Institut
Mines-Télécom

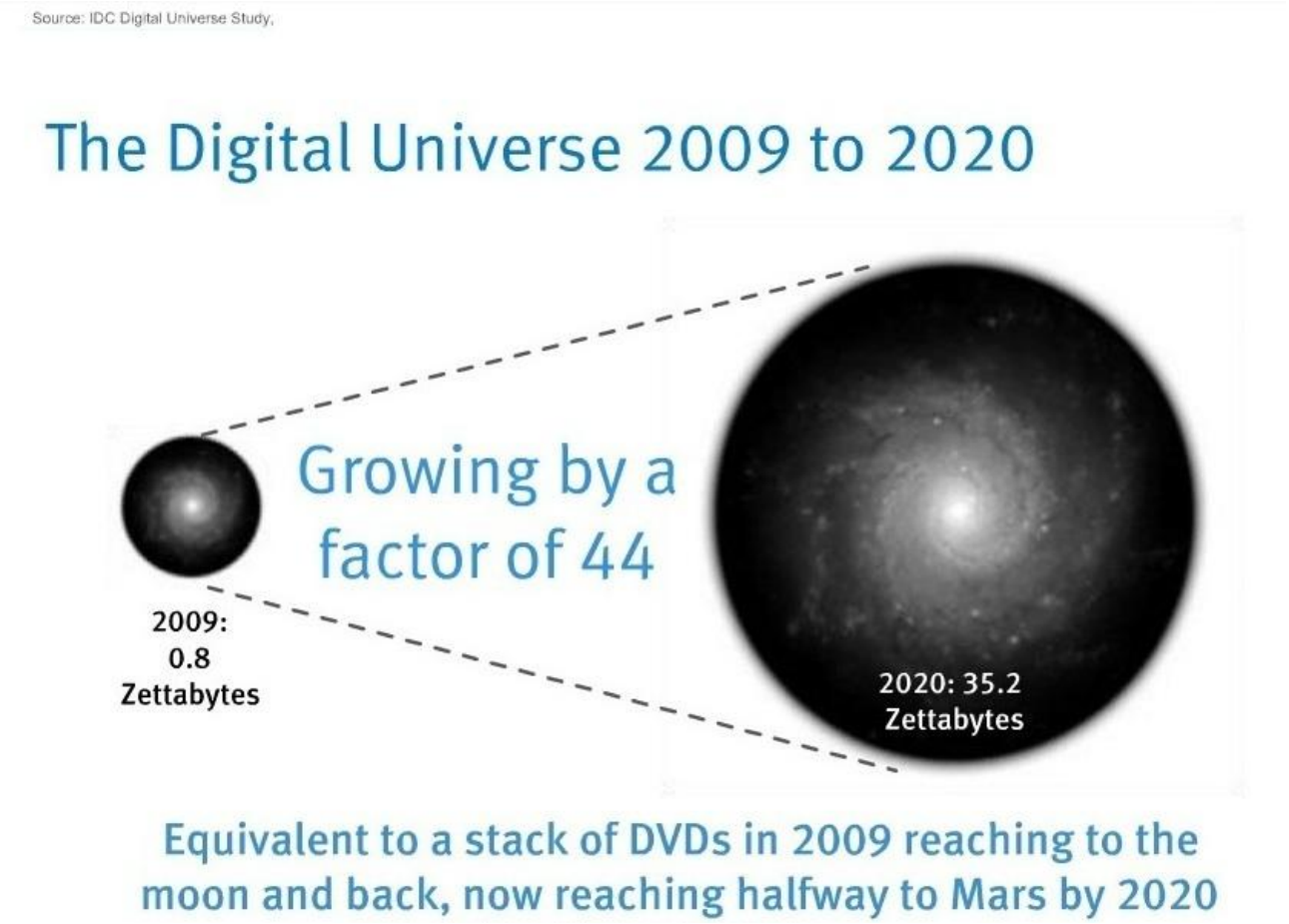
Big Data …
des enjeux et des opportunités

Yvon Kermarrec
Professeur en Informatique

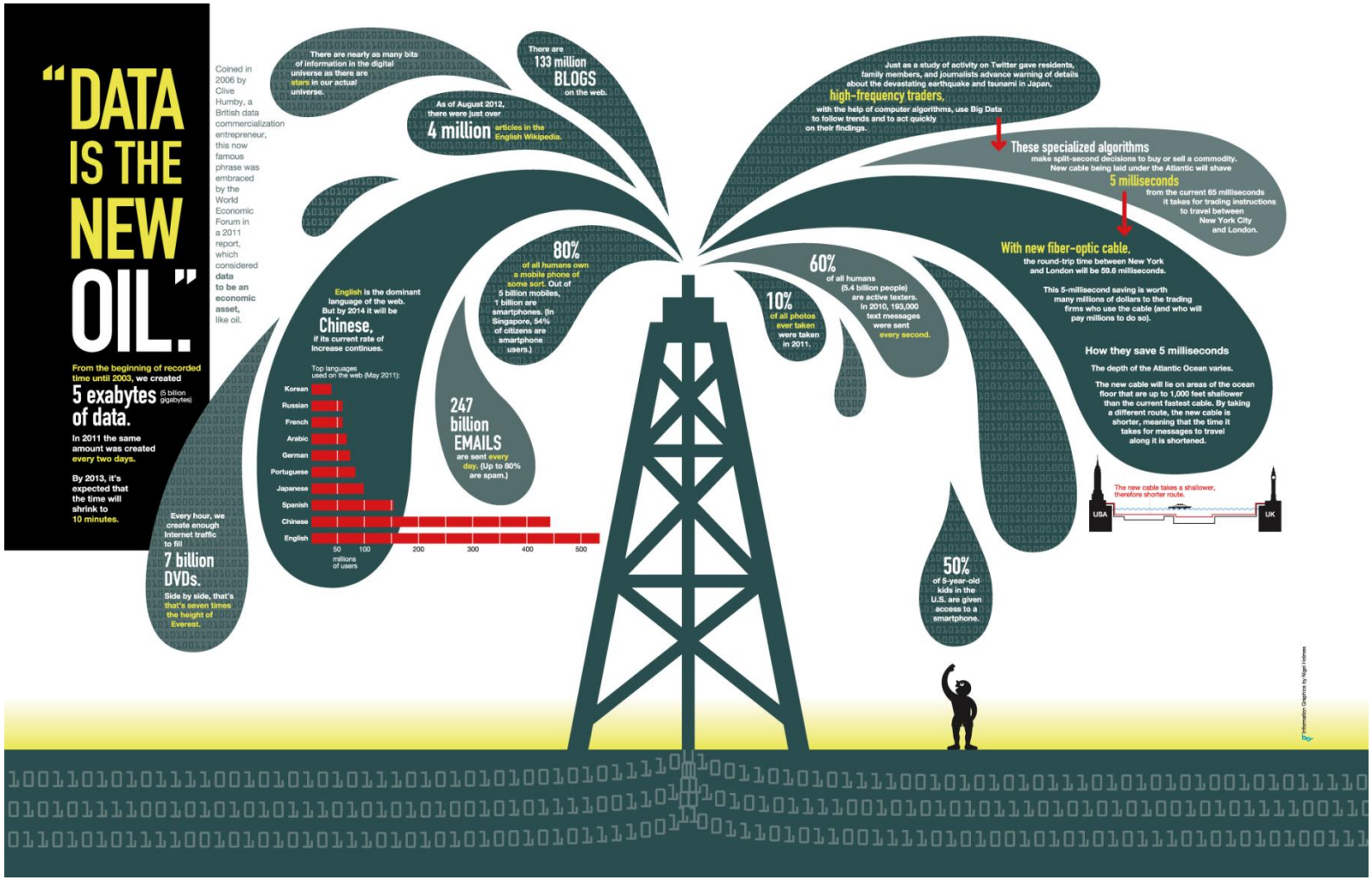
Agenda

- Introduction et contexte
- Une vue globale de Hadoop
- Un peu de technique avec HDFS et Map Reduce
- Comment se lancer et créer son propre cluster?
- Pour aller plus loin avec l'écosystème
- Des exemples en cours à Télécom Bretagne
- Conclusions et perspectives

Les données : un déferlement



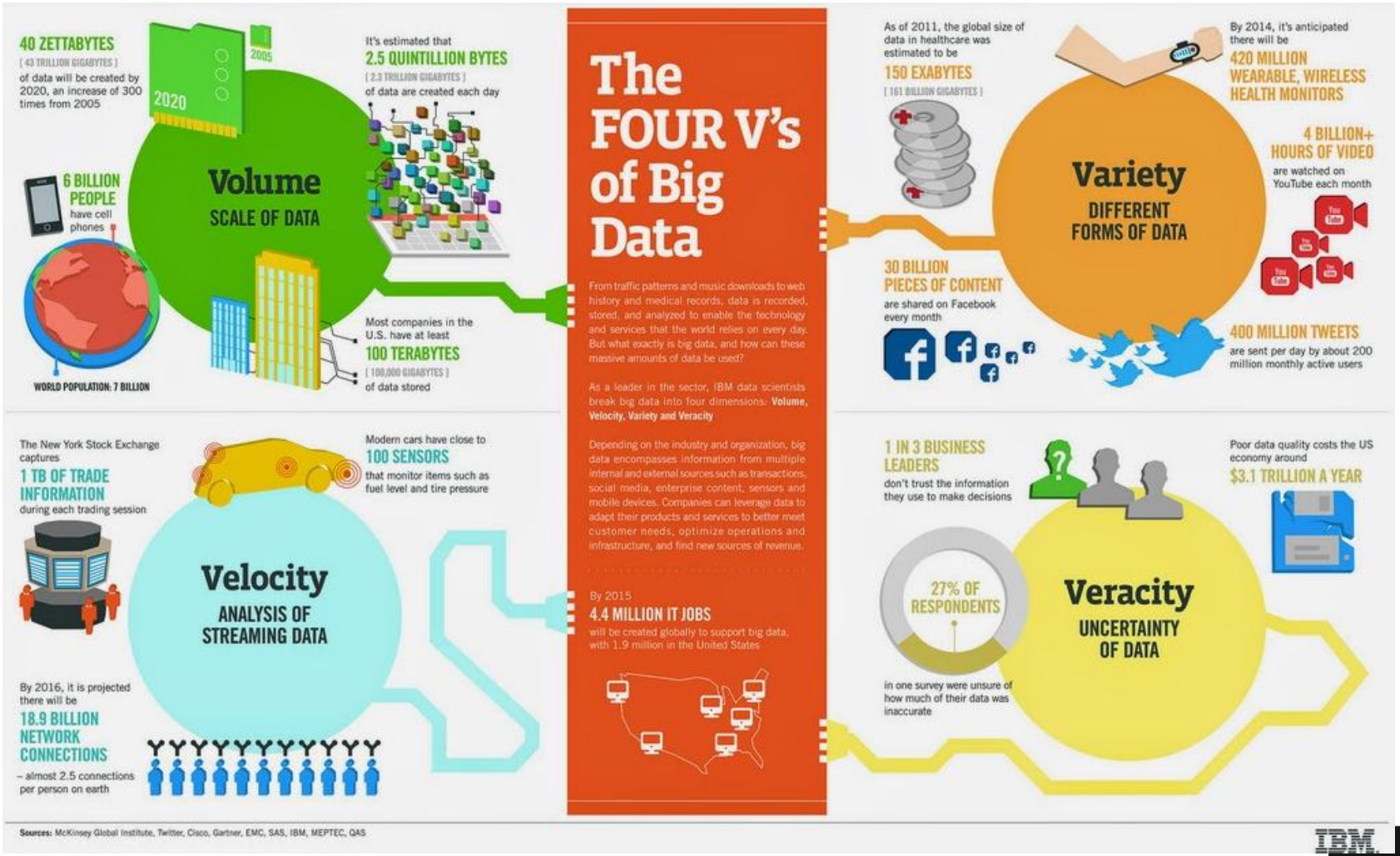
Les données : une valeur à exploiter



Pourquoi sommes nous arrivés à cette situation ?

- **L'informatique connectée et ses services les moteurs de recherche, le e-commerce, les applications ….**
  - Des traces de leur utilisation sont générées
  - Des données sont présentées: flightradar, waze, réseau mobile, …
  - Des informations sont couplées avec d'autres : GPS, réseaux sociaux, informations multi canales, etc.
- **Les objets « intelligents » sont avec nous et en plein développement**
  - Smartphones, tablettes, voitures connectées, chaussures de sports, capteurs, moteurs d'avion, …
- **Le coût du stockage est en chute libre**
  - 1 Go sur disque dur: de $147 en 97 à $0,01 en 2013

Les 4 ou 5 V des big data

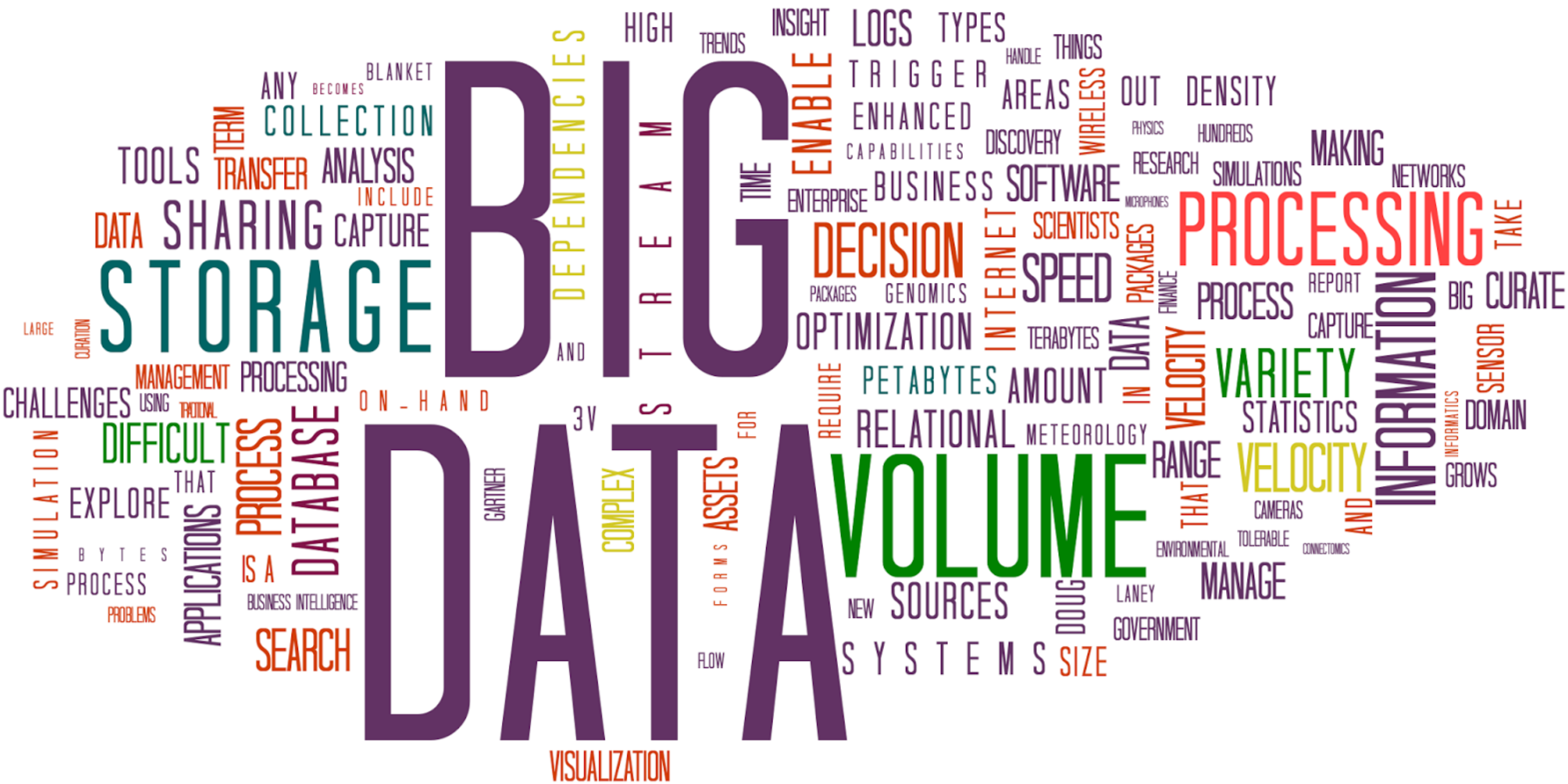


Comment stocker et exploiter ces données? Ou l'échec des approches classiques

- **Les supers calculateurs**
  - http://www.top500.org/
  - Leurs couts sont rédhibitoires (sauf pour certains..)
- **Les SGBD sont incapables de traiter des gros volumes de données**
- **Les latences et temps de transferts de disque sont élevés**
- **Les machines distribuées avec des centaines ou des milliers de processeurs / machines classiques**
  - La machine parallèle du pauvre
  - Complexité qui s'exprime en nombre de messages

Des entreprises s'y mettent

- Google, FaceBook etc… c'est dans leur ADN
- Amazon ou Netflix et leurs systèmes de recommandation
- La grande distribution: Walmart, Target, Carrefour et l'optimisation logistique et la connaissance de ses clients
- Boeing 787 : 0,5 TB de données et par vol – exploitées par les compagnies aériennes
- Les opérateurs de télécoms
- ….
- Des start ups : waze
- Et nos états et agences


BIG
DATA
HIGH
TRENDS
INSIGHT
LOGS
TYPES
HANDLE
THINGS
BLANKET
ANY
BECOMES
TRIGGER
AREAS
WIRELESS
OUT
DENSITY
COLLECTION
ENHANCED
PHYSICS
HUNDREDS
TERM
DISCOVERY
MAKING
TOOLS
TRANSFER
ANALYSIS
CAPABILITIES
RESEARCH
SIMULATIONS
NETWORKS
INCLUDE
ENABLE
TIME
ENTERPRISE
BUSINESS
SOFTWARE
MICROPHONES
DATA
SHARING
CAPTURE
DEPENDENCIES
STREAM
DECISION
SCIENTISTS
PROCESSING
TAKE
STORAGE
INTERNET
SPEED
PACKAGES
FINANCE
PACKAGES
GENOMICS
PROCESS
REPORT
BIG
CURATE
LARGE
CURATION
OPTIMIZATION
TERABYTES
DATA
CAPTURE
INFORMATION
AND
MANAGEMENT
PROCESSING
REQUIRE
PETABYTES
AMOUNT
IN
VELOCITY
VARIETY
SENSOR
CHALLENGES
USING
TRADITIONAL
ON-HAND
3V
FOR
RELATIONAL
METEOROLOGY
STATISTICS
INFORMATICS
DOMAIN
DIFFICULT
PROCESS
DATABASE
GARTNER
COMPLEX
ASSETS
VOLUME
RANGE
VELOCITY
GROWS
SIMULATION
EXPLORE
THAT
THAT
CAMERAS
AND
BYTES
IS A
FORMS
TOLERABLE
ENVIRONMENTAL
CONNECTOMICS
PROCESS
APPLICATIONS
BUSINESS INTELLIGENCE
NEW
SOURCES
DOUG
LANEY
MANAGE
PROBLEMS
GOVERNMENT
SEARCH
FLOW
SYSTEMS
SIZE
VISUALIZATION

Agenda

- Introduction et contexte
- Une vue globale de Hadoop
- Un peu de technique avec HDFS et Map Reduce
- Comment se lancer et créer son propre cluster?
- Pour aller plus loin avec l'écosystème
- Des exemples en cours à Télécom Bretagne
- Conclusions et perspectives

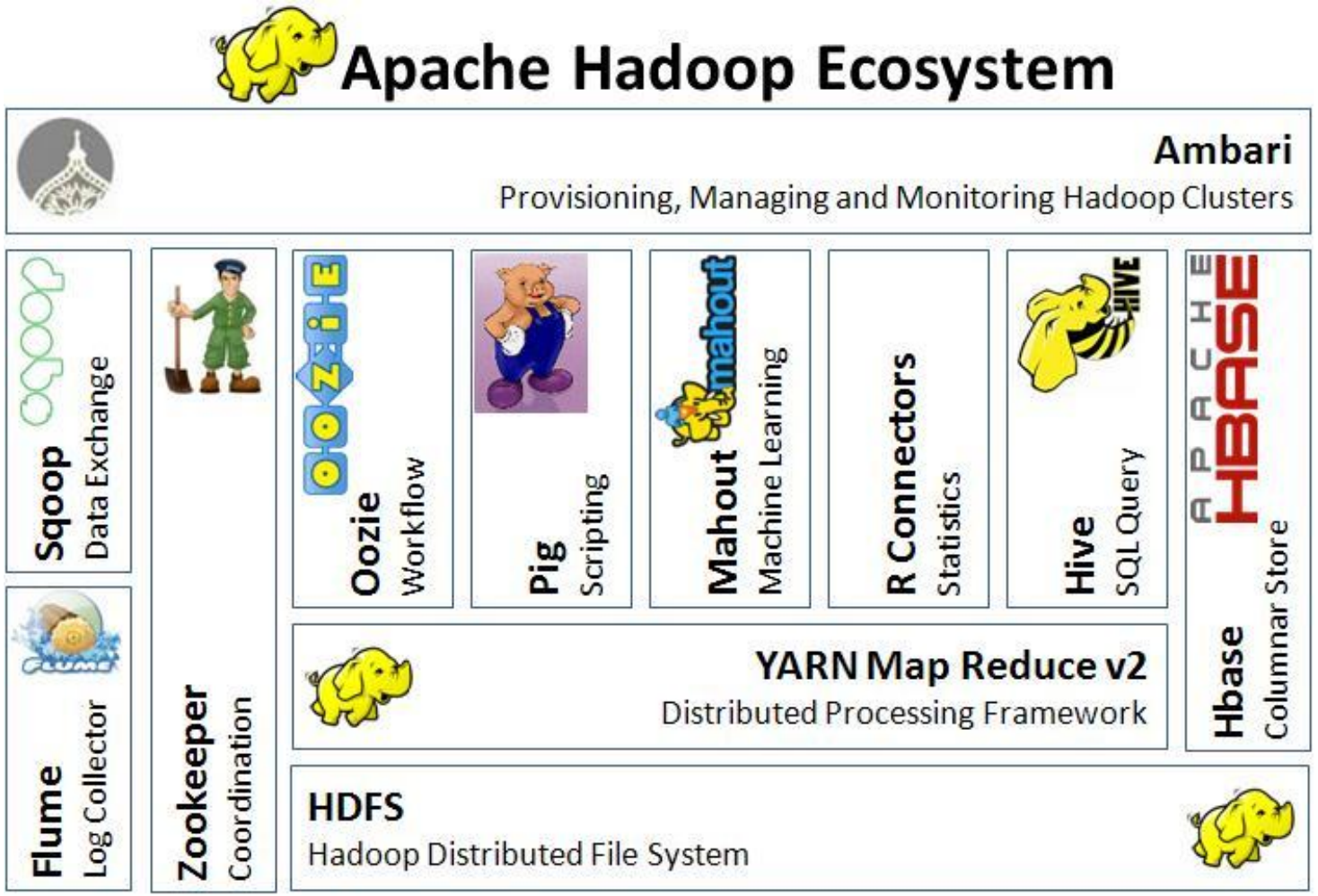
Hadoop

- **Une approche proposée dans la suite des travaux de Google**
- **« Un framework Java libre destiné à faciliter la création d'applications distribuées et échelonnables (scalables), permettant aux applications de travailler avec des milliers de nœuds et des péta-octes de données » (wikipedia)**
- **Un socle pour un éco-système riche**

Cahier des charges pour la conception d'Hadoop

- **Un cluster Hadoop doit pouvoir regrouper plusieurs dizaines, centaines ou milliers de nœuds: chaque nœud permet d'offrir du stockage et une puissance de calcul**
- **Un cluster Hadoop doit pouvoir stocker et traiter des gros volumes de données dans des délais et couts acceptables**
- **Si un nœud tombe, cela ne doit pas entrainer l'arrêt du calcul ou la perte de données**
- **Une machine peut être rajoutée dans le cluster et ceci conduit à une amélioration des performances**

Un éco système en expansion



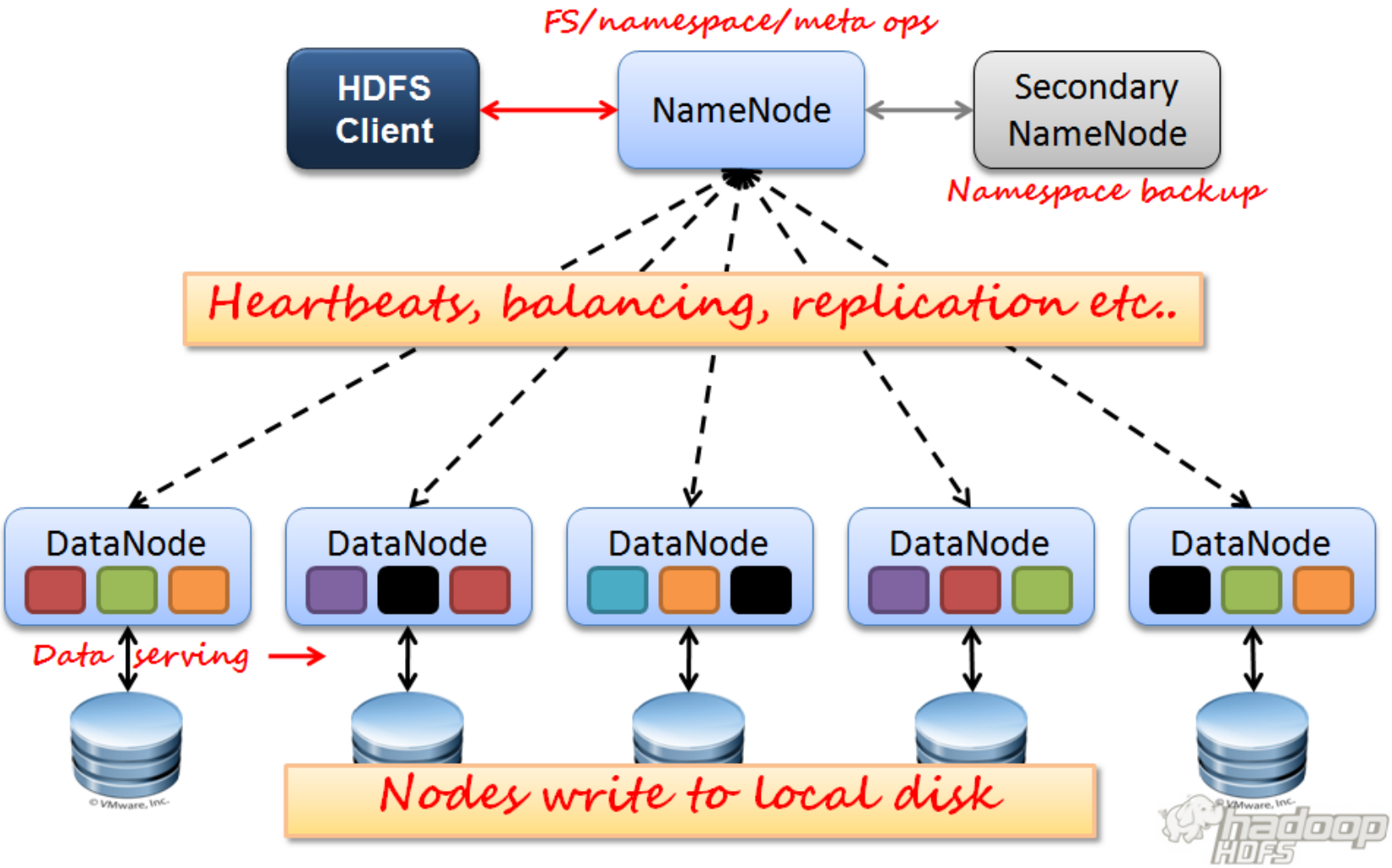
Agenda

- Introduction et contexte
- Une vue globale de Hadoop
- Un peu de technique avec HDFS et Map Reduce
- Comment se lancer et créer son propre cluster?
- Pour aller plus loin avec l'écosystème
- Des exemples en cours à Télécom Bretagne
- Conclusions et perspectives

HDFS : Hadoop Distributed File System

- Un ‘nouveau’ système de gestion de fichiers (SGF) pour lire et écrire des données sur le cluster
- Des blocs de taille importante : 64 MO par exemple au lieu de 4KO pour NFS
- Un SGF particulier de type « write once » adapté au stockage de flux de données
- Chaque bloc est sauvegardé 3 fois, au moins, pour augmenter le disponibilité et la sécurité des données
- HDFS repose sur des SGF classiques et donc des disques standards sont utilisés

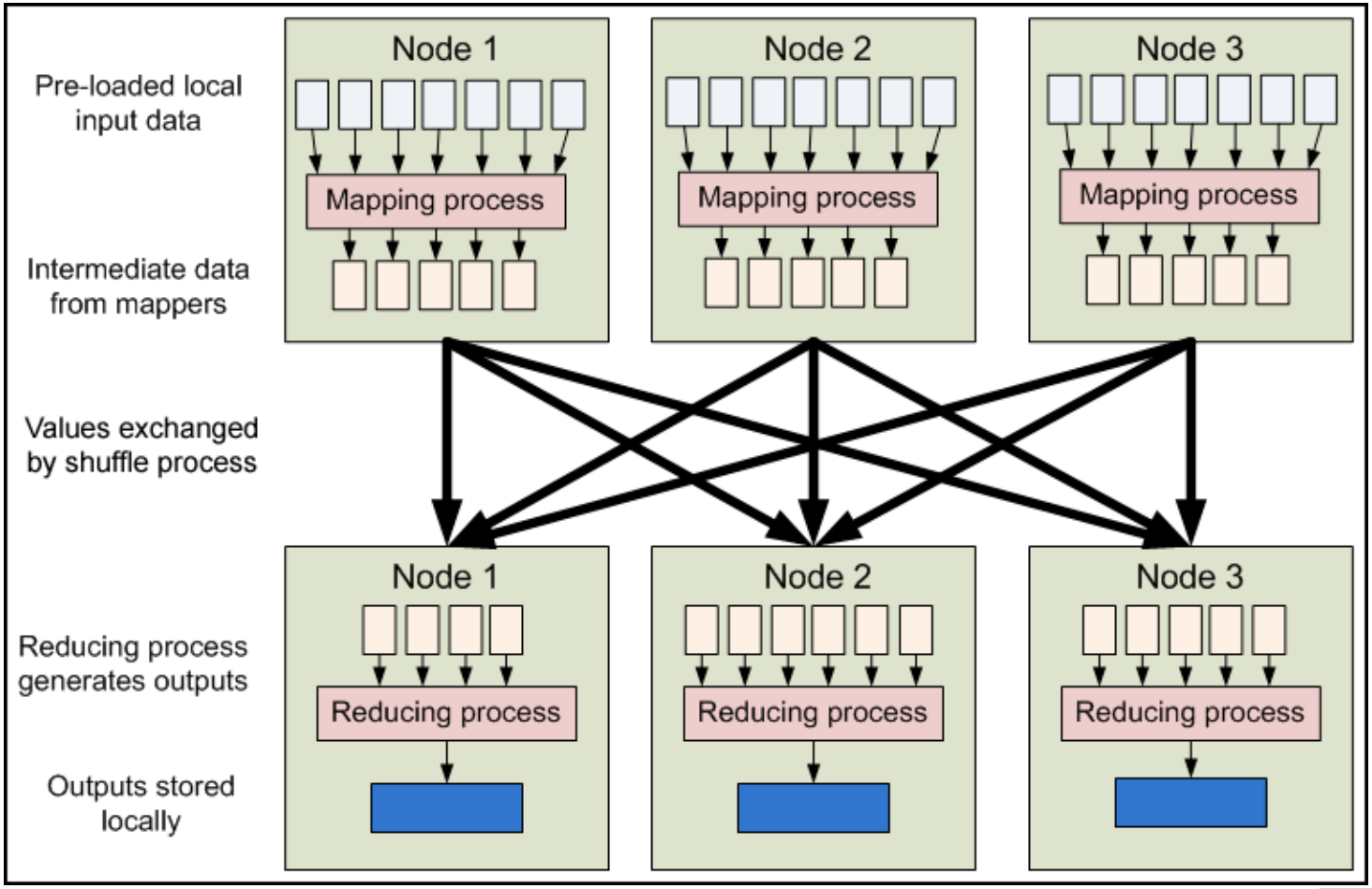
HDFS architecture



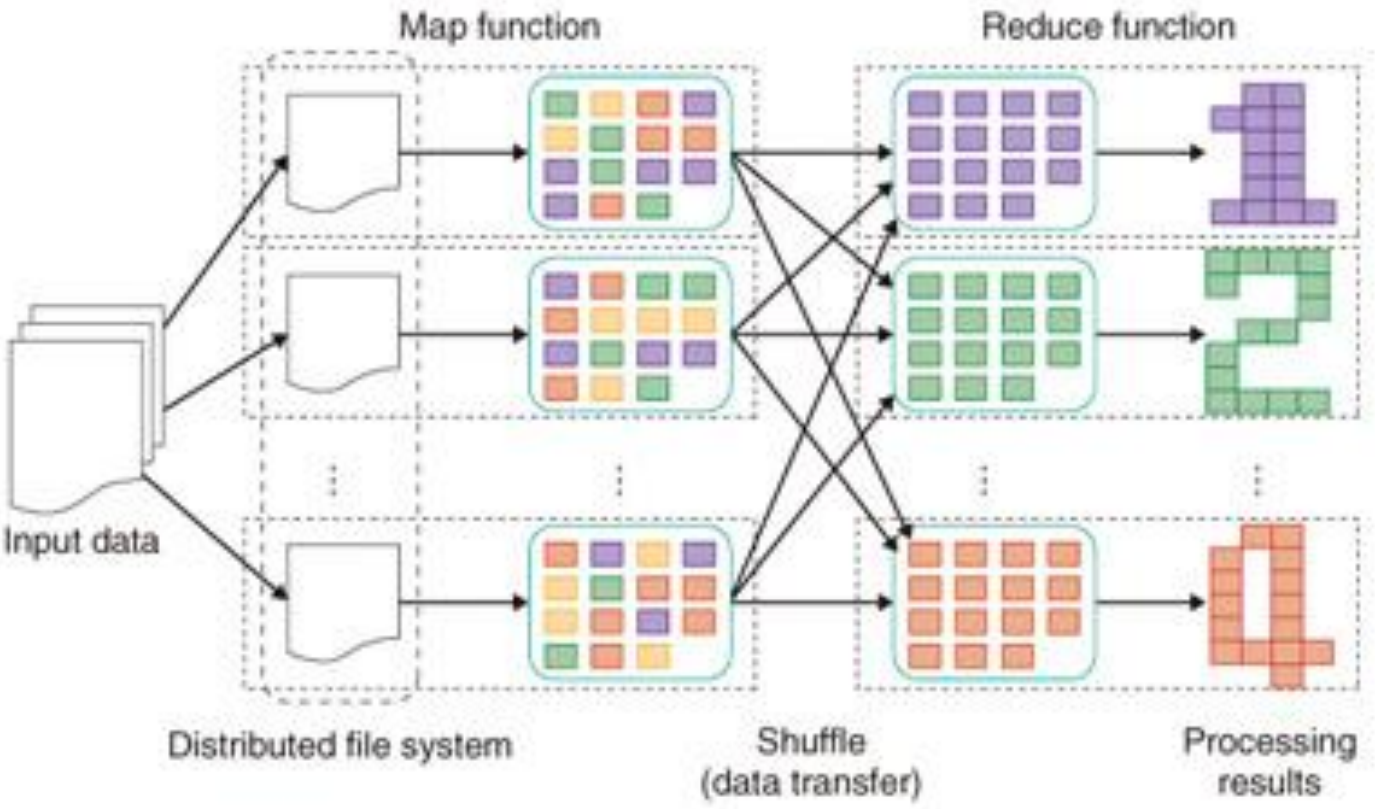
Map Reduce de Google

- **« High performance computing » sur plusieurs milliers de machines**
  - Réduit le déplacement des données entre machines qui est la source de la complexité en distribué
- **Fournit un très haut niveau de transparences aux utilisateurs**
  - Masque la parallélisation des traitements
  - Prend en charge la tolérance aux pannes
  - Gère l'équilibrage des charges et la coordination
- **Fournit un modèle « relativement » simple à comprendre et à programmer et une puissance d'expression**

Map Reduce de Google



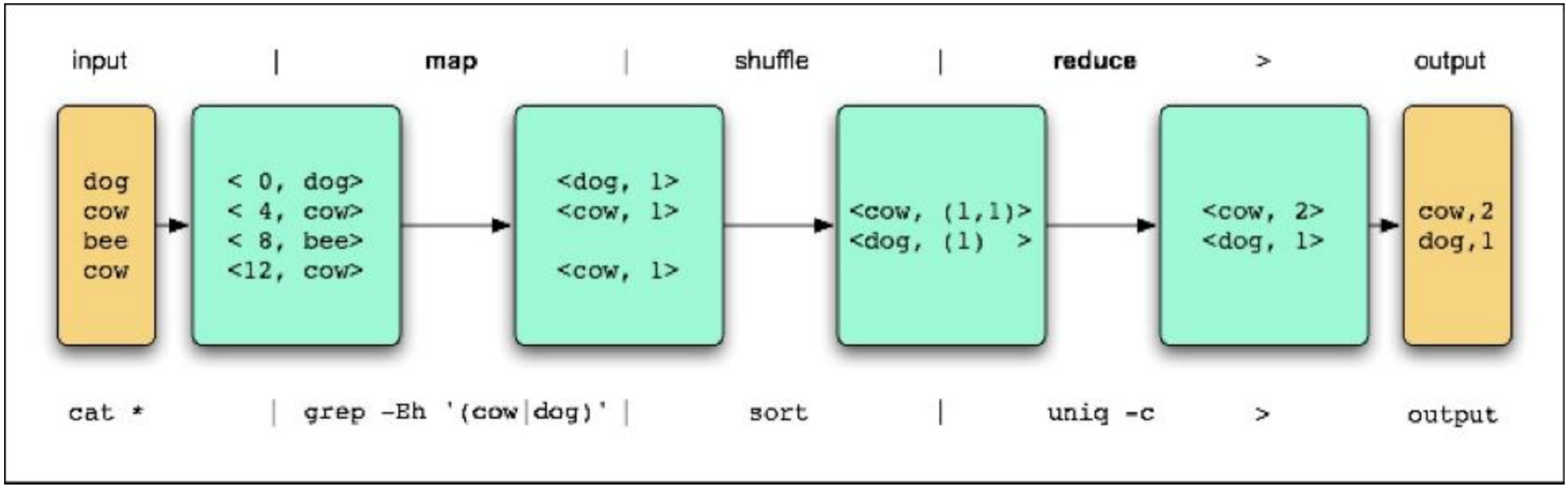
Map Reduce de Google



Map reduce en action

- **L'entité de base : la paire (clé, valeur)**
- **Une fonction de *map* qui traite les entrées et produit des couples (clé, valeur)**
- **Une fonction de *reduce* qui applique un traitement sur les données locales de type (clé, valeur)**
- **Une fonction intermédiaire *shuffle & sort* entre les fonctions *map* et *reduce***
- **Des traitements qui peuvent être enchainés**

Map reduce en action : compter des mots



Le programme associé

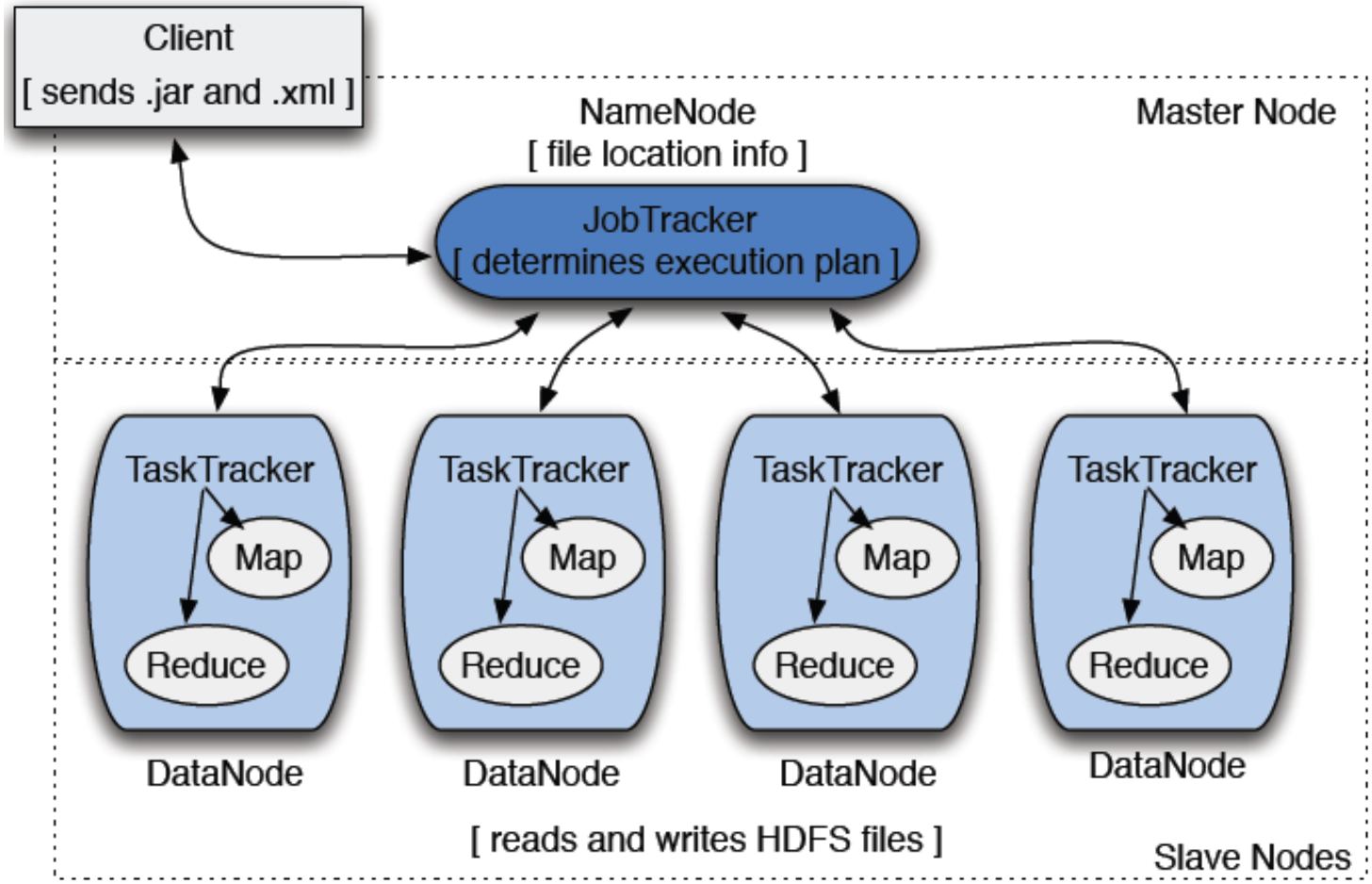
- **map(String key, String value):**

```
// key: document name
// value: document contents
for each word w in value:
        EmitIntermediate(w, "1");
```

- **reduce(String key, Iterator values):**

```
// key: a word
// values: a list of counts
int result = 0;
        for each v in values:
                result = result + ParseInt(v);
        Emit(key, AsString(result));
```

Derrière la scène : une infrastructure puissante



Derrière la scène : une infrastructure puissante

- **L'architecture d'Hadoop est distribuée et de type maitre esclaves**
- **Le *JobTracker***
  - Assure l'interface avec le client
  - Gère et coordonne les jobs s'exécutant sur le cluster
  - Interagit avec le Name Node pour localiser les données
  - Nécessite beaucoup de mémoire et de CPU
  - Composant critique pour cette architecture
- **Le *TaskTracker***
  - Exécute les tâches d'un job sur chaque noeud du cluster. Il s'agit des *Map* et des *Reduces*
  - Signale tout problème au *JobTracker*

Agenda

- Introduction et contexte
- Une vue globale de Hadoop
- Un peu de technique avec HDFS et Map Reduce
- Comment se lancer et créer son propre cluster?
- Pour aller plus loin avec l'écosystème
- Des exemples en cours à Télécom Bretagne
- Conclusions et perspectives

Comment se lancer ?

- **Apprendre avec une machine virtuelle de type Cloudera :**
  - Utilisation de HDFS et premiers programmes avec map reduce
- **Utiliser ensuite les différents modes de fonctionnement d'Hadoop**
  - Modes local, pseudo-distribué, totalement distribué ou virtualisé
  - Bâtir progressivement son cluster à partir de matériels recyclés puis achetés spécifiquement
- **Approches « idéales » en environnement d'enseignement recherche**

Pour une entreprise : un choix du DSI

- **Un cluster dédié : installé physiquement dans l'entreprise ou chez un hébergeur**

  + confidentialité des données, maitrise du cluster, totale liberté

  - investissement, administration du cluster et « tuning »

- **Un cluster dans le cloud**

  + tarif compétitif, fiabilité et disponibilité élevés

  - Limitations techniques, configurations imposées, confidentialité des données

Notre retour d'expérience à Télécom Bretagne

- **Des machines Dell de type PowerEdge à hauteur de 10 k€ et des machines « de base »**
- **Une configuration initiale doublée par la réutilisation d'une machine similaire et son intégration dans notre cluster**
- **Le projet TerraLab de l'Institut Mines Télécom**
  - Une machine prévue pour les traitements intensifs
  - 4 TB de mémoire
- **Un environnement logiciel complet et « free » distribué par la Fondation Apache : HDFS, MapReduce, Hbase, Hive, ZooKeeper, ….**

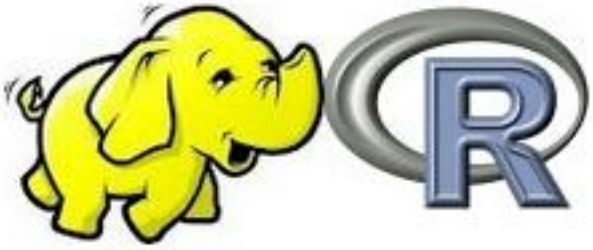
Agenda

- Introduction et contexte
- Une vue globale de Hadoop
- Un peu de technique avec HDFS et Map Reduce
- Comment se lancer et créer son propre cluster?
- Pour aller plus loin avec l'écosystème
- Des exemples en cours à Télécom Bretagne
- Conclusions et perspectives

L'Ecosystème

- **Apache Mahout**
  - Package open-source pour Hadoop pour le data mining et l'apprentissage (machine learning)
  - Un projet Apache – avec une communauté dynamique

- **Revolution R (R-Hadoop)**
  - Extensions du package R pour fonctic Hadoop



Ce que l'on peut faire avec Mahout

- **4 grandes classes d'applications**
  - Classification, clustering, data mining, recommandation
- **Algorithmes nombreux et disponibles pour les programmeurs : un environnement riche**
- **Une performance dans les traitements du fait de l'architecture sous-jacente**

Agenda

- Introduction et contexte
- Une vue globale de Hadoop
- Un peu de technique avec HDFS et Map Reduce
- Comment se lancer et créer son propre cluster?
- Pour aller plus loin avec l'écosystème
- Des exemples en cours à Télécom Bretagne
- Conclusions et perspectives

Applications en cours à Télécom Bretagne

- **Analyse des réseaux sociaux**
- **Recommandations et marketing**
- **Big data ocean**
- **Détection d'intrusion (cyber sécurité)**
- **Anticipation de mobilités de clients en téléphonie**
- **Métrologie des réseaux**
- **Anonymisation des grands volumes de données**
- **Fragmentation des grands volumes de données et impossibilité de reconstituer les données initiales**
- **Le tee shirt connecté**

Agenda

- Introduction et contexte
- Une vue globale de Hadoop
- Un peu de technique avec HDFS et Map Reduce
- Comment se lancer et créer son propre cluster?
- Pour aller plus loin avec l'écosystème
- Des exemples en cours à Télécom Bretagne
- Conclusions et perspectives

Extraction de valeurs à partir des données

- **La donnée : le nouvel or noir**
  - Systématiser la collecte issues de différentes sources
  - Croiser les données entre elles et identifier de nouvelles règles et connaissances
  - Favoriser les croisements thématiques : sciences, humanités, marketing etc.
- **L'innovation et la mise à disposition des clients des services à forte valeur ajoutée**
  - Des innovations et services à proposer autour des données
  - Des nouveaux services à imaginer et concevoir

Enjeux et menaces

- **La méga puissance des grands acteurs du web**
  - Place centrale et incontournable
- **Vie privée et données personnelles**
  - Qui peut en disposer ? Les limites et l'éthique ?
- **La souveraineté nationale**
- **Peu de formations et de professionnels ayant des compétences en données**

Opportunités

- **Des solutions techniques viables et de complexité raisonnable**
  - Bâtir son propre cluster à un cout raisonnable
  - Réaliser des traitements inenvisageables auparavant
- **Des recrutements de spécialistes des données à anticiper**
  - Un marché à fort potentiel
  - Des croisements thématiques à anticiper
- **Les logiciels disponibles via Apache**
  - Des codes sources disponibles
  - Un développement important de la suite logicielle et d'outils d'administration / monitoring

Des domaines à explorer pour la Cote d'Ivoire ?

- **Innovation et start-ups**
- **Le trafic routier**
  - Identification des zones de danger, de ralentissement, des goulots d'étranglements, …
  - Attendus : fluidité, gain de temps et de carburant
- **La santé**
  - Remonter les données des dispensaires
  - Identifier des anomalies et anticiper des risques
  - Attendus : modélisation des épidémies, zones à risque
- **L'environnement et la pollution**
  - Des capteurs sur un territoire
  - Vers une infrastructure de collecte et de stockage
  - Attendus : des anticipations de crues, mouvement de terrains, …

рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
mersi
barka
welalin
tack
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
kiitos
dankie
dhanyavad
thank you
gracias
хвала
asante
obrigada
enkosi
bedankt
nanni
nandri
bayarlalaa
gràcie
hvala
mauruuru
köszönöm
akun dankon
djiere dieuf
tau
mochchakkeram
дякую
mamnun
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
děkuji
obrigado
sukriya
kop khun krap
taiku
arigatô
takk
dakujem
trugarez
mèsi
sagolun
ありがとう
tanemirt
grazie
najis tuke
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
merce
мерси
kam sah hamnida
rahmat
감사합니다
xièxie
ευχαριστώ
merci