

Big Data : noSql, Map&Reduce, RDBMS

Quoi utiliser et quand ?

- Romain Chaumais – roman.chaumais@ysance.com
Directeur du pôle Business Intelligence

Les origines du phénomène Big Data



Big Data ?



1 203 900 000

+



164 500 000



1 368 400 000

**Le nombre de transactions aux péages
des autoroutes de France en 2010**

41,5M de véhicules / 84,1Mrd de Km

Evolution d'un péage autoroutier



Analogique

Anonyme

Product Centric

Vers un télépéage et le Big Data



Numérique

**Historisé et
analysé**

**Customer
Centric**



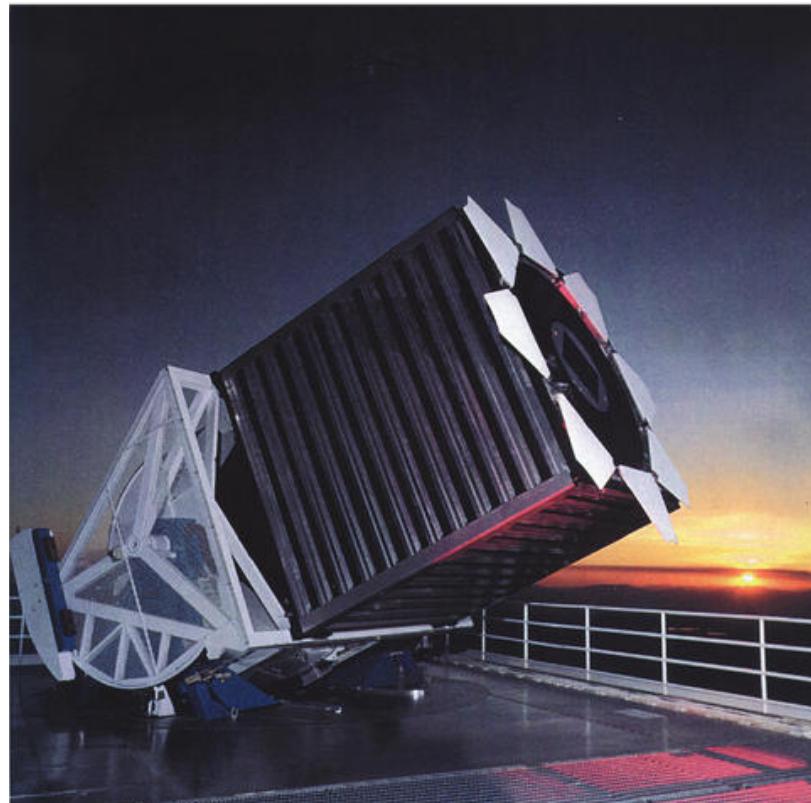
L'ampleur du déluge de données

Web-Scale



- Plus de 1 million de transactions clients / heure
- 2 500 To de données en base
- 500 millions de visiteurs / jours
- 50 milliards de photos stockées
- 90 milliards de contenus partagés chaque mois
- 7,2 milliards de pages vues / jour
- 88 milliards de recherches / mois
- 20 Po de données traitées / jour

Volume de données



Sloan Digital Sky Survey - Nouveau Mexique

Démarré en 2000

En quelques semaines, a collecté plus de données que dans toute l'histoire de l'astronomie

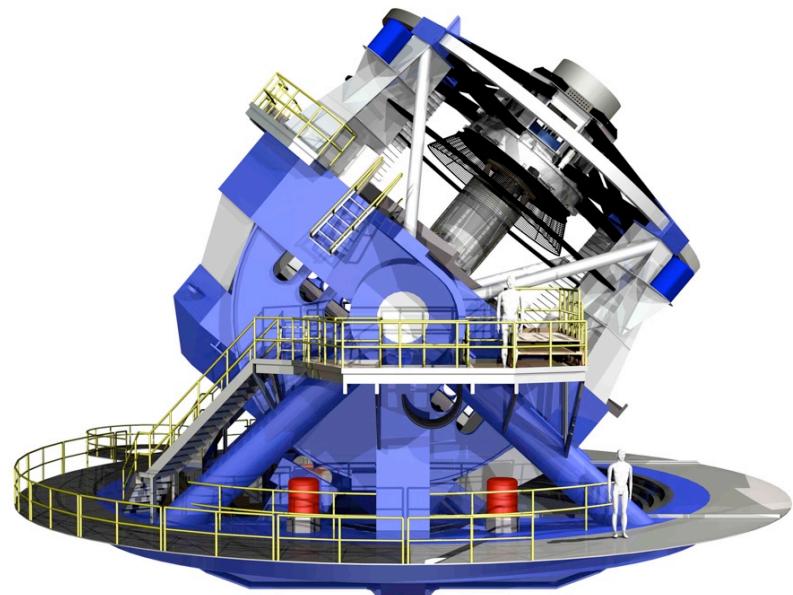
A ce jour, a généré plus de 160 To de data archivées

Volume de données

Démarrera en 2016

Génèrera plus de 160 To de
data en ...

4 jours !!

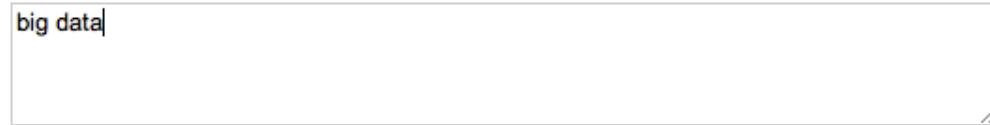


Large Synoptic Survey Telescope - Chili

Raw Data ?

Traduire du texte, des pages Web ou des documents

Saisissez du texte ou l'URL d'une page Web, ou [importez un document](#).



Langue source : ↗
Langue cible :

Traduction (anglais > français)

 grandes données

Google documents Fautes d'orthographe

Fichier Modifier Afficher Insertion Format Tableau Outils Aide



Fautes d'orthographe



Raw Data = Plus d'explorations

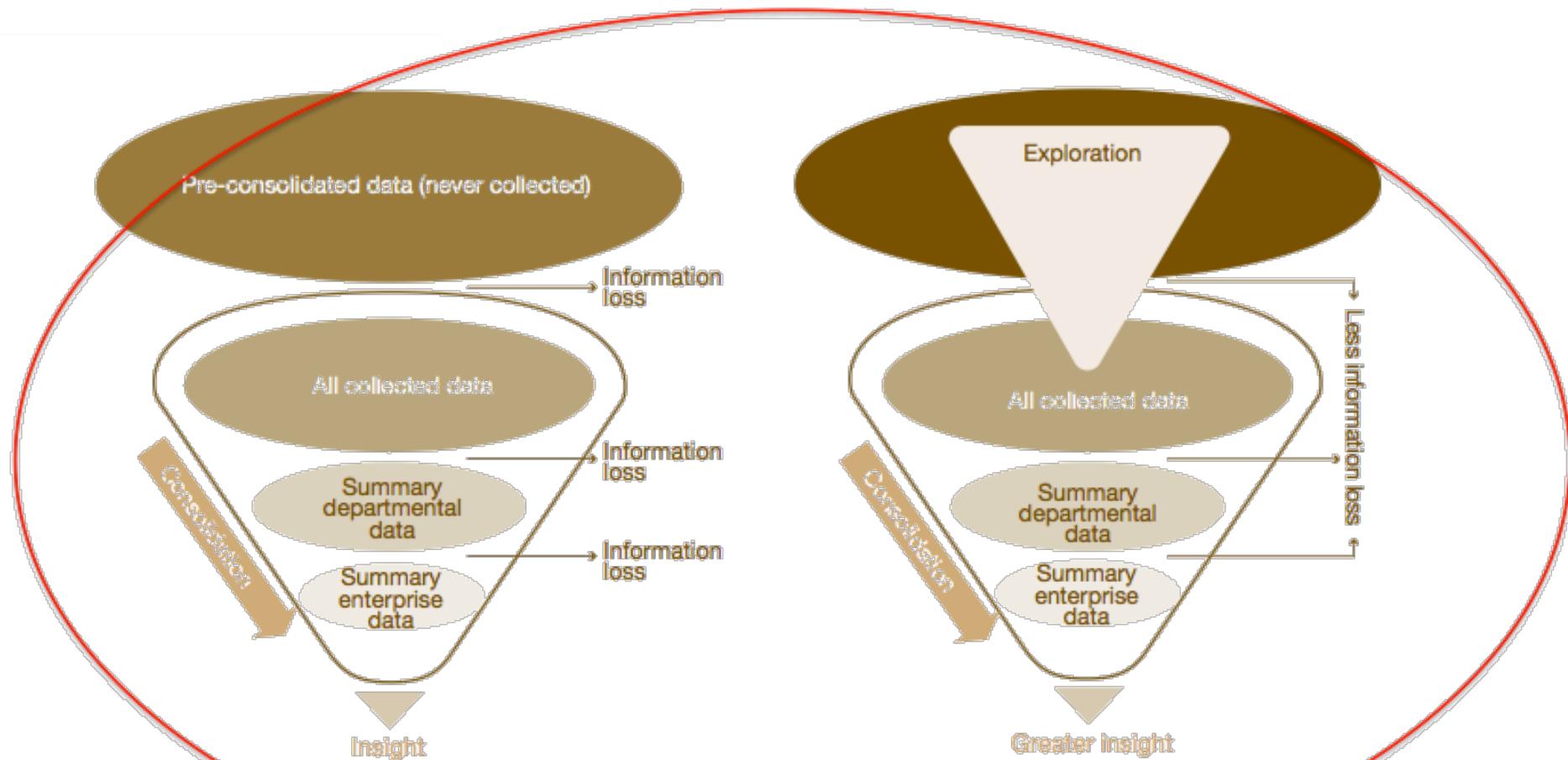


Figure 4: Information loss in the data consolidation process

Source: PricewaterhouseCoopers, 2010



Qui aujourd'hui produit et consomme la donnée ?

Ysance
simplifions les projets informatiques



Human generated data



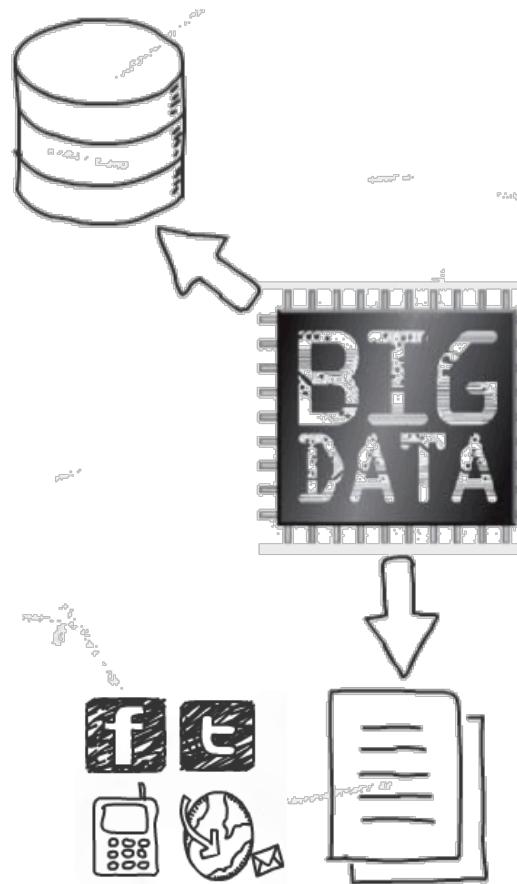
Machine generated data

Big Data : Proposition de définition

La règle des 3 V

Volume

Collecte & stockage d'un grand volume de données



Vélocité

Intégration, traitement et restitution en temps limité

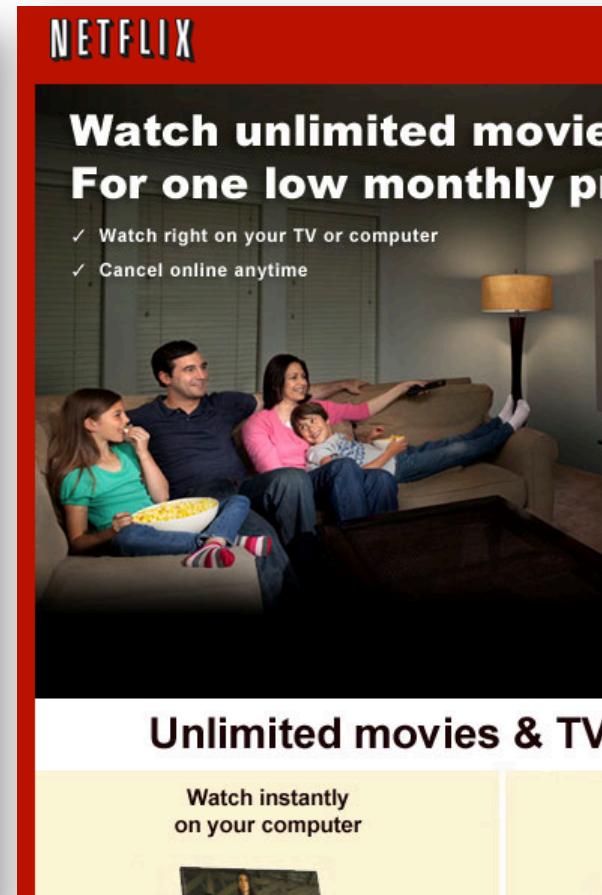
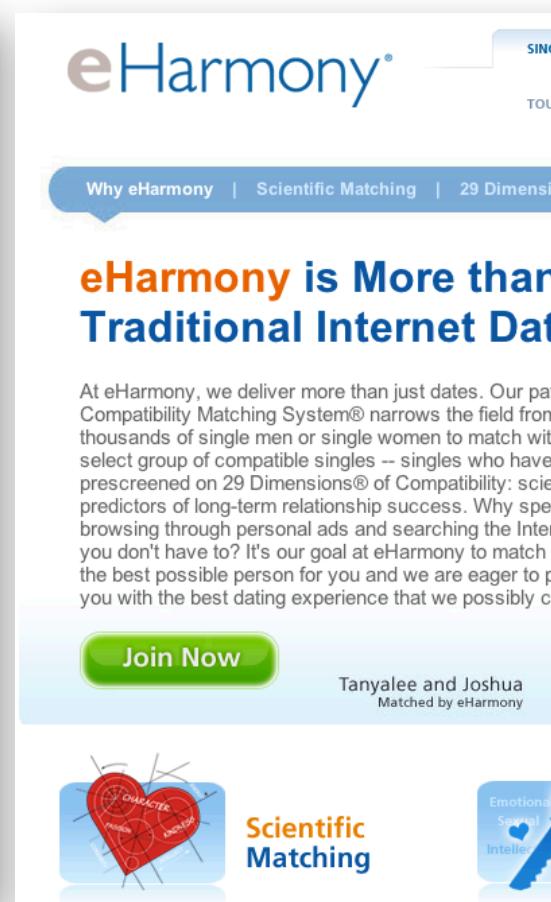
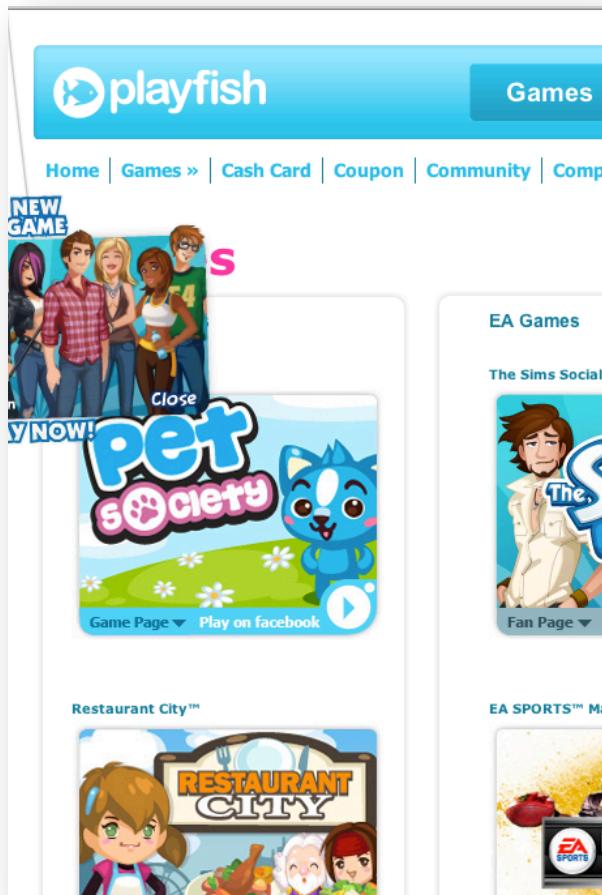
Variété

Données structurées mais aussi non structurées (texte, audio, vidéos, click streams, log files, etc.)

Exemples de projets Big Data



Exemples Map & Reduce



Exemple projet Big Data Ysance



Casual Game consistant à échanger des objets avec ses amis

- 1,3 millions de joueurs mensuels

Analyse des interactions entre les joueurs et leurs amis

- Analyse du graphe social des joueurs
- Catégorisation des joueurs selon le nombre d'amis jouant aussi à IsCool

Mise en place d'un parcours de jeux personnalisé selon le nombre d'amis

- Objectif : Avoir au moins 10 amis avec qui jouer à IsCool

>> Multiplication du CA

VSC : CHALLENGE TECHNO POUR UNE PROMESSE BUSINESS !

Comment offrir plus d'usage B2C aux clients sur un legacy limité au B2B ?

Des clients qui exprimaient leur mécontentement

« trop d'informations "accrocheuses", décalage avec ce que l'on trouve ensuite. Je trouve rarement un trajet en promotion à partir de ma ville (Tours) »

« J'attendais de pouvoir avoir un tarif très avantageux avec une date libre en complétant simplement départ et arrivée du train ou autre mode de locomotion choisi. »

SÉLECTIONNEZ VOTRE ALLER

Aller le 12/04/2011 entre 06h54 et 10h54 - prix total pour 1 passager

Départ à	06h54	07h24	07h54	08h54	09h54	10h54
A partir de	47.00 €	47.00 €	47.00 €	30.00 €	52.00 €	30.00 €
Durée	01h57	02h00	02h03	01h57	02h03	01h57
Voyez avec						

Flexible Trouvez directement le meilleur prix de la journée. 

 Trains précédents

Trains suivants 



CALENDRIER DES PRIX

- Vos critères de recherche MODIFIEZ VOS CRITÈRES DE RECHERCHE

Départ : Paris Arrivée : Lyon Votre voyage : Aller Simple
1 Passager 2e classe Carte de réduction : Sans carte

- Votre voyage

CHOISISSEZ VOTRE DATE ALLER

< 15 jours précédents Meilleurs prix disponibles 15 jours suivants >

MARS

lundi 28	mardi 29 23.90€	mercredi 30 23.90€	jeudi 31 23.90€
----------	---------------------------	------------------------------	---------------------------

AVRIL

lundi 04 22.00€	mardi 05 22.00€	mercredi 06 23.90€	jeudi 07 23.90€	vendredi 01 25.00€	samedi 02 39.90€	dimanche 03 24.90€
lundi 11 25.00€	mardi 12 25.00€	mercredi 13 22.00€	jeudi 14 22.00€	vendredi 15 25.00€	samedi 16 35.00€	dimanche 17 24.90€
lundi 18 22.00€	mardi 19 22.00€	mercredi 20 22.00€	jeudi 21 23.90€	vendredi 22 25.00€	samedi 23 36.00€	dimanche 24 24.90€
lundi 25 24.90€	mardi 26 23.90€	mercredi 27	jeudi 28	vendredi 29	samedi 30	

CHOISISSEZ VOTRE HORAIRE ALLER

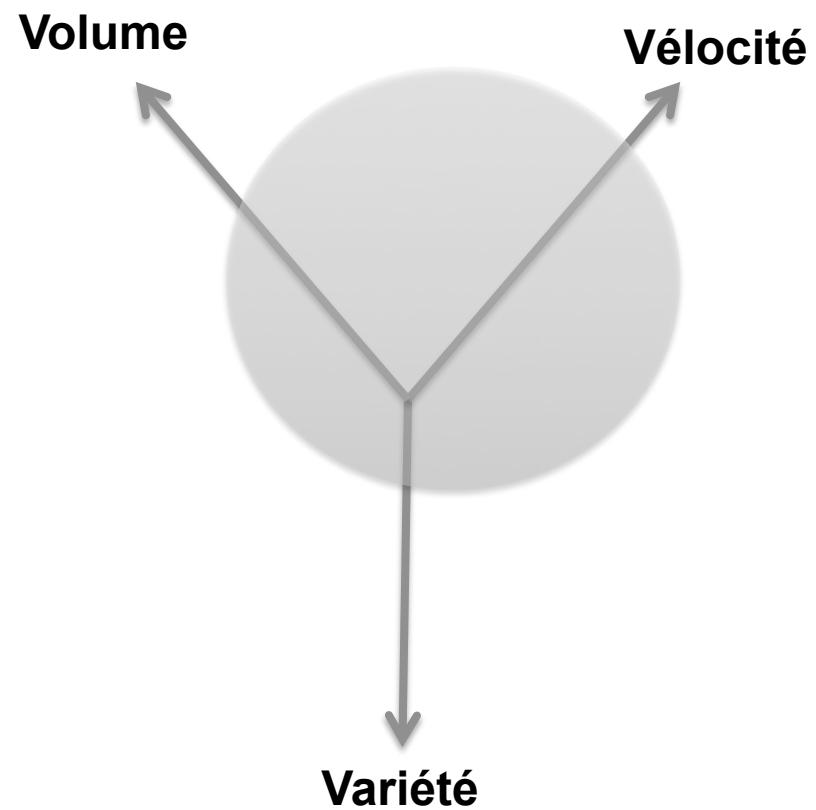
mardi 12 avr. 2011				
◆ Prix	◆ Durée	◆ Horaire	TGV 100% PREM'S	
25.00€	02h18	05h54 Paris Gare De Lyon 08h12 Lyon Perrache	2e classe - 100% Prem's	Billet non échangeable et non remboursable.
25.00€	02h05	05h54 Paris Gare De Lyon 07h59 Lyon Part Dieu	2e classe - 100% Prem's	Billet non échangeable et non remboursable.



Calendrier des prix : projet Big Data

The screenshot shows the Voyages-sncf.com homepage with the 'TRAIN France, Europe' tab selected. A large banner features autumn leaves and the text 'CALENDRIER DES PRIX'. Below it, a search form asks 'Où et quand souhaitez-vous partir?' and includes fields for 'Votre voyage *' (Aller Simple or Aller-Retour), 'Départ : *' and 'Arrivée : *' (both with dropdown menus for 'Gare de départ' and 'Gare d'arrivée'), 'Aller aux alentours du *' (date: 11/10/2011), and 'Retour aux alentours de'.

Voyages-sncf : Calendrier des prix



Comment le Big Data peut-il créer de la valeur ?



En apportant de la transparence et la suppression des silos de données

En simplifiant l'exploration des données, l'expérimentation, la compréhension de phénomènes, l'identification de nouvelles tendances : Data discovery / Data visualization

En permettant la mise en place d'un CRM hyper segmenté tendant vers du One-To-One

En offrant une véritable aide à la décision via des algorithmes riches exécutés automatiquement : Datamining / prédictif

En devant le socle actif de nouveaux produits, services et Business modèle orientés données



Paysage Big Data / No SQL

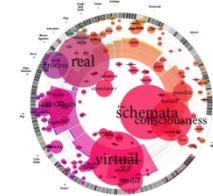


Les problématiques du Big Data

Capturer /
Acheminer



Stocker /
Organiser



Analyser /
Visualiser



Partager /
Sécuriser



Traiter /
Consolider

Des acteurs du Big Data



FuseSource
A PROGRESS SOFTWARE COMPANY



INFORMATICA®
The Data Integration Company™

talend*
*open data solutions

ORACLE®

Microsoft®
SQL Server® 2008 R2
Parallel Data Warehouse

TERADATA®

INFOBRIGHT

calpont

VERTICA

 www.ysance.com

wibiidata
developed by odiago



Neo4j
the graph database



IBM®
N NETEZZA



Windows Azure

amazon
web services™

Cassandra

Yellowfin
MicroStrategy®



MS PowerPivot

Datameer
Powerfully Simple™

DataStax

HADAPT

Zettaset

MAPR™
TECHNOLOGIES
EASY. DEPENDABLE. FAST.

cloudera

splunk™

APACHE
HBASE



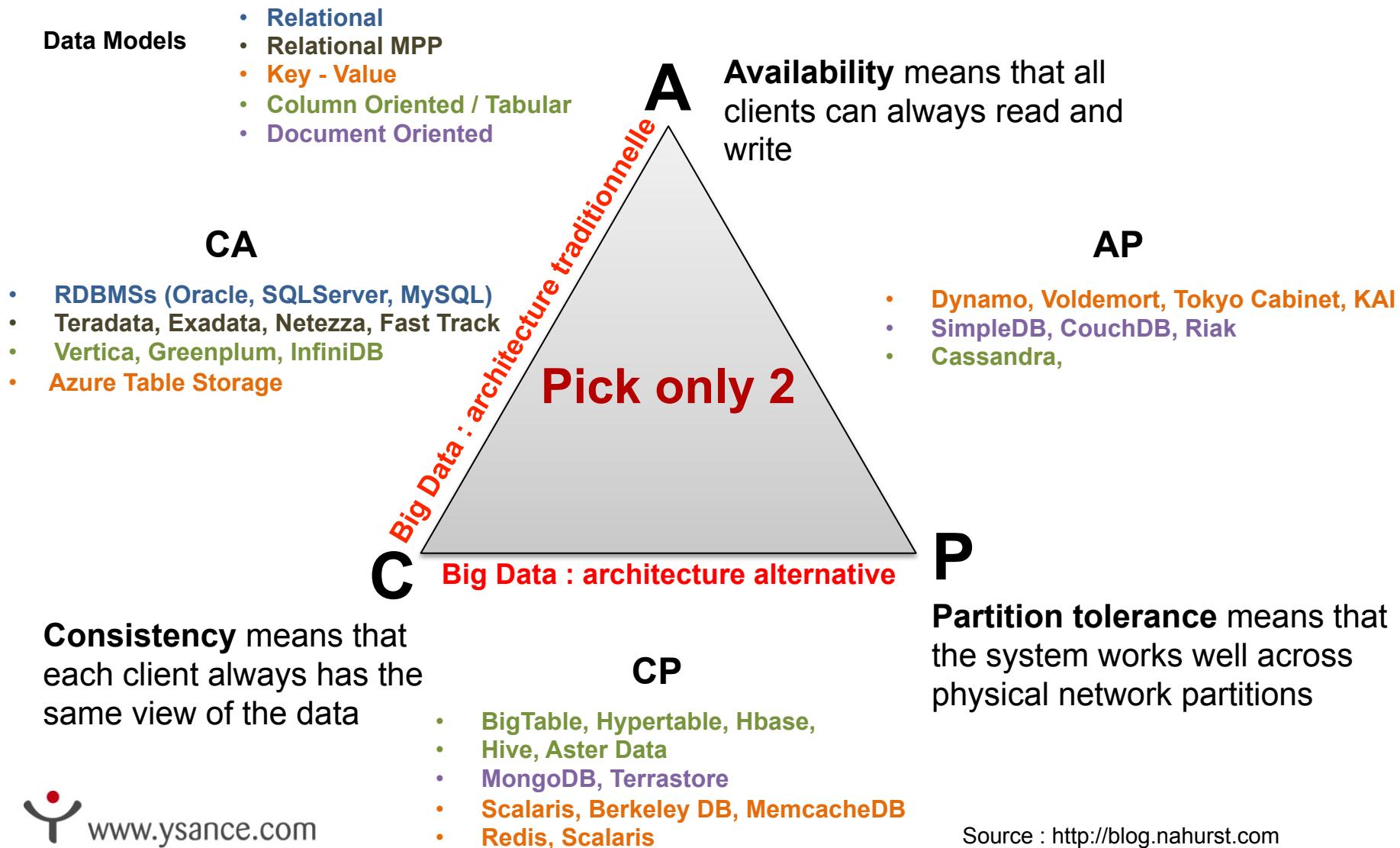
hadoop

hortonworks



Classification des « bases de données »

The CAP Theorem



Classification des moteurs de traitement des données de type Big Data



Moteur spécialisé

INFOBRIGHT
PARACCEL™

calpont
VERTICA

Sybase IQ

INGRES vectorwise



Appliance MPP

ORACLE
EXADATA

Microsoft®
SQL Server 2008 R2
Parallel Data Warehouse

TERADATA

IBM
N NETEZZA

Greenplum



Framework Map & Reduce

hadoop

splunk™

MAPR
TECHNOLOGIES
EASY. DEPENDABLE. FAST.

aster data
big data, fast insights.

IBM InfoSphere BigInsights
Bring the power of Hadoop to the enterprise.

+ - +

Performance / CPU

Volume de données traitées

Structuration des données

- + -

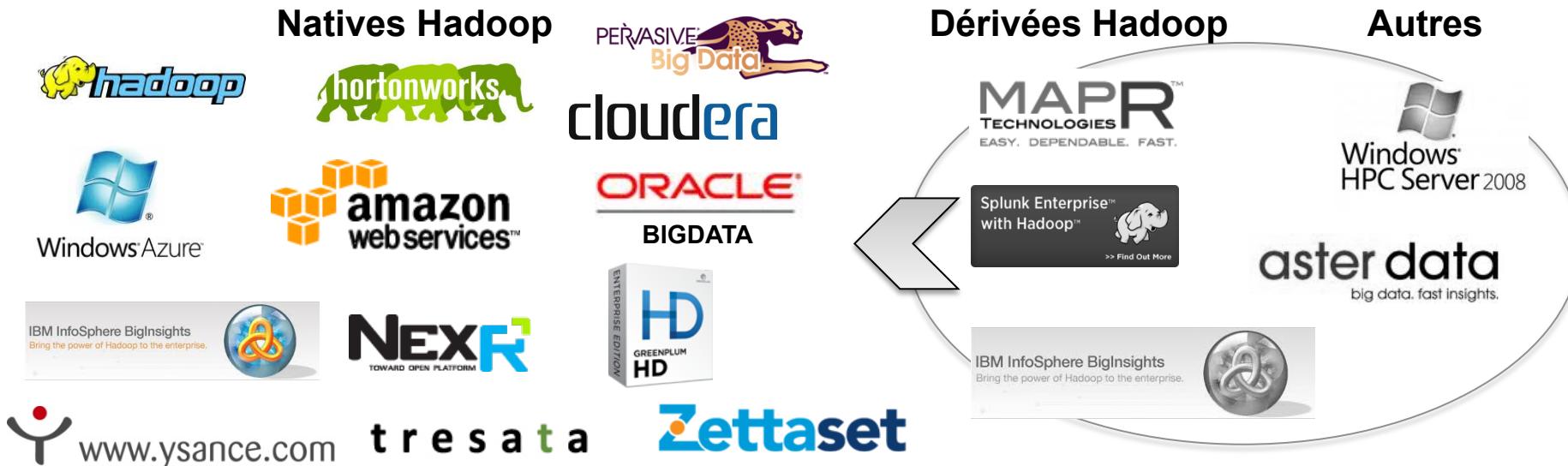
Hadoop : Plate-forme Big Data de référence

Hadoop : HDFS + MapReduce = Stockage + Traitement

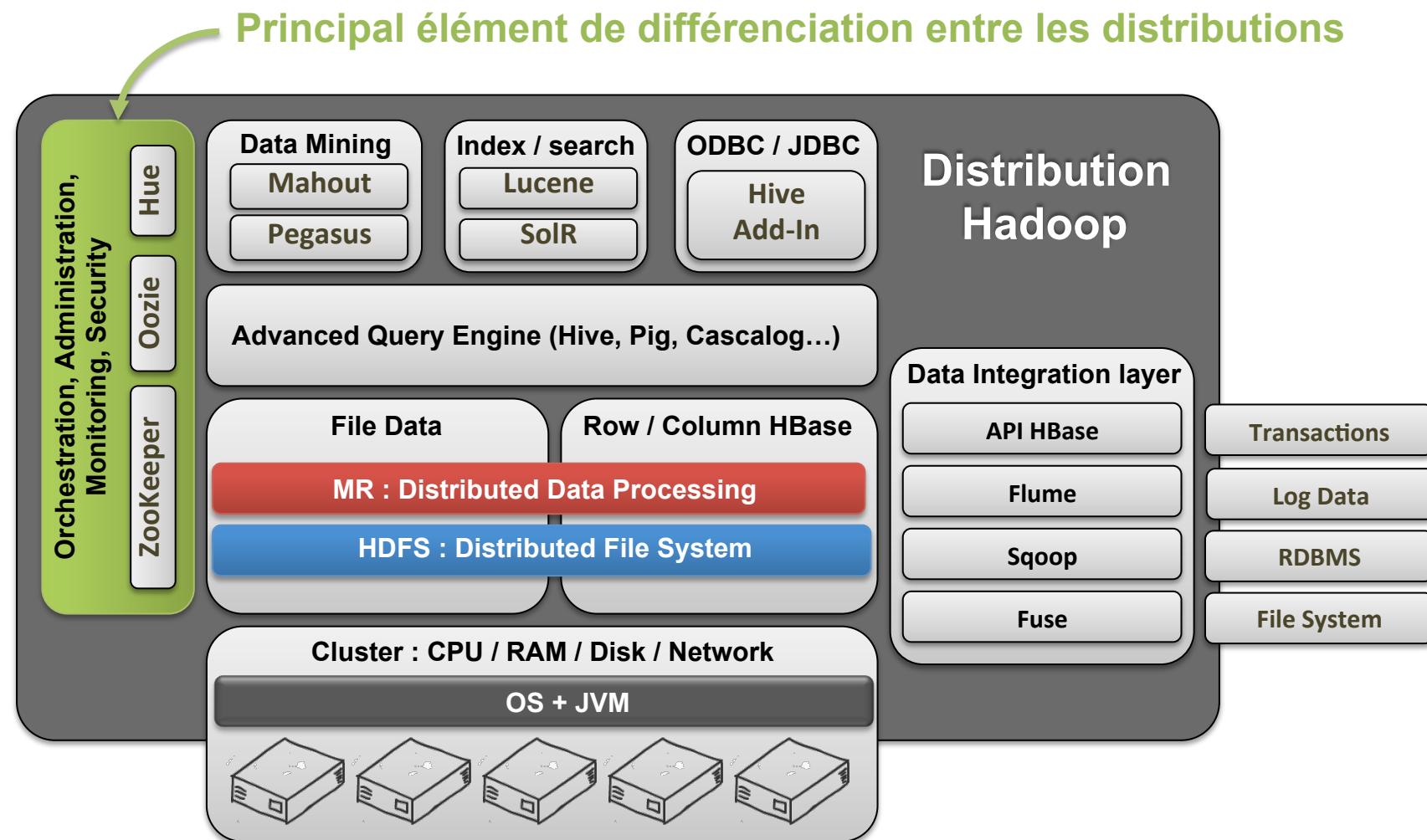
Historique d'Hadoop :

- Inspiré de Google Map Reduce - Première version en 2008 (Yahoo)
- Projet de la fondation Apache. Version 1.0 en janvier 2012.
- Utilisateurs : Yahoo, Facebook, Tweeter, LinkedIn, eBay, etc.

Hadoop : Leader des solutions de MapReduce



L'écosystème Hadoop = Distribution



Architectures Big Data



www.ysance.com

En complément ou en remplacement du SID traditionnel ?

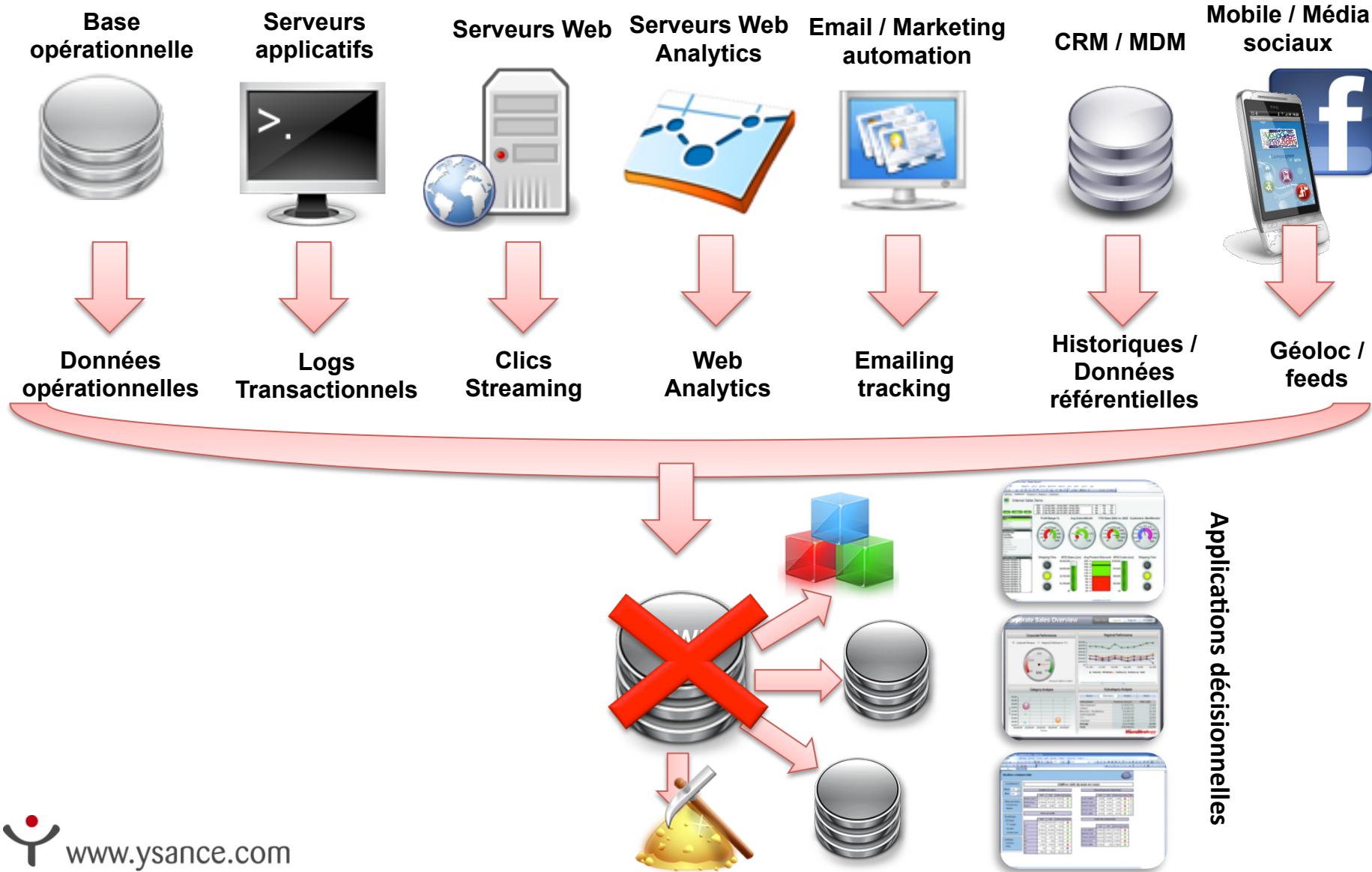


Le Big Data en **complément**
de la BI traditionnelle

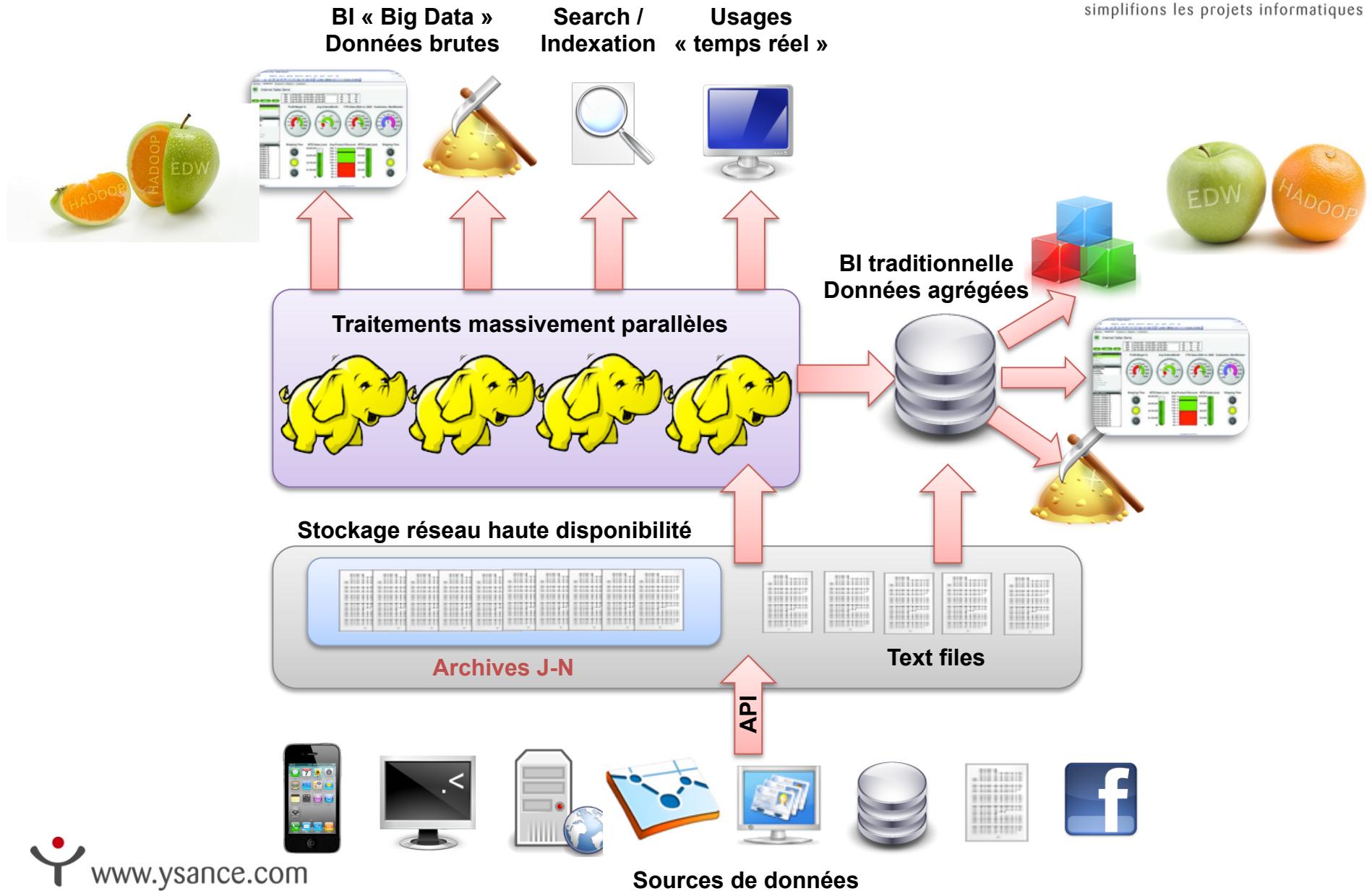
Le Big Data en **remplacement**
de la BI traditionnelle



Comment résoudre cette équation ?



Architecture de type Big Data

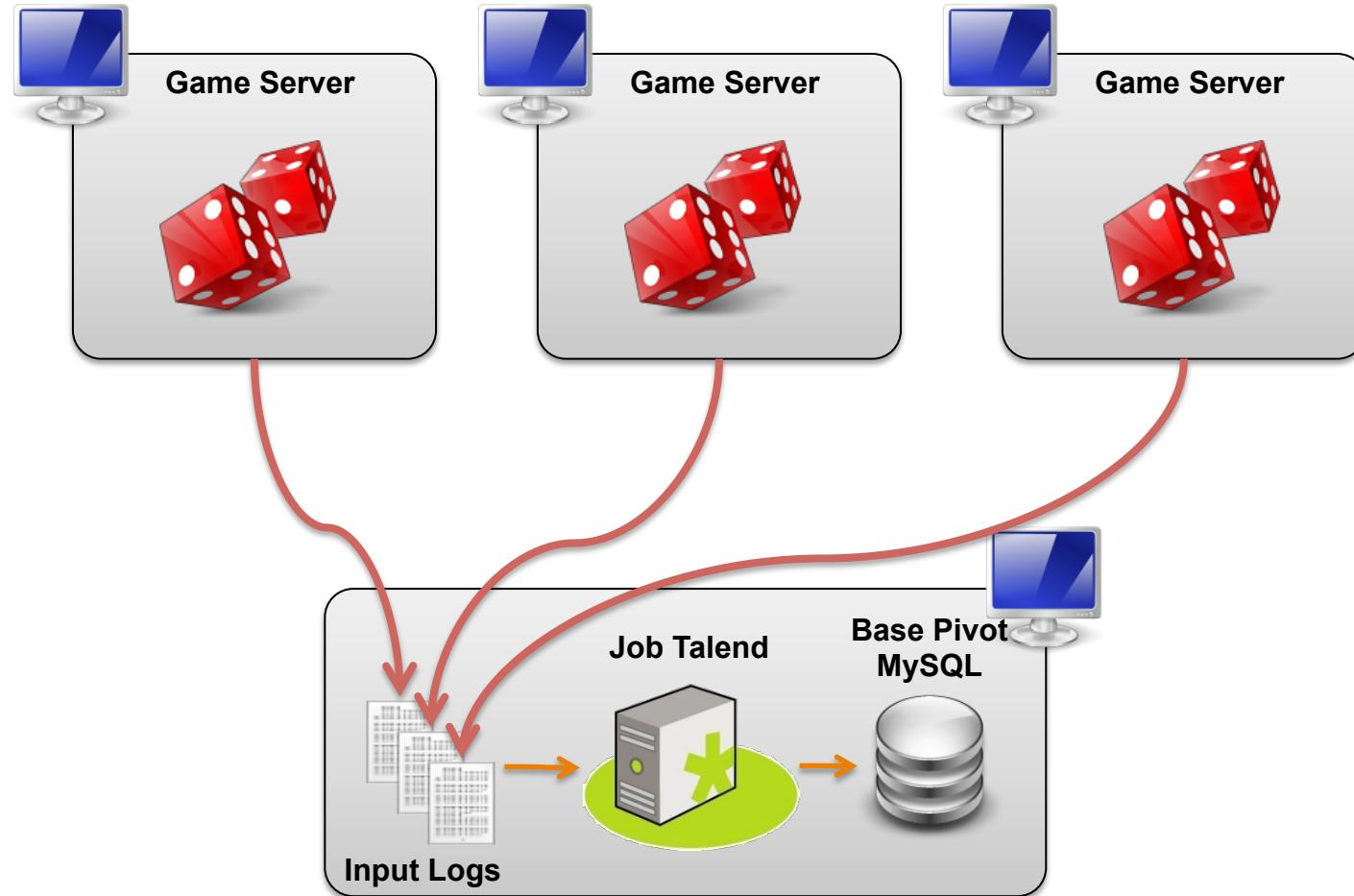


Cas client : Dans le monde du jeu



www.ysance.com

L'existant



Problématique

Problème de montée en charge avec la volumétrie effective

- L'augmentation des volumes de données sources dépasse les prévisions
- Ces volumes engendrent des problèmes bloquants de performance sur les calculs d'agrégation des indicateurs de type « patrimoine »
- Nécessité de désactiver des agrégations (18%) afin de retrouver des performances acceptables
- Les optimisations « standard » ont été réalisées, mais elles ne permettent pas de résoudre les problèmes (qui vont s'aggraver)

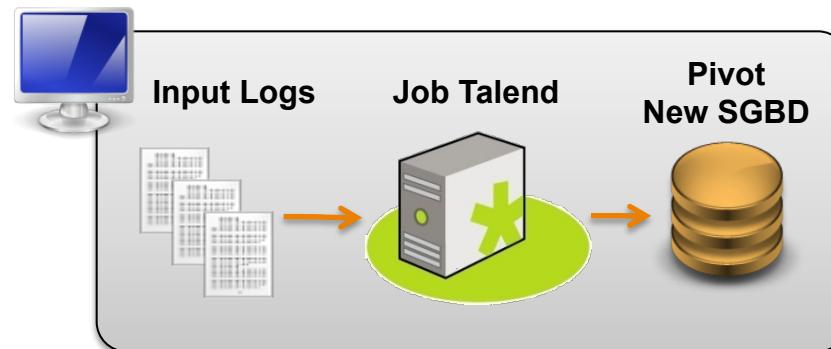
Basculer sur une architecture de rupture type « Big Data »



SGBD colonnes : Objectifs

S'appuyer sur un moteur de base de données taillé pour des grands volumes de données et des usages analytiques

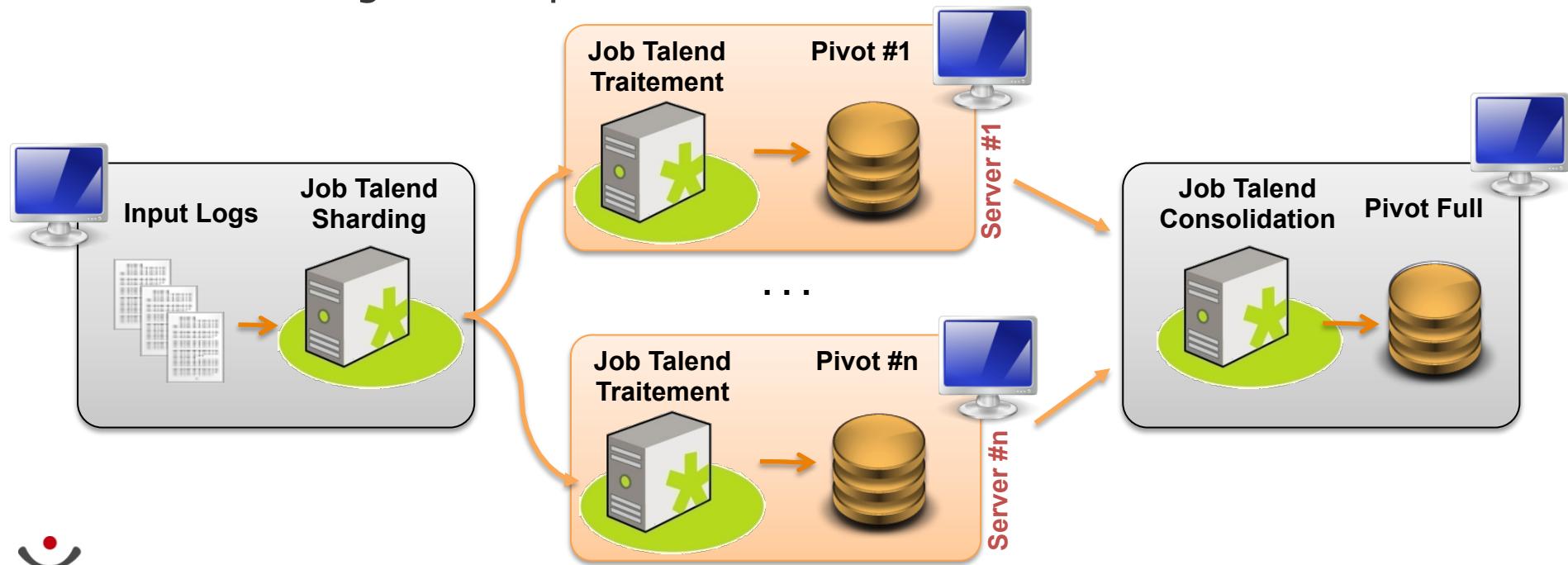
- Le stockage des données en colonnes permet un très haut niveau de compression des données et accélère les requêtes de type agrégation
- Les performances attendues devraient être très supérieures à celles de MySQL



Sharding : Objectifs

Bâtir une plate-forme de type Scale Out

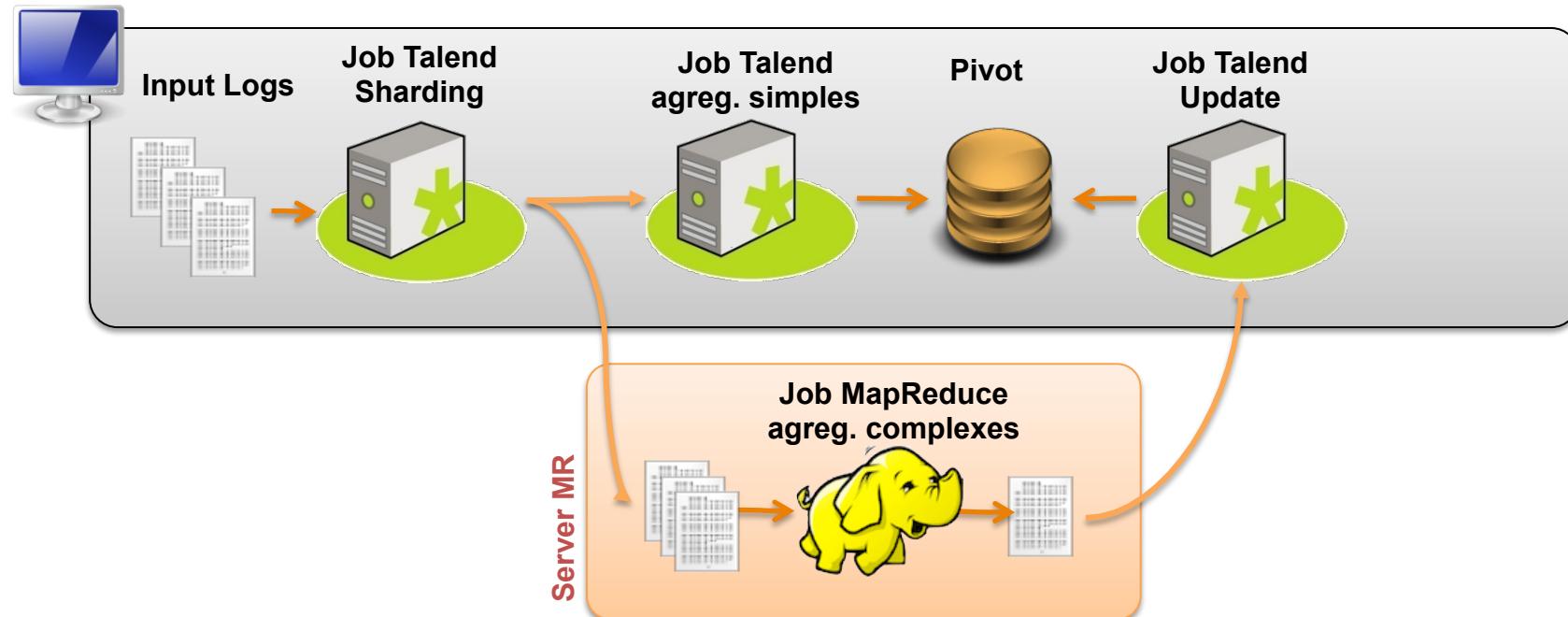
- Gérer la montée en charge consiste à ajouter de nouveau nœuds à la plate-forme
- Chaque nœud est techniquement identique et autonome en données
- Le sharding permet d'augmenter la tolérance aux pannes
- Le sharding est indépendant de la base de données



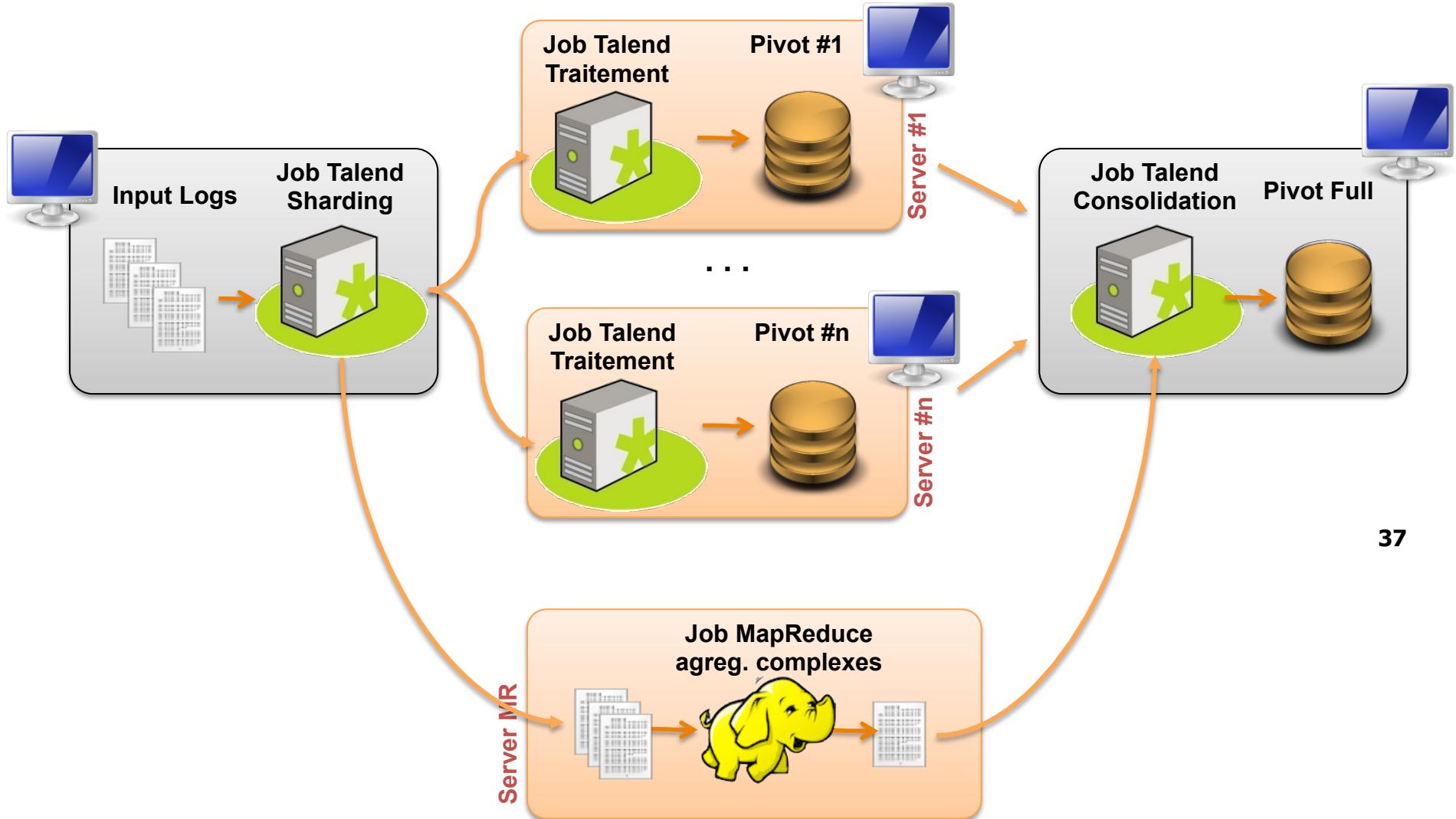
MapReduce : Objectifs

Apport de MapReduce : calculs parallélisés et distribués

- Changer l'approche et la méthode de calcul des agrégats complexes
- Linéariser la montée en charge
- S'appuyer sur le parallélisme et la distribution via AWS



Combination des techniques



37



Merci