LIVRE BLANC

Comprendre les data-lakes

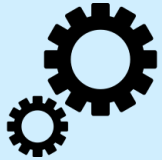
Les enjeux des nouvelles infrastructures de la donnée, pour une approche data-driven

Avril 2016

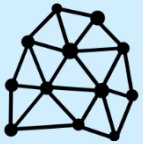
Converteo
ADLPerformance
SMART DIGITAL & DATA CONSULTANTS

Un data-lake en quelques mots c'est :

Espace de stockage de données



Avec des capacités de traitement



Virtuellement sans limite en s'appuyant sur une infrastructure big data
(approche distribuée potentiellement dans le cloud)

Permettant de stocker des données non structurées à moindre coût

Et de les retraiter en un temps record au moment de leur exploitation

Une opportunité technologique à mettre au service du business

Sommaire

1.

Révéler le potentiel business de vos data grâce aux data-lakes

Un data-lake pour exploiter et valoriser pleinement le potentiel data des organisations

Une pluralité de projets big data peuvent bénéficier de l'adoption d'un data-lake

	Travel – Leisure	Banque – Assurance	Retail	Transport – Industrie
Marketing – Expérience Client	Hotels.com *Personnalisation en fonction des données comportementales et de celles issues des réseaux sociaux*	Direct Assurance *Assurance auto connectée mesurant la qualité de la conduite pour proposer une assurance sur mesure*	picard *Identification des appétences canal (magasins vs web) pour optimiser les campagnes de couponing*	SNCF *Analyse des données wifi des gares, afin de suivre et d'optimiser en temps réel les flux voyageurs*
Marketing Média	Disneyland Paris *Amélioration du ROI en s'appuyant sur un modèle d'attribution connecté à 9 sources de données*	SOFINCO *Personnalisation des campagnes RTB pour optimiser les investissements médias*	Target *Algorithme déterminant la probabilité d'appartenance à un segment spécifique (femmes enceintes par ex) pour activation*	OSCARO *Analyse de toutes les données comportementales et d'acquisition au niveau le plus granulaire sur BigQuery de Google*
Optimisation industrielle	ACCOR HOTELS Feel Welcome *Plateforme réunissant 3600 hôtels, les informations de la concurrence, avis TripAdvisor … pour estimer le taux de remplissage et le pricing adéquat*	arvato BERTELSMANN *Solution basée sur les outils Big data qui fournit des analyses de fraudes en quasi temps réel (27millions d'€ sauvés / an)*	Pier 1 imports *Développement d'une solution d'analyse prédictive pour déterminer les futurs produits à la mode et préparer les campagnes marketing*	AIR FRANCE *Optimiser la gestion du stock de pièces détachées dans l'aviation en estimant finement la demande*
Innovation produit – Services	paradiski Les Arcs • Peisey-Vallandry • la Plagne *Application digitalisant le domaine skiable Paradiski pour personnaliser l'expérience ski des clients*	Crédit Mutuel ARKEA *Application d'aide prédictive à l'épargne basée sur 15 années d'historique de dépenses et revenus*	WEARING *Solution de « cognitive computing » e-commerce constituée d'un interface capable de répondre aux questions des internautes et de leur proposer des produits personnalisés*	RENAULT *Utilisation du big data pour réaliser de nombreuses simulations en phase de conception de nouveaux modèles*

Sources : Bigdataparis / Datafloq / Cap Gemini Consulting, « Big Data : où en est votre entreprise. Vraiment »

Le data-lake, terrain de jeu du big data

Un data-lake repose sur des outils permettant de traiter **rapidement d'importants volumes d'information - structurés ou bruts -** issus d'une **grande variété de sources**

*En 2015, 7.9 zo de données auraient été crées dans le monde, dont 80% non structurées**

EXHAUSTIVITE	Stocker tous les types de données, au format le plus granulaire, pour toujours pouvoir accéder au potentiel de leur forme non altérée
HISTORISATION	Conserver des données dans le temps pour établir des analyses de tendance ou comparatives
CONVERGENCE	Centraliser, joindre et comparer des données provenant de différentes sources (externes ou internes) pour réaliser des analyses exhaustives et transversales
QUALITATIVITE	Assurer la qualité de la donnée en amont de tout traitement et utilisation
ACCESSIBILITE	Accéder facilement aux données et les traiter en temps réel ou les requêter ponctuellement

** Source : Cap Gemini Consulting, « Big Data : où en est votre entreprise. Vraiment »*

2.

Intégrer le data-lake dans votre écosystème data

Le data-lake : une évolution du datawarehouse qui n'empêche pas les deux outils de conserver une certaine complémentarité

	Datawarehouse	Data-lake
Nature des données	Stocke **uniquement les données déjà structurées et considérées comme « utiles »** à l'entreprise	Stocke **toutes les données,** celles utiles aujourd'hui ou potentiellement dans le futur
Modèle de données	**Architecture qui repose sur des tables relationnelles** Structure **peu responsive** : espace de **stockage très structuré,** compliqué et chronophage à faire évoluer Stocke **certains types de données -** généralement des métriques quantitatives	**Architecture flexible** sans contrainte de forme ou de schéma particulier et a priori **Structure évolutive et non figée** ou de nombreux types et formes de données peuvent cohabiter et venir s'ajouter dans le temps Données stockées **quelle que soit leur forme** (brute et structurée ou non structurée)
Finalités	Son modèle de donnée très structuré rend le datawarehouse adapté à des analyses répétitives **Logique ETL** (Extract – Transform – Load)	Structure agile, les données sont configurées et traitées selon les besoins, via des séquencements parallélisés et indépendants **Logique ELT** (Extract – Load – Transform)

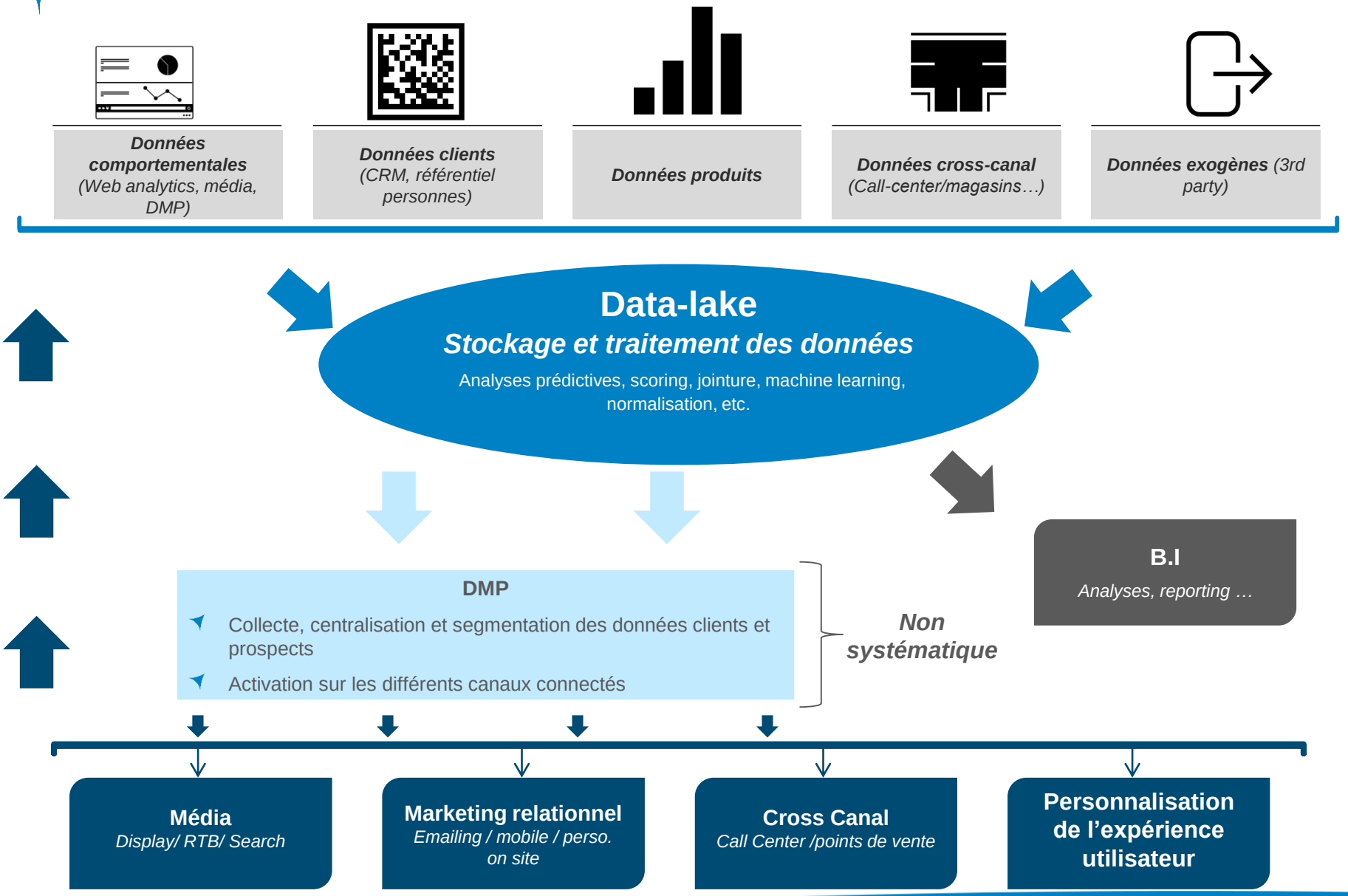
Le datawarehouse et le data-lake sont **complémentaires** et peuvent **cohabiter** : le data-lake stockant et traitant des données issues de nouvelles sources non configurées pour les datawarehouses

Data-lake et Data Management Platforms : quelles différences ?

	DMP	Data-lake
Stockage données	La DMP n'a pas vocation à stocker tout type de données, notamment les données personnelles des visiteurs, prospects et clients (PII) ainsi que les données sensibles de l'entreprise (marge, achats, etc.) – du moins de manière non cryptée	Le data-lake regroupe la totalité des données - y.c les données personnelles et sensibles -, dont celles issues de la DMP Le data-lake est un actif propre à l'entreprise, qui s'inscrit dans la durée
Connexions	La DMP est par nature connectée en temps réel à l'écosystème digital externe (DSP, 3rd party données providers, etc.)	Le data-lake n'a pas vocation à être connecté directement à l'écosystème externe. Les flux de données entrants et sortants ne sont pas nécessairement en temps réel
Activation de la donnée	La DMP est un outil d'identification (matching de cookies), de déduplication et d'enrichissement des données clients et prospects collectées et pré-calculées La DMP récupère l'intelligence pour l'activer sur les différents canaux avec lesquels elle est connectée - online et offline	Le data-lake agrège et traite la donnée pour la transformer en intelligence et la transmettre aux différentes briques d'activation, parmi lesquelles se trouve la DMP

Le data-lake est avant tout orienté connaissance et intelligence - c'est le socle exhaustif de la donnée, la DMP est résolument opérationnelle et orientée activation

Intégration d'un data-lake dans un dispositif data & digital



CASE STUDY

Refonte de l'infrastructure data

Converteo ADLPerformance – DIGITAL & CROSS-CHANNEL CONSULTING

Affini-TECH POUR MAISONS DU MONDE – MEUBLES & DÉCORATION

Les objectifs du projet

- Centraliser et faire converger une donnée de qualité pour la rendre accessible et exploitable pour tous les besoins métiers
- Automatiser les cas d'activations « classiques » de la donnée pour permettre à l'équipe data de consacrer du temps à de nouveaux projets à forte valeur ajoutée

Sources de données

Google Analytics

Google BigQuery

- Données GA/BigQuery Données acquisition
- Données produits
- Données CRM
- Données Magasins
- Données comptables – Achats/stocks
- Données exogènes

Google Cloud Platform

- Stockage en ligne
- Machines virtuelles scalables

Google BigQuery

- Requêtage et traitement de la donnée

Use cases principaux

Pilotage de la donnée

- Via des outils de data-viz directement connectés à l'infrastructure data-lake
- Les équipes métiers ont accès à des dashboards exhaustifs (toutes les sources et niveau de granularité possible

Cross-canal

- Mesurer la contribution du web à l'activité online en magasin : achats cross-canaux, RoPo, showrooming

Scénarios relationnels

- Optimiser les relances marketing pour adapter la pression 360° en fonction de la réaction des clients

3.

Appréhender les différents data-lakes possibles

Deux grands modes d'hébergement pour satisfaire des exigences différentes

	« On Premise »	Cloud Services
Ressources nécessaires	★☆☆	★★☆
Humaines	Des compétences très spécifiques d'une relative rareté sont nécessaires	Le cloud ne dispense pas des ressources de conception et d'administration de l'infrastructure
Hardware & Software	Même si le système peut fonctionner sur des machines banalisées, l'organisation doit se doter de son propre data-center Bien qu'étant open source et gratuit de base, une distribution payante du framework Hadoop est souvent à privilégier	Même si l'investissement direct dans du matériel hardware n'a pas lieu d'être le coût d'exploitation des machines est inclus dans la facturation globale. Les prestataires cloud facturent les services de manière packagée en fonction de la consommation des ressources machines
Sécurité de la donnée	★★★	★★★
	L'organisation impose à sa donnée ses propres exigences de sécurité	Les prestataires de cloud font bénéficier à leur client des mêmes engagements de sécurité qu'ils imposent à leur propre donnée
Facilité de déploiement	★☆☆	★★★
	Les frameworks nécessaires au fonctionnement du système impliquent du paramétrage relativement lourd et complexe	L'utilisation de plateformes et services managés permettent un déploiement très rapide
Gouvernance de la donnée	★★★	★★☆
	L'organisation est propriétaire exclusif de toute la chaîne de stockage et d'exploitation de la donnée	Le prestataire de service constitue un intermédiaire entre l'organisation et sa donnée qui lui appartient tout de même de manière exclusive
Scalabilité	★★☆	★★★
	La scalabilité est linéaire et l'unité d'ajustement est le serveur	La scalabilité est très granulaire et peut être gérée en fonction du volume stocké ou du temps d'utilisation des ressources

La dématérialisation de l'infrastructure et l'intégration native avec la plupart des outils digitaux favorisent le Cloud pour des projets agiles

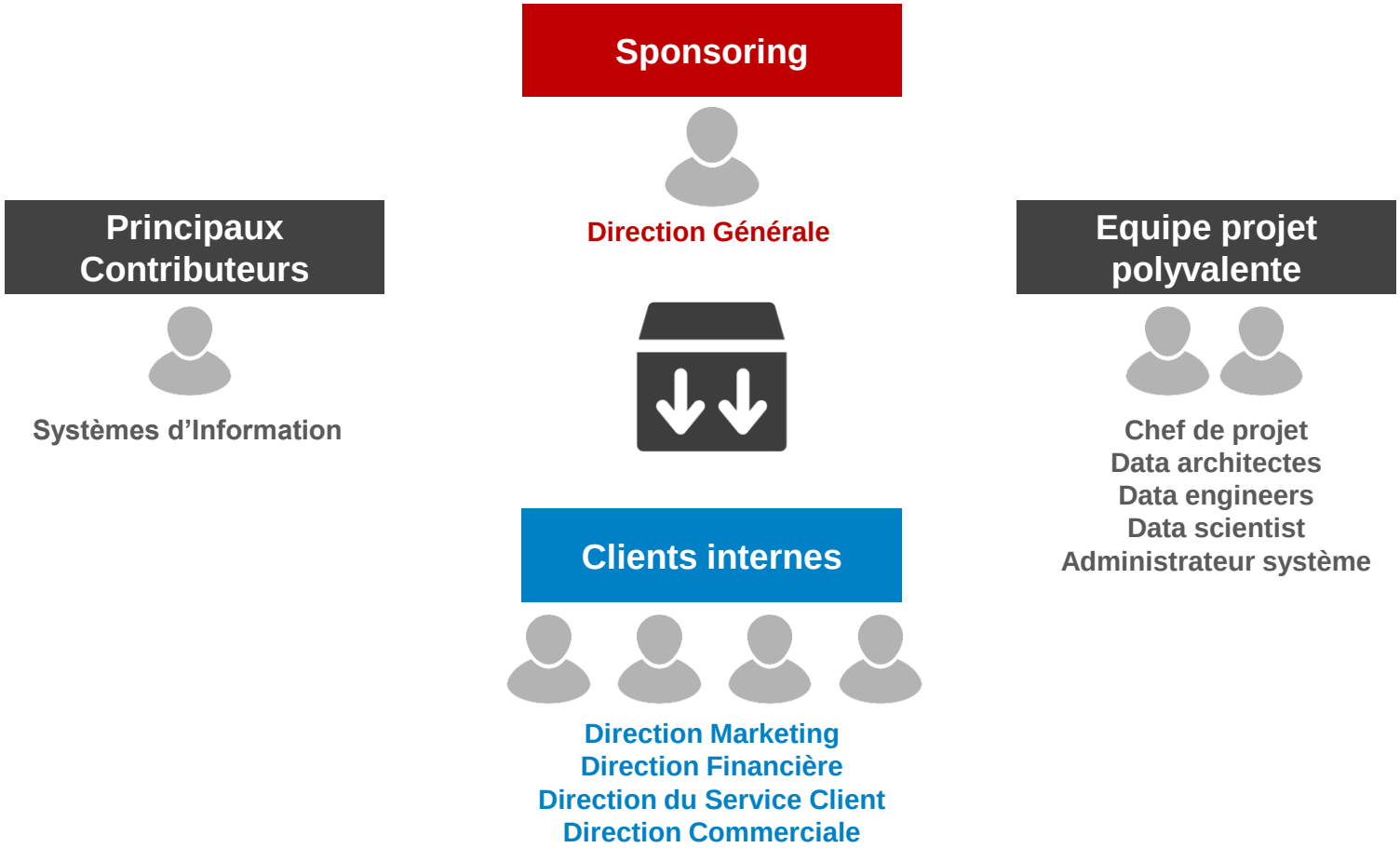
Collecte et stockage de la donnée	Traitement des données	Activation de la donnée
On premise		
La donnée est stockée sur des grappes de serveurs, avec des frameworks permettant le traitement distribué sur ces serveurs cloudera Hortonworks MAPR	L'analyse, le requêtage et la synthèse des données contenues sur Hadoop, se fait à travers des logiciels SQL-like ou java-like HIVE Spark apache PIG	L'intégration d'un Hadoop On premise avec l'éco-système SI/marketing d'une entreprise nécessite un ETL ou un développement spécifique
Cloud Services		
L'infrastructure de calcul et de stockage distribuée, est rapidement configurée sur des machines virtuelles et des espaces pré-paramétrés sur le cloud amazon web services S3 Microsoft Azure Active Directory Google Cloud Platform	Les suites Cloud offrent des outils de requêtage SQL-like très performants directement liés à la donnée stockée AMAZON REDSHIFT Azure data Lake analytics Google BigQuery	La plupart des outils digitaux du marché sont connectés nativement aux principales Suites Cloud (Google, Amazon, Microsoft …)

4.

Mener un projet data-lake

Le déploiement d'un data-lake est facilité par la constitution d'une équipe projet ad hoc bénéficiant d'un solide sponsoring

L'agilité nécessaire à un projet data-lake, la spécificité des ressources nécessaires à son exploitation ainsi que la transversalité de ses implications dans l'organisation nécessitent souvent sa construction en marge ou en parallèle d'un système d'information historique auquel il sera étroitement lié



Quels sont les utilisateurs d'un data-lake ?

Chef de projet

- Conçoit, développe et s'occupe au quotidien de la plateforme, c'est le maître d'œuvre du data-lake

Data Engineer

- Connecte le data-lake à toutes les sources de données, s'assure de la qualité de la donnée et lie la plateforme aux applications externes

Data Analyst

- Définit les KPIs et les métriques à des fins de reporting / dashboarding pour comprendre les processus métiers

Data Scientist

- Exploite en profondeur l'ensemble des données à des fins de prospection, pour déterminer les grandes tendances business et les opportunités que l'organisation devra saisir

Admin. système

- Responsable de l'infrastructure du data-lake et de la sécurité (à temps partiel sur le projet)

Les différentes étapes d'un projet data-lake

Un projet data-lake doit être mené selon une **méthodologie agile**, avec un **déploiement progressif**. **Tirer partie de la scalabilité** des infrastructures est la **clé de la réussite** d'un projet data-lake.

1 Etude d'opportunité/faisabilité et prise en considération des besoins

- Identification des cas d'usages pertinents
- Identification de la complexité de l'existant à rapatrier (mapping des sources de données et flux out à connecter)
- Evaluation des contraintes techniques, humaines, organisationnelles

2 Setup technico-fonctionnel du data-lake

- Définition de l'infrastructure et de l'environnement en fonction des besoins et contraintes
- Priorisation des cas d'usage et planning du déploiement
- Mise en place d'une équipe projet

3 Déploiement progressif des cas d'usage

- Lancement progressif des cas d'usages
- Etude de retour sur investissement
- Conduite du changement pour le shift vers du data-driven décision making

Les complexités et risques du projet data-lake

SPONSORSHIP et ADHESION – Un appui fort à un niveau hiérarchique élevé est requis pour obtenir l'adhésion et éviter les points de blocage

- Un projet data-lake est stratégique et concerne toutes les entités de l'entreprise, qui seront toutes parties prenantes du setup et de l'exploitation du projet
- La multiplicité des interlocuteurs, des technologies et des métiers complexifie de facto le projet et implique donc une forte priorisation du projet au niveau stratégique

PRAGMATISME – Le développement progressif d'un projet data-lake est un gage de réussite

- Avec des cas d'usage priorisés, qui imposent le rythme de raccordement des sources (flux in) et des activations (flux out)
- Tirer partie de la scalabilité de l'infrastructure : on monte/descend en charge en fonction de la nécessité

DATA LEAKING – Contrôle et ownership : attention à la fuite de donnée

- Toute la donnée, au niveau le plus granulaire est contenue dans le data-lake : la sécurité autour du data-lake est un enjeu majeur, qui peut conditionner des choix de faisabilité, de technologie ou autres

Merci de votre attention

Converteo
ADLPerformance
SMART DIGITAL & DATA CONSULTANTS

Thomas Faivre-Duboz, *Directeur Associé*
Julien Ribourt, *Manager*
Paul Ghorra, *Consultant*
Arthur Fulconis, *Consultant*