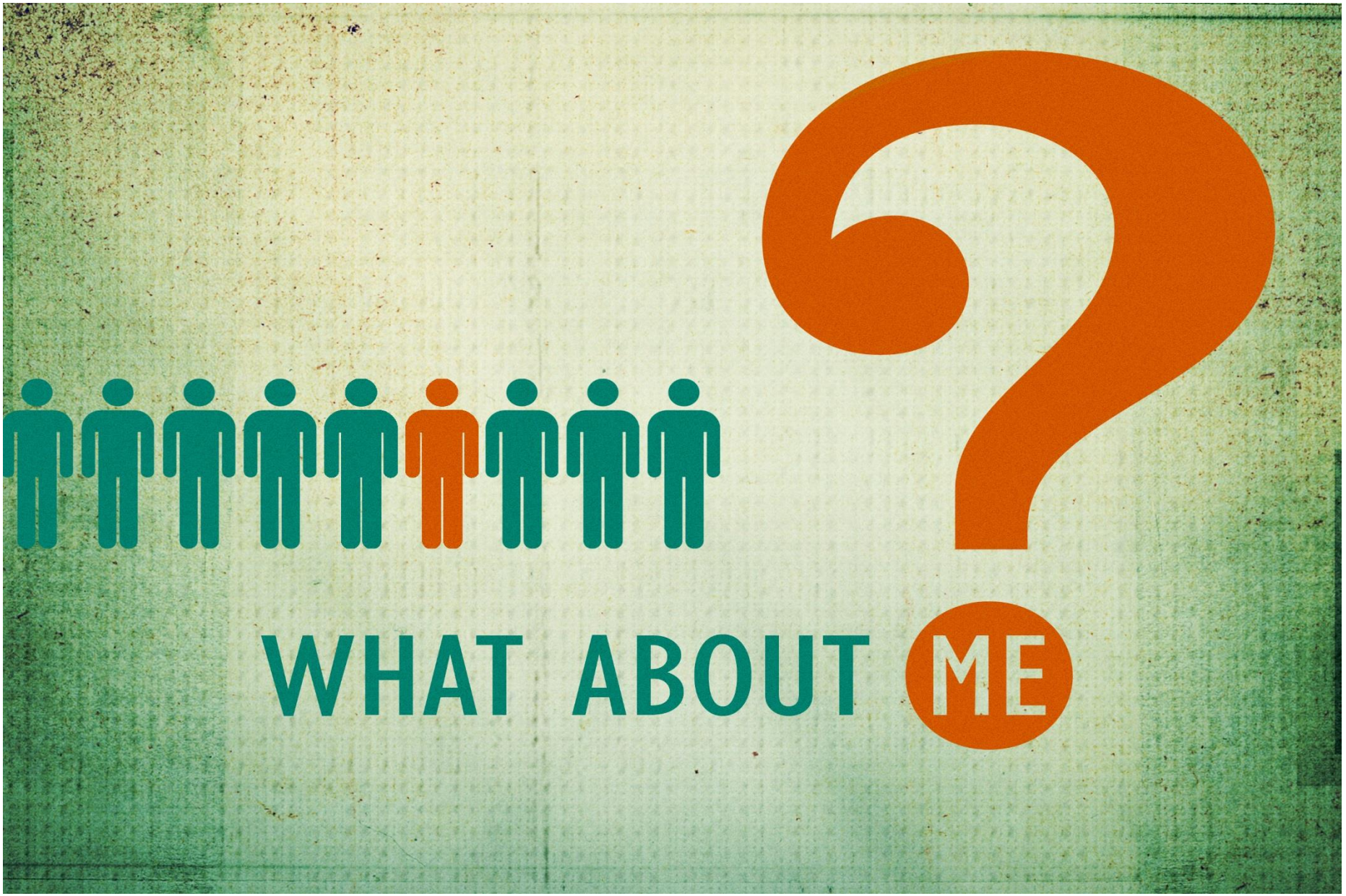
Introduction & concepts Big Data

Khamprasit LANPHOUTHACOUL, Responsable de Centre de Compétences OAB

UFR IM²AG , St Martin d'Hères – le 22 octobre 2015

Business Services



Business Services

orange™

Big Data : Facts

Sources : McKinsey et IDC, 2011/2013
* Trillion = 1000 Milliards

30x

Predicted growth

Of global data generated annually by 2020

28 M$

2016 market shares

(source IDC 2013)

1,2

Zetabytes

(1ZB = 1 Trillion* GB) of data created globally in 2010

31,7%

Worlwide Average Annual Growth Rate

2011-2016, on technologies & big data services

235

Terrabytes

of New Data collected by the US Library of Congress in 04/ 2011

Business Services

orange™

la plupart des entreprises sont convaincues par l'impact de la data intelligence

64% des entreprises pensent que les big data bouleversent les frontières traditionnelles du business

58% s'attendent à devoir faire face à des start-ups nées de la data

Big & Fast Data: The Rise of Insight-Driven Business, FreshMinds for Capgemini & EMC2, march 2015

la data intelligence : comment et pour faire quoi ?

- identifier
- extraire
- stocker
- croiser
- analyser

quelques exemples …

- éviter aux chaudières de givrer en croisant les données des chaudières (issues du client) et les données météo
- scorer le comportement à risques des automobilistes en croisant les données des boitiers sur les véhicules (issues du client) et les données météo, cartographie..
- détecter la fraude des assurés en croisant les données (issues du client) et un applicatif d'Emotional Analytics
- optimiser ses stocks en croisant les données de son SI (issues du client) et les données météo
- exécuter un service de recommandation en croisant des données de comportement du consommateur (issues du client) et des données externes

la data intelligence est un puissant vecteur de transformation pour :

saisissez les opportunités d'exploration de nouveaux territoires tout en maîtrisant les risques

Direction de l'Innovation

gagnez en performance pour répondre à vos enjeux opérationnels

Directions métiers

adaptez en continu vos environnements IT pour répondre - avec plus d'agilité aux besoins métiers

DSI

très peu d'entreprises réussissent à l'implanter durablement au sein de l'organisation

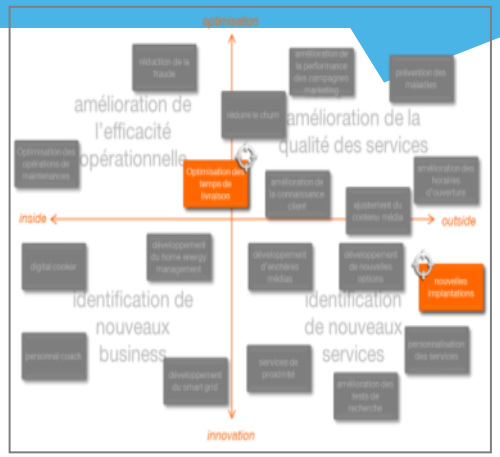
13% des entreprises réussissent implanter durablement les projets big data démarrés au sein de l'organisation

60% des projets d'analytics seront sanctionnés par des échecs d'ici 2017

selon Gartner, Janv 2015

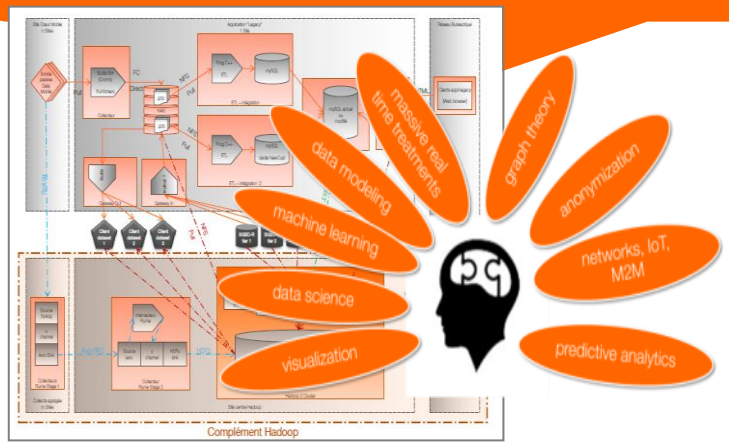
réussir un projet de Data Intelligence nécessite souvent de passer 4 barrières majeures

1 quel(s) cas d'usage(s) choisir ?



2 comment déterminer et atteindre les sources de données nécessaires ?

3 comment réunir les bonnes compétences, ainsi que les bonnes technologies ?



4 comment s'assurer du retour sur investissement ?

¥€$

Big Data : une définition

- Pas de définition « standard » encore

- Wikipedia dit :
  - « expression anglophone utilisée pour désigner des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information. L'on parle aussi de *datamasse* en français par similitude avec la biomasse. »

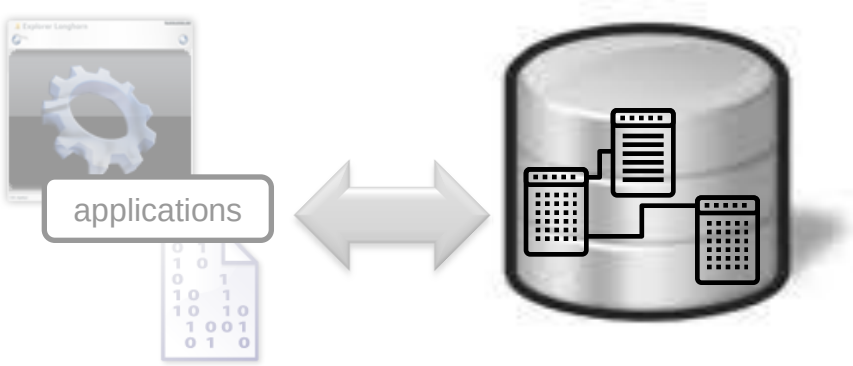
L’enjeu du Big Data

- Valorisation des données (les 5 V du Big Data)
  - Volume x Variété x Vélocité x Visualisation = Valeur
- Explosion des données par la croissance des usages numériques
  - Mobilité (usage permanent), mode SaaS, réseaux sociaux
- Multiplication des supports
  - Données numériques, texte, image , voix, son, vidéo
- Besoin de réactivité
  - Optimiser le temps de réalisation et le recours à l’IT (cf mise en œuvre BI d’entreprise, BI d’équipe, Self BI)
- Rendre visible l’ensemble des données à l’ensemble des utilisateurs
  - seul 30% des consommateurs d’analyses ont accès aux outils
  - Accès au données quelques soit la source, mais avec une gestion des droits

Le Big Data : alternative au stockage traditionnel en BDD

Base de données

Enregistrement suivant un modèle



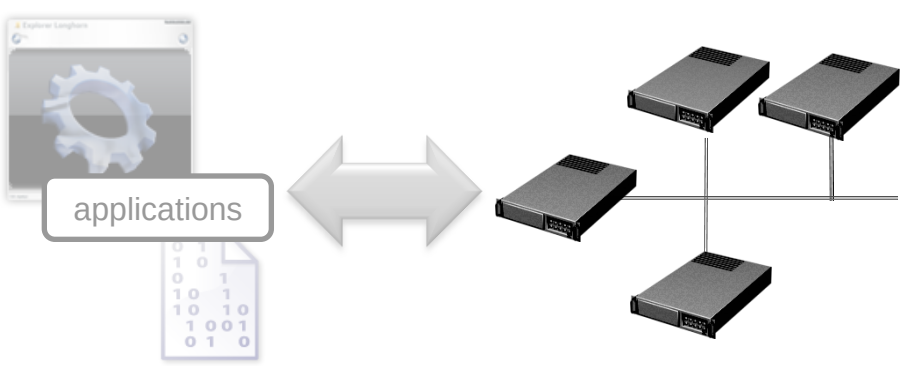
Adaptation en entrée des données brutes à un modèle prédéfini

Relations logiques gérées dans le modèle

Accès aux données par le langage structurée (SQL)

Big data

stockage des données brutes



Stockage des données brutes dans un cluster de serveurs, HDFS, enregistrement des formats d'entrée

Relations logiques gérées dans les applications

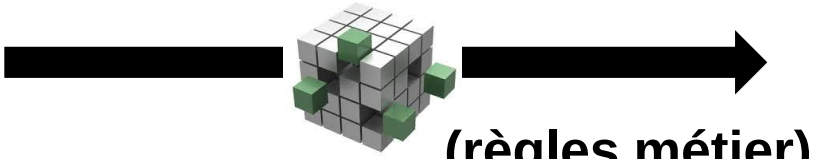
Accès avec langage NO SQL (Not Only SQL), requêtes MapReduce

Business Services orange

Big Data
de la BI traditionnelle à l'analyse prédictive

Données massives & multistructurées

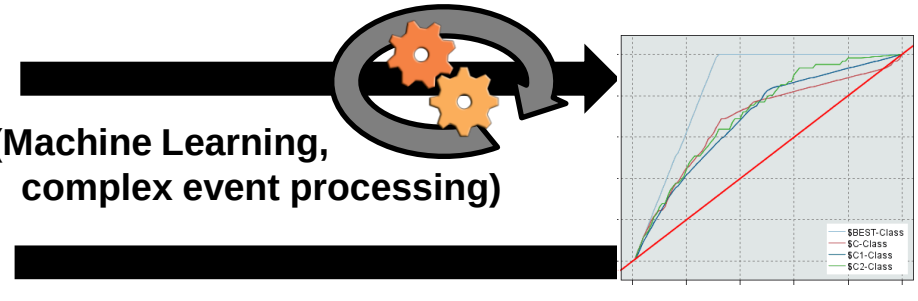
Analyse descriptive multidimentionnel



(règles métier)

KPI (valeur constatée)

Analyse prédictive



(Machine Learning, complex event processing)

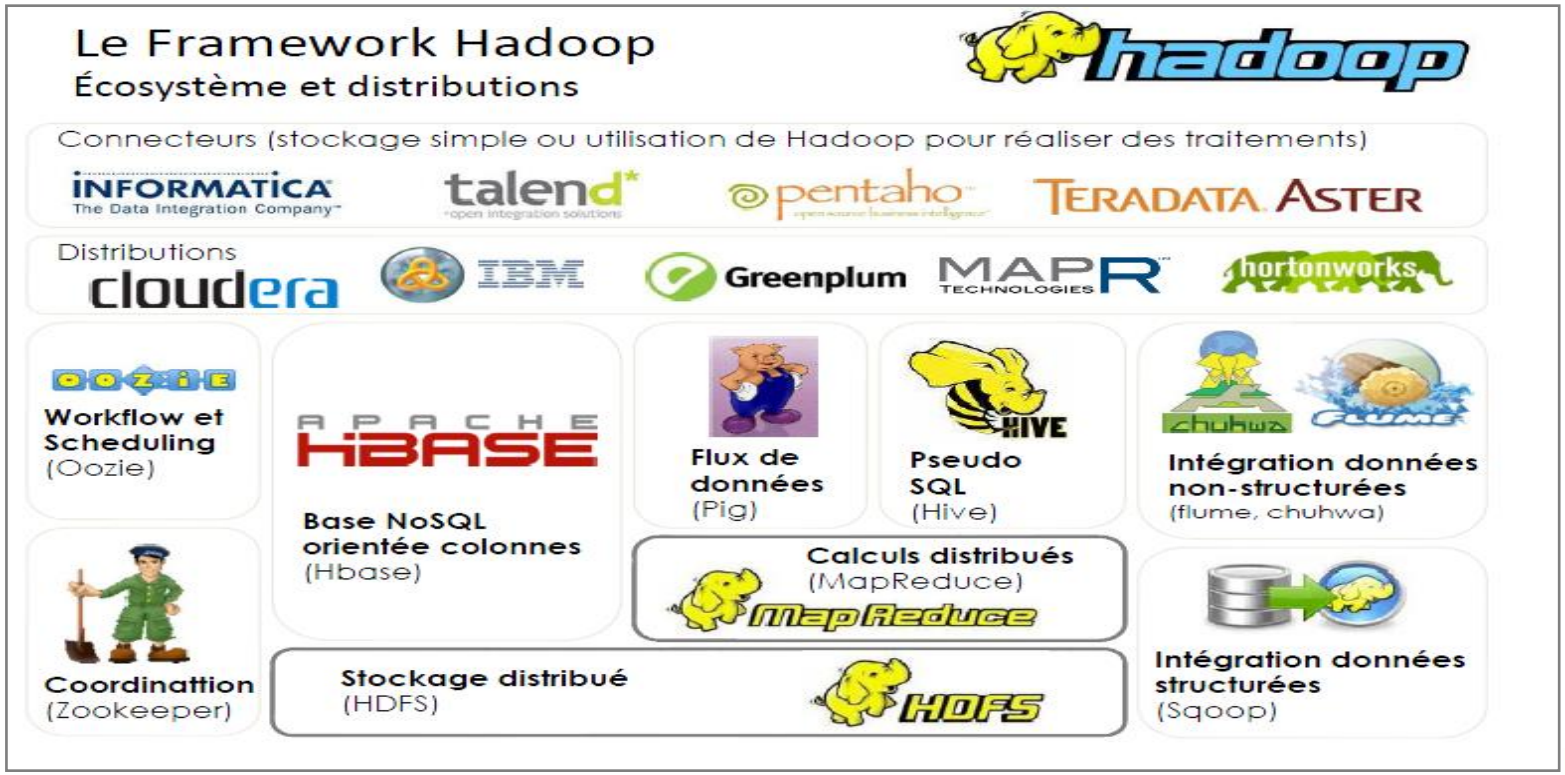
(modèle statistique)

Score (probabilité)

Business Services orange

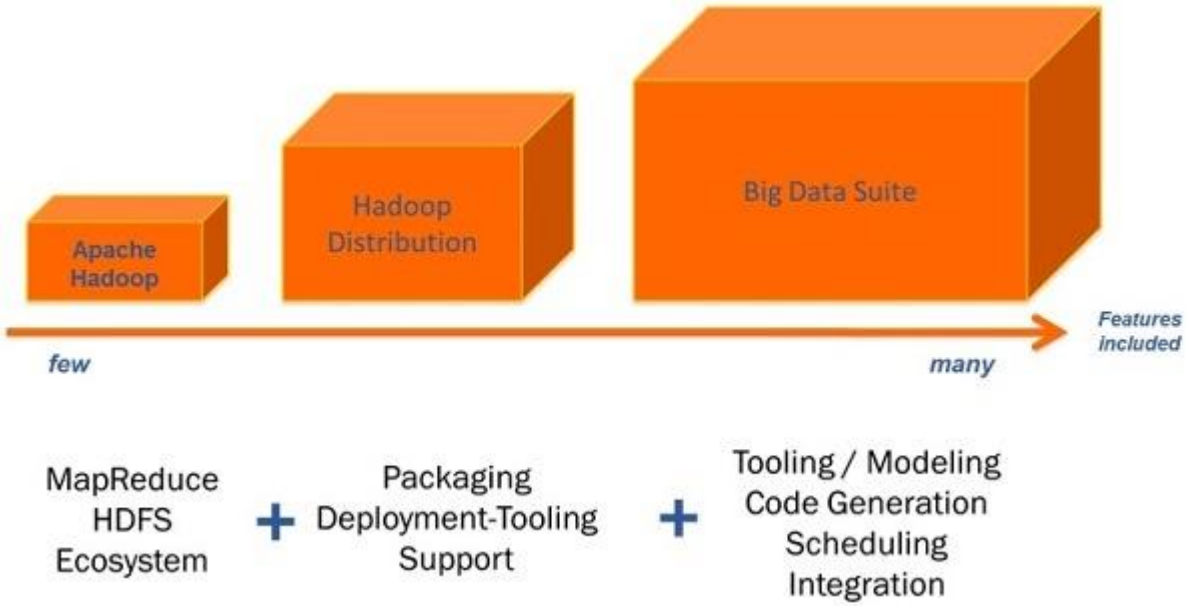
Le cluster

- Infrastructure :
  - Un réseau efficace (débit, disponibilité, ..)
  - Des nœuds dédiés au stockage : accès rapide au stockage
  - Des nœuds dédiés à la gestion du cluster
  - Des nœuds avec de la capacité CPU pour les traitements
- Pile logicielle



Hadoop, distributions et suites

- Il existe différentes alternatives de plates-formes Hadoop.
  - version open source de Apache
  - différentes distributions proposées par différents fournisseurs
  - ou package Big Data d'un éditeur
- Il est important de comprendre que chaque distribution contient Apache Hadoop, et que presque chaque package Big Data contient ou utilise une distribution



Hadoop (v2)

- Projet Open Source : pas de support commercial

- **Hadoop Common**: les utilitaires communs qui supportent les autres modules d'Hadoop.
- **Hadoop Distributed File System (HDFS)**: un système de fichiers distribués qui fournit un accès haut-débit aux données de l'application.
- **Hadoop YARN**: un framework pour la planification des tâches et la gestion des ressources du cluster
- **Hadoop MapReduce**: un système basé sur YARN pour le traitement parallèle des gros volumes de données.

Nouveaux usages : Directions métiers & DSI peuvent tirer partie des données et analytiques pour optimiser leurs performances opérationnelles

DSI

optimiser l'architecture de données

anticiper & optimiser les coûts de stockage des données

collecter & traiter de nouvelles sources de données

adresser & anticiper les besoins des directions métiers

traiter les données en (quasi) temps réel

permettre d'explorer les données en mode 'test & learn'

Directions métiers

monitorer

explorer

analyser

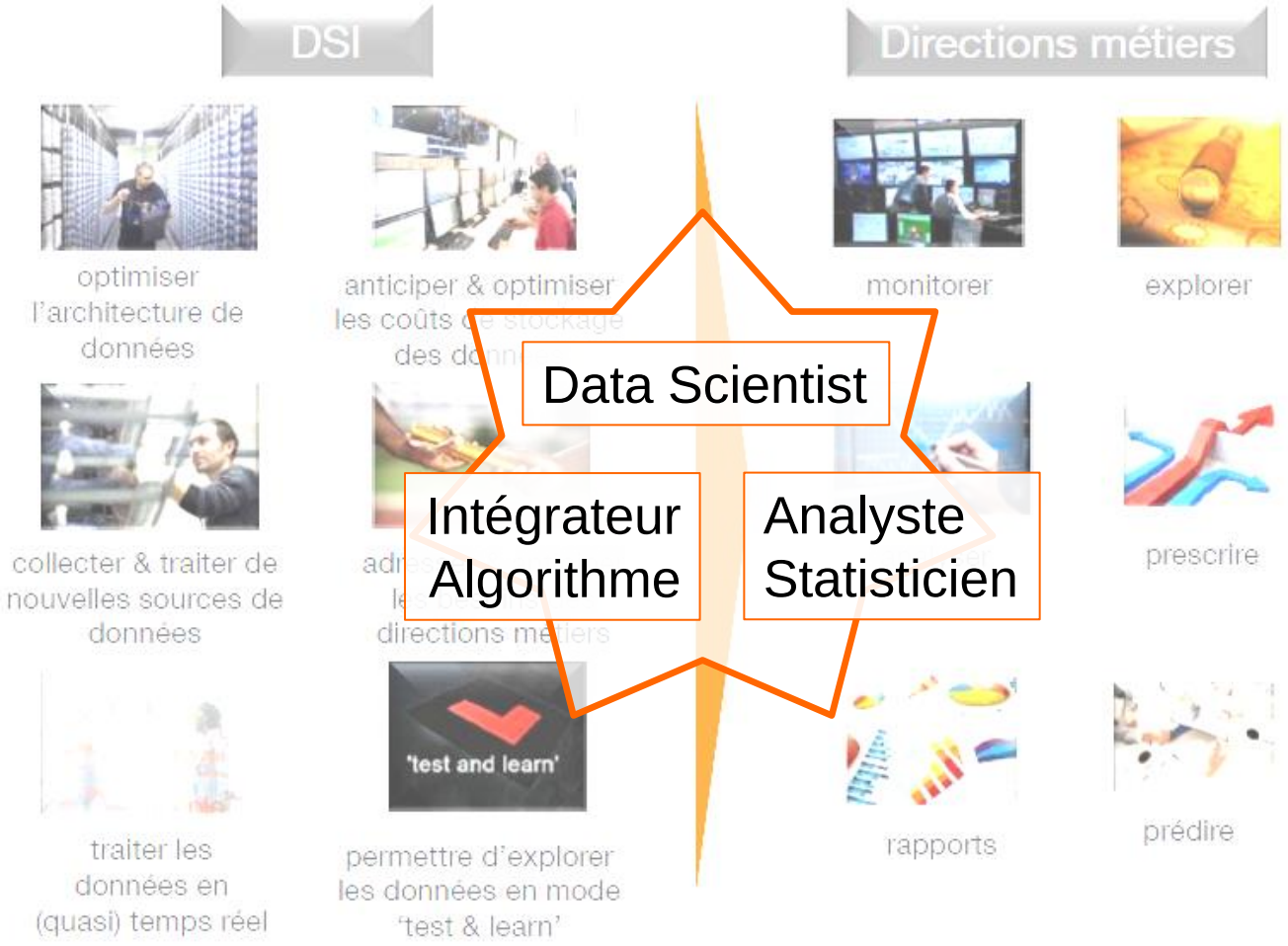
prescrire

rapports

prédire

Cas d'usages réels + Patrimoine de données + Compétences & moteurs de traitement de données & analytiques

Emergence de profils métiers à l'interface



Cas d'usages réels + Patrimoine de données + Compétences & moteurs de traitement de données & analytiques

BIG, FAST et OPEN Data

- FAST Data : Analyse temps réel
  - A partir de résultats d'analyse en temps différé, définir des traitements faits sur les flux de données
  - Génération d'indicateurs générés en continu
  - Génération d'alertes
  - Déclenchement de traitements

- OPEN Data : la donnée en accès libre
  - Quelles informations mettre à disposition, sous quelle forme (nature de la donnée), avec quelle fraicheur (mise à jour)
  - Quelles utilisations : étude sur les données seules ou association avec des données privées pour enrichir l'information, voire la monétiser ?

Parenthèse : le droit dans tout ça

- ATTENTION au droit de la personne
  - La législation porte sur le stockage et les traitements effectués
  - Les traitements statistiques ne doivent pas être réversibles, afin de ne donner aucune indication qui puisse être associée à un individu.
  - Dépend du code pénal
  - http://www.cnil.fr/vos-obligations/vos-obligations/
  - Europe
    - Convention 108 du Conseil de l'Europe
    - EU
      - Le règlement (CE) n° 45/2001 du Parlement européen et du Conseil du 18 décembre 2000 (uniquement pour les Instituions Européennes)
      - Directive 95/46/CE sur la protection des données personnelles (ne concerne pas le Troisième pilier de l'Union européenne, c'est-à-dire les fichiers de police de la coopération policière et judiciaire en matière pénale)
      - Directive 2002/58 dite "e-Privacy" directive.
      - Directive 2006/24/CE sur la conservation des données, etc.
  - France
    - La Loi informatique et libertés (LIL) de 1978, modifiée en 2004, qui prévoit, entre autres, la création de la CNIL s'inscrit dans ce cadre.
    - La loi pour L'Économie Numérique (LEN) de 2004 transpose également des mesures de la "e-privacy" directive.

Le Big Data : ce qui est à retenir

1. La volumétrie n'est pas le fondement du Big Data, mais plus la capacité à adapter l'entité de stockage

2. Cette capacité d'adaptation répond aux besoins de flexibilité et de réactivité des systèmes d'information

3. Ces technologies permettent de répondre plus efficacement aux besoins de valorisation des données (fourniture d'informations accessibles)

Usage Big & Fast Data : gestion des boitiers (LiveBox) foudroyés

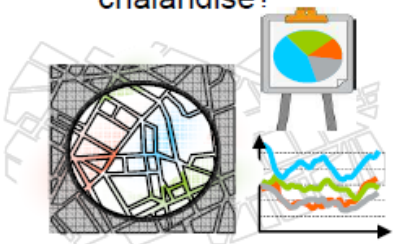
Usage Big Data : optimisation des couts

Mise en œuvre d'un stockage temporaire pour analyse circonstancielle

Usage Big Fast Data : création de services

Zone → Audience

Quels profils se trouvent dans ma zone de chalandise?



Déterminer la segmentation (socio démo, comportementale, intérêts etc..) de la population présente sur une zone de chalandise, et son évolution dans le temps

Path → Audience

Quels profils empruntent cette route ?

Déterminer la segmentation (socio démo, comportementale, intérêts etc.) de la population présente sur un trajet spécifique, et son évolution dans le temps

Audience → Zone

Où puis-je trouver ma cible ?

Déterminer les zones où se trouve un profil donné (socio démo, comportemental, intérêts etc.)

Monétisation des informations issues de l'analyse des données collectées.

Usage Big Data : Challenge Data for Development, D4D

Concours open innovation dans le domaine du big data.
Des échantillons de données du réseau mobile rendus anonymes ont été extraits conjointement par Sonatel et Orange, en accord avec les recommandations de la Commission des données Personnelles du Sénégal, ont été communiquées à plus de 150 laboratoires de recherche internationaux parmi les 250 universités inscrites.
Près de 60 projets ont été soumis pour la compétition finale.
Le challenge a été organisé autour de cinq thèmes principaux : la santé, l'agriculture, le transport/urbanisme, l'énergie et les statistiques nationales.

Merci, questions ?

Business Services

orange™