Centrale
CASABLANCA

DATA ENGINEERING

CONCEVOIR DES ENTREPOTS DE DONNÉES « RESPONSABLES »

PLAN DE L'INTERVENTION

Présentation générale

Atelier
Data Engineering

Conclusions

Data et gouvernance

Expositions

Collecte et stockage

Traitement et raffinage

UN HÉRITAGE …

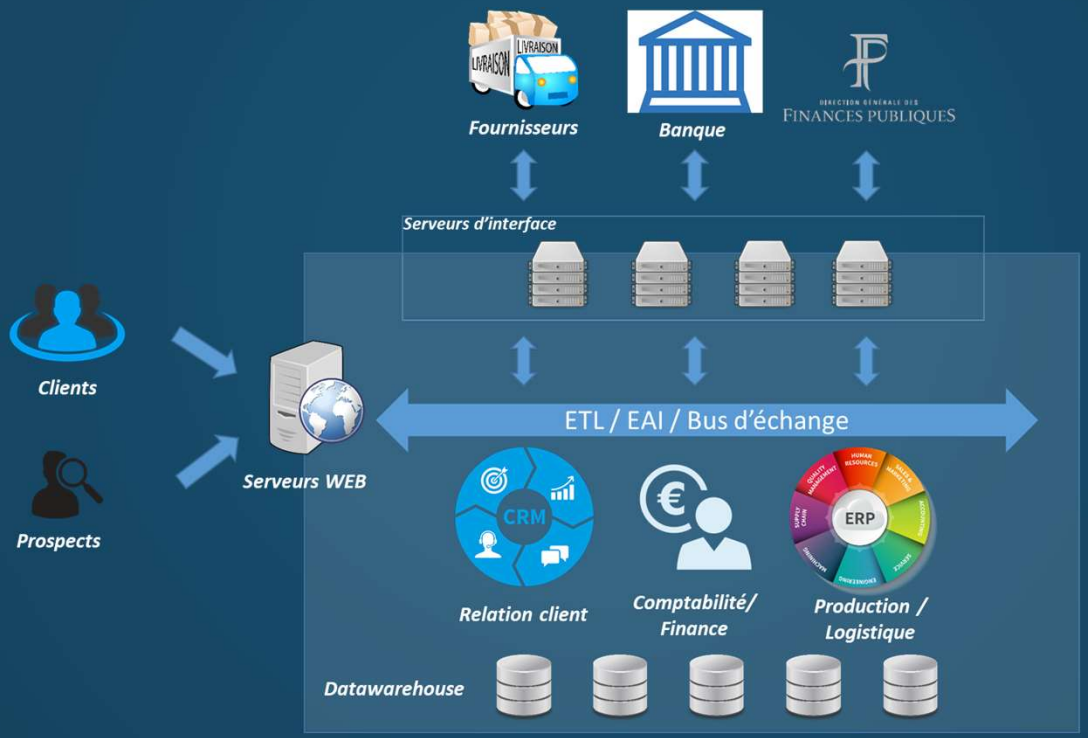
L'intelligence autour des données existe depuis longtemps dans l'industrie, les banques, les assurances, …

Mettre en commun les données des différents domaines de l'entreprise dans un seul et même « endroit » .. Le Datawarehouse (Patrimoine « data » de l'entreprise)

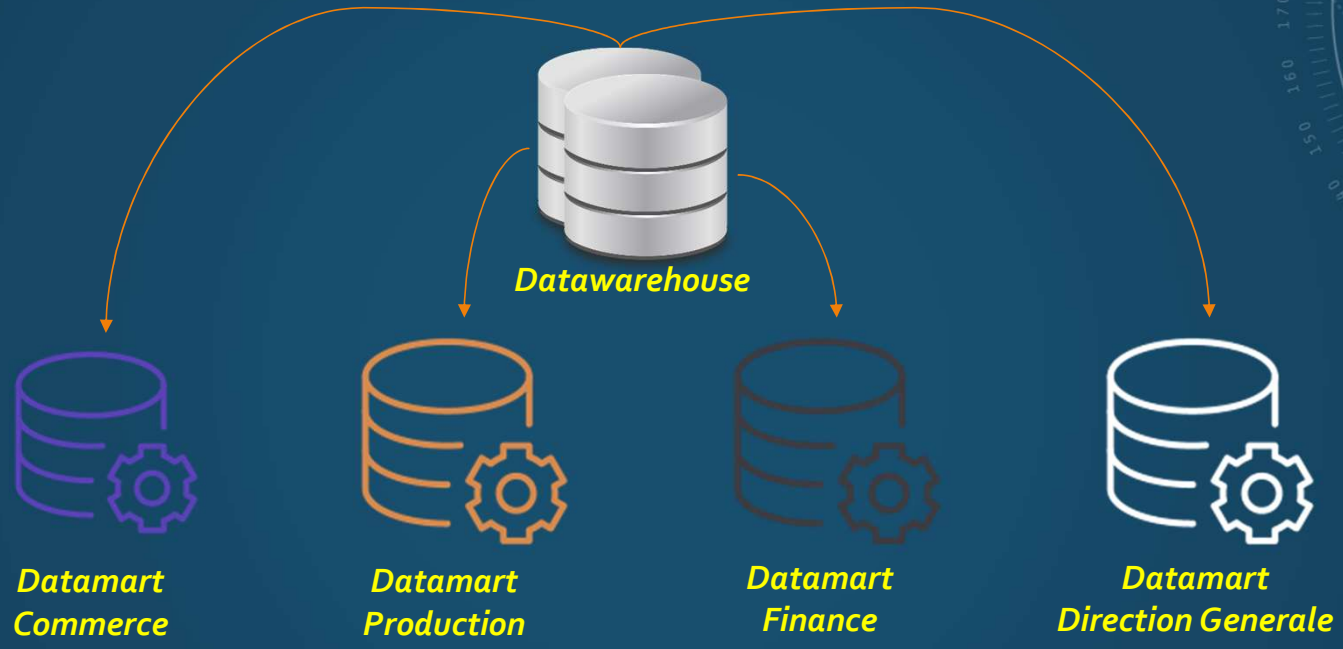
Faire circuler les données inter-applications, permet un pilotage transverse de l'entreprise

UN HÉRITAGE EXISTANT – EXEMPLE DE STRATÉGIE DATA

Un modèle « classique » de circulation de l'information en entreprise

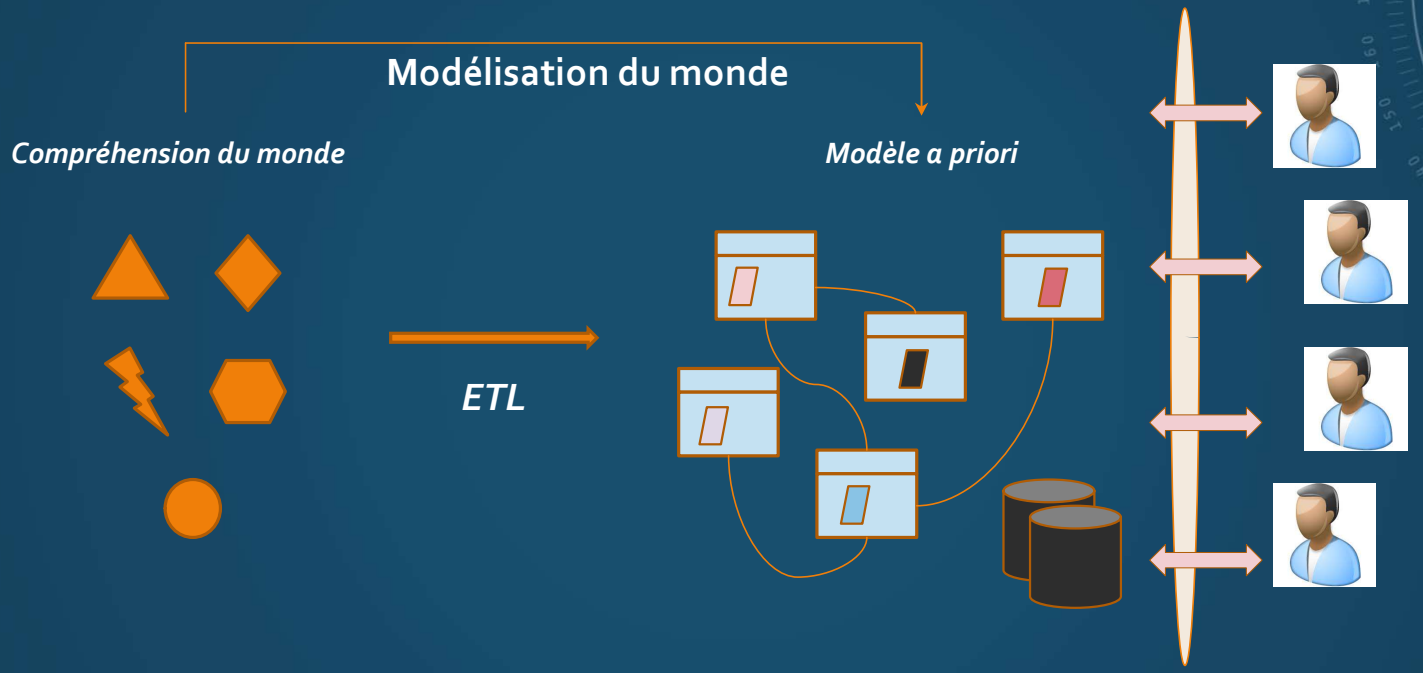


UN HÉRITAGE EXISTANT – EXEMPLE DE STRATÉGIE DATA



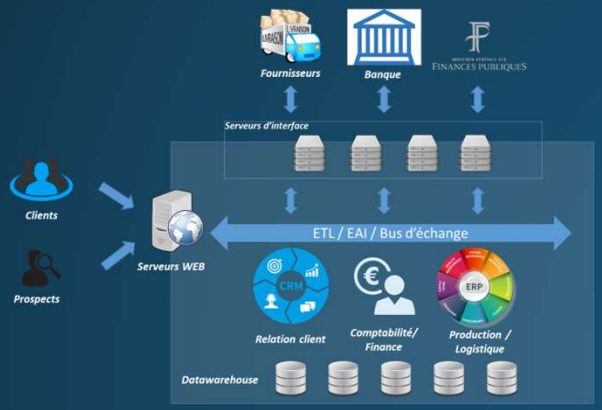
Une description du « monde » a priori
Une représentation « statique » des données
avec des capacités d'évolutions limitées.

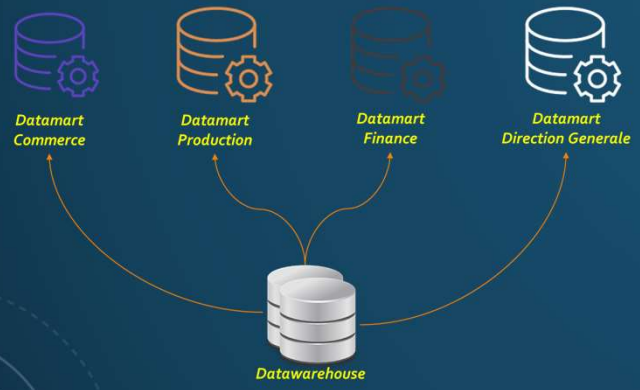
UN MONDE STATIQUE – SCHEMA « ON WRITE »



Les utilisateurs finaux sont « liés » au modèle prévu initialement. La vision du « monde » est figée et les évolutions sont « délicates » car elles entraînent une modification d'un grand nombre d'applicatifs et de structures de données

QUE MANQUE T'IL ?





- *Des données externes*
- *Des modèles flexibles et adaptables*
- *Des métriques calculés à la demande*
- *Une collaboration de toutes les équipes*
- *Des données accessibles à tous*
- *Un travail d'élaboration de toute l'entreprise*
- *Une adaptation des moyens techniques*
- *Un espace de R&D pour « concevoir » les modèles*
- *Des cas d'usages liés à la stratégie*

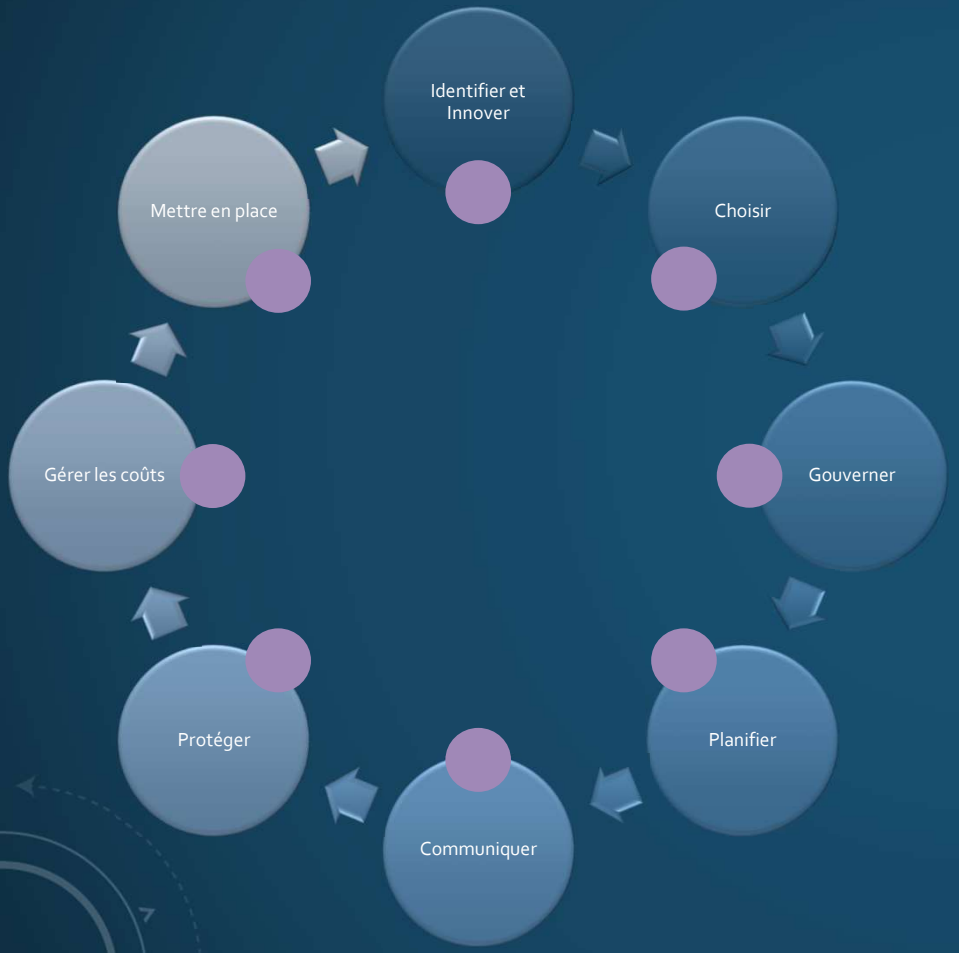
UNE STRATÉGIE « DONNÉES » POUR L'ENTREPRISE

Pourquoi une stratégie data est « critique » pour l'entreprise ? Pourquoi créer une vision « data » de l'entreprise ?
…

Comment l'entreprise va-t-elle être « modifiée » ?
Comment faire évoluer les mentalités ?
…

Quels résultats attendre ?
Quels sont les risques possibles ?
…

UNE STRATÉGIE « DONNÉES » POUR L'ENTREPRISE



Identifier les cas d'usage et **innover** dans leur conception en s'appuyant sur des données

Choisir les cas d'usage pertinents et en phase avec la stratégie

Gouverner l'intégration et le stockage des données dans le SI

Planifier la mise en œuvre des chantiers et suivre leurs évolutions

Communiquer sur l'avancement et sur les bénéfices des projets en cours

Protéger par construction toutes les données privées et personnelles. **Protéger** les données sensibles

Gérer les coûts en mettant en place un contrôle de gestion strict et efficace

Mettre en place et mesurer les résultats obtenus

IDENTIFIER ET INNOVER

Identifier les cas d'usage :

- Améliorer la prise en charge des clients par le service commercial
- Réduire l'érosion de la clientèle
- Déterminer les motifs des volumes de vente (saisonnalité, variations exogènes, …)
- Prévision des tendances produits (couleur, forme, matière, …)
- Calcul du niveau d'usure sur des pièces mécaniques subissant des contraintes
- Analyse comportementale de la concurrence
- …

Identifier les stocks de données :

- Imaginer les données nécessaires pour chaque cas d'usages
- En interne, déterminer au cas par cas les données existantes
- En externe, identifier les sources de données , leur qualité et leur pertinence

Innover :

- Penser sans contraintes et sans limites, sur les données et sur les processus
- Imaginer comment l'entreprise pourrait se différencier des autres
- Concevoir librement une histoire du cas d'usage.

CHOISIR

Pour l'entreprise :

- Etre en accord avec la stratégie de l'entreprise
- Prendre en compte l'existant et favoriser les transitions « douces »
- Ajouter de la valeur aux produits ou services de l'entreprise
- Minimiser les risques
- Analyser si le projet est « réalisable »
- Mettre en place place des KPI de mesure du ROI Projet

Pour les équipes :

- Définir des objectifs ambitieux mais atteignables
- Respecter les sensibilités et les niveaux de formation
- Travailler dans un espace propre à l'innovation et à la création
- Favoriser le travail inter départements et l'esprit de construction

GOUVERNER

Pour encadrer :

- La collecte des données
- Respecter les réglementations sur la vie privée et le droit à l'oubli
- L'utilisation des données faite par les utilisateurs identifiés et par les applications
- La disponibilité des données dans tous les départements de l'entreprise
- L'intégrité des données dans les entrepôts de stockage

Pour définir :

- Les règles d'intégration des données
- Les niveaux de sécurité et de confidentialité
- Les moyens de contrôles et d'audit
- Les processus de transformation et de mesure de la qualité des données
- Les socles logiciels à mettre en place pour mettre en œuvre les processus définis
- Les KPI (Key Performances Indicators)

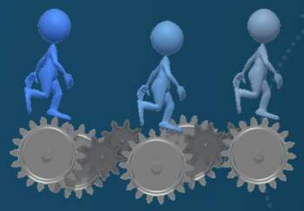
PLANIFIER

Pour prévoir:

- Les différentes étapes des projets (Roadmap)
- La participation des membres de l'équipe projet
- L'utilisation des moyens de l'entreprise (techniques, fonctionnels)
- Les dépassements de délais en fonction des problèmes rencontrés

Pour réussir:

- En se donnant des jalons visibles dans le temps
- En adaptant les besoins durant le parcours (agilité)
- Et atteindre les objectifs financiers du projet



COMMUNIQUER

Pour annoncer :

- Quelle est la nature du ou des projets « data »
- Quels seront les bénéfices pour les utilisateurs
- Comment l'entreprise va « grandir » en innovant
- L'achèvement des différents jalons des projets

Pour échanger :

- Entre les membres de l' équipe projet
- Avec les différents niveaux hiérarchiques , parties prenantes du projet
- Et faciliter la prise de décision
- Sur les objectifs atteints et les livrables
- Sur la culture de la donnée en entreprise et les bénéfices associés

PROTEGER

Pour garantir :

- En interne et en externe la mise en place d'une éthique d'entreprise
- L'utilisation raisonnée et responsable des données à caractère personnel
- Le droit à l'oubli et le respect de la vie privée
- L'utilisation conforme des données par les utilisateurs
- L'absence de diffusion de données confidentielles à l'extérieur de l'entreprise

Pour sauvegarder :

- De manière pérenne le patrimoine données de l'entreprise
- Le travail réalisé par les collaborateurs
- La propriété intellectuelle de l'entreprise en matière d'algorithmes
- Les investissement de l'entreprise

GERER LES COÛTS

Pour anticiper :

- La charge nécessaire (employés [TJM], moyens techniques,…)
- Les besoins de financements internes (acquisition de logiciels, développement, …)
- Les risques inhérents aux projets « data » (qualité des données, disponibilité, …)

Pour suivre :

- L'écart entre les estimations et le réalisé (dérives)
- Le P&L du projet , les dépenses et les charges consommées
- Les risques financiers extérieurs aux projets
- Les indicateurs financiers du projet (Contrôle de gestion)

METTRE EN PLACE

Avant :

- Mesurer l'achèvement du projet à mettre en place
- Préparer une checklist de lancement du projet
- Vérifier que tous les intervenants sont informés du lancement
- Préparer un document de suivi de lancement

Pendant :

- Déclencher les différents étages du projet dans l'ordre prescrit
- Toujours avoir une solution de repli et/ou de retour arrière
- Vérifier que les circuits de données sont bien alimentés
- S'assurer auprès des intervenants que les étapes sont correctement réalisées
- Rédiger un rapport de lancement

Après :

- Vérifier le bon fonctionnement du projet pendant une période de validation (VABF)
- Communiquer le rapport de lancement et signaler les dysfonctionnements
- Mesurer l'efficacité du projet en calculant régulièrement des KPI définis au départ

LES DONNÉES AU CENTRE

Le Datawarehouse est la propriété des « informaticiens » et se trouve ainsi sanctuarisé

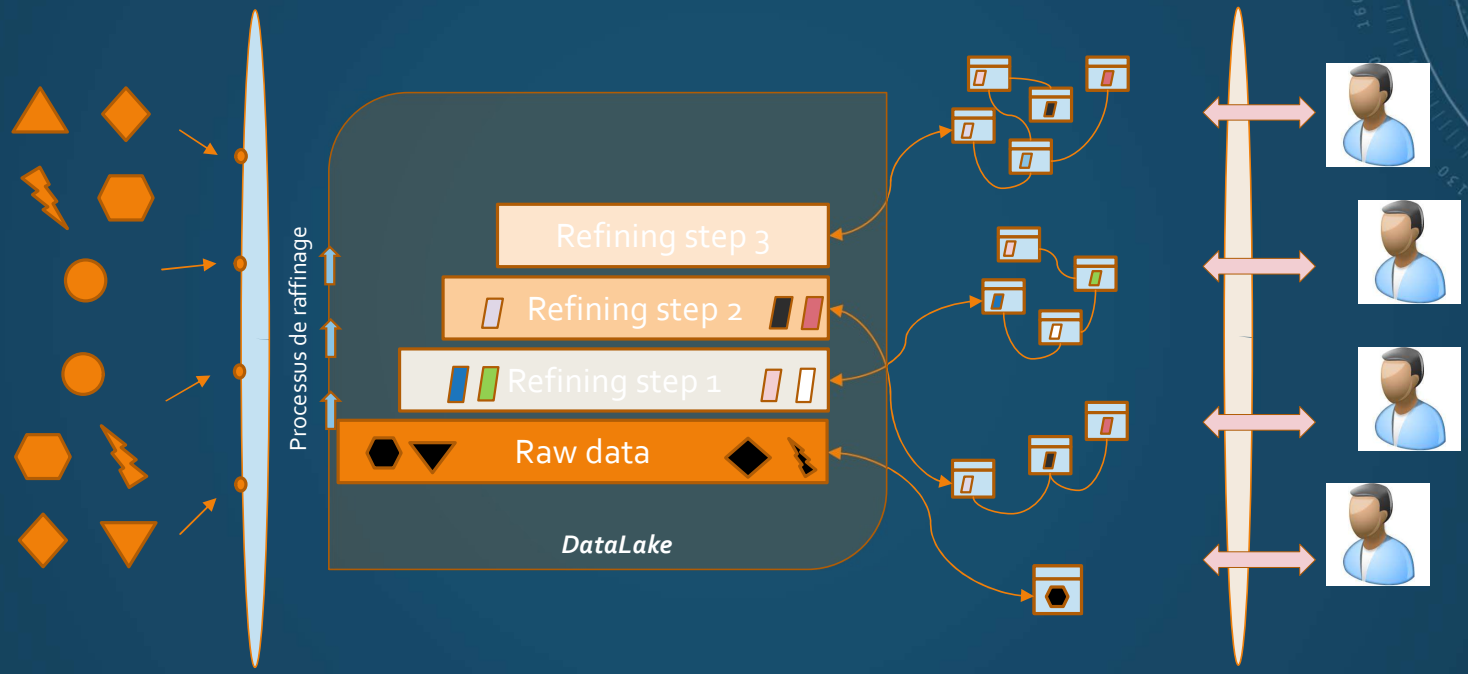
La vision du monde est celle de techniciens et beaucoup moins des opérationnels

Une stratégie orientée « data » consiste donc pour l'entreprise à repenser la manière de les exploiter

Le patrimoine de l'entreprise doit être préservé et la nouvelle stratégie doit composer avec l'existant

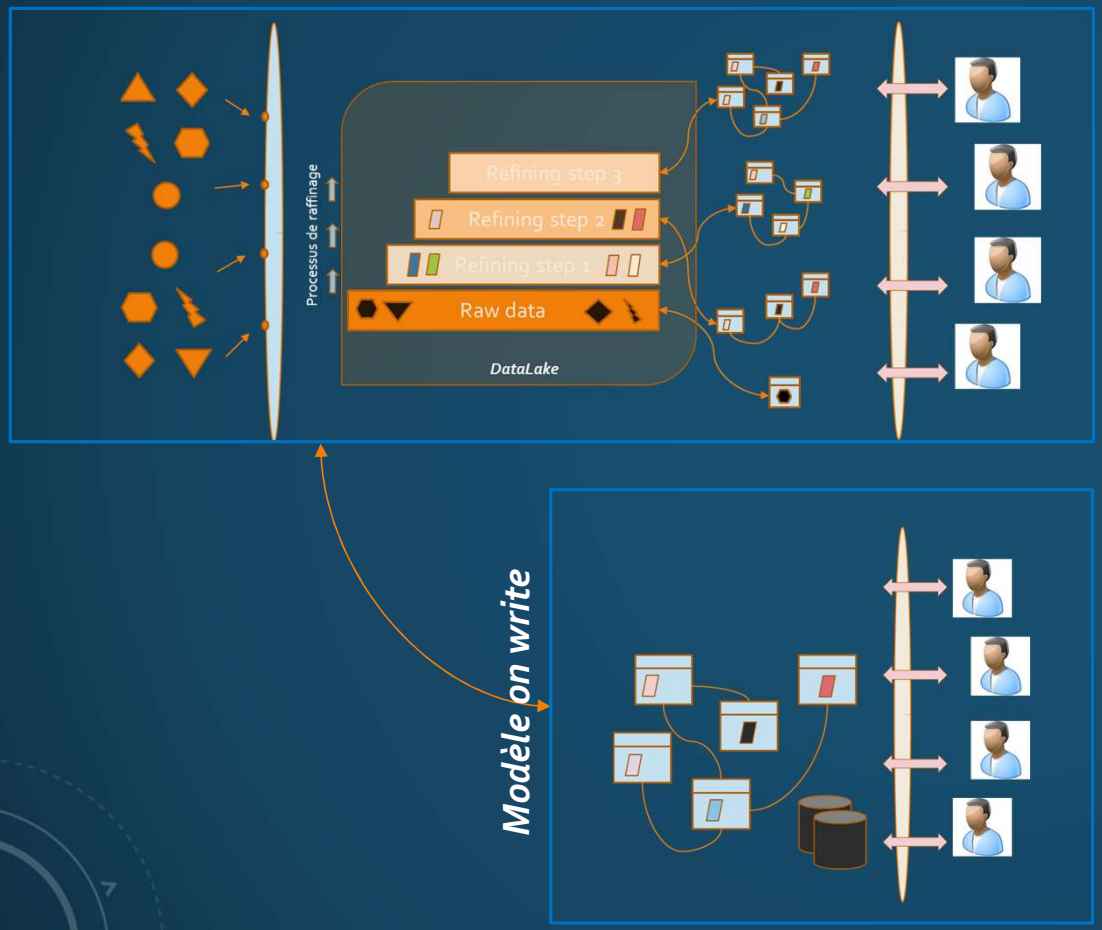
Il existe une autre manière, plus agile, plus souple pour mettre les données au centre

BOUGER AVEC LE MONDE – SCHEMA « ON READ »



Les utilisateurs finaux recherchent les données en fonction du modèle qu'ils veulent voir appliquer et ce compte tenu du niveau de raffinage souhaité (où imposé)

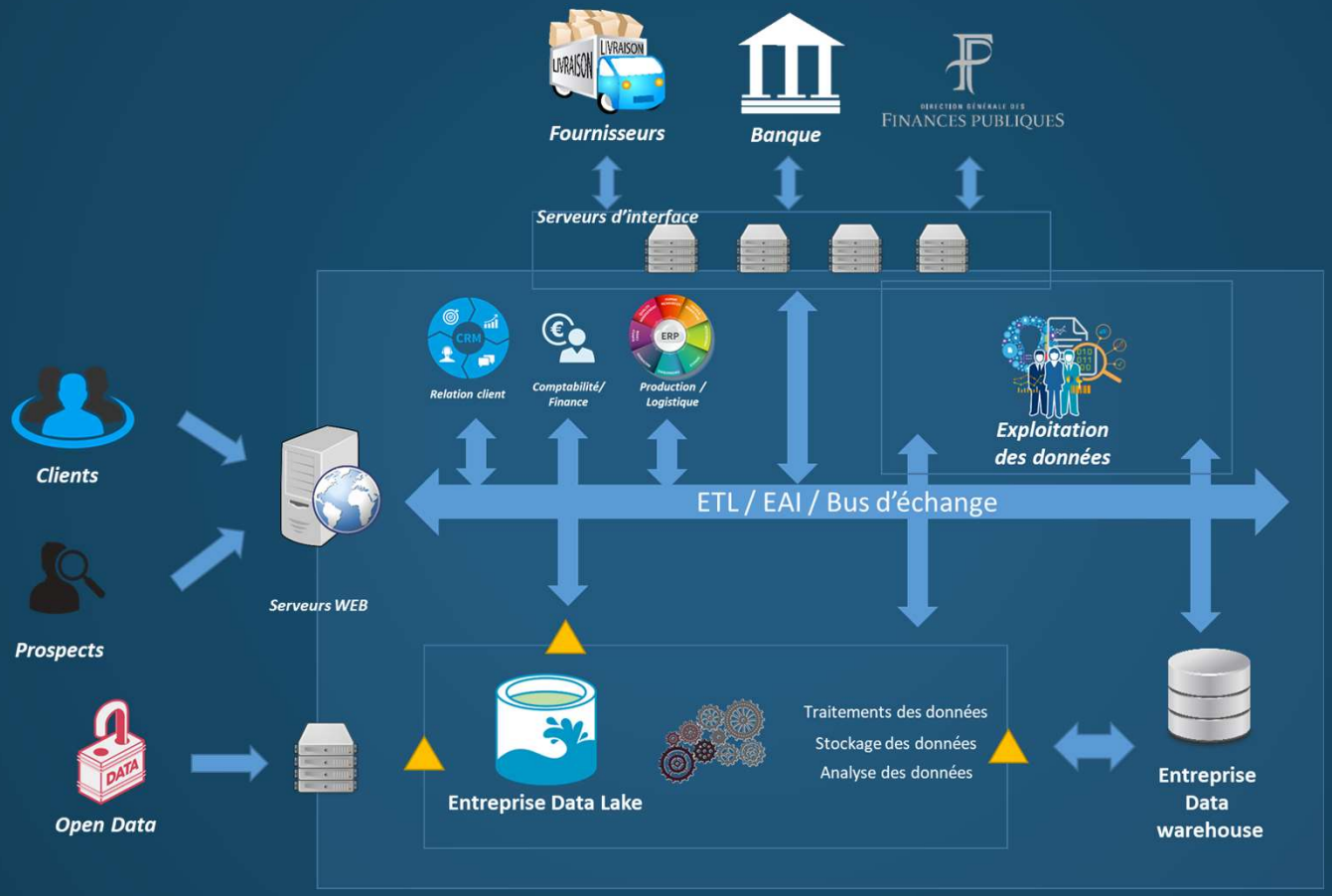
COEXISTENCE



Faire coexister les deux « mondes », permet :

- *De s'appuyer sur un patrimoine existant*
- *D'éviter les ruptures dans le traitement des données*
- *De lisser la charge de travail et les coûts*

VERS UNE NOUVELLE ORGANISATION « DATA CENTRIC »

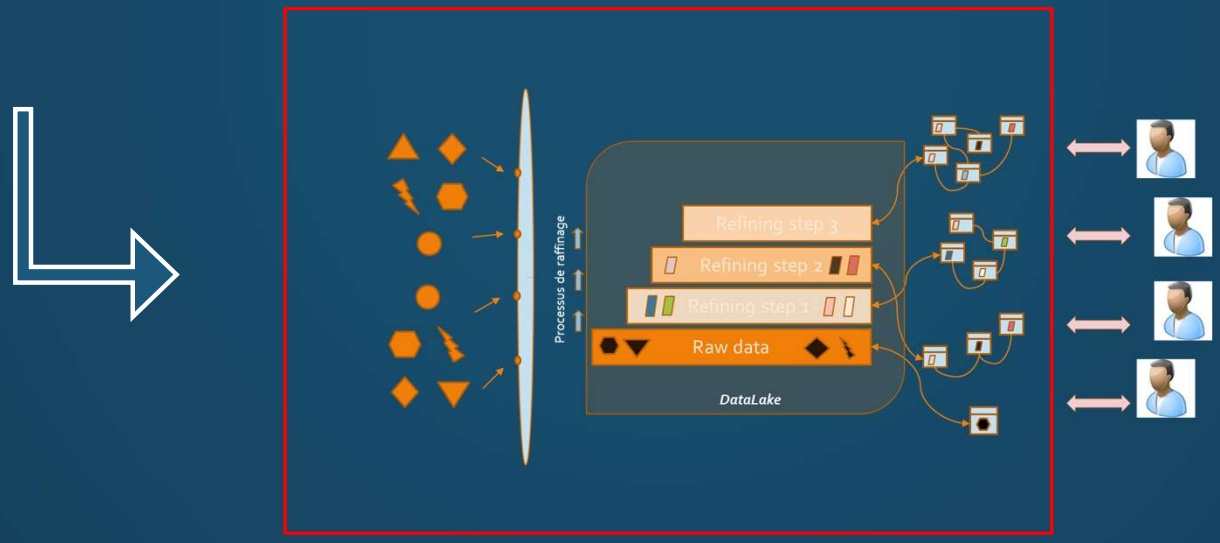


LE PROBLÈME RESTE ENTIER

La stratégie décrite permet à l'entreprise de s'adapter de manière plus efficace au monde extérieur

La coexistence des manières de gérer la donnée permet une adoption « douce » et « raisonnée » par l'entreprise

Mais comment gérer cela ??



DATA LAKE : LE RÉSERVOIR DES DONNÉES

- Les données doivent être stockées au fur et à mesure de leur arrivée dans l'entreprise
- Indexer les données doit être une priorité absolue
- Les traitements sur les données seront effectués a posteriori

- L'organisation du DataLake (Lac de données) doit être « robuste » et « pérenne »
- Le DataLake ne doit pas devenir un « marais » (respect des DLU)
- Le « minage » des données doit se faire sur des sous ensembles identifiés du DataLake

- Un processus d'audit du DataLake doit être prévu dès la conception
- Un tableau de bord (« carnet de santé » du DataLake) doit être mis en place et actualisé
- Des méthodes de corrections, de mises à jour doivent être prévues

D'UN FLUX MAITRISÉ A UN FLOT PERMANENT ?

Les données de l'entreprise présentent un volume maîtrisé et des formats normalisés

L'adaptation à cet environnement faiblement évolutif est « simple » et « maîtrisable »

L'ouverture vers de nouvelles sources de données est une « aventure » qu'il faut préparer

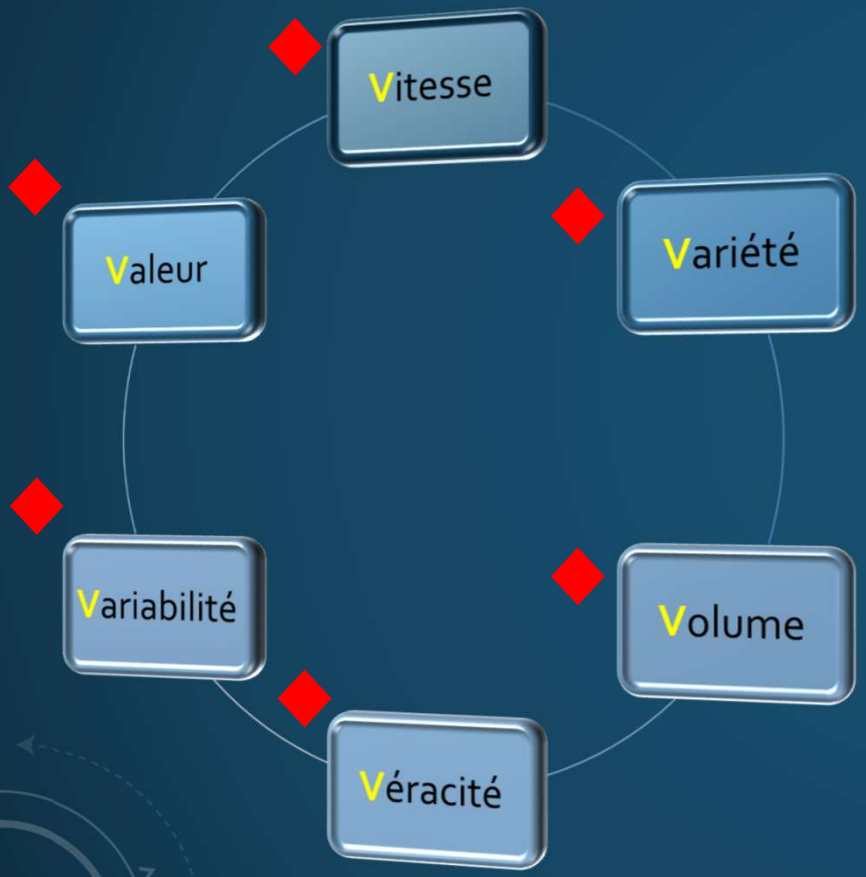
Les moyens techniques et humains doivent anticiper ce changement radical de paradigme

Parler de « Big Data » oui, mais de manière pragmatique et maîtrisée

Le « Big Data » n'est pas un « miracle » , c'est un nouveau socle technique

L'avalanche de données doit donner lieu à de nouvelles méthodes de travail

« BIG DATA » , POURQUOI ?



- **Vitesse** : Les données arrivent à des vitesses et à des fréquences toutes différentes
- **Variété** : Les types de données sont aussi variés que les données elles mêmes
- **Volume** : Les volumes sont très importants à l'intégration et au stockage
- **Véracité** : Les données viennent de tous les horizons, contrôlées ou non.
- **Variabilité** : Les sources de données peuvent être saisonnières ou se tarir complètement
- **Valeur** : Les données peuvent ou non représenter un intérêt pour l'entreprise

« BIG DATA » EST CE LA SOLUTION POUR TOUT ?

- Les volumes de données sont au-delà d'un volume de plusieurs centaines de Ti (Tebioctet) sur les 2 à 3 prochaines années (provisionning)
- Les formats de données sont très hétérogènes , et arrivent en temps réels et/ou différés
- Les « cas d'usage » sont identifiés et les flux de données « servent » à quelque chose
- L'évaluation du ROI est prête à être évaluée
- L'entreprise a préparé l'arrivée des données (formation, socle technique, socle logiciel)
- Le Datawarehouse arrive à saturation de volume et de pertinence

- Le « Big Data » c'est à la mode !!!
- Les données sont celles de mon entreprise et en volume peu important dans les 2 à 3 prochaines années
- Aucune stratégie autour de la donnée n'existe et les équipes ne sont pas formées
- Le Datawarehouse correspond aux besoins et n'a pas vocation à évoluer

« BIG DATA » OU STOCKER LES DONNÉES ?

Data Center de l'entreprise
(On premise)

Platform As A Service

Infrastructure As A Service

Data Center as a Service

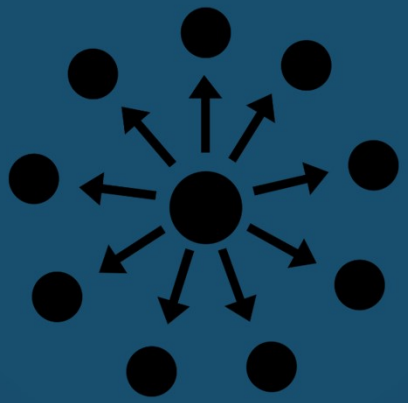
Modèle hybride
(Mix On premise / Cloud)

BIG DATA : ÉLASTICITÉ DES MOYENS

Adapter les capacités de stockage aux besoins

Mixer On Premise et Cloud (zone de débordement)

Augmenter les performances par la distribution des calculs



Faire disparaître les « goulots d'étranglements »

Supprimer les SPOF (Single Point Of Failure)

Rendre les changements d'échelle transparents

« BIG DATA » ET CLOUD

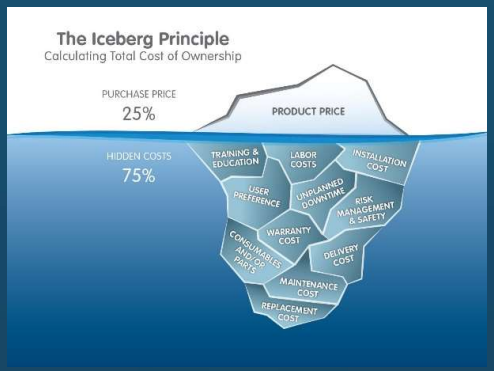
Si les capacités de stockages internes ne permettent plus de garder les données importantes pour la bonne marche de l'entreprise
Si les capacités de calculs internes deviennent un « goulot d'étranglement » pour les analystes de l'entreprise
Si les solutions techniques internes ne couvrent plus les besoins des data scientists

Investir dans une solution « cloud »

EUROPEAN OPEN SCIENCE CLOUD

BIG DATA » ET CLOUD : UN COUT DIFFICILE À ESTIMER

- Le TCO du SI d'une entreprise est très compliqué à calculer
- Les besoins en spécialistes, en machines (calcul, stockage), en logiciels difficiles à estimer
- Une solution « cloud » rajoute donc une complexité dans la gestion du SI (technique, financière)
- Le calcul des coûts proposés par les grandes plateformes « cloud » est très complexe
- L'utilisation d'une solution « cloud » nécessite un pilotage financier très fin
- Le ROI d'une solution « cloud » doit être estimé régulièrement
- L'estimation et le suivi des coûts doivent être effectués par une équipe multi-domaine

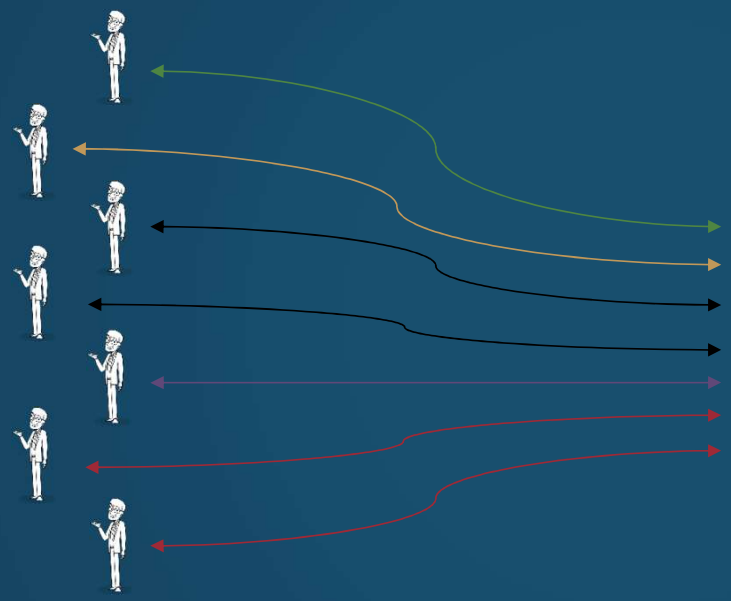


BIG DATA » ET CLOUD : INDÉPENDANCE ?

- Les données sont l'une des richesses de l'entreprise, les « confier » à une société externe représente un risque financier à évaluer
- Les flux de données montants sont « gratuits », les flux « descendants » payants
- Lorsqu'une entreprise s'engage dans le « cloud » , elle est liée à son fournisseur !
- **TOUJOURS** envisager le « retour arrière » lorsque l'on confie ses données à un fournisseur
- **TOUJOURS** rester maître du destin « data » de l'entreprise !

Reste à mesurer la dépendance de l'entreprise vis-à-vis de la plateforme de stockage des données !

UNE AUTRE SOLUTION DE STOCKAGE MASSIF : LA BLOCKCHAIN



http://www.scilogs.fr/complexites/la-puissance-de-la-blockchain/
[Jean-Paul Delahaye]

Un livre ouvert à tous , en écriture et en lecture, infalsifiable et indestructible

Un livre consultable depuis n'importe quel endroit du monde, sans organe de contrôle

UNE AUTRE SOLUTION DE STOCKAGE MASSIF : LA BLOCKCHAIN

La blockchain :

- Principe de fonctionnement utilisé par les crypto-monnaies (BitCoin, Ethereum,…)
  - Pas de banque et pas de mécanisme de régulation
  - Inventé par Satoshi Nakamoto, suite à la crise des subprimes en 2009
  - Le nombre de bitcoin est limité à 21 millions d'unités
  - Logiciel de minage open source est ouvert à tous
- Décentralisation, protocole ouvert pour la lecture et l'écriture
- Modification impossible au-delà d'un délai d'une heure. Pas de falsifications possibles

- ***Pour aller plus loin :***
  - https://blockchainfrance.net/
  - https://www.multichain.com/
  - https://courscryptomonnaies.com/bitcoin

BIG DATA : TOTAL COST OF OWNERSHIP

Projets

Formation

Infrastructures, Energie, locaux

R&D

Personnel

Maintenance

Sécurité

BIBLIOGRAPHIE

- https://www.canada.ca/content/dam/pco-bcp/documents/clk/Data_Strategy_Roadmap_ENG.pdf
- https://infolabs.io/sites/default/files/livret1culturedesdonneesvf_weba4.pdf
- https://www.damachicago.org/wp-content/uploads/2012/01/2015-Building-a-Data-centric-Strategy-and-Roadmap.key-1.pdf
- https://news.sap.com/wp-content/blogs.dir/1/files/SAP_Data-2020-Study_Infographic.pdf
- https://www.piloter.org/business-intelligence/big-data-definition.htm
- https://medium.com/existek/big-data-solutions-example-of-the-development-cost-e3d173d97064
- https://dzone.com/articles/redefining-scalability-in-the-era-of-big-data-anal
- https://www.ngdata.com/the-importance-of-scalability-in-big-data-processing/
- https://cloud.google.com/products/calculator/
- https://azure.microsoft.com/fr-fr/pricing/tco/calculator/
- https://awstcocalculator.com/
- https://ec.europa.eu/eurostat/statistics-explained/index.php/Cloud_computing_-_statistics_on_the_use_by_enterprises#Methodology_.2F_Metadata
- http://blog.buyq.org/2017/07/get-to-know-tco-why-the-wisest-purchases-dont-always-come-at-the-lowest-cost/
- http://www.scilogs.fr/complexites/la-puissance-de-la-blockchain/