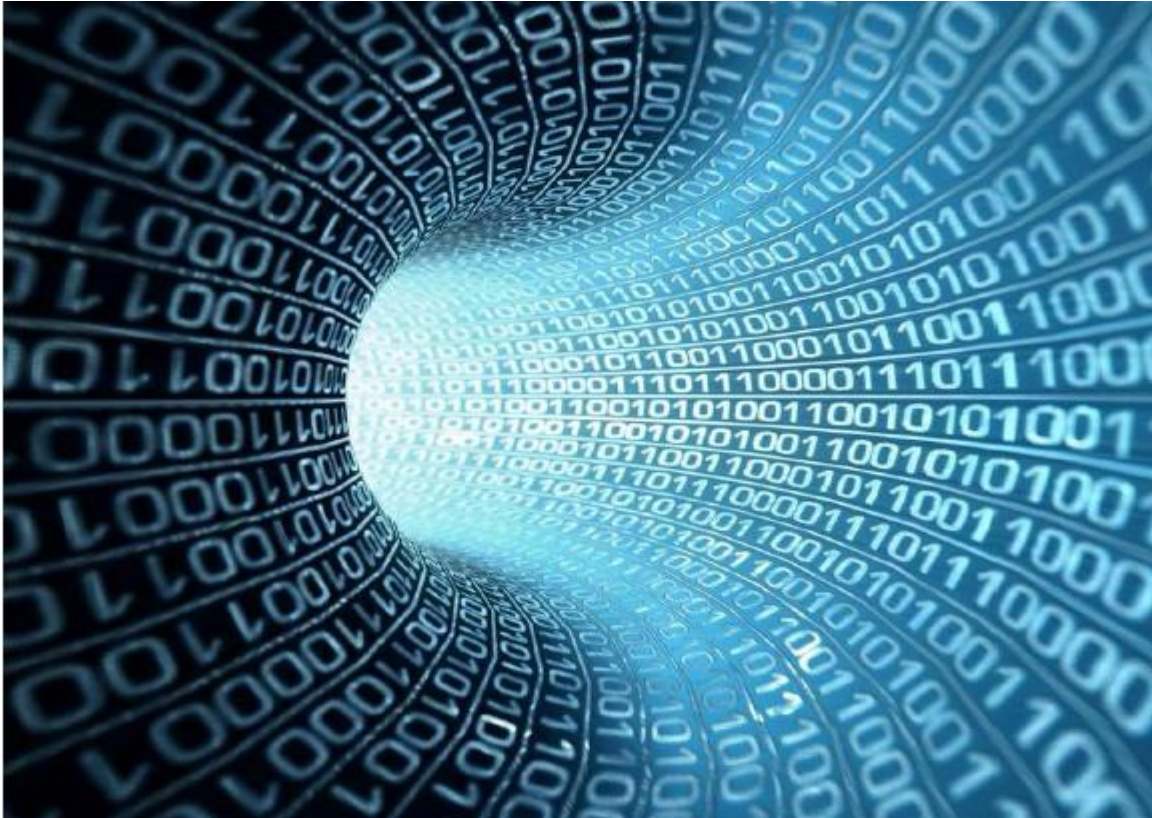


# Présentation Générale Big Data

Guide Share France



Olivier JOUANNOT

# Information On Demand 2013



## BIG DATA

### L'ENGOUEMENT MEDIATIQUE

Buzz des éditeurs pour propulser leurs nouvelles offres ou prochaine révolution informatique ?  
Quoi qu'il en soit, le « BigData » n'est pas une tendance en devenir mais des concepts et des technologies déjà largement éprouvées.



# Information On Demand 2013



## BIG DATA : DEFINITION

### ► LITTERALEMENT

grosse données ou volume massif de données structurées ou non.  
On parle aussi de Datamasse par similitude avec la biomasse.

### ► CONCEPTUELLEMENT

Ce terme vulgarise à la fois la représentation du volume des données mais aussi les infrastructures liées au traitement de ces données.



# Information On Demand 2013

## BIG DATA : GENERALITES

### EVOLUTION DES TECHNOLOGIES DE STOCKAGE DES DONNES

#### ➤ LE STOCKAGE EN 1956

IBM 305 RAMAC

5Mb de stockage sur disque

50 disques 24 pouces

temps d'accès de 10 caractères par seconde.

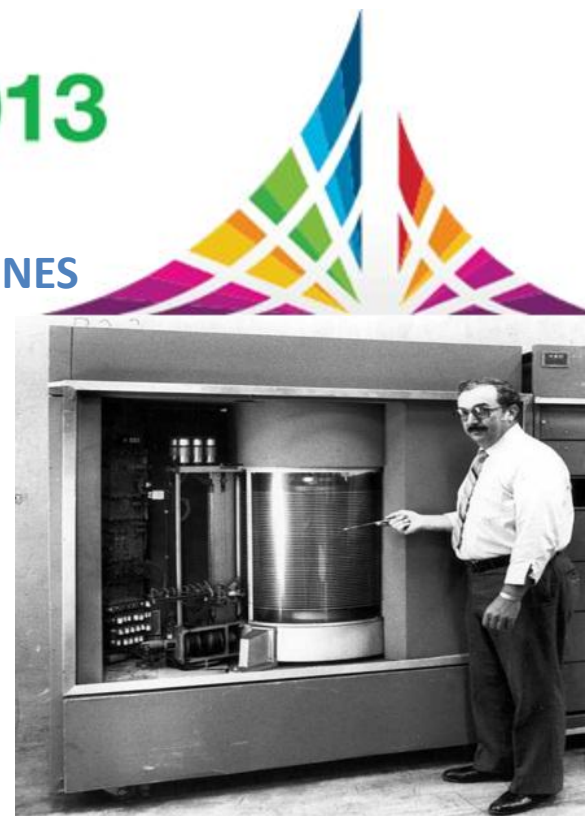
#### ➤ LE STOCKAGE EN 2013

IBM FlashSystem

1 Petabyte

22 millions d'IOPS

Temps de latence en microseconde ( $\mu = 10^{-6}$  secondes)



Multiples d'octets tels que définis par IEC 60027-2					
Préfixe SI			Préfixe binaire		
Nom	Symbole	Valeur	Nom	Symbole	Valeur
kilooctet	ko	$10^3$	kibioctet	Kio	$2^{10}$
mégaoctet	Mo	$10^6$	mébioctet	Mio	$2^{20}$
gigaoctet	Go	$10^9$	gibioctet	Gio	$2^{30}$
téraoctet	To	$10^{12}$	tébioctet	Tio	$2^{40}$
pétaoctet	Po	$10^{15}$	pébioctet	Pio	$2^{50}$
exaoctet	Eo	$10^{18}$	exbioctet	Eio	$2^{60}$
zettaoctet	Zo	$10^{21}$	zébioctet	Zio	$2^{70}$
yottaoctet	Yo	$10^{24}$	yobioctet	Yio	$2^{80}$



#### IBM FlashSystem

##### 1 Rack

- 1 Petabyte: 1 Floor Tile
- 100 microsecond latency
- 22 Million IOPS
- 210 GB/s
- 12.6 KW power  
*Less power than the average 200TB array*



# Information On Demand 2013

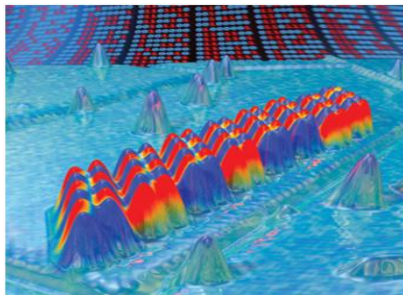


## BIG DATA : GENERALITES

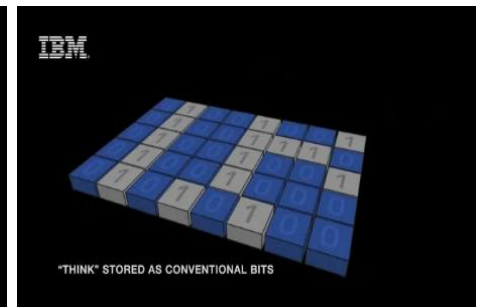
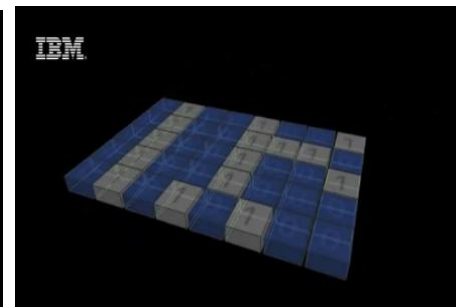
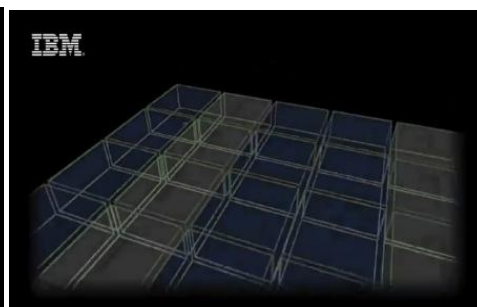
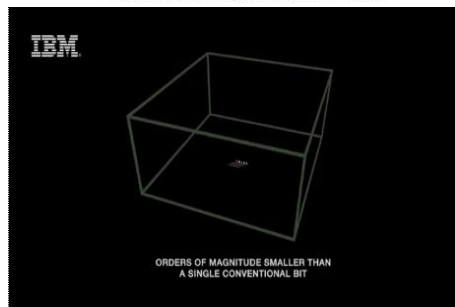
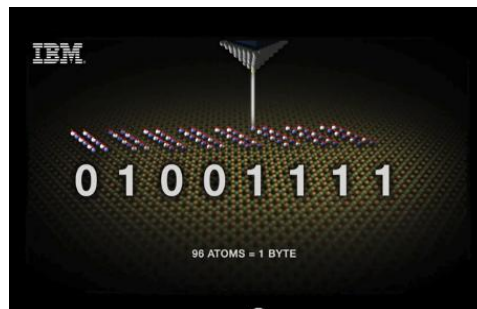
### EVOLUTION DES TECHNOLOGIES DE STOCKAGE DES DONNES

#### ➤ LE STOCKAGE AU NIVEAU ATOMATIQUE ATTEINT EN 2012 PAR IBM

Les chercheurs d'IBM ont réussi à stocker 1 bit sur seulement 12 atomes à l'aide d'un microscope à effet tunnel très puissant.



En passant par l'échelle atomique atteinte par IBM en 2011 (12 atomes pour stocker 1 bits !)



# Information On Demand 2013



## BIG DATA : GENERALITES

### DIVERSITE ET VOLUME DES SOURCES DE DONNEES

#### VOLUMES

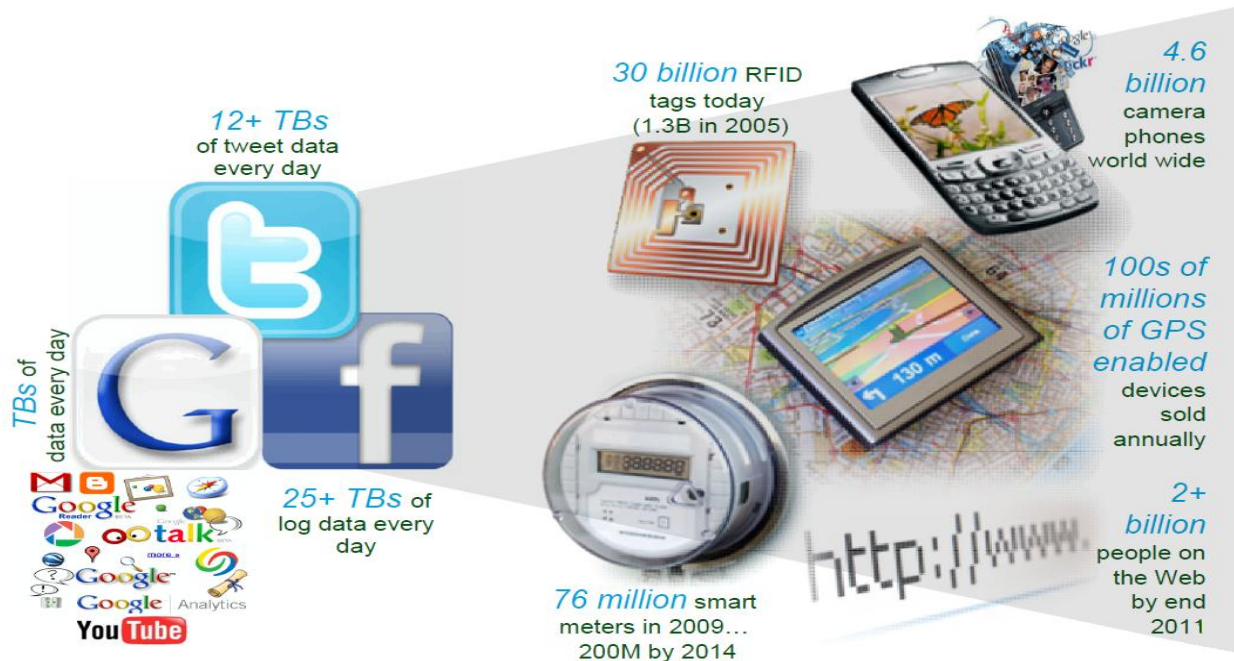
Chaque jour, 2,5 trillions d'octets de données sont générés.

90% des données créées dans le monde l'ont été au cours des 2 dernières années.

Prévision d'une croissance de 800% des quantités de données à traiter d'ici à 5 ans.

#### DIVERSITE DES SOURCES

capteurs, medias sociaux, images, videos, achats en lignes, signaux GPS ...





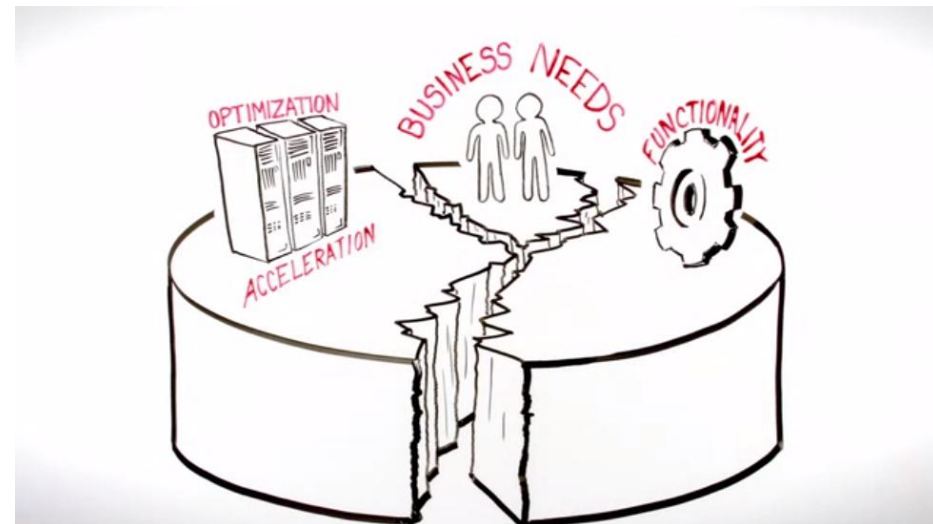
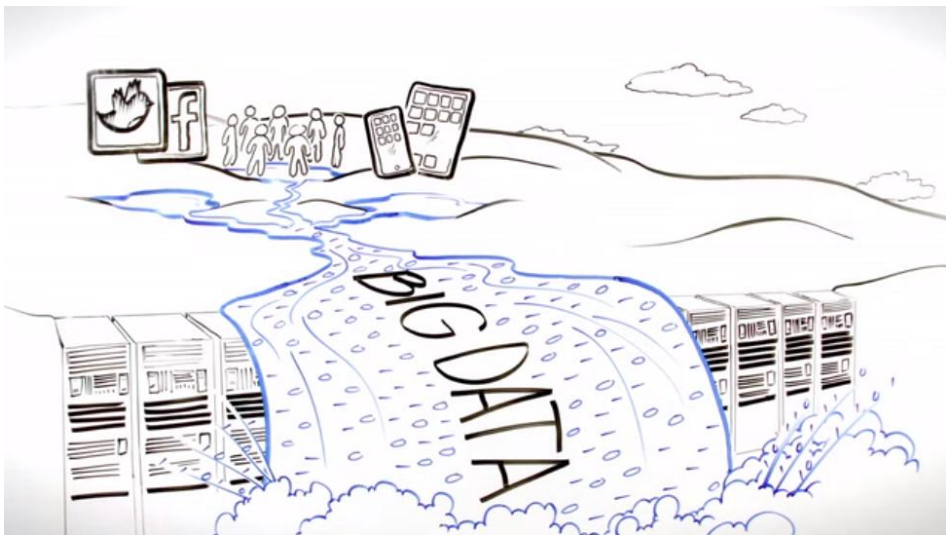
# Information On Demand 2013



## BIG DATA : GENERALITES

### ADAPTABILITE

Dans ce nouveau contexte, les méthodes de traitement de ces données (capture, stockage, recherche, partage, analyse, visualisation) doivent être redéfinies car l'ensemble de ces données deviennent difficilement manipulables par les outils classiques.



# Information On Demand 2013



## BIG DATA : GENERALITES

### PERSPECTIVES ET DOMAINES D 'APPLICATION

Les perspective d'utilisations de ces données sont énormes, notamment pour l'analyse d'opinions politiques, de tendance industrielles, la génomique, la lutte contre la criminalité et la fraude, les méthodes de marketing publicitaire et de vente etc ...





# Information On Demand 2013



## BIG DATA : GENERALITES

### COUVERTURE DE QUATRE DIMENSIONS



#### ► VOLUME

Croissance sans cesse des données à gérer de tout type, souvent en téraoctets voir en pétaoctets

#### ► VARIETE

Traitement des données sous forme structurées et non structurées mais devant faire l'objet d'une analyse collective (databases, textes, données de capteurs, sons, vidéos, de parcours, fichiers journaux etc ...)

#### ► VELOCITE

Utilisation des données en temps réel (pour la détection de fraudes ...)

#### ► VERACITE

Gestion de la fiabilité et de la véracité des données imprécises et prédictives



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

Les grands acteurs du web tel que **Google, Yahoo, Facebook, Twitter, LinkedIn ...** ont été les premiers à être confrontés à des volumétries de données extrêmement importantes et ont été à l'origine des premières innovations en la matière portées principalement sur deux types de technologies :

- Les bases de données (NoSql)
- Les plateformes de développement et de traitement des données

La majorité de ces entreprises ont décidés d'ouvrir ces développements internes au monde **Open Source**.

Un certains nombre de ces technologies comme « **hadoop** » font partie de la fondation Apache et ont été intégrés aux offres de « **Big Data** » des grands acteurs tel que **IBM, Oracle, Microsoft, EMC ...**

# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE


Société	Technologie développée	Type de technologie	
Google	Big Table	Système de base de données distribuée propriétaire reposant sur GFS ( <i>Google File System</i> ). Technologie non <i>open source</i> , mais qui a inspiré HBase qui est <i>open source</i>	
	MapReduce	Plate-forme de développement pour traitements distribués	
Yahoo	Hadoop	Plate-forme Java destinée aux applications distribuées et à la gestion intensive des données. Issue à l'origine de Google BigTable, MapReduce et Google File System	
	S4	Plate-forme de développement dédiée aux applications de traitement continu des flux de données	
Facebook	Cassandra	Base de données de type NoSQL et distribuée	
	Hive	Logiciel d'analyse de données utilisant Hadoop	
Twitter	Storm	Plate-forme de traitement de données massives	
	FlockDB	Base de données distribuée de type graphe	
LinkedIn	Kafka	Système distribué de gestion des messages	
	SenseiDB	Base de données temps réel distribuée et semi-structurée	
	Voldemort	Base de données distribuée destinée aux très grosses volumétries	

Tableau 4.3. Quelques technologies open source du big data.

© Lavoisier

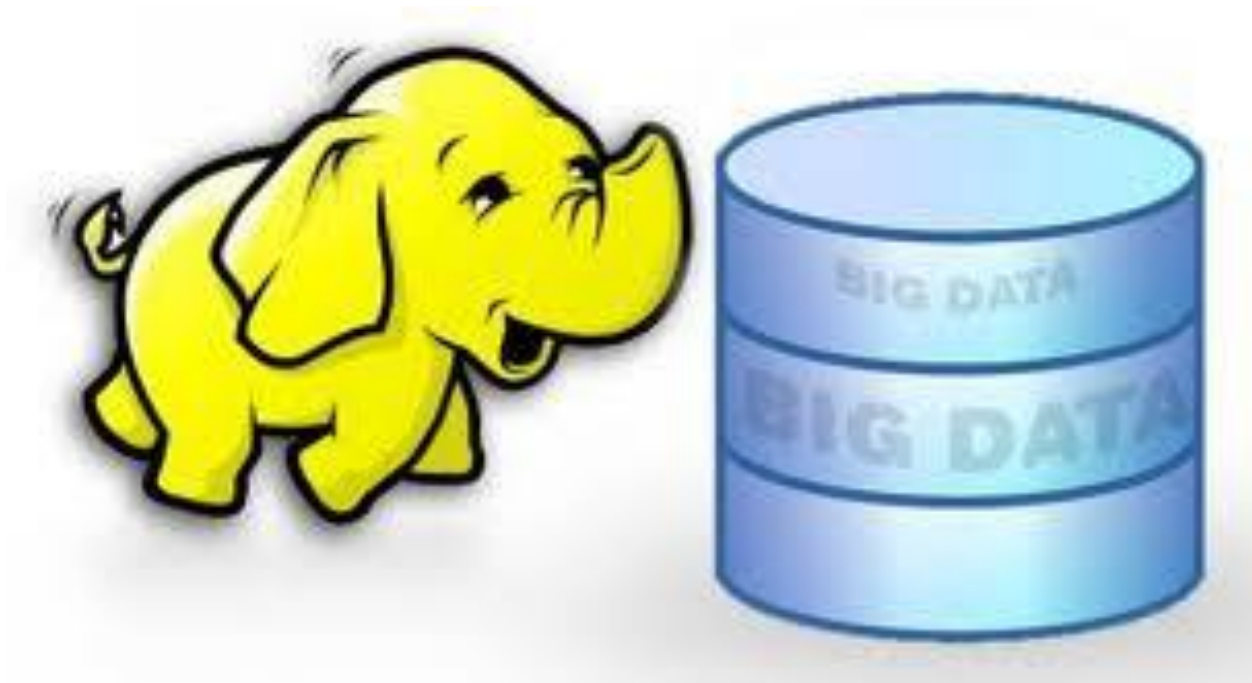


# Information On Demand 2013



**BIG DATA : LES ACTEURS DE L'OPEN SOURCE**

**HADOOP / MAPREDUCE : LES BASES DU BIG DATA**





## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

### HDFS / MAPREDUCE OU LES BASES DU BIG DATA

#### ► HDFS (Hadoop Distributed File System)

C'est un système de fichiers distribué, extensible et portable développé par Hadoop et basé sur le principe MapReduce à partir du GoogleFS. Écrit en Java, il a été conçu pour stocker de très gros volumes de données sur un grand nombre de machines peu coûteuses équipées de disques durs banalisés. Il permet l'abstraction de l'architecture physique de stockage, afin de manipuler un système de fichiers distribué comme s'il s'agissait d'un disque dur unique.

De très nombreuses entreprises utilisent Hadoop parmi lesquelles on peut citer Amazon, Adobe, AOL, Bing (Microsoft), Cornell University, eBay, Facebook, Fox Audience Network, Google, Hotels & Accommodation, IBM, Last FM, LinkedIn, Rakuten, Sling Media, Spotify, StumbleUpon, Telefonica Research, The New York Times, Twitter, Web Alliance, Yahoo.

#### ► MapReduce

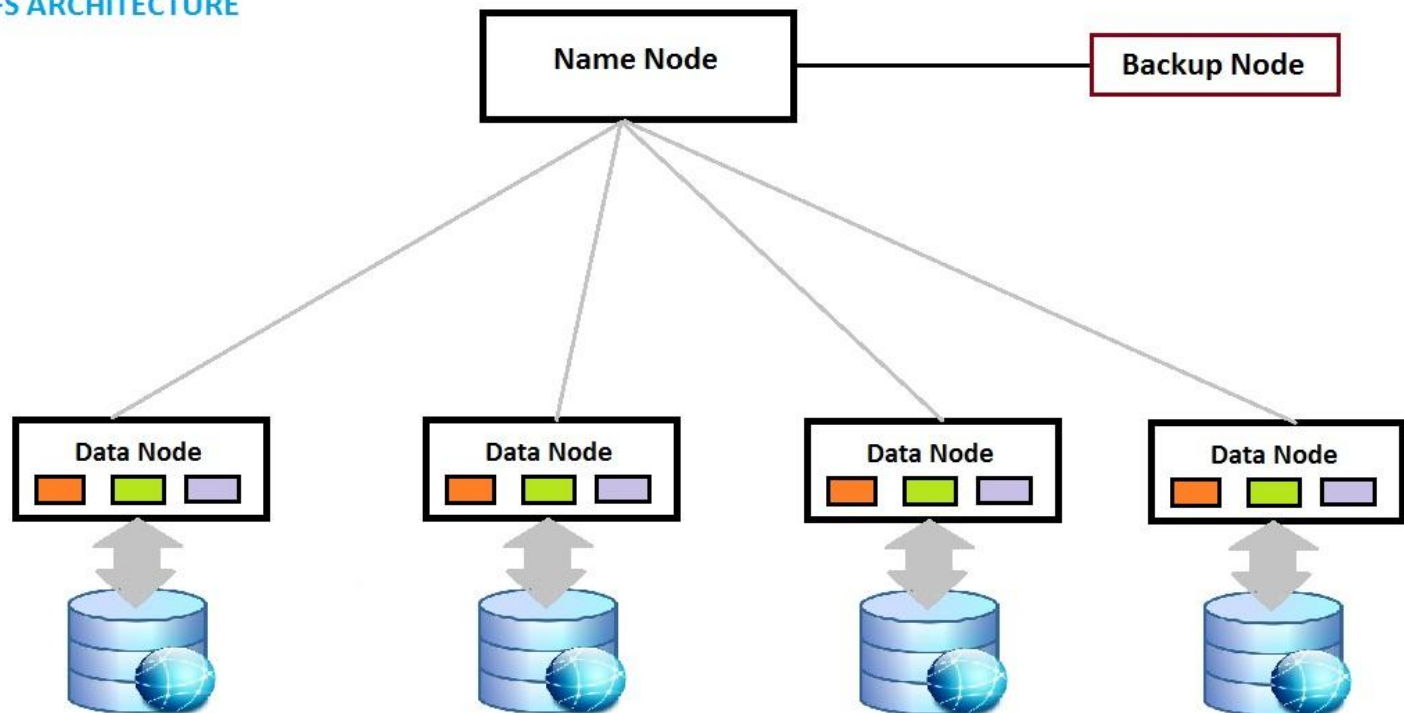
Il joue un rôle majeur dans le traitement des grandes quantités de données. La distribution des données au sein de nombreux serveurs permet le traitement parallélisé de plusieurs tâches portant chacune sur des morceaux de fichiers. La fonction Map accomplit une opération spécifique sur chaque élément. L'opération Reduce combine les éléments selon un algorithme particulier, et fournit le résultat. Soulignons que le principe de délégation peut être récursif : les nœuds à qui sont confiées des tâches peuvent aussi déléguer des opérations à d'autres nœuds.



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

### HDFS

#### HDFS ARCHITECTURE







## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

### HDFS : SPECIFICATIONS TECHNIQUES

Une architecture de machines HDFS (aussi appelée cluster HDFS) repose sur deux types de composants majeurs :

#### **NameNode (nœud de nom) :**

ce composant gère l'espace de nom, l'arborescence du système de fichier et les métadonnées des fichiers et des répertoires. Il centralise la localisation des blocs de données répartis dans le cluster. Il est unique mais dispose d'une instance secondaire qui gère l'historique des modifications dans le système de fichiers (rôle de backup). Ce NameNode secondaire permet la continuité du fonctionnement du cluster Hadoop en cas de panne du NameNode d'origine.

#### **DataNode (nœud de données) :**

ce composant stocke et restitue les blocs de données. Lors du processus de lecture d'un fichier, le NameNode est interrogé pour localiser l'ensemble des blocs de données. Pour chacun d'entre-eux, le NameNode renvoie l'adresse du DataNode le plus accessible, c'est-à-dire le DataNode qui dispose de la plus grande bande passante. Les DataNodes communiquent de manière périodique au NameNode la liste des blocs de données qu'ils hébergent. Si certains de ces blocs ne sont pas assez répliqués dans le cluster, l'écriture de ces blocs s'effectue en cascade par copie sur d'autres.



## **BIG DATA : LES ACTEURS DE L'OPEN SOURCE**

### **HDFS : SPECIFICATIONS TECHNIQUES**

Chaque DataNode sert de bloc de données sur le réseau en utilisant un protocole spécifique au HDFS.

Le système de fichier utilise la couche TCP/IP pour la communication. Les clients utilisent le Remote Procedure Call pour communiquer entre eux. Le HDFS stocke les fichiers de grande taille sur plusieurs machines. Il réalise la fiabilité en répliquant les données sur plusieurs hôtes et par conséquent ne nécessite pas de stockage RAID sur les hôtes. Avec la valeur par défaut de réplication, les données sont stockées sur trois nœuds : deux sur le même support et l'autre sur un support différent. Les DataNodes peuvent communiquer entre-eux afin de rééquilibrer les données et de garder un niveau de réplication des données élevé.

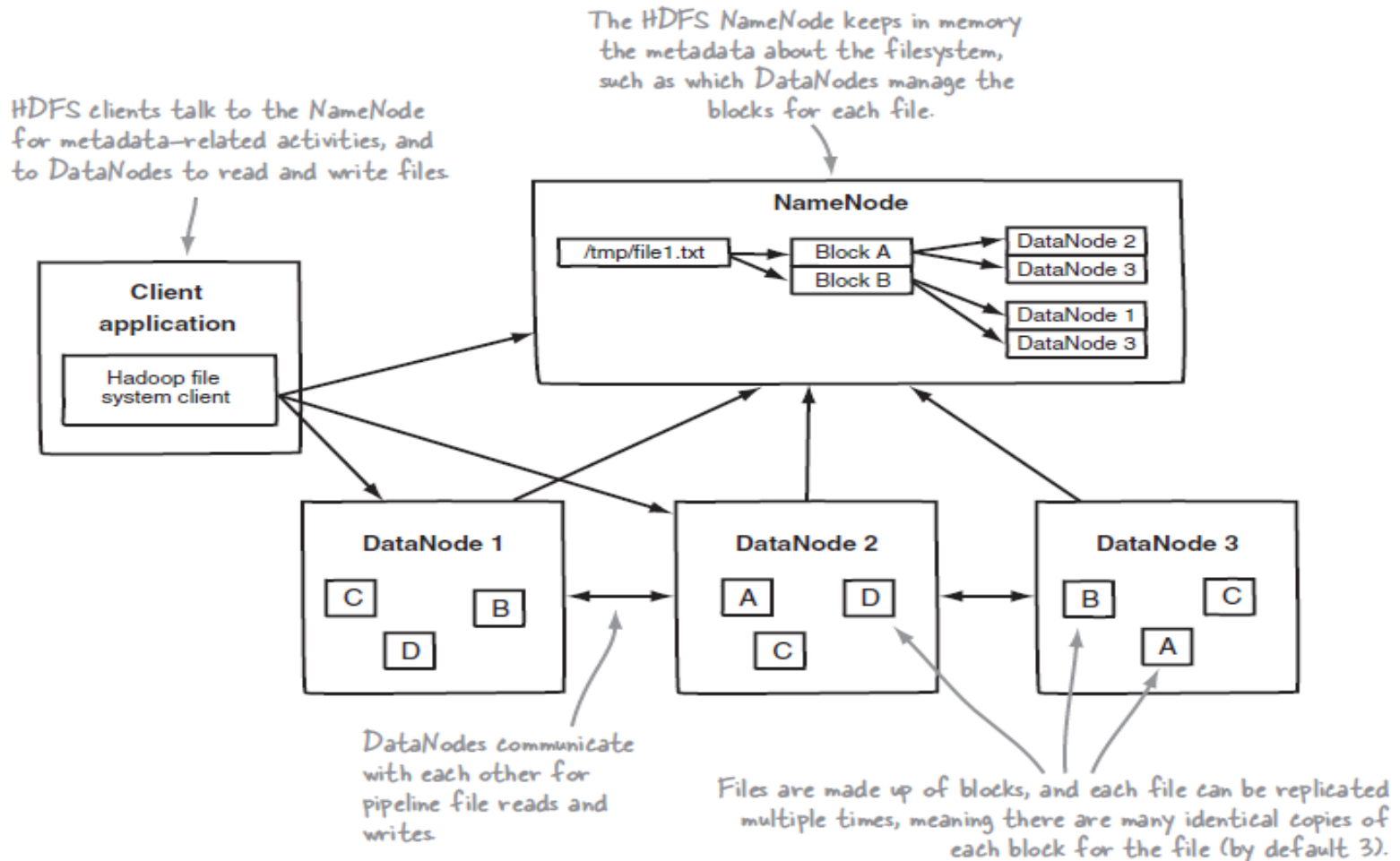
Le HDFS n'est pas entièrement conforme aux spécifications POSIX, en effet les exigences relatives à un système de fichier POSIX diffèrent des objectifs cibles pour une application Hadoop. Le compromis de ne pas avoir un système de fichiers totalement compatible POSIX permet d'accroître les performances du débit de données. Le HDFS a récemment amélioré ses capacités de haute disponibilité, ce qui permet désormais au serveur de métadonnées principal d'être basculé manuellement sur une sauvegarde en cas d'échec (le basculement automatique est en cours d'élaboration). Les NameNodes étant le point unique pour le stockage et la gestion des métadonnées, ils peuvent être un goulot d'étranglement pour soutenir un grand nombre de fichiers, notamment lorsque ceux-ci sont de petite taille. En acceptant des espaces de noms multiples desservis par des NameNodes séparés, le HDFS limite ce problème.

# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

### HDFS



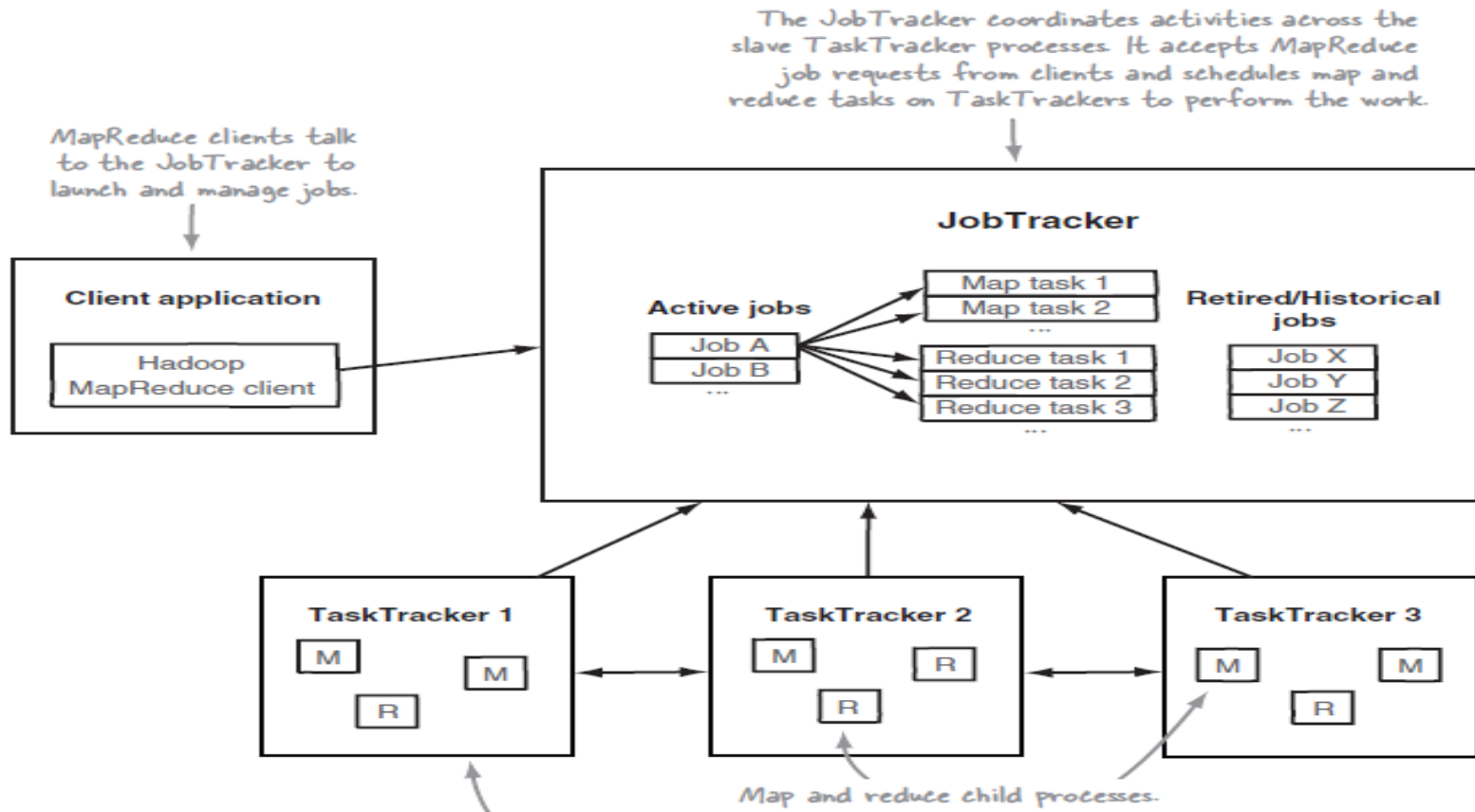
**HDFS architecture shows an HDFS client communicating with the master NameNode and slave DataNodes.**



# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE MAPREDUCE



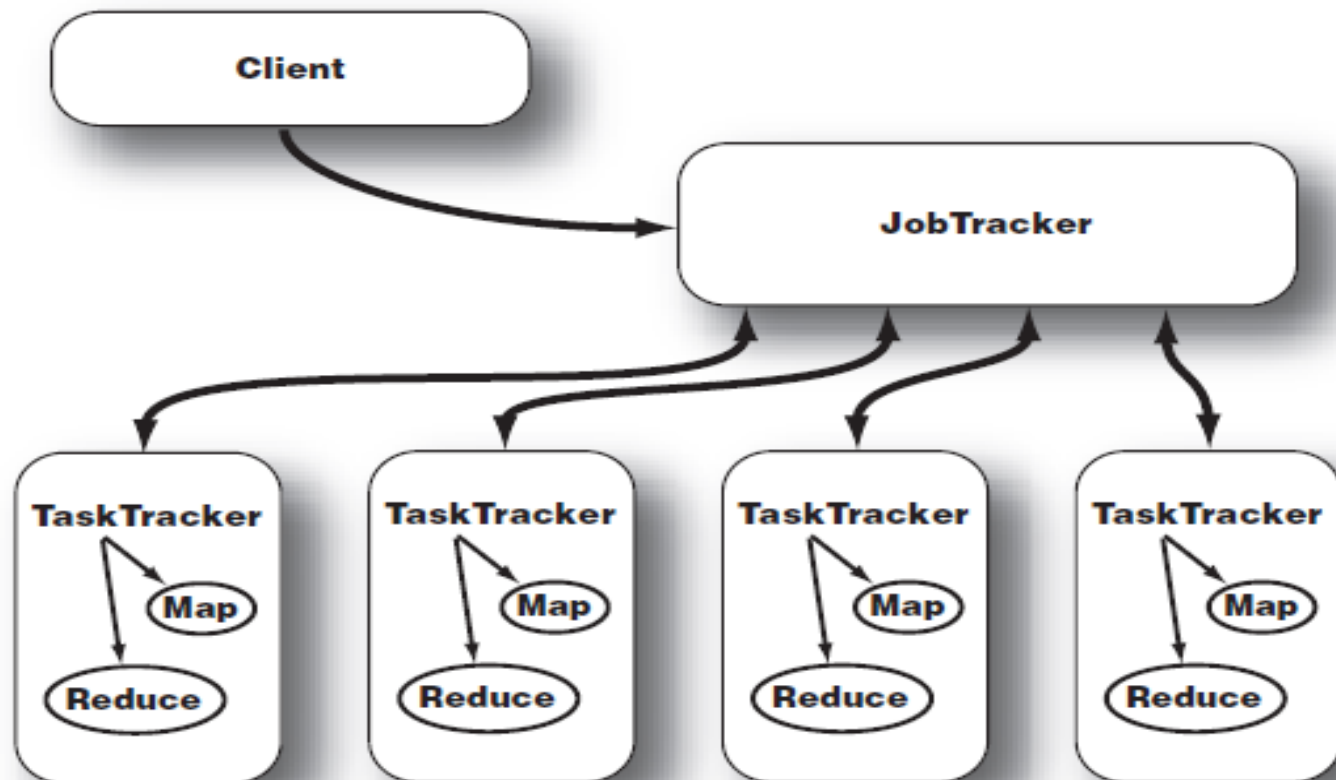
The TaskTracker is a daemon process that spawns child processes to perform the actual map or reduce work. Map tasks typically read their input from HDFS, and write their output to the local disk. Reduce tasks read the map outputs over the network and write their outputs back to HDFS.

MapReduce logical architecture

# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE MAPREDUCE



JobTracker and TaskTracker interaction. After a client calls the JobTracker to begin a data processing job, the JobTracker partitions the work and assigns different map and reduce tasks to each TaskTracker in the cluster.

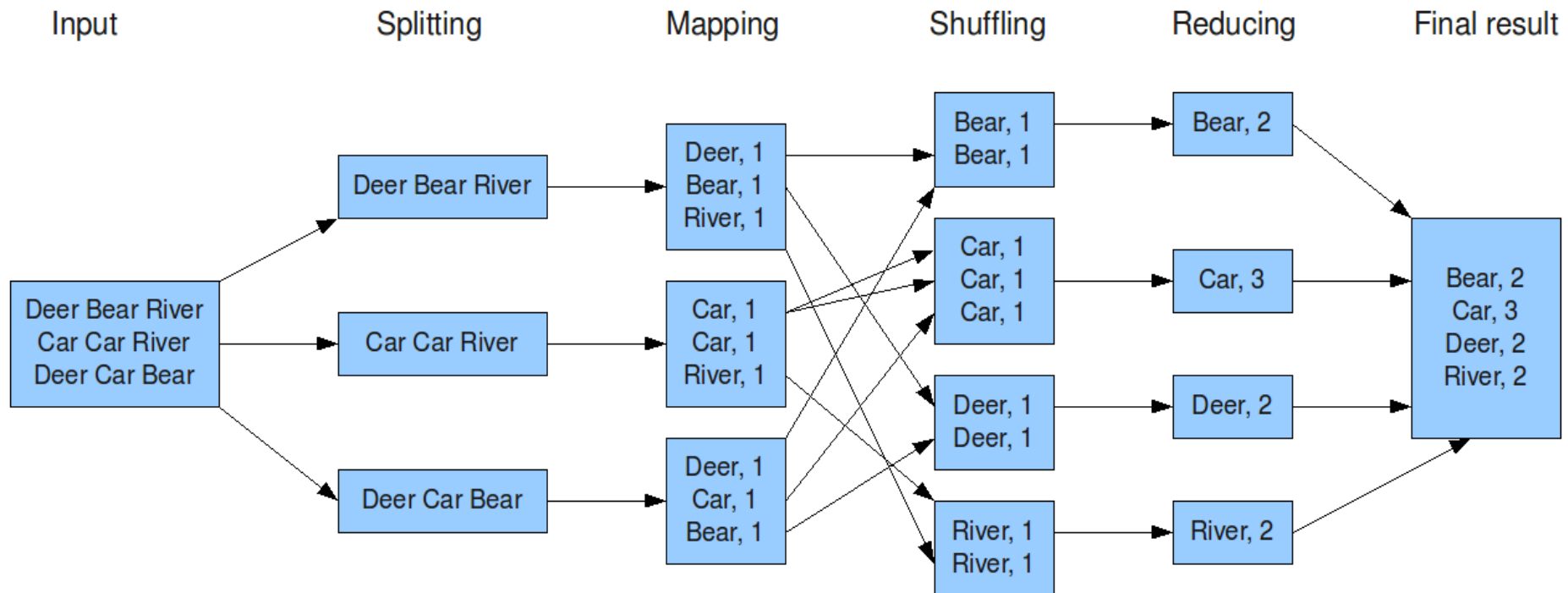
# Information On Demand 2013

## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

### MAPREDUCE LE FRONTAL D'HADOOP



The overall MapReduce word count process





## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

### MAPREDUCE : POURQUOI CE MODEL ?

Tout s'articule sur le découpage de vos programmes en deux parties distinctes dont les exécutions vont être successives : la phase 'map' et la phase 'reduce'.

Le **mapper** est là pour **filtrer** et transformer les entrées en une sortie  
le **reducer** pourra **agréger** une fois la première phase terminée, aboutissant alors au résultat souhaité, que ce soit un simple calcul statistique ou un traitement métier plus complexe.

Chaque phase est en fait une simple méthode, écrite en Java ou éventuellement dans votre langage préféré (Python, Ruby...), de traitement de données à implémenter.

Lors de la première phase, MapReduce reçoit les données et donne chaque élément à traiter à chaque mapper (sur chaque noeud de votre cluster, soit de 1 à n machines).  
A l'issue de cette phase les données traitées sont redistribuées à chaque reducer (idem, chaque noeud de votre cluster, de 1 à n machines) pour arriver au résultat final.

Ces deux phases ne sont pas issues de l'imaginaire des développeurs, mais bien des retours terrains constatés par les Googlers qui travaillaient sur ces problématiques.



# Information On Demand 2013



## BIG DATA : SOLUTIONS

### HADOOP, SPECIFICATIONS TECHNIQUES

#### ➤ AVANTAGES

HDFS n'est pas le premier système de fichier distribué existant mais il a quelques particularités à noter :

- Conçu pour être déployé sur des clusters hétérogènes constitués de machines tout à fait communes
- Tolérance aux pannes
- améliore considérablement le volume des données instantanées fourni aux applications.
- dédié au stockage de gros fichiers, c'est à dire plusieurs centaines de Mégaoctets, Gigaoctets, Téraoctets...
- conscient de la topologie réseau sous-jacente dans laquelle il évolue, il optimise ainsi l'emplacement des blocs.

#### ➤ INCONVENIENTS

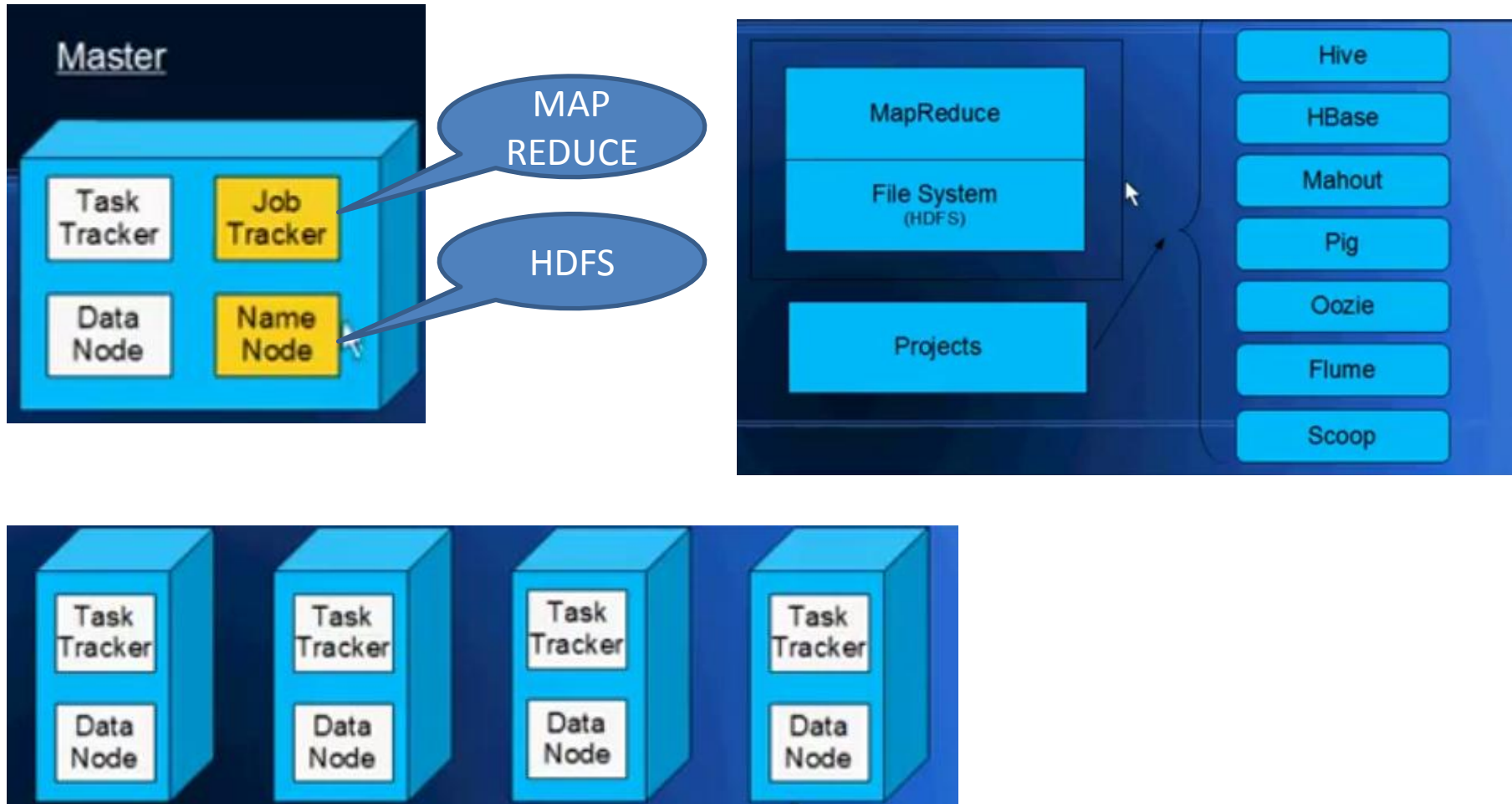
- pas dimensionné pour travailler avec des petits fichiers.
- Les écritures uniquement permises en fin du fichier
- HDFS est un système de fichier en espace utilisateur a contrario des systèmes de fichier directement inclus dans le noyau de systèmes d'exploitation (Ext3, NTFS...). Cette caractéristique ne lui permet pas d'être "monté" à l'égal de l'Ext3 ou du ReiserFs. Il peut cependant l'être grâce à l'outil WebDAV.

# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

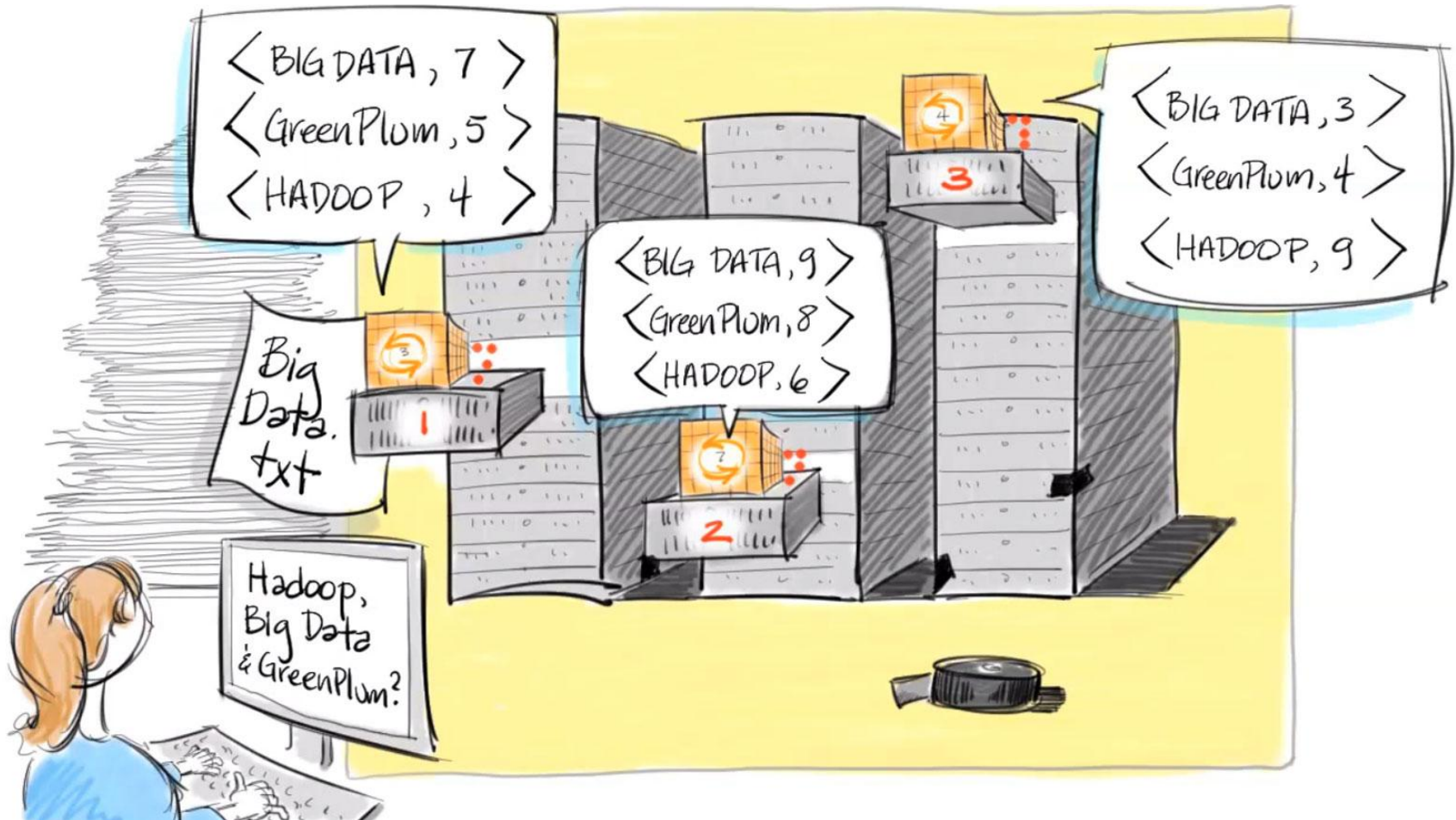
### HADDOP / MAPREDUCE ET « L'ECOSYSTEM »



# Information On Demand 2013



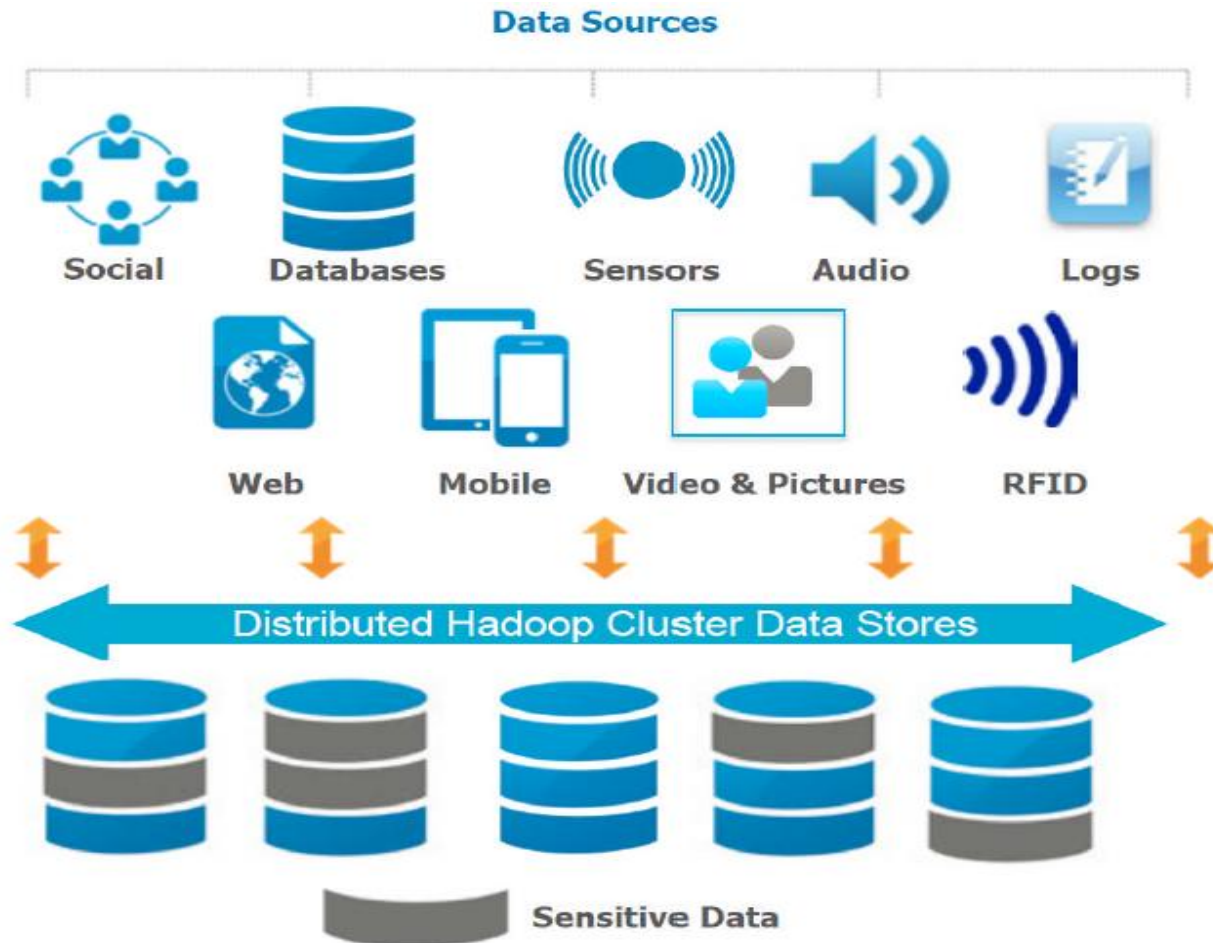
## BIG DATA : LES ACTEURS DE L'OPEN SOURCE HADDOP / MAPREDUCE PAR L'EXEMPLE



# Information On Demand 2013



**BIG DATA : LES ACTEURS DE L'OPEN SOURCE**  
**HADDOP / MAPREDUCE PAR L'EXEMPLE**





# Information On Demand 2013



**BIG DATA : LES ACTEURS DE L'OPEN SOURCE**  
**HADDOP / MAPREDUCE PAR L'EXEMPLE**

## NCDC Sensor Data Format

Year Temperature (C) Quality

```
0029029810999991901110806004+59500+020350FM-12+002699999V0203201N022119999999N00000001NS+00441-9999909823:
0035029810999991901110813004+59500+020350FM-12+002699999V0203201N022119999999N00000001NS+00061-9999909835:
0035029810999991901110820004+59500+020350FM-12+002699999V0203401N021119999999N00000001NS+00111-9999909858:
0029029810999991901110906004+59500+020350FM-12+002699999V0203201N020119999999N00000001NS+00171-9999909910:
0029029810999991901110913004+59500+020350FM-12+002699999V0203401N022119999999N00000001NS+00221-9999909976:
0035029810999991901110920004+59500+020350FM-12+002699999V0203401N014919999999N00000001NS+00281-9999910044:
0029029810999991901111006004+59500+020350FM-12+002699999V0202701N014919999999N00000001NS+00331-9999910019:
0029029810999991901111013004+59500+020350FM-12+002699999V0202701N012919999999N00000001NS+00391-9999909984:
0029029810999991901111020004+59500+020350FM-12+002699999V0202701N008719999999N00000001NS+00331-9999909946:
0029029810999991901111106004+59500+020350FM-12+002699999V0203601N007219999999N00000001NS+00331-9999909872:
0029029810999991901111113004+59500+020350FM-12+002699999V0200201N011819999999N00000001NS+00111-9999909907:
0029029810999991901111120004+59500+020350FM-12+002699999V0200501N008719999999N00000001NS+00111-9999909958:
0029029810999991901111206004+59500+020350FM-12+002699999V0203401N015919999999N00000001NS+00111-9999909995:
0029029810999991901111213004+59500+020350FM-12+002699999V0203201N018019999999N00000001NS+00001-9999910006:
```

# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

### HADDOP / MAPREDUCE PAR L'EXEMPLE

A screenshot of the Eclipse IDE interface. The title bar reads "Java - ClimateAnalysis/src/MaxTemperature.java - Eclipse". The menu bar includes File, Edit, Source, Refactor, Navigate, Search, Project, Run, Window, and Help. The toolbar contains various icons for file operations, running, and debugging. The editor shows three tabs: "MaxTemperature.java", "MaxTemperatureMapper.java", and "MaxTemperatureReducer.java". The "MaxTemperature.java" tab is active, displaying the following code:

```
1 // cc MaxTemperature Application to find the maximum temperature in the weather dataset
2
3 import org.apache.hadoop.fs.Path;
4 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
5 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
6 import org.apache.hadoop.io.IntWritable;
7 import org.apache.hadoop.io.Text;
8 import org.apache.hadoop.mapreduce.Job;
9
10 public class MaxTemperature {
11
12     public static void main(String[] args) throws Exception {
13         if (args.length != 2) {
14             System.err.println("Usage: MaxTemperature <input path> <output path>");
15             System.exit(-1);
16         }
17
18         Job job = new Job();
19         job.setJarByClass(MaxTemperature.class);
20         job.setJobName("Max temperature");
21     }
22 }
```

# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

### HADDOP / MAPREDUCE PAR L'EXEMPLE

A screenshot of an Eclipse IDE window. The title bar reads "Java - ClimateAnalysis/src/MaxTemperature.java - Eclipse". The menu bar includes "File", "Edit", "Source", "Refactor", "Navigate", "Search", "Project", "Run", "Window", and "Help". The toolbar contains various icons for file operations, running, and debugging. A "Quick Access" search bar is visible on the right. Overlaid on the IDE is a table with two columns: "Input Format" and "Description".

Input Format	Description
TextInputFormat	Default behavior for FileInputFormat. Each line of text is a single record, passed as the Value.
KeyValueTextInputFormat	Key & Value are Text, separated by a delimiter (\t by default).
SequenceFileInputFormat<K, V>	Key & Value types are user-defined. Data is stored in a compressed binary format, optimized for Hadoop jobs.
NLineInputFormat	Guarantees each split to have N lines.
Custom	User-defined class to specify record format of input data files.



# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE HADDOP / MAPREDUCE PAR L'EXEMPLE

A screenshot of the Eclipse IDE interface. The title bar reads "Java - ClimateAnalysis/src/MaxTemperatureMapper.java - Eclipse". The menu bar includes "File", "Edit", "Source", "Refactor", "Navigate", "Search", "Project", "Run", "Window", and "Help". The toolbar contains various icons for file operations, running, and debugging. The "Quick Access" search bar is visible on the right. The left sidebar shows a project explorer with three files: "MaxTemperature.java", "MaxTemperatureMapper.java" (selected), and "MaxTemperatureReducer.java". The main editor displays the code for "MaxTemperatureMapper.java". The code includes imports for "java.io.IOException", "org.apache.hadoop.io.IntWritable", "org.apache.hadoop.io.LongWritable", "org.apache.hadoop.io.Text", and "org.apache.hadoop.mapreduce.Mapper". The class "MaxTemperatureMapper" extends "Mapper<LongWritable, Text, Text, IntWritable>". It has a static final int "MISSING" set to 9999. The "map" method is annotated with "@Override" and throws "IOException" and "InterruptedException". It processes a "LongWritable" key and a "Text" value, extracting a substring and parsing it as an integer, skipping lines that start with a plus sign. The code is as follows:

```
1 // cc MaxTemperatureMapper Mapper for maximum temperature example
2 import java.io.IOException;
3
4 import org.apache.hadoop.io.IntWritable;
5 import org.apache.hadoop.io.LongWritable;
6 import org.apache.hadoop.io.Text;
7 import org.apache.hadoop.mapreduce.Mapper;
8
9
10 public class MaxTemperatureMapper
11     extends Mapper<LongWritable, Text, Text, IntWritable> {
12     //FORM: <InKeyType, InValueType, OutKeyType, OutValueType>
13
14     private static final int MISSING = 9999;
15
16     @Override
17     public void map(LongWritable key, Text value, Context context)
18         throws IOException, InterruptedException {
19
20         String line = value.toString();
21         String year = line.substring(15, 19);
22         int airTemperature;
23         if (line.charAt(87) == '+') { // parseInt doesn't like leading plus signs
```



# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE

### HADDOP / MAPREDUCE PAR L'EXEMPLE

```
Java - ClimateAnalysis/src/MaxTemperatureReducer.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help

MaxTemperature.java  MaxTemperatureMapper.java  MaxTemperatureReducer.java

1 // cc MaxTemperatureReducer Reducer for maximum temperature example
2 import java.io.IOException;
3
4
5 import org.apache.hadoop.io.IntWritable;
6 import org.apache.hadoop.io.Text;
7 import org.apache.hadoop.mapreduce.Reducer;
8
9 public class MaxTemperatureReducer
10     extends Reducer<Text, IntWritable, Text, IntWritable> {
11     //FORM: <InKeyType, InValueType, OutKeyType, OutValueType>
12
13     @Override
14     public void reduce(Text key, Iterable<IntWritable> values, Context context)
15         throws IOException, InterruptedException {
16
17         int maxValue = Integer.MIN_VALUE;
18         for (IntWritable value : values) {
19             maxValue = Math.max(maxValue, value.get());
20         }
21         context.write(key, new IntWritable(maxValue));
22     }
23 }
```

# Information On Demand 2013



## BIG DATA : LES ACTEURS DE L'OPEN SOURCE HADDOP / MAPREDUCE PAR L'EXEMPLE

```
root@node-36:/maprcluster/user/ClimateAnalysis
Desktop Cluster
root@node-36:/maprcluster/user/ClimateAnalysis 94x28
ClimateAnalysis.jar input_samplefile input_sampleset
[root@node-36 ClimateAnalysis]# cat input_samplefile/sample.txt
0067011990999991950051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9+00001+
9999999999
0043011990999991950051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9+00221+
9999999999
0043011990999991950051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9-00111+
9999999999
0043012650999991949032412004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9+01111+
9999999999
0043012650999991949032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9+00781+
9999999999[root@node-36 ClimateAnalysis]#
[root@node-36 ClimateAnalysis]#
[root@node-36 ClimateAnalysis]# hadoop jar ClimateAnalysis.jar MaxTemperature /user/ClimateAna
lysis/input_samplefile /user/ClimateAnalysis/maxtemp_out1
12/12/14 11:37:37 INFO fs.JobTrackerWatcher: Current running JobTracker is: node-37.lab/10.10.
20.37:9001
12/12/14 11:37:37 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. A
pplications should implement Tool for the same.
12/12/14 11:37:38 INFO input.FileInputFormat: Total input paths to process : 1
12/12/14 11:37:38 WARN snappy.LoadSnappy: Snappy native library not loaded
12/12/14 11:37:38 INFO mapred.JobClient: Creating job's output directory at /user/ClimateAnaly
sis/maxtemp_out1
12/12/14 11:37:38 INFO mapred.JobClient: Creating job's user history location directory at /us
er/ClimateAnalysis/maxtemp_out1/logs
12/12/14 11:37:38 INFO mapred.JobClient: Running job: job_201212131811_0009
12/12/14 11:37:39 INFO mapred.JobClient: map 0% reduce 0%
```

# Information On Demand 2013

## BIG DATA : LES ACTEURS DE L'OPEN SOURCE HADDOP / MAPREDUCE PAR L'EXEMPLE

A screenshot showing a terminal window on the left and a Mozilla Firefox browser window on the right. The terminal displays a command prompt and a list of data. The browser window shows the MapR Control System interface with tabs for JobTracker Status, TaskTracker Status, and HBase Master. The active tab displays a table of data.

mapr@mapr-desktop: ~/eclipse\_workspace/ClimateAnalysis

mapr@mapr  
/home/map  
mapr@mapr  
bin Build  
mapr@mapr  
r/user/Cl  
mapr@mapr  
is/maxtem  
[1] 23101  
mapr@mapr

1901 239  
1902 156  
1903 172  
1904 172  
1905 178  
1906 283  
1907 272  
1908 283  
1909 256  
1910 239  
1911 239  
1912 278  
1913 300  
1914 300  
1915 278  
1916 278  
1917 272  
1918 267  
1919 283  
1920 272  
1921 233  
1922 267  
Done

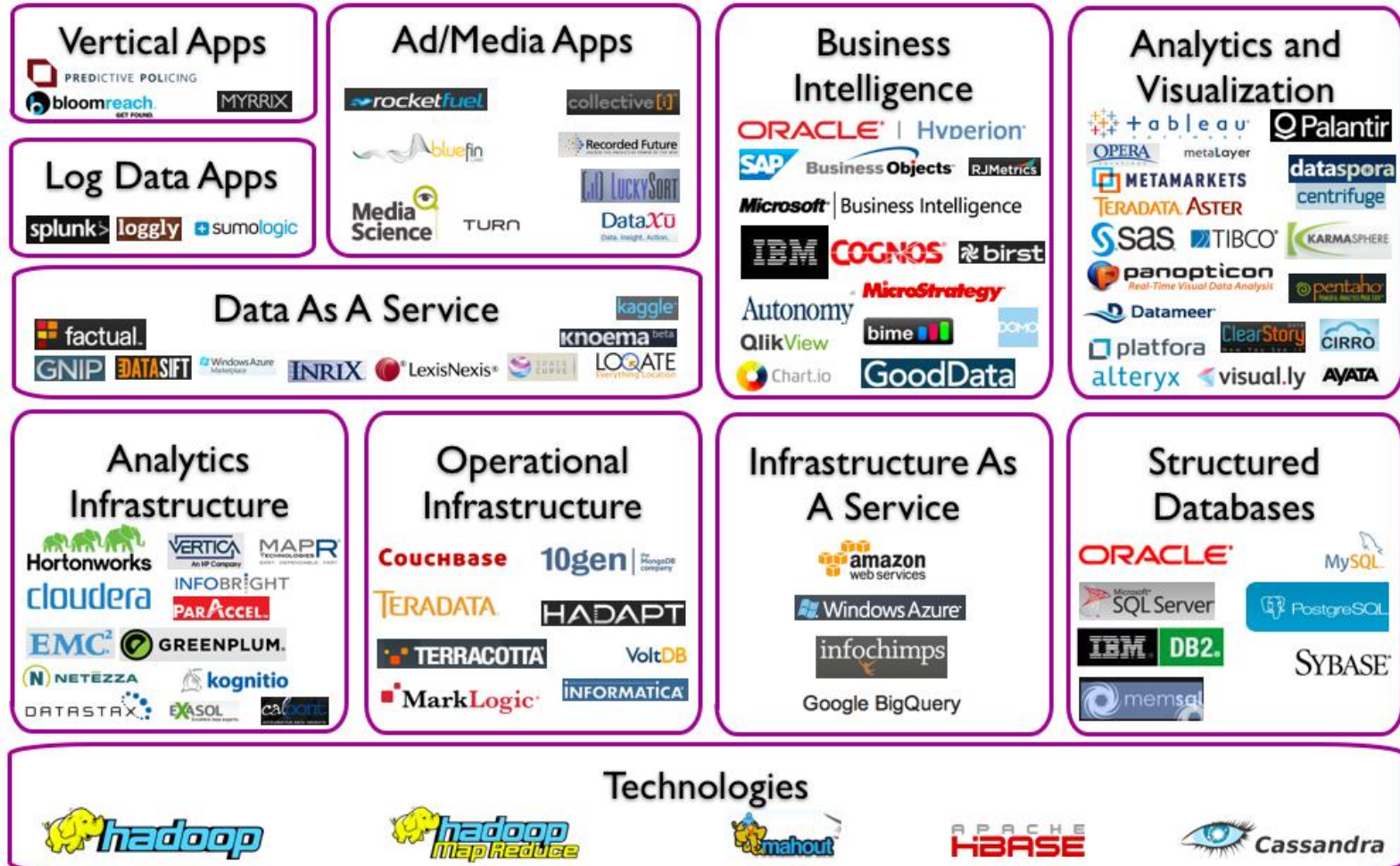
Mozilla Firefox  
File Edit View History Bookmarks Tools Help  
file:///maprcluster/user/ClimateAnalysis  
MapR Control System JobTracker Status TaskTracker Status HBase Master  
file:///maprclus...ut2/part-r-00000



# Information On Demand 2013



## BIG DATA – HADOOP : C'EST TOUT ?





# Information On Demand 2013



**BIG DATA – HADOOP : LES DISTRIBUTIONS**





## BIG DATA – HADOOP : LES DISTRIBUTIONS

Apache Hadoop est un projet open source. Cela a beaucoup d'avantages.

Cependant, un réel problème est d'obtenir un support commercial pour un projet open source tel qu'Apache Hadoop. Habituellement, les entreprises offrent uniquement du support sur leurs produits, pas sur un projet open source (ce n'est pas un problème qui concerne seulement Hadoop, mais beaucoup de projets open source).

La version courante du projet Apache Hadoop inclut ces modules :

**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.

Cependant, l'écosystème Hadoop ne contient pas uniquement Hadoop, mais beaucoup d'autres projets Apache tels que :

**Pig, Hive, Hbase, Sqoop, Flume, Zookeeper** et bien d'autres



## BIG DATA – HADOOP : LES DISTRIBUTIONS

**Pig:** une plate-forme pour analyser des ensembles de gros volumes de données. Cela consiste en un langage haut-niveau pour l'expression de programmes d'analyse de données, couplé à une infrastructure pour évaluer ces programmes.

**Hive:** un système d'entrepôt de données pour Hadoop qui offre un langage de requête de type SQL pour faciliter les agrégations, le requêtage ad-hoc et l'analyse de gros volumes de données stockés dans des systèmes de fichiers compatibles Hadoop.

**Hbase:** un stockage de données distribué et scalable dédié au big data avec un accès direct et une lecture/écriture temps réel.

**Sqoop:** un outil conçu pour transférer efficacement une masse de données entre Apache Hadoop et un stockage de données structuré tel que les bases de données relationnelles.

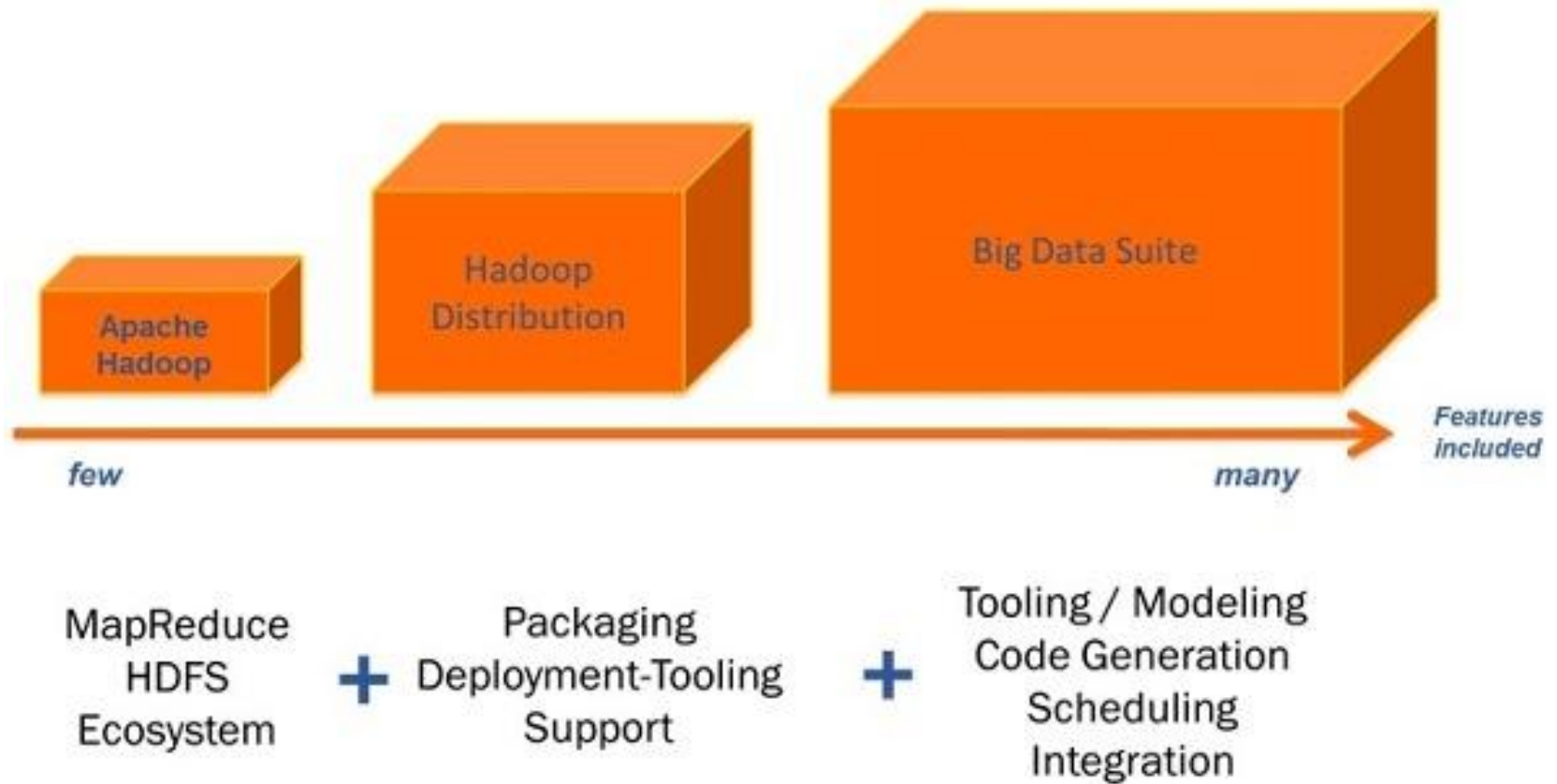
**Flume:** un service distribué, fiable et disponible pour collecter efficacement, agréger et déplacer une grande quantité de logs.

**Zookeeper:** un service centralisé pour maintenir les configurations, la nomenclature, pour fournir une synchronisation distribuée et des services groupés.

# Information On Demand 2013



## BIG DATA – HADOOP : LES DISTRIBUTIONS

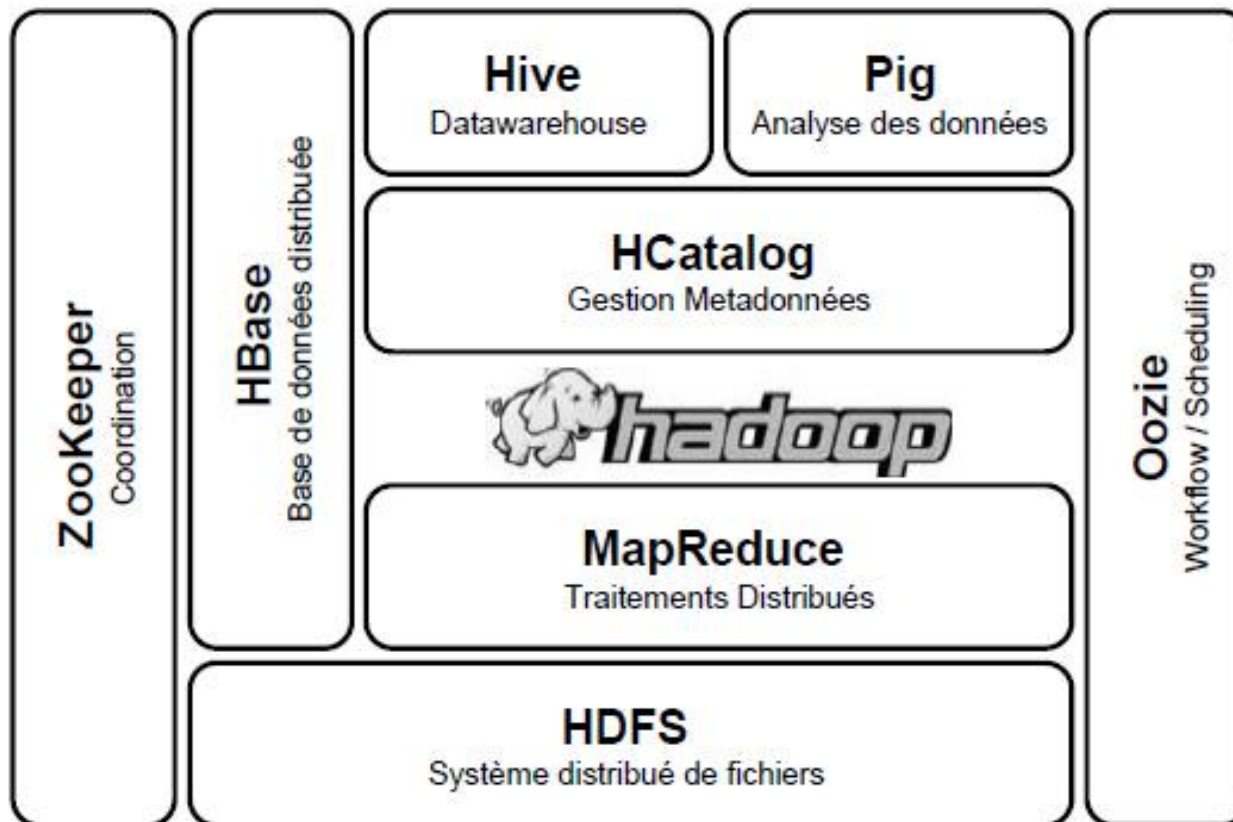




## BIG DATA – HADOOP : LES DISTRIBUTIONS

### DISTRIBUTION APACHE HADOOP

Ces projets doivent être installés et intégrés manuellement dans Hadoop en portant une grande attention aux différentes versions et releases. Malheureusement, toutes les releases ne fonctionnent pas parfaitement bien ensemble. Vous devez comparer les notes de livraisons et vous débrouiller par vous-même.





# Information On Demand 2013



## BIG DATA – HADOOP : LES DISTRIBUTIONS LES DISTRIBUTIONS PACKAGEES

Elles offrent des outils et un support commercial, ce qui réduit beaucoup les efforts à mettre en oeuvre, pas seulement pour le développement mais aussi pour l'opérationnel.

Une distribution contient différents projets de l'écosystème Hadoop. Ceci assure que toutes les versions utilisées fonctionnent ensemble sans problèmes.

Il y a des releases régulières avec des versions mises à jour de différents projets.

En plus du package, les fournisseurs de distribution offrent des outils graphiques pour le déploiement, l'administration et le monitoring des clusters Hadoop. De cette façon, il est beaucoup plus facile d'installer, de gérer et de surveiller les clusters. L'investissement est beaucoup réduit.

Il y a plus ou moins trois grandes distributions Hadoop qui en ce moment se distinguent : HortonWorks, Cloudera et MapR.

Bien que dans le même temps, d'autres distributions Hadoop voient aussi le jour.





## **BIG DATA – HADOOP : LES DISTRIBUTIONS**

### **LES PACKAGE BIG DATA**

Au dessus d'Apache Hadoop ou d'une distribution Hadoop, vous pouvez utiliser un package Big Data. Ce dernier supporte souvent différentes distributions Hadoop. Cependant, certains fournisseurs implémentent leur propre solution Hadoop. De toute façon un package Big Data ajoute plusieurs autres caractéristiques aux distributions pour le traitement des données.

Un package Big Data offre tout un panel d'outils graphiques :

- De modélisation, de débogage et d'optimisation du code Mapreduce.
- De planification et la surveillance des jobs
- Des connecteurs de toutes sortes permettant l'intégration de données SQL, NoSQL, des médias sociaux tels que Twitter ou Facebook, des middleware de messagerie ou des données de produits B2B tels que Salesforce ou SAP etc ...
- De traitement en temps réel de vos données
- De cartographie
- De monitoring
- De réalisation de vos reporting
- D'analyse de textuel et prédictive

.

# Information On Demand 2013



## BIG DATA : LES DISTRIBUTIONS IBM LES PACKAGES BIGDATA IBM



### BigInsights Basic Edition

- Apache Hadoop
- Web-based mgmt console
- Jaql
- Integrated install



### BigInsights Quick Start Edition

- Big Sheets
- Text Analytics
- Big SQL
- Workload optimization
- Pre-built apps
- Enhanced security
- Dev tools
- RDBMS Connectors
- Mgmt tools



### BigInsights Enterprise Edition

- Quick Start features  
PLUS:
- Accelerators
  - Enterprise Integration
  - Production support
  - Production-ready features



### PureData System for Hadoop

- Appliance simplicity
- Faster deployment than custom-built solutions<sup>1</sup>
- First appliance with built-in analytics accelerator<sup>2</sup>
- Only Hadoop system with built-in archiving tools<sup>2</sup>

# Information On Demand 2013



## BIG DATA : LES DISTRIBUTIONS IBM LES PACKAGES BIGDATA IBM



### IBM InfoSphere BigInsights

#### Visualization & Discovery

BigSheets

Dashboard &  
Visualization

#### Applications & Development

Apps

Text Analytics

MapReduce

Workflow

Pig & Jaql

Hive

#### Administration

Admin Console

Monitoring

#### Integration

JDBC

Netezza

DB2

Streams

DataStage

Guardium

Platform  
Computing

Cognos

Flume

Sqoop

#### Advanced Analytic Engines

Adaptive Algorithms

Text Processing Engine &  
Extractor Library)

R

#### Workload Optimization

Integrated  
Installer

Enhanced  
Security

Splittable Text  
Compression

Adaptive  
MapReduce

ZooKeeper

Oozie

Jaql

Flexible  
Scheduler

HCatalog

Lucene

Pig

Hive

Index

#### Runtime / Scheduler

MapReduce

Symphony

Symphony AE

#### Management

Security

Audit & History

Lineage

#### Data Store

HBase

#### File System

HDFS

GPFS FPO

Open Source

IBM

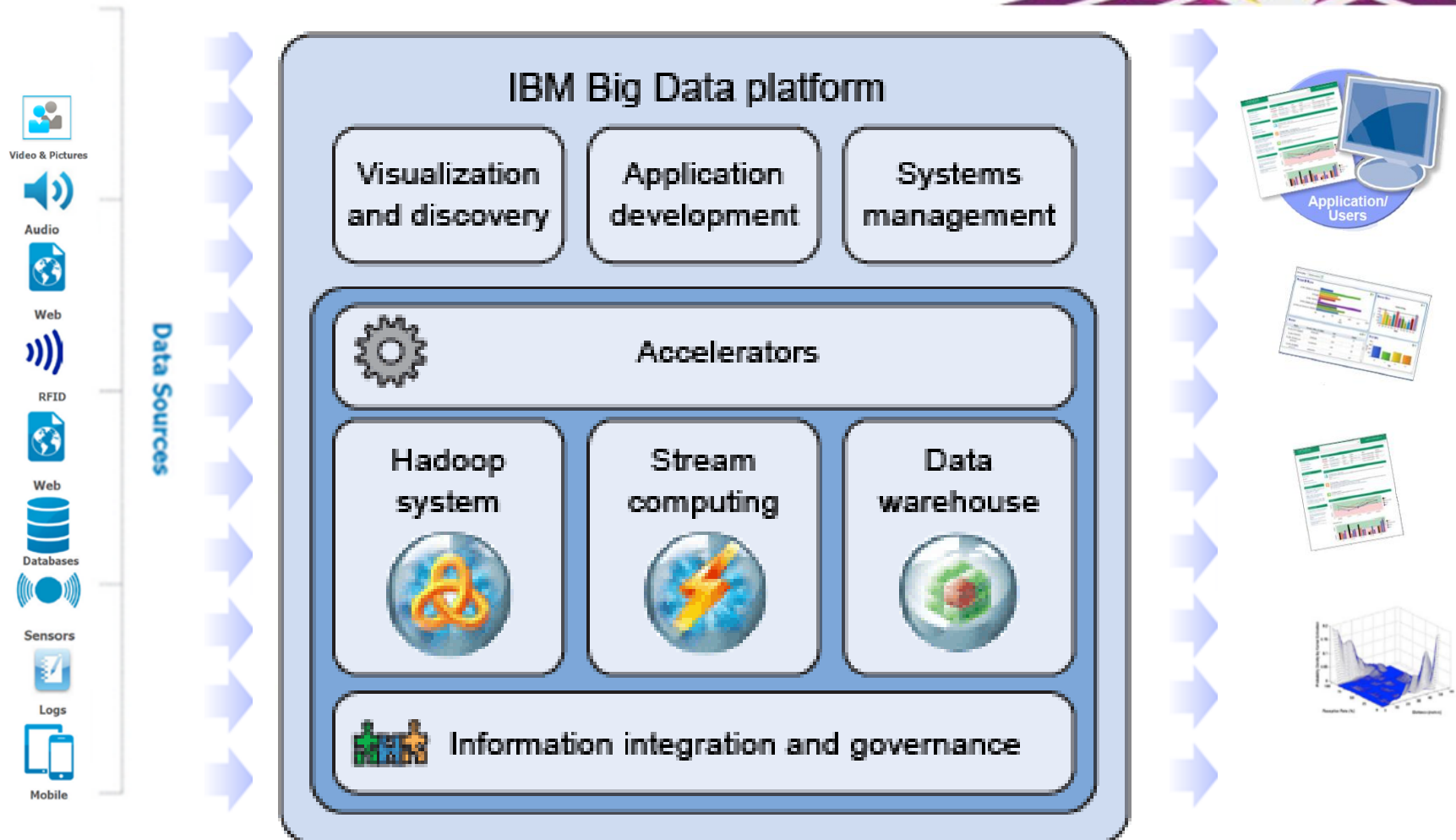
Optional



# Information On Demand 2013

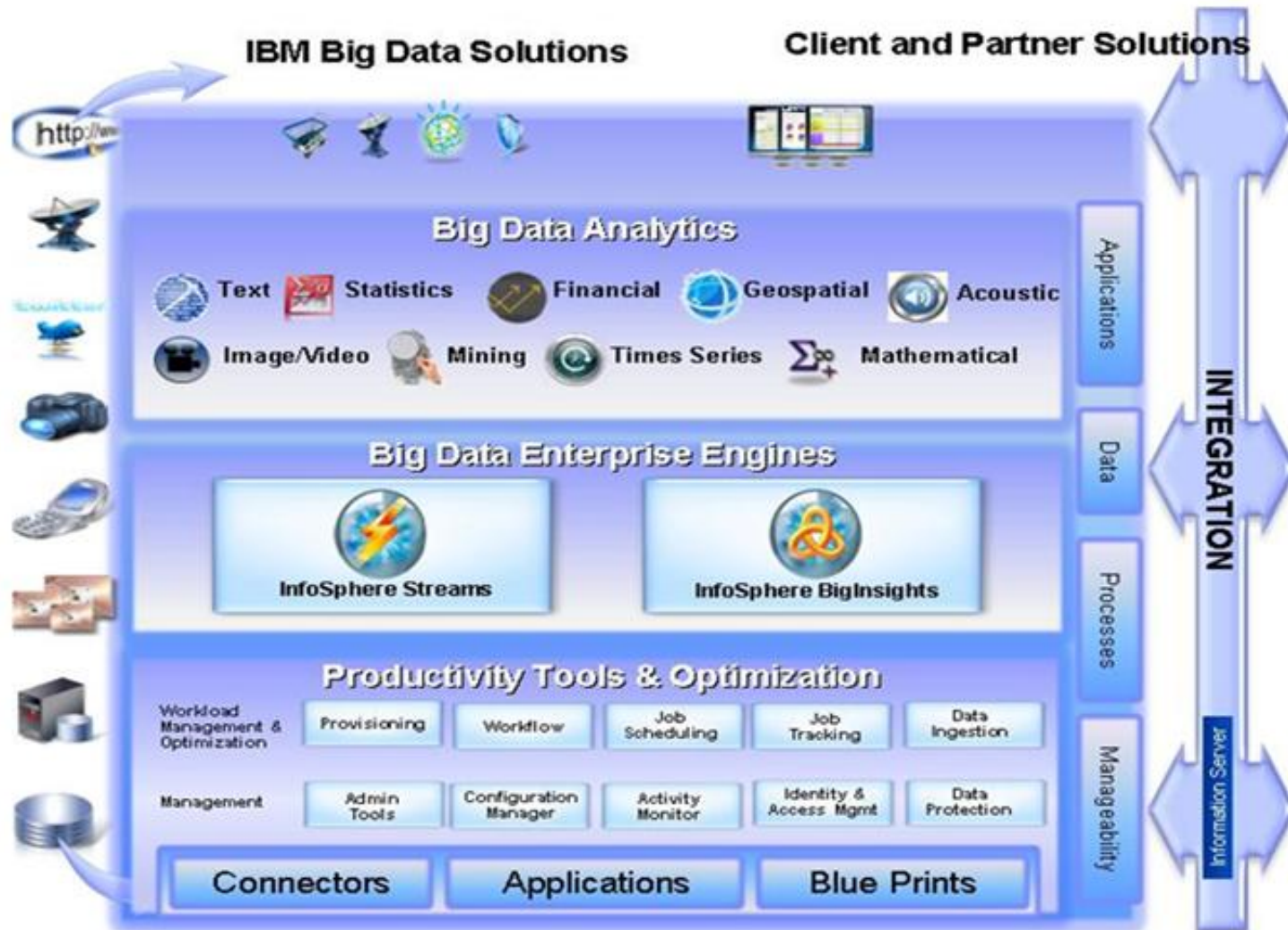


## BIG DATA : LES DISTRIBUTIONS IBM LES PACKAGES BIGDATA IBM



# Information On Demand 2013

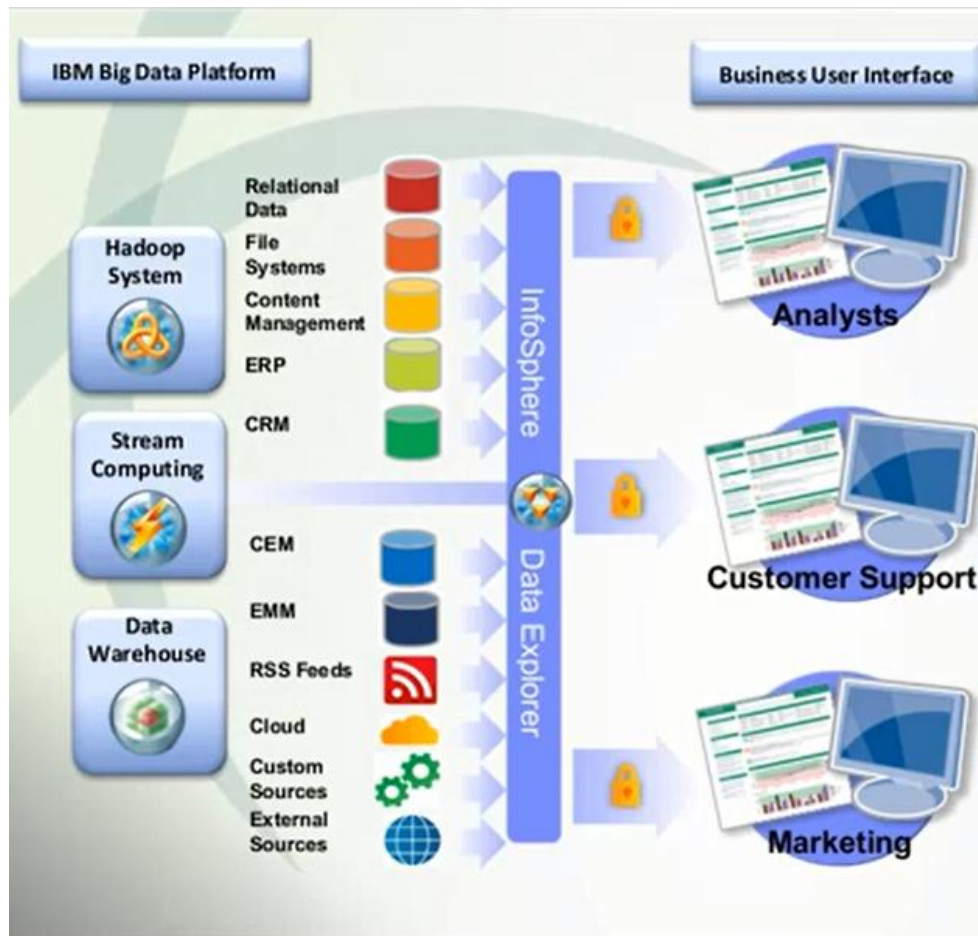
## BIG DATA : LES DISTRIBUTIONS IBM LES PACKAGES BIGDATA IBM



# Information On Demand 2013



## BIG DATA : LES DISTRIBUTIONS IBM



- **Discover the Data**
  - Provide **discovery and navigation**
  - **Connect securely** to the applications that manage the data—regardless of location
- **Analyze Structured & Unstructured Data**
  - **Reveal themes & visualize relationships**
  - Identify the **value** of the data
  - Recognize **users** of the data
  - Establish **context** of data usage
- **Collaborate on the Data**
  - Augment the data with **user knowledge**
  - Create **personalized views** of the data
  - Identify ongoing **user and system integration points**

# Information OnDemand 2013



## **BIG DATA : LES DISTRIBUTIONS IBM**

### **IBM InfoSphere BigInsights**

IBM InfoSphere BigInsights Enterprise Edition fournit des fonctions d'analyse de volumes massifs de données sur une plateforme destinée aux entreprises.

Il associe la solution open source Apache Hadoop à des fonctionnalités d'entreprise et à l'intégration et fournit une analyse à grande échelle, caractérisée par sa résilience et sa tolérance aux pannes.

Le logiciel prend en charge les données structurées, non structurées et semi-structurées dans leur format natif, offrant ainsi une flexibilité maximale.

Conçu pour la performance et la facilité d'utilisation grâce à des fonctions optimisées de performance, de visualisation, des outils de développement riches et des fonctions analytiques puissantes.

Offre des fonctions de gestion, de sécurité et de fiabilité qui prennent en charge des déploiements à grande échelle et accélèrent la réalisation de la valeur.

S'intègre à IBM et à d'autres solutions d'information pour simplifier et améliorer les tâches de manipulation de données.



# Information On Demand 2013



## BIG DATA : LES DISTRIBUTIONS IBM

### IBM Infosphere Streaming Data

Avec plus de 6 milliards d'abonnés mobiles dans le monde, les opérateurs doivent analyser les grandes quantités de données issues de leurs réseaux. Les laboratoires IBM Research ont donc mis au point la solution InfoSphere Stream. L'outil est capable d'analyser des pétaoctets de données en continu pour comprendre comment les utilisateurs emploient les différents services et quelles sont leurs préférences. L'application intègre également des analyses des réseaux sociaux. Sprint, troisième opérateur de téléphonie mobile aux Etats-Unis, utilise ces technologies pour capturer et interpréter les données réseaux (localisation, appels coupés, interruption de service, performances,...) et améliorer l'expérience utilisateur.

InfoSphere Streams contient également une interface graphique qui permet, par simple glisser/déposer de sets de données, de visualiser graphiquement des flux de processus complexes.

# Information On Demand 2013



## BIG DATA : LES DISTRIBUTIONS IBM

### IBM Digital Analytics Accelerator

Les responsables marketing (CMO) doivent aujourd'hui analyser les demandes des clients provenant des medias sociaux, des terminaux mobiles et des canaux traditionnels, et aligner ces demandes avec les développements produits.

IBM Digital Analytics Accelerator, intégré à la nouvelle plateforme PureData System, leur vient en aide en analysant le ressenti d'un client pour lui adresser des campagnes et des promotions personnalisées et prédire ses besoins futurs afin de réduire le risque de le perdre. La nouvelle solution va, pour cela, analyser les médias sociaux, le trafic web de la société et les communications des clients. Elle intègre les technologies Netezza et Unica qui permettent, selon IBM, d'analyser plusieurs pétaoctets de données en quelques minutes.

# Information On Demand 2013



## **BIG DATA : LES DISTRIBUTIONS IBM**

### **IBM InfoSphere Data Explorer**

IBM InfoSphere BigInsight est un outil d'analyse de données structurées et non structurées, basé sur la plateforme Hadoop. Il intègre désormais le nouvel InfoSphere Data Explorer qui offre les capacités de fédération de données issues du rachat de Vivisimo. Le soft repère et explore automatiquement les données disponibles quel que soit leur emplacement et en ressort les relations, identifie leur valeur et les replace dans leur contexte d'usage.

<http://www.ibm.com/developerworks/library/bd-exploration/>

# Information On Demand 2013



## BIG DATA : LES DISTRIBUTIONS IBM

### IBM SmartCloud Analytics Answer

Les technologies d'analyse prédictive d'IBM ne sont pas à la portée de toutes les entreprises. Le géant américain admet d'ailleurs volontiers que la grande majorité des déploiements se font dans de grandes organisations disposant des moyens financiers pour investir dans le hardware, les logiciels et la formation. L'offre Analytics Answer est donc un service d'analyse à la demande s'adressant aux divisions métiers des PME, qui souhaitent bénéficier des avantages de ces outils sans avoir à réaliser d'investissements trop importants.

Les utilisateurs pourront accéder au service via le portail web SmartCloud, entrer leurs données et leurs questions sur une problématique précise, et obtenir une réponse rapide. Une société d'assurance peut, par exemple, demander à déterminer quels sont ses clients les plus susceptibles de renouveler une police particulière.



# Information On Demand 2013



**IBM BIG DATA : LES DISTRIBUTIONS IBM**  
**EXEMPLE D'APPLICATION : LE BINGO ONLINE**



# Information On Demand 2013



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE GENERALITES

- ▶ TRES POPULAIRE EN EUROPE
- ▶ PLUS DE 3 MILLIONS DE JOUEURS EN ANGLETERRE (UK)
- ▶ SOCIAL : ENORMEMENT DE TCHAT ET D'INTERACTION ENTRE JOUEURS
- ▶ MARCHE TRES CONCURRENTIEL
- ▶ OFFRE DE PUBLICITAIRES
- ▶ UN SENS DE LA COMMUNAUTE CONTRIBUANT A LA FIDELISATION DE LA CLIENTELE

# Information On Demand **2013**



## **IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE LE BESOIN DU CLIENT**

- ▶ MOYEN FLEXIBLE ET EVOLUTIF DE DEPLOIEMENT ET D'ANALYSE EN TEMPS REEL DES DONNES DES SALLES DE TCHAT, DES DONNES DE TYPE SOCIALES, DES LOGS ET DONNES NON STRUCTUREES.
- ▶ CONTRÔLE SUR L'AMBIANCE DES SALLES DE TCHAT

# Information On Demand **2013**



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE ANALYSE DU LANGUAGE DES TCHAT ROOMS

- ▶ HIYA Hello Everyone
- ▶ WB Welcome Back
- ▶ LTNS Long Time No See
- ▶ ROOMIE Another Member In Chat Room
- ▶ XOXOXO Hugs and Kisses
- ▶ GM Good Morning
- ▶ GL Good Luck!
- ▶ GL e1Good Luck Everyone!
- ▶ WTG Way To Go!
- ▶ 1TG 1 To Go!
- ▶ 2TG 2 To Go!
- ▶ 3TG 3 To Go!
- ▶ TUVN Thank You Very Much
- ▶ • TTFN Ta Ta For Now
- ▶ • TTYL Talk To You Later
- ▶ • TC Take care
- ▶ • CYA See You Later
- ▶ • LOL Laugh Out Loud
- ▶ • L8R Later
- ▶ • ROFL Rolling On Floor Laughing



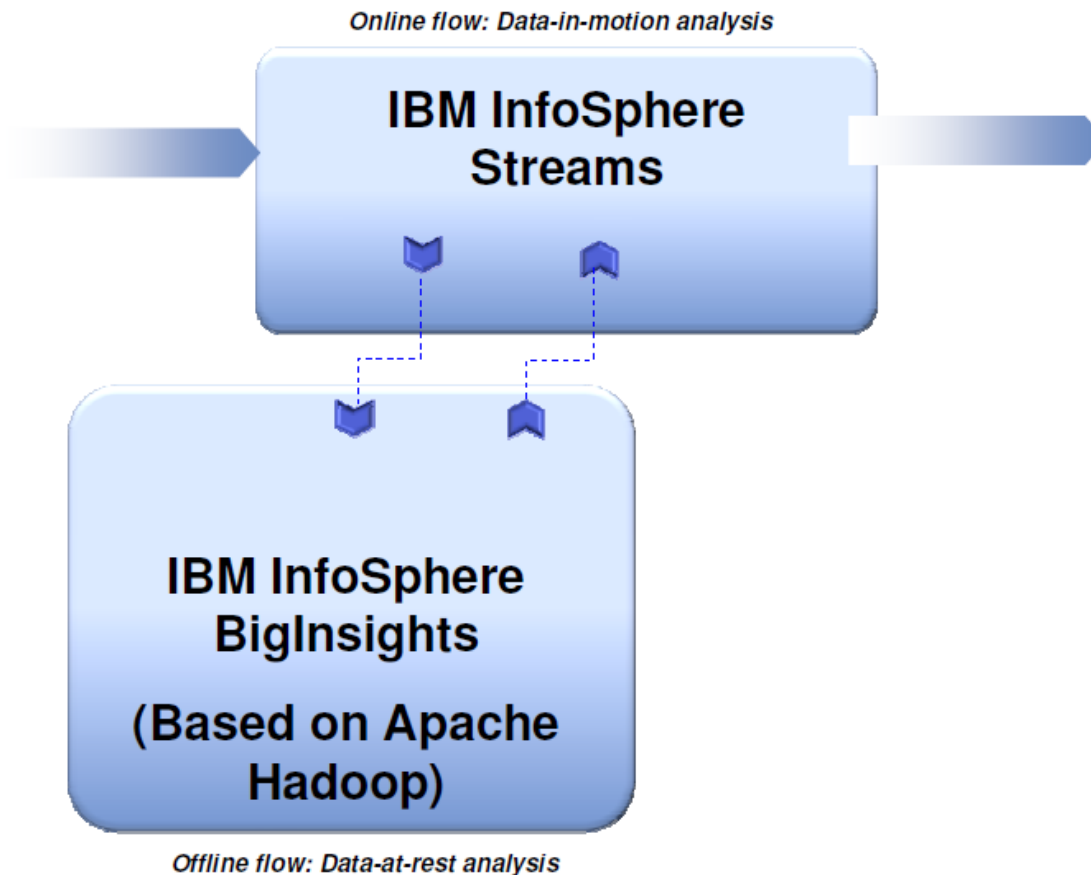
# Information On Demand 2013



IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE

## STREAMS

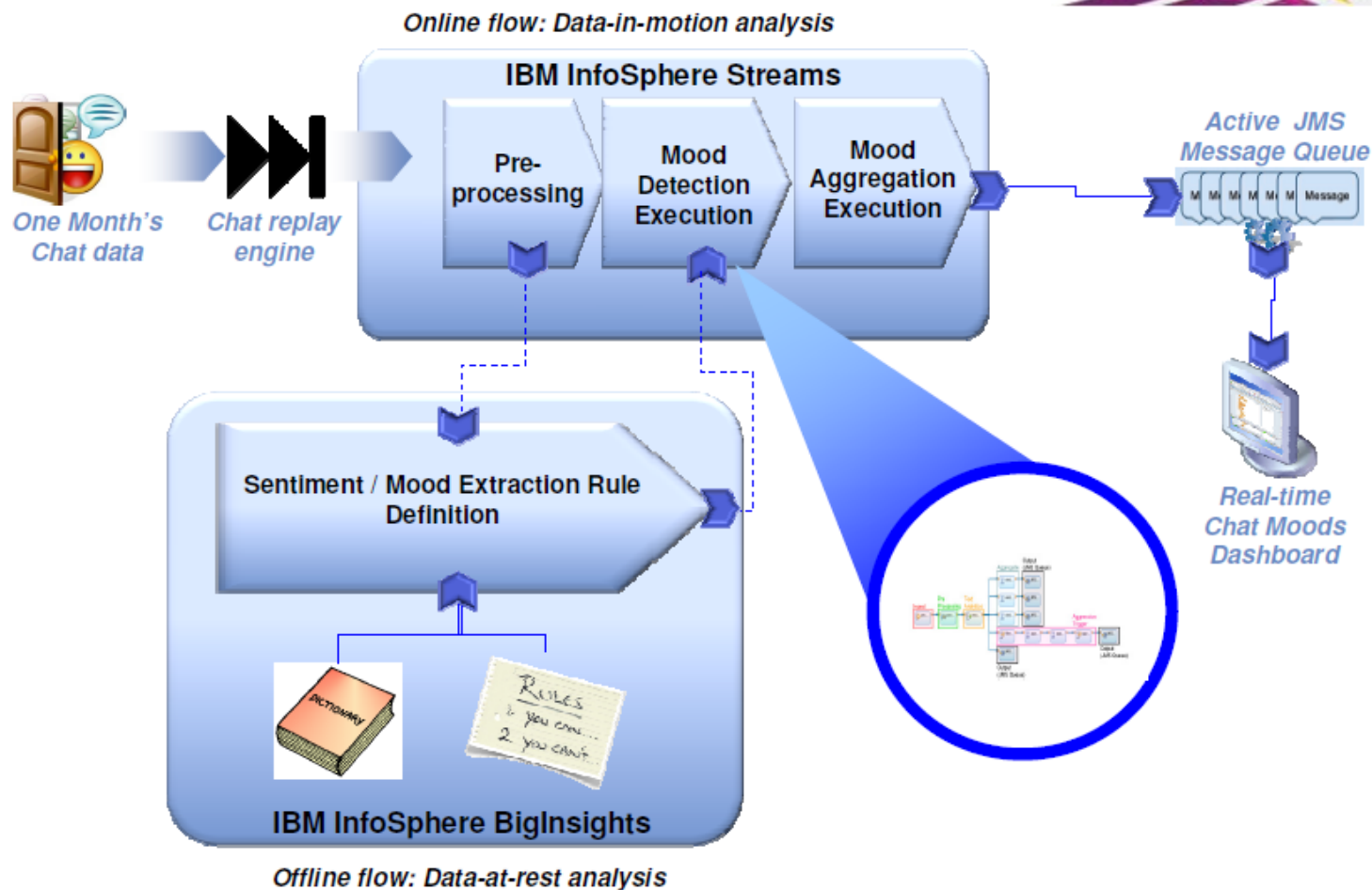
### IBM Big Data Platform



# Information On Demand 2013



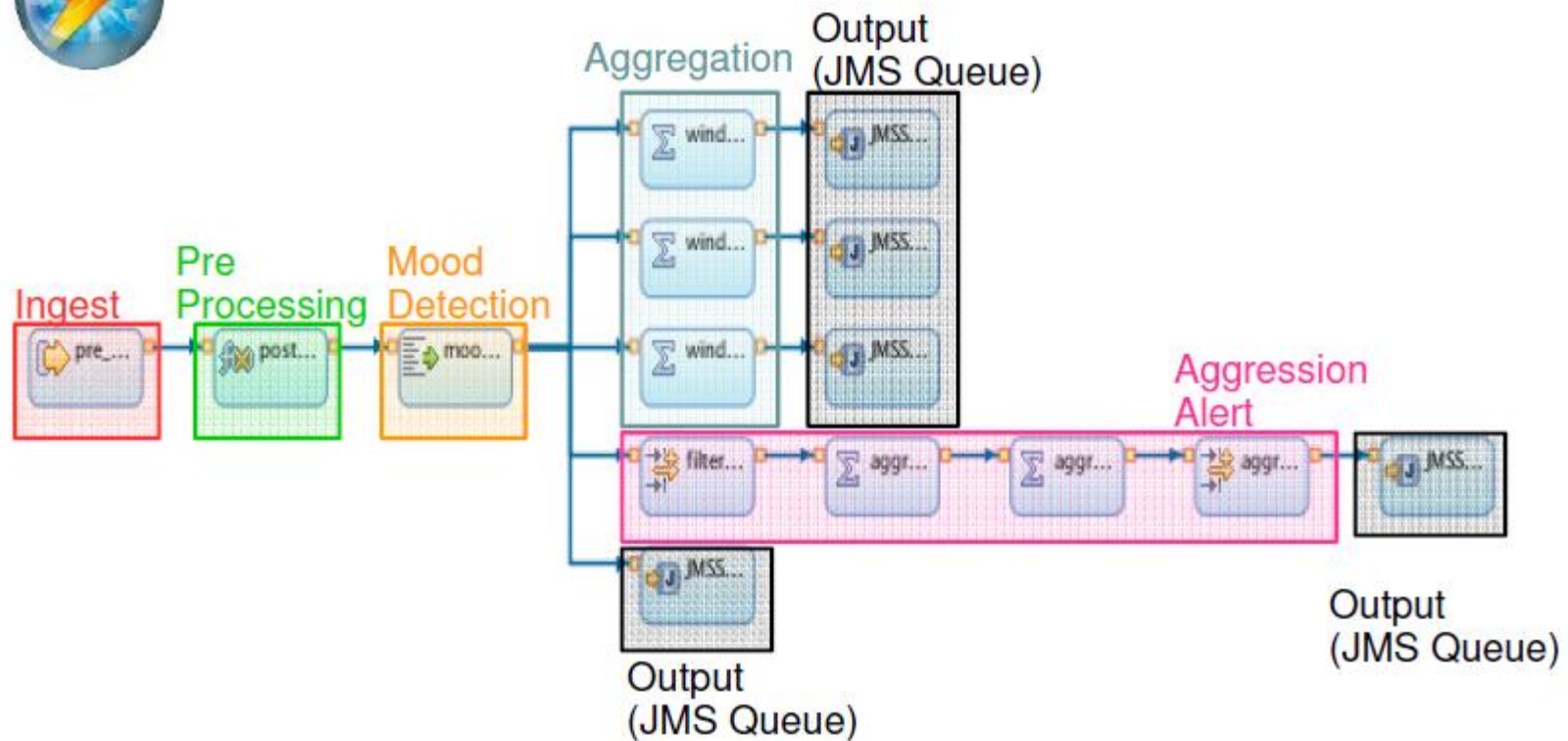
## IBM BIG DATA : CAS D'ECOLE AVEC LE BINGO ONLINE ANALYSE DE L'HUMEUR EN TEMPS REEL



# Information On Demand 2013



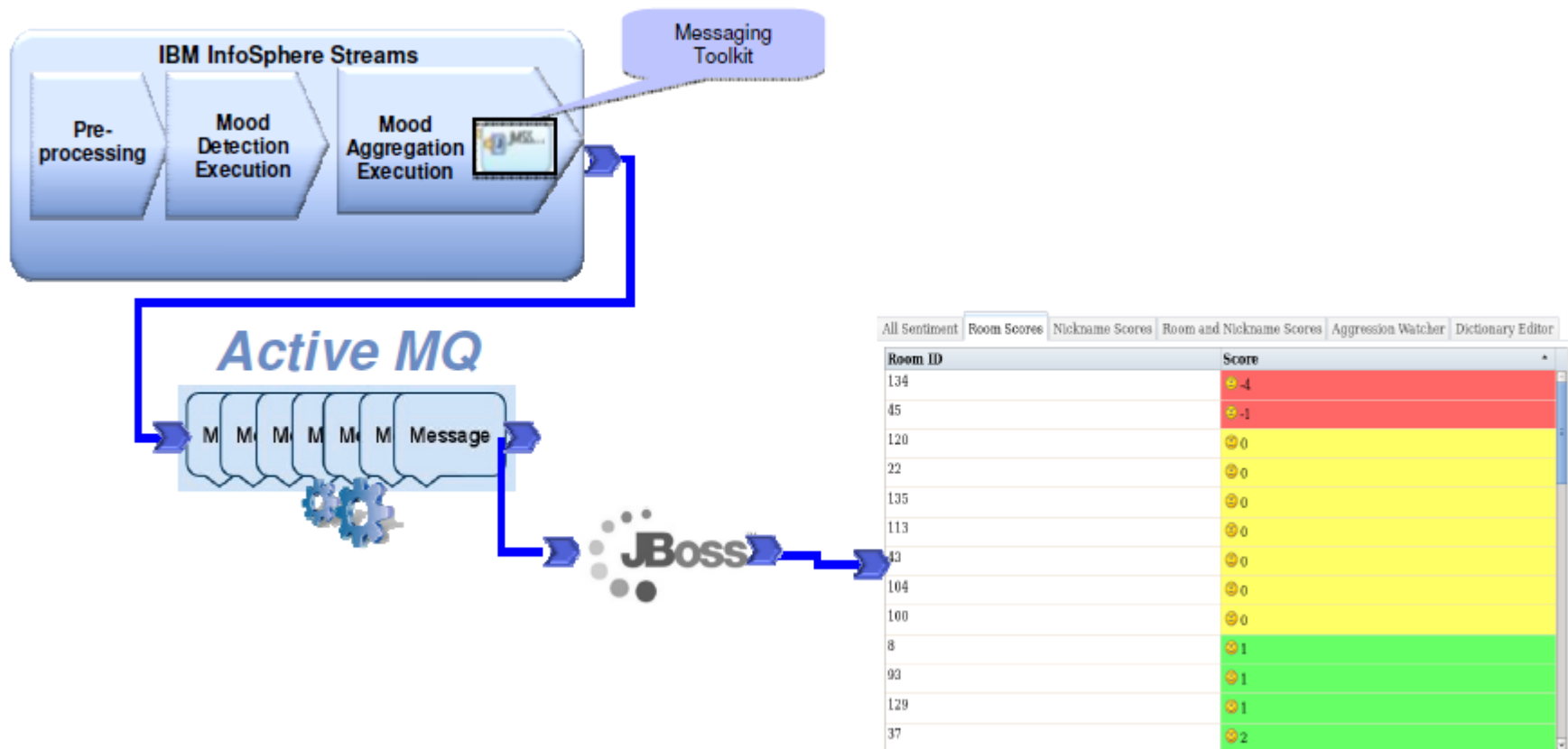
## IBM BIG DATA : CAS D'ECOLE AVEC LE BINGO ONLINE INFOSPHERE STREAMS APPLICATION



# Information On Demand 2013



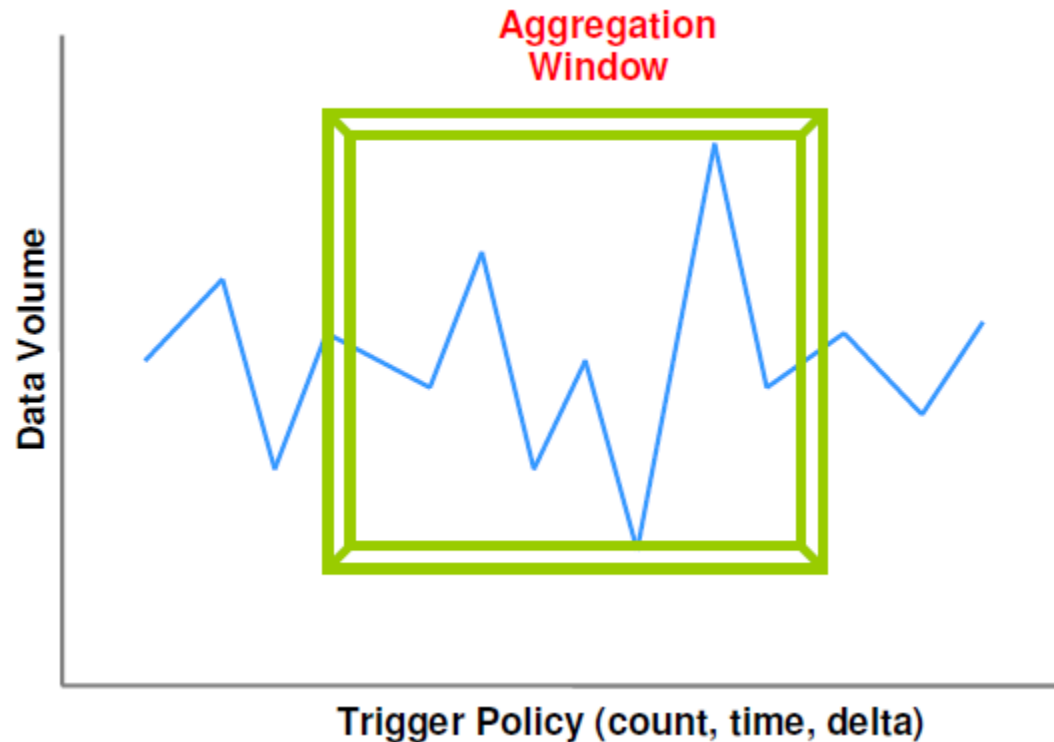
## IBM BIG DATA : CAS D'ECOLE AVEC LE BINGO ONLINE TABLEAU DE BORD SUR L'HUMEUR DES JOUEURS



# Information On Demand 2013



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE AGREGATION DES DONNES

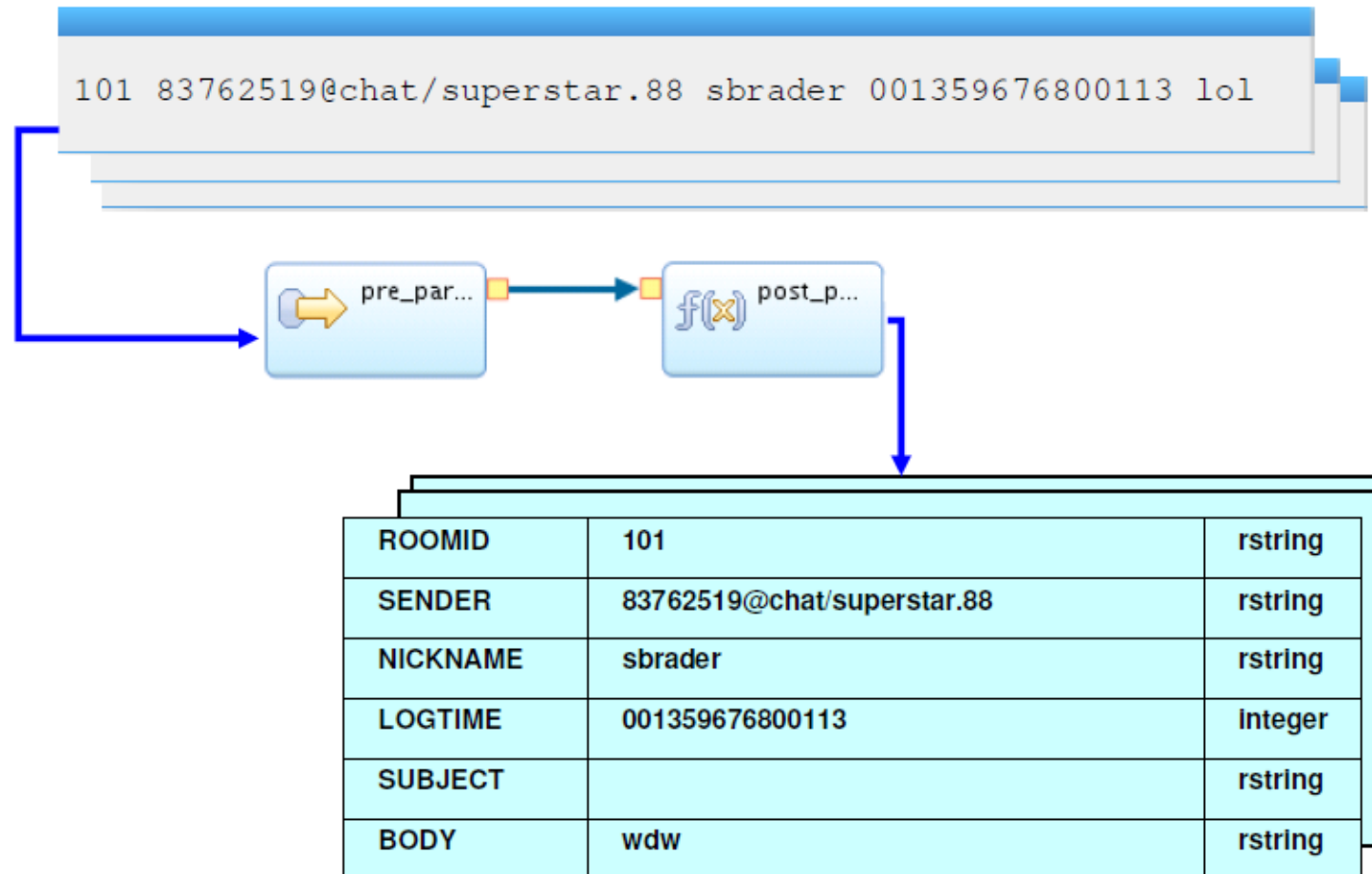




# Information On Demand 2013



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE PRE PROCESSING



# Information On Demand 2013



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE ACCELERATORS : DEVELOPPEMENT ACCELERE

### Relational Operators

Filter	Sort
Functor	Join
Punctor	Aggregate

### Adapter Operators

FileSource	UDPSource
FileSink	UDPSink
DirectoryScan	Export
TCPSource	Import
TCPSink	MetricsSink

### Utility Operators

Custom	Split
Beacon	DeDuplicate
Throttle	Union
Delay	ThreadedSplit
Barrier	DynamicFilter
Pair	Gate
JavaOp	Switch
Parse	Format
Decompress	CharacterTransform

### XML Operator

XMLParse

### IBM Supported Operators

Database	DataStage
Big Data	Data Explorer
Messaging	Internet
Text Analytics	Mining
SPSS	CEP
Time Series	Geospatial
Financial	

### Open-Source Operators

JSON	HTTP/REST
OpenCV	Accumulo
HBase	...

### Big Data Accelerators

Social Data Analytics  
Machine Data Analytics

### User-Defined Operators

Extend the language by adding user-defined operators, types, and functions

☐ Operator used in this PoC

# Information On Demand 2013



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE TEXT ANALYTICS EN ACTION

- ▶ **DICTIONNAIRES** : mots positifs/négatifs, mots indiquant des sentiments, emoticons
- ▶ **Paquets de négations et autres cas particuliers**
  - « i nerver win ! » vs « I win again »
  - « bloody game ! » vs « bloody awesome game »
- ▶ **Positionner un poids sur chaque mot**
- ▶ **Agrégation, afin de donner un score global sur le sentiment du message.**
- ▶ **Attribuer une note globale sur l'ambiance générale basée sur les dictionnaires.**



Aggressive



Bored



Concerned



Happy



Light Hearted

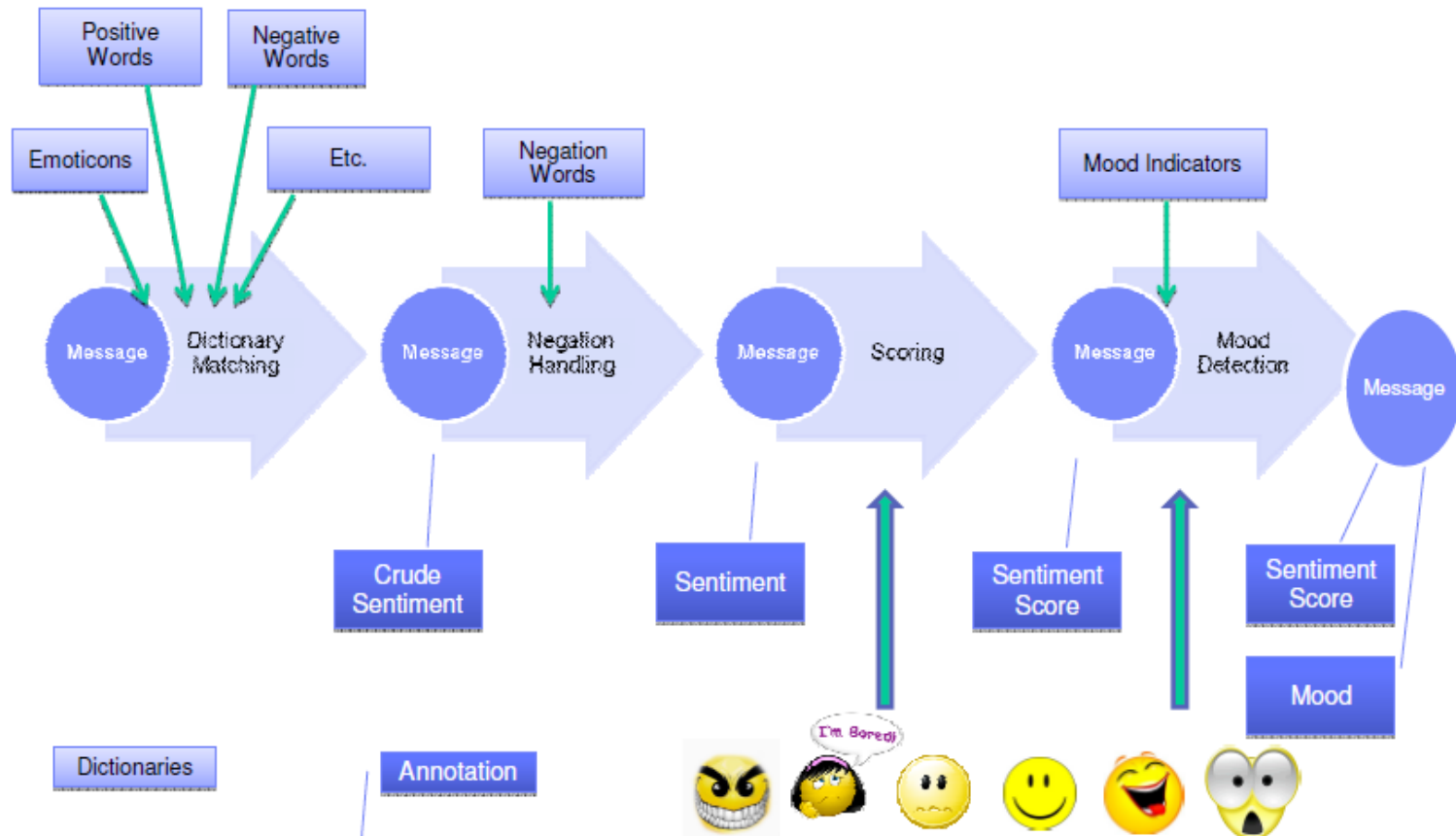


Shocked

# Information On Demand 2013



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE DATA FLOW SUR LA DETECTION DE L'HUMEUR



# Information On Demand 2013



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE EXEMPLES DE SENTIMENTS POSITIFS

Chat Message	Mood	Score
"my first win :-)"	"Happy"	5
"won small Pj bubbles :-)"	"Happy"	5
"just won 50 spins ty"	"Happy"	4
"Ive won 12 tickets wonderful :-D:-D"	"Happy"	6
"hi all won 50 spins n got 17.10 lol gl all xxx"	"Happy"	5
"thanx dora im always saying i never win but i did so gd luck to everbody with a big :GL:and a big :-D from me"	"Happy"	4



# Information On Demand 2013



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE EXEMPLE DE SENTIMENTS NEGATIFS

Chat Message	Mood	Score
"havent seen win for long time :-/"	"Negative"	-4
"never seen anyone on chat win,,always seems to be someone thats never on chat..whys that?"	"Negative"	-3
"This is rubbish tonite no win!!!:-/"	"Negative"	-4
"never buying pre game tickets never won once in years"	"Negative"	-3
"last game for me been here all night not won really need a win plzseeeeeeee"	"Negative"	-2
"cant win at all dont know the last time i won"	"Negative"	-3
"i bet no one mwon that game"	"Negative"	-1
"Host there is a problem in lounge, Game did not play and no-one won"	"Negative"	-3

# Information On Demand 2013



## IBM BIG DATA. EXEMPLE D'APPLICATION : LE BINGO ONLINE EXEMPLE D'AGRESSION

Chat Message	Mood	Score
"ffs i never win on this site what c r a p"	Aggression	-6
"ffs do one... sick of gamseys sites taking all time. only one winner. sorry to offend if i did. but not everyone can bet big and win every spin. us poorpers got no chance."	Aggression	-9
"that chit cost me 90p"	Aggression	-4
"you can still report them to council"	Aggression	-4
"this site is so fixed,im going to trading standards"	Aggression	-4
"contact the Gaming Commission Lyn, they are the regulators of bingo sites"	Aggression	-4
"slots are a rip off a 10r dissapears in no time"	Aggression	-5
"ooo they havent changed the cashback scam then :("	Aggression	-6
"wat a xxxxing shit site see ya"	Aggression	-8

# Information On Demand 2013

**BIG DATA : GENERALITES**

