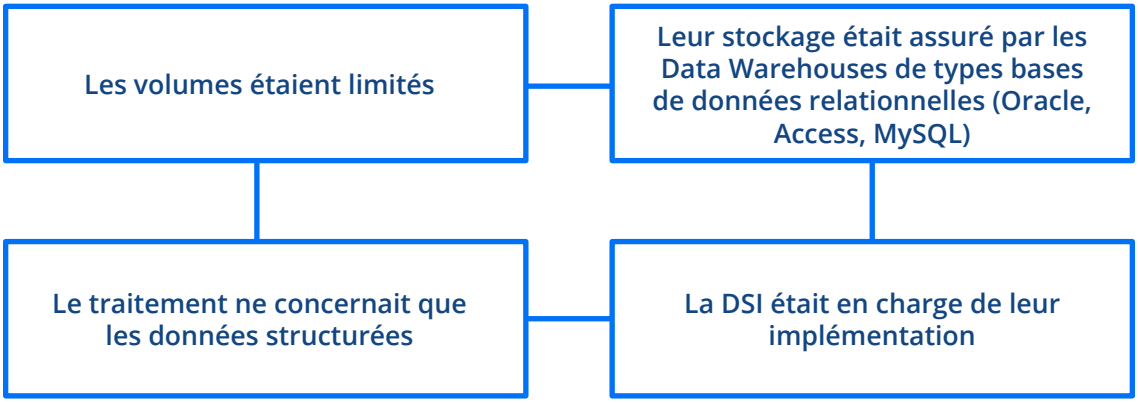
LIVRE BLANC

DataOps :
le DevOps de la data au secours de vos projets

Saagie©

Sommaire

Il n'y a pas si longtemps, traitement de données rimait seulement avec IT. Les projets ressemblaient alors à cela :



En 2010, l'arrivée des Data Lakes (lacs de données) bouleverse le système en place. Le but n'est plus simplement d'alimenter en données les outils, mais d'aller chercher la donnée où elle se trouve. Ils permettent ainsi de centraliser ("dé-siloter") plus efficacement en permettant le stockage des données brutes, les Data Warehouses de l'époque imposant un pré-formatage. L'ajout de serveurs de commodités facilite aussi la mise à l'échelle. Les sources se multiplient et la donnée prend une place centrale. Les outils se perfectionnent et permettent d'en traiter en plus grande quantité et dans des formats beaucoup plus diversifiés, voire déstructurés (images, sons, textes...).

Avec les Data Lakes arrivent de nouveaux profils comme le Data Scientist ou le Data Engineer, et donc la formation des premiers Data Labs. Ces changements apportent plus de liberté, les projets ne dépendant plus uniquement des contraintes parfois strictes de l'IT (processus de sécurité, conformité, validation, etc.). Cependant, les difficultés persistent et en particulier celle de la mise en production, l'écosystème analytique n'apportant pas les mêmes standards de livraison IT que les frameworks hérités du développement logiciel. Transformer une idée en POC devient réalisable, mais du fait que les scripts data sont peu compatibles avec l'IT, la déployer dans un environnement de production est tout de suite plus complexe.

> *Seuls 53% des POCs ont été déployés en production, et cela en 9 mois en moyenne.*
>
> **Gartner, 2020**

Alors comment rapidement déployer ces projets ? Les mêmes analystes parlent de plus en plus d'une approche, celle du DataOps. Comparée au DevOps, son penchant dans le développement informatique, elle se distingue pourtant sur de nombreux aspects intrinsèquement liés aux problématiques uniques des initiatives en matière de donnée.

PARTIE 1

Les challenges des projets data

Aujourd'hui, les entreprises qui souhaitent lancer des initiatives Big Data & Analytics doivent affronter la dure réalité du terrain dès le démarrage du projet :

28%
des responsables de projets ne parviennent pas à sécuriser et gouverner leurs données

31%
des chefs de projets considèrent que le manque de processus organisationnels tel que le DevOps est la première barrière à la mise en production des projets

près de 40%
des responsables d'entreprise considèrent que le ROI d'un projet est difficile à démontrer

moins d'1/3
des décideurs estiment que leur projet rapporte de la valeur

Le principal obstacle à la création de valeur est donc celui de l'industrialisation des projets Big Data & IA. Pour surmonter cet obstacle, les entreprises doivent faire face à trois challenges : le défi humain, méthodologique et technologique.

Le défi humain

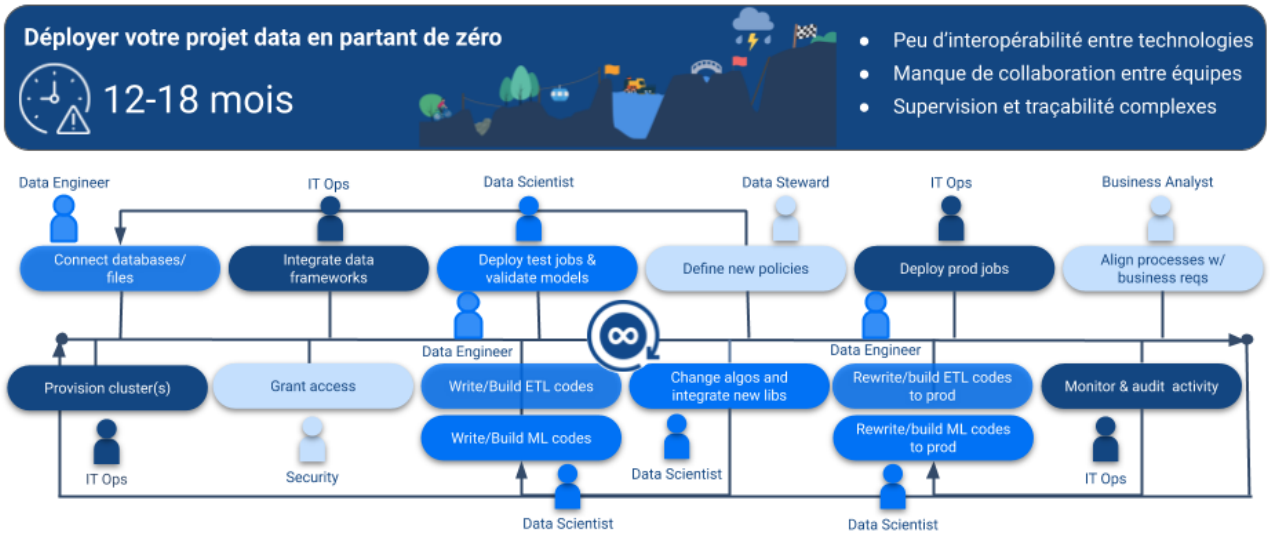
Dans les projets IA et Big Data, trois départements sont directement impliqués et amenés à travailler en étroite collaboration : **l'IT, les profils data et les profils métiers. Souvent, ils ne partagent pas les mêmes objectifs et enjeux**.

Par ailleurs, des divergences de culture et des compétences hétérogènes rendent difficile l'alignement des équipes et impactent la productivité.

Les profils métiers cherchent de la réactivité et des insights facilement actionnables. Les profils Data ont besoin d'un environnement agile et manageable dans lequel ils ont leurs outils préférentiels à jour et disponibles afin de répondre au mieux aux attentes des métiers. Enfin, et cela peut souvent s'avérer être en contradiction avec ces objectifs, **l'IT souhaite maintenir une infrastructure stable et sécurisée**.

Ces besoins et modes de fonctionnement n'étant pas alignés dès le départ du projet, les initiatives ont souvent du mal à être portées au-delà du POC et perdent de l'intérêt à mesure que des tensions se créent entre les équipes.

Le défi méthodologique



Les entreprises ont souvent tendance à minimiser au démarrage du projet les processus au profit du choix des technologies. Mais pour faire coïncider équipes et outils, il faut implémenter des méthodologies ou procédés clairs et connus de tous. Et, outiller les équipes, extraire les données, écrire du code / scripts et les amener d'un environnement de test à celui de production nécessite des rouages solides, mais aussi bien huilés.

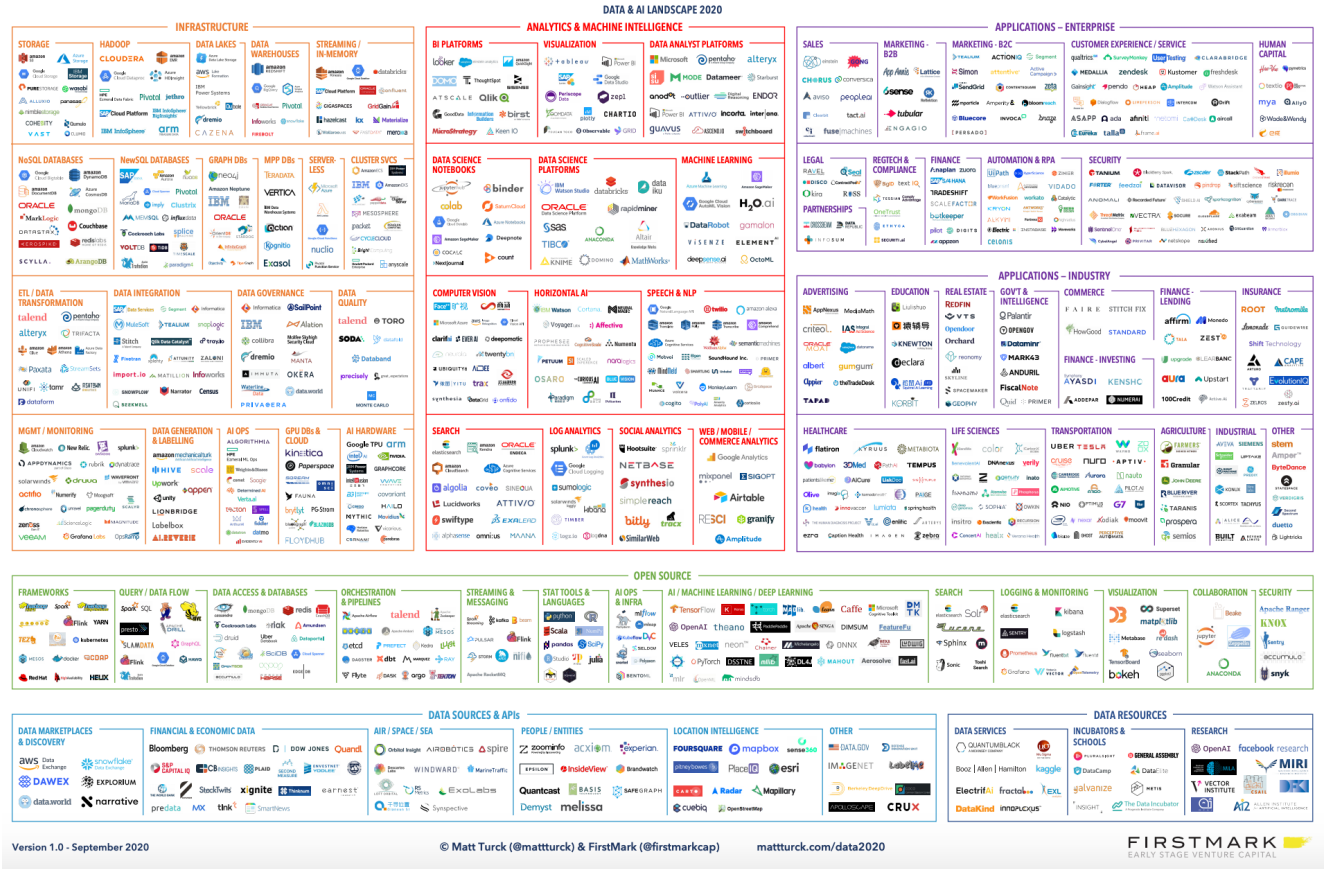
Cela concerne un nombre d'étapes important, du fait de provisionner les serveurs, de permettre l'intégration de technologies et librairies, d'écrire avec des langages différents pour correspondre à des standards de déploiement divers, jusqu'à la prise de feedbacks des utilisateurs métiers.

Ces différentes étapes peuvent parfois représenter plus d'un an de travail et ne sont pas toujours répétables d'un projet à l'autre. Il faut donc trouver une solution qui permette aux équipes de collaborer et d'itérer et de déployer les travaux en continu, et de façon agile.

Le défi technologique

Si vous avez déjà démarré votre projet ou êtes sur le point de le démarrer, vous ne pouvez que constater la diversité de l'écosystème Big Data et IA et la complexité du choix qui en résulte. Les technologies sont presque innombrables, apparaissent aussi vite qu'elles disparaissent et évoluent continuellement afin de coller aux attentes d'un marché en mouvement.

Ci-dessous, la preuve en image avec le panorama de Matt Turck. Traditionnellement représenté annuellement, il en sort désormais deux versions par an, tant l'évolution du marché est rapide.



https://mattturck.com/data2020/

De plus, les frameworks qui se sont imposés ne cessent d'être mis à jour et modifiés par les développeurs pour suivre la cadence du marché. Dans le contexte d'un projet Data, **maintenir une stack technologique d'ampleur devient alors un véritable casse-tête** pour les équipes concernées. Le choix des technologies devient une étape primordiale et parfois critique pour certains projets. Elle représente un risque pour l'entreprise, autant sur son aspect financier que sur ses conséquences sur l'infrastructure en place en interne.

Répondre à ces trois challenges implique de nombreux changements, qui vont s'opérer selon les 3 axes cités au-dessus : l'aspect humain, organisationnel et technologique. Et pour cela le DataOps, identifié comme la nouvelle tendance de la Hype Gartner, semble s'imposer comme un candidat sérieux.

PARTIE 2

Qu'est-ce que le DataOps ?

Gartner définit le DataOps comme “une pratique collaborative de gestion des données axée sur l’amélioration de la communication, de l’intégration et de l’automatisation des flux de données entre les métiers de la data et leurs consommateurs.”

Plus concrètement, nous définissons cette approche comme un modèle technologique et organisationnel dérivé du DevOps. **La méthodologie utilise la technologie pour automatiser la conception, le déploiement et la gestion de la livraison des données**. Le DataOps vise à amener l’agilité, l’automatisation et le contrôle entre les différents porteurs de projet, dont l’IT (exploitants des SI, développeurs d’applications, architectes), les équipes analytiques (Responsable Produit, les Data Scientists et Data Engineers, Data Stewards) ainsi que les métiers (Marketing, Finance, Production, Logistique, RH, etc.). Le but est d’industrialiser les processus analytiques en tirant parti d’un large et divers écosystème Big Data et de la complémentarité des compétences des profils concernés.

Les principaux piliers du DataOps

La promesse du DataOps

Améliorer et optimiser en continu le cycle de vie des projets Data & Analytics via plus de rapidité et de qualité

Les grands principes du DataOps

Le DataOps a beau porter un nom qui résonne encore peu, les grands principes sur lequel il se base sont pourtant bien connus. L'approche s'appuie sur des méthodes de travail éprouvées dans le monde DevOps, les rassemble et les applique au monde de la gestion de données et à ses contraintes. Parmi elles, on trouve :

L'agilité

Il s'agit du fait de mettre en place des cas d'usage rapides à mettre à l'échelle. Le but est d'apporter de la valeur au fil de l'eau afin, notamment, de renforcer la confiance au sein des équipes.

Ces pratiques ont pour avantage de favoriser la communication et la collaboration de différentes équipes, ce qui permet un déploiement accéléré des projets et donc des coûts réduits. On pourra citer pour ce qui est de la gestion d'environnements :

- la gestion de versions
- les pratiques de métrologie et de supervision

L'idée est de contrôler la disponibilité et la performance de l'infrastructure et des traitements qui sont exécutés quotidiennement. Cela permet également de mesurer la performance des algorithmes de machine learning qui sont élaborés dans l'environnement.

Le DevOps

C'est l'approche qui a permis de faire le pont entre les équipes de développement qui cherchaient à mettre en place un projet innovant en toute liberté, et les équipes d'exploitation qui pensaient davantage à un déploiement structuré et stable.

Comment ?

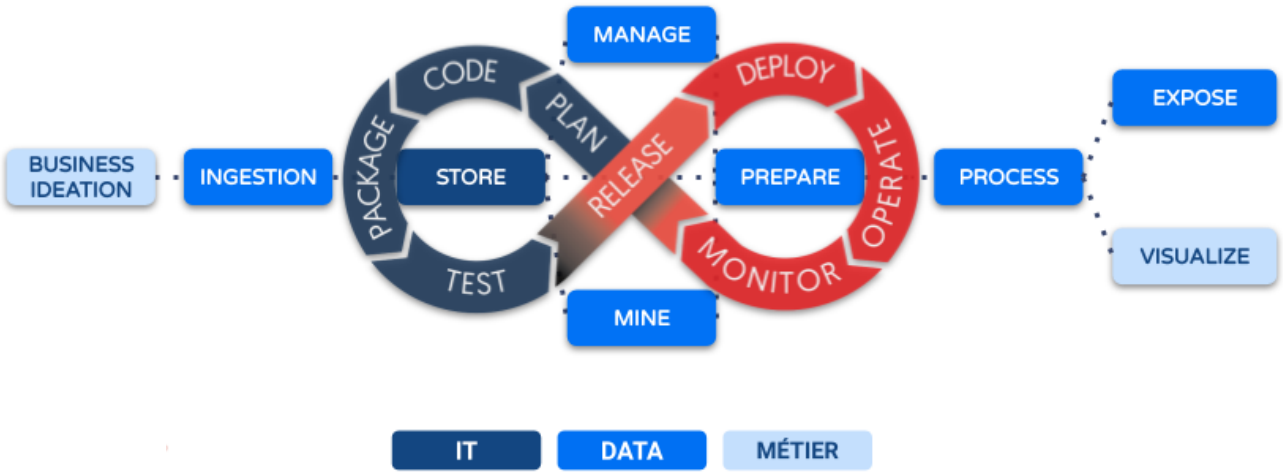
Le DevOps repose sur deux concepts fondamentaux : l'intégration continue (CI) et le déploiement continu (CD).

- **L'intégration continue** consiste à construire, intégrer et tester de nouveaux codes de façon répétée et automatisée. Cette méthode permet d'identifier et de résoudre rapidement les potentiels problèmes, mais aussi de mettre à jour («commiter» ou soumettre) rapidement des codes complexes impliquant de nombreux développeurs.

- **Le déploiement continu** automatise le déploiement ou la livraison de logiciels. Une fois qu'une application a passé l'ensemble des tests de qualification, le DevOps permet son passage en production.

En résumé, l'approche DevOps permet l'alignement entre les équipes de développement / exploitation et l'automatisation de chacune des étapes de la du cycle de vie applicatif, de sa conception à son déploiement, jusqu'à son administration.

Appliquée au projet Data, l'approche concerne donc l'ensemble du cycle de vie de la donnée, de son extraction jusqu'à sa visualisation par l'utilisateur final. Chacune des pratiques la composant doivent donc être répétées à chaque étape de ce cycle.



Quel lien avec le DataOps ?

Cela fait donc du DataOps l'héritier du DevOps, puisqu'il reprend ces deux principes, mais est encore plus difficile à appliquer puisqu'il implique des profils qui vont au-delà de la DSI. Adapté aux projets Data, son objectif consiste donc à faciliter et à accélérer la livraison tout en intégrant de nouveaux profils, en gérant différents environnements et en orchestrant un nombre important de technologies.

Quels sont les profils concernés ?

IT Team

IT Ops Manager

Maintenir et monitorer un ensemble de technologies, tout en assurant un certain niveau de gouvernance et de sécurité. Tout cela dans divers environnements.

Architecte IT

Concevoir et mettre en place des architectures IT. Installer des bases de données et assurer les flux de données.

Développeur

Concevoir et maintenir les codes informatiques nécessaires au bon fonctionnement des applications métiers.

Data Team

Data Product Owner

Concevoir les expressions de besoins, et aider à la collaboration entre l'IT et l'équipe Data afin d'atteindre les objectifs métiers.

Data Engineer

Mettre en place les traitements d'ETL, en particulier les flux de données multi-sources, préparer les échantillons de données pour les data scientists et aider à l'optimisation de déploiement de leurs algorithmes.

Data Scientist

Apporter de la valeur aux données par la création d'algorithmes de préparation (nettoyage, enrichissement, feature engineering...) et de modélisation (statistiques, machine learning, NLP). Aider à la visualisation et à l'interprétation des résultats.

Business Team

Data Steward

S'assurer de la qualité des données au sein de l'ensemble de l'organisation en identifiant les responsables de ces données et en les documentant.

Business Analyst

Trier et traiter les données afin de les rendre facilement visualisables et compréhensibles sous forme de tableaux de bords. Suggérer des recommandations métiers sur la base de données raffinées.

Business Experts

Contribuer à la qualification des données. Tirer parti des données pour apporter de la valeur business aux différents départements concernés : Marketing, Finance, Ressources Humaines, Logistique.

PARTIE 3

Comment le mettre en place ?

Avant son implémentation concrète, il faut d'abord se pencher sur les aspects spécifiques et challenges auxquels vous serez confronté en prenant part à un projet Data & Analytics, ceux que nous avons évoqué plus haut. La réponse à l'un de ces défis consiste à construire et maintenir un data pipeline (ou flux de données).

On entend par data pipeline une séquence de traitements de données, de la phase d'extraction à celle de la visualisation.

Le pipeline de données est l'aspect « Ops » de l'analyse des données. Les données entrent continuellement d'un côté du pipeline, progressent à travers une série d'étapes et sortent sous forme de rapports, de modèles et de tableaux de bord.

Un autre aspect différenciant du DataOps est lié à des spécificités de projets de Data Science :

- **la reproductibilité des résultats**, quel que soit l'environnement d'exécution ;
- **le monitoring des performances du modèle** : notez qu'un modèle de prédiction bon aujourd'hui, ne le sera pas forcément demain. Et si le modèle de données change…
- **la mise à disposition et exposition de modèles** dans une application (ou par API) pour utilisateurs finaux.

Ces contraintes, auxquelles, le DataOps peut répondre, vont donc demander des changements, et ce à tous les niveaux de l'entreprise.

Un chamboulement culturel

Il n'est pas rare de penser qu'un projet Data & Analytics repose essentiellement sur les choix technologiques et que la bonne combinaison d'outils pourrait à elle seule résoudre les problèmes associés à ce type de projet.

Même s'il est vrai que trouver les outils adaptés est d'une grande aide, il faut pouvoir les assembler et les associer à un changement culturel d'ampleur qui bouleversera les processus en place. Un terme revient très souvent pour décrire ce changement : le fait de devenir une entreprise data-driven (centrée sur la donnée).

Selon le rapport "Big Data and AI Executive Survey 2019" de NewVantage, au-delà du terme tendance se cache réellement une difficulté à considérer la donnée et les bénéfices qu'elle peut apporter :

72%
des responsables interrogés ne considèrent pas avoir adopté une culture data

69%
considèrent que la donnée n'est pas un facteur central dans leur stratégie

+ de 50%
ne considèrent pas la donnée comme un avantage compétitif

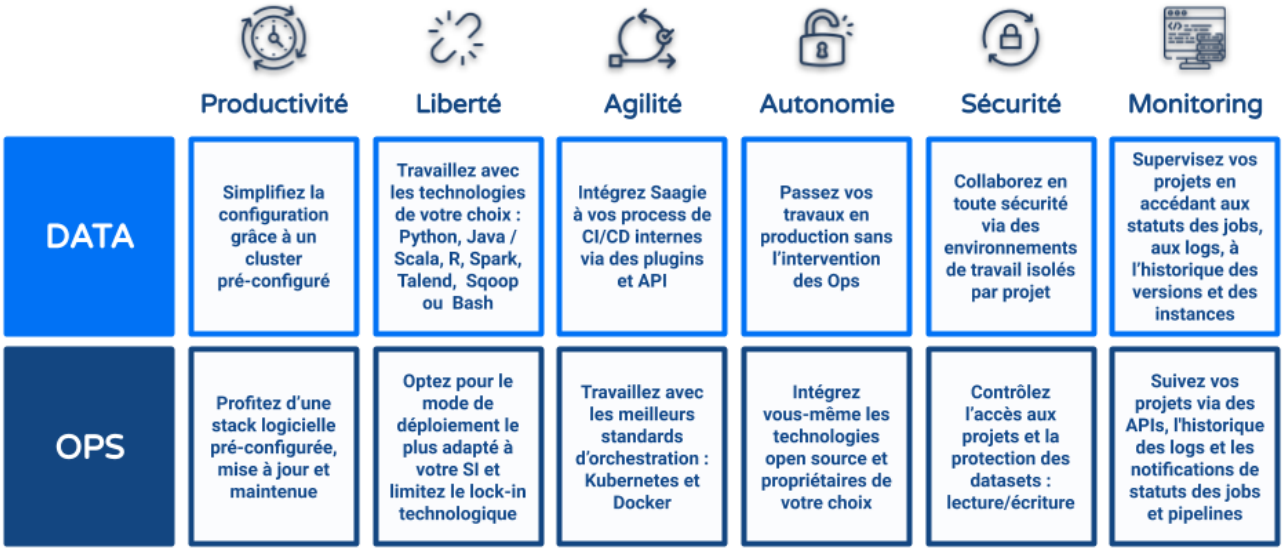
Pour en revenir à la culture, **le DataOps implique un état d'esprit de collaboration** entre les équipes qui doit se faire naturellement. Seul un alignement des équipes concernées par le projet pourra garantir son succès et maintenir des résultats sur le long terme. Dans la pratique, l'organisation est impactée. Un niveau de gouvernance supplémentaire pourra aussi être mis en place afin de s'assurer de la qualité des données et du niveau de sécurité requis.

Le développement d'applications de Machine Learning ou Data Science à usage intensif de données demande généralement beaucoup de données de qualité qui peuvent parfois être sensibles. Leur déploiement soulève de nouvelles questions d'accès et de gouvernance que l'IT doit prendre en considération. Tout cela sans même parler d'infrastructure : cloud public, on-premise, multi-clouds, cross-clouds... Les choix sont multiples et chacun d'entre eux apporte ses avantages et contraintes. De plus, il reste encore à trouver les profils cités ci-dessus et parvenir à les faire travailler ensemble. En plus d'induire un changement de culture important, ces choix d'urbanisation requièrent un outil unique, transversal et partagé si vous souhaitez implémenter le DataOps dans vos projets. Il devra donc permettre l'accès à diverses sources de données par l'ensemble des profils concernés, leur stockage, leur préparation, leur traitement et leur visualisation.

La plateforme DataOps

> *On entend souvent "Data is the new oil" : la donnée est le nouvel or noir ; mais l'enjeu n'est pas tellement la donnée elle-même, mais ce qui est autour. Ce qu'il faut c'est une raffinerie, car ce que les gens attendent, ce n'est pas du pétrole, mais bien de l'essence.*
>
> **Adrien Blind, VP Product & Technology chez Saagie**

Et si la raffinerie, le relai technologique nécessaire, était donc un outil que l'on associe au DataOps ? Ce que nous appelons : **la plateforme DataOps**. Un outil qui aidera vos équipes Data et Ops :



Comment choisir sa plateforme DataOps ?

Afin d'amener agilité et collaboration, l'orchestrateur doit être adaptable et modulable, aussi bien dans ses modes de déploiement que dans son utilisation. Cinq piliers sont majeurs pour y parvenir :

1. Il doit pouvoir **s'adapter au mode de déploiement** choisi par l'entreprise, être prêt-à-l'emploi pour ne pas perdre de temps et être facile à maintenir.

2. Il doit pouvoir intégrer harmonieusement des technologies qui s'adressent à **l'ensemble du cycle de la donnée** (préparation, traitement, valorisation et visualisation) tout en étant immédiatement opérationnel et robuste (infrastructure stable).

3. Le fait de rassembler un grand nombre de technologies ne suffit pas, **il faut qu'elles soient orchestrées** par tâches qui pourront être ordonnancées (extraction, préparation, traitement...)

4. Des fonctionnalités de **monitoring et de traçabilité** doivent aussi être présentes afin de superviser l'ensemble de la chaîne, de pouvoir intervenir rapidement et donc d'améliorer les aspects de performance et de sécurité.

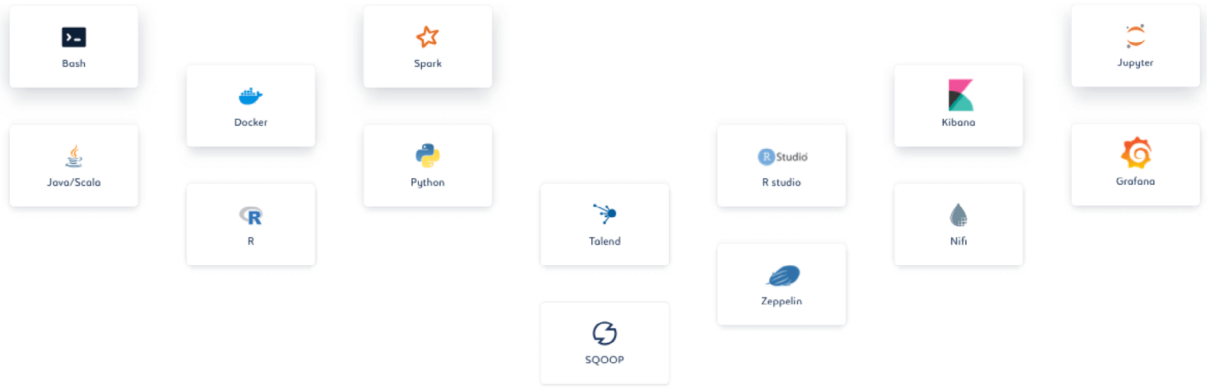
5. Enfin, **l'ergonomie et l'accessibilité** ont leur importance puisque, comme nous en avons parlé un peu plus haut, des profils métiers sont impliqués dans le projet. Il faut donc une interface compréhensible et intuitive, qui puisse parler à tous, peu importe leur niveau technique.

L'intérêt d'un tel outil est d'avoir des jobs reliés par des pipelines de traitements qui permettent de **déployer de façon ordonnancée, programmatique et automatisée**, et tout cela en pouvant facilement passer d'un environnement à un autre.

La plateforme de Saagie rassemble le meilleur des technologies du marché data pour permettre aux data engineers de faciliter, d'accélérer et de fiabiliser la mise en production de leurs projets.

Simplifier l'accès aux technologies

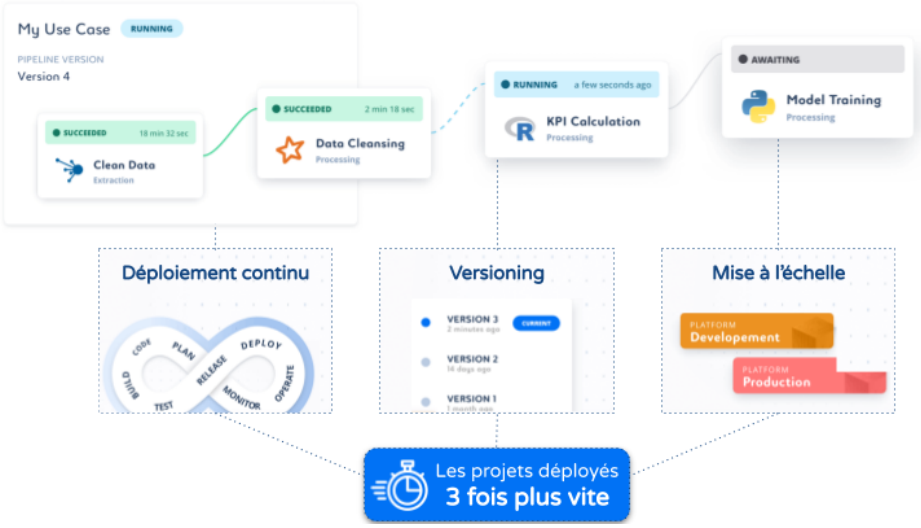
Construire soi-même sa propre plateforme peut sembler attractif : les Ops peuvent garder la main dessus, les équipes data peuvent la faire customiser pour répondre à leurs besoins. Mais en réalité, Gartner estime que cela prend entre 12 et 18 mois.



Nous mettons à la disposition de nos clients une plateforme prête à l'emploi qui rassemble et orchestre les standards du marché afin d'offrir un point d'entrée unique au meilleur de l'écosystème data. Le tout est infogéré afin de leur permettre de se concentrer sur ce qui leur apporte de la valeur et non plus sur la configuration ou la maintenance de l'outil qui le permet.

Accélérer la mise en production

Saagie est une plateforme conçue pour la production afin que vos projets tournent aussi bien une fois déployés en production qu'ils le faisaient en phase exploratoire. Il vous suffit de créer un projet, de choisir les technologies adaptées, de lancer vos traitements et de les orchestrer en pipelines automatisés. Vous pouvez ensuite reproduire ces jobs et pipelines et les promouvoir d'un environnement à l'autre.



Plus simplement, nous vous donnons les moyens de mettre en place une chaîne de production entièrement automatisée qu'il vous sera possible de répliquer pour vos futurs projets.

Fiabiliser la gestion des projets

Chez Saagie, nous tâchons de faire rimer performance et sécurité. Nous orchestrons les technologies qui vous permettront de gérer l'ensemble du cycle de vie de la donnée et nous ajoutons à cela les fonctionnalités afin de tracer et de superviser chaque étape de ce cycle. Vous pouvez ainsi isoler vos projets, en gérer les accès, superviser les statuts de vos jobs mais aussi centraliser et stocker les logs pour suivre votre activité.

Notre plateforme est robuste et simplifie la configuration et la maintenance gérées par les Ops. Et une fois vos projets en production, ils peuvent ainsi se concentrer sur l'amélioration continue de la sécurité et de la performance de vos flux de données.

Références

Gartner, CIO Survey, 2018
Capgemini & Informatica, The Big Data Payoff: Turning Big Data into Business Value, 2016
BCG, Putting Artificial Intelligence to Work, September 2017
Gartner, Innovation Insight for DataOps, December 2018
Gartner, Market Share: all software markets, worldwide, 2018
Gartner, Your Data Culture Is Changing — Do You Need DataOps?, 2019
Gartner, 3 Ways to Deliver Customer Value Faster Using DataOps, 2020
Gartner, Introducing DataOps Into Your Data Management Discipline, 2020
Gartner, Innovation Insight for DataOps, 2020

Saagie©

Vous souhaitez en savoir plus sur notre plateforme DataOps ? Contactez-nous !

+33 (0)2 72 88 31 69

Demandez une démo

www.saagie.com/fr

PARIS

Saagie
10 rue Lincoln
75008 Paris
FRANCE

ROUEN

Seine Innopolis
72 rue de la République
76140 Le Petit Quevilly
FRANCE

NEW YORK

WeWork
315W 36th St, 2nd Floor
New York, NY 10018
USA