GRIDPOCKET

PERSONAL SMARTGRID SOLUTIONS

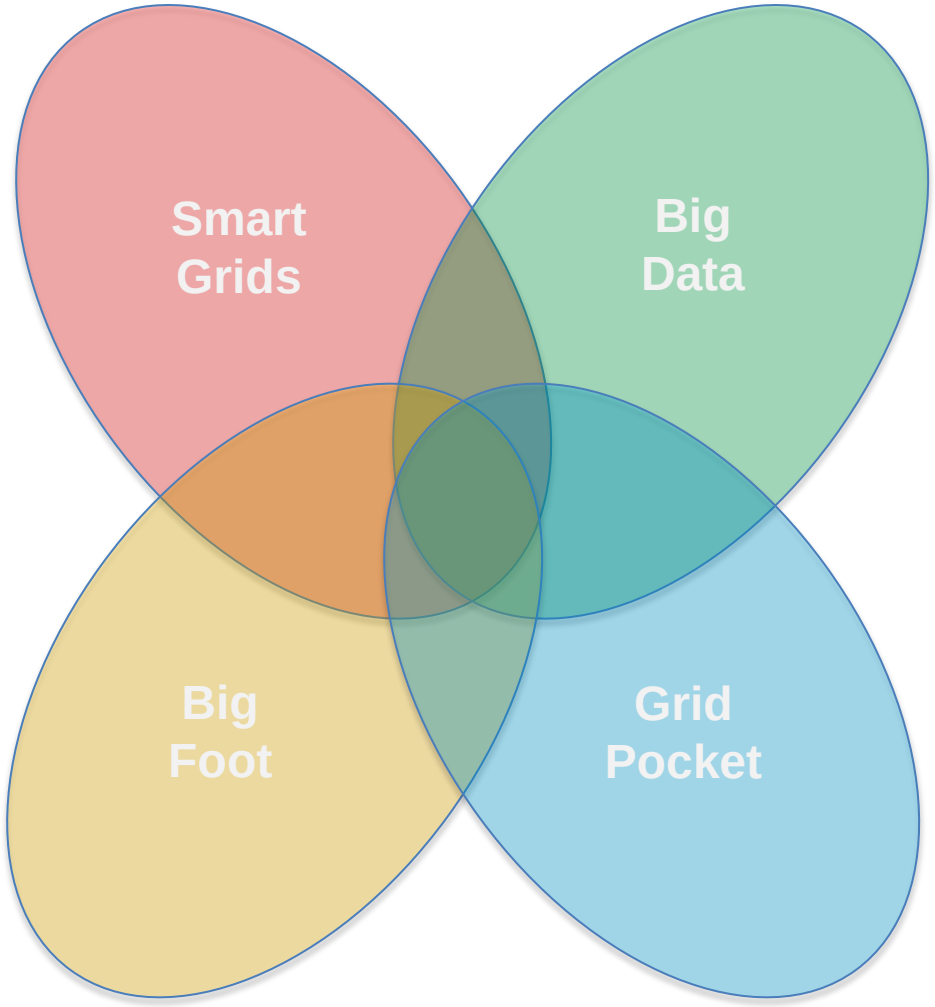
Le BigData avance à grands pas

Un exemple dans les Smart Grids

03/07/2013

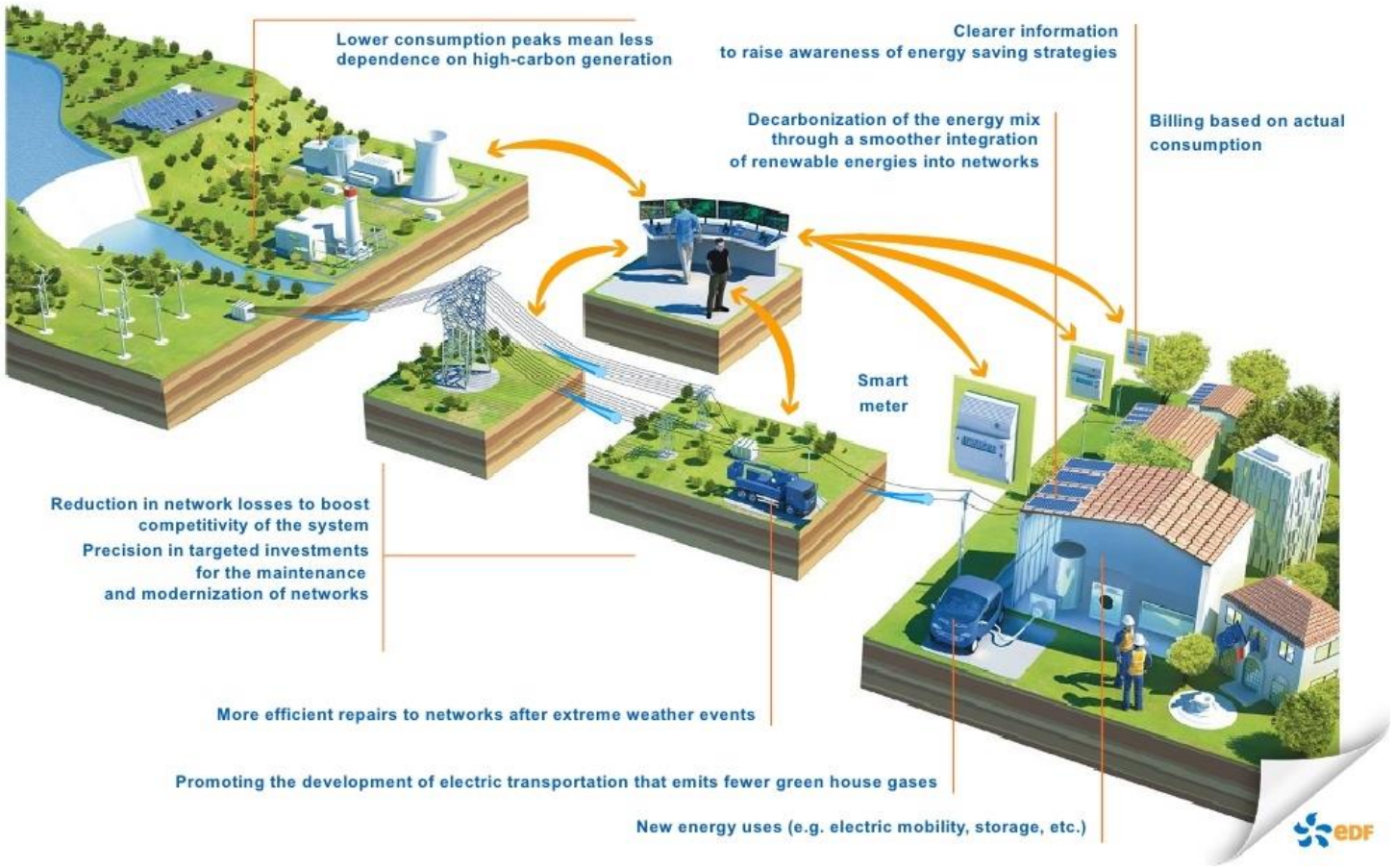
Michael Defoin-Platel

Outline



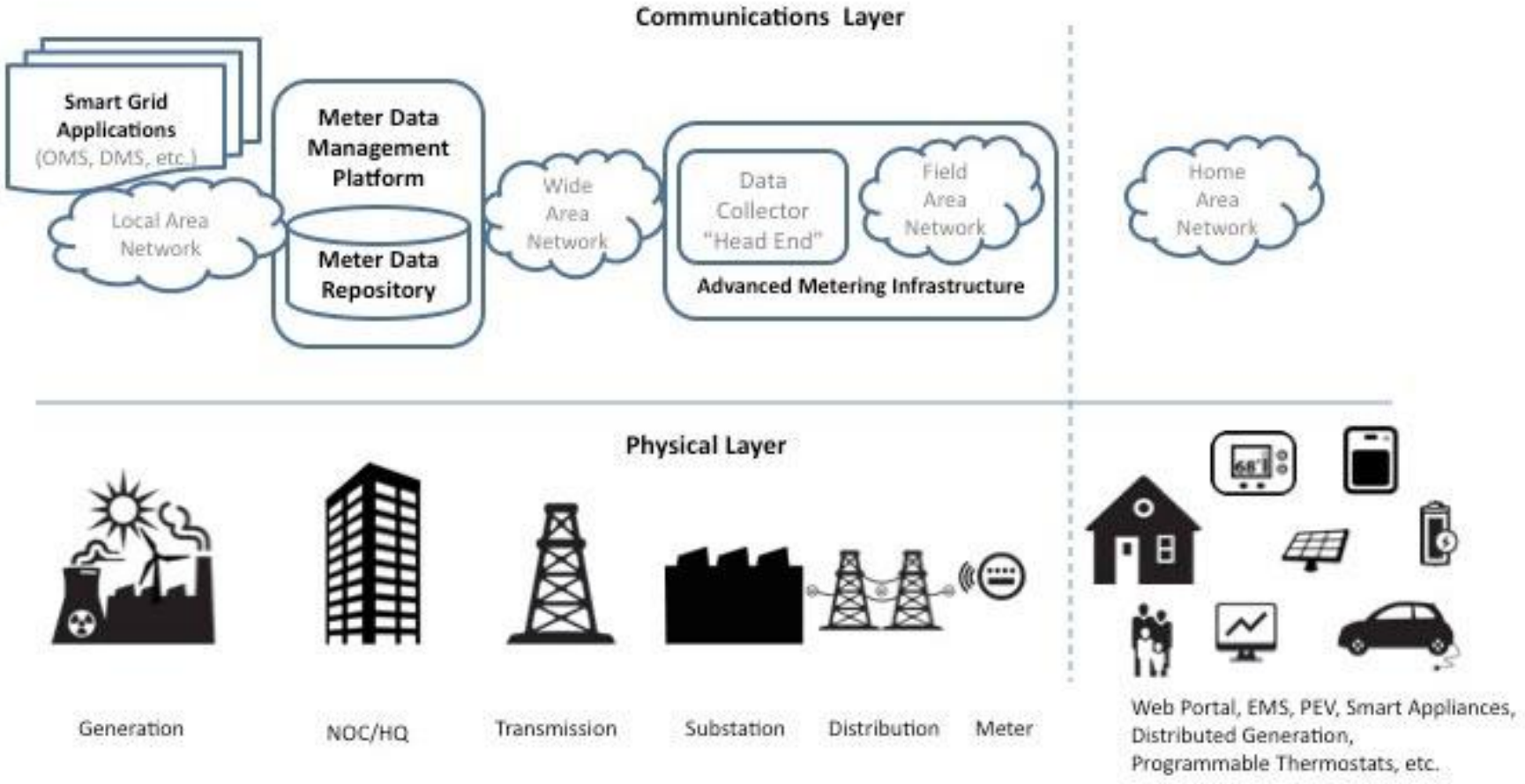
SMART GRIDS

What is a Smart Grid ?



→ **"An electricity network that can intelligently integrate the actions of all users connected to it (generators, consumers and those that do both) in order to efficiently deliver sustainable, economic and secure electricity supplies"**

Smart Grid infrastructures



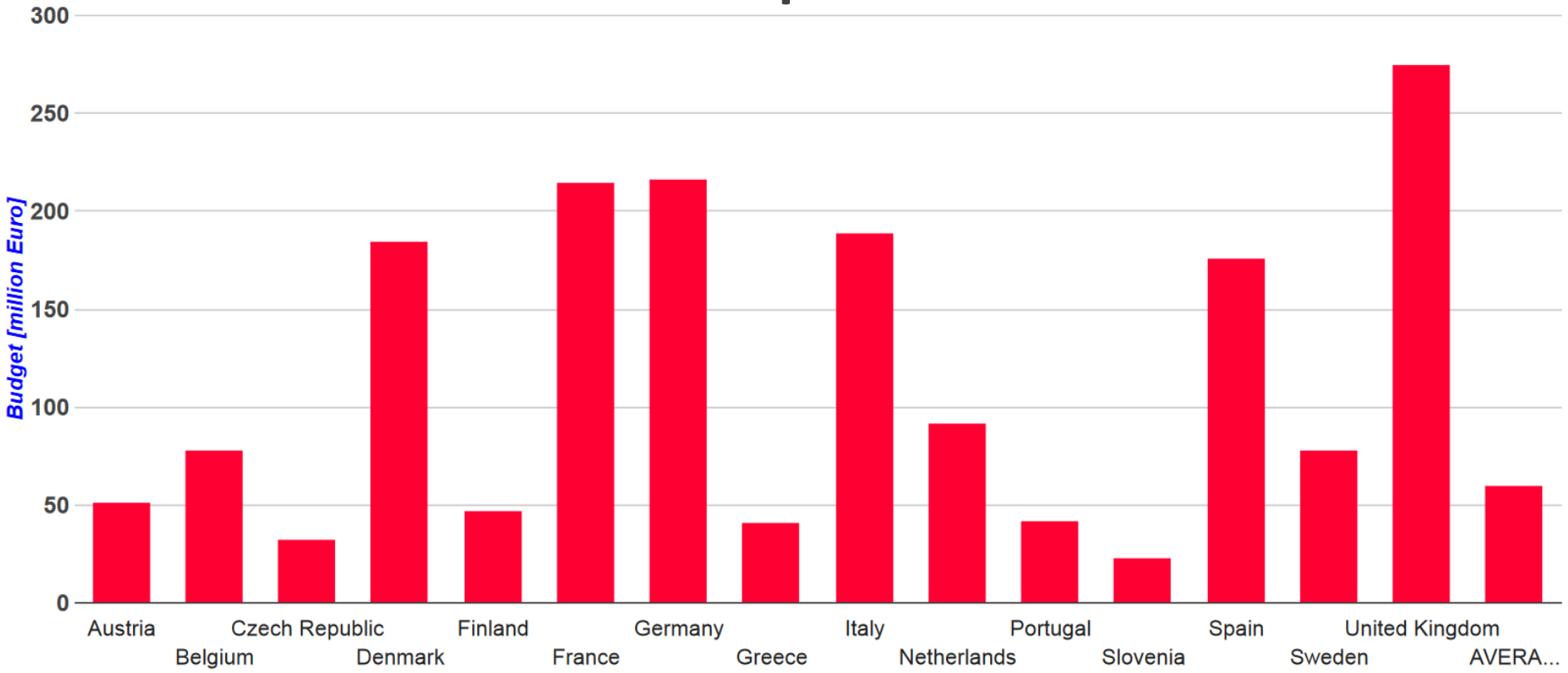
Smart Grid benefits

- Overall goals (European perspective)
  - Cutting greenhouse gases by 20%
  - Reducing energy consumption by 20% through increased energy efficiency
  - Meeting 20% of the EU's energy needs from renewable sources
  - Enabling the set-up of an internal European market

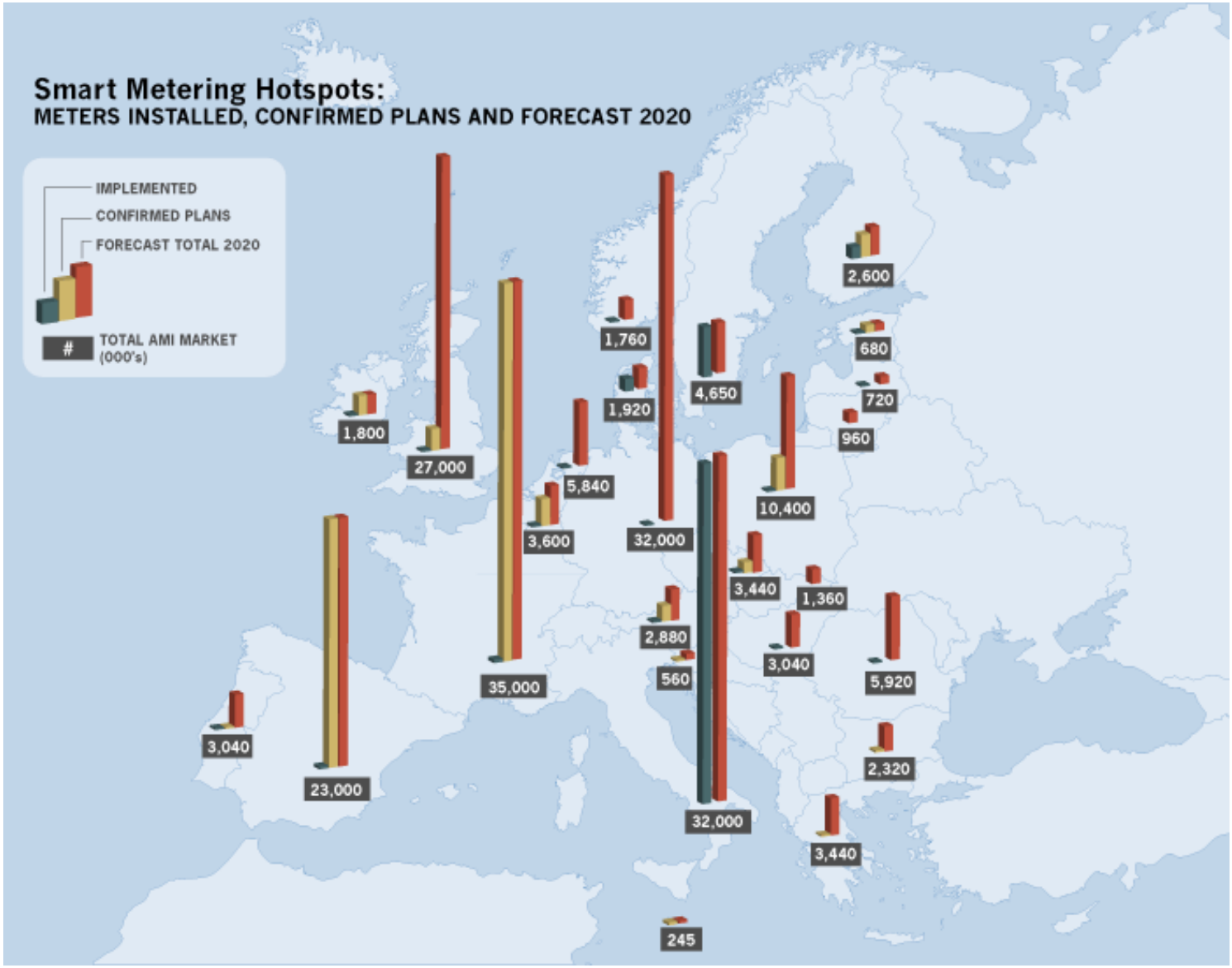
Smart Grid benefits

- Detailed goals
  - Meter reading costs reduction and accuracy increase
  - Billing improvement and complaints on meter reading reduction
  - Collection time and rate improvement
  - Remote connection/disconnection
  - Complex tariff system
  - Network planning and operation improvement
  - Demand-side management
  - Maintenance costs reduction and reliability improvement

Smart Grid projects in Europe (2005-2013)



→ **56 billion Euros by 2020 (estimation Pike Research)**



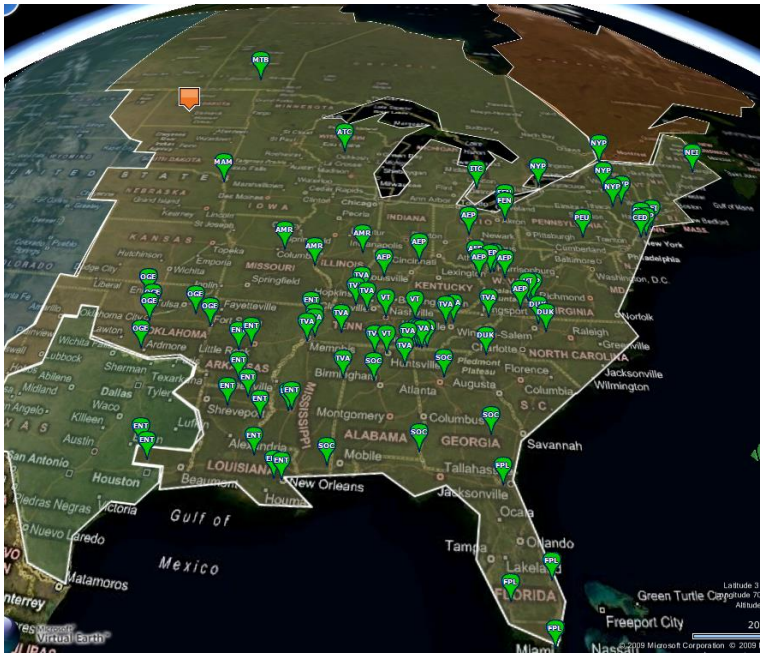
Source: The Smart Grid in Europe, 2012-2016: Technologies, Market Forecasts and Utility Profiles (GTM Research)

SMART GRIDS AND BIGDATA

GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

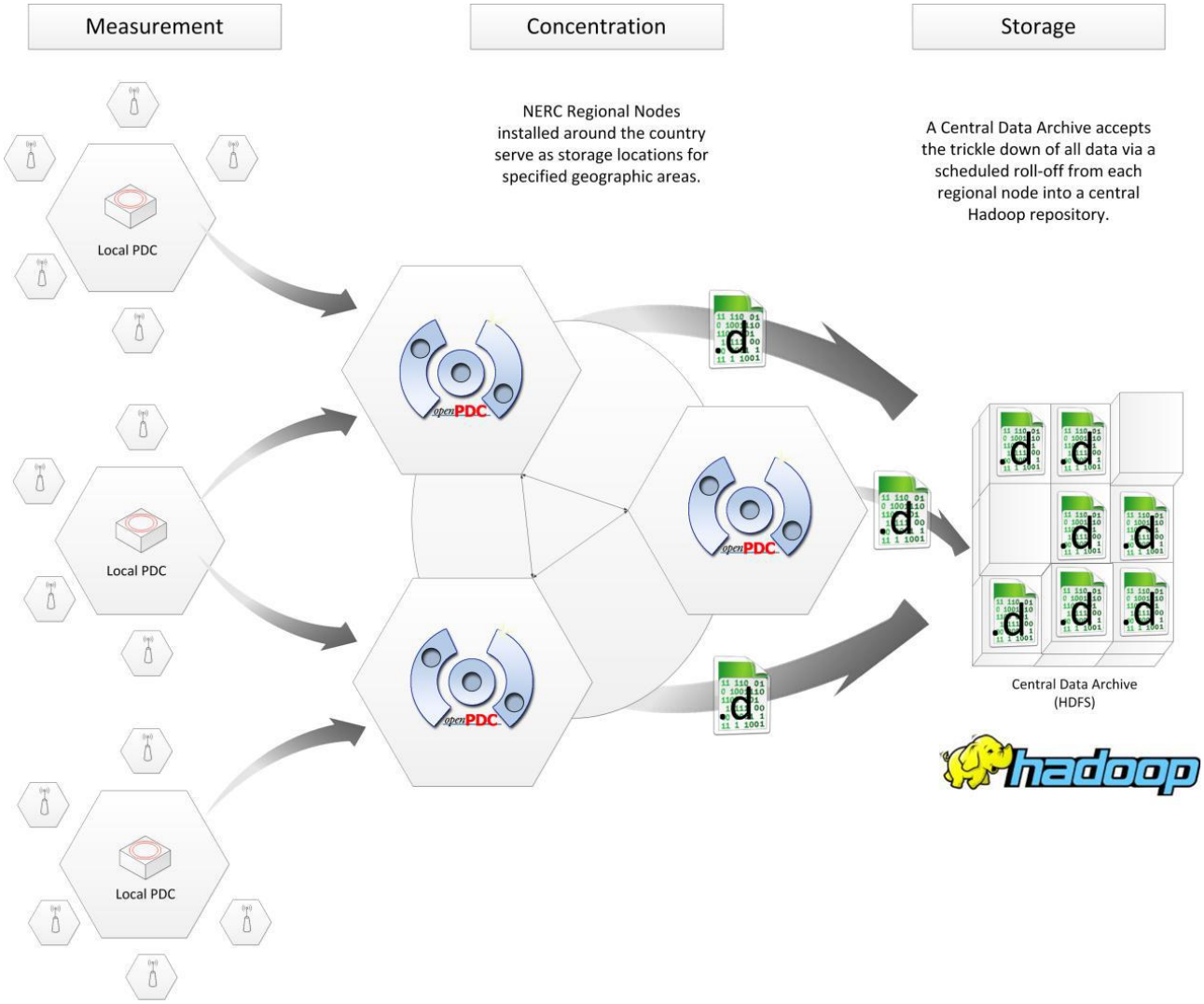
Example of OpenPDC

- Complete set of applications for processing streaming time-series data in “real-time”
  - Measured data is gathered with GPS-time from multiple input sources, time-sorted and provided to user defined actions, dispersed to custom output destinations for archival
  - Started at the Tennessee Valley Authority (TVA) following 2003 blackout



- 120 Sensors
- 30 samples/second
- **4.3B** Samples/day
- Housed in Hadoop

Example of OpenPDC



Courtesy of Josh Patterson, Cloudera

Example of OpenPDC

- Why Hadoop ?
  - 4.3 billions sample a day
  - Cost of SAN storage became excessive
  - Little analysis possible on SAN due to poor read rates on large amounts (TBs) of data
  - Scale Out, not Up
  - Linear scalability in cost and processing power
  - Robust in the face of hardware failure
  - Not fans of vendor lock-in
  - The “Haystack” in PMU data typically involved in scanning through TBs of info to find the one particular event we were interested in
  - RDBMs simply do not work with high resolution time series data
  - Need for Ad-Hoc processing on data to explore network effects and look at how events cascade across the grid

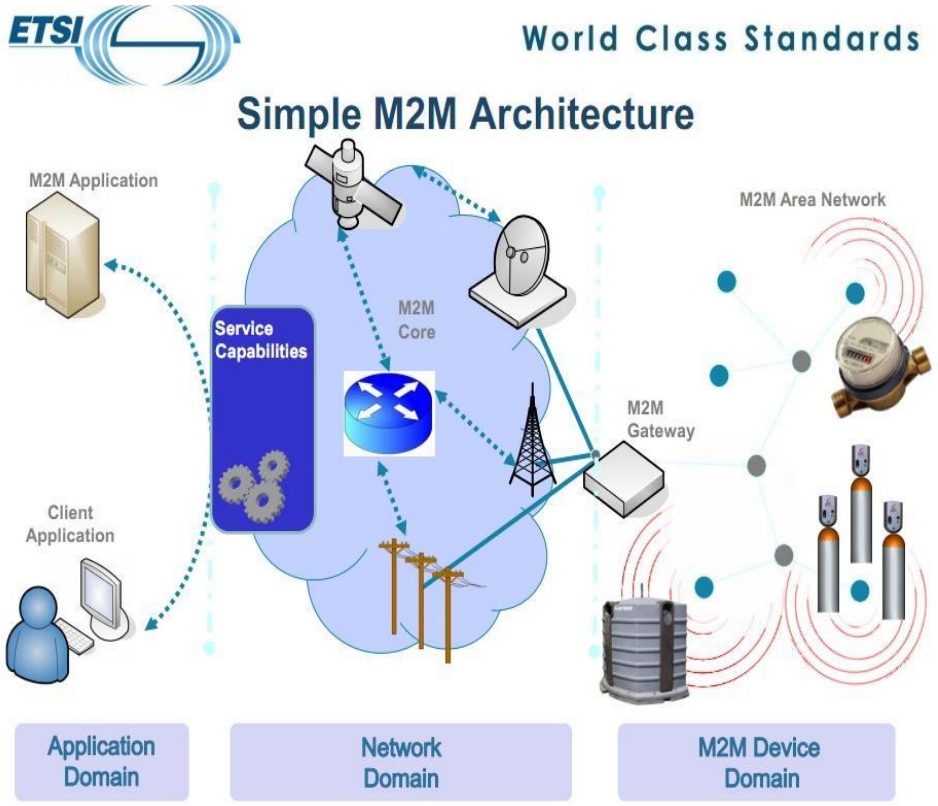
Courtesy of Josh Patterson, Cloudera

Smart Grid related data

- Power data
  - power consumption, power quality, voltage and many more
- But also many other relevant information
  - Meteorological parameters, such as temperature, humidity, cloud cover or wind
  - Indoor parameters, such as temperature, ambient noise or brightness
  - Events / Signal / Alerts possibly coming from various entities in the grid
  - Information about the buildings or about the customers
  - Data from Geographical Information Systems
  - Electricity prices from the market or from forecasting systems
  - Rate plans from utility companies
  - Behavioural data collected by CRM

Power Data Collection

- ETSI-M2M
  - Provides an architecture with multiple service capabilities
  - Structures data exchange
  - Standardisation of data handling
  - Exposes data through HTTP Rest interfaces
  - Enables an easy creation of M2M applications



Power Data Collection

- The GreenButton initiative
  - launched in January 2012
  - 35 utilities and electricity suppliers in US
  - 36 million homes and businesses
  - data exchange format and protocol

→ **provide utility customers with easy and secure access to their energy usage information in a consumer-friendly and computer-friendly format**

GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

Data Volume

- 35 millions of meters, 1 power measure every 10 min
  - Big size utility such as EDF in France
  - 120 Tb per year, around 40 Gb a day
- 3,5 millions of meters, 1 power measure per min
  - Medium size utility such as Energa in Poland
  - 120 Tb per year, around 40 Gb a day
- 35 000 meters, 1 power + 1 T° measures per sec
  - Service provider such as GridPocket in France
  - 144Tb per year, around 50 Gb a day

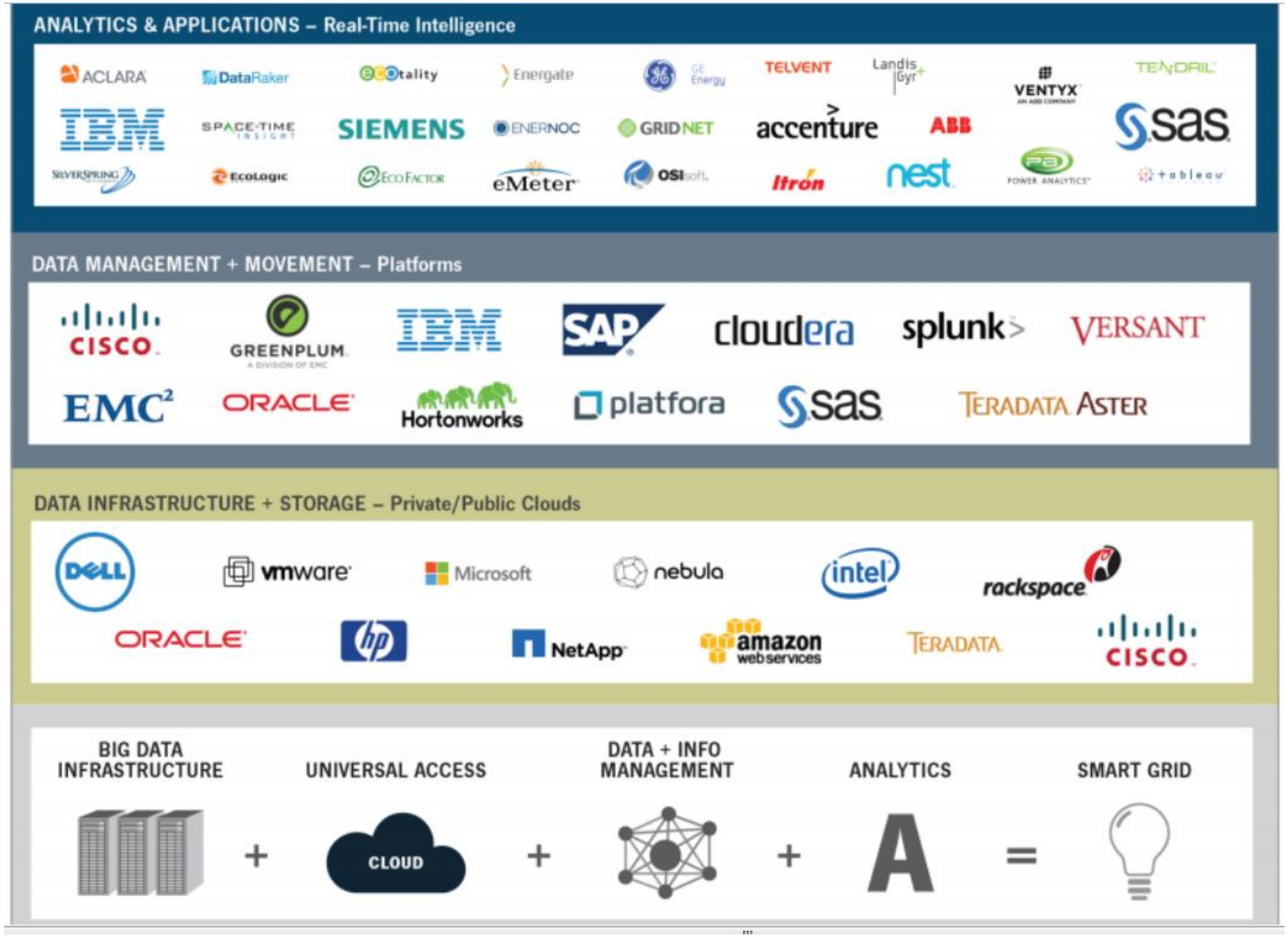
Data Integration

- Roberta Bigliani, head of IDC Energy Insights for Europe, commenting on the Smart Grid data:

  > "We could be in a situation where we are creating silos [n.b. data in a silo remains sealed o from the rest of the organization] of data rather than making more consistent availability of the data, data needs to be validated and translated into a meta-data model, to create something that is usable by multiple applications. IT people need to work with the line of business to define a master data sort of approach and try to create a layer where all the data coming from meters or operational systems, are transformed into pieces of data that different applications can call."

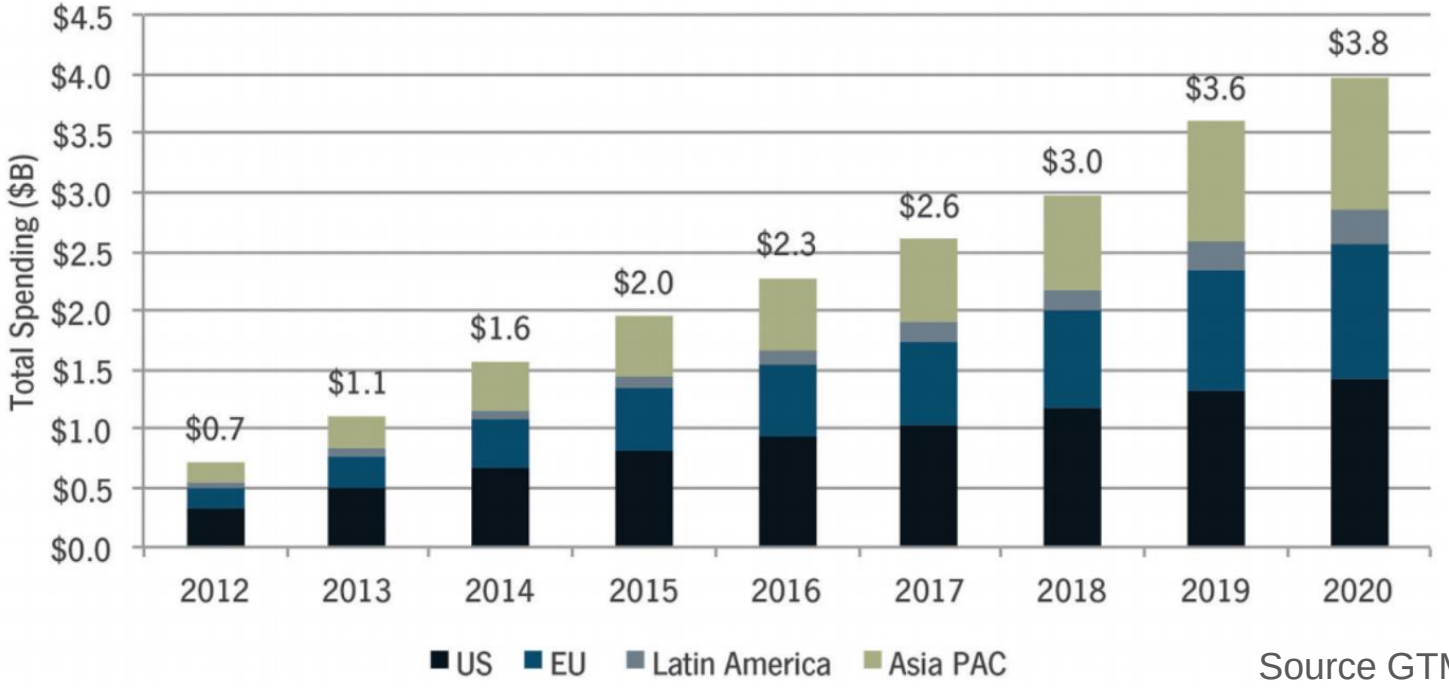
→ Building systems capable of collecting and integrating all these data is one of the main challenges of the Smart Grid.

Vendors landscape



Source GTM Research

Smart Grid analytics



Source GTM Research

→ **Electric utilities will spend $322.5 million on analytics in the U.S. alone in 2012, a figure that will reach $1.4 billion by 2020.**

Smart Grid data analytics

- Three main categories of application
  - Revenue Management

    Load forecasting, theft detection, advanced rate plans, demand/response

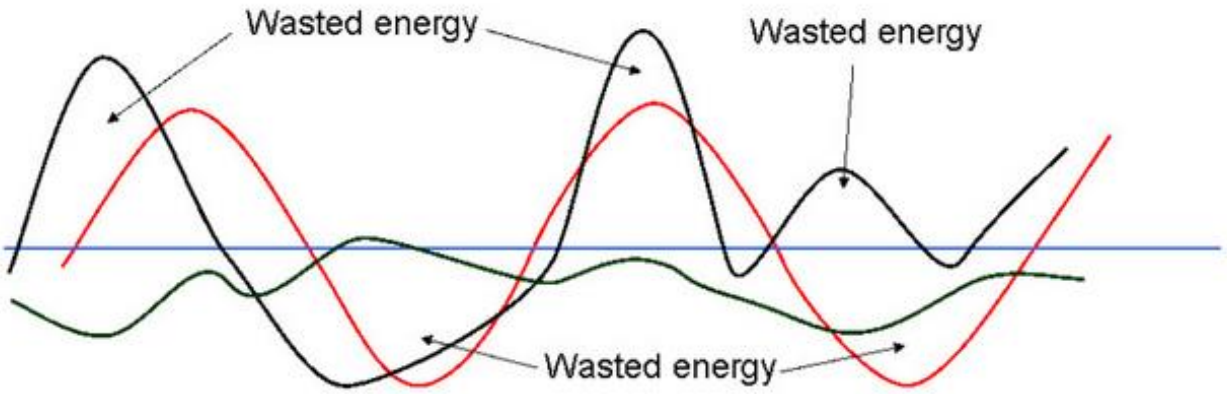
  - Consumer Engagement

    Conservation tips, optimal plan rate selection

  - Distribution Optimization

    Outage management, distribution network planning

Demand/Response



Base Load Supply: ——— Energy supplies that cannot be quickly varied. Includes coal and nuclear options.

Intermittent Supply: ——— Energy supplies that vary with natural phenomena such as waves, wind or sunlight available.

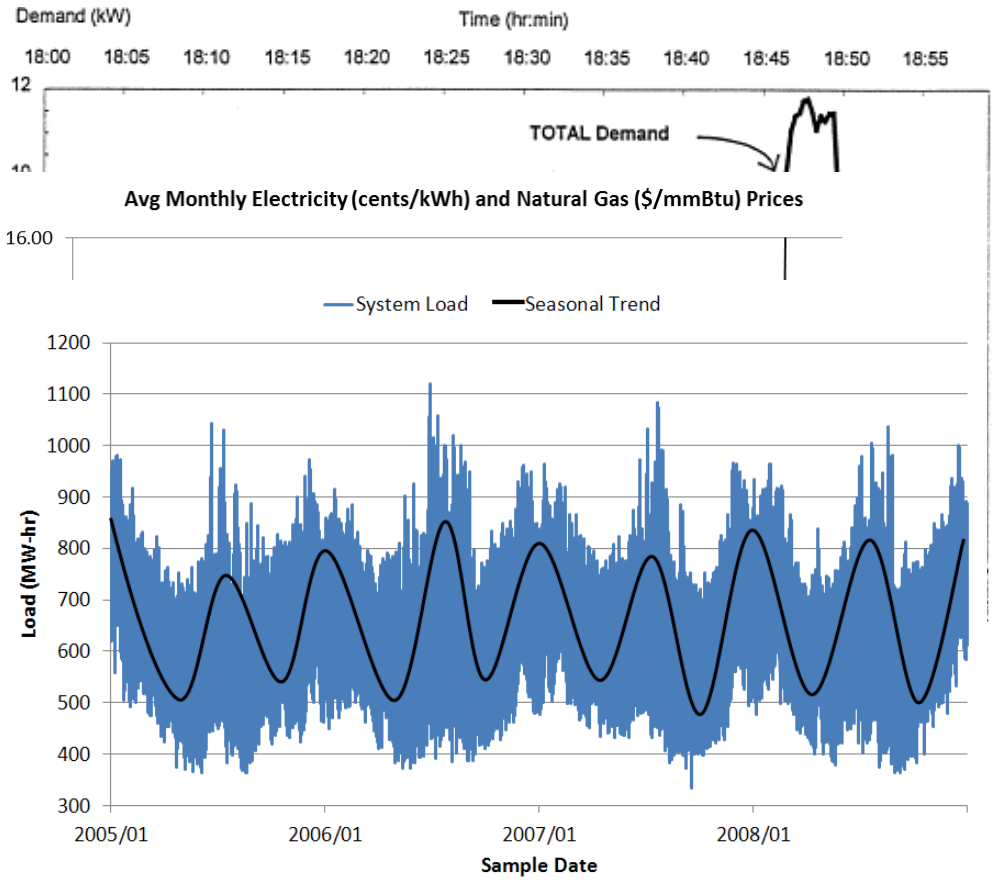
Adjustable Supply:——— Energy supplies that can be easily adjusted such as from water (hydro).

Demand: Generally demand follows a sinusoidal pattern that peaks during the day and troughs at night. It is usually predicable depending on the weather and economic circumstances.

source: http://www.tececo.com

Smart Grid data analytics

- Analytics based on fundamental operations
  - Aggregation and disaggregation
  - Correlation
  - Trends and exceptions detection
  - Clustering and Forecasting

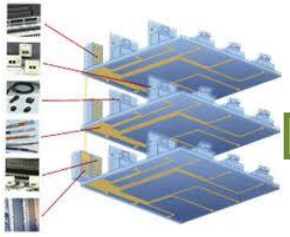


GRIDPOCKET

Relational Energy Services

Smart Meters

0110100101



Intelligent buildings

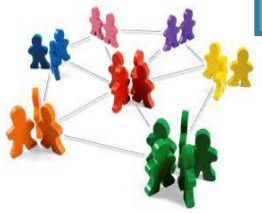
0110100101

Other data sources

0110100101

GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

Behavioral consumer engagement



M2M apps

Energy search engine

Semantic M2M (ETSI)

Technology protected by patents, copyrights and trade secret

$$$

End Users

- Gratification
- New services
- Energy Savings

$$$

Utility

- Increased customer profit
- CO2 reduction
- Customer retention

$$$

Manufacturers

- Embedded software
- Compatibility
- New features

Problem with energy efficiency

50% of people find it difficult to read energy graphs

65% find financial incentives too low to be interested in efficiency

92% do not really understand underlying environmental challenges

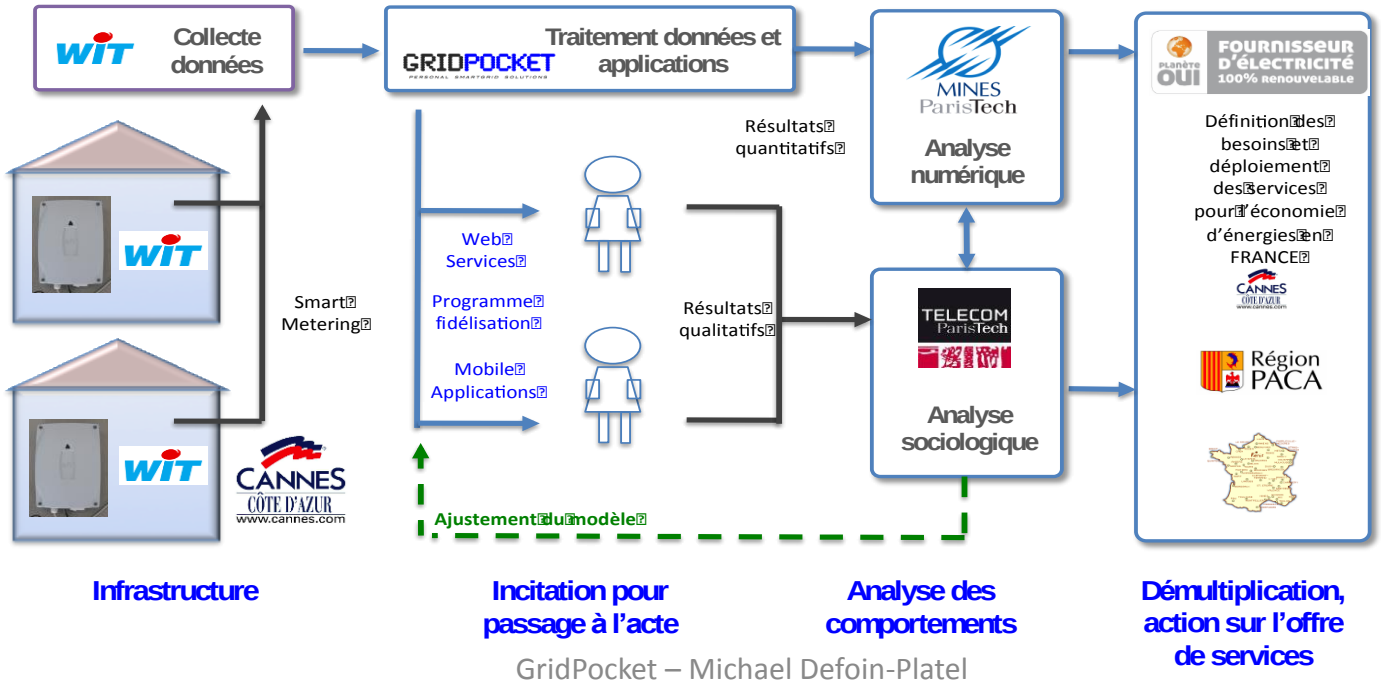
Sources : EDF PACA 2011, Health Literacy Inst.

GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

GRID Teams

- Goals
  - Energy management through residential customer empowerment
  - Personalised long term incentive plans

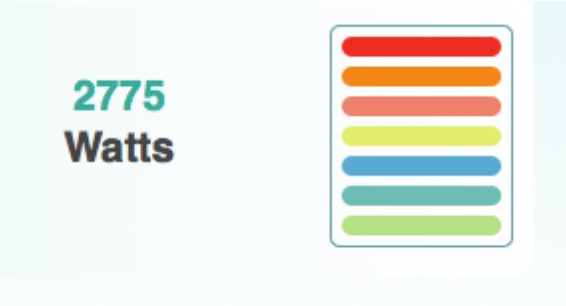
- Setup
  - 30 households in the Cannes area
  - One year of data
    - power, temperature

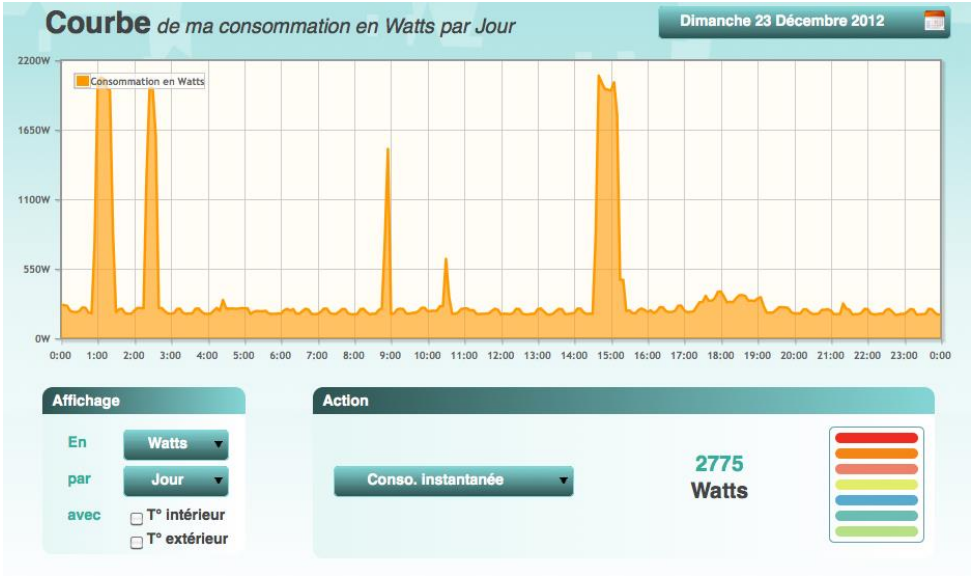


GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

- Energy services

Real-time consumption



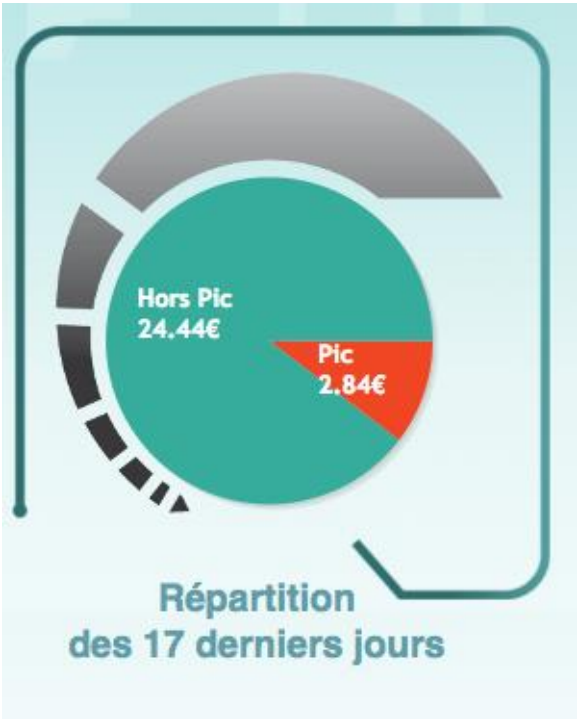


Power load

GRIDstreams

- Energy services

Weekly spendings

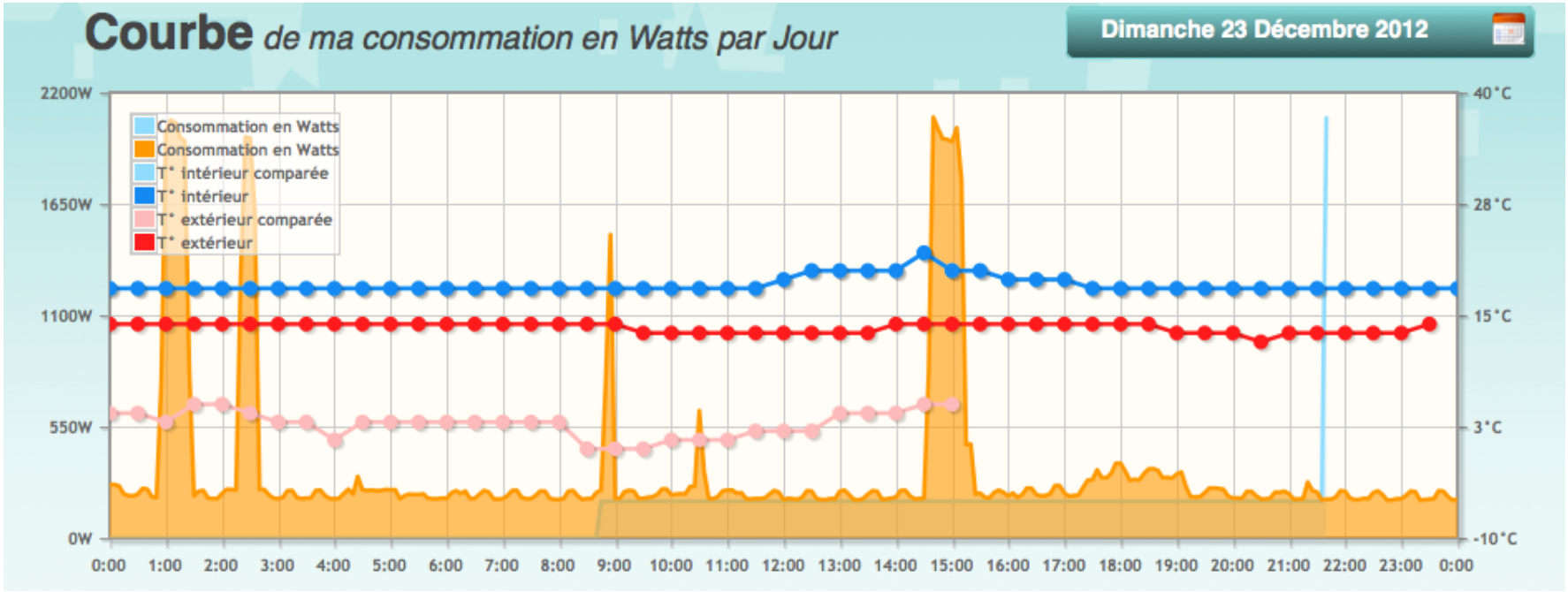


Peak / off Peak

GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

GRIDstreams

- Energy services



Historical consumption vs environment

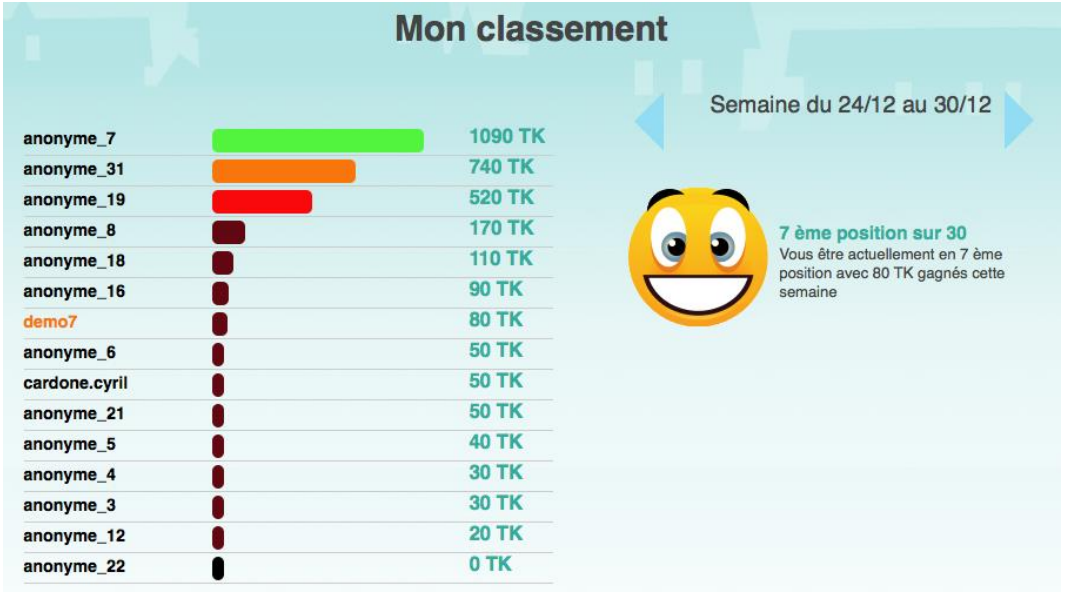
GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

GRIDstreams

- Energy services

Consumption targets

Comparison with peers

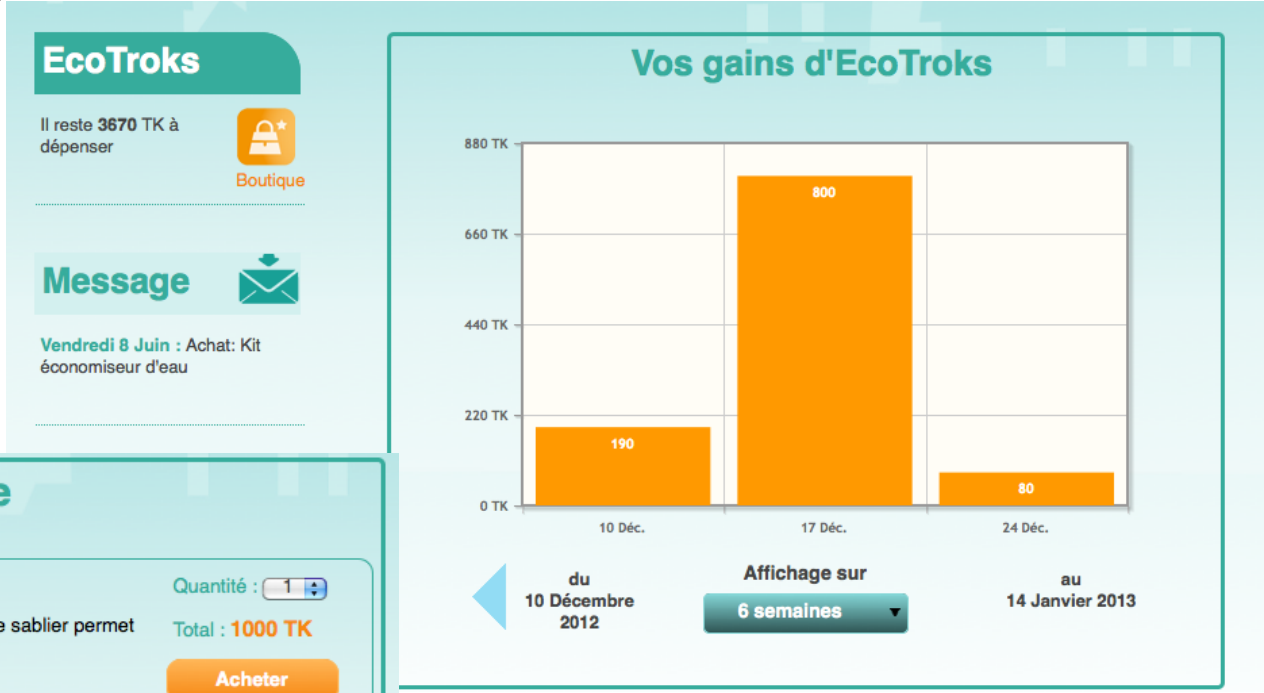


GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

GRIDstreams

- Energy services

Rewards



Unique EcoTroks engagement system

Personal computer WEB

Tablets

Mobile apps

From : Grid-Teams
To : You
Subject : Congrats!

You have won 1200 eco-troks last week.

Email

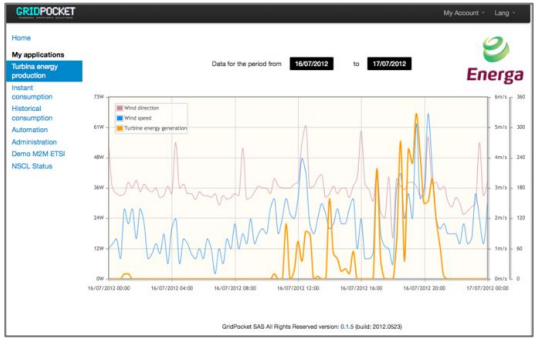
SMS Alert*

Letter / invoice*

Projects of GridPocket

Home control system

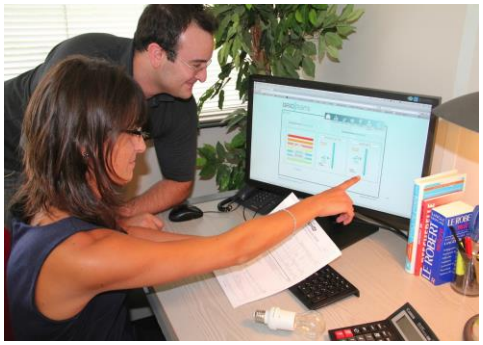
Embedded energy systems



Renewable energy monitoring

Electric car charging

Energy BigData Analytics



Behavioral metering applications

International customers and partners

Utilities and operators

Energa, PGNiG, Fournisseur d'électricité 100% renouvelable, GDF SUEZ, EDF, CA Store, orange, ERDF – Électricité Réseau Distribution France

Research

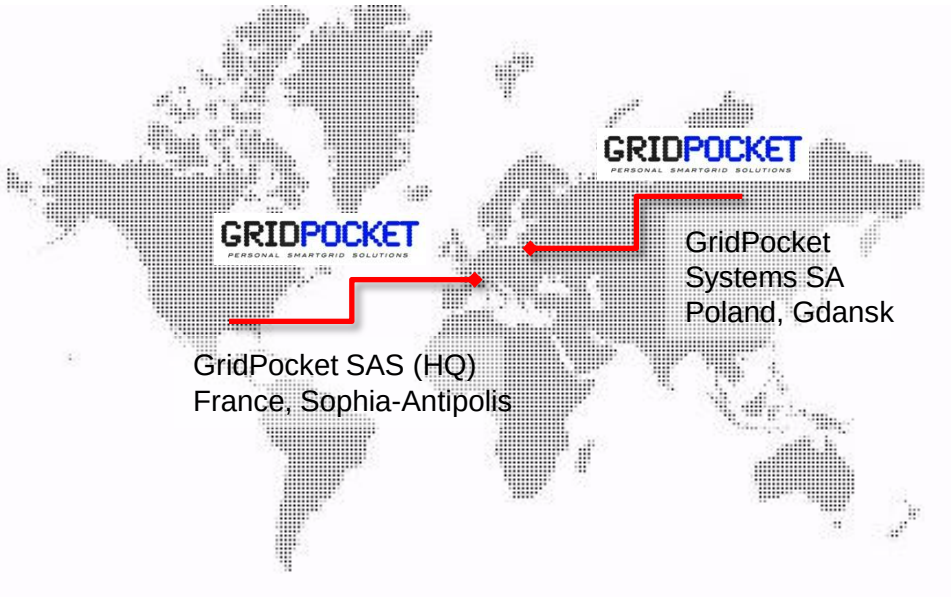
TELECOM ParisTech, EURECOM, MINES ParisTech, Inria, EPFL École Polytechnique Fédérale de Lausanne, Warsaw University of Life Sciences, Technische Universität Berlin, LAAS-CNRS

Industry

Digi, actility, WIT, VEDIA, Itron, Sagemcom, EMC², Sierra Wireless, InterSystems, Symantec

Institutional

Région Provence Alpes Côte d'Azur, capenergies, Commission de régulation de l'énergie, oséo, ETSI World Class Standards, ADEME, CCI Nice Côte d'Azur, Seventh Framework Programme, Narodowe Centrum Badań i Rozwoju, Clean Tech Open France



Recent awards and honors

- CapEnergies Cluster
  → **Innovative SME Label 2010**
- CRE (Energy Regulation Commission)
  → **Reference publications**
- Digital Green Growth
  → **Best research project award 2011**
- CleanTech Open
  → **Best start-ups award 2012**
- EDF Intelligent Energy
  → **Award for the best consumer service 2012**
- Japan External Trade Organization
  → **Selection top 14 global SMEs at Expo 2013**

BIGDATA@GRIDPOCKET

GridPocket offers scalable solutions for Smart Grid

- Hadoop support
  - Hardware and Architecture specification
  - Platform profiling and tuning
  - Virtualization
  - Code development and optimization
  - Ecosystem: Hive/PIG/Oozie/HDFS/Hbase and more
- Product development
  - Proof of concept
  - Saas

GridPocket offers scalable solutions for Smart Grid

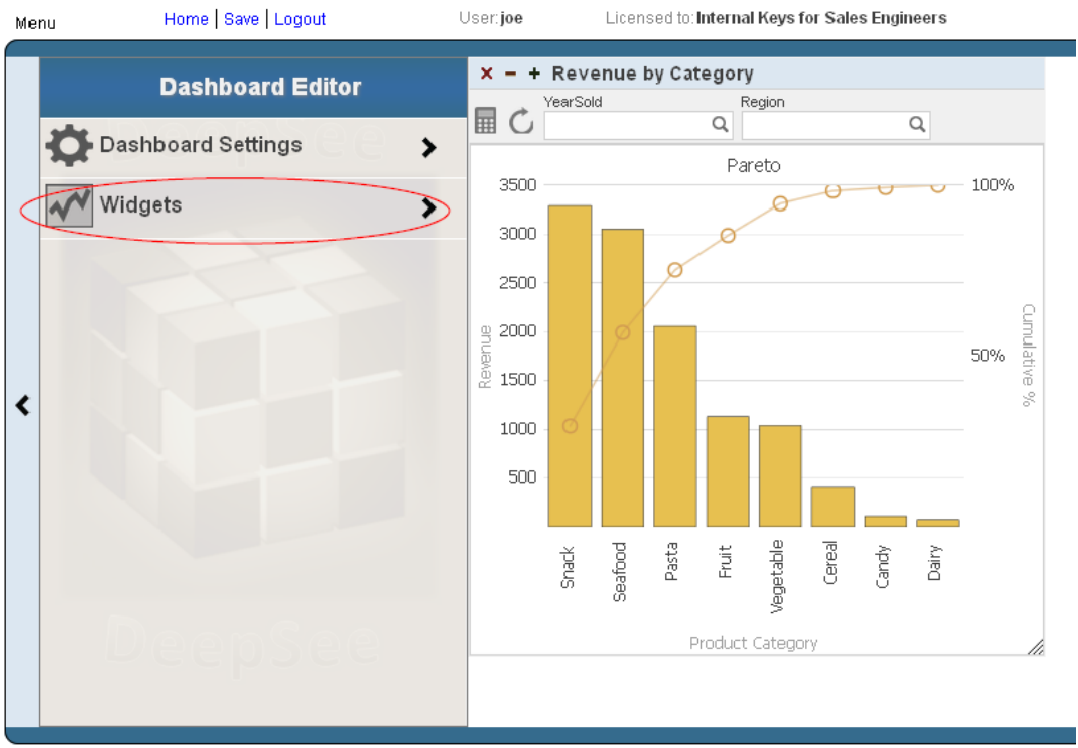
- Scalable Smart data Grid Analytics
  - Complex graph-based data integration, if needed
  - Data exploration
  - Data mining , Graph mining
  - Dashboards
  - Deliver reports or workflows + maintenance

- Benchmarking and research activities
  - Investigate new technologies, e.g. in memory computation
  - Compare architectures
  - Profiling of algorithms
  - The BigFoot project, (Eurecom, EPFL, TU Berlin, Symantec and GridPocket)

Proof-Of-Concept with InterSystems

GRIDPOCKET PERSONAL SMARTGRID SOLUTIONS

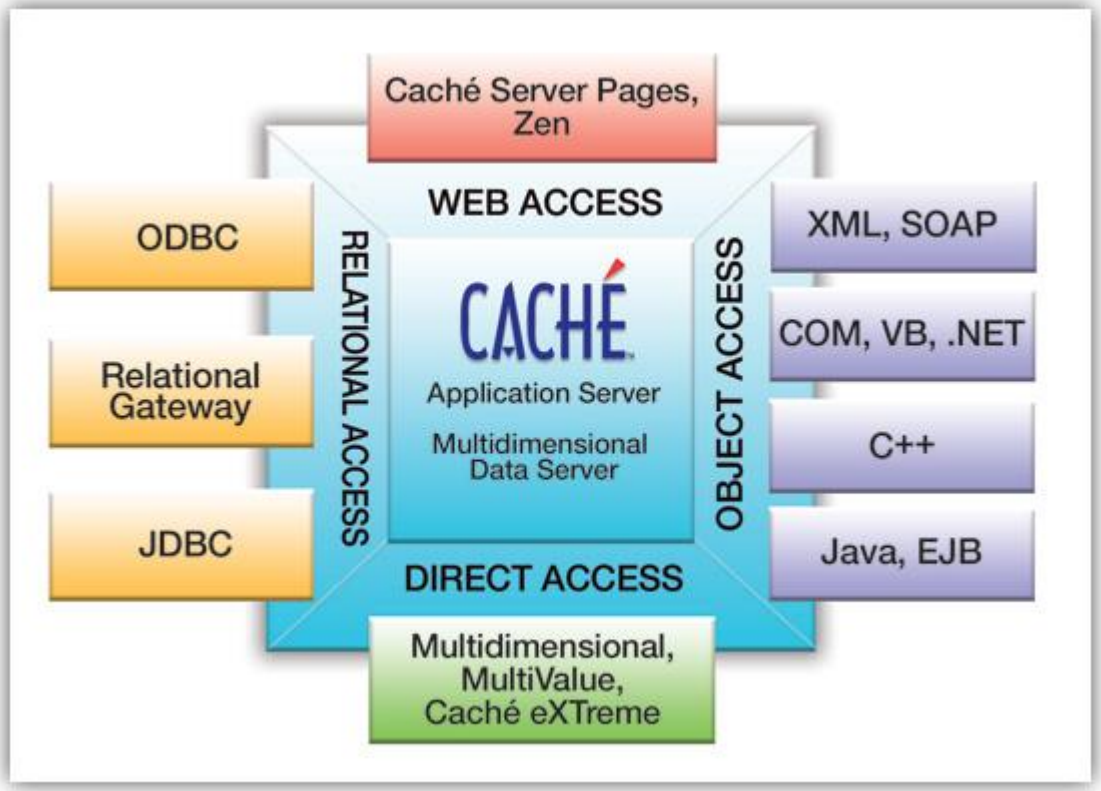
- US Company
  - +1300 employees, 25 offices, 22 countries
  - +2000 partners (editors, integrators, final customers)
  - +300 000 system operated by IS in 88 countries
  - +5 000 000 users



Proof-Of-Concept with InterSystems

GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

- Original technology
  - Caché : noSQL DB (since 70's)
  - Ensemble : dev. Studio, integration, connectors, …
  - Manage and process:
    - structured data
    - unstructured data
    - events



European Space Agency

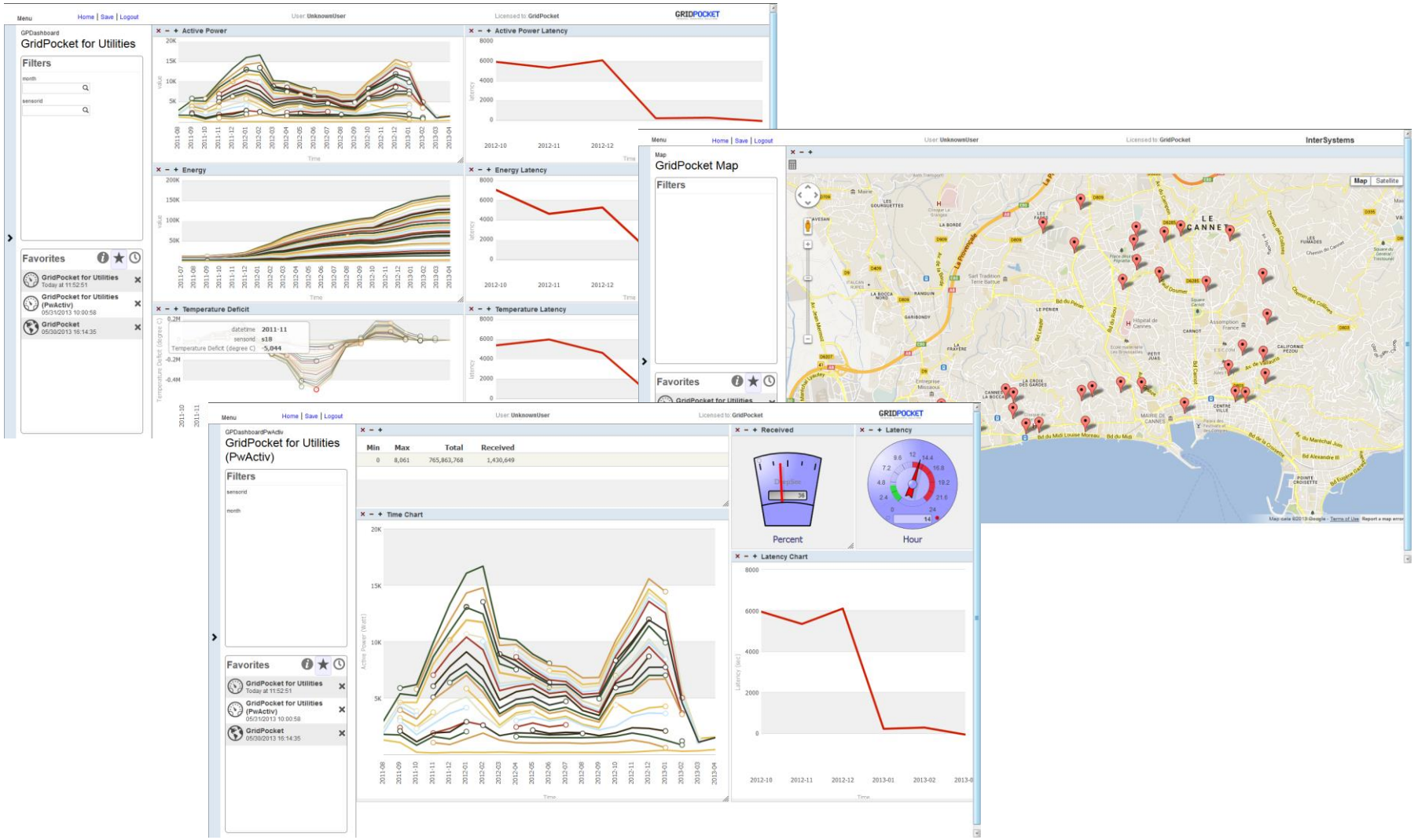
- Goal: Make the largest, most precise 3-D map of our Galaxy
- Expected

Challenge: Capture data for 1 billion celestial objects

1,000,000,000	objects
X 100	observations per obje
X 600	bytes per observatio
60,000,000,000,000	(6

Gaia

GP4U powered by IS



GridPocket and Hadoop

1 platform : 3 use cases

1. "Smart services for smart customers"

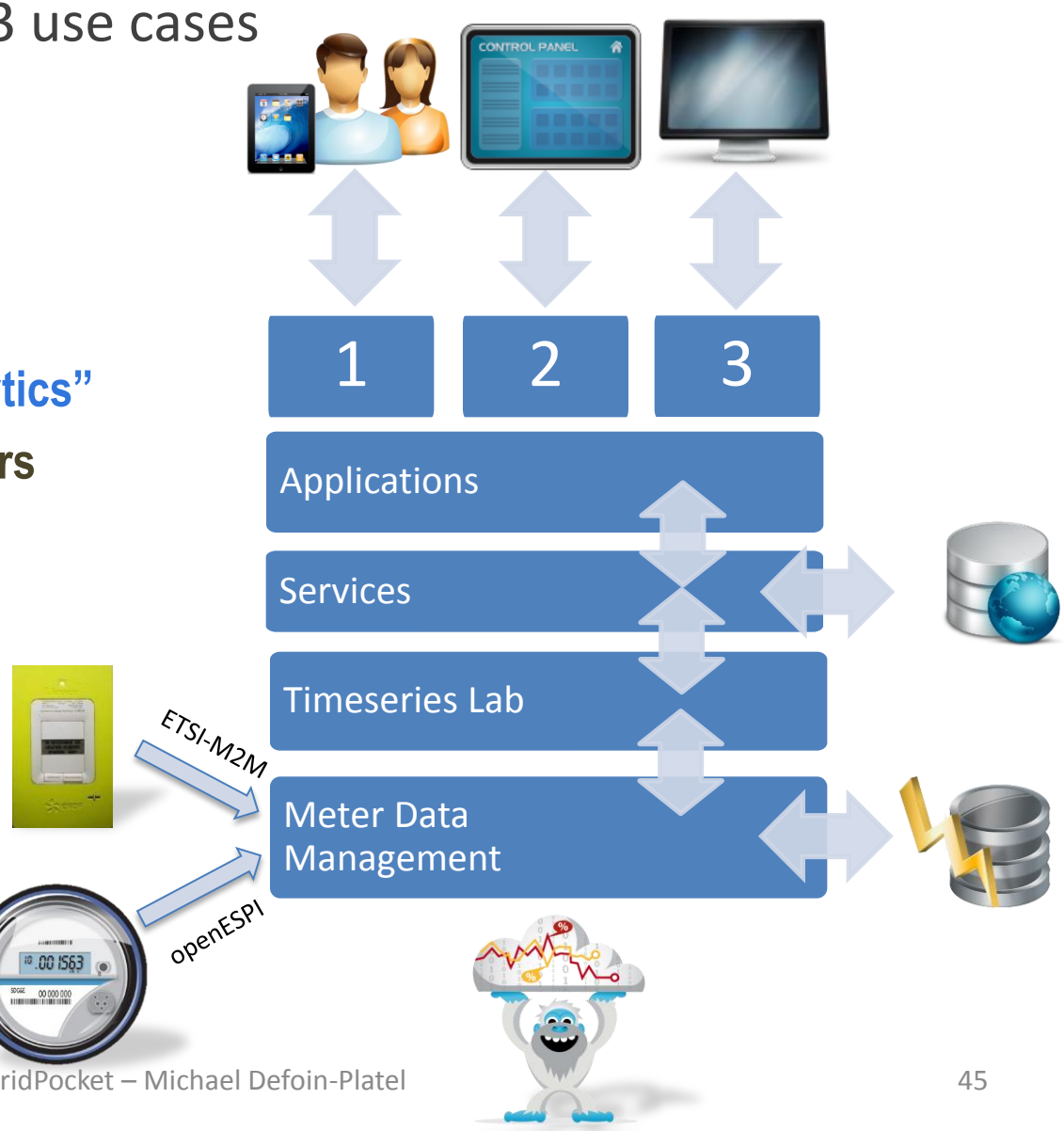
→ **dedicated to customers**

2. "Business Intelligence analytics"

→ **dedicated to utility managers**

3. "Mining smart grid data"

→ **dedicated to data scientists**



GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

Meter Data Management

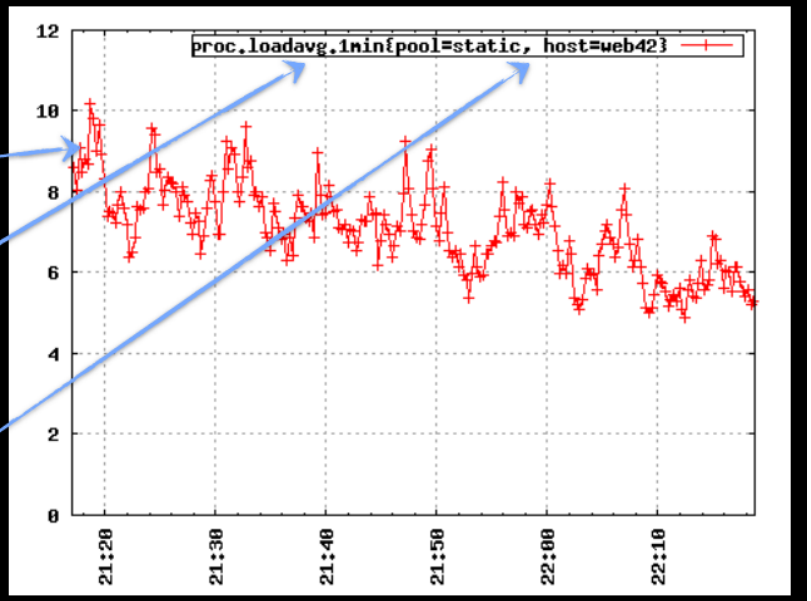
- GridPocket MDM specifications
  - Built upon Hadoop ecosystem for scalability
  - Collect meter data in multiple format, from multiple sources and multiple technologies
  - Handle batches (PIG) and streams (FLUME)
  - Basic validation, estimation and editing (VEE) functionalities
  - Deliver data in the appropriate format to the layer above with low latency and high availability (Hbase / openTSDB)

OpenTSDB

GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

Distributed, Scalable, Time Series Database

- Data Points
  (time, value)
- Metrics
  proc.loadavg.1m
- Tags
  host=web42 pool=static
- Metric + Tags = Time Series

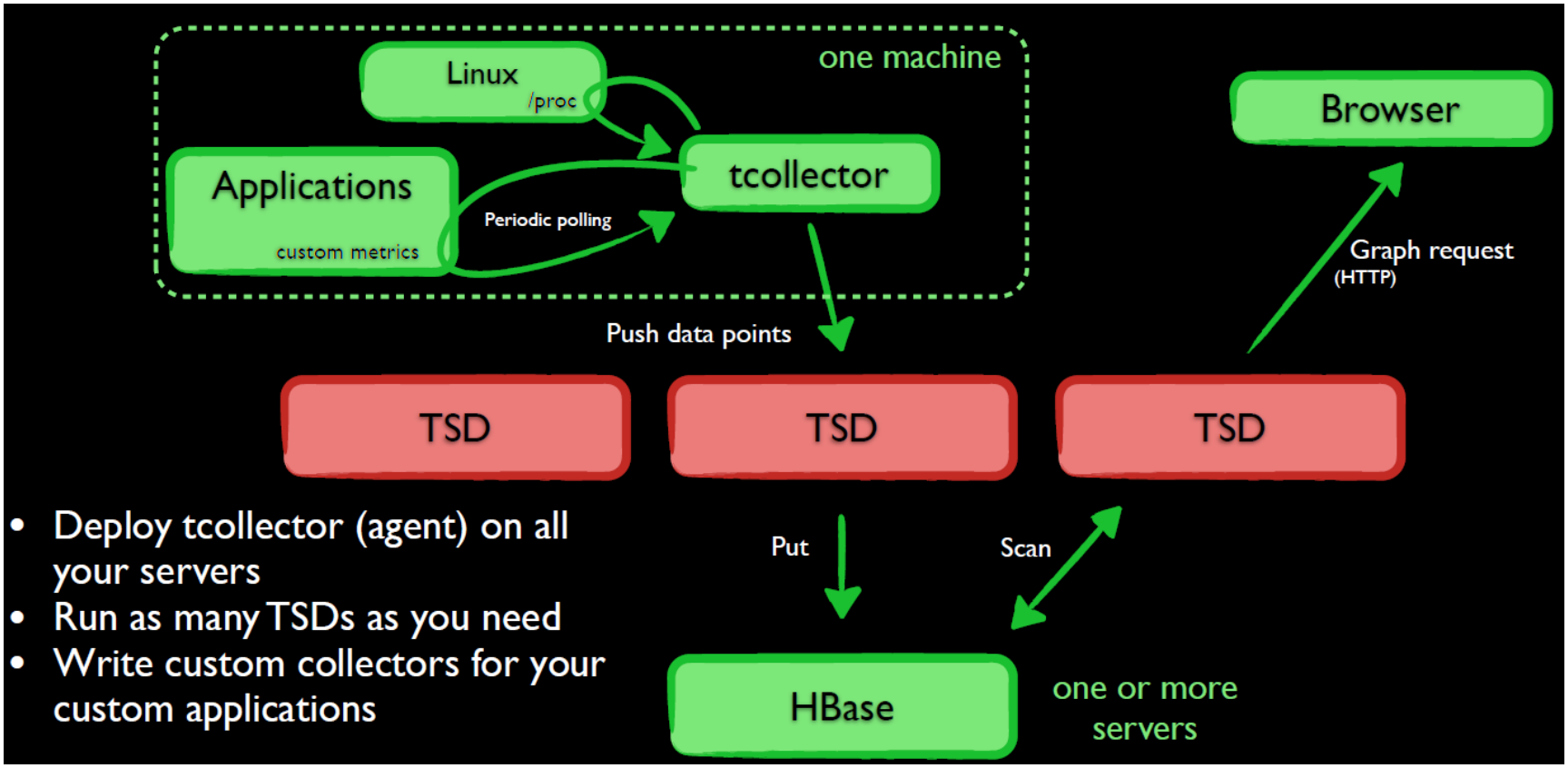


- At StumbleUpon
  - 600 M data points/day
  - 70 Billion data points stored
  - > 2000 data points / sec /core

Courtesy of Benoit Sigoure – Stumble Upon

OpenTSDB

Distributed, Scalable, Time Series Database



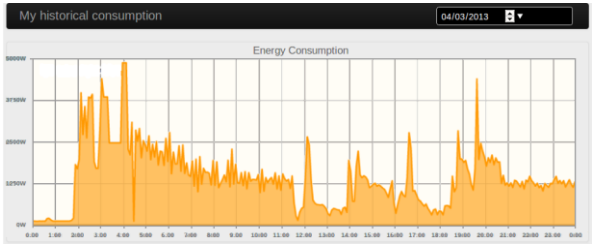
GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

Timeseries Lab

- Examples of Basic functionalities
  - Compression
    SAX, Piecewise Linear Approximation
  - Frequency
    Sampling, Moving Average, Interpolate, Slice
  - Aggregation
    Temporal, Spatial, Multi-criteria
  - Summarize
    Avg, Std, Max, Min, Sum
  - Distance
    Correlation, RMS, Dynamic Time Wrapping
  - Transform
    Normalize
- Examples of Advanced functionalities
  - Trending / Exception detection
  - Clustering / Forecasting
  - Disaggregation

→ **Library of a library of PIG UDFs**

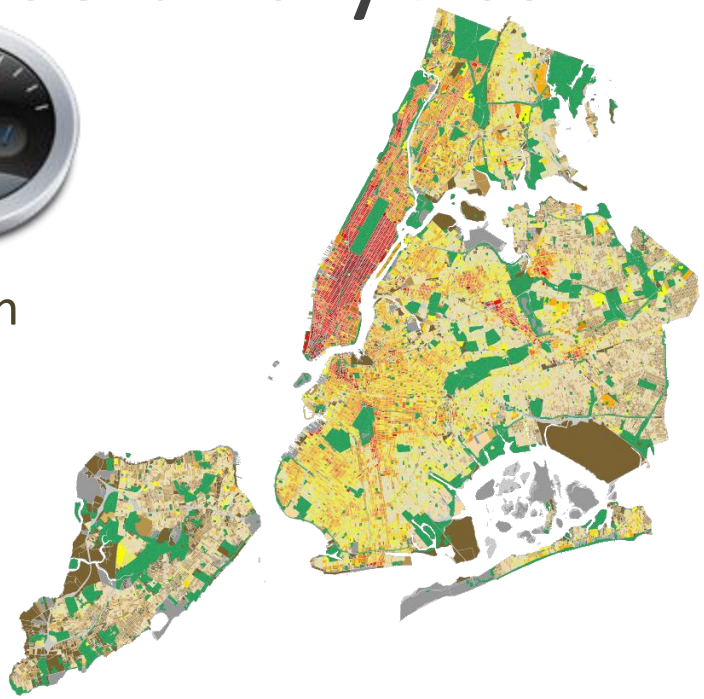
1. Smart services for smart customers



- Empowerment of customers
  - Historical load curve, Summary
- Provide personalized
  - conservation tips
    - E.g. based on correlation with weather or detection of excessive baseline
  - advanced power-related services
    - Comparison with “similar” users
    - Low frequency disaggregation

- Evaluate efficiency of Tips and Quality of services
  - Detect changes and correlate with IHM logs
  - Detect degradation of service

2. Business intelligence analytics

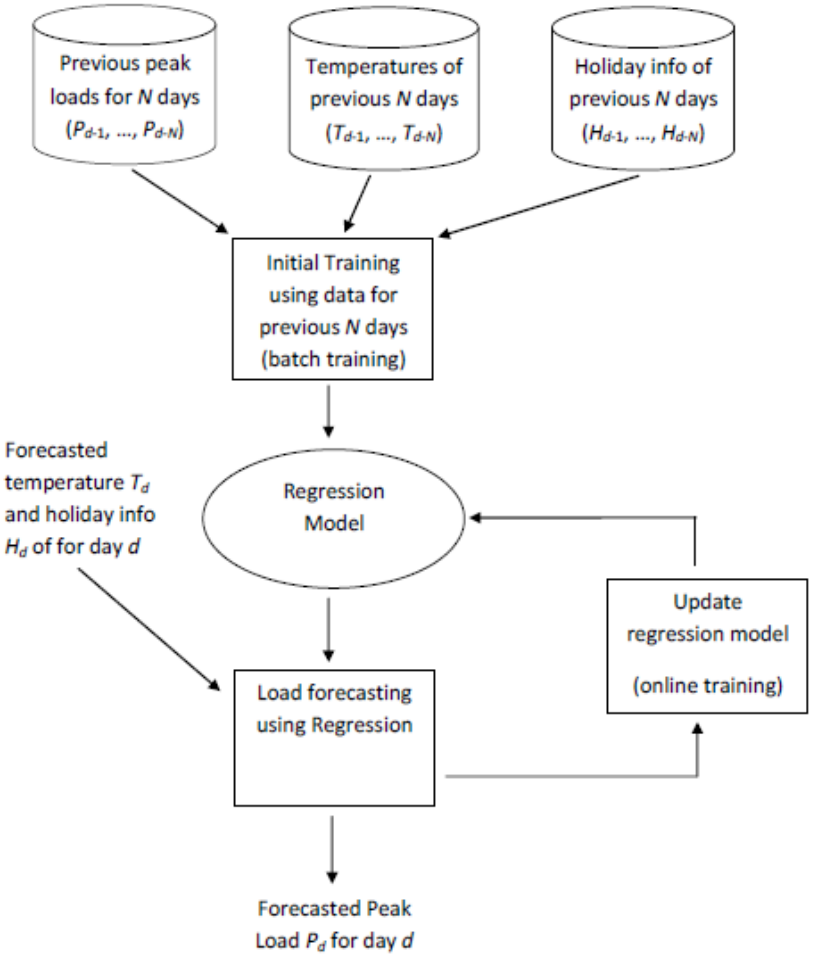


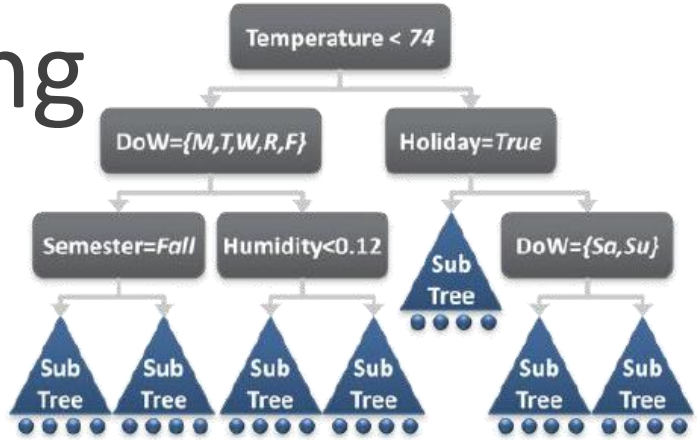
- Create virtual meters
  - Multi-criteria selection and aggregation
  - Historical load curve, Summary
- Identify
  - Exceptions, e.g. using state automata
  - Trends, e.g. days/nights, weekend, seasons
  - Multi-dimensional clusters
- Energy demand simulator

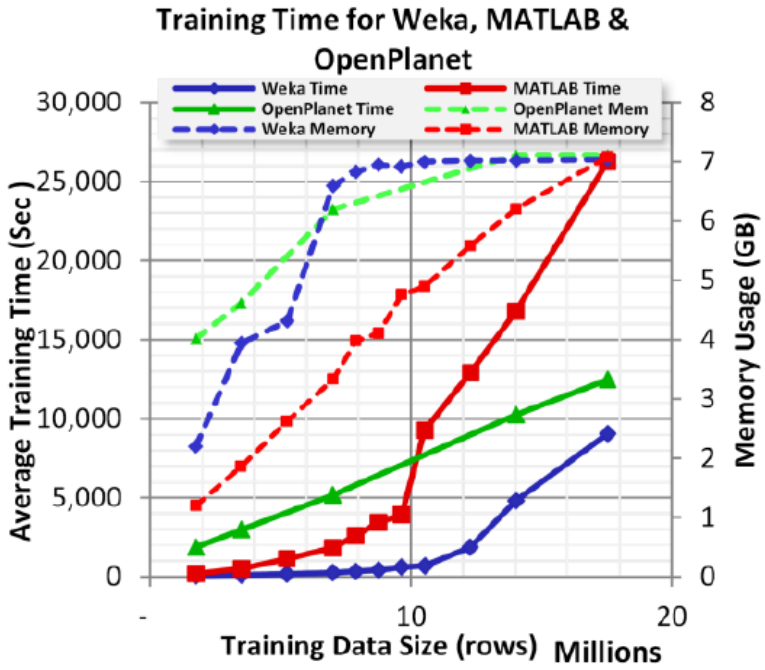
3. Mining smart grid data

- Implement complex algorithms
  - Examples
    - Forecasting daily consumption and peak load
    - Disaggregation using low frequency data
- Benchmark complex algorithms
  - Select the appropriate data
    - Publically available : REDD, BLUED, UMASS Smart*, …
    - GridPocket datasets
    - Simulated data
  - Prepare the data
    - Normalisation, Discretization, Sampling
  - Deployment on hybrid cloud
    - Virtualization
    - Privacy issue
  - Run and profil algorithms
    - Accuracy , Time, Scalability, Energy cost

Forecasting



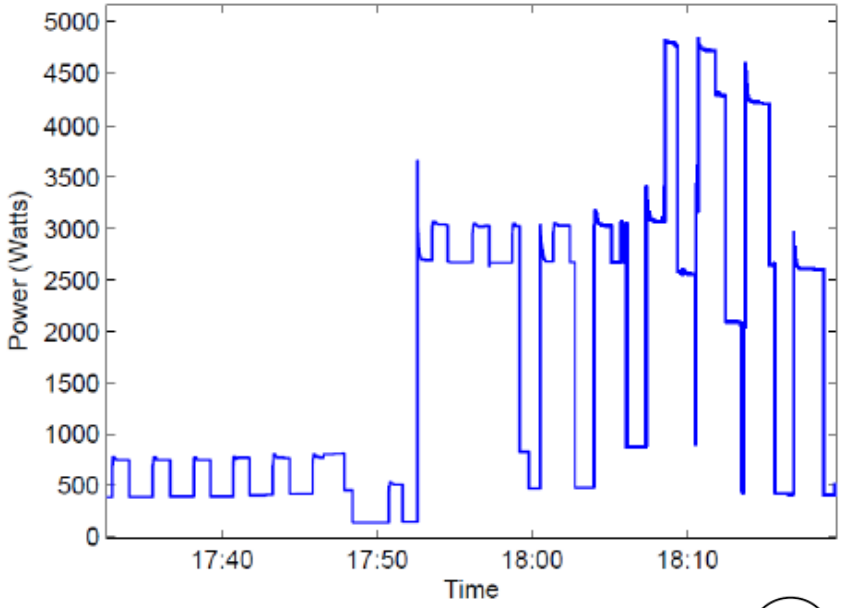




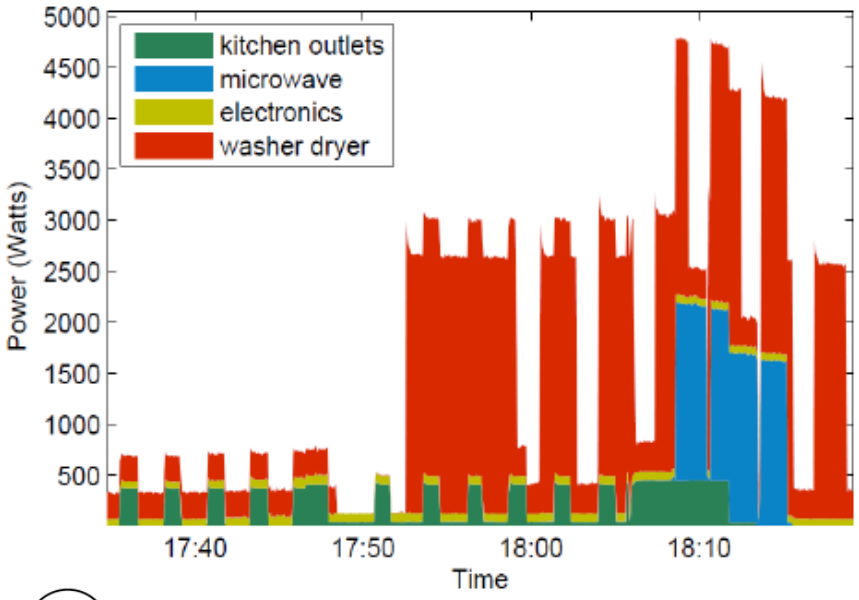
→ **Scalable Regression Tree Learning on Hadoop using OpenPlanet**

Disaggregation

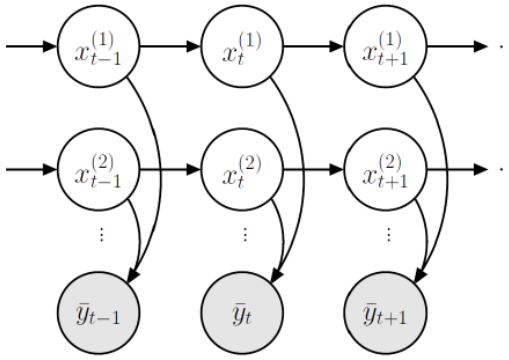
Total Power



Disaggregated Power



→ **Factorial Hidden Markov Model**



BIGFOOT PROJECT

BigFoot team

- EURECOM
  - Ass. Prof. Pietro Michiardi
  - Ass. Prof. Marko Vukolic
  - Dr. Matteo Dell'Amico
- SYMANTEC
  - Olivier Thonnard, Sr. Researcher
  - Marc Dacier, Sr. Director of Research
- TECHNISCHE UNIVERSITAT BERLIN
  - Georgios Smaragdakis, Sr. Researcher
  - Prof. Anja Feldmann
- EPFL
  - Manos Athanassoulis, Research Assistant
  - Prof. Anastasia Ailamaki
- GRIDPOCKET
  - Filip Gluszak, CEO
  - Michael Defoin-Platel, PhD, Senior Data Scientist

BigFoot is driven by real world applications

- Project philosophy
  - Top-down approach
    'Application' requirements and workloads drive our R&D
  - Real Data, real use-cases, real problems

- Two practical use cases
  - Data analytics in Cyber Security (Symantec)
  - Meter data management and analytics (GridPocket)

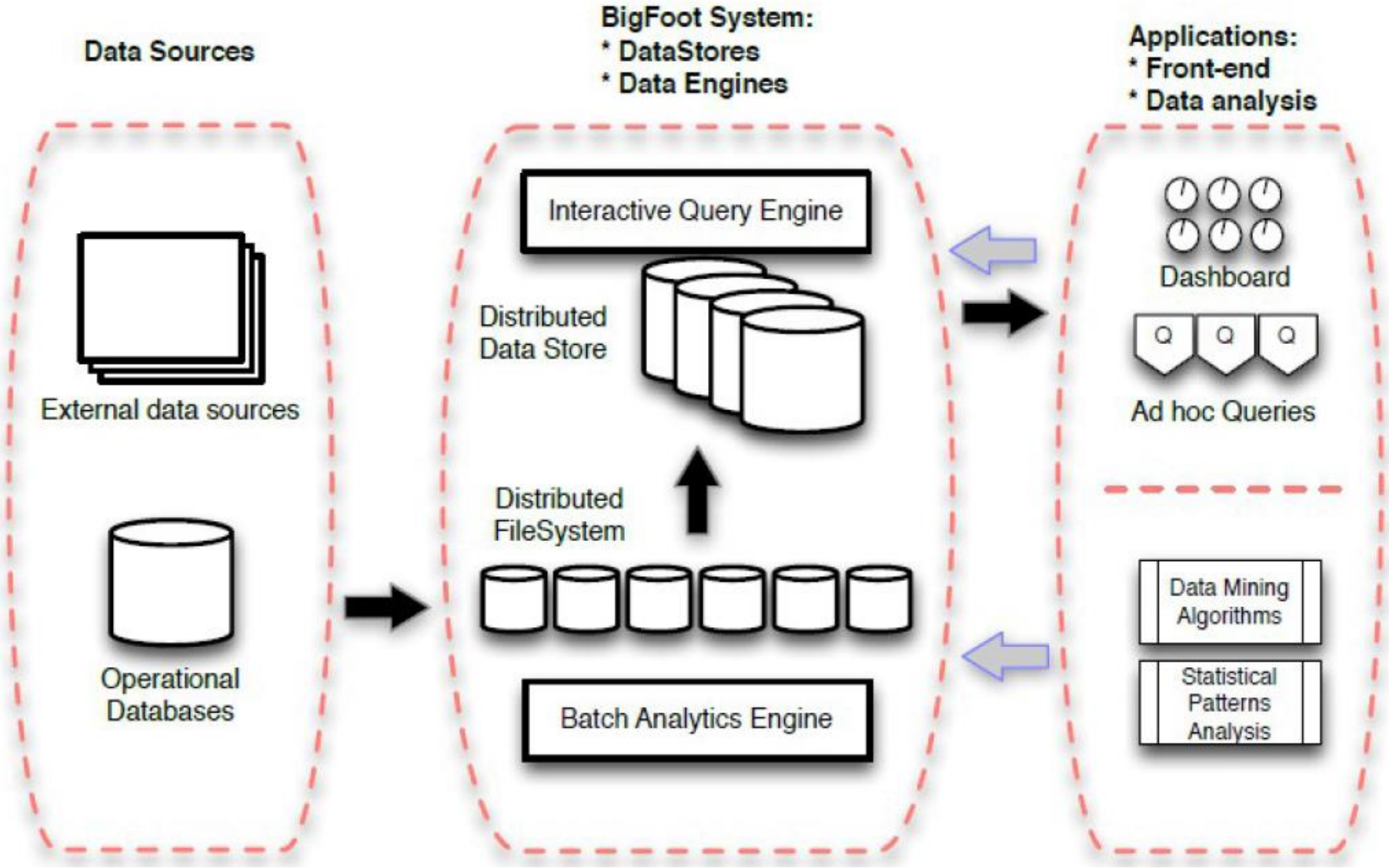
Courtesy of Prof. Pietro Michiardi, EURECOM

GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

What is BigFoot ?

- Hadoop on steroids !
  - Custom distribution + new components
  - Complete API compatibility with Hadoop
  - Modular design, use only what you need

- Features
  - Self-tuned deployments on private clouds
  - Several optimization over vanilla Hadoop
  - New interactive query engine


Data Sources
BigFoot System:
* DataStores
* Data Engines
Applications:
* Front-end
* Data analysis
External data sources
Operational Databases
Interactive Query Engine
Distributed Data Store
Distributed FileSystem
Batch Analytics Engine
Dashboard
Q
Q
Q
Ad hoc Queries
Data Mining Algorithms
Statistical Patterns Analysis

Benefits for BigFoot Users

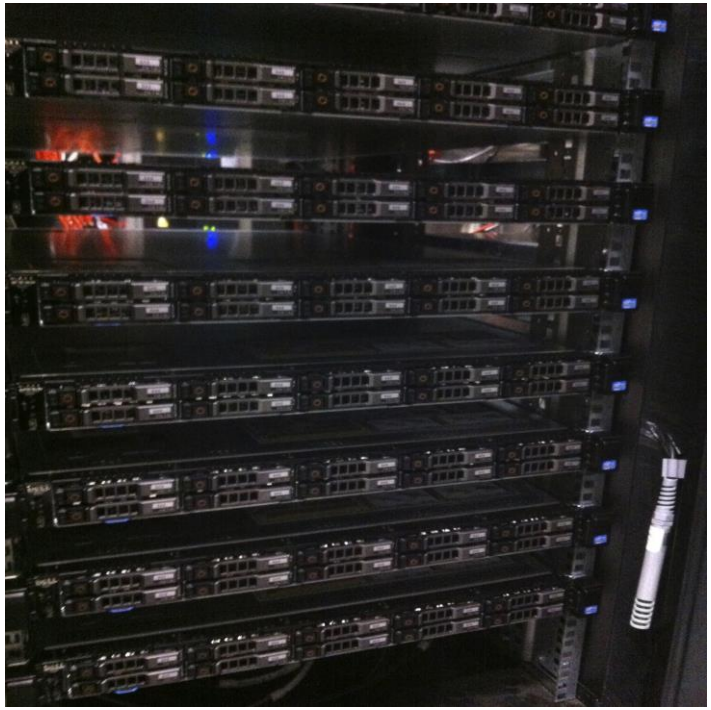
- For the Data Scientist
  - High-level language support
    - Optimizations at compile time
    - Machine learning and time-series analysis libraries
  - Data interaction made easy
    - New scheduler that avoids starvation
    - New interactive query engine

- For the IT
  - Hardware + data consolidation through virtualization
  - Performance enhancements to mitigate bottlenecks
  - Multi-site add-ons for geo-replication

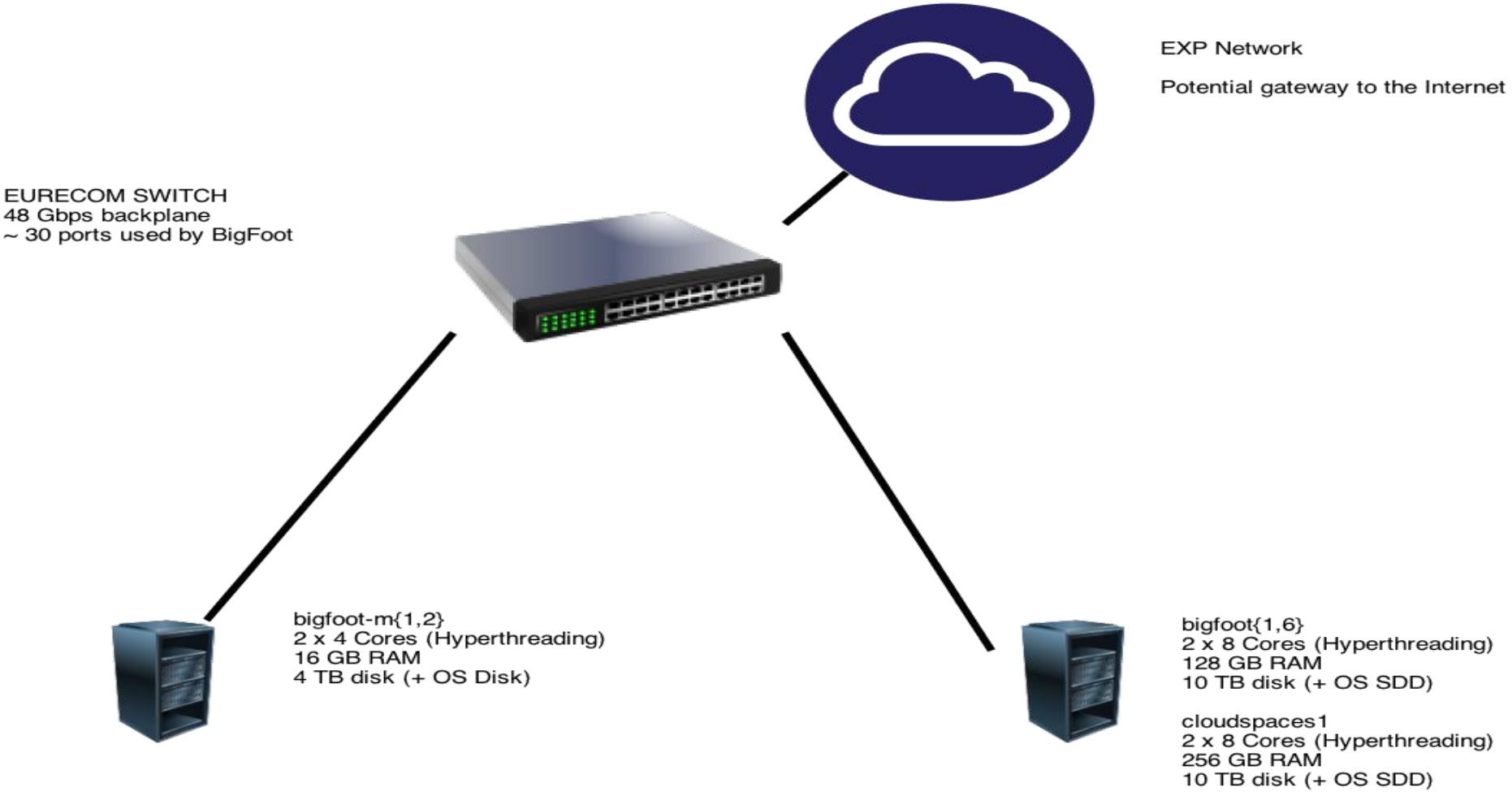
GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

The BigFoot cloud

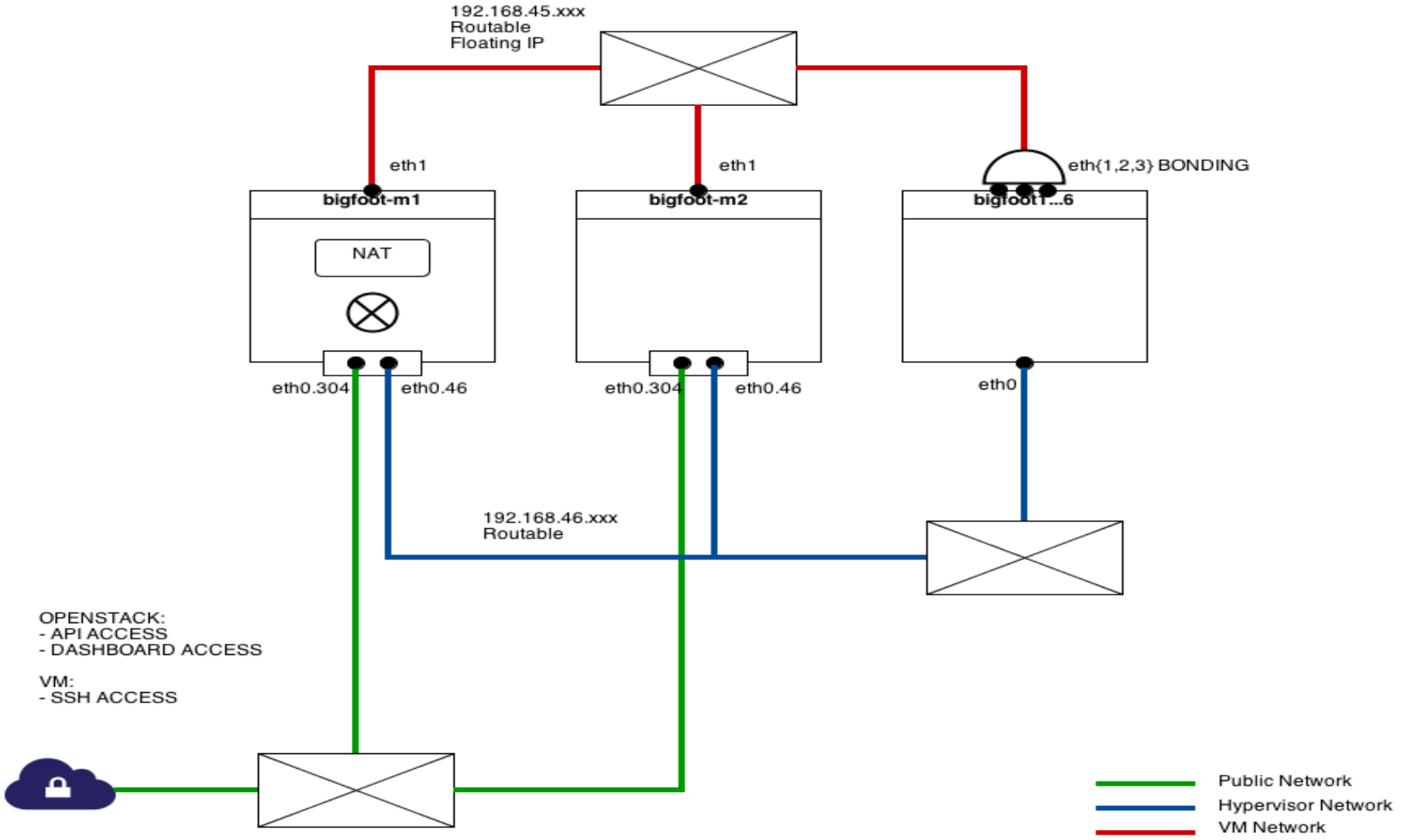
- Hardware specification
  - 224 + 32 cores
  - 1 TB RAM
  - 64 + 10 TB of storage
  - 1 Gbps Open(V)Switch interconnect
- Multi-sites
  - Sophia-Antipolis
  - Lausanne
  - Berlin



PHYSICAL PLAN

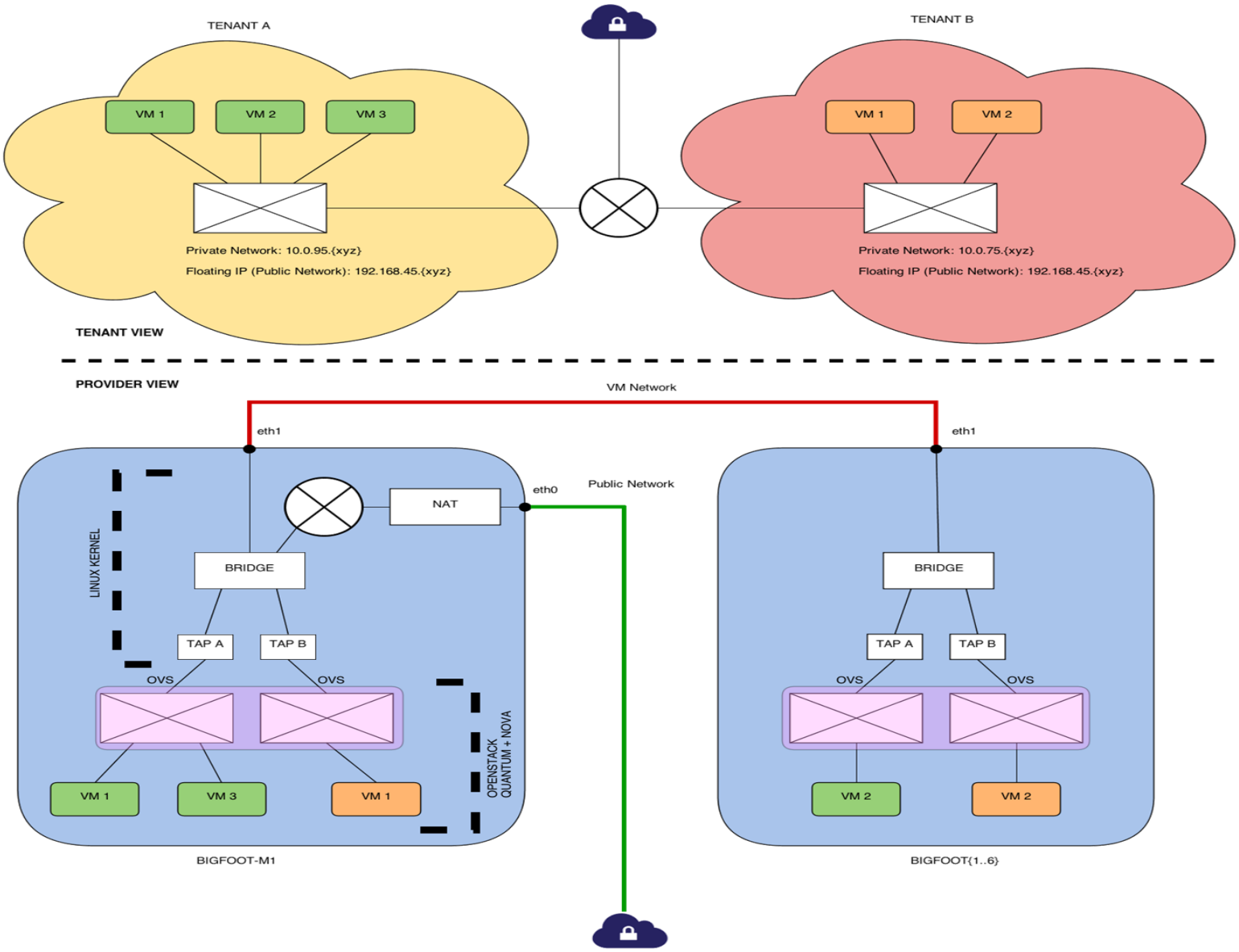


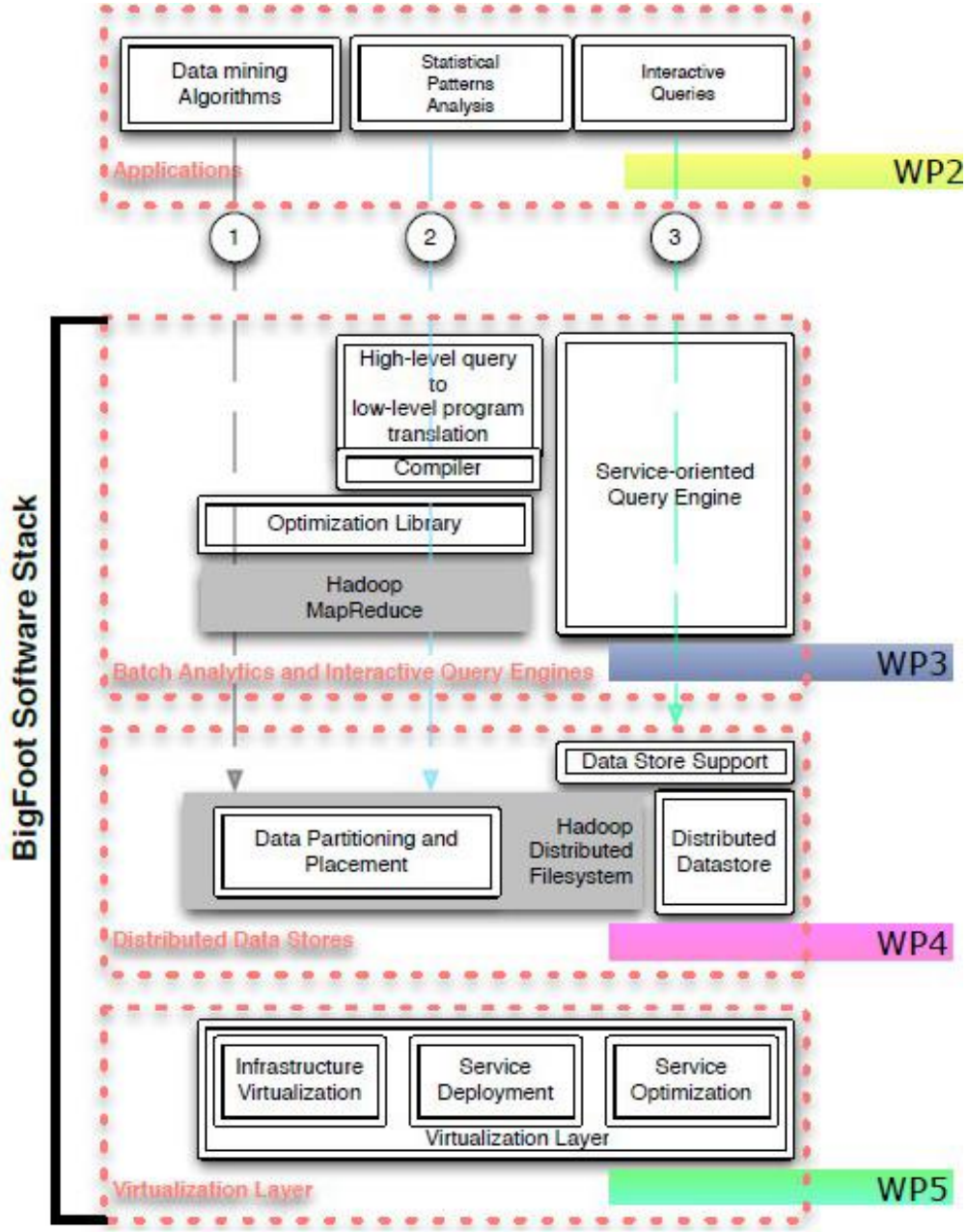
LOGICAL PLAN



Virtualization and SDN

- Cloud Management Systems: OpenStack
  - Automatic and self-tuned ‘application’ deployments
  - Amazon EMR-like PaaS
- Resource allocation
  - The curse of elasticity: Resource (De-)fragmentation
  - ‘Application’-aware VM Scheduling
- Performance evaluation and Networking
  - Distributed monitoring, measurements
  - The curse of data locality: Datacenter Networking, Flat Topologies and Virtual Networks

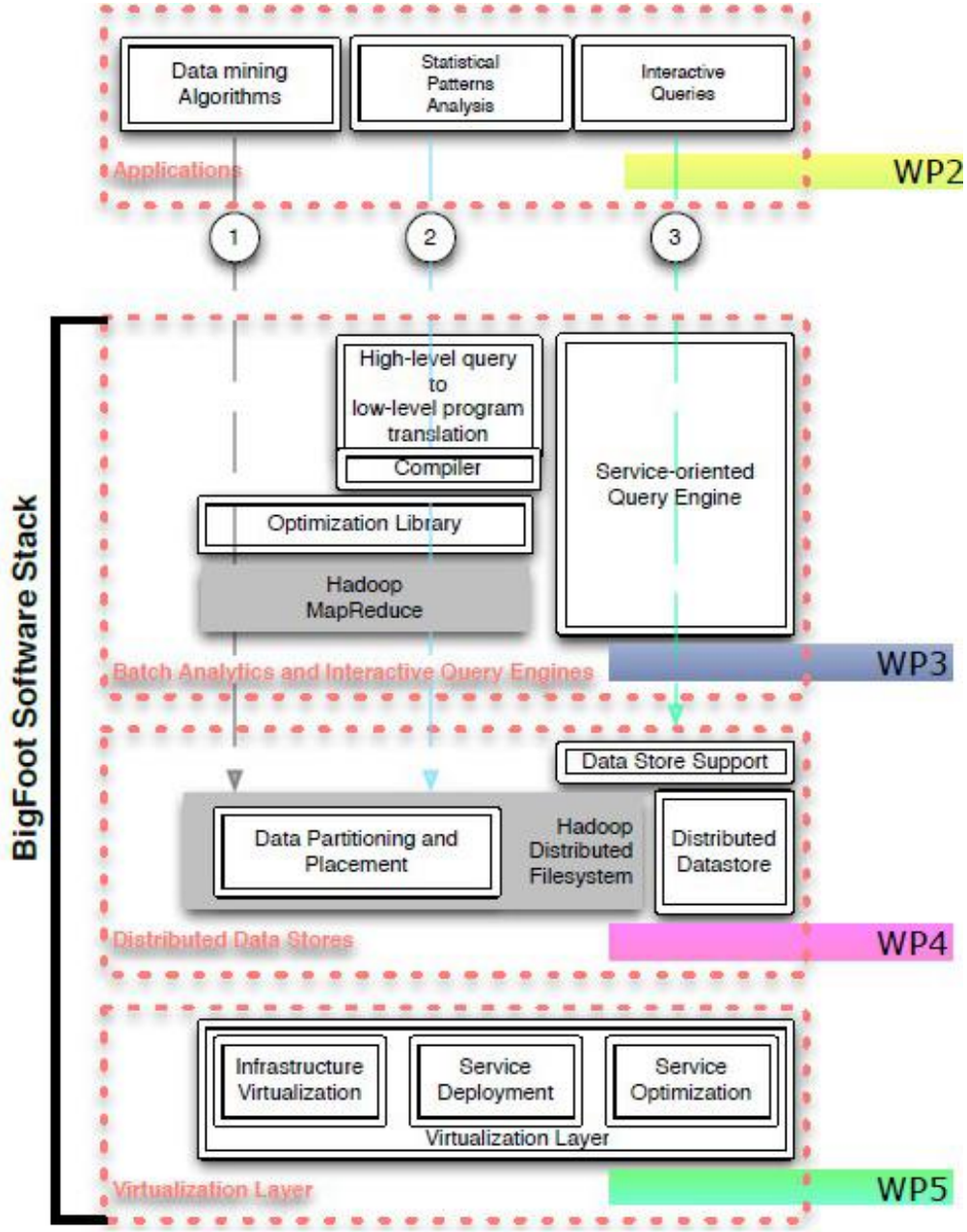

TENANT A
VM 1
VM 2
VM 3
Private Network: 10.0.95.{xyz}
Floating IP (Public Network): 192.168.45.{xyz}
TENANT VIEW
TENANT B
VM 1
VM 2
Private Network: 10.0.75.{xyz}
Floating IP (Public Network): 192.168.45.{xyz}
PROVIDER VIEW
VM Network
eth1
eth1
eth0
Public Network
NAT
LINUX KERNEL
BRIDGE
TAP A
TAP B
OVS
OVS
VM 1
VM 3
VM 1
OPENSTACK QUANTUM + NOVA
BIGFOOT-M1
BRIDGE
TAP A
TAP B
OVS
OVS
VM 2
VM 2
BIGFOOT{1..6}


Data mining Algorithms
Statistical Patterns Analysis
Interactive Queries
Applications
WP2
1
2
3
BigFoot Software Stack
High-level query to low-level program translation
Compiler
Optimization Library
Hadoop MapReduce
Service-oriented Query Engine
Batch Analytics and Interactive Query Engines
WP3
Data Store Support
Data Partitioning and Placement
Hadoop Distributed Filesystem
Distributed Datastore
Distributed Data Stores
WP4
Infrastructure Virtualization
Service Deployment
Service Optimization
Virtualization Layer
Virtualization Layer
WP5

BigFoot Workpackages

- WP5: Virtualization
  - Hadoop is not designed for Virtual Environments
    - HDFS is not elastic (remove/add nodes)
    - Mapping virtual vs real topology (cf VMWare)
    - IO limitation : virtual switchs, virtual disks

→ **resource allocation to optimize VM placement for Hadoop**

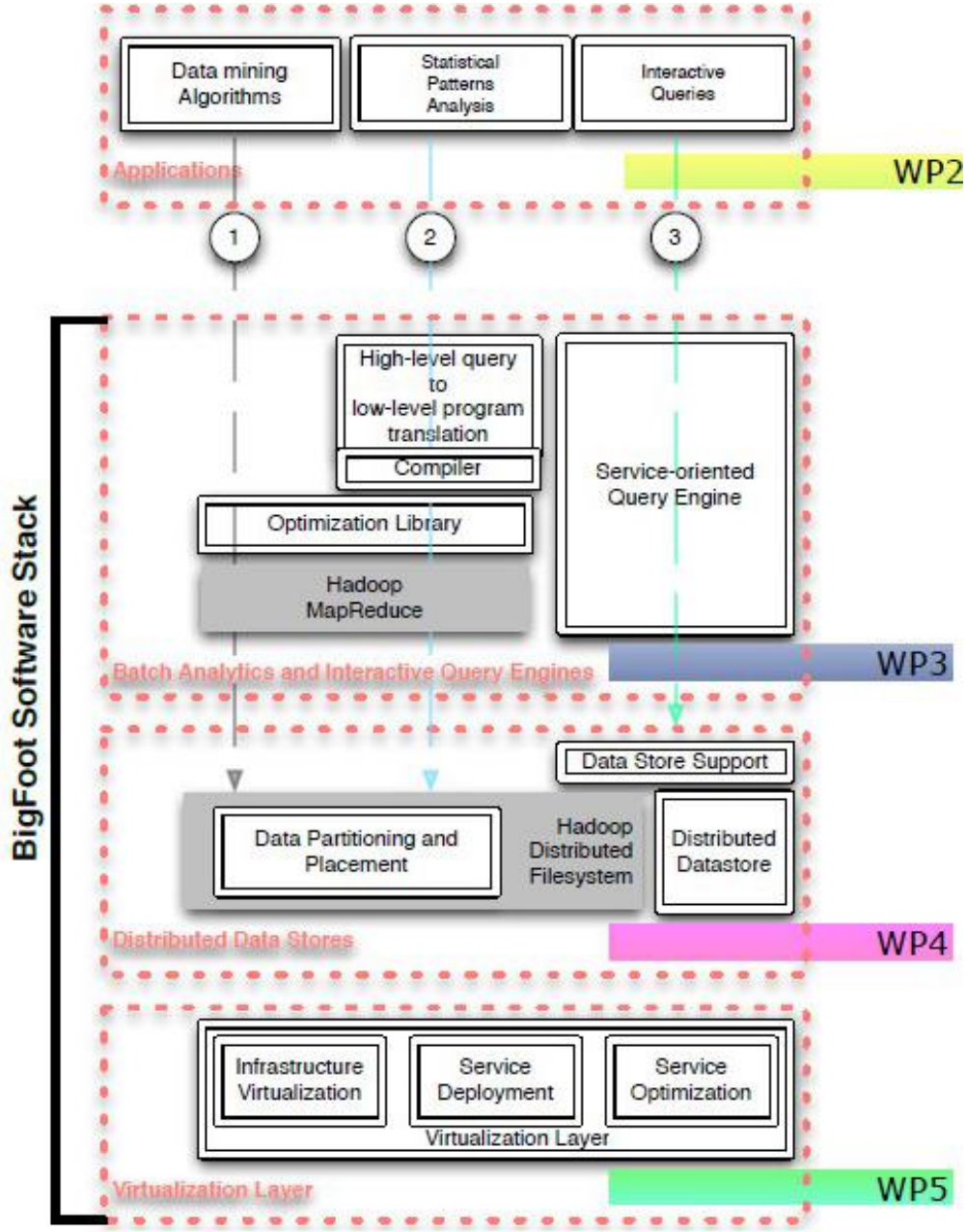

Data mining Algorithms
Statistical Patterns Analysis
Interactive Queries
Applications
WP2
1
2
3
BigFoot Software Stack
High-level query to low-level program translation
Compiler
Optimization Library
Hadoop MapReduce
Service-oriented Query Engine
Batch Analytics and Interactive Query Engines
WP3
Data Store Support
Data Partitioning and Placement
Hadoop Distributed Filesystem
Distributed Datastore
Distributed Data Stores
WP4
Infrastructure Virtualization
Service Deployment
Service Optimization
Virtualization Layer
Virtualization Layer
WP5

BigFoot Workpackages

- WP4: Distributed Data store
  - Data storage and distribution are not optimized for the workloads
  - Coherence of the data across nodes / sites ?

→ **NoDB: adaptive indexing to provide efficient access to raw data**

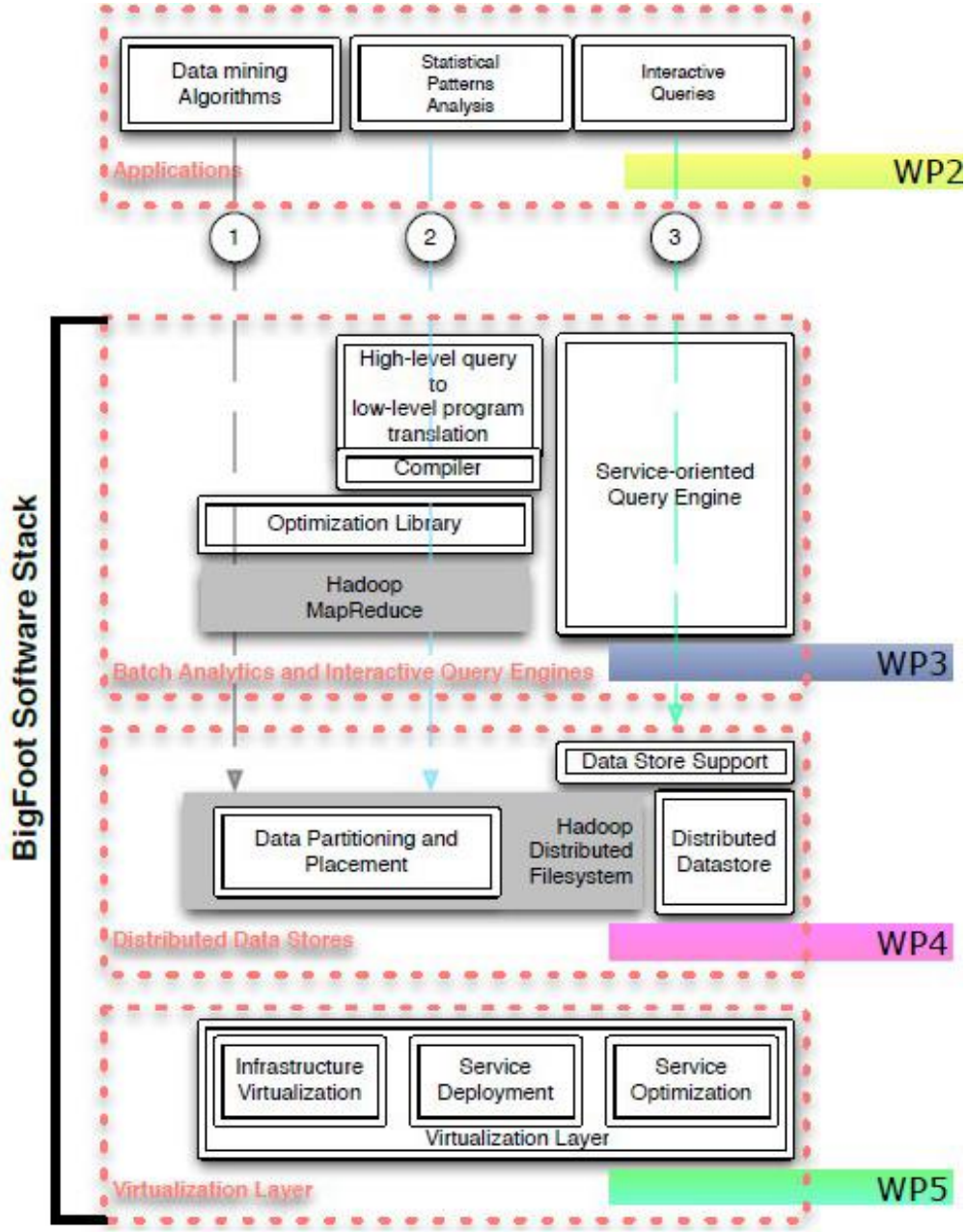
→ **data partitioning and placement optimizations**


Data mining Algorithms
Statistical Patterns Analysis
Interactive Queries
Applications
WP2
1
2
3
BigFoot Software Stack
High-level query to low-level program translation
Compiler
Optimization Library
Hadoop MapReduce
Service-oriented Query Engine
Batch Analytics and Interactive Query Engines
WP3
Data Store Support
Data Partitioning and Placement
Hadoop Distributed Filesystem
Distributed Datastore
Distributed Data Stores
WP4
Infrastructure Virtualization
Service Deployment
Service Optimization
Virtualization Layer
Virtualization Layer
WP5

BigFoot Workpackages

- WP3: Analytics and Interactive Query Engines
  - Scheduling is crucial (at FB 180,000/w)
  - Schedulers are not fair in terms of Sojourn time
  - Work is not shared

→ **New scheduler to avoid starvation**
→ **New compiler (high level language to MapReduce jobs)**
→ **Work-sharing across jobs and workflows**


Data mining Algorithms
Statistical Patterns Analysis
Interactive Queries
Applications
WP2
1
2
3
BigFoot Software Stack
High-level query to low-level program translation
Compiler
Optimization Library
Hadoop MapReduce
Service-oriented Query Engine
Batch Analytics and Interactive Query Engines
WP3
Data Store Support
Data Partitioning and Placement
Hadoop Distributed Filesystem
Distributed Datastore
Distributed Data Stores
WP4
Infrastructure Virtualization
Service Deployment
Service Optimization
Virtualization Layer
Virtualization Layer
WP5

BigFoot Workpackages

- WP2: Applications
  - Use cases driven
  - Need to port Cyber-security and Smart Grid analytics over to MapReduce

→ **Hadoop time series library**

→ **Scalable machine learning library but not like Mahout !**

→ **Simple Pig UDF + Ensemble learning**

BigFoot Impact

- Research
  - Advance state-of-the-art in Big Data technologies and Large-scale Data Mining
- Open Source Software (Apache License)
  - Contribute to the Hadoop ecosystem
  - Contribute to the OpenStack cloud software
- Big Data Market
  - Key enabler for new Big Data solutions
  - Data Scientist friendly approach

CONCLUSIONS

Conclusions

- In France/EU, data volumes are still small but will grow
- Value will come from crossing multiple data sources and efficient mining in line with business needs
- Ok, but what are the business needs ?
- Interested in BigFoot, join the IAB !
- Hadoop and real-time / event processing
- Multiple scale distributed computing
- Cost (money/energy) of the energy efficiency services ?

GRIDPOCKET
PERSONAL SMARTGRID SOLUTIONS

Thanks