DATA AS A SERVICE

POURQUOI AUTOMATISER VOS PLATEFORMES DE DONNÉES ?

aws | D2SI by devoteam

DATA AS A SERVICE : POURQUOI AUTOMATISER VOS PLATEFORMES DE DONNÉES ?

Nous sommes convaincus que l'entreprise Data-Driven est celle d'aujourd'hui et non celle de demain. Poussé par les entités digitales, le sujet est omniprésent. Les DSI et les directions métiers en ont compris l'importance, mais le chemin est long entre cette prise de conscience et la concrétisation d'une stratégie Data au service des métiers.
Quelles données exploiter, quelles technologies utiliser, et comment valoriser ses Data tout en maîtrisant ses coûts ? Pour clarifier ces sujets, il nous a semblé pertinent de décoder les enjeux de la donnée et les bénéfices d'une plateforme Data as a Service.

aws partner network

Premier
Consulting Partner

DevOps

Big Data

À PROPOS DE D2SI

Société en pleine croissance, convaincue que l'automatisation et les plateformes Cloud impactent en profondeur la valorisation des données, D2SI est **spécialisée dans l'accompagnement des DSI sur le Cloud**. La confiance accordée par nos clients et l'implication de nos équipes nous ont permis de co-construire de nombreux projets en particulier avec Renault Digital, Bonduelle et Adways. Nous avons ainsi obtenu le statut de **Partenaire Premier AWS** ainsi que les compétences **Big Data et DevOps d'AWS**. Spécialistes du Cloud, nos collaborateurs totalisent plus de **200 certifications AWS** et nous sommes **APN Training Partner**.

TABLE DES MATIÈRES

INTRODUCTION

Le phénomène n'est pas nouveau, les outils non plus. Pourtant le constat est sans appel : très peu d'acteurs sont en capacité d'ajouter une réelle valeur à leurs données. Les technologies sont nombreuses et hétérogènes, avec un niveau d'automatisation très faible malgré de fortes attentes sur le sujet.

Pour la DSI, l'enjeu des projets Data est d'accompagner ses métiers à se projeter sur leur utilisation de la donnée. Elles doivent adopter une plateforme scalable, simple, rapide à mettre en oeuvre et surtout adaptable en fonction des différents contextes projets. Dans l'incapacité de répondre à ces critères, les solutions de Big Data On Premises sont obsolètes by design et les TCO associés prohibitifs. Couplé à l'automatisation, le Cloud s'impose comme la plateforme d'exploitation des données. Les facteurs de succès des plateformes Data proviennent de :
+ La maturité des équipes sur la Data et les technologies associées (Big Data, Machine Learning, AI...) d'où l'importance de l'expérimentation, pour bénéficier rapidement de retours terrain,
+ L'utilisation d'une plateforme Cloud avec ses services managés, pour réduire le coût de l'innovation, accélérer le time to market et passer à l'échelle,
+ L'automatisation de bout en bout en intégrant l'interconnexion au SI, pour gagner en délai et disposer de plateforme adaptée pour chaque projet Data.

En couplant les services managés d'AWS et l'automatisation (Infra as Code et chaîne de déploiement continu) D2SI implémente des solutions de Big Data As a Service. Le nombre d'environnements, la taille et le contenu sont paramétrables à la demande par les équipes clients. Chaque équipe projet Data dispose d'une plateforme :
+ Adaptée à ses besoins, avec les outils propres à ses projets et la puissance souhaitée,
+ Instanciable plusieurs fois pour des besoins de recette, ou paralléliser les travaux,
+ Facturable à l'usage en supprimant à la volée les clusters non utilisés,
+ Disponible immédiatement et raccordée au SI,
+ Mise à jour by design à chaque déploiement, profitant des évolutions de la plateforme.

Les retours d'expériences clients :
 # Chez un grand acteur de l'industrie automobile, chaque projet peut instancier plusieurs fois son propre cluster Hadoop avec l'outillage qui lui est adapté (Amazon EMR, SPARK, Presto, Hbase...).
 # Le projet Big Data chez Euler Hermes est capable d'instancier plusieurs clusters Hadoop à la demande par environnement.

L'engouement autour du Big Data ne semble pas près de s'arrêter : le volume de données généré annuellement devrait atteindre les 163 ZB* d'ici 2025 selon Forbes. Nous produisons une quantité toujours croissante de données en particulier de qualité hétérogène (garbage in - garbage out), d'où un travail de transformation et nettoyage de la donnée qui nécessite d'être stockée et analysée. Nous parlons du **paradigme de computation ubiquitaire** : les données sont produites en tous lieux, par tous types de devices, et sous tous types de formats. Les objets connectés, Internet of Things, fournissent un flux constant de données qui peut varier avec une très grande vélocité. Cela pose de nouveaux challenges en terme de collecte et d'insertion dans les pipelines d'analyse.

Collecter massivement les données en temps réel fiabilise le résultat produit par les modèles de prédictions. Les dernières évolutions technologiques (i.e. Machine Learning, Deep Learning) ont d'ailleurs introduit la réalisation de nouveaux use cases métiers : customer churn (prédiction de la réduction de client), détection de fraude, optimisation de process industriel, maintenance prédictive...

*ZB = Zettabyte : 10^{21} bytes

1

LA COLLECTE DES DONNÉES

LES 4 PRÉREQUIS +

1.1 LA DONNÉE VALORISABLE À PLUSIEURS TITRES

La valeur extraite de l'analyse des données intéresse différents acteurs de l'entreprise :

// LA DSI

L'implémentation d'une plateforme de Big Data et l'analyse des données met souvent en lumière certains aspects de l'état d'une infrastructure ou d'une application, et ce, bien avant le traitement en production. Les cas d'usage sont nombreux tels que prédire les pannes sur le système d'information, ou analyser le temps passé sur les applications des collaborateurs.

RETOUR D'EXPÉRIENCE CLIENT

→ **D2SI a accompagné Engie dans la fiabilisation de sa plateforme Cloud AWS** avec la mise en place d'une solution d'extraction et de corrélation des logs (plateforme, infra et applicatif).

// LES MÉTIERS

La plateforme Big Data permet d'orienter la stratégie à la lumière des tendances et informations extraites de la donnée.
L'analyse du comportement des clients permet par exemple de définir de nouveaux produits pertinents, de les cibler sur les bons prospects, d'améliorer constamment la qualité de service. Ce n'est pas forcément une idée nouvelle avec le Big Data, mais l'ordre de grandeur des analyses a changé et la pertinence du choix est améliorée avec les progrès technologiques. Le Big Data confirme ou infirme des hypothèses en valorisant l'approche quantitative. On qualifie les entreprises matures dans leurs analyses, d'entreprises Data-Driven.

RETOUR D'EXPÉRIENCE CLIENT

→ Pour prédire les pannes à 48h avec un taux de réussite à 90%, un grand acteur de l'environnement analyse les données de ses vannes d'usine au travers d'une **solution Serverless sur le Cloud**.

DÉCRYPTAGE DE L'ENTREPRISE DATA-DRIVEN

Plusieurs bonnes pratiques doivent guider l'entreprise Data-Driven :

DÉCENTRALISER LES INTERROGATIONS :

Chaque collaborateur métier de l'entreprise est partie prenante sur la définition de ses KPI pour produire de la valeur directement. Les spécialistes de la Data se concentrent sur l'amélioration continue de la plateforme et l'optimisation des process Data. Commerciaux, managers et autres métiers, consommateurs, sont formés pour interroger directement les données qui répondront à leurs questions et faciliteront leur prise de décision.

RATIONALISER LES INTERPRÉTATIONS :

Nous recommandons de favoriser la confrontation des hypothèses aux enseignements de la Data, plutôt que de chercher des confirmations. L'approche neutre de la méthode rationnelle valorise efficacement les données. Les utilisateurs doivent être sensibilisés à la lecture non biaisée des données.

METTRE EN PLACE UNE STRATÉGIE DE GOUVERNANCE DES DONNÉES :

Une politique d'exposition de la Data doit être instaurée et doit évoluer pour accompagner son utilisation par des tiers : classification de données, profils d'utilisation, mesures de surveillance. Les méthodologies et les outils utilisés pour la gestion des données de références (Master Data) sont essentielles pour asseoir la stratégie de gouvernance des données. Les solutions MDM (Master Data Management) permettent de comprendre la Data de manière plus holistique :

- **Identifier** quelle vue métier sera impactée par un changement dans une Data Source du système d'information,
- **Analyser** l'impact des décisions avant application de celles-ci,
- **Suivre** le cycle de vie de la Data.

SOURCER LA DATA :

La collecte de données s'effectue en interne mais également en externe. L'entreprise Data-Driven a la capacité à détecter des signaux faibles en questionnant des données externes à l'entreprise tels que les réseaux sociaux.

DÉCIDER GRÂCE À LA DATA :

D2SI préconise d'itérer vers une cible où "toute la stratégie de l'entreprise est éclairée par les tendances extraites de la Data".

1.2 LE STOCKAGE DE DONNÉES : TOUT STOCKER

L'entreprise doit stocker l'ensemble de ses données sur une durée la plus longue possible pour ne pas se priver de futures analyses.

// Quelles données stocker ?
// Comment les stocker ?
// Dois-je les archiver ?

Les coûts de stockage dans le Cloud sont faibles (Amazon S3, par exemple) mais ces questions restent nécessaires pour optimiser les performances, la sécurité et les coûts. En bénéficiant du Cloud et d'une bonne politique de gestion de la Data, il est simple de conserver des Tera et Péta de données sans que cela représente un coût prohibitif.

> **Centraliser les données permet également de les traiter à l'échelle de l'entreprise et non plus d'un seul service.**

1.3 LES ASPECTS LÉGAUX À BIEN ANTICIPER

Les questions juridiques restent complexes, suivant le déploiement sectoriel (financier, politique, santé...), la localisation et le type de données, la législation diffère.
→ Il est indispensable de travailler avec les métiers pour connaître ses données et ses processus business.

Les lois principales en France (liste non exhaustive) sont les suivantes :
> Loi Informatique et Libertés
> RGPD
> Cloud Act
> Loi de programmation militaire (OIV)

Par exemple : La CNIL impose la conservation des données relatives à la gestion de la paie ou au contrôle des horaires des salariés pendant 5 ans ou la suppression des données d'un prospect qui ne répond à aucune sollicitation pendant 3 ans.

La négligence dans le respect des lois sur le type de données traitées peut entraîner une sanction administrative et pénale, ou impacter l'image de l'entreprise (cf. le scandale Cambridge Analytica qui a provoqué la cessation immédiate des activités de cette dernière, et la sanction de l'UE envers Facebook suite au croisement des données utilisateurs après le rachat de WhatsApp).

Les principales obligations légales portent sur les points suivants :

LES TYPES DE DONNÉES COLLECTÉES ET TRAITÉES

En France, légalement, il y a deux types de données : celles à caractère personnel, et celles qui ne le sont pas. Une donnée personnelle permet d'identifier **directement ou indirectement** une personne physique. Les articles 2 de la loi "Informatique et Libertés" et 4 de la RGPD donnent une définition plus juridique et complète.

→ Dans un contexte international, les aspects juridiques sont à initier en amont des projets, les lois des pays régissent l'utilisation des données.
Ex : La Russie, la Chine interdisent l'export de données hors de leur pays. L'Angleterre a également renforcé en début d'année ses lois en la matière.

LES TRAITEMENTS EFFECTUÉS SUR LES DONNÉES

L'entreprise doit maîtriser la sécurisation des processus de traitement (collecte, enregistrement, consultation, suppression...). L'anonymisation des données, toujours souhaitable, n'est pas applicable dans tous les contextes. La donnée peut être utilisée à des fins business. On parlera d'anonymisation irréversible ou de pseudo anonymisation (possibilité d'identifier la personne indirectement).

→ Les partenaires, personnes ou organismes interagissant également avec les données de l'entreprise devront aussi être conformes aux lois "Informatique et Libertés" et RGPD (chapitres III et IV de la RGPD).

LE CAS SPÉCIFIQUE DES DONNÉES DE SANTÉ

Les données de santé, définies aussi dans l'article 4 de la RGPD, sont des données à caractère personnel sensibles, pour lesquelles les réglementations sont plus strictes. L'une des principales spécificités des données de santé est la notion d'Hébergeur de Données de Santé (HDS). Un hébergeur de données, peut-être "RGPD compliant", sans avoir la certification HDS.

DÉCLARATION DES ACTIVITÉS À LA CNIL

Tout traitement de données personnelles doit faire l'objet de formalités préalables (déclaration, demande d'autorisation) auprès de la CNIL.

→ Travailler de pair par itération avec l'équipe sécurité de l'entreprise et le DPO (Data Protection Officer, nouveau métier introduit par la RGPD) est nécessaire pour bénéficier de retours terrain rapidement et construire une cible répondant aux exigences de sécurité. Le DPO s'occupera de toutes les questions de conformité et l'équipe sécurité de celles relatives à la gouvernance des données dont l'entreprise est en possession.

1.4 LES NOUVEAUX MÉTIERS DE LA DATA

L'ensemble des contraintes et opportunités apportées par le Big Data font émerger trois nouveaux métiers spécialisés dans le traitement et l'exploitation de grands volumes de données :

DATA SCIENTISTS

Une entreprise qui croît rapidement doit impérativement comprendre la valeur de ses données recueillies qui sont souvent désordonnées et massives. La popularisation du métier de Data Scientist est un résultat naturel du boom de la donnée et du changement de vision des entreprises.

Les Data Scientists sont à la fois mathématiciens, statisticiens et informaticiens. Ils ont la capacité de dégager des tendances à partir des données et sont l'interface entre le business de l'entreprise et l'IT. Ils accompagnent les métiers pour identifier la valeur contenue dans les Data, comment l'extraire et définir quelles données corréler pour être plus pertinent.

Chargés de développer des modèles d'apprentissage automatique pour l'analyse des données, les Data Scientists doivent avoir :
→ une expérience des outils statistiques (R ou autres),
→ la capacité à créer et à évaluer des modèles prédictifs complexes,
→ des compétences en storytelling pour être capable de restituer et vulgariser des informations complexes,
→ une compréhension complète du domaine de l'entreprise.

DATA ENGINEERS

Les Data Engineers sont chargés de la préparation de l'infrastructure Data pour permettre aux Data Scientists de mener leurs analyses et de développer des applications orientées Data.

Les Data Engineers conçoivent et construisent des infrastructures à grande échelle intégrant des données de tous types, souvent hétérogènes, en provenance de plusieurs sources différentes. Ils peuvent également mettre en place des pipelines d'ETL (Extract, Transform and Load) pour alimenter des entrepôts de données. Ces derniers sont utilisés pour le reporting ou l'analyse par les Data Scientists et les Business Analysts.
D'avantage concentrée sur la conception et l'architecture, leur culture des outils de Machine Learning et d'analyse est une réelle plus value pour les projets.

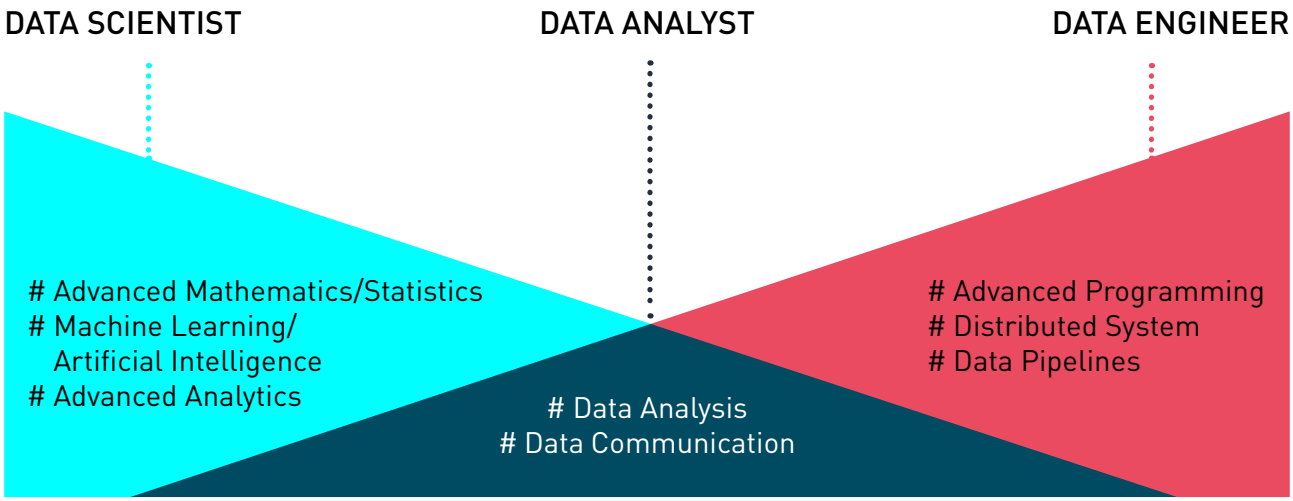
→ Les architectures élaborées par les Data Engineers sont conçues pour être automatiquement scalables. La tolérance aux pannes ou ce qu'on appelle parfois la haute disponibilité est un des volets à mettre en place impérativement dans une architecture Big Data.

DATA ANALYSTS

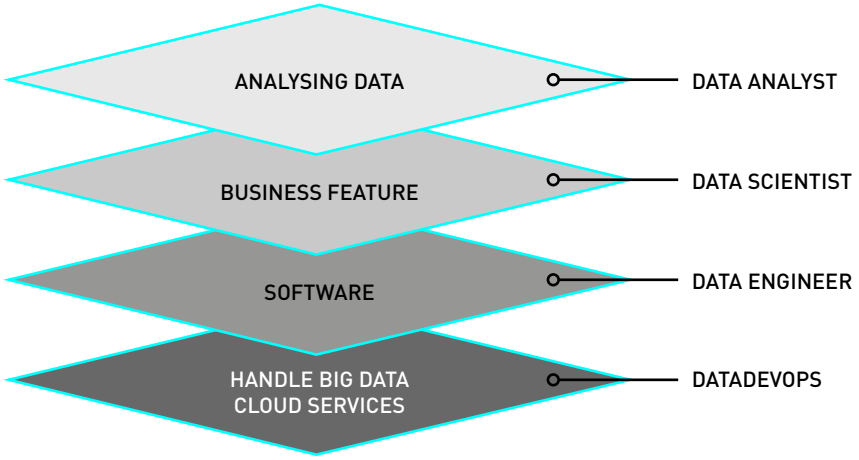
La principale responsabilité des Data Analysts est d'intervenir au profit des métiers.

Comment un responsable marketing peut-il utiliser des données analytiques pour lancer sa prochaine campagne ? Comment un représentant peut-il mieux identifier les données démographiques à cibler ? Comment un PDG peut-il mieux comprendre les raisons sous-jacentes de la croissance récente de l'entreprise ?
Les Data Analysts fournissent des réponses à toutes ces questions en effectuant une analyse et en présentant les résultats. Ils entreprennent le travail complexe d'exploiter des données pour apporter de la valeur à leur organisation.

→ Le rôle du Data Analyst semble similaire à celui du Data Scientist mais ils sont en réalité complémentaires.
Par exemple : le Data Engineer implémente la plateforme de données dans laquelle un Data Analyst peut extraire un nouveau jeu de données à l'aide de l'API et commencer à identifier les tendances intéressantes dans ces données. L'analyste résumera et présentera ses résultats de manière claire, ce qui permettra aux équipes non techniques de mieux comprendre leurs données et comment celles-ci se comportent. Enfin, le Data Scientist s'appuiera probablement sur les résultats initiaux de l'analyste et cherchera plus de possibilités pour en tirer des conclusions. Que ce soit en formant des modèles de Machine Learning ou en effectuant des analyses statistiques avancées, le Data Scientist fournira une nouvelle perspective prédictive sur l'avenir.



Source : bigdatainstitute.io



2

DE L'AUTOMATISATION À LA DATA DANS LE CLOUD

2.1 L'AUTOMATISATION D'INFRASTRUCTURE : UN PASSAGE OBLIGÉ

De l'automatisation d'infrastructure à l'automatisation du traitement des données dans le Cloud.

Si les entreprises collectent des données depuis longtemps, l'explosion des volumes de données fait que cette collecte doit être **automatisée**.

L'automatisation de l'infrastructure, rendue possible avec les plateformes Cloud, apporte de nombreux bénéfices pour le traitement des données :
→ **La flexibilité** d'avoir une plateforme Big Data par besoin (on parle de Big Data As a Service),
→ **L'optimisation des coûts** en supprimant à la volée les clusters non utilisés,
→ **La simplification** des déploiements des évolutions d'une plateforme,
→ **La maintenabilité** des ressources et des services managés,
→ **La scalabilité** du pipe de projets Big Data.

2.2 LE CLOUD : UN ENVIRONNEMENT PROPICE POUR LE BIG DATA

Le Cloud représente un nouveau modèle pour l'IT, dans lequel tout est automatisable. La résilience et la performance sont portées par les applications et non plus par les infrastructures.

Les entreprises souhaitant mettre en place des systèmes Big Data n'ont pas souvent une vision claire de leurs objectifs.
→ **L'expérimentation et l'itération** vers une cible sont nécessaires pour démontrer l'intérêt du Big Data dans un sous-domaine spécifique.

Si les infrastructures dédiées au Big Data peuvent être déployées On Premises, les entreprises n'ont aucun intérêt à supporter leur propre infrastructure. Les infrastructures internes sont chères, limitées en capacité matérielle, complexes à maintenir et à exploiter, peu flexibles et trop longues à mettre en œuvre. Devant les échecs de la mise en place de plateformes Big Data On Premises, nous observons que les DSI se retranchent trop souvent derrière la phrase suivante : "Le client ne sait pas ce qu'il veut".

Pour nous, c'est aussi leur rôle d'accompagner les métiers dans la définition de leurs besoins.

// Le Cloud est l'environnement naturel des systèmes Big Data, notamment pour les gains :

 # D'agilité, afin de répondre plus rapidement et plus précisément aux besoins métier,
 # De simplification de construction des plateformes grâce aux services managés et au Serverless qui permet de se concentrer sur les problématiques métier plutôt que l'infrastructure,
 # De réduction des coûts avec le paiement à l'usage, favorisant l'expérimentation et l'innovation,
 # D'élasticité afin de s'adapter sans effort et sans contrainte au volume croissant de données en conservant les performances,
 # De sécurité des données en abaissant le coût de son implémentation,
 # De richesse fonctionnelle grâce aux briques de collecte, stockage, analyse et visualisation directement prêtes à l'emploi.

// L'intérêt du Serverless dans les architectures Data est triple :

 # Simplifier la maintenance de la plateforme pour se concentrer sur des tâches à valeur ajoutée pour les métiers,
 # Construire de nouvelles architectures "Event-Driven" (les traitements d'analyses sont déclenchées et facturées uniquement lors d'événements sur la plateforme),
 # Paralléliser les traitements sur le même jeu de données.

RETOUR D'EXPÉRIENCE

L'UTILISATION DU CLOUD PAR ADWAYS

Partant du constat que tous les supports (mobile, tablette, desktop) sont interactifs et que les vidéos ne le sont pas, Adways propose un outil SaaS permettant à l'utilisateur d'éditer ses vidéos en y ajoutant des hotspots associés à une fonction : une fiche produit, une vidéo dans la vidéo...

Avec le déploiement de la solution en France et à l'étranger, le volume de trafic explose. En publicité, la statistique est un élément de facturation, d'où son importance pour Adways. Afin de mieux mesurer et analyser les interactions vidéos, Adways souhaitait évoluer d'une plateforme analytics basique basée sur des compteurs vers une véritable plateforme Big Data. L'objectif était notamment de répondre à la montée en charge du trafic (enjeu de scalabilité), et de doter la plateforme de nouvelles fonctionnalités, comme par exemple pouvoir requêter tous les événements passés et non seulement des compteurs.

D'un point de vue fonctionnel, il s'agissait de fournir les KPI métier au plus proche du temps réel par une solution de collecte et de traitement sur un volume accru de données. Cette plateforme devait également s'adapter aux contraintes d'une start-up : il fallait une solution simple à développer et à maintenir, peu onéreuse et surtout scalable.

→ **La solution retenue est un pipeline utilisant les services managés Amazon Kinesis, AWS Lambda, Amazon S3 et Amazon Athena d'AWS.** L'intérêt de cette solution est double : reposant sur des services managés, elle est **simple à exploiter** et surtout permet d'**avoir le même mécanisme, la même base de données**, pour collecter des statistiques tous formats confondus (interactive, publicité, éditorial).

ARCHITECTURE HYBRIDE : UNE STRATÉGIE CONTRADICTOIRE

Certaines entreprises envisagent de mixer les solutions Big Data :

→ **Cloud Native** pour bénéficier de la puissance de calcul
→ **et On Premises** pour conserver une partie de ses données (de nature critique ou sensible).

Si une architecture "en débordement" sur le Cloud est techniquement possible, cette stratégie est contradictoire :

Dans le contexte de la Data, il est indispensable de concentrer ses données pour éviter un écroulement des performances.
La consommation des outils Big Data Cloud sera calquée sur les pratiques internes. **Les solutions hybrides ne bénéficieront pas des services managés et de l'automatisation**.

→ **Ces solutions hybrides deviennent rapidement complexes.**
La sécurité des données reste nébuleuse, la gouvernance est à deux vitesses et la facturation ne profitera pas du paiement à l'usage.

Nous privilégions l'adoption d'une PLATEFORME DE SERVICES BIG DATA SUR LE CLOUD et le traitement des impacts sur l'organisation de la DSI.

2.3 L'EXPLOSION DES SERVICES MANAGÉS

Principalement basés sur des outils open source de la stack Hadoop, la quasi totalité des outils Big Data, ont un équivalent en terme de fonctionnalités dans le Cloud.

Les services managés facilitent la gestion des infrastructures en :
→ Fournissant des mécanismes simples et rapides pour modifier l'infrastructure, suivre son état de santé et permettre le self-healing et la haute disponibilité,
→ Simplifiant le déploiement de nouvelles stacks et la mise à jour des stacks existantes,
→ Mettant à disposition des APIs permettant l'interopérabilité avec les outils d'automatisation,
→ Libérant les utilisateurs de la complexité de la gestion des patchs et en mettant en place des règles d'accès fines aux éléments de l'infrastructure.

PRENONS POUR EXEMPLE LE SERVICE AMAZON EMR
(Elastic Map Reduce : infrastructure Hadoop managées sur AWS).

Le temps moyen de création d'un cluster Amazon EMR avec les outils les plus communs du monde Hadoop tels que Hive, Spark, Hue, Zeppelin, Hbase, **est d'environ 12 minutes.**
La mise en place d'un cluster Hadoop On Premises **se mesure en mois.**

Ces services managés disposent de l'interopérabilité des services AWS et permettent des design patterns tels que les "Event-Driven Architecture".
→ **Il est d'une facilité déconcertante de créer une solution d'ingestion de données basée sur de l'architecture "Event-Driven" qu'elle soit batch ou near-realtime.**

Certains besoins autour de l'accréditation et de la gouvernance de la donnée sont difficiles à mettre en application sur Amazon EMR. Les outils sur l'écosystème Hadoop tels que Ranger, Atlas ne sont pas encore supportés. Un effort supplémentaire d'industrialisation et d'automatisation des composants Ranger ou Atlas doit être réalisé pour répondre à ces besoins dans l'architecture.

→ Il est alors envisageable de lancer rapidement des clusters Hadoop Vendors tels que Hortonworks ou Cloudera depuis la marketplace AWS.

LES SERVICES CLOUD ORIENTÉS DATA DANS L'INDUSTRIE AUTOMOBILE

Un grand acteur de l'industrie automobile dispose d'un pipe projet conséquent, les équipes ont besoin de passer à l'échelle. L'usage du Cloud AWS et plus particulièrement des services managés d'AWS ont contribué au succès des projets Data métiers.

L'objectif était de fournir un Cluster Hadoop séparé, avec ses outils et ses données sur AWS par équipes de Data Scientists et Data engineer. L'équipe DevOps Big Data a fourni l'assistance dans le design et l'automatisation de cette infrastructure :
→ **Déploiement automatisé d'un cluster Hadoop**
→ **Utilisation de l'authentification LDAP du client** pour l'accès aux outils interne du cluster Hadoop comme Zeppelin et Hue
→ **Optimisation des coûts**
→ **Chiffrement des données**

Dans cette optique, l'équipe a conçu une architecture Stateless permettant de fournir un cluster Amazon Elastic Map Reduce clé en main avec les outils Spark, Hive, Zeppelin et Hue etc. Cette architecture stateless s'appuie sur plusieurs types de stockage :
+ Stockage des données métier sur Amazon S3
+ Externalisation des métadonnées Hive, Hue sur Amazon RDS
+ Stockage notebook zeppelin sur Amazon S3

Cette réalisation optimise l'usage des clusters Amazon EMR : nous avons la flexibilité de détruire puis de reconstruire les clusters la nuit et le week-end sans perte de données et de livrer la dernière release Amazon EMR aux équipes projets avec un temps d'interruption de service minimum.

AMAZON EMR AS-A-SERVICE

L'automatisation est au cœur de la plateforme permettant à ce client de disposer de Big Data As a Service. La taille et le contenu (outils de l'écosystème Hadoop fournis avec le cluster Amazon EMR : Spark, Hive, View, Zeppelin, Hbase) des clusters fournis sont configurés en fonction des besoins de l'équipe.

→ **Ces clusters sont fournis de façon à être immédiatement utilisables** : les outils tels que Zeppelin, Hue sont paramétrés pour consommer les outils d'entreprise (GitLab, Nexus et les solutions d'authentification LDAP du client). L'enjeu n'est pas simplement de consommer le service Amazon EMR, mais de l'interconnecter, de le mettre au centre des outils et de la stratégie client.
L'on-boarding d'un Data Scientist ou d'un Data Engineer disposant immédiatement de l'ensemble du pipeline pour travailler sur un cluster Hadoop directement interconnecté au SI a apporté énormément de **flexibilité et de time to market** pour les équipes projets.

TOUCHE D2SI

D2SI a procédé à une ingénierie financière afin de réduire les coûts de la plateforme autant que possible.
→ Une partie du cluster est lancée sur des instances spot, le reste sur des instances à la demande ou réservées.
→ L'auto-scaling est également appliqué pour éviter la surconsommation d'instances
et l'usage d'Amazon S3 est privilégié pour les données au repos.
→ Toutes les couches sont construites suivant les principes de l'infrastructure as code, avec Terraform et Ansible.
→ Côté sécurité, les données des clusters au repos et en transit sont chiffrées à la demande,
via le "security configuration" d'Amazon EMR créé par la stack Terraform.

LES SERVICES CLOUD ORIENTÉS DATA DANS L'INDUSTRIE

L'un de nos clients dans l'industrie mène une importante activité exploratoire de la donnée. Pour certains projets, il est nécessaire de réaliser un "pipeline d'ingestion de données" afin de mettre à disposition les données pour les équipes de Data scientists (siloté par projet).

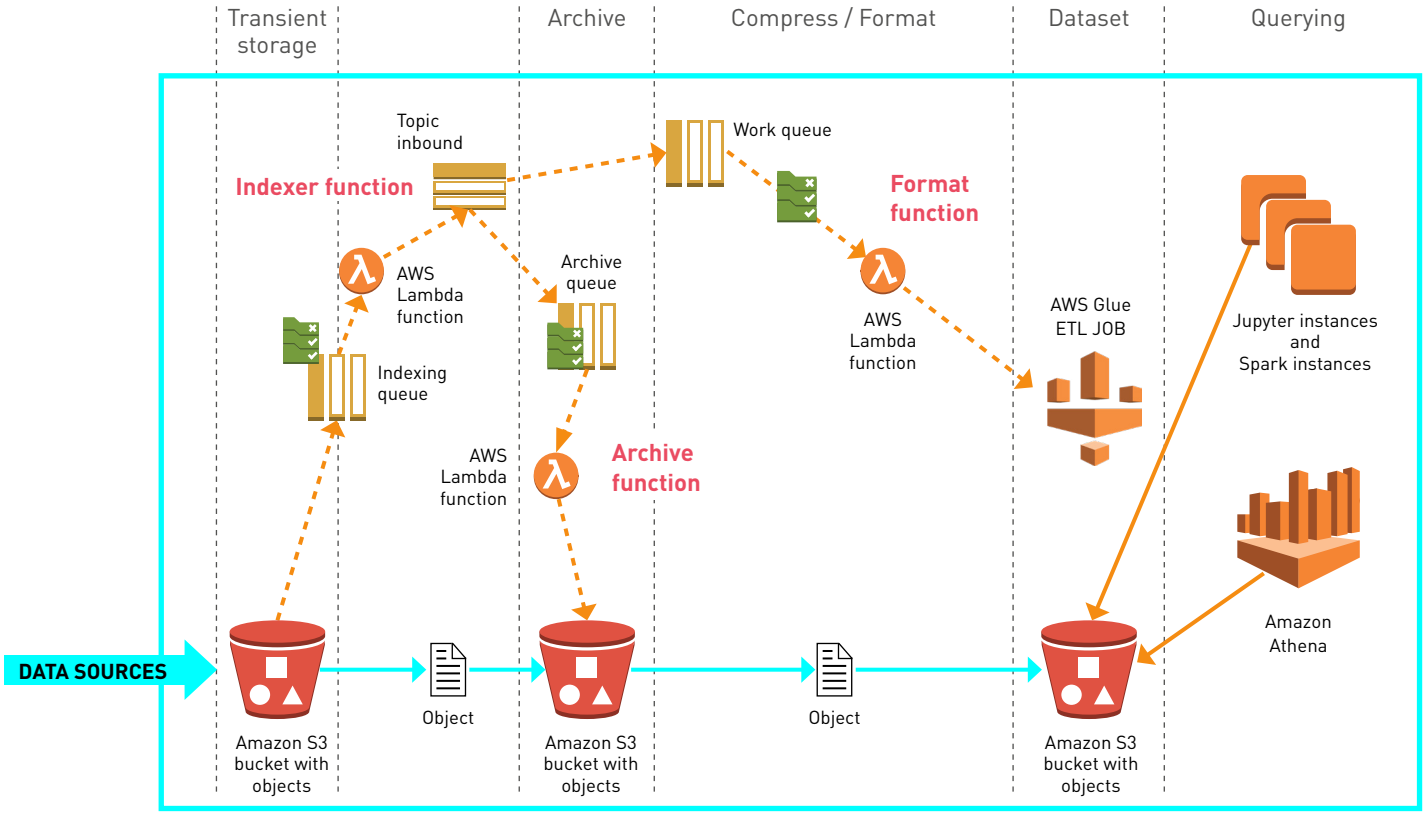
Nous avons réalisé un pipeline d'ingestion avec une architecture Serverless "Event-Driven". Ce client pousse l'utilisation des services managés dans sa stratégie Cloud.
Voici les concepts clés du pipeline :

→ **Serverless**
→ **Cheaper**
→ **Ingestion At Scale**
→ **Fault tolerant**
→ **Optimize output format (Apache Parquet)**

Pour répondre à ces concepts clés, nous avons implémenté une plateforme bénéficiant de l'interopérabilité entre les services managés Amazon S3, Amazon SNS, Amazon SQS, AWS Lambda, AWS Glue et l'outil d'infrastructure as code AWS CloudFormation.

Amazon Athena interrogeant directement les données optimisées au format Parquet, nous avons évité de construire une solution de "SQL query engine MPP". Ce fut un gain de temps projet considérable.

Un pool de machine customisées Amazon EC2 avec les outils de Data science tels que Jupyter, Spark, sont à disposition des équipes Data Scientists pour les activités exploiratoires de la donnée.



3

QUELS MODÈLES DE DONNÉES POUR QUELS BESOINS ?

3.1 MISE EN PLACE D'UN SOCLE DE DONNÉES COMMUN

La centralisation des données est la clé de voûte des projets Data. La mise en œuvre d'un Data Lake doit être priorisée mais dans une approche itérative. Démarrer avec un référentiel commun dans le SI permet de commencer à "cruncher" la donnée et de trouver des corrélations.

La centralisation des données permet de :

→ **Simplifier la stratégie de gouvernance des données** et la mise en oeuvre de mécanismes de sécurité tels que le chiffrement ou l'audit permanent.
→ **Maximiser les performances** pour le processing des données batch ou temps réel.

3.2 PRISE EN COMPTE DU CYCLE DE VIE DES DONNÉES POUR GAGNER EN AGILITÉ

Pour répondre aux nouveaux besoins tout en bénéficiant de modèles de données agiles, à chaque étape de transformation, une partie ou la totalité des données est dupliquée puis conservée afin de régénérer et refaire plus tard des calculs de données dérivées.

Les données peuvent-être froides, chaudes et ultra-chaudes. Chaque niveau de données doit avoir son propre support de stockage afin que celles les plus consultées (ultra-chaudes) soient stockées sur des supports adaptés à une lecture à très grande vitesse (RAM, SSD...) et de très faible latence.

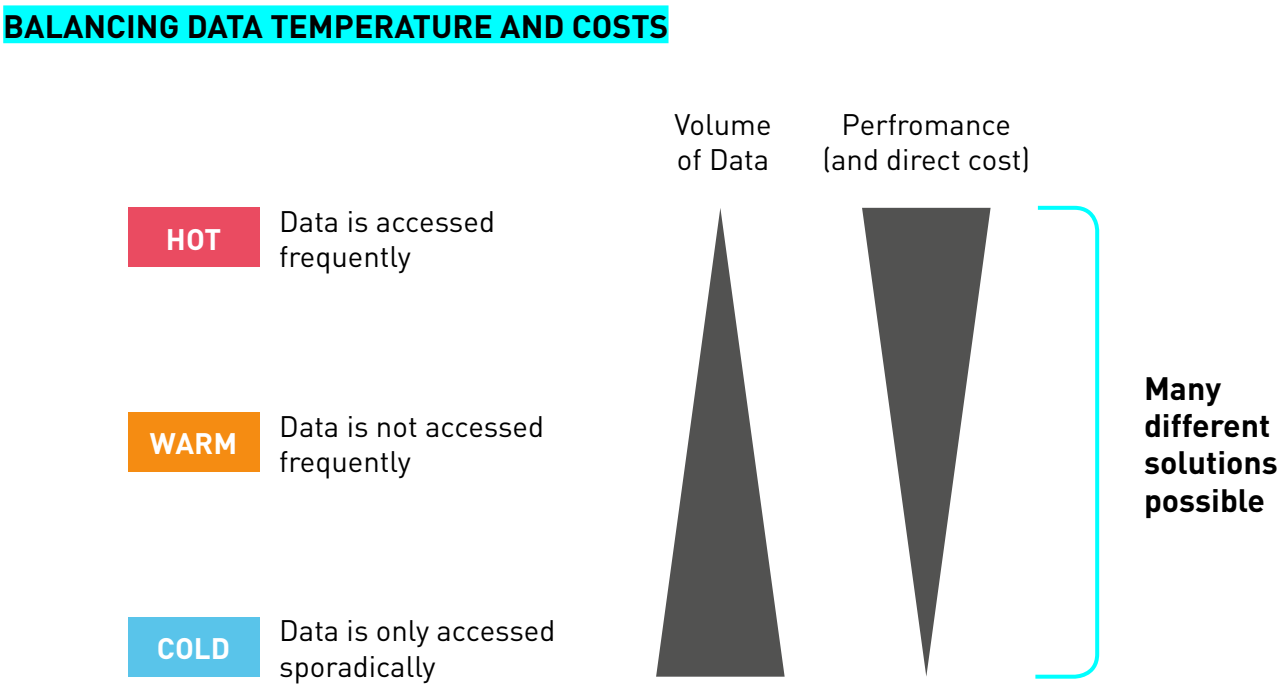
→ Le découpage de données peut se faire sur plusieurs critères comme la criticité, la fréquence de lecture et d'écriture, le volume et le facteur temps.

HDFS, dans l'écosystème Hadoop, joue le rôle d'un système de stockage distribué, tolérant aux pannes, scalable et durable. HDFS est la proximité des données avec les nœuds de calcul - on parle de Data-Locality - mais impose l'ajout des nœuds et donc de la puissance si le volume de données augmente. Cela peut générer des coûts importants et des contraintes de scalabilité et d'opérations sur les nœuds HDFS.

Amazon S3, système de stockage objet, supporte nativement la gestion de cycle de vie d'objets et **s'intègre parfaitement dans l'écosystème Hadoop** grâce à EMRFS qui expose Amazon S3 pour un cluster EMR comme si c'était du HDFS.Il permet de s'abstraire des problématiques de stockage et de configuration avancée que réclament les solutions comme HDFS et se configure de façon à archiver les données froides sur Amazon Glacier pour optimiser les coûts. Amazon S3 permet de séparer la partie calcul de la partie stockage, d'éteindre le cluster de nœuds en cas d'inactivité (à des intervalles régulières comme les nuits et les week-ends, ce qui réduit les coûts).

Une architecture bénéficiant de plusieurs modèles de stockage en fonction de la température de la donnée est pertinente (stockage mémoire, HDFS, Amazon S3, Amazon Glacier). Ce modèle permet de tirer pleinement profit de la puissance du Cloud et de dépasser les limites opérationnelles et budgétaires des infrastructures locales ou non managées.

Plusieurs niveaux de données :



Le schema-on-read est un mécanisme important dans le cycle de vie de données.

→ **Cette solution d'architecture est importante parce que le socle de données Big Data devrait avoir un premier niveau de données stockées dans un format brut qui servira plusieurs cas d'utilisation au sein de l'organisation.** Par la suite, chaque utilisateur devrait appliquer des transformations sur les données en suivant un modèle de données adapté à ses besoins à la volée pendant la lecture. Ce mécanisme permet de booster l'agilité dans le développement Big Data et le traitement de données et d'avoir un système réactif aux changements. Il est ainsi possible de tester plusieurs modèles, de revenir en arrière et de simplifier le cycle de vie des données.

3.3 FAST DATA PROCESSING : UN SYSTÈME EN TEMPS RÉEL

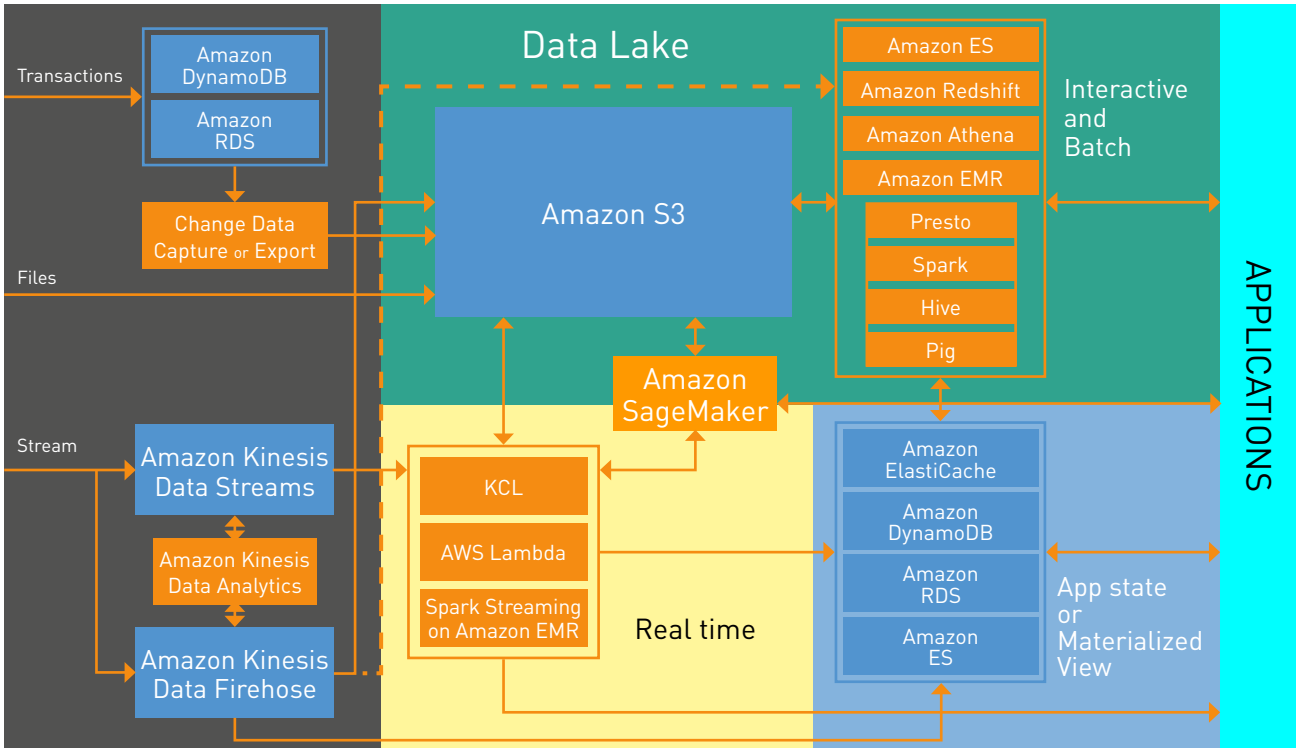
Le Fast Data concerne principalement le traitement des données en mouvement (rapide), alors que le Big Data fait souvent allusion à des traitements en batch sur des terabytes ou petabytes de données au repos.

Les approches du Big Data se reposant sur des Data Warehouses traditionnels posent souvent des problèmes de latence, ce qui n'est pas adapté aux nouveaux cas d'utilisation de la donnée.

Parmi ces cas d'usages, on peut citer les systèmes de recommandation online, la personnalisation de services selon le pattern d'utilisation de l'utilisateur, la détection de fraude, etc. Cette émergence de nouveaux cas d'usages a engendré un changement radical de paradigme. Aujourd'hui, il est primordial de prendre en compte tout le flux (Stream) des événements entrants au lieu de ne garder que le résultat des ces événements.
→ Dans un contexte d'application e-commerce par exemple, on va garder des informations sur les étapes par lesquelles le client est passé avant d'annuler sa commande, ce qui permet de détecter d'éventuels problèmes avec la procédure d'achat et d'améliorer la plateforme e-commerce.

La Fast Data repose sur des systèmes réactifs qui donnent beaucoup d'importance à la résilience et à la scalabilité et qui sont dits **Message-Driven**. On parle alors de **temps réel**.

ARCHITECTURE MULTI-COUCHES : FAST-DATA ET AI

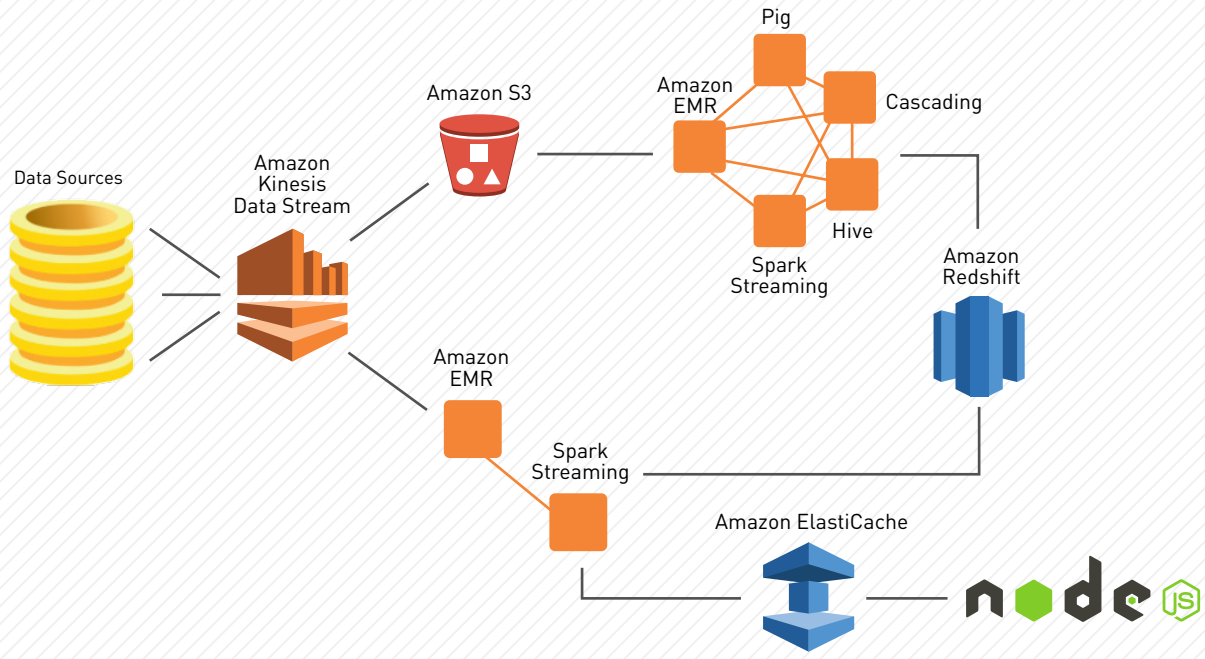


EXEMPLE D'ARCHITECTURE SERVERLESS

L'éco-système AWS offre de nombreux outils interopérables qui facilitent la mise en place de plusieurs architectures.

→ Il est ainsi possible d'intégrer Amazon Kinesis dans la stack applicative Big Data pour écouter instantanément les sources des données et les injecter directement dans le référentiel de données et/ou en faire un traitement en temps réel.

→ Certaines architectures sont ainsi basées sur les solutions Amazon Kinesis / Amazon EMR / Amazon Redshift (Cf retour d'expérience Adways en page 14).



CONCLUSION

Le Big Data est une réponse logique à l'afflux toujours croissant de données produites.

L'augmentation du nombre de points de mesures, que ce soit au niveau de l'activité numérique ou dans le monde physique, notamment avec la démocratisation de l'IoT, apporte de nombreuses contraintes mais également des opportunités de développement.

Pour débuter un projet Data, il faut garder en tête ces principes :
→ L'enjeu de la DSI est d'accompagner ses clients à découvrir quelles sont les données à exploiter et comment. L'expérimentation et l'itération avec le client sont indispensables.

→ Les infrastructures On Premises sont complexes et obsolètes by design.

→ Il n'existe pas un Big Data pour la DSI, mais «n» façons de consommer la Data.
Les projets métiers ont besoin de solutions simples et adaptées : on parle de "Data As a Service".

→ Le Big Data est un levier pour mettre la DSI au service d'une stratégie digitale et expérimenter sur de nouveaux modes d'exploration des données tel que le Machine Learning ou l'Intelligence Artificielle.

Que ce soit de l'exploitation directe, comme dans le cas de la facturation de publicité chez Adways, ou de la détection d'activités malveillantes, il est indispensable d'itérer sur une cible des KPI métiers à calculer/afficher en bénéficiant de retours terrain.

 # Le paiement à l'usage
 # Les faibles coûts de stockage et de sécurité
 # L'utilisation de services managés
 # La résilience by design
 # L'élasticité des ressources pour garantir les performances
 # L'automatisation de bout en bout

Ces éléments sont autant de raisons qui nous poussent à voir dans le Cloud **l'unique plateforme d'hébergement valable** pour les entreprises qui souhaitent concrétiser une stratégie Data.

// Vers une organisation DataDevOps ?

La stratégie Data d'une entreprise est capitale dans le monde digital d'aujourd'hui. Si elle a déjà été adoptée par les géants du numériques, nos clients ne savent pas toujours comment s'y prendre et ont besoin d'expérimenter.

→ La bonne méthodologie sera de traiter la Data avec le même processus que les applications en mettant en place l'agilité et un pipeline automatisé de CI/CD - Continuous Integration Continuous Delivery.
On parlera de DataDevOps.

GLOSSAIRE

Les services Cloud orientés Data

// STOCKAGE ET BASE DE DONNÉES :

Amazon Simple Storage Service (S3) : est un service de stockage objet permettant de stocker toute sorte de données sur le Cloud AWS et ce de manière illimitée et à moindre coût.

Amazon S3 propose plusieurs classes de stockage adaptées à chaque pattern d'accès:

→ **Classe Amazon S3 Standard :** meilleure offre en termes de durabilité (99,999999999%), de disponibilité (99,99%) et de redondance (la donnée est répliquée sur 3 zones différentes). Grâce à sa faible latence et son débit important, cette classe permet de répondre à une large variété de problématiques : les applications Cloud native, l'analyse de données, etc...

→ **Classe S3 Standard-Infrequent Access :** cette classe est prévue pour les accès moins fréquents tout en offrant la même durabilité qu'Amazon S3 Standard. Le coût de stockage pour cette classe est inférieur à la classe standard, mais l'accès aux données induit des frais supplémentaires. Cette classe est souvent utilisée pour stocker des backups.

→ **Amazon Glacier :** cette classe extrêmement durable et à très bas coût de stockage est principalement destinée à l'archivage de données. Contrairement aux autres classes, l'accès aux données n'est pas instantané et prend des minutes ou même des heures.

L'utilisation d'Amazon S3 comme point central dans le Data Lake d'entreprise est hautement conseillé par AWS car il s'intègre parfaitement avec le reste de services orientés Data.

Amazon RDS (Relational Database Service) : permet de créer des bases de données à la demande. Amazon RDS supporte plusieurs moteurs de base données (Amazon Aurora, MySQL, MariaDB, Oracle, SQL Server et PostgreSQL database engines).
Amazon Aurora est compatible MySQL ou PosgreSQL et est 5 fois plus performante que la version MySQL Open Source. Amazon Aurora est également disponible en version Serverless.

Amazon RDS gère automatiquement les aspects de sécurité, de backup et de mises à jour. De plus, Amazon RDS permet des créer des bases de données répliquées de manière synchrone sur plusieurs zones de disponibilité, ou des read replicas asynchrones. L'utilisation d'Amazon RDS nous permet donc de gagner en temps et effort en termes d'exploitation.

Amazon ElastiCache : service managé permettant de déployer des Data Stores en mémoire reposant sur Redis ou Memcached sous forme de service managé.

Amazon Redshift : service managé permettant de monter son propre Data Warehouse pour des utilisations telles que la BI (Business Intelligence).
En interne, Amazon Redshift stocke les données dans un format orienté colonnes, permettant ainsi de diminuer radicalement la durée d'exécution des requêtes.
Redshift Spectrum, qui est une fonctionnalité d'Amazon Redshift, peut être utilisé pour faire des requêtes SQL sur des données stockées dans Amazon S3.

// TRAITEMENT ET ANALYSE DES DONNÉES :

Amazon EMR (Elastic Map Reduce) : service semi-managé, permettant de monter un cluster Hadoop avec différents services de la stack Hadoop tels que Spark, Presto, Tez, HBase, etc. Il est très utile pour monter un cluster éphémère à la demande, effectuer certains traitements, renvoyer le résultat et terminer le cluster.
Amazon EMR permet un scale up/down automatique selon des métriques Cloud Watch prédéfinies ou personnalisées.

Amazon Elasticsearch : moteur de recherche open source ElasticSearch, déployé et managé par AWS. Il a l'avantage d'inclure Kibana et d'offrir des intégrations avec Logstash, pour de la visualisation et du Data Processing.

Amazon Athena : service AWS basé sur le moteur de traitement Presto. Il permet de faire des requêtes SQL rapides sur les données présentes sur Amazon S3. Pour cela, il se base sur les catalogues de données internes ou le catalogue AWS Glue.

AWS Glue : service managé d'AWS permettant de faire de l'ETL (Extract, Transform, Load) sans gestion de serveurs.
AWS Glue propose la notion de crawling qui permet d'inférer automatiquement la structure de données depuis plusieurs sources homogènes ou hétérogènes pour ensuite alimenter le catalogue avec les métadonnées inférées.
Le catalogue AWS Glue peut ensuite être utilisé sur d'autres technologies de requêtage comme Hive, Spark SQL, etc...
AWS Glue peut également générer le code permettant la transformation de la donnée ou appliquer la logique fournie par l'utilisateur.
AWS Glue permet des gains importants en termes d'exploitation (pas de serveurs à gérer) et en termes de développement (la logique de transformation est générée automatiquement).

Amazon Kinesis : similaire à Kafka, ce service permet la collecte, le traitement et l'analyse des données en temps réel. Il est capable de recevoir un débit très important de données de plusieurs sources différentes et peut s'intégrer avec plusieurs destinations en même temps.

Amazon Kinesis propose plusieurs sous-services spécifiques à des cas d'utilisations différents :

→ Amazon Kinesis Data Stream : permet d'ingérer des téraoctets de données par heure. Les données ingérées sont mises à disposition sous forme de streams pour les applications consommatrices.

→ Amazon Kinesis Data Firehose : ce service permet d'ingérer des données reçues en temps réel et de les insérer automatiquement dans différents Data Stores (Amazon S3, Amazon Redshift, ElasticSearch).

→ Amazon Kinesis Data Analytics : similaire à KSQL, il permet de traiter les données en temps réel avec des requêtes SQL.

→ Amazon Kinesis Video Streams : ce service permet d'ingérer, de chiffrer et d'indexer des données vidéos.

→ Amazon Kinesis Client Library (KCL) permet d'utiliser et de traiter des données depuis un Kinesis Data Stream. Elle s'occupe d'un grand nombre de tâches complexes associées à l'informatique distribuée, telles que l'équilibrage de charge entre plusieurs instances, la réponse aux pannes d'instance, les points de contrôle des enregistrements traités et la réaction au repartitionnement.

// BUSINESS INTELLIGENCE :

Amazon QuickSight : solution AWS permettant à partir de différentes sources de données (Amazon Athena, Amazon Redshift, Redshift Spectrum), de faire de la visualisation, des analyses et autres actions liées à la BI.

// MACHINE LEARNING ET ARTIFICIAL INTELLIGENCE :

Amazon Comprehend : service de traitement du langage naturel, identifie la langue du texte, extrait les phrases, lieux, personnes, marques ou événements clés, comprend la nature positive ou négative du texte, analyse le texte à l'aide de jetons et de parties du discours (PoS).

Amazon Lex : chatbot (robot conversationnel).

Amazon Polly : text to speech (conversion texte vers voix).

Amazon Rekognition : service de reconnaissance visuelle, permettant de reconnaître des éléments au sein d'une image ou d'une vidéo.

Amazon SageMaker : couvre l'ensemble du cycle de vie du modèle de machine learning, permettant de concevoir, entraîner, puis déployer rapidement et facilement un modèle en production. Il propose à ce jour 14 algorithmes intégrés parmi lesquels Linear Learner, XGBoost, seq2seq, K-Means, etc. Il est également possible d'utiliser ses propres algorithmes ou d'importer et d'exporter des modèles.

Amazon Transcribe : speech to text (conversion voix vers texte).

Amazon Translate : service de traduction.

// INTERNET OF THINGS (IOT) :

AWS IoT et Greengrass : services AWS permettant aux objets connectés d'exécuter des fonctions AWS Lambda, de communiquer entre eux et également avec d'autres services AWS.

Apache Parquet : format de stockage en colonnes présentant les caractéristiques suivantes :

→ Permet un stockage efficace des données en colonnes par rapport aux fichiers en ligne tels que le format CSV.

→ Conçu à partir de zéro avec des structures de données imbriquées complexes en tête.

→ Prend en charge des schémas de compression et d'encodage très efficaces.

→ Réduit les coûts de stockage des données et optimise l'efficacité des requêtes de données avec des technologies Serverless telles qu'Amazon Athena, Redshift Spectrum et Google Dataproc.

Apache Parquet est un format de données auto-descriptif qui intègre le schéma ou la structure dans les données elles-mêmes. Cela se traduit par un fichier optimisé pour les performances des requêtes et la réduction des E/S. Apache Parquet prend également en charge efficacement des schémas de compression et d'encodage.

Suite à la lecture de cet e-Book, contactez-nous pour :

- Continuer d'échanger : **marie.vanhaecke@d2-si.eu**
- Être accompagné sur un projet : **maxime.rivals@d2-si.eu**
- Suivre une formation appropriée : **sarah.foubert@d2-si.eu**
- Rejoindre notre équipe Big Data : **olivia.blanchon@d2-si.eu**

Co-rédigé par nos consultants techniques :

Selim Chergui
Zied Ezzar
Romain Gros
Adrien Houllier
Jonathan Ki
Sathiya Kumar
Maël Louvet
Othmane Nahyl
Alex Palesandro
Nicolas Peray
Damien Thauvin

Piloté par :

Maria Betbeze Dufrenoy - Alliance Manager
Julien Lemarchal - Journaliste
Marie Van Haecke - Social & Innovation Transformer

Mise en page :

Sandrine Anberrée - Graphiste

CLOUD IS THE NEW PLAY -GROUND

aws | D2SI by devoteam