies data dans le but d'automatiser les processus de l'entreprise et déployer des applications métiers intelligentes à grande échelle.



Voir la vidéo

Saagie, une plateforme de bout en bout



Saagie

Data Fabric

Ouverture

- ◆ La combinaison de technologies open-source et de modes de déploiement la plus large du marché

Simplicité

- ◆ Solution prête à l'emploi pour démarrer vos projets data dès le premier jour

Contrôle

- ◆ Gouvernance des données, sécurité et gestion des ressources

Ils nous font confiance

Caisse d'Epargne



GROUPE
CAISSE D'EPARGNE

« C'est une somme de plusieurs technologies qui permet à la fois l'extraction de données, leur stockage et leur visualisation. »

Guillaume Cordelier
Directeur de l'innovation, Caisse d'Epargne Normandie
[Lire l'article >](#)

BANQUES ET ASSURANCES



ÉNERGIES ET INDUSTRIES



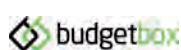
SANTÉ



SECTEUR PUBLIC



SERVICES & E-COMMERCES



Intéressé par nos offres ?

Rencontrons-nous à Paris et Rouen.

+33 (0)2 72 88 31 69
72 Rue de la République,
76140 Le Petit Quevilly, FRANCE

WeWork
92 Av. des Champs-Élysées,
75008 Paris,
FRANCE



Contactez-nous



LA DATA FABRIC POUR INDUSTRIALISER LES DATA LABS

Comment passer les initiatives
Big data / IA de expérimentation
à la production ?

Livre Blanc



Saagie

Data Fabric