TECH TRENDS

#6
MARS
2015

DATA LAB

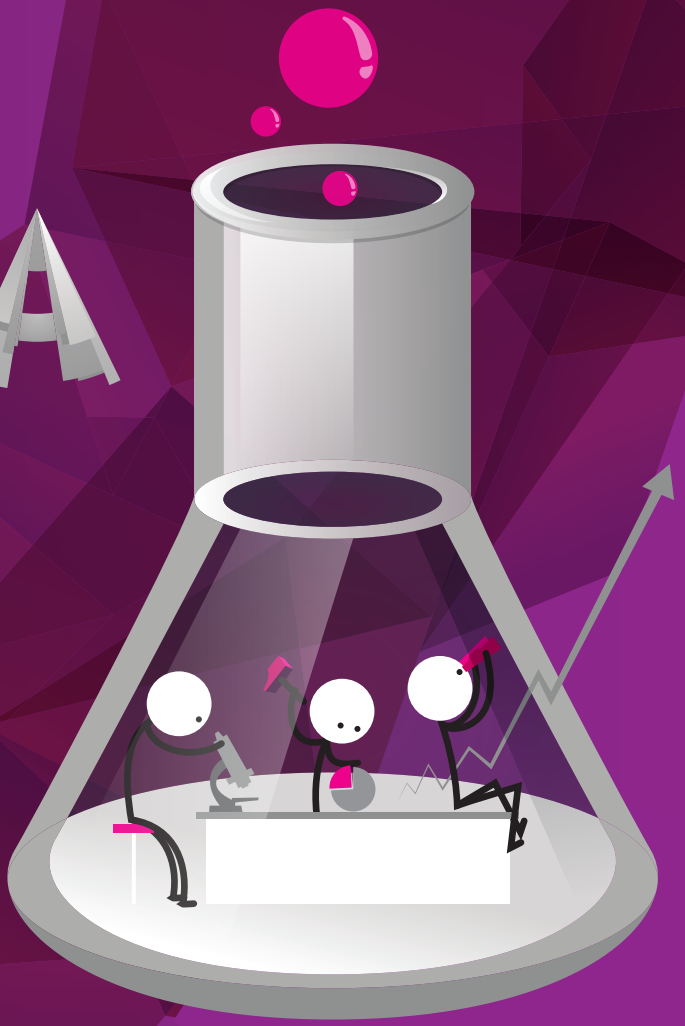
TechTrends - Publication de Xebia IT Architects

DATA LAB

Imaginer

Matérialiser

Exploiter



Xebia
SOFTWARE DEVELOPMENT DONE RIGHT
THIGA
Hungry & Foolish
UX
REPUBLIC

Introduction

Depuis quelques années maintenant, les « Big Data » font le buzz : impossible de passer à côté de ce phénomène né des nouveaux usages du Web, de l'explosion des données personnelles et des avancées technologiques dans le calcul distribué. Au-delà du mot et de la tendance, comment en tirer profit au sein de votre entreprise ? En construisant un Data Lab.

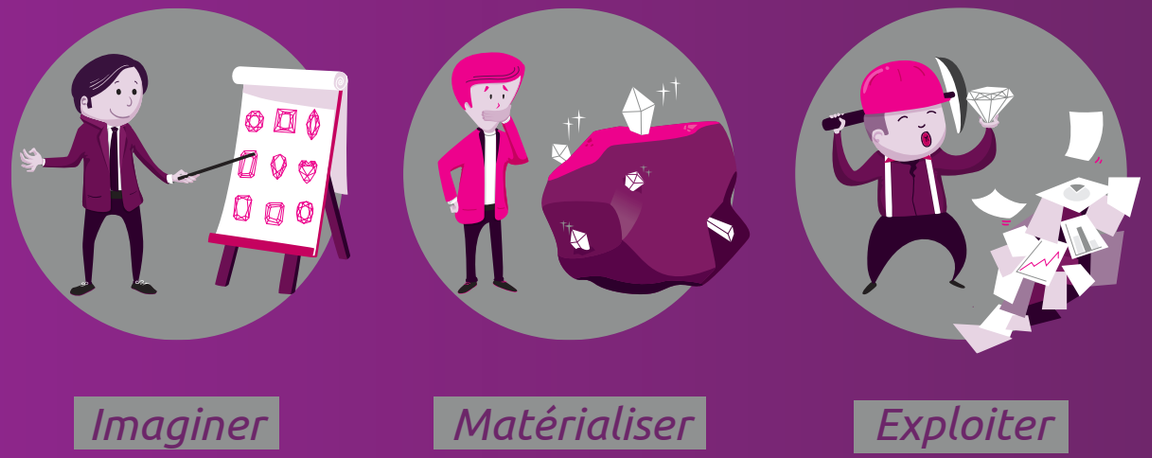
Dassault Systèmes, Ford, AXA, BNP, Google, etc. Ces grandes entreprises sont nombreuses à avoir mis en place des « Labs ». Ces initiatives, bien que labellisées sous un même titre, couvrent des réalités diverses et variées : cellule de veille dédiée à la détection de signaux faibles, « kick starter » d'idées en rupture, vitrine de communication à destination de l'externe, Fabrication Lab, etc.

Cette course à l'innovation se fait le plus souvent sur trois axes :

- La compétence interne : comment tirer le meilleur de votre savoir-faire actuel pour conquérir de nouveaux marchés ?
- La connaissance client : comment mieux comprendre les besoins de vos clients pour leur proposer de nouvelles offres et les fidéliser ?
- Les tendances : comment exploiter les nouvelles technologies pour faire émerger des opportunités business ?

Parmi les tendances technologiques à exploiter, on retrouve les importants progrès techniques de ces dernières années en matière de sauvegarde et de traitement des données. Collecter, stocker et traiter la donnée n'a jamais été aussi peu coûteux. Un nouvel axe d'innovation est ainsi en train d'émerger : l'exploitation des données.

Un Data Lab repose sur trois piliers : une équipe, un lieu et une approche. Monter un Lab, c'est faire travailler des personnes triées sur le volet, dans un lieu dédié en les dotant de l'outillage et des méthodologies les mieux adaptées afin de tirer avantage des données qui les entourent.



> *Un Data Lab repose sur trois piliers : une équipe, un lieu et une approche.*

Intégrer une telle structure dans une entreprise représente cependant un changement de philosophie et nécessite d'adopter une nouvelle approche. En effet, le caractère expérimental et l'aspect très mathématique du travail sur la donnée peuvent parfois faire penser à la « découverte accidentelle ». Dans ce cas, l'approche la plus souvent rencontrée est :

- Mettre en place une infrastructure Big Data « couteau-suisse », capable d'adresser une vaste palette de problèmes fonctionnels et techniques.
- Laisser les Data Scientists expérimenter sur les données.
- Attendre qu'ils « trouvent » quelque chose.

Guidée par la technologie (techno-push), cette approche ne donne que très rarement des résultats satisfaisants.

A contrario, adopter une approche dite « Business Driven » s'avère souvent fructueux. Cette dernière s'inspire des frameworks de Product Management des grands acteurs du Web et emprunte des outils et principes à des courants de pensée comme le Design Thinking, le Lean Startup et, bien sûr, des méthodologies agiles telles que Scrum et Kanban. Elle prône également la recherche de use cases sous un angle business. En effet, le faible niveau de maturité des approches « Data Lab / Big Data » nécessite de rassurer en permanence les directions générales sur le bien-fondé de leur investissement.

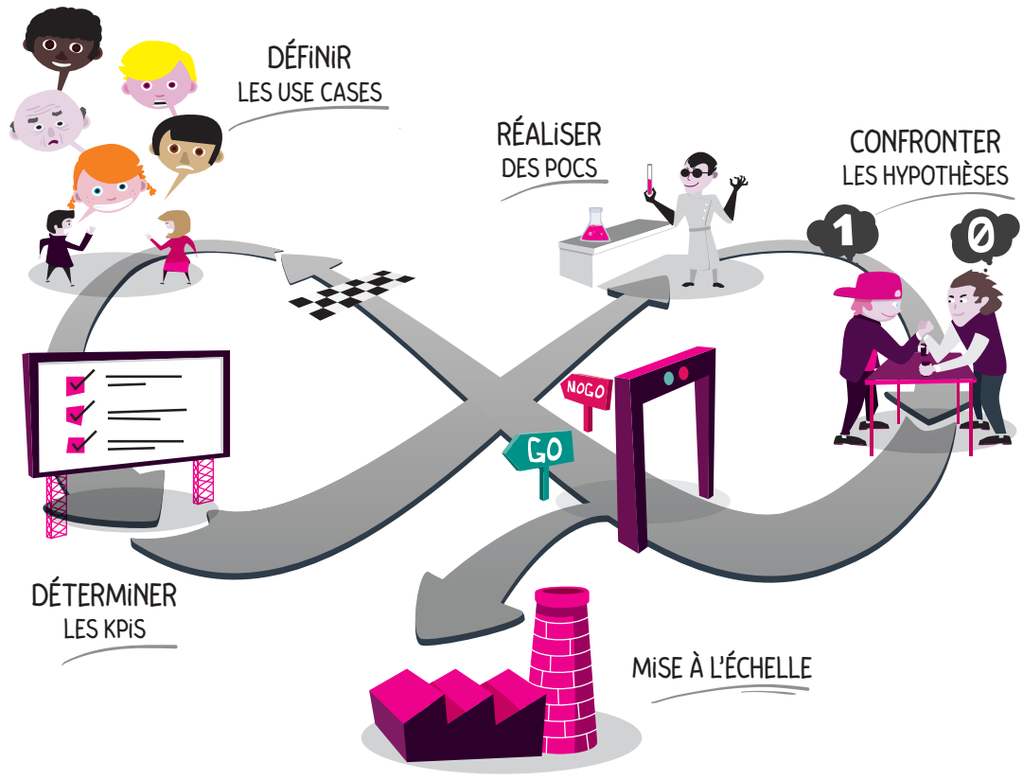
Afin de donner de la valeur à vos données, nous préconisons une démarche en trois étapes :

Imaginer : faire émerger des use cases et définir des KPIs.

Matérialiser : monter une équipe, une architecture et collecter des données.

Exploiter : réaliser une courte démonstration de faisabilité via un Proof of Concept (PoC), valoriser les données puis communiquer à travers la data visualisation.

Le Data Lab vous donnera ainsi les clés pour exploiter différemment vos données sans perdre vos impératifs business de vue.



Les étapes clés du processus d'un Data Lab

> *Adopter une approche dite « Business Driven » s'avère souvent fructueux. Elle prône la recherche de use cases sous un angle business.*

Imaginer

Se jeter à corps perdu dans les données

dès les prémisses d'un Data Lab est une erreur courante. Un projet Data nécessite un important travail de réflexion préalable et d'analyse des besoins. Il se doit d'avoir des objectifs clairs et les moyens de mesurer l'atteinte de ces derniers : la définition de use cases et de leurs bénéfices attendus est donc un passage obligé.

CONSTRUIRE VOTRE ÉQUIPE

Afin de mener cette phase d'imagination, une équipe avec des compétences et des rôles bien définis sera mobilisée.

La pluridisciplinarité est un critère très important dans le succès d'une démarche d'innovation (qu'elle soit liée ou non à la donnée). Constituer des équipes avec plusieurs compétences, c'est garantir des angles de vue multiples et donc, une plus grande créativité. Ainsi, nous suggérons de constituer une équipe composée :

• d'experts du métier de l'entreprise,

• de responsables IT,

• de responsables de la BI,

• de commerciaux,

• de membres du marketing,

• et d'experts techniques.

L'équipe mise en place pour la formalisation des use cases n'est pas la même que celle composée pour leur réalisation. Même si certaines personnes doivent être présentes dans les deux étapes, les profils de l'équipe de réalisation seront beaucoup plus techniques.

Afin de donner plus de poids à la démarche, il est également important de faire participer des sponsors de haut niveau et de leur présenter régulièrement les avancées dans la construction des use cases.

Une fois cette équipe constituée, il est temps d'entrer dans le vif du sujet !

> *La pluridisciplinarité est un critère très important dans le succès d'une démarche d'innovation (qu'elle soit liée ou non à la donnée).*

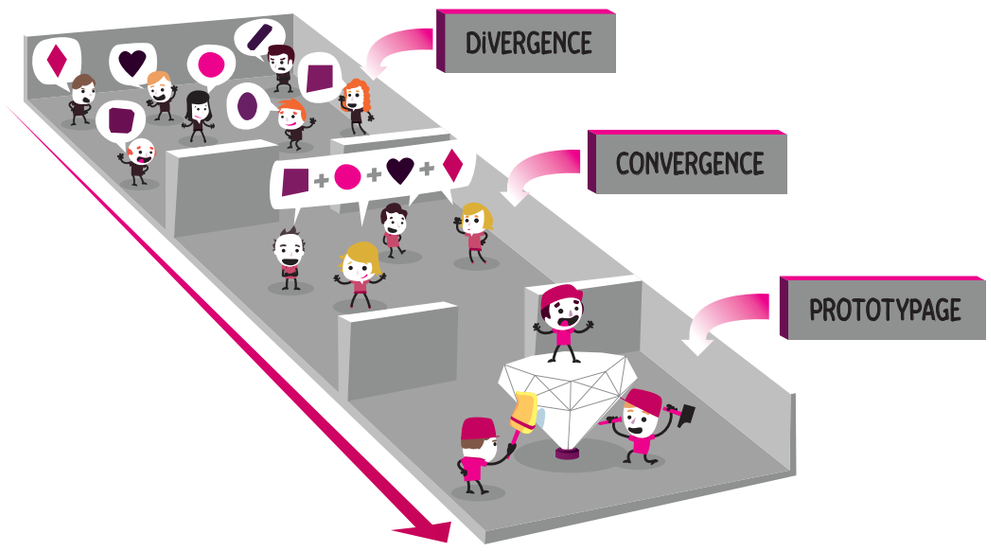
FAIRE ÉMERGER VOS USE CASES BUSINESS

Les approches inspirées du Design Thinking sont une manière idéale de faire émerger les besoins métier de vos projets Data. Détailler toutes les étapes d'un processus d'innovation reposant sur cette méthode n'est pas l'objet de ce TechTrends. Il faut cependant retenir que le Design Thinking est une alternance de phases de divergence et de convergence intellectuelles s'appuyant sur des principes tels que l'empathie et la pluridisciplinarité. En effet, avant même de proposer des solutions, il est indispensable de comprendre les problèmatiques de vos utilisateurs (internes ou externes). Einstein illustre d'ailleurs cette nécessité dans une célèbre citation : *« Si on me donne une heure pour résoudre un problème, je passerais 55 minutes à comprendre la problématique et 5 minutes à lui chercher des solutions. »*

Mener un projet Data, c'est répondre par la technologie à une problématique clairement identifiée. Nous insistons beaucoup sur ce point car il est fréquent de voir des équipes « Data Lab » travailler sur des use cases qui ne répondent à aucun objectif et qui seront abandonnés par la suite, faute d'une utilité réelle. L'expérience nous a montré que cette approche donne des résultats plus pertinents qu'une approche traditionnelle où l'on se contenterait de simples ateliers de brainstorming.

Elle comporte deux grandes phases :

- Une phase de divergence permettant de faire émerger en grand nombre des use cases potentiels.
- Une phase de convergence visant à sélectionner les use cases les plus pertinents au regard d'un ensemble de critères (faisabilité, viabilité, cohérence avec la stratégie de l'entreprise, etc.).



Initier la phase de divergence

L'objectif d'une phase de divergence est de générer un maximum d'idées en stimulant le potentiel créatif des participants. Un atelier de divergence dure typiquement de 1h30 à 2h et respecte les préceptes suivants :

- Ne pas juger : aucune idée ne doit être écartée.
- Construire sur les idées des autres : il n'y a pas de « mais », seulement des « et ».
- Encourager les idées excentriques : explorer les pistes les plus saugrenues.
- Privilégier la quantité : chercher à générer le plus d'idées possibles (jusqu'à 100 en 90 minutes).
- Rendre les idées visuelles : matérialiser les idées autant que possible (dessiner à main levée, utiliser des couleurs, etc.).

La clé du succès consiste à ne pas introduire un raisonnement analytique et rationnel dans la phase de divergence et ce, par opposition à une séance de brainstorm traditionnel (pendant laquelle on va chercher dans le même laps de temps à diverger et converger). L'analyse et la rationalisation dès la phase de convergence conduisent à tuer l'innovation dans l'oeuf. Ce mode de pensée est à réserver pour la phase de convergence avant laquelle il est préférable de laisser s'écouler quelques jours.

Poursuivre avec la phase de convergence

Selon le nombre d'idées qui ont été générées en atelier de divergence, il peut être intéressant de faire un point intermédiaire entre les deux phases afin de réduire le backlog. Selon votre contexte, une partie plus restreinte de l'équipe ou d'autres parties prenantes pourront effectuer ce filtre.

Il est ensuite temps d'entamer la phase de convergence visant à sélectionner les meilleures idées du backlog intérmediaire. Pour cela, on fait appel à la logique, au pragmatisme et à la critique (constructive). L'équipe impliquée dans cette phase est la même que celle de la phase de divergence.

Pendant l'atelier de convergence, il faut balayer l'intégralité des use cases en les soumettant à une grille d'évaluation qui rassemble des critères à la fois business, techniques et organisationnels :

<u>Business :</u>

- Le use case produit-il une valeur ajoutée métier ? Des KPIs peuvent-ils être définis pour mesurer la valeur métier créée ?
- Le PoC que je vais réaliser pourra-t-il intégrer des processus métiers opérationnels ?

Techniques :

- Les sources de données cibles sont-elles facilement récupérables ?
- Quelles sont la véracité et la qualité des sources de données idoines ?
- Quelle est la complexité d'analyse de la donnée (matching, algorithmes, etc.) ?
- Sera-t-il facile de proposer une visualisation efficiente de la donnée et des résultats produits ?

Organisationnels / méthodologiques :

- Le PoC peut-il être mis en œuvre rapidement (i.e. en moins de 10 semaines) ?
- Le PoC peut-il être produit en mode itératif avec une démonstration des résultats intermédiaires à la fin de chaque itération ?
- Le PoC pourra-t-il être industrialisé à terme ? Quels efforts faudra-t-il fournir ?
- Les compétences nécessaires sont-elles disponibles (en interne ou en externe) et mobilisables dans un délai court ?

L'importance de chaque critère devra être pondérée en fonction de votre contexte, de vos moyens et de vos priorités. L'évaluation systématique, sur ces critères, de l'ensemble des use cases issus de la phase de divergence permettra d'en extraire un ou plusieurs à prototyper.

Ainsi, le Data Lab dispose d'un backlog d'idées priorisées. Les développements d'un premier PoC peuvent débuter. La phase d'idéation pourra pour sa part recommencer afin de maintenir un backlog à jour (l'échelle de temps est à définir selon votre contexte).

Étude de cas

Premiers pas d'un transporteur français, accompagné de Thiga et Xebia, dans l'univers du Big Data.

Un premier atelier, celui de divergence, est organisé dans le but de faire émerger des use cases « Big Data » à forte valeur ajoutée business. Le grand challenge est d'amorcer la créativité. En effet, dans le cadre traditionnel d'une salle de réunion, il est difficile pour les participants de « se lâcher ». Il est donc conseillé de sortir du cadre de travail habituel et de se retrouver dans un lieu inconnu et convivial. D'une durée de 3h, l'atelier réunit une dizaine des personnes de l'entreprise. Plusieurs métiers sont représentés. Ainsi, ce sont des commerciaux grands comptes, des ingénieurs BI et des experts du transport qui alimentent le backlog d'idées.

Afin d'ouvrir le champ des possibles, les réflexions du groupe sont, en préambule, alimentées par des retours d'expérience de Thiga et Xebia.

Bien qu'une des règles d'un atelier de divergence soit de ne pas se fixer de limites, il est décidé de cadrer la séance en imposant trois thèmes : l'efficacité opérationnelle, la relation client et les nouveaux business. Cette restriction a été faite pour deux raisons :

- la population présente à ces ateliers n'étant pas familière à ce type d'exercice, une liberté totale peut être contre productive,
- la seconde raison est tout simplement liée à une contrainte de temps. En effet, nous ne disposons que d'une seule séance dédiée à la divergence.

À l'issue de cet atelier, une trentaine d'idées constituent notre backlog.

> *L'objectif d'une phase de divergence est de générer un maximum d'idées en stimulant le potentiel créatif des participants.*

Durant la période de transition entre les deux ateliers, toutes les idées sont passées en revue par une équipe réduite composée d'experts du métier et d'experts techniques. Pour

chacune d'entre elles, son niveau de faisabilité (à la fois technique et organisationnelle) et son intérêt business sont mesurés. Cette étape intermédiaire répond à un des enjeux majeurs de notre client : la compréhension par une population non technique des possibilités et des contraintes que représente le Big Data. Il est donc important de vulgariser au maximum les concepts qui y sont associés. Cette revue intermédiaire a permis de disqualifier les use cases non réalisables et d'avoir une première idée de hiérarchie entre ceux restant.

Quelques semaines plus tard, lors de l'atelier de convergence, tous les use cases sélectionnés, ainsi que leurs critères de faisabilité estimés, sont présentés aux experts métier. Ils doivent alors choisir les use cases prioritaires. Cet atelier est réalisé via l'exercice « Buy a feature » : de faux billets sont distribués aux intervenants qui peuvent alors « investir » sur les use cases qui leur semblent les plus pertinents. Trois idées ressortent comme étant prioritaires en vue d'un prototypage via la réalisation de PoC.

Grâce à cette phase d'idéation, l'entreprise est passée d'un objectif très flou (« Embrasser le mouvement Big Data ») à trois livrables actionnables à court terme.

> *La phase de convergence vise à sélectionner les meilleures idées du backlog intérmediaire. Pour cela, on fait appel à la logique, au pragmatisme et à la critique (constructive).*

Take away Data Lab

IMAGINER

- *Construire une équipe avec divers profils et compétences.*
- *Faire émerger un maximum de cas business.*
- *Les évaluer selon des critères business, techniques et organisationnels.*
- *Sélectionner les use cases prioritaires classés au top du backlog qui feront ensuite l'objet d'un prototypage.*

Matérialiser

Maintenant que vous avez sélectionné

un ou plusieurs use cases à prototyper, confrontez vos idées à la réalité du terrain et des données. Pour ce faire, il faut constituer une équipe plus technique, l'outiller de technologies adaptées et lui offrir un environnement technique lui permettant d'héberger et de manipuler de nombreuses données. Chacun de ces points joue un rôle important dans la construction d'un Data Lab, et donc, dans la démonstration de la valeur ajoutée de vos use cases.

CONSTRUIRE VOTRE ÉQUIPE

Pour créer un Data Lab opérationnel, vous devez impérativement constituer une équipe réunissant trois types de compétences distinctes :

- Une connaissance approfondie 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 et au développement d'API et de reporting.
- Une expertise mathématique et statistique afin de choisir les meilleures méthodes pour construire un modèle, puis interpréter les résultats d'une analyse.

Il est préférable de mettre en place une équipe avec des profils variés plutôt que de tout miser sur un « Super Data Scientist » idéal, un mouton à 5 pattes qui allierait ces trois compétences. D'une part, des profils capables d'adresser tous ces sujets sont extrêmement rares – si tant est qu'ils existent –, d'autre part, il sera très compliqué pour une seule personne de faire face à toutes les problématiques rencontrées.

La mise en place d'une équipe diversifiée représente un investissement important mais, nous en sommes convaincus, les résultats obtenus apporteront une très forte valeur ajoutée à l'entreprise.

Afin de réunir ces compétences, l'équipe à constituer comptera au maximum 6 profils :

- L'administrateur : son rôle est de mettre en place l'infrastructure technique, de faciliter la mise en place des processus de collecte et d'être le garant de la disponibilité du système, de sa sécurité et des choix d'infrastructure.
- Le Data Scientist : il conçoit des algorithmes pour répondre aux besoins, en se basant sur des outils de programmation pour l'analyse de données, notamment sur des méthodes statistiques liées au Machine Learning (discipline visant à construire des algorithmes auto-apprenants permettant d'automatiser des tâches complexes de prédiction ou de prise de décision).
- Le Data Engineer : ses compétences en développement et en architecture Big Data lui permettent de fournir au Data Scientist, à l'aide de l'administrateur, la plate-forme de travail nécessaire mais aussi de l'orienter sur la faisabilité en production des modèles proposés. Il joue également un rôle majeur dans la collecte et l'extraction des données.

- L'expert métier : il a une connaissance approfondie des problématiques de l'entreprise, ainsi que des enjeux business qui les sous-tendent. De plus, comprenant le travail du Data Scientist et du Data Engineer, il facilite la communication entre ces derniers et les directions métier.
- Le Product Owner : il maîtrise le processus de création et de delivery d'un produit numérique. En complément des experts métiers, le Product Owner va ainsi jouer un rôle majeur en portant la vision du produit au sein du Data Lab mais aussi à l'extérieur. Il aidera l'expert métier dans la formalisation du besoin et la priorisation des réalisations.
- L'expert Data Visualisation : il est au coeur de la communication des résultats obtenus par l'équipe du Data Lab à destination des décideurs de l'entreprise et de l'équipe marketing.

Cette liste est évidemment modulable et adaptable selon vos ressources et votre contexte. Certaines personnes physiques peuvent tenir plusieurs rôles. Par exemple, dans certains cas, le besoin en Data Visualisation est moindre ; le Data Engineer pourra prendre en charge cette fonction. Dans d'autres situations, l'expert métier sera davantage présent tout au long du processus et prendra alors le rôle de Product Owner. Considérez qu'a minima votre équipe doit comprendre un Data Scientist, un Data Engineer et un expert métier.

L'équipe complète d'un Data Lab

METTRE EN PLACE UNE DÉMARCHE AGILE

La qualité exploratoire d'un Data Lab implique un mode de fonctionnement suffisamment souple pour que les use cases puissent évoluer au cours du temps tout en définissant un cadre de travail pour l'équipe. En effet, définir un cahier des charges précis en amont des développements est une mission impossible. Au contraire, la définition de petits incréments réalisables rapidement et pouvant être confrontés au besoin permet d'explorer de manière tangible ces use cases. La démarche agile est donc une méthodologie très à propos pour une équipe Data Lab.

Il faut cependant éviter l'écueil d'une application dogmatique de l'agilité ou d'en attendre des miracles. L'efficacité de celle-ci nous semble indiscutable. Même si la prédictibilité ne sera pas optimale - n'oublions pas que nous parlons d'un laboratoire d'innovation autour de la donnée donc d'un dispositif très exploratoire - l'agilité permettra de mettre en lumière des problèmes potentiels au plus vite et de faire des arbitrages.

De plus, sur cette typologie de projet, il est important de mettre en place des outils de management visuel tels qu'un tableau Kanban *(voir TechTrends #3)*. En effet, ces projets étant souvent stratégiques, tous les niveaux hiérarchiques sont concernés par la situation et l'avancement de chacun des use cases.

Ainsi, privilégiez la mise en place d'une approche agile pragmatique qui consiste à profiter de chaque itération pour valider les hypothèses définies en amont. Les use cases sont découpés en items mesurables, puis développés. À l'issue de l'itération, chaque item est confronté au besoin :

- si les KPIs sont bons, alors l'item est conservé et le développement continue autour des hypothèses qu'il validait,
- si les KPIs sont mauvais, alors l'item n'est pas conservé et les hypothèses rattachées sont revues.

Ce mode de fonctionnement permet d'effectuer un Go / No Go après chaque itération afin de rationaliser les efforts de l'équipe. D'autre part, l'avancement des projets du Lab est visible de tous.

RÉFÉRENCER VOS SOURCES DE DONNÉES

Base de travail essentielle pour la réalisation des use cases, les données soulèvent plusieurs questions : d'où viennent-elles ? Comment les regrouper ? Quelles sont les difficultés liées à leur utilisation ? Pour y répondre, il est nécessaire de référencer les données disponibles qui seront ensuite explorées. 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 nombreuses bases de données utilisées quotidiennement représentent une grande quantité d'informations exploitables. Les traces (logs) que les utilisateurs laissent lors de l'utilisation de vos applications sont également un riche puits d'informations. Par exemple, des données de navigation web peuvent être exploitées dans vos use cases. L'historique des navigations utilisateurs permet d'obtenir une meilleure connaissance du comportement de ces derniers sur le site de l'entreprise et ainsi d'améliorer, entre autres, sa stratégie de ciblage.

Dans un deuxième temps, il est tout à fait envisageable, voire nécessaire, d'utiliser des sources de données non structurées internes telles que les e-mails ou autres documents textuels pour enrichir les données déjà à disposition. Le travail de nettoyage et de transformation sur ce type de données est certes contraignant mais les informations apportées méritent d'être exploitées de manière plus systématique.

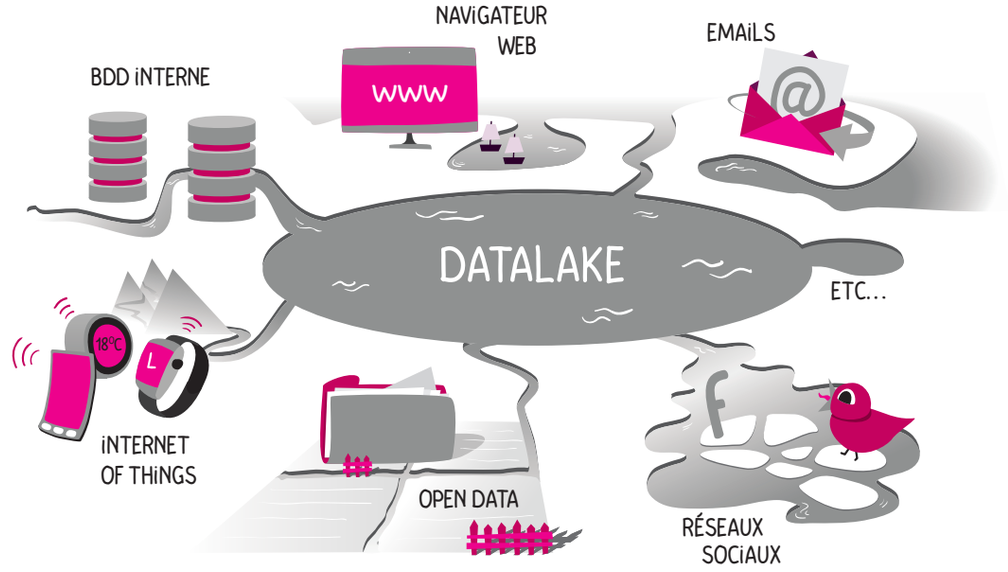
Ouvrez-vous également aux sources externes afin d'obtenir des données supplémentaires. Les réseaux sociaux, notamment, sont une mine d'informations précieuses lorsqu'ils sont bien exploités. Ces données peuvent être utilisées pour mesurer la réputation de l'entreprise ou récolter des points de vue de clients.

Attention cependant à ne pas vous limiter à ces seuls réseaux sociaux. Certaines entreprises établies dans le commerce de la donnée peuvent vous apporter énormément : informations géographiques, topologiques, sociologiques, etc. Ces Open Data issues d'entreprises et d'instituts privés ou publics diffusant des données peu sensibles et utilisables gratuitement représentent un moyen supplémentaire d'enrichir votre base. Il faut cependant s'assurer de leur qualité et de leur pertinence pour une utilisation optimale. Leur recherche et leur cartographie peuvent aussi s'avérer fastidieuses, faute de catalogue disponible.

À ne pas oublier, et ne surtout pas négliger, les données issues du monde de l'Internet des Objets (IoT) ont déjà une part importante à jouer dans une démarche autour de la donnée et auront un rôle primordial dans les mois et années à venir. La quantité de données que va générer les objets connectés sera telle qu'il sera bientôt impossible de passer à côté. Que ces données soient disponibles pour votre entreprise (via vos propres objets ou devices mobiles) ou bien achetées auprès d'entreprises tierces, elles vous fourniront de précieuses informations sur les personnes auxquelles elles sont rattachées.

METTRE EN PLACE UN DATA LAKE

Vous l'aurez compris, les sources de données sont nombreuses et disparates. Le travail de nettoyage et d'intégration fluctuera selon la qualité et l'accessibilité des données mais il est toujours possible de les « faire parler ». La question est de savoir comment regrouper et exploiter toutes ces données de manière simple sur une même et unique plate-forme. Une solution consiste à mettre en place un Data Lake.



Les sources de données disponibles à intégrer au Data Lake

L'idée est relativement simple : toutes les données à disposition sont regroupées et accessibles depuis un même endroit. Ce Data Lake pourra être complété par de nouveaux imports tout au long de la vie des projets. L'intérêt d'un Data Lake est qu'il devient très simple de croiser les données et de les enrichir, afin de répondre efficacement aux problématiques soulevées par les use cases. De plus, étant séparé du reste du système d'informations de l'entreprise, il permet un mode de fonctionnement du Data Lab ne compromettant ni le quotidien de ce dernier ni le travail des autres équipes du SI.

Afin de rapatrier toutes ses données dans le Data Lake, le Data Lab va être amené à tester et utiliser un large panel de technologies.

Pour l'alimenter, la fréquence des imports de données, la nature de ceux-ci (imports incrémentaux, remplacement ou fusion de données), ainsi que le type des données rapatriées (données provenant d'une base de données, données événementielles) devront être pris en compte dans vos choix technologiques. **Sqoop, Flume** ou encore **Kafka** pourront être choisis en fonction de vos contraintes.

Côté persistance, ce Data Lake peut reposer sur Hadoop Distributed File System (HDFS), devenu un standard de facto pour le stockage de données volumineuses sur un cluster de machines. Cette technologie ne répondra cependant pas à elle seule aux problématiques que rencontreront les membres du Data Lab. En effet, pour des données où il est nécessaire de faire une recherche aléatoire et rapide, privilégiez des bases de données relationnelles (SGBDR) ou NoSQL (Cassandra, HBase, Couchbase, MongoDB, etc.).

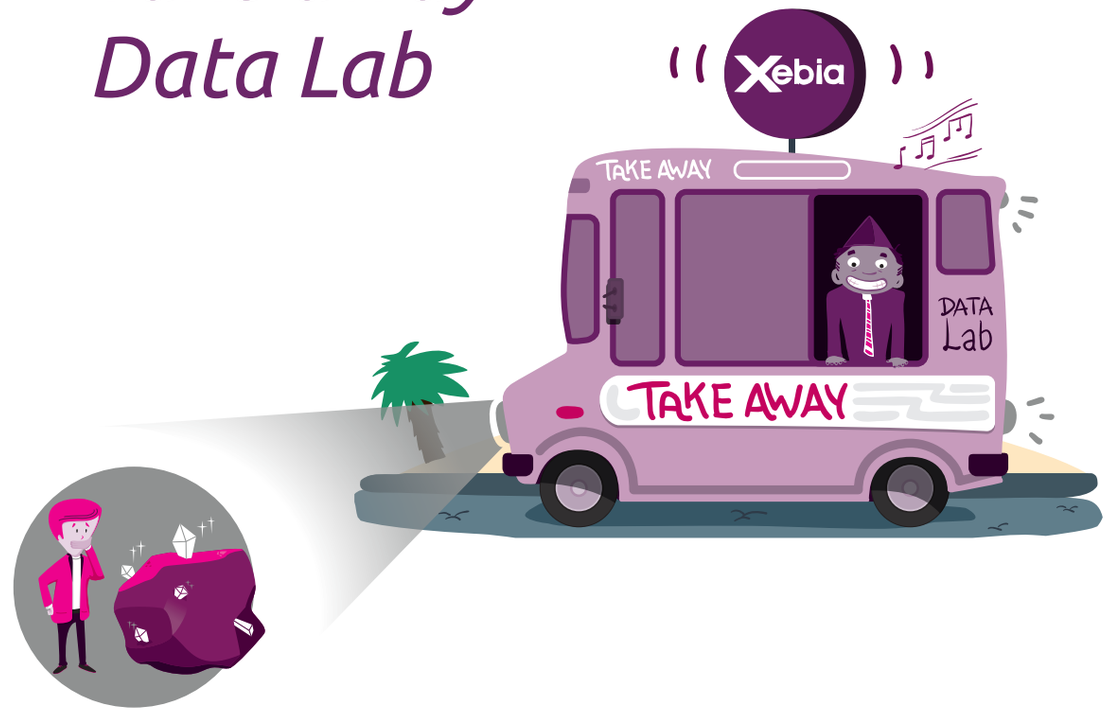
Une attention particulière doit être apportée au format de stockage des fichiers. Des formats binaires récents, tels qu'**Avro, Parquet** ou **ORC** pourront vous faire bénéficier d'une meilleure performance dans certains traitements des données ou encore vous permettre de faire évoluer facilement le schéma de celles-ci.

Construire un Data Lake signifie qu'une grande partie de vos données sera stockée et accessible à de nombreuses personnes depuis une source unique. Une partie de ces données peut s'avérer sensible (données personnelles, financières ou stratégiques). Il est alors nécessaire, pour des raisons de sécurité ou d'obligations légales, de limiter et de sécuriser l'accès à ces données.

Toutes les composantes du Data Lake doivent être protégées, ce qui nécessite d'introduire de multiples frameworks d'authentification tels que **Kerberos** pour Hadoop, **Sentry** pour Hive et **Impala, ACL** pour HDFS, etc.

Il est dans certains cas nécessaire d'aller jusqu'à la mise en place d'un système d'audit. On se penchera alors sur **Apache Falcon** ou **Cloudera Navigator.**

Take away Data Lab



MATÉRIALISER

- *Constituer une nouvelle équipe plus orientée technique et la doter des technologies les plus pertinentes.*
- *Construire une architecture adaptée à vos use cases.*
- *Déterminer les données disponibles et pouvant être collectées, tout en se protégeant des problèmes de sécurité liés à la donnée.*
- *Centraliser les données dans un Data Lake.*

Exploiter

Maintenant que votre équipe

a identifié et regroupé les données pertinentes, il est temps de les exploiter dans le cadre des use cases retenus lors de la phase d'idéation. Commencent alors les phases d'exploration et d'analyse permettant d'éprouver la faisabilité de ces derniers.

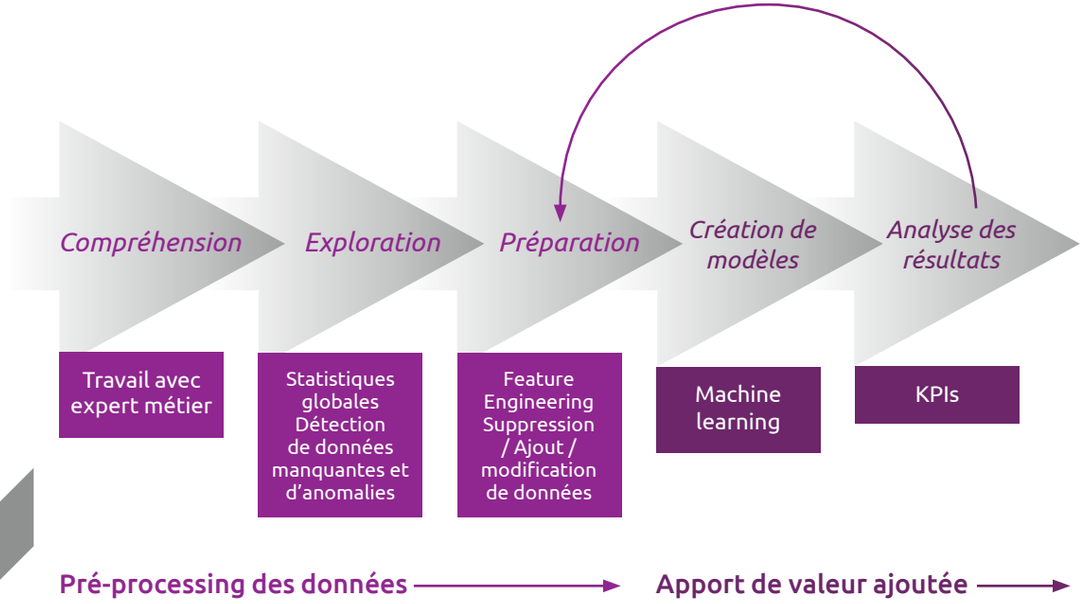
Dans le cadre de la démarche agile mise en place au sein du Data Lab, les use cases ne sont pas réalisés d'un bloc, mais par succession de petits incréments permettant de tester, valider ou écarter, puis affiner les concepts et hypothèses élaborés pour y répondre. Ces PoCs seront industrialisés progressivement s'ils s'avèrent concluants.

RÉALISER VOS POCS

Produire de la valeur métier à partir des données centralisées dans le Data Lake ne se fait pas, contrairement à certaines idées reçues, d'un coup de baguette magique du Data Scientist. Lors de la réalisation des PoCs, les rôles du Data Scientist et du Data Engineer sont prépondérants. Leurs travaux peuvent être scindés en deux grandes phases :

- Le pré-processing des données dédié à la compréhension, à l'exploration et au nettoyage des données du Data Lake dans le cadre du use case traité. Ces travaux sont fastidieux et chronophages mais indispensables à l'obtention de bons résultats.
- L'ajout de valeur métier qui se fait ensuite par modélisation des problématiques et élaboration d'algorithmes permettant de les résoudre.

À la fin de chaque itération, les résultats obtenus sont mesurés en fonction des KPIs définis afin d'orienter les travaux de la suivante dans une optique d'amélioration continue.



Tout au long du processus, le choix des outils utilisés sera piloté par les données.

Si les données sont exploitables sur une seule machine (autrement dit, de taille raisonnable), les langages R et Python seront privilégiés.

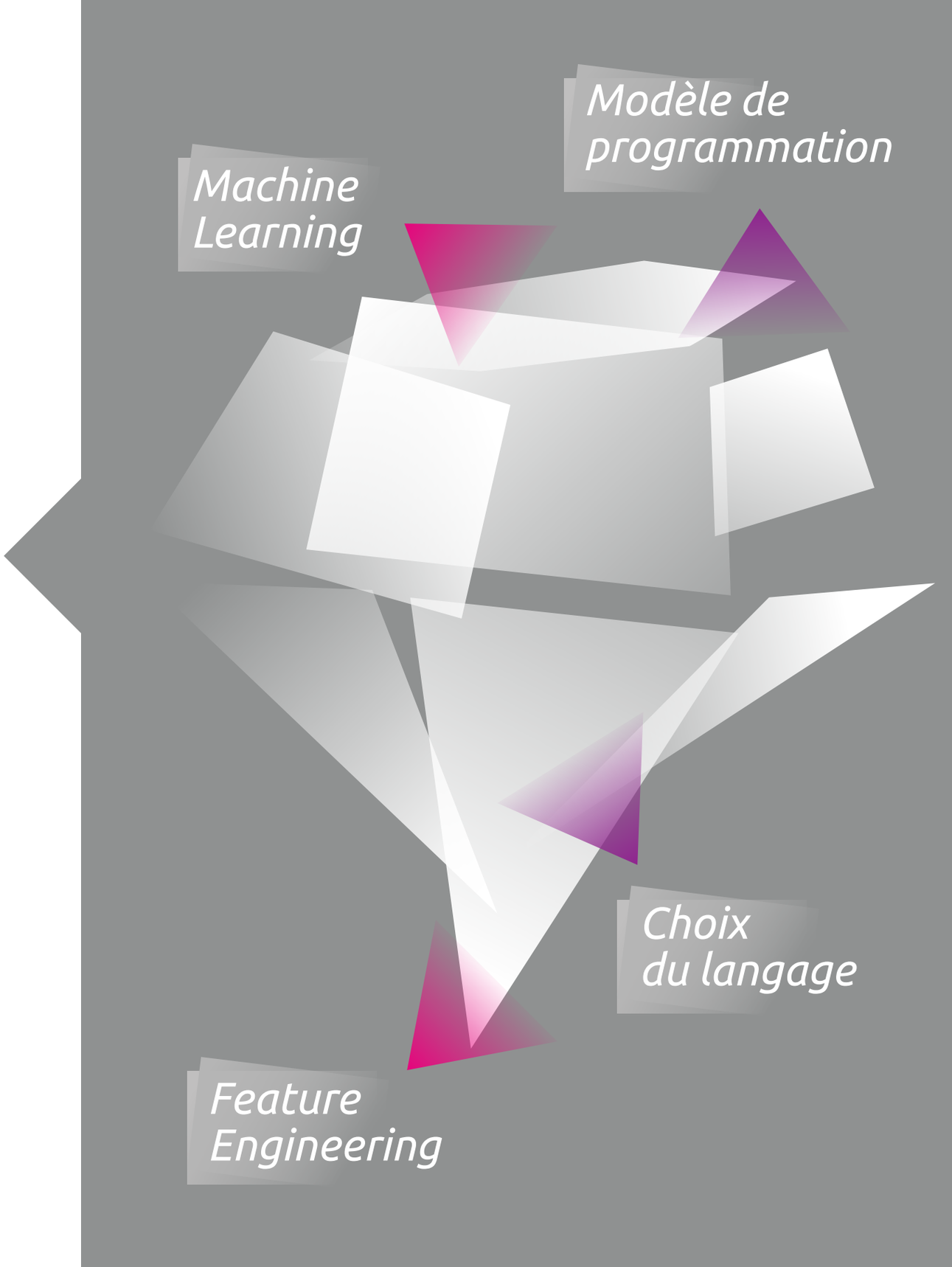
R est un langage d'analyse statistique parmi les plus utilisés chez les statisticiens. Il se repose fortement sur des packages développés et maintenus par la communauté. Ces librairies (plus de 6 000 à ce jour) fournissent un accès rapide à de très nombreux algorithmes statistiques.

Python et son écosystème fournissent également un environnement de travail plébiscité par les Data Scientists à travers :

- la librairie **Scikit-learn** proposant une API unifiée pour de nombreux algorithmes de Machine Learning,
- la librairie **Pandas** donnant la possibilité de manipuler de grands volumes de données structurés de manière simple et intuitive,
- les **notebooks d'IPython** qui offrent un environnement de travail dans une interface web permettant de mélanger exécution de code, texte et graphiques.

Si les données ne sont pas exploitables sur une seule machine (soit parce que le volume des données d'apprentissage atteint les limites physiques de la machine, soit parce que les capacités de calculs de cette dernière engendrent des temps d'apprentissage trop longs), deux approches sont possibles :

- utiliser des méthodes d'échantillonnage pour pouvoir faire les premières analyses sur un volume de données traitables sur une seule machine,
- se tourner vers des librairies de Machine Learning utilisant des frameworks de calcul distribué. Précurseur historique dans ce domaine, **Mahout,** une librairie d'algorithmes de Machine Learning en MapReduce, cède progressivement sa place à des librairies utilisant des paradigmes mieux adaptés pour des algorithmes itératifs, tels que **Vowpal Wabbit, MLlib de Spark** ou encore **H2O.**



Le pré-processing des données

Cette phase est de loin la plus longue mais sûrement la plus importante car elle conditionne l'obtention de bons résultats lors de la conception d'algorithmes. Elle se décompose en 3 étapes.

La première étape consiste à comprendre les données. Inutile, en effet, de se lancer dans l'élaboration de modèles complexes sans avoir assimilé la signification, le rôle et l'utilité des différentes données à disposition. L'implication de l'expert métier (travaillant régulièrement avec les données en question) et sa bonne interaction avec le Data Scientist, sont indispensables.

Vient ensuite l'étape d'exploration qui vise à déterminer à quel point les données sont porteuses d'informations exploitables.

Elle fait souvent intervenir des statistiques globales : Quelle est la moyenne, la variance, les quartiles d'une caractéristique numérique ? En fonction d'une variable catégorielle (homme / femme par exemple), quelle est la répartition d'une autre variable (le poids, la taille, etc.) ? Est-ce que les données ont un caractère saisonnier ? Sont-elles sparses, c'est-à-dire avec beaucoup de zéros ?

Cette phase exploratoire permet de mettre en évidence des manques dans les données, des valeurs aberrantes dues à une faute de frappe ou encore la présence de réelles anomalies.

Visualiser simplement les données permet de découvrir et de vérifier certains insights. De manière classique, les frameworks de développement proposent des librairies graphiques permettant de réaliser cette exploration visuelle. Citons, par exemple, **Matplotlib, Pyplot** et **Bokeh** pour Python, **ggplot** pour R.

La dernière étape représente la préparation (ou nettoyage) des données afin de maximiser les chances de réussite des modèles qui seront élaborés par la suite.

Cela consiste à supprimer les variables qui ne sont pas significatives pour le use case traité et celles comportant trop de données manquantes. Parfois, l'absence d'information est porteuse de sens, mais dans d'autres cas, il est nécessaire de combler au mieux les manques. Enfin, on peut retirer ou modifier les outliers (valeurs aberrantes au vu de la répartition des autres valeurs de la même caractéristique) pour éviter de créer de mauvaises prédictions.

Pour compléter la préparation des données, il est souvent nécessaire de pratiquer du Feature Engineering, c'est-à-dire de calculer de nouvelles caractéristiques utiles aux futurs modèles sur la base des données déjà à disposition. L'objectif est d'optimiser les traitements du modèle mathématique qui sera conçu. Par exemple, si vous avez à disposition des données de distance et de temps et vous savez que la vitesse sera importante par la suite, il est préférable de calculer la caractéristique "vitesse" en amont et de la fournir en entrée du modèle.

Enfin, avant d'utiliser les données pour créer des modèles prédictifs, il est souvent nécessaire de les transformer vers un format optimisé pour les algorithmes. Historiquement, cette transformation était réalisée via le modèle de programmation **MapReduce** et ses différentes abstractions (Pig, Cascading / Scalding), de manière distribuée, sur un cluster Hadoop. Depuis peu, la communauté porte plutôt son intérêt vers de nouveaux frameworks, utilisant la mémoire vive du cluster, tels que **Spark.** En complément, il est aussi possible d'utiliser un moteur de requêtes massivement parallèles (MPP) tel qu'**Impala** ou **Drill,** pour exécuter des requêtes interactives, à la manière du SQL. Enfin, des moteurs de traitement en temps réel, tels que **Storm, Spark Streaming** ou **Samza,** pourront préparer et traiter les données au fur et à mesure de leur arrivée dans le Data Lake.

> *Le pré-processing se décompose en 3 étapes : compréhension, exploration et préparation des données.*

La valorisation des données

Cette phase est la plus enrichissante. Les données vont, enfin, générer de la valeur métier.

Les techniques issues du Machine Learning permettent d'élaborer et d'implémenter des modèles mathématiques donnant un nouveau sens aux données. Des méthodes de régression, de classification ou de segmentation mettent en lumière des prédictions, des groupages, qui répondent aux use cases métier.

Parmi les plus classiques, on retrouve :

- La prédiction : à partir des données historiques, il est possible de prédire ce qui est susceptible de se produire dans le futur (augmentation ou diminution du volume de vente, du trafic, etc.).
- La détection d'anomalies : le principe consiste à repérer dans un jeu de données celles qui sortent de l'ordinaire et, par exemple, identifier des cas de fraude.

• La segmentation : cette méthode permet de séparer les données en ensembles ayant des caractéristiques communes. La segmentation permet, par exemple, de proposer une expérience utilisateur personnalisée à ses clients en fonction du segment auquel ils appartiennent.

Itérer

Les tâches de préparation des données, de construction des modèles et d'analyse des résultats s'inscrivent dans une boucle itérative. Chaque itération vise à améliorer les modèles et algorithmes élaborés. La progression de leur pertinence est suivie en mesurant, à la fin de chaque itération, les résultats obtenus en fonction des KPIs définis.

Pour ce faire, proposer un dashboard de pilotage du projet visuel grâce auquel chacun peut comprendre (en quelques graphiques simples) les avancées et les résultats obtenus à chaque itération représente un support de communication efficace pour promouvoir le ROI du Data Lab. Il constitue un outil d'aide à la décision et permet ainsi de valider un GO / NO GO pour une mise en production.

> *Les techniques issues du Machine Learning permettent d'élaborer et d'implémenter des modèles mathématiques donnant un nouveau sens aux données.*

PASSER EN PRODUCTION

C'est, bien sûr, la fin rêvée pour tout PoC. Si la valeur métier du projet est démontrée, les utilisateurs voudront rapidement bénéficier des informations produites. Cependant, pour ne pas tomber dans le syndrome 'PoC to Prod', qui conduit souvent à de retentissants échecs, il est nécessaire de transformer les expérimentations du Data Lab en un réel projet inscrit au sein du SI. Plusieurs paramètres seront alors à déterminer : l'orchestration, le partage et la visualisation des données finales.

Orchestrer

Un projet Data implique plusieurs étapes de traitement des données : récupération, validation, transformation, conversion, agrégation, analyse, etc. Ces différentes actions devront être planifiées, coordonnées, automatisées et monitorées grâce à un outil d'orchestration de workflows de jobs, choisi pour s'intégrer au reste du SI et en fonction des appétences des équipes. La palette des outils disponibles est large, allant de **scripts shells** codés en interne et schedulés à l'aide de Crontab, à des outils plus évolués comme **Oozie, Azkaban, Luigi** ou encore **Chronos**.

Outre l'orchestration, la gestion des ressources matérielles est capitale. Des frameworks sont d'ailleurs exclusivement dédiés à optimiser l'utilisation des ressources du cluster et à permettre aux différents utilisateurs et technologies qui partagent ces ressources de travailler en bonne entente. On citera les deux projets les plus actifs en ce moment : **Yarn,** introduit avec la version 2.0 de Hadoop, et **Mesos.**

Partager les résultats obtenus

Dernière action capitale pour la réussite du projet, il faut rendre les données accessibles et utilisables par les processus et les acteurs de l'entreprise. Plusieurs voies sont envisageables, de l'accès le plus simple au développement d'applications de restitutions complexes. Ces actions ne sont pas antinomiques et peuvent être menées les unes après les autres ou de concert, en fonction du niveau d'adoption et de la valeur du projet Data.

<u>Mettre à disposition les résultats obtenus</u>

La manière la plus évidente et la plus simple de proposer les résultats pour exploitation est de les stocker directement dans le file system distribué du Data Lake et de l'ouvrir. Les résultats peuvent alors être utilisés :

- de manière brute grâce à la récupération des fichiers sur le file system distribué,
- via des APIs mises en place par le Data Lab au dessus du file system distribué : **Hive, Impala** ou **Hue** par exemple,
- au travers d'outils graphiques. De nombreux éditeurs issus du monde de la BI proposent aujourd'hui de connecter directement leurs outils historiques sur des sources de données Big Data. On citera par exemple **Tableau, QlikView** et **Pentaho.**

Afin d'en faciliter l'usage, il est également courant d'adjoindre au Data Lake un ou plusieurs stockages alternatifs mieux adaptés à certaines utilisations :

- pour le parcours des résultats via des requêtes complexes, on favorisera un stockage orienté document ou clé/valeur comme **CouchBase** ou **MongoDB,**
- pour des recherches rapides sur un grand nombre de critères, on privilégiera l'indexation des données sur un cluster **Elasticsearch** ou **SolR.**

Enfin, implémenter et exposer des APIs de plus haut niveau permettent de proposer aux consommateurs des données une couche d'abstraction fonctionnelle. Se posent alors les problèmes classiques d'intégration et d'exposition d'un sous-système au sein d'un SI, par nature hétérogène. Dans le cadre de l'établissement d'un modèle prédictif, comment exposer ce dernier (implémenté en R, Python, etc.) aux applications métiers (écrites en Java, .Net, PHP ou autre) ?

- Une première stratégie consiste à sérialiser les modèles prédictifs, en utilisant les outils natifs de chacun des frameworks (**RData** pour R, **pickle** pour Python, **Serialization Java** pour Spark MLlib, etc.).
- Une autre stratégie consiste à utiliser une représentation "universelle" de ces modèles, telle que celle proposée par PMML (à noter : cette dernière est assez neuve et ne permet pas encore l'expression de tous les algorithmes).

<u>Visualiser les résultats obtenus</u>

Certaines données issues du Data Lab constituent un résultat « final » que l'on souhaite présenter de manière graphique. La Data Visualisation permet de répondre à ce besoin. Il s'agit ici, avant tout, de construire un outil de communication proposant une présentation optimale des résultats obtenus dans un objectif de prise de décision éclairée.

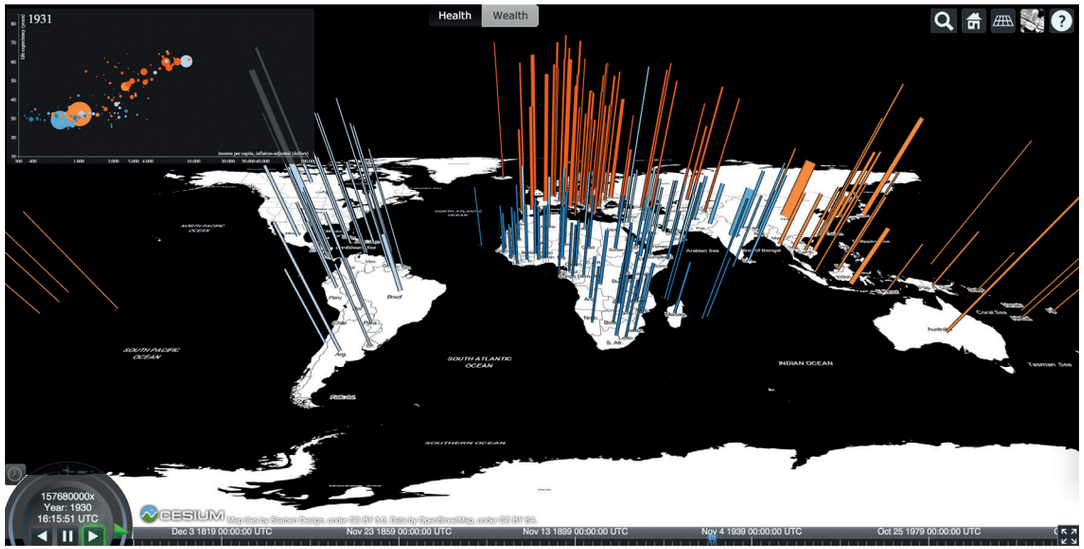
Pour être efficace, une Data Visualisation doit :

- Offrir aux utilisateurs une information rapidement compréhensible. Dans ce cadre, les applications multi-écrans sont à proscrire au profit de dashboards d'un seul tenant.
- Proposer une interactivité simple rendant facile l'application de filtres sur les données et la navigation au sein de ces dernières (en mode drill-down par exemple).

Nous pouvons citer **Kibana,** un outil simple et accessible offrant ces fonctionnalités et permettant d'arriver rapidement à un résultat exploitable.

Pour aller plus loin, la réalisation d'une application web dédiée à l'exploitation des résultats devient nécessaire. Cette approche permet de répondre à des enjeux tels que la granularité des données visualisées, les contrôles d'accès à ces dernières ou encore la personnalisation des tableaux de bord. Elle offre donc plus de richesse et de finesse dans la mise en valeur de vos données.

Les librairies de manipulation et de traitement de la donnée permettent aujourd'hui d'atteindre des résultats réellement probants. Par exemple, **D3.js** offre une manière simple de manipuler et d'afficher la donnée sous forme de représentations graphiques dans une application web.

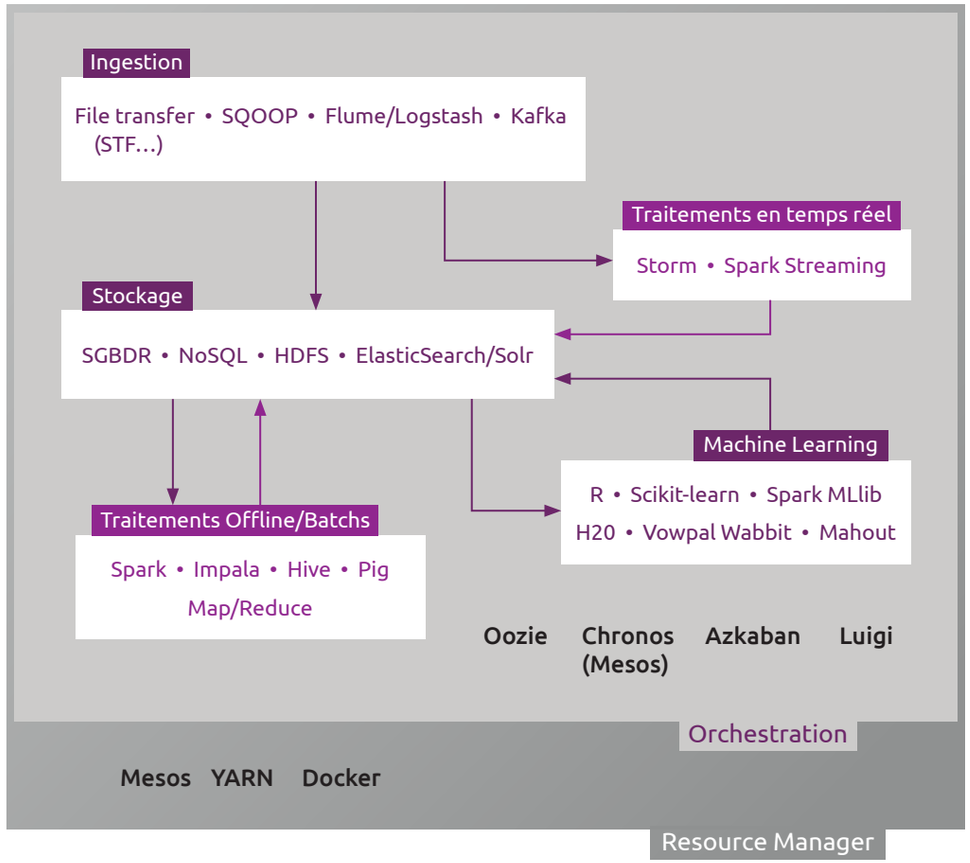


Visualisation de l'évolution de la température à l'échelle mondiale depuis 1819 à partir de D3.js

RESTER EN VEILLE

Le monde des Big Data est en perpétuelle mutation. Les changements vont en s'accélérant. Les technologies évoluent et de nouveaux paradigmes et frameworks apparaissent régulièrement. La veille technologique est donc indispensable afin que votre Data Lab soit en mesure de tirer le meilleur parti des technologies émergentes.

Bien que prendre une photo figée des forces en présence à l'heure actuelle représente le risque d'être dépassé dans quelques mois, nous nous sommes prêtés à l'exercice. Voici les technologies sur lesquelles nous vous recommandons de vous pencher dans les mois à venir :



TechRadar des technologies d'un Data Lab

Take away Data Lab

EXPLOITER

- *Réaliser votre PoC via les étapes incontournables que sont l'exploration, le nettoyage et l'analyse des données. Itérer.*
- *Pour passer ce PoC en production, gérer finement les ressources de votre cluster de données et exposer les résultats aux consommateurs finaux, via des APIs ou une Data Visualisation pertinente.*
- *Le monde de la Data est en perpétuelle évolution, effectuer une veille technologique régulière.*

Conclusion

Le Big Data apporte depuis des années son lot de promesses quant à la capacité à collecter puis corréler de nombreuses données. Des projets épars, souvent portés par les DSI, ont vu le jour sans forcément apporter la révolution promise, et ce, en grande partie à cause de l'approche techno-push adoptée dans la majorité des sociétés. Aujourd'hui, seule la valeur métier qu'on extrait des données a une importance, tout le monde utilisant peu ou prou les mêmes outils.

Le Data Lab constitue une approche permettant d'aborder les projets autour de la donnée de manière innovante et orientée business, en se basant sur les besoins des utilisateurs. Un Data Lab s'appuie avant tout sur des individus et des méthodologies.

Le choix et la complémentarité des équipes, aussi bien dans les phases d'idéation que dans les phases de prototypage, sont des facteurs clés de succès. L'implication des acteurs métier, alliée à l'expertise mathématique et technologique des spécialistes de la donnée, génère beaucoup plus de valeur que la simple mise en place d'une plate-forme Big Data, sans but business.

Le Data Lab a pour vocation de réaliser un produit numérique, traitant la donnée et la transformant en information pour l'entreprise. De facto, il reprend le cycle classique de création : idéation, prototypage, réalisation.

Les méthodologies issues du Design Thinking, avec une approche très centrée sur le métier, donnent de très bons résultats dans les phases amont du projet, lors de la définition des use cases. L'organisation d'ateliers, certains très créatifs, d'autres plus pragmatiques, débouche sur des idées de projets détaillées et mesurables, via des KPIs métier.

Une approche itérative dans les phases de prototypage permet de démontrer en continu la valeur ajoutée des initiatives. Les itérations portent aussi bien sur les données en elles-mêmes (identification, nettoyage), sur les traitements qui leur sont appliqués (algorithmiques, Machine Learning) que sur les choix technologiques qui supportent ces traitements.

Enfin, la phase d'industrialisation est, elle, plus classique, avec l'inscription du produit au sein du SI. En effet, les résultats obtenus sont souvent le point de départ de projets de plus grande ampleur, plus intriqués au sein des systèmes opérationnels et décisionnels en place. Que ce soit via le biais d'APIs ou d'applications de Data Visualisation, la restitution des résultats aux autres acteurs du SI sera l'achèvement des projets issus de la cellule d'innovation que constitue le Data Lab.

Take away Data Lab

IMAGINER

- *Construire une équipe avec divers profils et compétences.*
- *Faire émerger un maximum de cas business.*
- *Les évaluer selon des critères business, techniques et organisationnels.*
- *Sélectionner les use cases prioritaires classés au top du backlog qui feront ensuite l'objet d'un prototypage.*

MATÉRIALISER

- *Constituer une nouvelle équipe plus orientée technique et la doter des technologies les plus pertinentes.*
- *Construire une architecture adaptée à vos use cases.*
- *Déterminer les données disponibles et pouvant être collectées, tout en se protégeant des problèmes de sécurité liés à la donnée.*
- *Centraliser toutes les données dans un Data Lake.*

EXPLOITER

- *Réaliser votre PoC via les étapes incontournables que sont l'exploration, le nettoyage et l'analyse des données. Itérer.*
- *Pour passer ce PoC en production, gérer finement les ressources de votre cluster de données et exposer les résultats aux consommateurs finaux, via des APIs ou une Data Visualisation pertinente.*
- *Le monde de la Data est en perpétuelle évolution, effectuer une veille technologique régulière.*

Merci à

Paali Tandia, Laetitia Janné, Héloise Guyot et Chloé Desault

Les auteurs Xebia

Yoann
Benoit

Matthieu
Blanc

Pablo
Lopez

Julien
Buret

Thomas
Ounnas

Anne
Beauchart

Christophe
Heubès

Antoine
Michaud

Marina
Tracco

Les auteurs THIGA Hungry & Foolish

Hugo
Geissmann

Audrey
Pedro

L'auteur UX REPUBLIC

Yann
Cadoret

Design graphique : micheletmichel.com