



DATA LAKE INDUSTRIEL

**COMMENT
PASSER
LE CAP DU
DÉMONSTRATEUR ?**

CGI

La force de l'engagement^{MD}

SOMMAIRE

Q	Avant-propos.....	3
Q	Industrialiser : quels enjeux ?.....	7
Q	La démarche	10
Q	Urbanisation & gouvernance	12
Q	Les bonnes pratiques industrielles	15
Q	Industrialiser la sécurité.....	19
Q	Mise en place d'un Data Lake chez un retailer dans le luxe.....	22

POURQUOI L'INDUSTRIALISATION EST UN FACTEUR CLÉ DE SUCCÈS DES DATA LAKE ?

Toutes vos données ont de la valeur. Pas seulement celles structurées et déjà exposées à vos usages analytiques, mais aussi et surtout celles qui échappent à votre contrôle et à des exploitations innovantes. Quelles sont ces données ? On parle ici des données brutes - bureautique, messagerie, images, vidéos, etc. Ces informations non-structurées possèdent une valeur qui ne doit plus être négligée, notamment lorsqu'elles sont croisées avec des informations issues de vos référentiels de données d'entreprise.

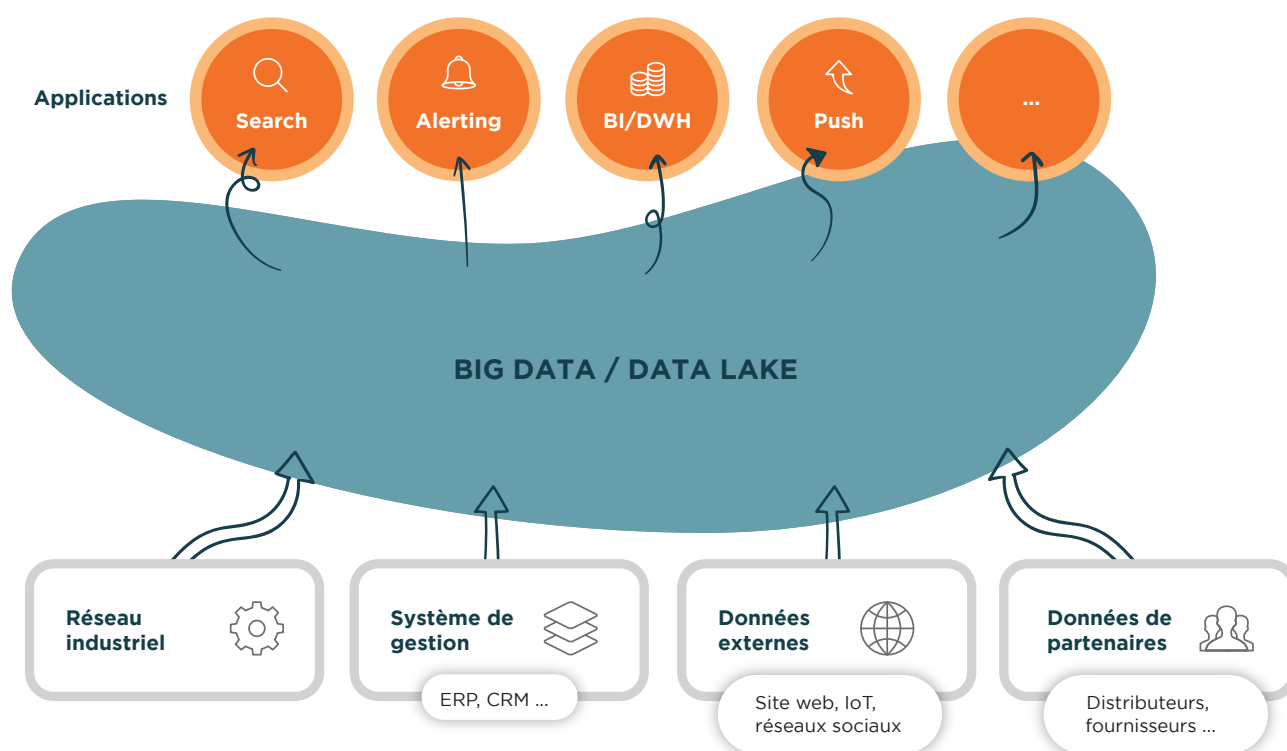
C'est la raison pour laquelle, en 2010, sont apparus les Data Lake (ou lacs de données en français). Le Chief Technical Officer (CTO) de Pentaho*, James Dixon, définissait à l'époque le Data Lake comme « la somme des données à usage analytique présentes dans l'entreprise ». Le Data Lake constitue ainsi une opportunité pour l'entreprise de devenir réellement Data centric. Grâce à de nouvelles technologies de stockage et de traitement, les organisations vont pouvoir positionner l'ensemble de leurs données, de provenance interne ou externe, au centre de leur système d'information pour produire et diffuser des informations stratégiques vers l'ensemble de leurs applications à partir de données issues de multiples sources.

L'approche privilégiée consiste à n'ajouter de nouvelles données qu'en fonction des besoins identifiés pour les usages. Néanmoins, la pratique montre qu'une approche plus volontariste se dessine sur les projets Data Lake à plus ou moins long terme pour ajouter des données en grand nombre sans avoir identifié d'usage *a priori*, ce qui peut engendrer de la dette technique *a posteriori*. Le développement de projets Data Lake s'est avéré particulièrement dynamique ces dernières années, même si certains acteurs ont connu des échecs. Avec le recul, les causes de ces échecs ont pu être identifiées. Chez CGI, nous sommes convaincus que l'industrialisation permet d'assurer le succès d'un Data Lake.

*Pentaho est éditeur d'une suite logicielle pour le décisionnel

Data Lake versus Data Warehouse

Le Data Lake constitue une brique complémentaire du Data Warehouse. L'objectif principal de ce dernier est de produire des indicateurs opérationnels et stratégiques pour l'entreprise, avec parfois quelques incursions dans le prédictif, afin d'aider à la prise de décision. Le Data Lake se distingue par la diversité et la complétude des données déversées, avec pour objectif de les utiliser in situ, mais également de les préparer à la diffusion vers d'autres briques applicatives. Le Data Warehouse se positionne ainsi comme un des nombreux consommateurs du Data Lake, et conserve les objectifs pour lesquels il a été construit.



Malgré cette cohabitation clairement définie, la frontière entre ces deux zones du système d'information devient parfois ténue lorsque vient le choix de basculer des usages portés par le Data Warehouse vers les serveurs et les technologies exploités par le Data Lake. Pour autant, le bien-fondé des systèmes d'information décisionnels et des Data Warehouse n'est pas remis en cause, et demeure au cœur du pilotage des entreprises.

Un espace de liberté dans les Data Lab

Quelles sont les différentes manières d'exploiter les potentialités du Data Lake et ainsi générer de la valeur à partir de la totalité des données présentes dans l'entreprise ?

Chez CGI, nous croyons aux Data Lab, ces espaces de travail agiles au sein desquels vos analystes vont consulter, explorer et mettre en corrélation les données :

- Soit par requêtage simple à travers l'utilisation d'interfaces natives telles qu'Ambari Views ou Hue,
- Soit par construction de modèles à travers des outils tels que Zeppelin et Anaconda,
- Soit via des outils de dataviz et d'exploration connectés aux espaces exposés comme Tableau, Qlik, ou Microstrategy.

Grâce à cet accès facilité aux données, les utilisateurs peuvent alors générer des modèles prédictifs en mode machine/deep learning. Des cas d'usage métiers plus innovants sont envisageables, les expérimentations n'étant désormais plus nécessairement contraintes par la mise en œuvre d'un nouveau projet à chaque nouveau besoin. La gestion efficace de ces espaces nécessite toutefois une réflexion en amont pour industrialiser l'offre de service afin de raccourcir le délai de réponse à une nouvelle demande, faciliter l'exploitation, et garantir la sécurité des données.

A quoi ressemble le cycle de vie d'un espace Data Lab ? La première étape consiste à attribuer des droits d'accès aux données en créant les groupes et rôles adaptés. La deuxième réside dans l'attribution des ressources de traitement afin de sélectionner, traiter, qualifier, et définir les données. Il faut être vigilant quant à la consommation des ressources par un groupe afin de ne pas impacter négativement les autres utilisateurs, voire l'alimentation globale du Data Lake. Enfin, la troisième étape revient à créer un espace de stockage dédié au Data Lab disposant des sécurités, des données et de l'outillage requis. Cette approche permet d'isoler un usage et ne pas surcharger le reste du cluster. L'automatisation partielle ou complète de l'ensemble de ces étapes est un gage de réactivité lors d'une nouvelle demande de Data Lab.

De l'intérêt d'industrialiser le Data Lake...

Pour assurer les services techniques et fonctionnels relatifs à l'alimentation et l'exposition des données, il est nécessaire de mettre en place très tôt des processus clairs et efficaces autour des contraintes réglementaires (comme le respect de la protection des données personnelles issu du règlement général sur la protection des données – RGPD), des enjeux de gouvernance et de gestion du cycle de vie des données.

C'est pourquoi nous sommes convaincus que la bonne approche pour réussir un Data Lake revient à industrialiser sa mise en place, à savoir formaliser une démarche et une organisation spécifiques à même d'anticiper les chantiers structurants pour poser les bases de la réussite.

Cadrer le projet, entre contraintes technologiques et gouvernance

Les premières briques de l'industrialisation du Data Lake obéissent à des contraintes spécifiques. Nous avons identifié cinq grands enjeux techniques :

- Un socle technique, suffisamment flexible pour évoluer en fonction des orientations de la feuille de route,
- Un outillage d'alimentation, porté par un framework afin d'industrialiser les développements,
- Des processus d'exécution, capables d'activer les composants du socle technique et de l'outillage d'alimentation pour réussir les chaînes de traitement et d'exposition des données,
- Des processus de fabrication, pour s'intégrer au mieux aux socles de développement existants et à venir,

- La gestion des environnements, pour déployer rapidement les composants sur des infrastructures correctement dimensionnées.

Évidemment, la question de la gouvernance est aussi capitale. Dans de nombreux projets Data, nous remarquons que les principales causes d'échec se situent autour de cette problématique de gouvernance. Clairement, il s'agit d'anticiper au maximum l'organisation autour de la donnée avant de s'engager sur le Data Lake. Une gouvernance bien en place vous permet de connaître la composition de vos données, de les retrouver et d'en garantir la qualité à vos consommateurs, tout en assurant la sécurisation requise.

Chez CGI, nous avons identifié quatre grandes phases au cours desquelles, à travers une urbanisation par couches, la gouvernance des données peut être assurée :

- La phase d'ingestion des données brutes,
- Le raffinement pour aboutir à des données préparées,
- L'enrichissement pour créer des données compréhensibles et partagées,
- Et l'exposition pour fournir des données d'usage.

Un guide pratique pour vos équipes

Le Data Lake s'impose progressivement comme l'espace nouvelle génération pour des pratiques analytiques avancées, originales, innovantes. Afin de générer de nouveaux cas d'usage métiers, de comprendre plus finement vos consommateurs, d'identifier les possibles optimisations de processus, votre entreprise a tout intérêt à se lancer dans une approche Data qui ne se limite plus à des référentiels établis (et finalement utilisés par tous), mais ouvre les horizons et supporte réellement votre transformation.

Pour vous permettre d'exploiter réellement toutes vos données, nous avons souhaité vous présenter les fondamentaux de l'industrialisation du Data Lake qui symbolisent,

à notre sens, la condition de la réussite d'un tel projet. Dans ce guide, vos équipes IT pourront en apprendre davantage sur les enjeux de l'industrialisation, notamment d'un point de vue organisationnel, identifier la démarche et les prérequis techniques à ne pas négliger, les bonnes pratiques de mise en œuvre et de sécurité, et nos préconisations en matière de gouvernance. Nous avons également souhaité partager un témoignage d'experts CGI relatif à leur expérience sur le projet d'un de nos clients, issu du secteur Retail.

Bonne lecture.

Vincent AMBROISE



INDUSTRIALISER : QUELS ENJEUX ?

Quel que soit le secteur de marché, la concurrence entre nouveaux acteurs « Digital native » et acteurs historiques a accentué la nécessité de définir une stratégie digitale, seule solution pour rester dans la course... Et tirer profit des transformations en cours. La Data est au cœur de cette dynamique. Elle représente l'or noir de la révolution digitale et l'énergie nécessaire à la croissance au XXI^{ème} siècle. Son exploitation doit forcément être industrialisée.

CGI interroge chaque année ses clients partout dans le monde et analyse leurs réponses à la fois au niveau géographique et sectoriel. Cette analyse permet à CGI de se positionner en fonction des enjeux clés exprimés par ses clients. En 2018, toutes géographies et tous secteurs confondus, les priorités identifiées étaient les suivantes :

1. Augmenter l'efficacité opérationnelle,
2. Définir de nouveaux services,
3. Assurer la mise en conformité réglementaire.

Le Data Lake, en tant que réservoir et concentrateur des données de l'entreprise, est devenu le standard de facto pour répondre à ces enjeux. Il rassemble à la fois tous les espoirs et toutes les craintes. Les espoirs viennent de ceux qui ne l'ont pas encore mis en place ou qui ont réussi l'expérience. Les craintes, de ceux qui se posent des questions liées au coût de mise en place, à la sécurisation des données, ou tout simplement les craintes de ceux qui associent des échecs d'expériences vécues au manque de maturité des technolo-



L'industrialisation est souvent le parent pauvre des projets informatiques du fait des coûts et des délais additionnels engagés, parfois sans assurance de gains à venir. Pour mesurer le retour sur investissement, il convient déjà de bien saisir les enjeux.

gies au moment de la réalisation de ces projets, ou du manque d'expérience des équipes internes lors de leur mise en œuvre.

Pour éviter ces échecs et lever les freins au sein des équipes métiers et des directions opérationnelles, le maître mot est l'industrialisation du Data Lake. Industrialiser, c'est améliorer de manière durable et mesurable la performance. C'est la garantie que toutes les ambitions, tous les projets pourront dépasser le stade des démonstrateurs et soutenir durablement la croissance.

LES ENJEUX ORGANISATIONNELS

Les enjeux de la transformation digitale obligent les organisations à faire évoluer leurs systèmes informatiques en les architecturant autour d'une approche Data centric pour casser les silos et démocratiser l'accès à l'information à leurs collaborateurs, clients et partenaires. Cette démarche doit s'accompagner d'une politique de sécurité d'accès aux données. Rendre disponible l'ensemble des données que possède l'entreprise, les intégrer et les regrouper, qu'elles soient internes, externes, venant des réseaux sociaux, ou de l'Open Data. A l'heure actuelle, seul un faible pourcentage de Data des entreprises est valorisé. Néanmoins, les potentiels sont énormes, et tous les métiers sont concernés.

Le Data Lake est la technologie clé de cette transformation. Au-delà des aspects techniques, sa mise en œuvre nécessite une prise de conscience de l'importance de la Data à tous les niveaux de l'entreprise. L'organisation, les métiers, la gouvernance et les méthodologies doivent évoluer en même temps que le projet avance. Il s'agit en effet de revoir en profondeur, à travers le cheminement des flux de données et de leur exploitation, le rôle de chaque entité. Il s'agit

de casser les silos métiers, de redéfinir la frontière entre métiers et IT devenue floue. Au sein même de la direction informatique, il s'agit de revoir l'organisation entre les équipes exploitantes et les équipes d'étude et de développement.

Cela signifie aussi une transformation majeure des compétences. Le Data Lake s'appuie sur des technologies Big Data encore insuffisamment présentes dans les entreprises.

La jeunesse de ces technologies, leur manque de maturité nécessitent un surcroît d'expertise pour compenser le manque de recul.

Sans mener ces transformations profondes, le Data Lake restera un environnement de bac à sable, et ne pourra pas accéder à sa version industrielle, la seule à même de supporter les transformations majeures de l'entreprise numérique.

LES ENJEUX TECHNOLOGIQUES

La mise en place d'un Data Lake pour collecter toutes les données utiles, internes ou externes n'est plus l'exception mais la règle. Le potentiel de valorisation de données a été intégré par la plupart des organisations, et fait maintenant partie intégrante de la stratégie Data des entreprises.

L'ère du Proof of Concept (POC) est à présent révolue. Les tests à petite ou grande échelle ont démontré l'intérêt de ces nouveaux écosystèmes.

L'enjeu est désormais de faire du Data Lake le socle d'une architecture Data centric, fiable, performante, transverse et générique à même de supporter toutes les attentes liées à la transformation digitale et aux besoins croissants en matière d'analyse de la donnée.

Pour atteindre cet objectif, le socle technologique se doit de répondre à des caractéristiques majeures, telles qu'une construction multi-tenant, l'utilisation de composants pérennes, stables et ouverts, et l'intégration aux standards de l'entreprise.

C'est dans cette perspective que les choix d'architecture et du socle doivent se faire, et ce guide technique a pour objectif de vous permettre de répondre à vos questions.

LES ENJEUX POUR DEMAIN

Le Data Warehouse a suivi l'évolution de la Business Intelligence (BI) en adaptant les bases de données existantes, majoritairement relationnelles (Oracle, DB2, Teradata, etc.), leurs structures de stockage, de requêtage (SQL), leurs performances et optimisations, au reporting et divers besoins de pilotage.

Demain, le Data Lake doit jouer le même rôle que les bases de données d'hier pour l'intégration et l'utilisation de la Data Science et le développement de l'Intelligence Artificielle. Il s'agit donc de construire un socle ouvert aux évolutions des usages et des technologies futures, et de mettre en œuvre, au-delà du

socle lui-même, un outillage complet et un ensemble de frameworks, normes et procédures réutilisables qui garantiront sa pérennité et son évolutivité.

Pour jouer pleinement son rôle de concentrateur (hub), et permettre la diffusion de données vers l'ensemble des applications du système d'information de l'entreprise, le Data Lake se doit également d'exposer des interfaces applicatives (API) et se doter d'outils de transfert et de transformation de données (ETL/ELT) propres à fournir les services attendus par les fournisseurs et les consommateurs de données.



LA DÉMARCHE

Le passage à l'échelle industrielle s'entreprennd par la formalisation d'une démarche et d'une organisation orientées vers l'industrialisation, à même de maximiser la performance des réalisations pour en tirer les bénéfices attendus. Pour répondre à cet enjeu, l'appréhension de l'écosystème du SI et l'anticipation des chantiers techniques sont des facteurs clés de succès de la mise en place d'un Data Lake.

La démarche d'industrialisation vient supporter la volonté de lancer un projet Data Lake à grande échelle pour accueillir de nouveaux cas d'usage et satisfaire à la construction Data centric du SI. La mise à l'échelle industrielle ne peut s'appuyer uniquement sur la capitalisation issue de la phase de prototype. Ce changement de paradigme nécessite l'activation des moyens techniques, humains et méthodologiques requis.

La stratégie d'industrialisation s'appuie sur une démarche composée d'éléments fonctionnels, techniques, méthodologiques et organisationnels. Sans cette approche globale, l'industrialisation a peu de chance d'adresser les objectifs fixés.

Une industrialisation solide adresse les caractéristiques techniques intrinsèques des projets Data Lake, à savoir :

- L'assemblage de nombreux composants logiciels à évolution rapide,
- Le déploiement sur d'imposants clusters de serveurs,
- La présence de nombreuses données variées et volumineuses portant des aspects réglementaires à prendre en compte,
- L'architecture multi-tenant impliquant une gestion complexe des sécurités d'accès,
- La construction d'espaces de travail en mode Data Lab sur de courts délais.

Dans ce contexte, une phase de cadrage doit ébaucher les premières briques indispensables de l'industrialisation :

- **Le socle technique** : le choix des composants logiciels du socle technologique parmi une distribution présélectionnée (Horton works, Cloudera) ou par choix de composants spécifiques hors distribution (« Best-of-Breed »), en s'assurant de leur pérennité et des évolutions à prévoir dans le cadre d'une feuille de route.
- **Un outillage d'alimentation** : la mise en place des premières fonctionnalités d'un framework, construit pour évoluer en fonction des nouveaux besoins identifiés sur la base de codes spécifiques ou de briques logicielles disponibles sur le marché.
- **Des processus de fabrication** : l'industrialisation des processus de fabrication articule les développements à produire avec l'outillage en place, particulièrement pour la gestion des sources, l'automatisation des déploiements et des tests.
- **Des processus d'exécution** : la description des processus d'exécution met en musique les composants du socle technique et de l'outillage d'alimentation pour élaborer les chaînes de traitement, en particulier depuis les producteurs de données jusqu'au Data Lake. Ces processus tiennent compte de l'écosystème du SI, et une vigilance particulière est à porter concernant l'outillage d'ordonnancement et de supervision.
- **La gestion des environnements** : une gestion industrielle des environnements doit permettre de déployer rapidement des socles technologiques sur des infrastructures correctement dimensionnées, et dans des configurations qualifiées et maîtrisées. L'industrialisation via un hébergement en mode Cloud est également à considérer si elle est envisageable dans le contexte. Une approche DevOps contribue également, à travers l'organisation et l'outillage, à faciliter la livraison continue, limitant ainsi « l'effet tunnel » et diminuant le délai de mise à disposition.

Une fois ces premières briques industrielles posées, chacune doit faire l'objet d'une feuille de route, en cohérence avec les besoins exprimés pour les cas d'usage, et maintenue tout au long du projet. L'adéquation des composants industrialisés avec les cas d'usage est renforcée par l'adoption d'une méthodologie agile qui ne cloisonne pas la communication entre les responsables de produits techniques et fonctionnels, et favorise une organisation d'équipes constituées de compétences diversifiées.

URBANISATION & GOUVERNANCE

Un Data Lake industriel ne peut s'abreuver des rivières de données sans maîtrise. L'urbanisation et la gouvernance sont la garantie de pouvoir cartographier les données, d'en connaître la composition et d'en assurer la qualité aux consommateurs. Au même titre que les chantiers techniques, l'instruction de ces sujets s'entreprind dès le démarrage du projet.

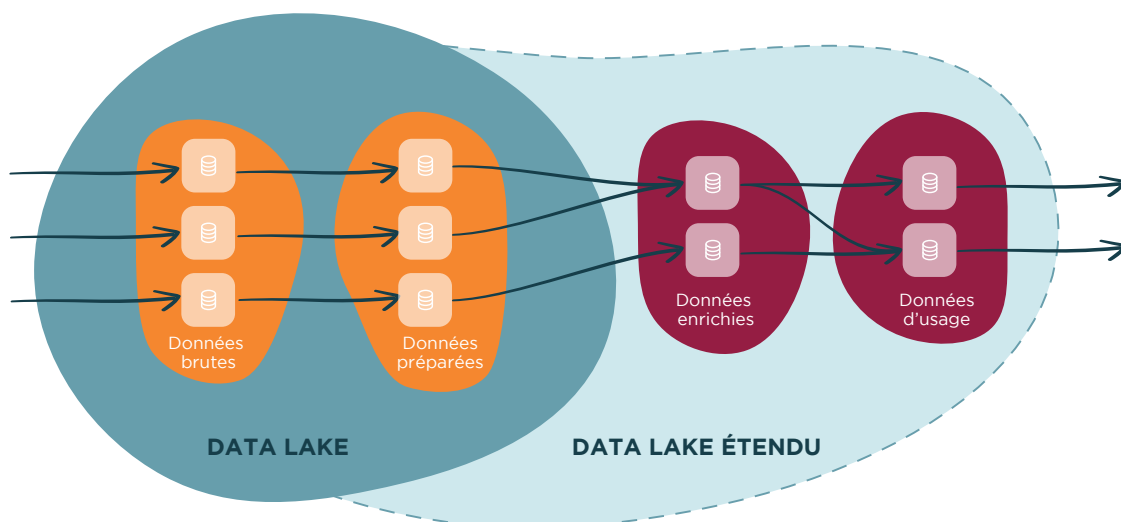
URBANISATION PAR COUCHES

Dans notre définition, le Data Lake n'est pas un lieu de consommation privilégié de données.

Les données consommables prennent place dans un espace logique distinct, **le Data Lake étendu**, afin :

- D'être capable de distinguer les données brutes de celles qui ont été travaillées, et pouvoir transformer les mêmes données source pour plusieurs usages différents.
- D'autoriser l'usage d'autres technologies répondant au mieux au besoin métier.

Déverser un océan de données de nature variée signifie que vous devez mettre en place une organisation rigoureuse de rangement et de classification, sinon la noyade est garantie.



L'alimentation des Data Lake et Data Lake étendu se matérialise à travers une succession d'étapes :

- **Ingestion des données brutes** : aucune transformation n'est réalisée vis-à-vis des données fournies par le producteur,
- **Raffinement vers des données préparées** : transformation légère des données brutes pour avoir une cohérence d'ensemble, par exemple en normalisant les formats de dates,
- **Enrichissement des données** : issues ou non de croisements de données préparées et de calculs sur la base de règles métiers, les données enrichies sont construites et mutualisables pour les usages,
- **Exposition des données d'usage** : nous recommandons de distinguer la zone de l'enrichissement de la zone d'exposition des données dans des espaces dédiés aux usages afin d'optimiser la mutualisation des objets enrichis et d'adresser des contraintes de sécurité spécifiques. Ces données peuvent être hébergées sur des supports technologiques hors Data Lake.

GOVERNANCE DES DONNEES

Les zones de l'urbanisation étant posées avec les règles associées, l'objectif est désormais d'organiser et de gouverner les données. Car si ingérer des données dans un Data Lake est le premier objectif, l'enjeu est surtout de savoir gérer le cycle de vie de la donnée dans la durée pour en garantir la qualité et le suivi.

La qualité de données, une des clés de la réussite du Data Lake

La vraie difficulté d'un Data Lake réside dans la façon dont il est construit. La tentation a été forte durant les premiers temps de décider de descendre sans discernement l'ensemble des données d'un périmètre sur la simple demande d'une direction. La justification principale étant la peur du manque...

Or, la stratégie d'envoyer tous les flux de données et de réfléchir *a posteriori* à leur usage s'avère totalement contre-productive, voire même dangereuse. Si on envoie tout et n'importe quoi dans un Data Lake, ce dernier risque de se transformer très rapidement en marais de données (Data Swamp), en perdant au passage toute crédibilité auprès des différents consommateurs.

Une donnée n'a de valeur que si les personnes qui l'exploitent sont capables de la remettre

dans son contexte de production puis de traitement.

Elle ne pourra être exploitée à grande échelle, de façon industrielle, que si les personnes prennent la responsabilité d'en assurer la fiabilité. Ces mêmes personnes devront pouvoir traiter cette non-qualité, à chaud de façon curative, et à froid de façon préventive. Ces actions préventives pouvant aller jusqu'à l'adaptation de certains processus, voire uniquement un « rappel à la loi » des personnes chargées de la saisie de ces informations.

Enfin, la donnée repose sur un langage d'entreprise avec des référentiels qui nécessitent, pour être généralisés, l'appropriation par toute l'entreprise de la culture Data driven. Cela suppose de s'affranchir des silos traditionnels (marketing, production, supply chain, finance, etc.).

Le changement de mindset est indissociable d'un accompagnement au changement

Pour une gouvernance efficace, l'outillage technologique est loin d'être la problématique principale. Cette nouvelle culture Data se confronte à de nombreux freins organisationnels.



La présence d'un sponsoring fort est donc un prérequis. Il peut prendre la forme de la nomination d'un Chief Data Officer (CDO), membre du comité de direction. C'est lui qui sera en charge d'assurer la définition et la mise en place du cadre de gouvernance des données à l'échelle de l'entreprise. Il proposera et pilotera les projets stratégiques associés. Un travail en étroite collaboration avec les sponsors métiers et l'IT lui permettra de prioriser et résoudre les problématiques rencontrées dans le cadre de ses activités.

Le CDO pourra s'appuyer en cela sur :

- Un Data Manager par domaine fonctionnel qui est le garant de la collecte des données sur son périmètre, en s'assurant de leur exploitabilité au service des usages qu'il aura préalablement identifiés. Au même titre que le CDO, il est la déclinaison stratégique de la gouvernance.
- Un Data Owner qui est le référent métier au sein de l'entreprise sur un périmètre identifié de données. Il travaille donc à ce titre en étroite collaboration avec le Data Manager correspondant. Il est le garant des définitions métiers et des référentiels associés. Il porte la responsabilité de la qualité de la donnée car c'est lui qui en définit les règles. Enfin, il statue sur les autorisations d'accès à la donnée.
- Des Data Steward qui

portent la déclinaison opérationnelle de la gouvernance. Ils gèrent et administrent tous les aspects d'un sous-ensemble de données. Ils travaillent directement avec le Data Manager qui fixe le cap et portent la responsabilité des aspects de sécurité, de qualité et d'accessibilité des données.

Mais le CDO devra surtout pouvoir s'appuyer sur l'ensemble des collaborateurs de l'entreprise. Sa communication devra permettre que chacun se sente investi dans cette transformation.

La gouvernance ne doit donc pas s'appuyer sur un système pyramidal mais sur un modèle décentralisé avec une approche progressive et non-invasive basée sur une responsabilisation via les priorités.

LES BONNES PRATIQUES INDUSTRIELLES

La mise en place d'un modèle de normes et de bonnes pratiques est essentielle pour ajouter massivement de nouvelles données tout en conservant la maîtrise des temps de développement et de la qualité des informations délivrées. Ce chapitre livre quelques pratiques structurantes issues de nos projets, à appliquer au plus tôt.

FRAMEWORK D'ALIMENTATION

Contrats d'interface

Les contrats d'interface sont les entrants d'un framework d'alimentation et leur structuration est primordiale. Le modèle se construit de manière itérative en fonction de l'évolution des fonctionnalités de l'outillage d'alimentation.

Les contrats d'interface spécifient pour chaque source de donnée :

- Le format de la donnée brute et ses caractéristiques
- Le type de fichier ou de flux
- Le mode d'ingestion complet ou partiel, en temps réel ou différé
- Le format d'encodage et les délimiteurs
- La fréquence d'ingestion et de préparation
- Le partitionnement à appliquer
- Les règles de préparation
- Les règles d'anonymisation
- Le schéma des sources et de leurs datasets
- Les noms et types des colonnes
- Le schéma des fichiers ou messages structurés (XML, AVRO, JSON, YAML...)
- L'emplacement de la source avant ingestion (Protocole, URL, ...)

Certains choix techniques structurants orientent les réalisations d'un projet pour longtemps, avec de lourds travaux à la clé lorsque vient l'heure de les remettre en cause. D'où l'importance de formaliser des bonnes pratiques et de les suivre dès le cadrage du projet pour en assurer la réussite.

Ces contrats contribuent également à l'alimentation du référentiel de métadonnées, par exemple Atlas, en stockant les attributs complémentaires de l'interface via l'usage de tags.

Framework d'ingestion automatique

Le framework doit se déployer de façon automatisée, ainsi que les traitements d'ingestion et de préparation. Il doit également implémenter un mécanisme de reprise automatique pour pallier les retards liés à la disponibilité des sources en mode batch.

Métriques applicatives & qualité de la donnée

Le framework doit mettre en place des métriques d'alimentation. Ces informations sont utiles à la supervision et à la création d'une météo journalière à destination des exploitants et des utilisateurs. La qualité de service doit être visualisable dans des tableaux de bord donnant accès à des statistiques telles que la disponibilité de la plateforme ou la qualité de la donnée (volumes de rejet), en vision courante ou historique.

Librairies

Afin de systématiser et de normaliser la production des métadonnées, l'adoption de librairies communes aux équipes de développement est indispensable. Elles sont développées, maintenues, mais également utilisées par l'équipe responsable du socle technique en charge du framework.

L'adoption de ces composants communs passe par la mise en place de librairies suffisamment documentées, de squelettes de projets pour faciliter le démarrage et normaliser l'enveloppe. Cela nécessite aussi le partage d'exemples basés sur ces squelettes donnés en référence.

PLAN DE NOMMAGE

Un plan de nommage normalisé du système de fichiers garantit une facilité d'exploitation et un cloisonnement des accès aux données, crucial pour la gestion des sécurités. Il décrit l'arborescence des répertoires, leurs nommages et leurs contenus.

Le plan de nommage doit respecter l'urbanisation du Data Lake en reprenant dans ses racines son découpage en couches basé sur :

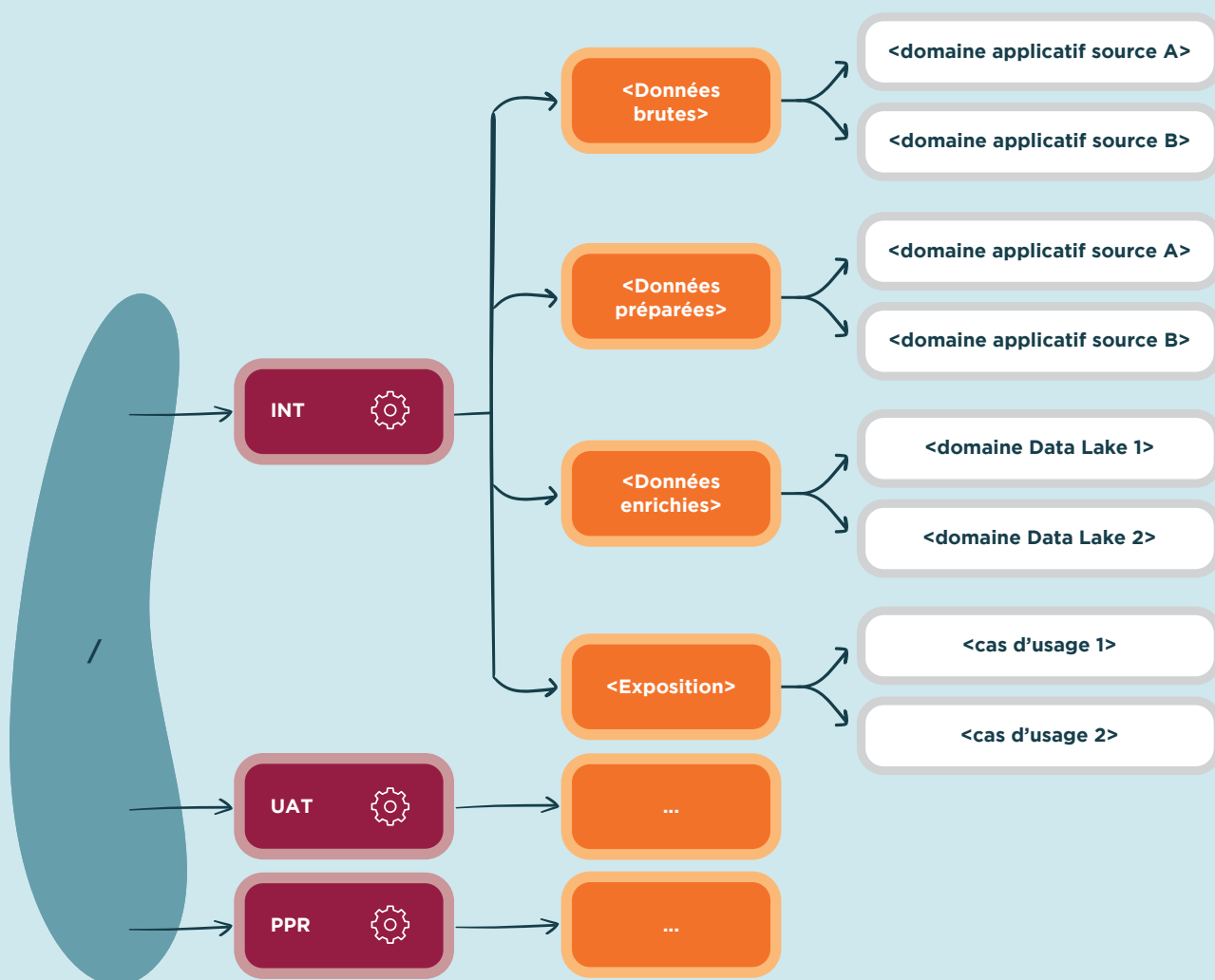
- L'urbanisation applicative des sources pour les données brutes et préparées,
- La définition d'une organisation des données enrichies par un urbaniste,
- Le nom des cas d'usage pour les données d'usage.

Pour veiller au cloisonnement des environnements, en particulier si plusieurs d'entre eux sont hébergés sur une même plateforme, le code environnement doit également être utilisé. Ce code environnement est variabilisé dans les scripts pour faciliter le déploiement

et garantir une cohérence entre les différents environnements. L'arborescence du plan de nommage est complétée par le code de l'application, du projet, ou du cas d'usage en fonction de la zone concernée du Data Lake.

Si plusieurs entités de la même entreprise se partagent un cluster, le code entité est un bon candidat au découpage des files de traitement (queues YARN). En étanchéifiant les ressources allouées, ce cloisonnement permet également une refacturation des ressources allouées ou consommées par entité.

Exemple de plan de nommage



PARTITIONNEMENT

L'ingestion doit normaliser le **format de stockage** des données. Si les données brutes peuvent arriver sous différents formats, il est primordial de standardiser le format de stockage des données ingérées et préparées, et que celui-ci soit choisi en fonction du mode d'ingestion et d'exposition des données.

Le **partitionnement** tire parti des fonctionnalités d'optimisation natives d'HDFS. Il améliore grandement les performances en limitant un retraitement systématique et massif des données.

Dans la chaîne d'alimentation, le partitionnement doit être mis en œuvre dès la source. L'outil d'ingestion des données brutes doit être le premier traitement à partitionner dans HDFS. Le partitionnement est ensuite hérité par les données préparées. L'enrichissement et l'exposition peuvent toutefois donner lieu à de l'agrégation de partitions en fonction des usages ciblés, par exemple en regroupant des données partitionnées au quart d'heure dans des partitions horaires.

Pour que le partitionnement soit efficace, encore faut-il que le choix des attributs de partitionnement soit judicieux. Ce choix est avant tout guidé par le mode d'ingestion, via l'acquisition d'une photo complète ou la capture des changements, la stratégie de

stockage avec ou sans historisation, et les types d'accès aux données pour les usages.

Dans la majorité des cas, les données portent sur un partitionnement temporel complété éventuellement par d'autres clés de partitionnement tels que des critères géographiques ou organisationnels.

Le partitionnement temporel se construit impérativement sur la base du « temps métier » de la donnée fournie par l'émetteur, et non pas une référence temporelle technique de traitement. Car si la date fonctionnelle de la donnée est immuable et apporte de la valeur métier, la date de traitement a uniquement une valeur d'exploitation et peut varier en fonction des aléas de fonctionnement de la production.

La simplicité de suppression d'une partition accélère également la gestion des purges de données. Le partitionnement doit donc également être pris en compte dans les stratégies de purge.



INDUSTRIALISER LA SÉCURITÉ

L'industrialisation de la sécurité garantit la rapidité et la robustesse des ouvertures de droits. Elle est la composante incontournable de la gestion fine des droits d'accès, et accompagne la prise en compte des contraintes de protection de la donnée (privacy by design).

L'industrialisation de la sécurité nécessite d'instruire les cinq piliers du domaine : **l'authentification, les habilitations, la traçabilité, le chiffrement, et l'anonymisation.**

Concernant **l'authentification**, le standard est la mise en place d'un service LDAP et de l'authentification par jeton Kerberos, pour augmenter le niveau de sécurisation et s'affranchir de la manipulation des mots de passe dans les scripts. En couplant les services LDAP avec un serveur de distribution de clés (Kerberos Distribution Center ou KDC), la gestion des comptes et de la stratégie de mots de passe associée devient centralisée.

Le stockage du mot de passe fourni par le KDC est assuré par un « keytab » dont les droits d'accès sont exclusivement donnés au compte propriétaire.

Concernant les comptes à définir, il convient de distinguer les comptes nominatifs, attribués à des personnes physiques, des comptes applicatifs qui se destinent aux traitements applicatifs exécutés sur les serveurs. Les comptes nominatifs se basent sur un code utilisateur unique alors que le compte applicatif est associé à un découpage en périmètres techniques et fonctionnels, généralement par code projet. La régénération des mots de passe des comptes applicatifs et le redéploiement des « keytabs » doivent être automatisés afin de prévenir les interruptions des traitements en cas de renouvellement des mots de passe.



Le Data Lake est souvent présenté à juste titre comme le pot de miel qui concentre en un même lieu toutes les données d'une entreprise. L'efficacité des procédures de sécurité est incontournable pour ne pas mettre en risque la valeur et la notoriété de l'entreprise.





Les **habilitations** définissent les privilèges d'accès aux ressources pour les groupes créés précédemment. Ces autorisations doivent être centralisées au maximum dans un même référentiel, par exemple Ranger ou Sentry. Dans certains cas, la définition des droits d'accès doit être complétée dans des référentiels séparés, par exemple pour Kafka et Nifi.

La mise en place d'un plan de nommage normalisé facilite grandement la gestion des règles d'autorisation. La base de l'industrialisation des habilitations est la création d'une matrice de sécurité, humainement lisible et exploitable, qui spécifie les habilitations des groupes ou des comptes sur les ressources du plan de nommage. L'automatisation de la mise en place des droits d'accès à partir de cette matrice facilite ensuite la gestion du cycle de vie des habilitations.

L'adjonction d'un processus et d'une organisation de gestion des comptes et des habilitations est également indispensable. Clarifiée au plus tôt, elle est le gage d'une bonne réactivité dans la prise en compte des changements et l'assurance de l'implémentation adéquate des règles d'accès.

La **traçabilité** répond au besoin de savoir qui a eu accès à quelle donnée et à quel moment. Cette notion de « qui » nécessite d'avoir une personne unique, et donc un compte nominatif, plus qu'un compte applicatif « générique » partagé par plusieurs traitements. Ces informations sont exploitées pour contrôler que les habilitations paramétrées sont effectives, vérifier les accès aux ressources à des fins d'audit, et tracer d'éventuelles manipulations de données pour exploiter la solution.

Le **chiffrement** vient en complément des habilitations pour se prémunir de la violation des règles d'accès aux données. La conservation des données en clair sur des dispositifs de stockage et leur transmission à travers des liens non-sécurisés laisse envisager la possibilité de les capturer par des processus de « bas niveau » inhérents aux socles techniques utilisés.

Le chiffrement des protocoles d'échange entre le Data Lake et l'extérieur est primordial. Les outillages d'ingestion, de visualisation, et les API exposées sont particulièrement concernés. Cependant, malgré le haut niveau de sécurisation généralement appliqué aux accès Data Lake, les échanges intra Data Lake et le chiffrement des données stockées ne doivent pas être négligés.

Concernant le stockage, l'ensemble des lieux de stockage doit être cartographié avec, en premier lieu, les zones persistantes telles que les Data Node, mais également les zones de stockage temporaires telles que les bus de messages Kafka.

Les mécanismes de chiffrement induisent la génération et le déploiement de certificats qui doivent être automatisés afin de limiter au maximum le risque d'interruption de service lors de leur expiration.

L'**anonymisation** appose la confidentialité aux données sensibles ciblées par des contraintes réglementaires comme le RGPD. Lorsque la réglementation interdit toute conservation de la donnée d'origine, l'anonymisation doit être prise en charge à la source dans le lieu de production, et l'information résultante n'est pas forcément communiquée au Data Lake. Lorsque les données brutes peuvent être conservées en clair, elles doivent être sanctuarisées et accessibles à un nombre restreint de personnes. Les règles d'anonymisation interviennent ensuite, lors de la diffusion dans les autres zones du Data Lake, afin de limiter les usages et la visibilité du contenu.

Les techniques d'anonymisation doivent être sélectionnées en fonction des contraintes adaptées à la donnée, réversibles ou non, et plus ou moins destructrices sur la capacité à visualiser et à lier les données.

L'identification et la mise en œuvre des techniques d'anonymisation est adressée dès le démarrage du projet pour éviter le risque de non-conformité à la réglementation et ne pas accumuler la dette technique associée à une correction tardive.

Les contrats d'interface identifient les colonnes concernées et indiquent la méthode d'anonymisation à utiliser par le framework d'alimentation.

MISE EN PLACE D'UN DATA LAKE CHEZ UN RETAILER DANS LE LUXE

Le cas pratique d'un retailer a été choisi ici pour illustrer la mise en place d'un Data Lake. Ce témoignage démontre bien les enjeux d'industrialisation de la mise à disposition de données massives dans un SI moderne.

La mise en place de ce Data Lake a été faite en plusieurs étapes.

Etape 1 : mise en place de l'architecture et de la sécurité du Data Lake

Le choix du socle du Data Lake s'est porté sur une distribution Hadoop Cloudera du fait de sa polyvalence de stockage (HDFS, IMPALA, Hbase). Les architectures Hadoop n'ont pas été pensées au départ pour être sécurisées. Néanmoins, les distributions intègrent aujourd'hui tous les outils pour avoir un fort niveau de sécurité.

Dans notre projet, nous avons activé les briques Active Directory et Kerberos afin de pouvoir finement administrer les accès.

Dans la dynamique du projet, voici les grandes étapes de mise en place du socle technique :

- Installation du socle technique et des composants Hadoop Cloudera,
- Sécurisation des accès périmétriques avec le lien Active Directory et le serveur Kerberos,
- Définition de la gouvernance des données et de leur sécurisation dans le Data Lake.

Etape 2 : mise en place de l'architecture de fabrication et d'exécution

Cette étape est majeure pour la bonne marche des projets, du développement à la mise en production. Il faut impérativement mettre en place un système PIC (Plateforme d'Intégration

Continue). En effet, les systèmes Big Data sont massivement distribués. Cela peut représenter plusieurs dizaines, voire centaines de serveurs à administrer lors des déploiements.

Pour la mise en place des outils, le client ayant choisi TALEND, nous avons mis en place la pile complète de composants pour le versionning (SVN), la compilation (Nexus Server) et le déploiement à l'aide de l'outil Ansible.

Cette étape doit être particulièrement soignée afin de gagner en productivité par la suite. Des playbooks Ansible de déploiement et de contrôle doivent être développés afin d'être prêts lors du premier développement. En effet, au démarrage d'un projet, l'outillage doit permettre de créer tous les éléments de bout en bout.

Les espaces de fabrication et de livraison ont été créés en réalisant les actions de création :

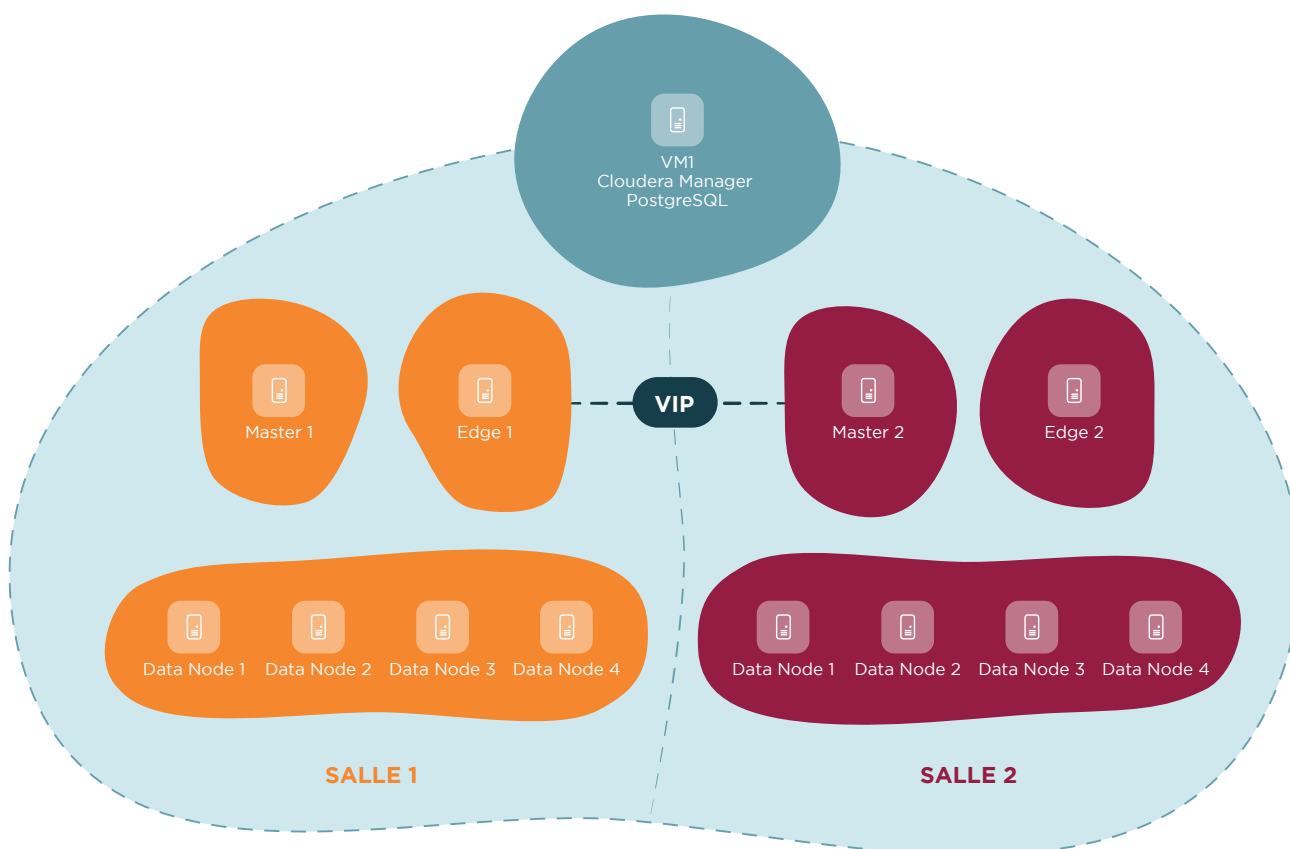
- des groupes dans l'AD,
- des zones Raw, Processing et Refined dans

le système HDFS,

- des ACLs (droits) dans tout le système en lien avec l'AD et Kerberos,
- des bases Hive, Impala, Hbase avec les bons droits de façon automatisée.

La mise en place des environnements de développement, de test et de recette doit suivre la même exigence. On peut imaginer la mise en place d'environnements physiques, mais cela peut représenter certains coûts d'investissement et un délai important. Dans notre cas, nous nous sommes contentés de déployer les environnements inférieurs sur des serveurs virtualisés. Les environnements de pré-production et de production sont les seuls à disposer d'architecture physique afin de permettre une réelle comparaison des performances sur les développements avant la mise en production.

Schéma physique de l'architecture Hadoop de production sur deux zones d'hébergement





CONTACTEZ NOS EXPERTS



Jean-Christophe ALLE
jean.christophe.alle@cgi.com



Véronique L'HELGUEN SMAHI
veronique.lhelguen.smahi@cgi.com



Vincent AMBROISE
vincent.ambroise@cgi.com



Jérôme MALARD
jerome.malard@cgi.com



Christophe CERQUEIRA
christophe.cerqueira@cgi.com



Olivier PRUDENT
olivier.prudent@cgi.com



Olivier DROUIN
olivier.drouin@cgi.com



Vincent RUANO
vincent.ruano@cgi.com

CGI

17, place des reflets
92400 Paris La Défense
www.cgi.fr/donnees