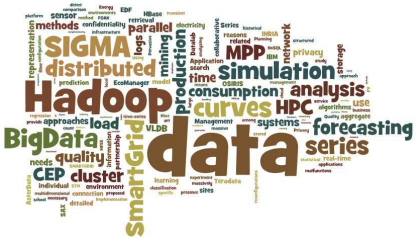
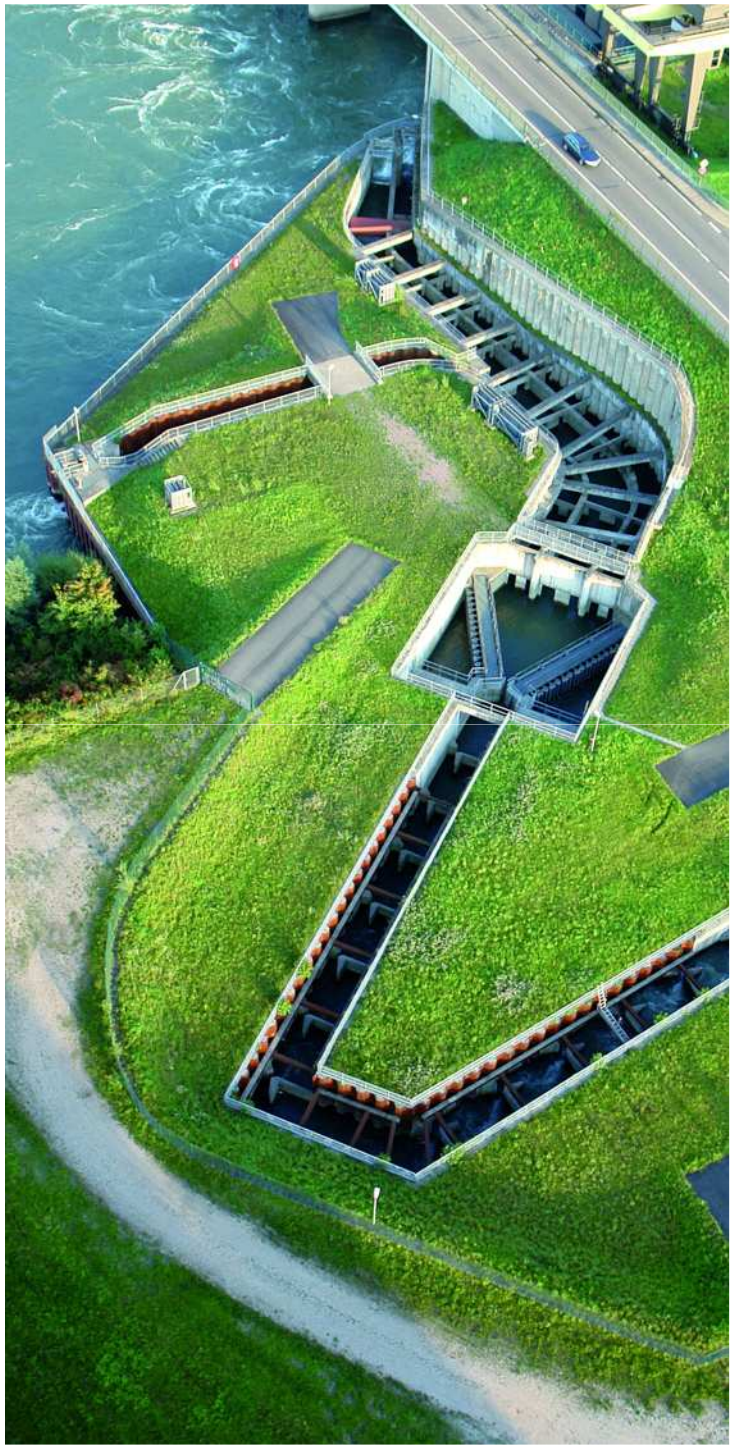
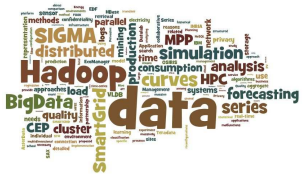
A SMART ELEPHANT FOR A SMART-GRID: (ELECTRICAL) TIME-SERIES STORAGE AND ANALYTICS WITHIN HADOOP

EDF R&D – SIGMA Project
Marie-Luce Picard

Forum TERATEC
June 26th 2013

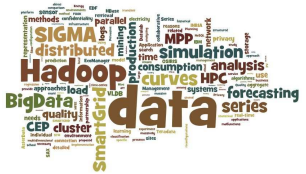






OUTLINE

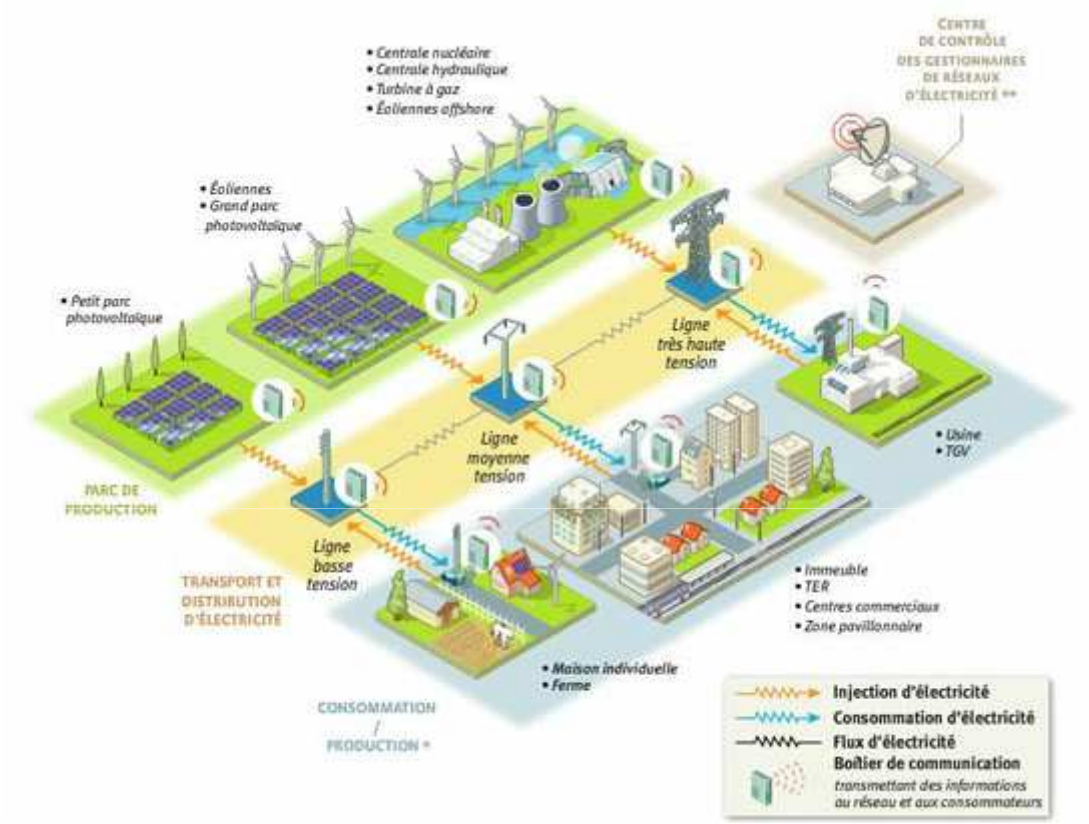
1. CONTEXT
2. A PROOF OF CONCEPT USING HADOOP
3. ADVANCED ANALYTICS
4. CONCLUSION AND PERSPECTIVES



OUTLINE

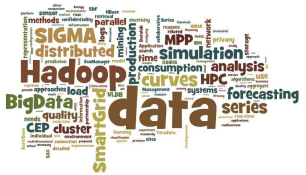
1. **CONTEXT**
2. **A PROOF OF CONCEPT USING HADOOP**
3. **ADVANCED ANALYTICS**
4. **CONCLUSION AND PERSPECTIVES**

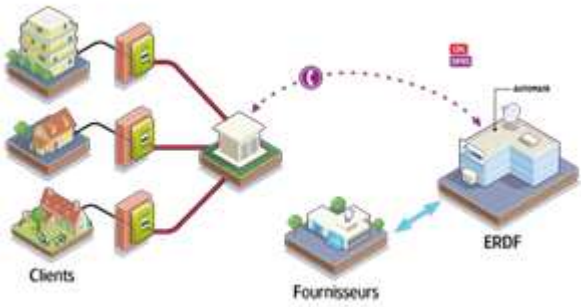
SMART GRIDS ⇨SMART METERS ⇨ SMART DATA

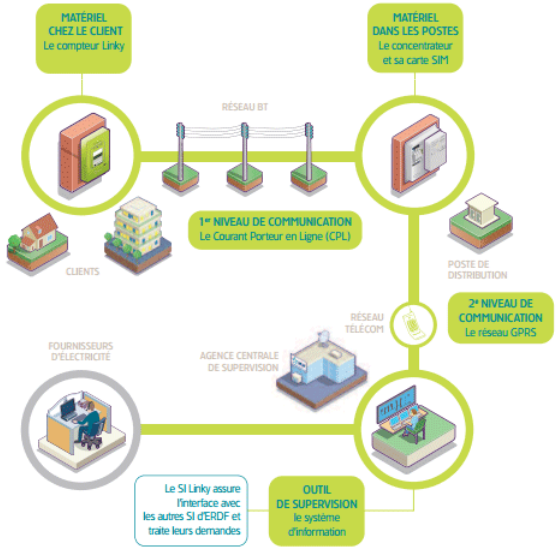


Motivated by economic or regulatory constraints, or environmental needs, smart-grid projects are spreading. New perspectives emerge for energy management as electric vehicles, and renewable energy generation will expand. A large amount of smart-meters and sensors will be deployed and they will provoke a data deluge utilities will have to face.

SMART METERING: A DATA DELUGE!





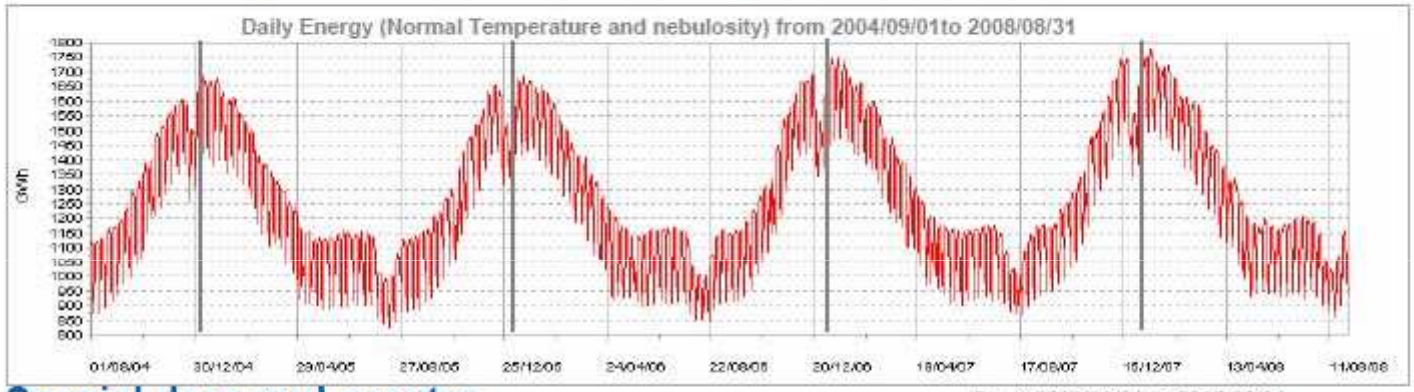


French project : 35+ millions of Smart Meters, ⇨ billions of metering data

Currently, pilot program with 300k smart meters deployed

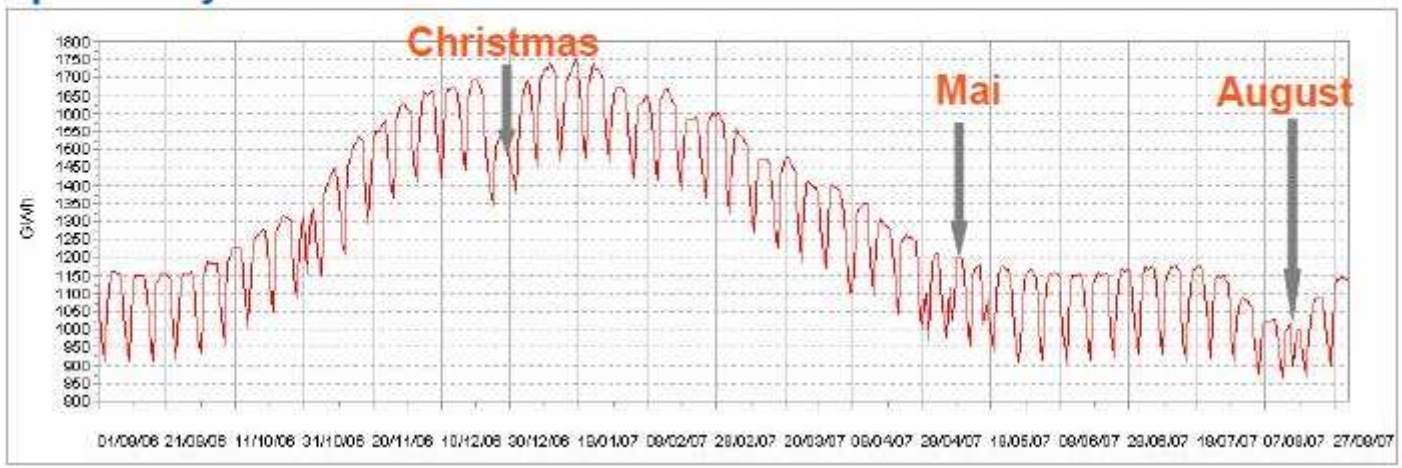
SMART METERING DATA: WHAT DOES A LOAD CURVE LOOK LIKE ?

Some characteristics of French electricity consumption : annual cycle



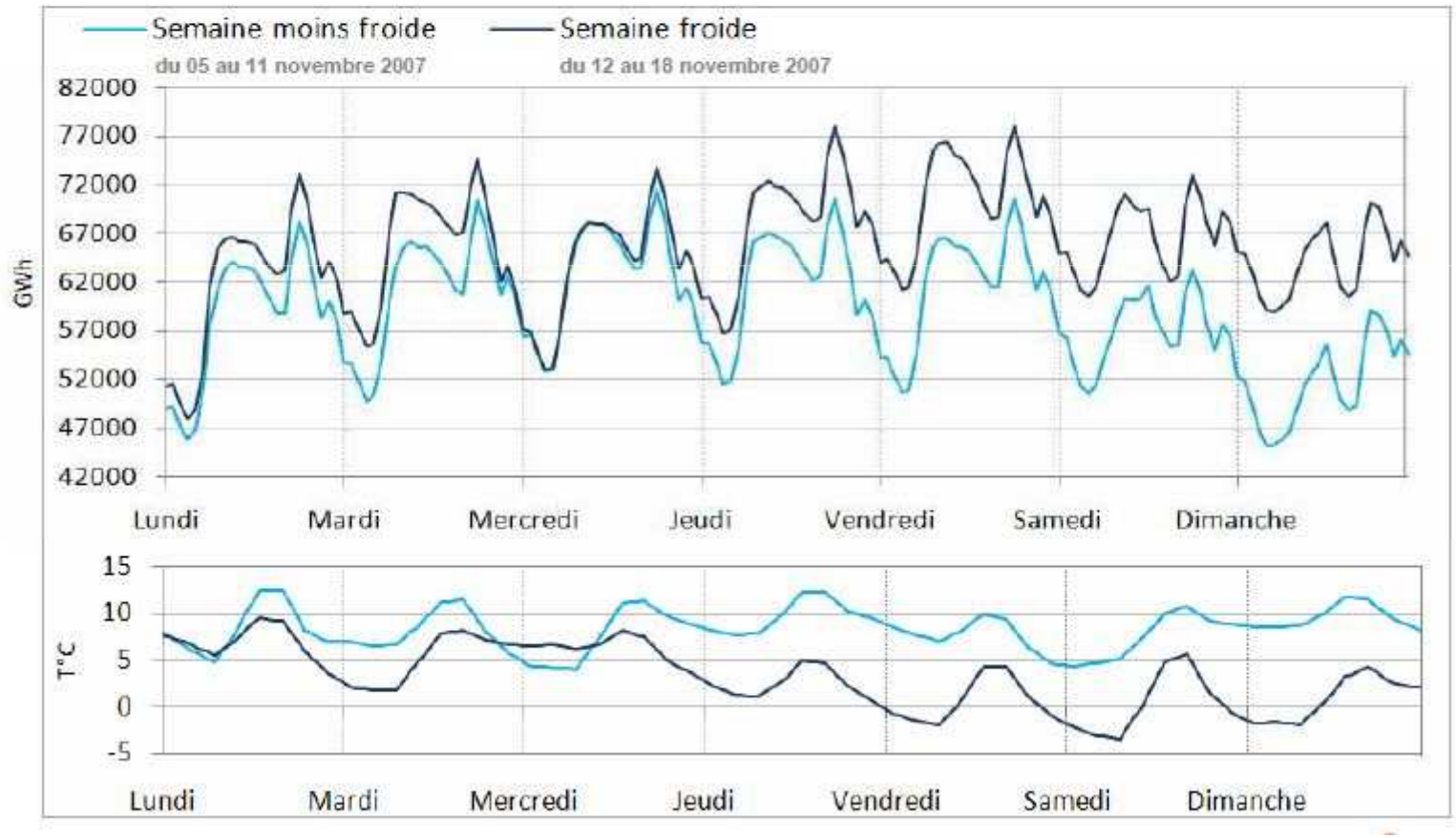
Special days and events:

From 2006/09/01 to 2007/08/31

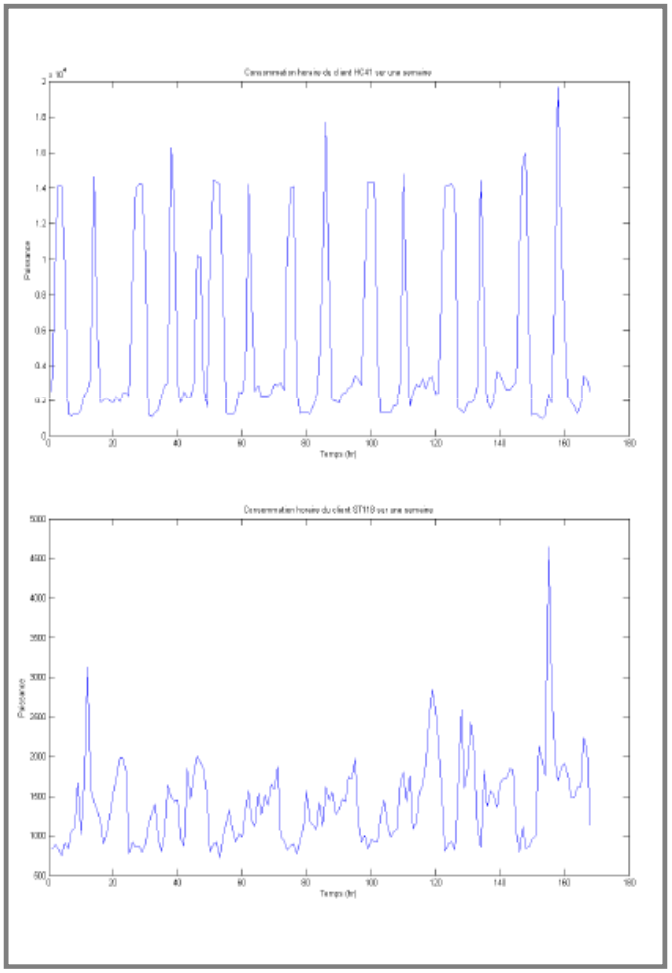


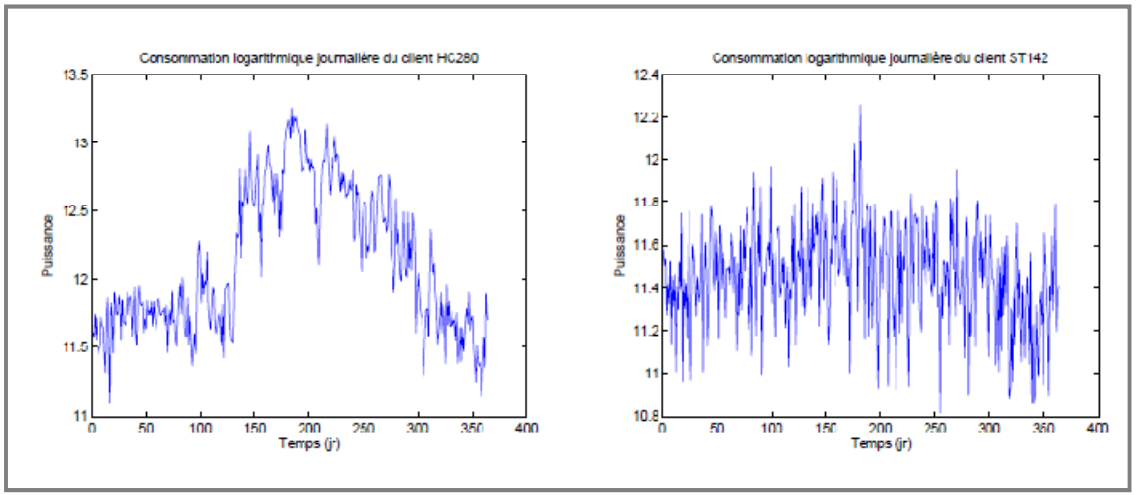
SMART METERING DATA: WHAT DOES A LOAD CURVE LOOK LIKE ?

Some characteristics of French electricity consumption : temperature's effect in winter



SMART METERING DATA: WHAT DOES A LOAD CURVE LOOK LIKE ?



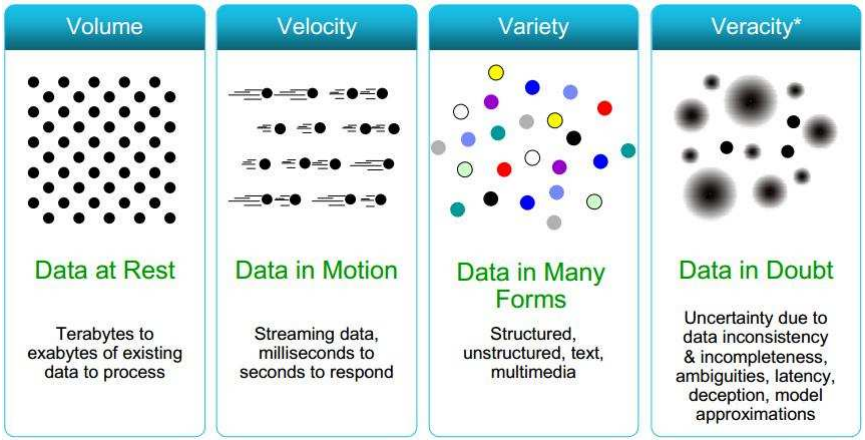


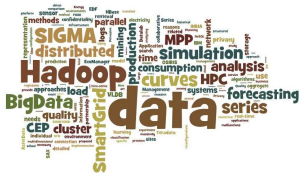
Individual load curves :
-Left : same customer, two different days
-Up: same day, two different customers

MASSIVE DATA MANAGEMENT IN THE ENERGY DOMAIN

- **Challenges:**
  - More complexity in the electric power system (distributed generation, demand response …)
  - Faster complexity of customer indoor equipment (smart meters and devices, Internet Of Things …)

  ➢ **Core business will involve more IT and data management**

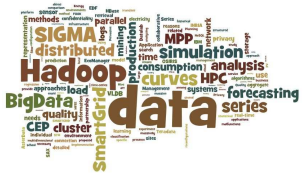




OUTLINE

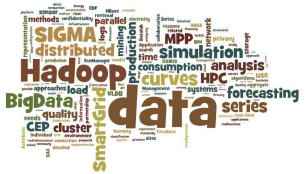
1. CONTEXT
2. A PROOF OF CONCEPT USING HADOOP
3. ADVANCED ANALYTICS
4. CONCLUSION AND PERSPECTIVES

MASSIVE TIME-SERIES STORAGE

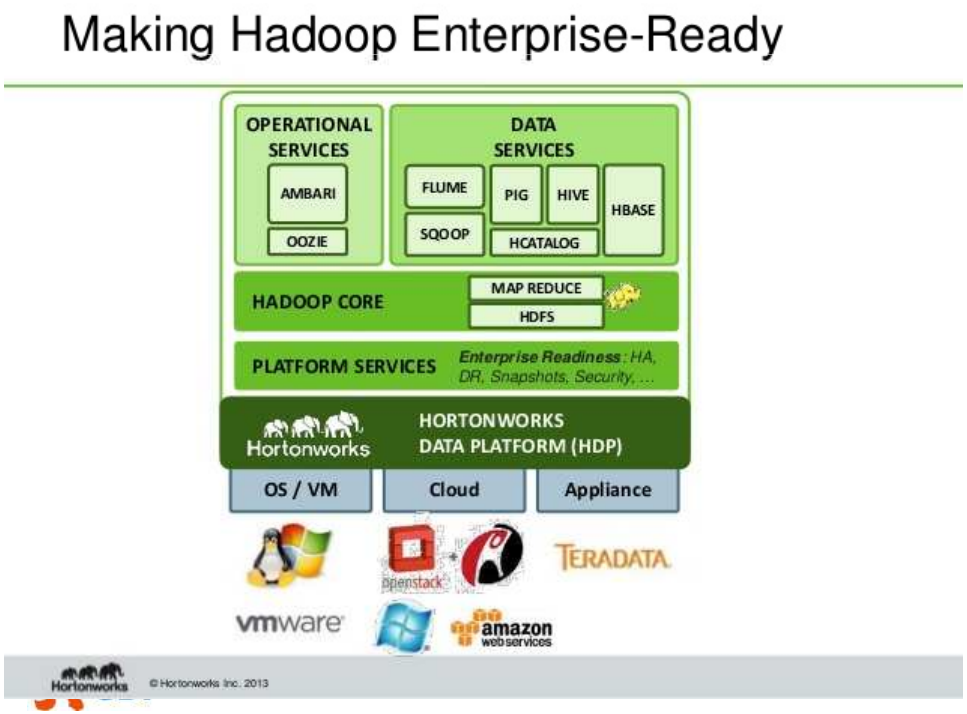


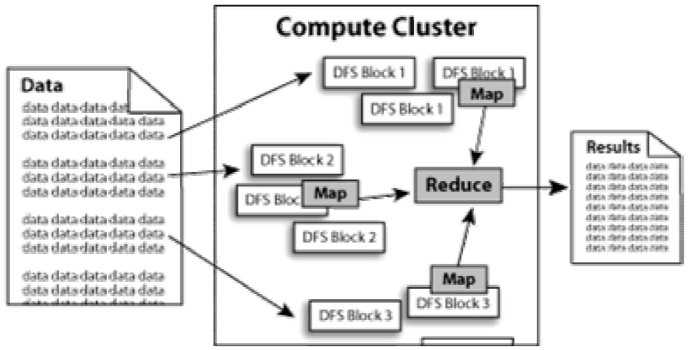
- **Objective: Proof Of Concept for running a large number of queries (variable levels of complexity with variable scopes and frequencies, variable acceptable latencies) on a huge number of load curves**
  - Data: individual load curves, weather data, contractual information, network data
    - 1 measure every 10 mn for 35 million customers
    - Annual data volume: **1800 billion records, 120 TB of raw data**
  - "Operational DataWareHouse" able to:
    - Supply a large volume of data
    - Ingest new incoming data (pre-processing)
    - Allow concurrent and simultaneous queries: tactical queries, analytical queries, ad-hoc queries
  - **Evaluation criteria**: competition and performances (QoS, SLA), convergence and agility
- **Traditional VLDB as well as Hadoop as a solution**

USING HADOOP FOR STORING MASSIVE TIME-SERIES



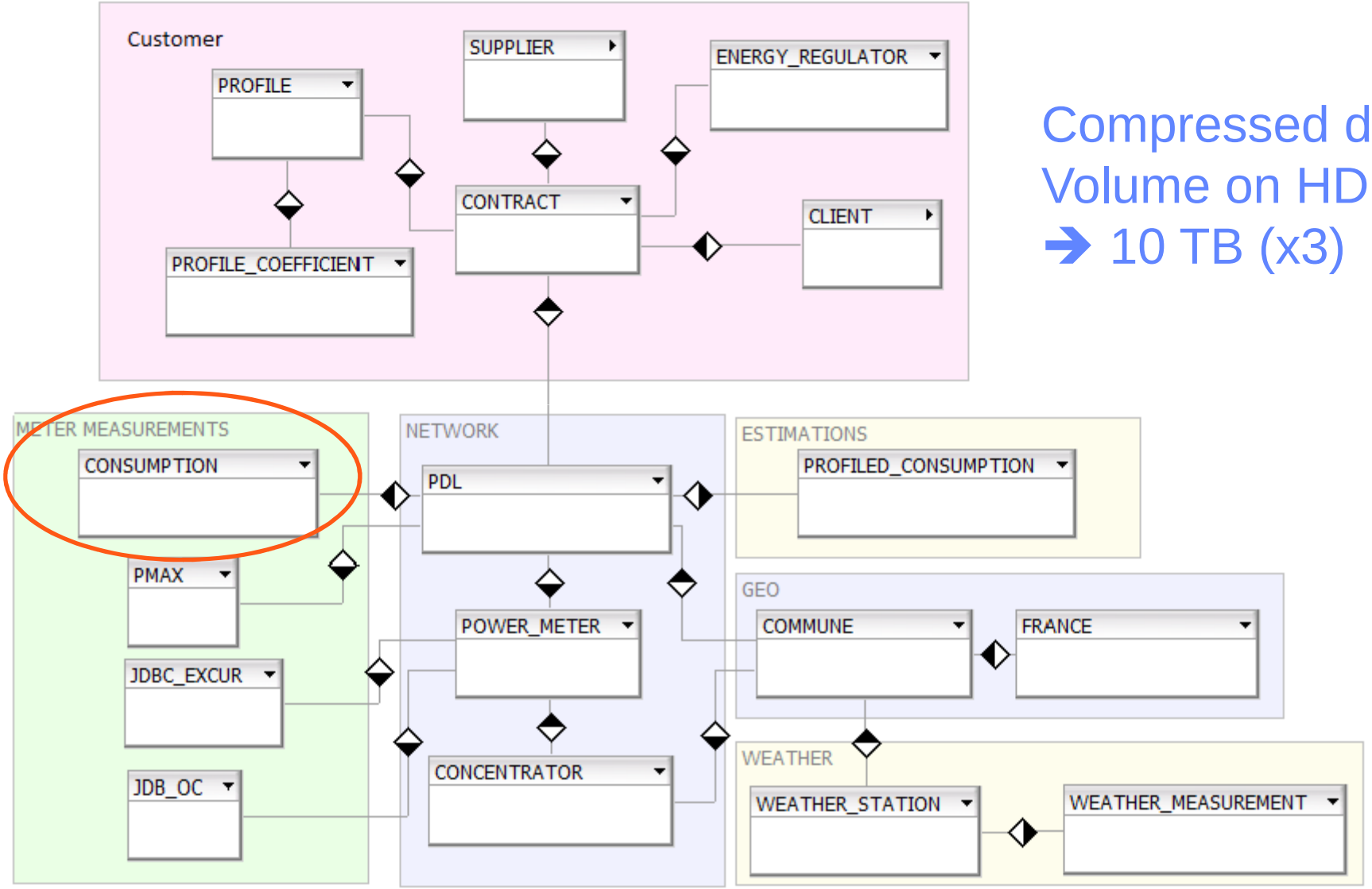
- **HADOOP:**
  - Native distributed file system (HDFS)
  - Distributed processing using the MapReduce paradigm
  - A large and fast evolving ecosystem
  - Use cases: mainly on unstructured data





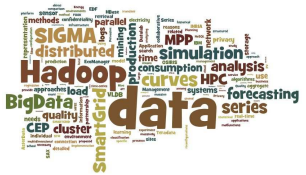
Divide and Conquer

POC HADOOP: THE DATA MODEL

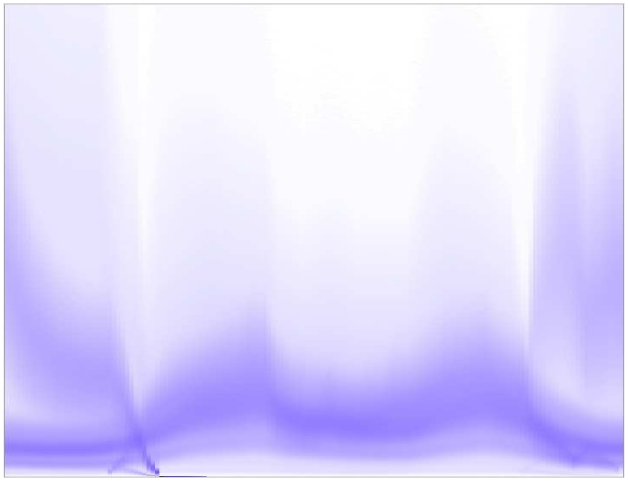


Compressed data Volume on HDFS :
➔ 10 TB (x3)

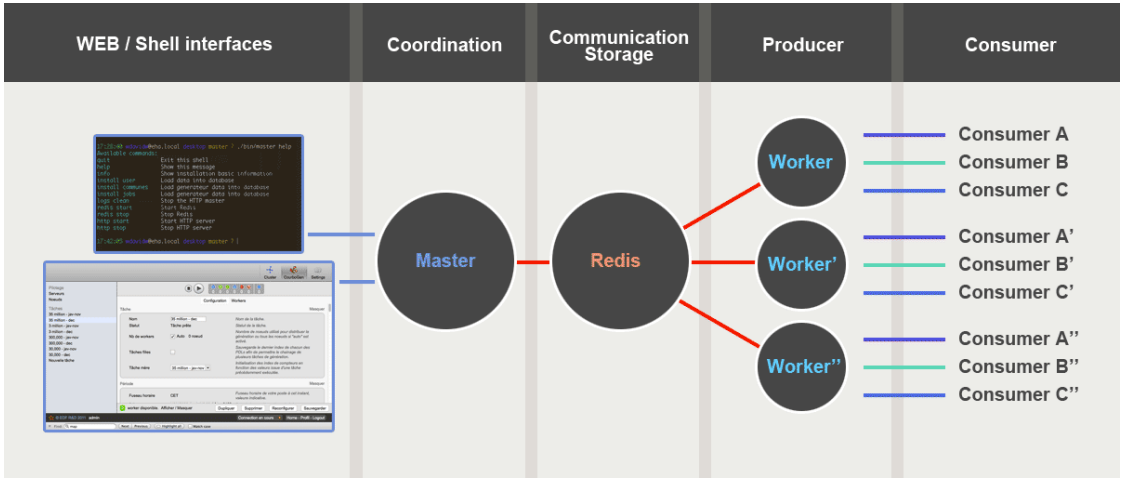
DATA GENERATOR – COURBOGEN ©



- **Courbogen can generate massive datasets (not too statistical realistic)**
  - Generates load curves and associated data
  - Customizable tool: interval, duration, data quality, noise on the curves
  - Distributed architecture (NodeJS, Redis)
  - Output as a data stream



One week data for 35M curves

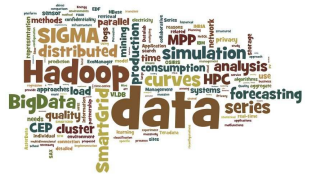


DESIGN

R&D 20 nodes cluster,
336 cores

- **All data stored in HDFS**
  - Use of HIVE for analytical queries (11 months of data)
  - Use of HBASE at the forefront of data access (tactical queries) – (3 months of data)

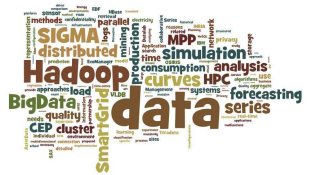
POC HADOOP: DATA UPLOAD



- **Daily data**
  - Raw data: 327 Go ; HDFS data: 50 Go, Hive/HBase : 25-28 Go
  - Initial upload : 15 minutes

Hadoop component	Representation models	Total volume of data (compressed)	Time for uploading and preparing the data
HIVE	Tuple mode	10 * 3 = 30 To	2h 30, 35' for uploading
HBASE	Array (daily data)	3 * 3 = 9 To	8 minutes (only upload, aggregated data is computed in Hive).
	TOTAL	**39 To**	**2H 38**

POC HADOOP: RESULTS

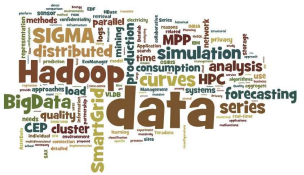


- **Overall GOOD results; for some types of queries, competitive with traditional approaches**
  - Tactical queries are successful handled by HBase, offering low latencies under a high concurrent load
  - Analytical queries are processed by Hive

- **Analytical queries on HIVE**
  - 11 months of data, 35 millions of curves, 10 minutes interval
  - Partitioning: day/tariff/power
  - Tuple representation

Query	Execution time (1 day)	Execution time (7 days)
Global France aggregated curve – 10mn	1 min, 56 sec	19 min, 24 sec
Aggregated curve by tariff and power	2 min, 21 sec	24 min, 33 sec
Average consomption by building type (ad-hoc)	1 heure, 18 sec	3 heures, 16 min, 50 sec

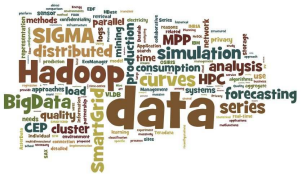
POC HADOOP: RESULTS



- **Tactical queries using HBase**
  - 3 months of data, 35 millions of curves, 10 minutes interval
  - 500 concurrent queries, SLAs asked to be respected for 90% of queries
  - Array representation

Durée	Pas de temps	NB requêtes simultanées demandées / obtenu		SLA demandé / obtenu par requête en sec + taux de réussite obtenu			Ecart type en sec	Nombre de requêtes effectuées / sec
Jour	10min	200	200	< 1	0,57	86,53	1,77	210,98
	30min	50	50	< 3	0,68	93,54	1,98	73,83
Semaine	10min	160	160	< 2	2,11	88,36	9,16	45,76
	horaire	40	40	< 7	2,89	92,24	11,43	13,83
Mois	horaire	32	32	< 7	2,45	94,09	11,06	13,09
	jour	8	8	< 10	1,68	94,67	7,35	4,79
Trimestre	jour	8	8	< 10	2,13	94,33	9,12	3,77
	hebdo	2	2	< 15	2,13	94,92	8,98	0,95

POC HADOOP: TIME SERIES REPRESENTATION MODELS



TUPLE

ID_CDC	DATE_RELEVE	P (Power)	DAY	OPTARIF	PSOUSC
136630	-128 (00.00 am)	453	2008-01-01	RES1	6
136630	-127 (00.10 am)	307	2008-01-01	RES1	6
...	...	...			
136630	15 (23.50 pm)	433		RES1	6

ARRAY

ID_CDC	Values	DAY	OPTARIF	PSOUSC
136630	[[-128,453], [-127,307],...,[15,433]]	2008-01-01	RES1	6

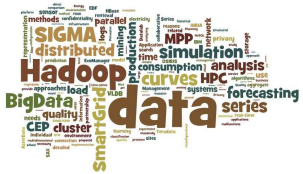
COLUMN

ID_CDC	P1 (-128)	P2 (-127)	...	P143 (-15)	DAY	OPTARIF	PSOUSC
136630	453	307	...	433	2008-01-01	RES1	6

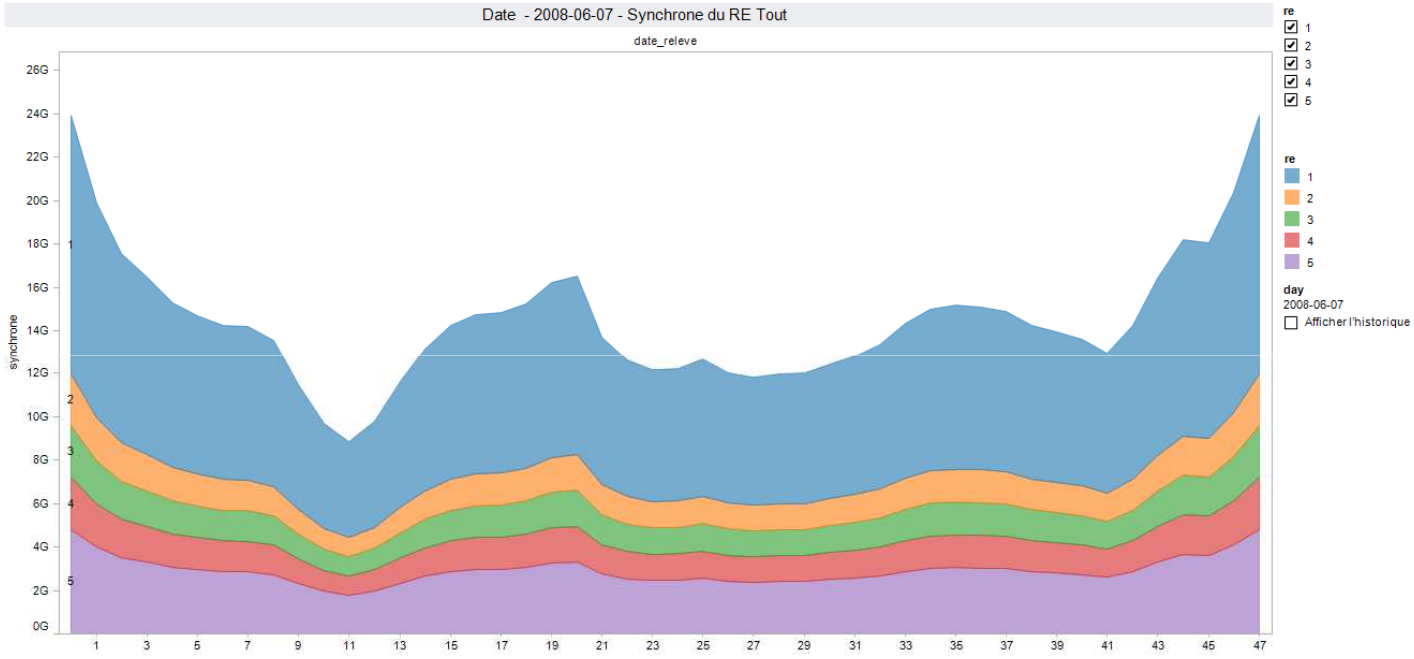
- Computing a global aggregated load curve for 1 day

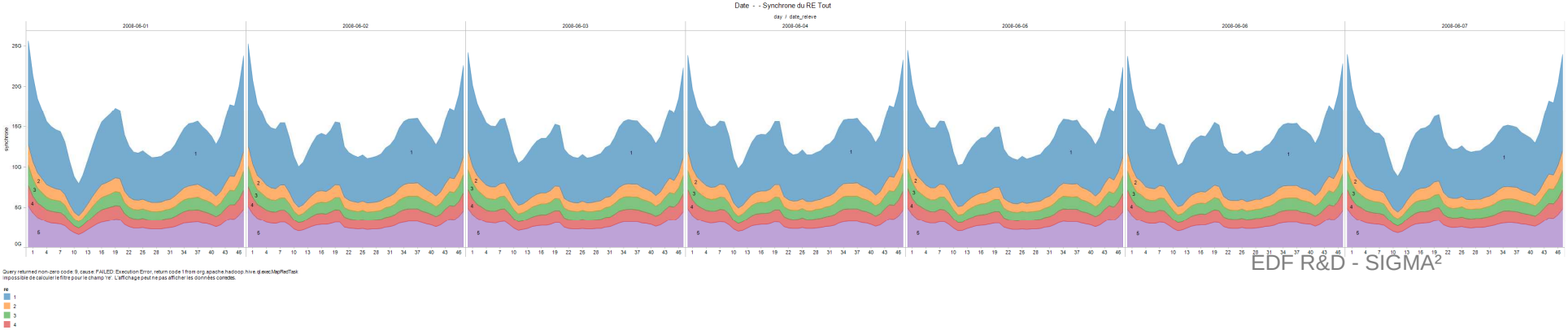
Representation model	Daily volume	Query execution time
Tuple	10.1 GB (x 3 replicas)	2 min 22 sec
Column	8.8 GB (x 3 replicas)	1 min 17 sec
Array	16 GB (x 3 replicas)	1 min 18 sec

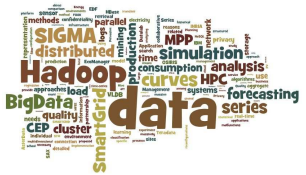
POC HADOOP: VISUALIZATION



Hadoop / Tableau Software : Synchrones par RE



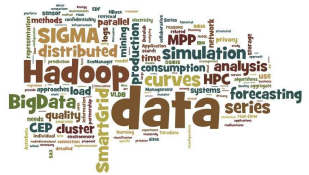




OUTLINE

1. CONTEXT
2. A PROOF OF CONCEPT USING HADOOP
3. ADVANCED ANALYTICS
4. CONCLUSION AND PERSPECTIVES

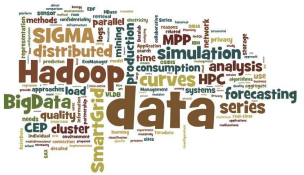
ADVANCED ANALYTICS WITH HADOOP



Strong Needs for advanced analytics on smart metering data: Segmentation based on load curves, Forecasting on local areas, Scoring for Non Technical Losses, pattern recognition within load curves, predictive modeling, …

- **Implementing classical data-mining tasks vs home-made methods**
- **Use of existing librairies and toolkits vs developing Hadoop On Time-series toolkit**

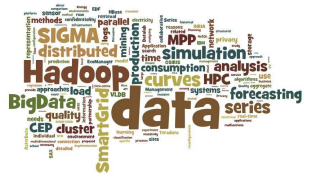
EXISTING ANALYTICS TOOLKIT



- **Rhadoop**
  - 3 R packages for analyzing data stored within Hadoop from R (rmr2,rhdfs,rhbase)
  - You need to re-implement methods using the rmr2 package
  - Small users community
  - Linked to Revolution

- mahout
  - A Java toolkit with data analysis methods implemented using MapReduce paradigm
  - Apache foundation (http://mahout.apache.org/)
  - Large users community
  - Evolving (and growing) very fast

POC HADOOP: DATA CLUSTERING



- Examples of tests run on a data sets of 35 millions of curves, one measurement by curve (daily average consumption)
- Script Rhadoop (execution time: 3.12 heures)

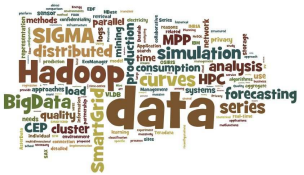
```
> tb_input_kmeans = mapreduce('/tmp/tb_kmeans_r.csv',
    input.format = make.input.format('csv', sep=','),
    structured = T,
    vectorized = T,
    map = function(k, v) keyval( v$V1 %% 35000 , v, vectorized = T),
    reduce = function(k,vv) keyval(k , vv, vectorized = F),
    backend.parameters =
    list(hadoop = list(D ="mapred.reduce.tasks=200",D="mapred.map.tasks=200")),
    verbose=T
)
> kmeans(tb_input_kmeans, ncenters = 20, iterations = 15, fast = T)
```

- Script Mahout (execution time: 17 minutes)

```
$ mahout org.apache.mahout.clustering.conversion.InputDriver --input
/user/hive/warehouse/sigma.db/tb_input_kmeans --output /user/sigma/outputvector/


$ mahout kmeans --input /user/sigma/outputvector/ -c clusters -k 20 --output
/user/sigma/output -dm org.apache.mahout.common.distance.EuclideanDistanceMeasure
--maxIter 15 --overwrite --clustering
```

FROST : A LIBRARY FOR TIME SERIES TRANSFORMATION WITHIN HIVE



- Analytics and Data Mining on time series : curse of dimensionnality

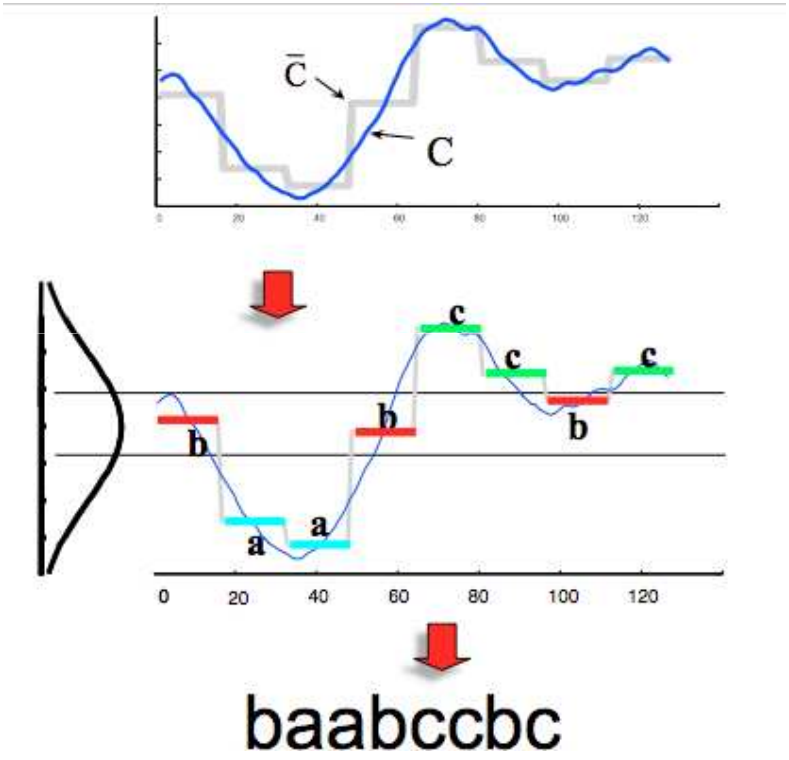
- FROST is a set of user defined functions in HIVE enabling simplification of time series

FROST.JAR (First Release Of time Series Toolkit)

Example :

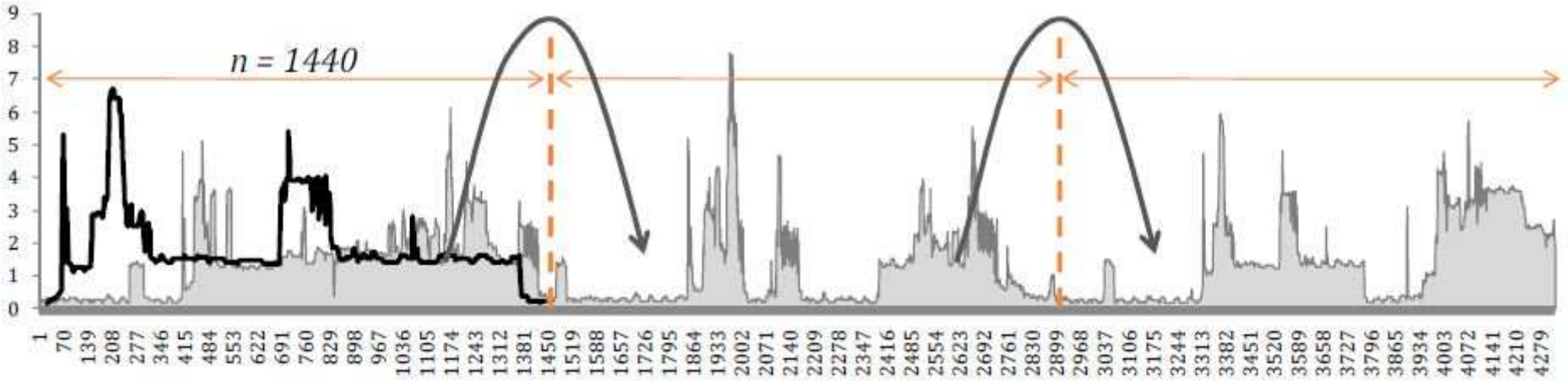
```
ADD JAR FROST.JAR;
…
SELECT ID,SAX(POWER,8,3)
 FROM BIG.DATA
 GROUP BY DAY;
```

- Other functions in FROST :
  - PAA : Piecewise Aggregate Approximation
  - DFT : Discret Fourier Transform
  - DWT : Discret Wavelet Transform
  - and other « home made » methods

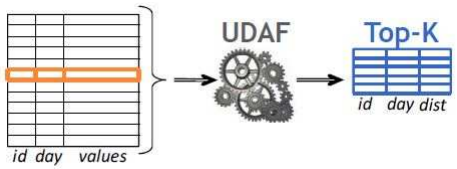


SAX principle
(Symbolic Aggregate approXimation)

SEARCHING TIME SERIES WITH HADOOP

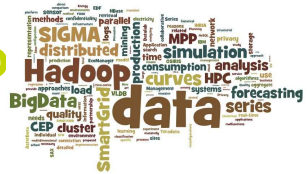


- **Objective: searching similar time-series within a huge sets of series**
  - Top-K or range queries based on a similarity measure
  - Jumping or sliding windows
  - Brutal force of distributed computations in the Hadoop environment: use of UDF Hive functions




```
CREATE TEMPORARY FUNCTION knn_saut
    as 'com.alice.UDAF.KnnSaut';
SELECT knn_saut(id_client, values,
    day)
FROM table;
```

SEARCHING TIME SERIES WITH HADOOP



Distance euclidienne, sur des données brutes normalisées, avec pour requêtes des séries n'existant pas dans la base

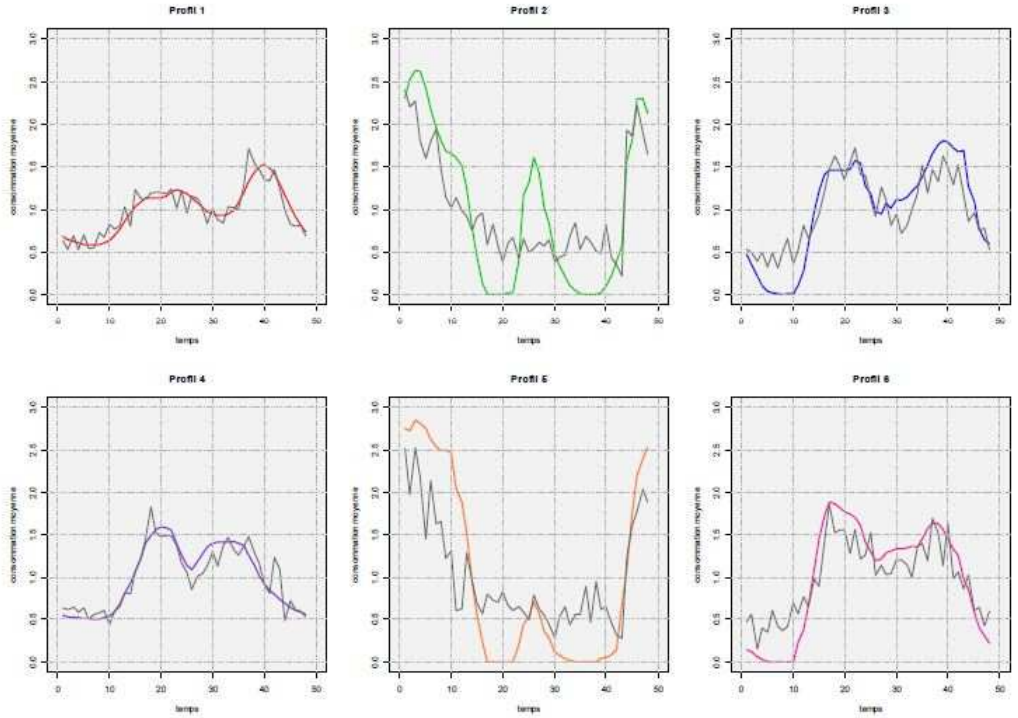
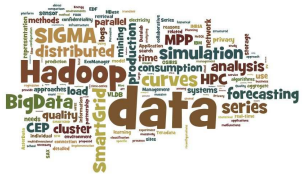


FIGURE 6.2 – Résultats de six Top-1, en distance euclidienne, avec pour requêtes les six profils-type à partir desquels sont dérivés les séries de la table `conso_releve_array`

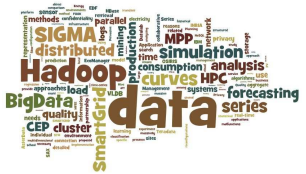
- **Data: 35 millions curves, 30 days**
  - Top-5 (or top-500), with UDAF functions on jumping windows
  - ~4 mn 45s



OUTLINE

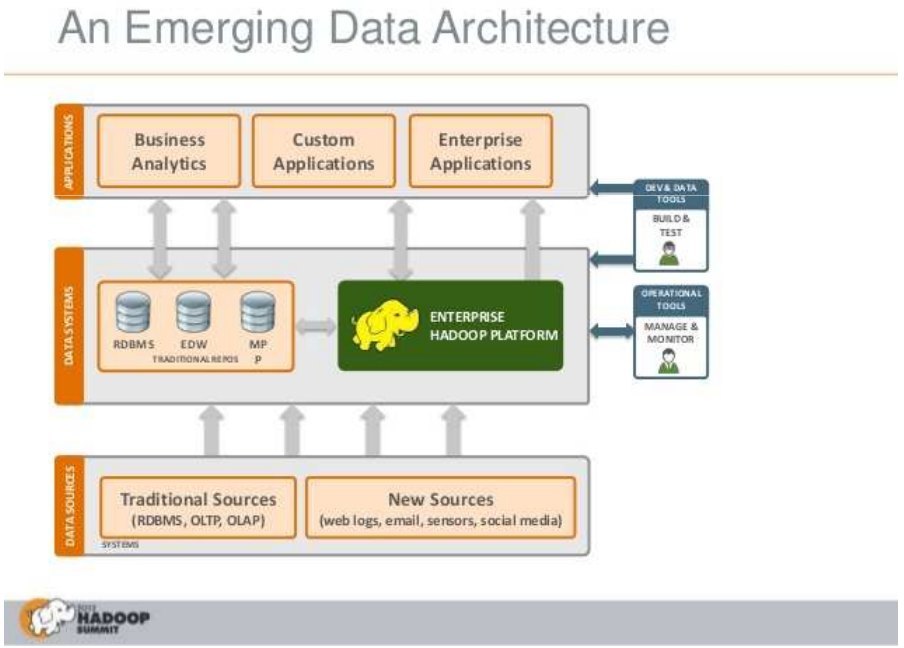
1. CONTEXT
2. A PROOF OF CONCEPT USING HADOOP
3. ADVANCED ANALYTICS
4. CONCLUSION AND PERSPECTIVES

POC HADOOP : CONCLUSIONS

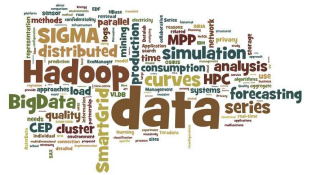


- Hadoop as an alternative approach for storing and analyzing massive time-series data (good results, some of them competitive for querying, ability to implement advanced analytics)

- Is Hadoop enterprise-ready ?
  - (+) : costs, commodity hardware, flexible, scalability, structured and non structured data, interoperability
  - (-) : skills (high complexity), maturity (security, operating, SQL-compliant)
  - Recommendation: should be limited to specific (non critical) usages such as Big ETL, unstructured data, BI sub-set, Big sand-box and analytical warehouse

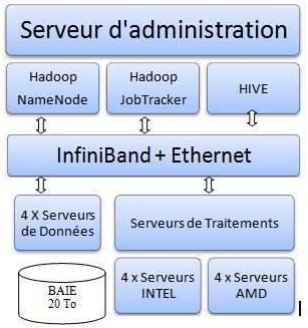


POC HADOOP : PERSPECTIVES

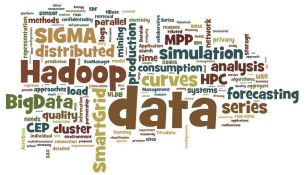


- **Hadoop for unstructured data**
  - Text-mining applications
  - Using Mahout library (more efficient, updatable) to implement text-mining functions (Lucent)
  - Needs tuning (format, stop words, …) and integration (text + clustering, results visualization)
  - Currently tested on tweets

- **Hadoop and HPC**
  - Impact on the location of data (Hadoop recommendations: data nodes should be close to the processing nodes)
    - Installation of Hadoop ecosystem on an HPC cluster
    - Impact on performances (r/w)
  - NetCDF files (coming from simulation codes) ingested by Hadoop:
    - SciHadoop: Hadoop plug-in for NetCdf data sets
    - Work in progress



REFERENCES



A proof of concept with Hadoop: storage and analytics of electrical time-series.
Marie-Luce Picard, Bruno Jacquin, *Hadoop Summit 2012*, Californie, USA, 2012.
présentation : http://www.slideshare.net/Hadoop_Summit/proof-of-concent-with-hadoop
vidéo: http://www.youtube.com/watch?v=mjzblMBvt3Q&feature=plcp

Massive Smart Meter Data Storage and Processing on top of Hadoop.
Leeley D. P. dos Santos, Alzennyr G. da Silva, Bruno Jacquin, Marie-Luce Picard, David Worms,Charles Bernard. *Workshop Big Data 2012*, Conférence VLDB (Very Large Data Bases), Istanbul, Turquie, 2012.
http://www.cse.buffalo.edu/faculty/tkosar/bigdata2012/program.php

Smart Metering x Hadoop x Frost: A Smart Elephant Enabling Massive Time Series Analysis.
Benoît Grossin, Marie-Luce Picard, *Hadoop Summit Europe 2013*, Amsterdam, Mars 2013
http://hadoopsummit.org/amsterdam/

- **Acknowledgments - Work achieved with:** Alice Bérard, Charles Bernard, Leeley Daio-Pires-Dos-Santos, Alzennyr Gomes Da Silva, Benoît Grossin, Georges Hébrail, Bruno Jacquin, Jiannan Liu, Vincent Nicolas, David Worms