Fast data meets HPC at TERALAB

Pierre PLEVEN, Direction de l'innovation, INSTITUT MINES TELECOM

INSTITUT Mines-Télécom

GROUPE-GENES

TERALAB DATA SCIENCE FOR EUROPE

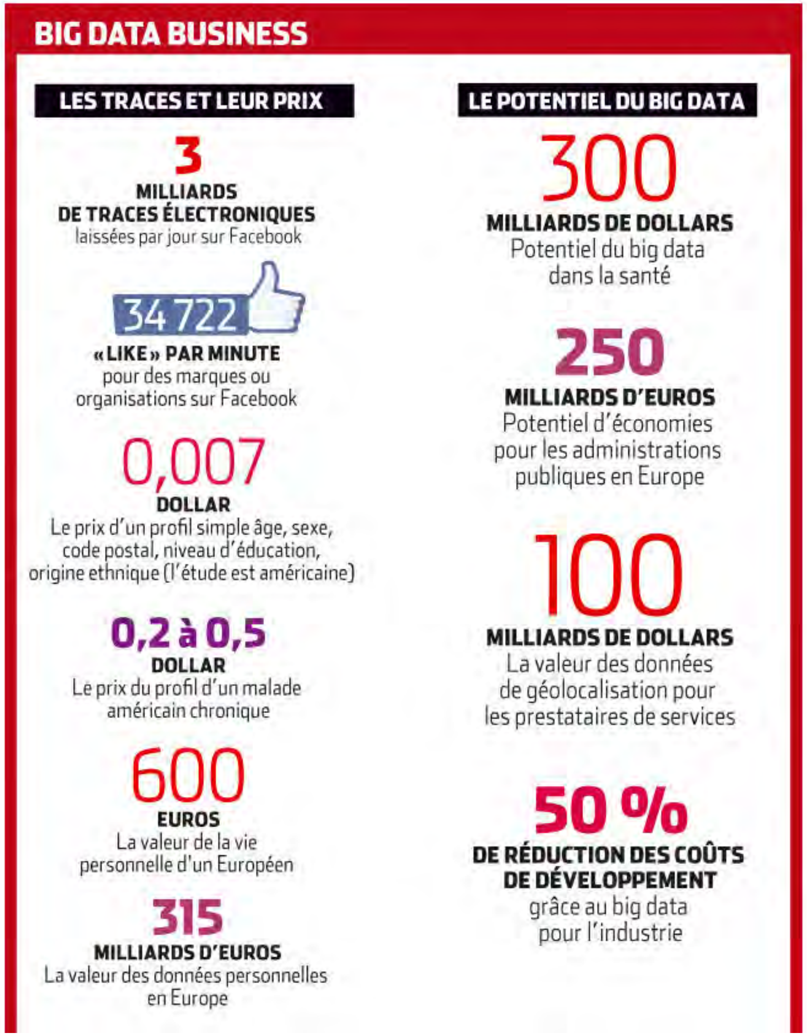
INVESTISSEMENTS D'AVENIR

cap·digital

Leibniz at the time of the invention the Printing Press :

“This huge mass of books that we can never read will bring us back to Barbarism or forward to Culture”

Michel Serres, on France Info makes the parallel with Big Data phenomenon



BIG DATA BUSINESS

LES TRACES ET LEUR PRIX

3
MILLIARDS
DE TRACES ÉLECTRONIQUES
laissées par jour sur Facebook

34 722
« LIKE » PAR MINUTE
pour des marques ou
organisations sur Facebook

0,007
DOLLAR
Le prix d'un profil simple âge, sexe,
code postal, niveau d'éducation,
origine ethnique (l'étude est américaine)

0,2 à 0,5
DOLLAR
Le prix du profil d'un malade
américain chronique

600
EUROS
La valeur de la vie
personnelle d'un Européen

315
MILLIARDS D'EUROS
La valeur des données personnelles
en Europe

LE POTENTIEL DU BIG DATA

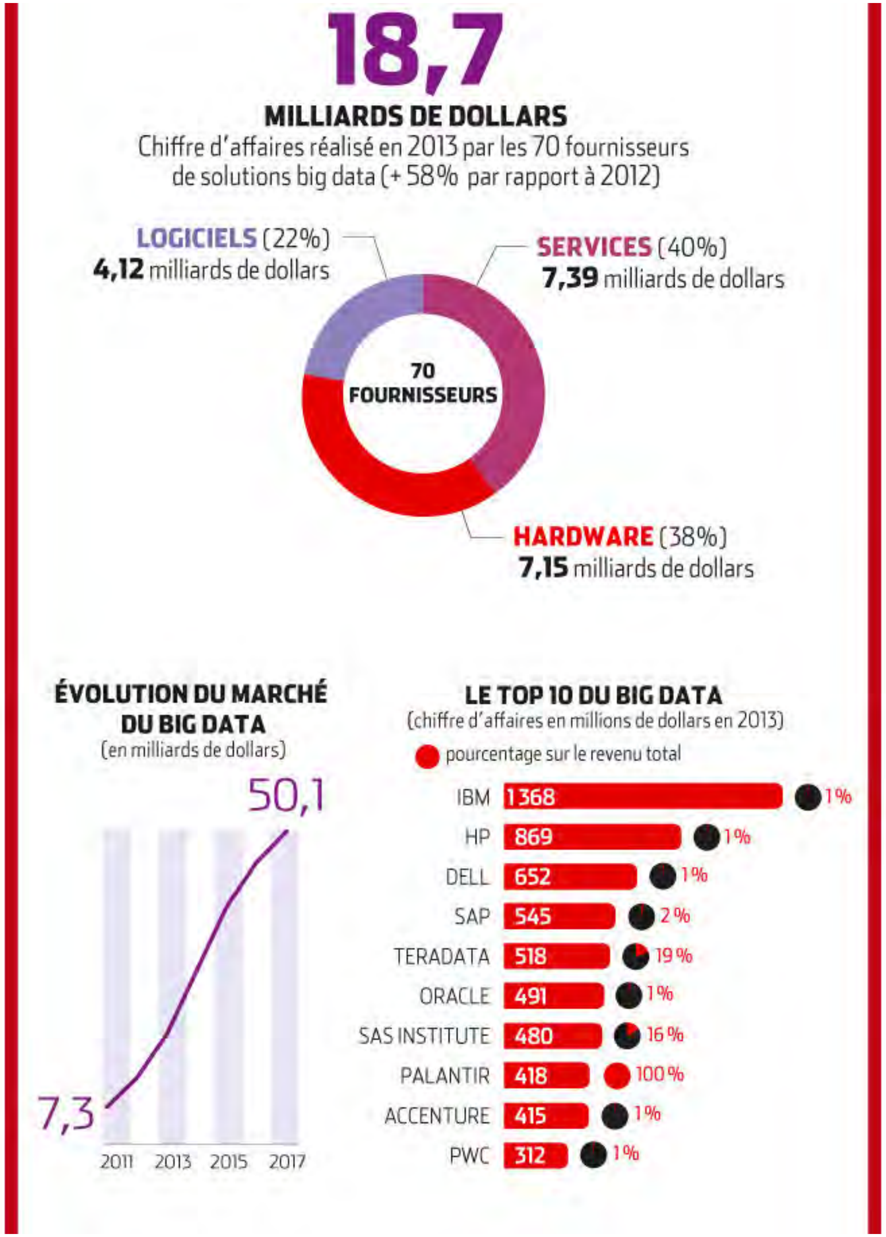
300
MILLIARDS DE DOLLARS
Potentiel du big data
dans la santé

250
MILLIARDS D'EUROS
Potentiel d'économies
pour les administrations
publiques en Europe

100
MILLIARDS DE DOLLARS
La valeur des données
de géolocalisation pour
les prestataires de services

50 %
DE RÉDUCTION DES COÛTS
DE DÉVELOPPEMENT
grâce au big data
pour l'industrie

Source : L'Usine Nouvelle May 2014



18,7

MILLIARDS DE DOLLARS

Chiffre d'affaires réalisé en 2013 par les 70 fournisseurs de solutions big data (+ 58 % par rapport à 2012)

ÉVOLUTION DU MARCHÉ DU BIG DATA

(en milliards de dollars)

LE TOP 10 DU BIG DATA

(chiffre d'affaires en millions de dollars en 2013)

● pourcentage sur le revenu total

IBM	1368	1 %
HP	869	1 %
DELL	652	1 %
SAP	545	2 %
TERADATA	518	19 %
ORACLE	491	1 %
SAS INSTITUTE	480	16 %
PALANTIR	418	100 %
ACCENTURE	415	1 %
PWC	312	1 %

Source : L'U[redacted]ne Nouvelle May 2014

Why data driven innovation now ?

- From the years 2000, important disruptions :
  - Surge of vast volume of data : Web , social networks, IOT
  - Software limits are pushed back by the NoSQL and advance in analytics, open source playing a major role
  - Drop of hardware cost de Teraflops process and of Terabyte storage , once Disk now Memory
- These advances allow economic processing of massive data in Volume Variety , Velocity with Value creation potential

Value Creation : which economic sectors?

Pioneers industries : because the ROI was proven ; ie

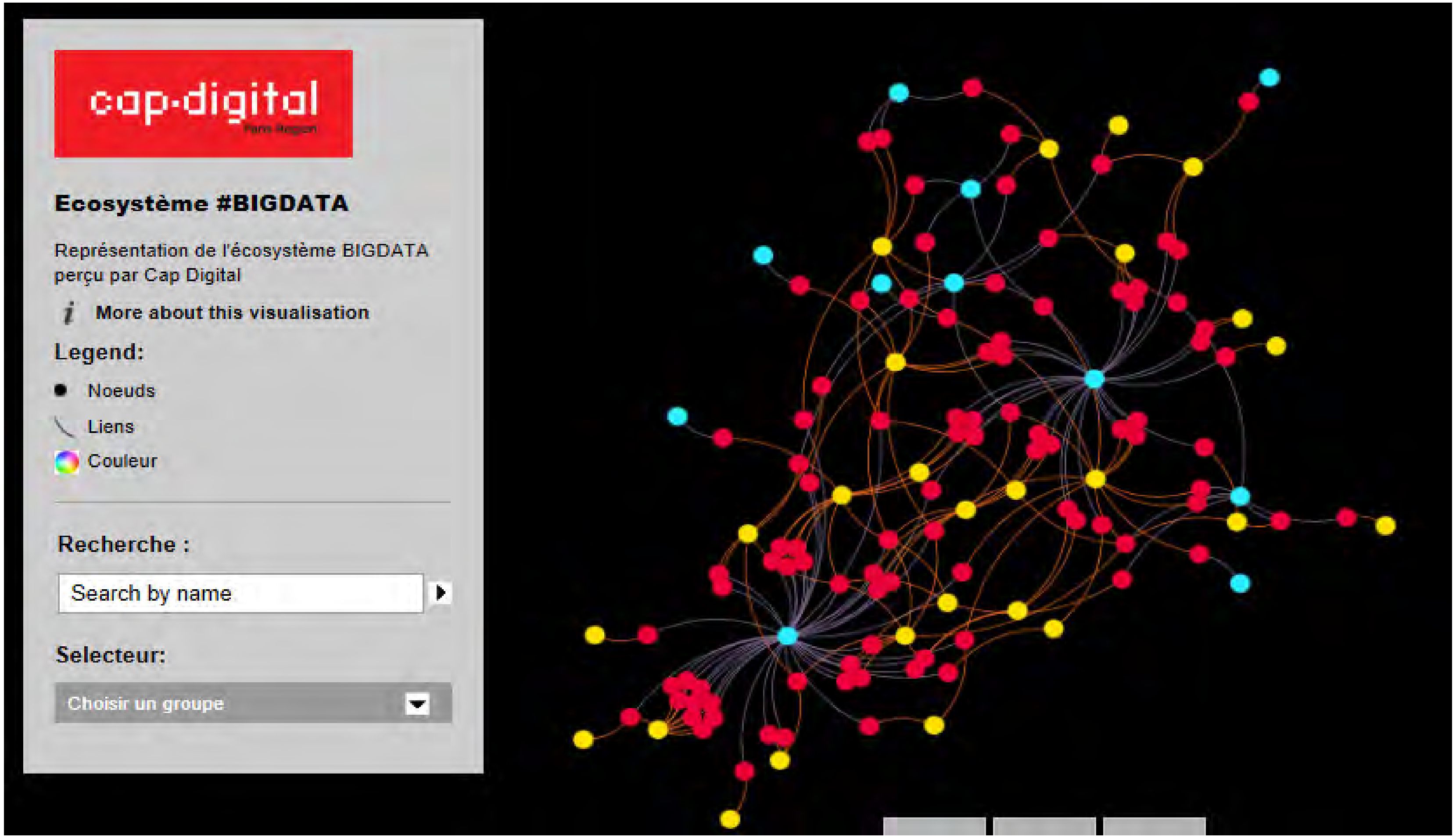
- « Pure web players » : e commerce , search , social networks…
- Banking & Insurance: Risks , High Frequency trading
- Oil&Gas : geophysical research

New Entrants : lower ROI thresholds , partly thanks to open source

- “Telco Operators : Software Defined Networks …
- Energy Operators : Smart Grids…
- Manufacturing Industries: Predictive maintenance , Manufacturing 4.0
- Smart City: Transport, City Planning , Crowd sourcing, Safety
- Agriculture & Environment : Micro agriculture “furrow to furrow”
- Brick & Mortar” Retail: Store to Web, Web to store
- Health : Epidemiology, Risks, (Genome, ie Adam project)
- Administrations: State and Regions

Example :Cap Digital Cluster , a vibrant SME community:150 Startups

http://connect.capdigital.com/wp-content/uploads/2014/ecosysteme-bigdata/



Ecosystème #BIGDATA

Représentation de l'écosystème BIGDATA perçu par Cap Digital

More about this visualisation

Legend:

● Noeuds

Liens

Couleur

Recherche :

Search by name

Selecteur:

Choisir un groupe

cap·digital Paris Region

LE PÔLE DE COMPETITIVITÉ
DES SERVICES ET CONTENUS NUMÉRIQUES

Semantics / NLP

KWAGA

{OWI} understand to succeed

viavoo smarter feedback

TEMIS

MONDECA

SyLabs

synomia

VIASEMA

troo click

armadillo editeur de technologie et logiciels

API / interopérabilité

Joul

webshell

APISpark by Restlet

Dataviz

linkurious visualize graph data easily

Dataveyes

qunb

bluenod

ARTICQUE Solutions

Data management platform

SCALED RISK

OpenDataSoft

squid SOLUTIONS

MULTIPOSTING

Dataiku

DASH

Machine learning

Dataiku

tinyclues

JÉRÔME

:snips

eXenSa

DATA&DATA CONSULTING

PredicSis

precogs

RTB - RTProcessing

alephd We help publishers

capptain PILOT YOUR APPS

Tradelab Trading Desk

Flaminem

adomik

Analytics

kxen

ISTHMA

PredictiveDB confidential data analysis

ProbaYes Mastering Uncertainty

estimages

Search

SINEQUA

Qwant

exalead connect the dots

pertimm

OpenSearchServer Search Engine API

antidot

ETL

talend open data solutions

hurence

Virtualisation

Virtual Scale

isi EXPERT

semsoft

E-réputation (écoute)

linkfluence synthesio
bluenod mention

Smart cities

:snips Joul

Relation client

davi {OWI} understand to succeed viavoo smarter feedback

Com / SEO

G⁴ interactive synomia qunb quantities & numbers
focusmatic ACCURATE DIGITAL REACH wisemetrics

RH

MULTIPOSTING
Meteojob mieux postuler, mieux recruter

Publishing

alephd We help publishers adomik advertising intelligence for publishers sanspapier.com

Banque / finance

SCALED RISK troo click arx CORPORATE FINANCE

Immobilier

PlaCE de l'IMMOBILIER PRO
homengo KelQuartier Trouver le quartier où habiter

Media

SKYROCK Free People Network mfg labs KILA SYSTEMS From Big Data to Smart Data
ALLOCINE.COM blitzr BETA

Marketing

tinyclues squid SOLUTIONS capptain PILOT YOUR APPS
kxen PredictiveDB confidential data analysis Tradelab Trading Desk
Captain DASH Dataiku

Logistique

precogs

Tourisme

SyJLabs

Prospection commerciale

DIGIT m Get more profitable leads! Corporama IMPROVE YOUR BUSINESS IKO SYSTEM
C-RADAR salezeo THE SALES COMMUNITY Get+ by webleads

Health

KAPPA SANTÉ VIDAL

Retail / e-commerce

Smart Flows
Graphinium LEVERAGE SOCIAL NETWORKS openPricer JÉRÔME

INSTITUT Mines-Télécom GROUPE-GENES TERALAB DATA SCIENCE FOR EUROPE INVESTISSEMENTS D'AVENIR cap·digital

Instruments for innovation: AMI Challenges Big Data

Under future investments program (development of the digital economy - heart of digital sector), State launches a call for expression of interest on the "Big Data Challenges"

- **Community building activities**
  - Industrial Data owners and potential use case
  - DataScientists: SME, Startups, Students , Individuals
  - Analytics Software providers and Secure cloud platforms
- **Preparing and Running the Challenges**
  - Specify Detailed challenges conditions with Data Owner
  - Scout for matching DataScientists talents
    - One to One, One to Few, One to Many
  - Run the challenges on secure platforms with analytics support

ANOTHER INSTRUMENT
BIG DATA PLATFORM FOR RESEARCH AND INNOVATION

TERALAB

Pierre Pleven

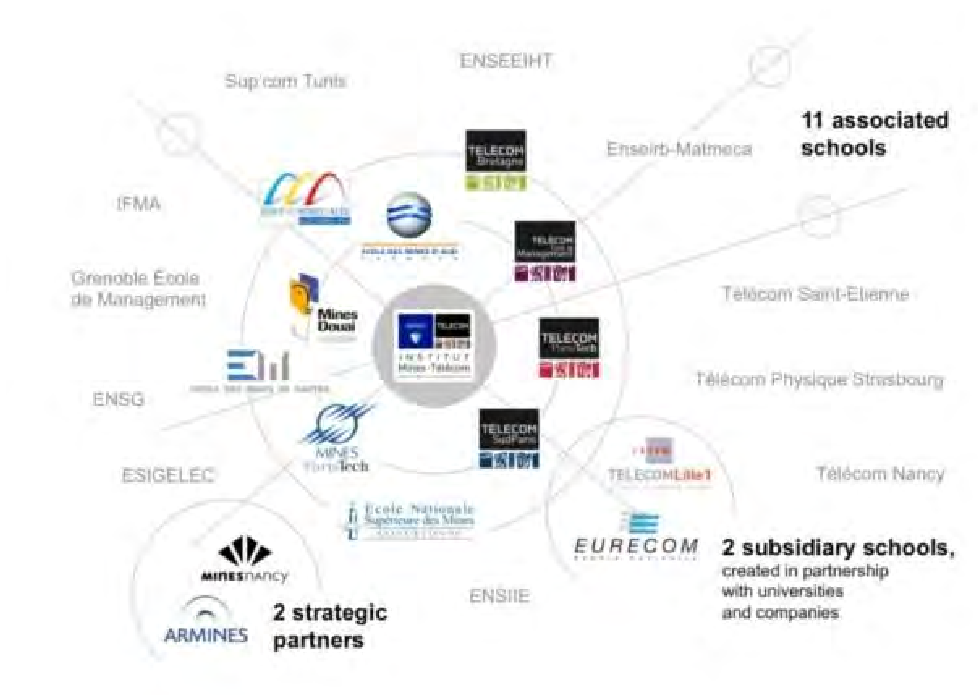
Direction de l'Innovation

Birth of the TeraLab project

- Call for projects “Cloud computing / Big Data” conducted by the French Government
- Proposal for the construction and operation of a Big Data platform,
  - For Innovation, Research and Education projects
  - Submitted by a consortium comprising
    - The IMT (Institut Mines-Télécom)
    - The GENES, particularly the CASD (secure remote access data center)
    - With INSEE partnership
- Project selected and launched
  - Budget of 5.7 M€
  - Over 5 years
  - Contract signed in December 2013

The ambition :
Accelerate « Data Science » innovation

Institut Mines-Télécom

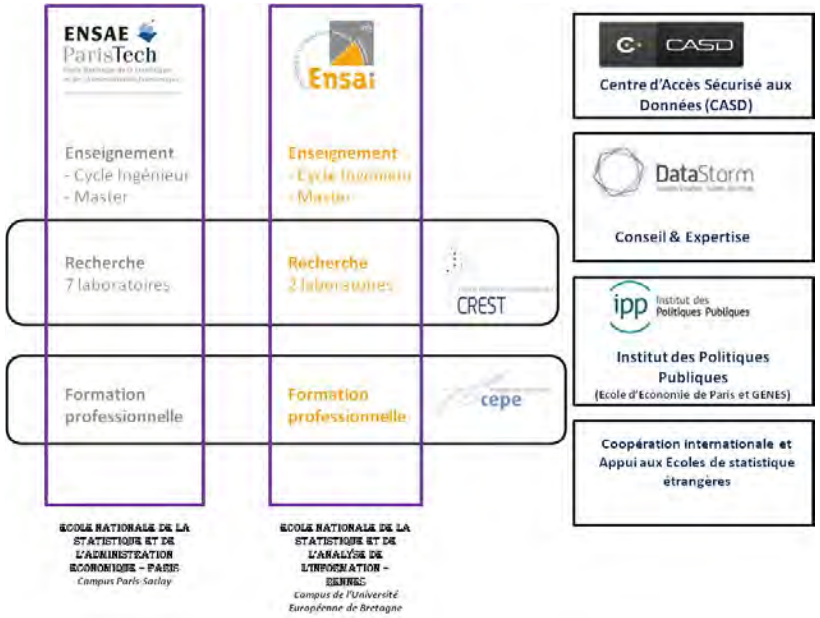


11 associated schools

2 subsidiary schools, created in partnership with universities and companies

2 strategic partners

GENES



ENSAE ParisTech	Ensai	
Enseignement - Cycle Ingénieur - Master	Enseignement - Cycle Ingénieur - Master	
Recherche 7 laboratoires	Recherche 2 laboratoires	CREST
Formation professionnelle	Formation professionnelle	cepe
ECOLE NATIONALE DE LA STATISTIQUE ET DE L'ADMINISTRATION ECONOMIQUE – PARIS *Campus Paris-Saclay*	ECOLE NATIONALE DE LA STATISTIQUE ET DE L'ANALYSE DE L'INFORMATION – RENNES *Campus de l'Université Européenne de Bretagne*	

Centre d'Accès Sécurisé aux Données (CASD)

Conseil & Expertise

Institut des Politiques Publiques
(Ecole d'Economie de Paris et GENES)

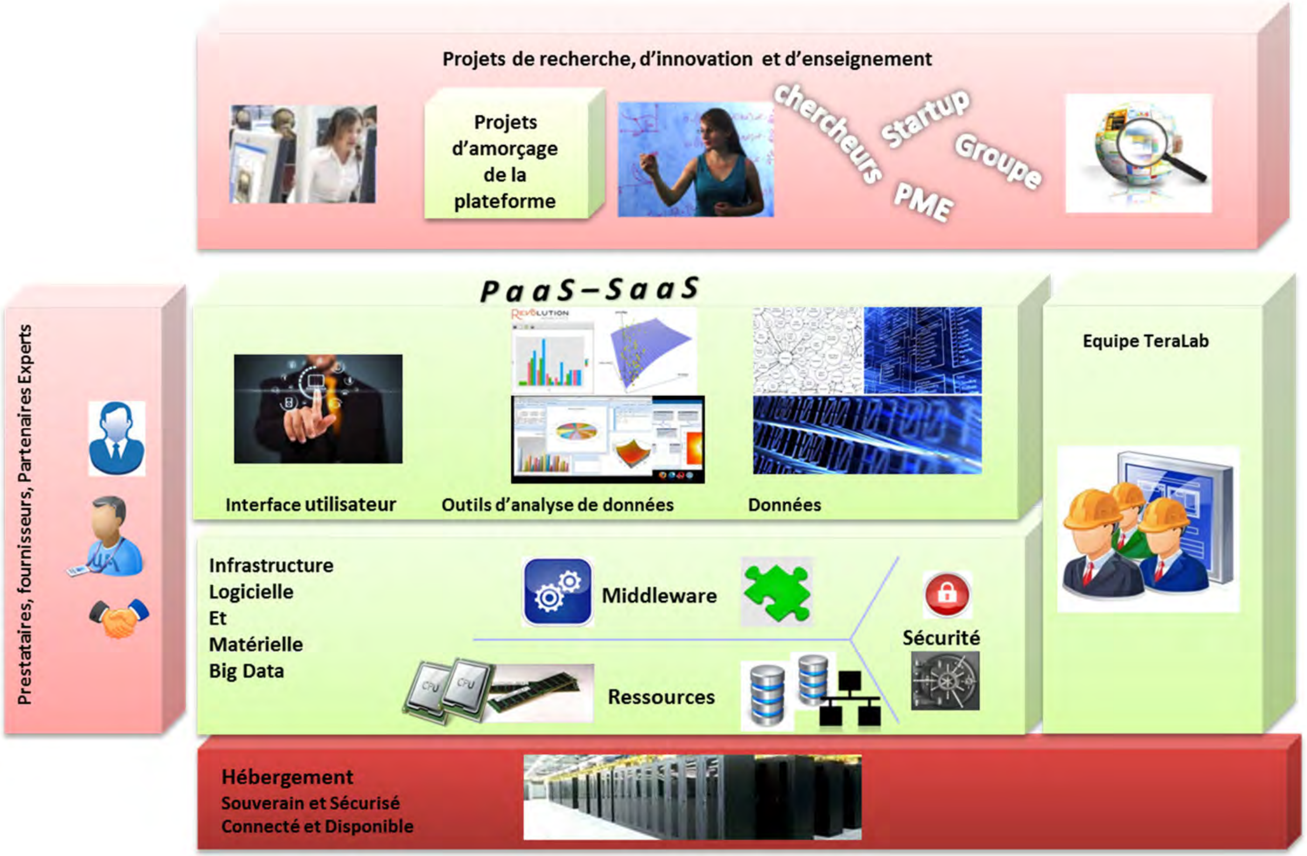
Coopération internationale et Appui aux Ecoles de statistique étrangères

The TeraLab platform

- A state-of-the-art technical infrastructure
  - Elastic distributed system + tera-memory server for "in -memory"
  - With unique security features
- A rich catalogue of software tools
  - Data storage (NoSQL)
  - Query, exploration, visualization (Pig, Hive, Mahout…)
  - Management and monitoring
- Data sets
  - Pre-installed (public data, open data…)
  - Brought by the projects, or acquired for them
- A dedicated team
  - 6 people
  - Platform configuration and operation
  - Project advisors

TERALAB

Technology ressources Center



Projets de recherche, d'innovation et d'enseignement
Projets d'amorçage de la plateforme
chercheurs
Startup
Groupe
PME
PaaS–SaaS
Interface utilisateur
Outils d'analyse de données
Données
Equipe TeraLab
Prestataires, fournisseurs, Partenaires Experts
Infrastructure Logicielle Et Matérielle Big Data
Middleware
Sécurité
Ressources
Hébergement
Souverain et Sécurisé
Connecté et Disponible

Opensource

Commercial

ACCESS PORTAL

Directories

Provisioning requests

Workspace management

DATA

Project waterproof Data

Shared Data

Public Data

ANALYTICS/ VISUALIZATION

Revolution Analytics

pentaho

KNIME

Tulip 4.0

mahout

sas

Gephi

python

StratoSphere Above the Clouds

DATA MANAGEMENT

SQL (Postgre, mySQL ..)

Hadoop

IMDB(Quartet ..)

INFRASTUCTURES

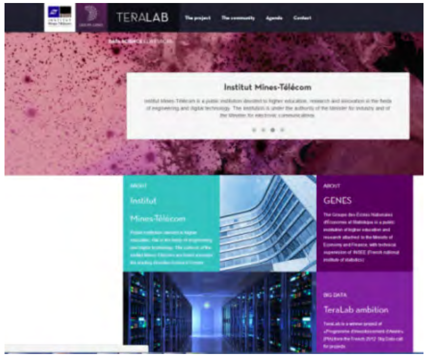
Private Cloud

Hybrid Cloud

Teramemory Server

TeraLab compartments

https://www.teralab-datascience.fr/fr/accueil



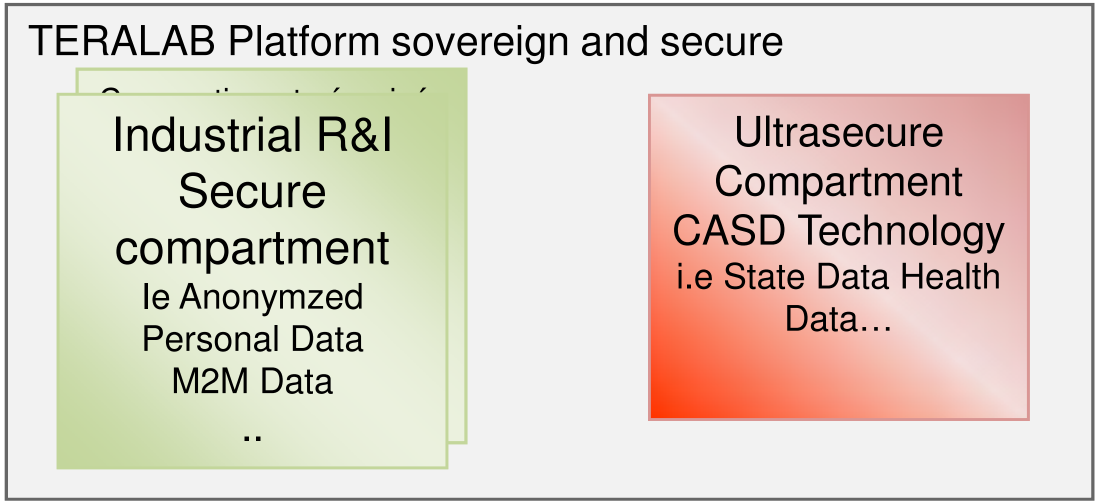
TERALAB The project The community Agenda Contact

Institut Mines-Télécom

Institut Mines-Télécom

GENES

TeraLab ambition



TERALAB Platform sovereign and secure

Industrial R&I Secure compartment

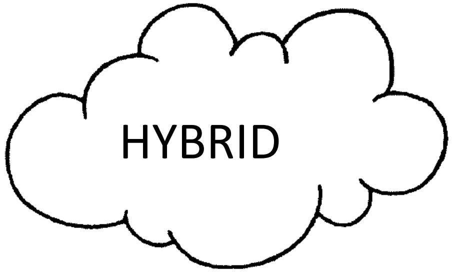
Ie Anonymzed Personal Data M2M Data

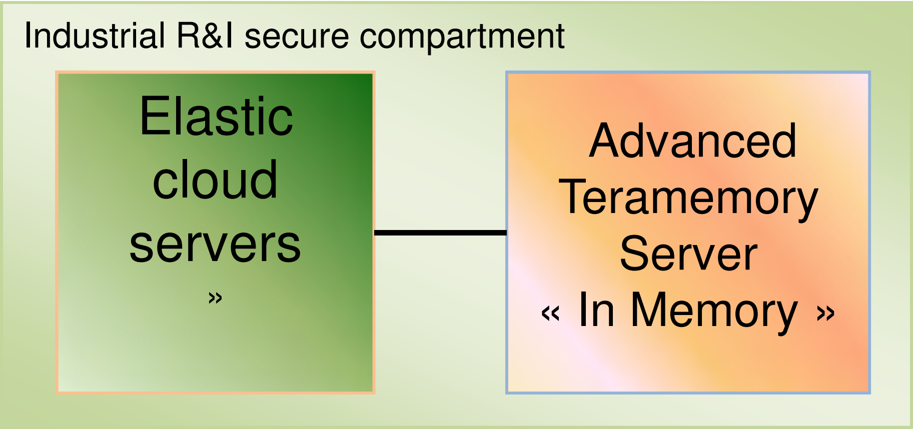
..

Ultrasecure Compartment CASD Technology

i.e State Data Health Data…

http://www.casd.eu/

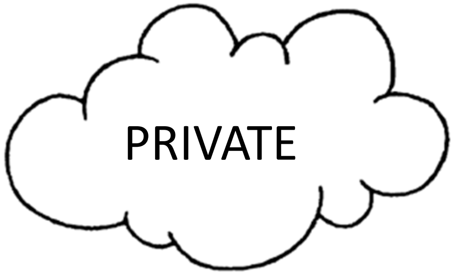




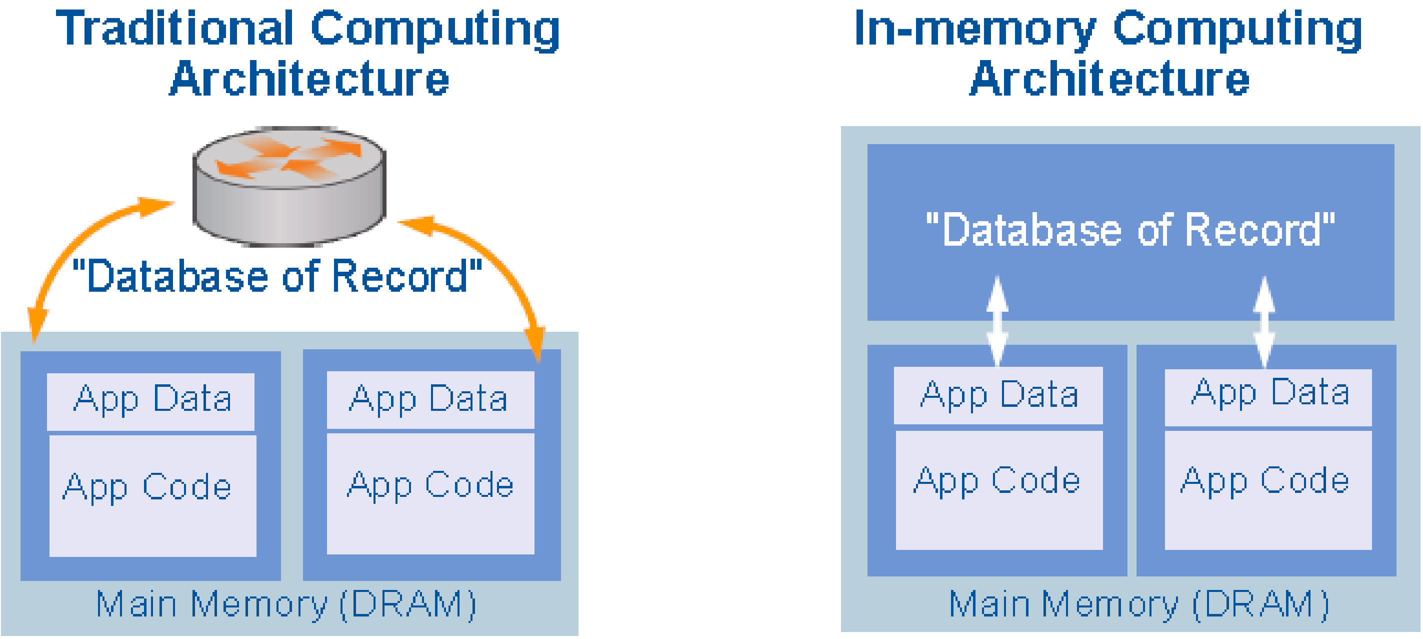
Industrial R&I secure compartment

Elastic cloud servers »

Advanced Teramemory Server « In Memory »



What is “in memory Computing”



Traditional Computing Architecture

"Database of Record"

App Data

App Code

App Data

App Code

Main Memory (DRAM)

In-memory Computing Architecture

"Database of Record"

App Data

App Code

App Data

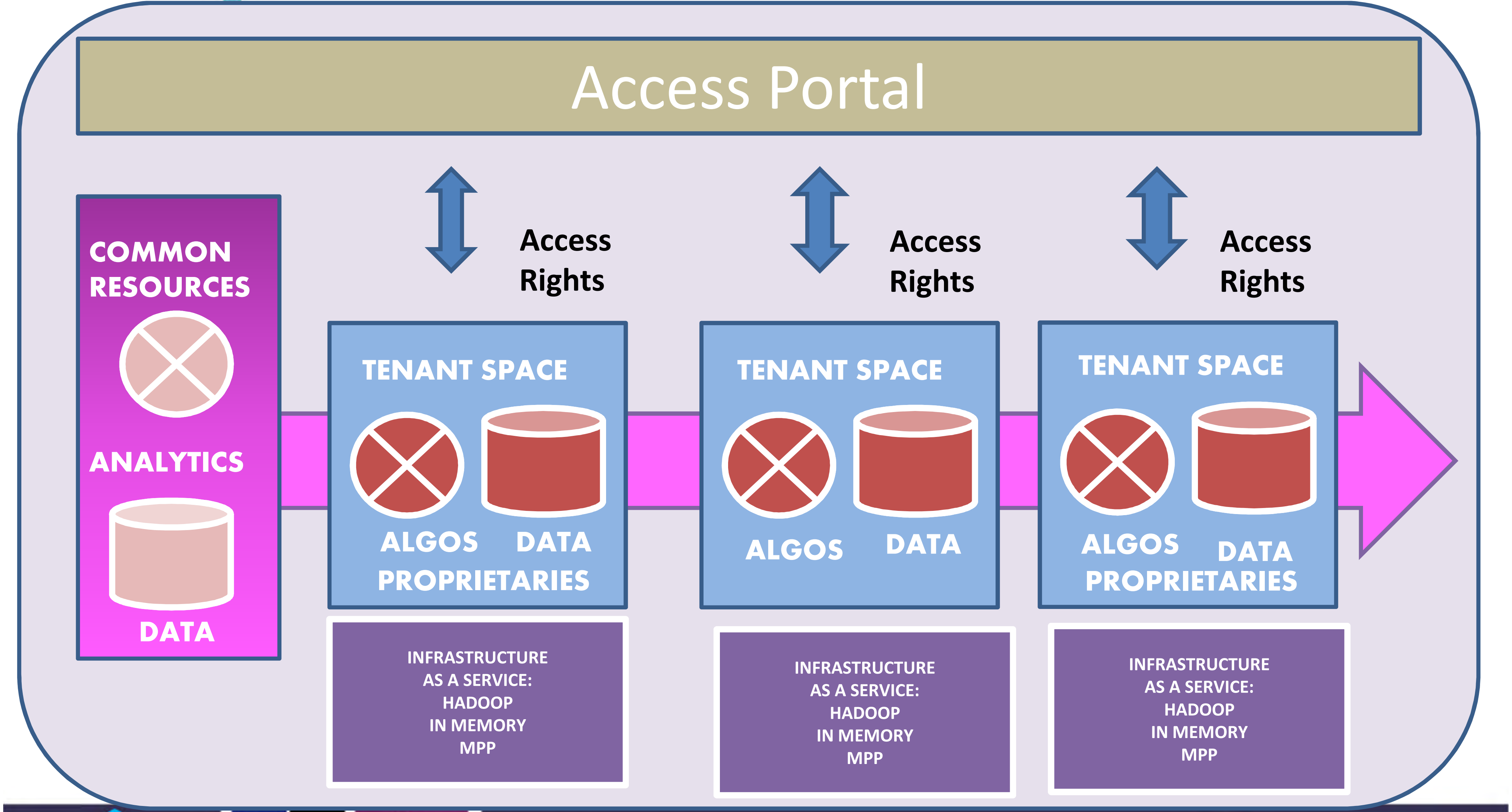
App Code

Main Memory (DRAM)

In-memory computing (IMC) is an architecture style where applications assume all the data required for processing are located in the main memory of their computing environment.

Gartner.

TeraLab Architecture



Industrial Uses Cases

CAP ITEA 2

INFORMATION TECHNOLOGY FOR EUROPEAN ADVANCEMENT

Virtual Metrology

Energy

Scenario 1: Statistical Business Environment in Wind Power Applications (Moventas)
Scenario 2: Wind Power Icing Atlas (VTT)
Scenario 3: Managed Service Provider Intelligence (Net Man)

Traffic (CCTV)

Scenario 1: Fast search of an object from a large volume of CCTV video data
Scenario 2: Searching for missing people
Scenario 3: Traffic Control and Reducing Traffic Accidents

Telecommunications

Scenario 1: Mobile Application Analytics
Scenario 2: Customer (Entity) Behavior Analytics

GeoIntelligence

Scenario: social Web monitoring for crisis management

Machine Manufacturing

Scenario 1: Reactive maintenance
Scenario 2: Pro-active and condition-based maintenance

Logistics

Scenario: logistics and marketing

Security

Scenario 1: Cyber Security Analysis using Network Traffic Classification on Huge Amounts of High Speed Network Traffic

DATAPIXEL QUALITY CONTROL ENGINEERING

Calibration Verification Engineering UNIMETRIK S.A.

VTT

moventas GEARED FOR NEW ENERGY

ETRI 한국전자통신연구원 Electronics and Telecommunications Research Institute

TURKCELL

ERICSSON

THALES

PICANOL GROUP

LA POSTE

Use cases in public statistics

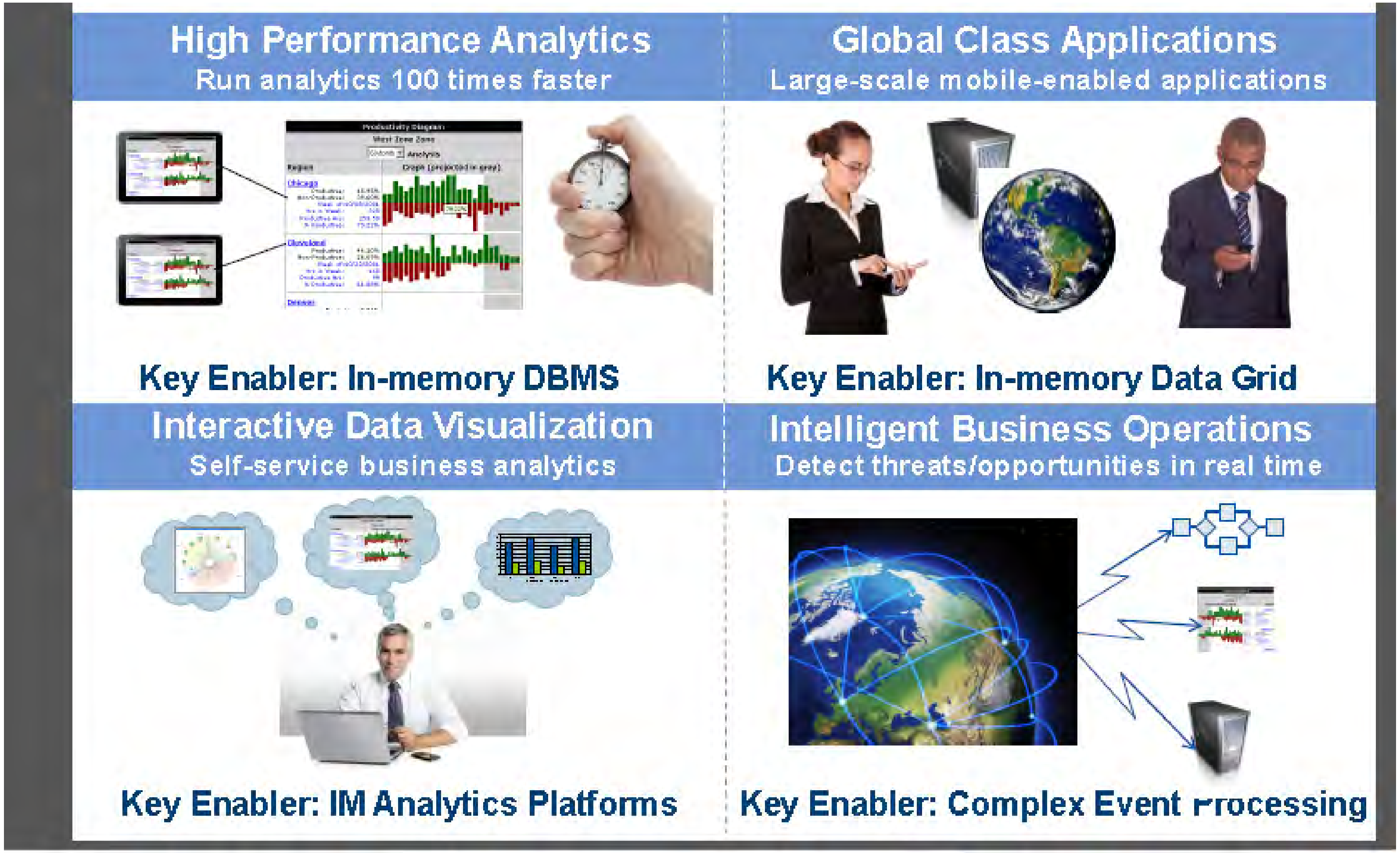
- A burning subject
  - The statistical community sees Big Data as a high-priority topic
  - A few experiences in some pioneer statistical institutes (Estonia,The Netherlands, etc.)
  - Several actions launched by international organizations (OECD, UNECE, Eurostat)
- How TeraLab fits in
  - Needs: methodological tests, exploration of data sources, process redesign
  - A presentation to the French official statistics system aroused much interest
  - Precise project on scanner data for the consumer price index
    - Currently a 7 terabytes relational database
  - Other ideas expressed
    - Telco data for tourism statistics
    - Web site log analysis
    - Next-generation social declarations

*Etude Deloitte 2013

Use case for health data

- French context
  - Everyone has a unique personal identifier (the NIR)
    - Allowing data matching
    - Longitudinal studies
    - **Using the NIR requires high confidentiality (organized by law)**
  - A central database with all the health services provided to every citizen
    - More than 1.2 billion records with more than a thousand variables
    - About 250 terabytes of data generated each year
    - Real time updates
- How TeraLab fits in
  - Able to meet the challenges
    - Huge volumes
    - Real-time analysis
  - While ensuring ultra-high security

Preferred UC for In memory: Gartner “low hanging fruits



High Performance Analytics

Run analytics 100 times faster

Key Enabler: In-memory DBMS

Global Class Applications

Large-scale mobile-enabled applications

Key Enabler: In-memory Data Grid

Interactive Data Visualization

Self-service business analytics

Key Enabler: IM Analytics Platforms

Intelligent Business Operations

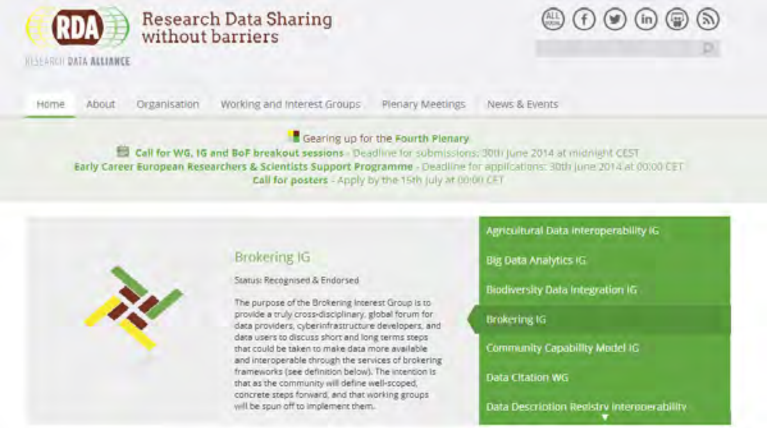
Detect threats/opportunities in real time

Key Enabler: Complex Event Processing

R&I USE CASES are WELCOME !

Pierre.pleven@mines-telecom.fr

EU Partnership in progress



Research Data Sharing without barriers

RESEARCH DATA ALLIANCE

Gearing up for the Fourth Plenary

Call for WG, IG and BoF breakout sessions - Deadline for submissions: 30th June 2014 at midnight CEST

Early Career European Researchers & Scientists Support Programme - Deadline for applications: 30th June 2014 at 00:00 CET

Call for posters - Apply by the 15th July at 00:00 CET

Brokering IG

Status: Recognised & Endorsed

The purpose of the Brokering Interest Group is to provide a truly cross-disciplinary, global forum for data providers, cyberinfrastructure developers, and data users to discuss short and long terms steps that could be taken to make data more available and interoperable through the services of brokering frameworks (see definition below). The intention is that as the community will define well-scoped, concrete steps forward, and that working groups will be spun off to implement them.

Agricultural Data Interoperability IG
Big Data Analytics IG
Biodiversity Data Integration IG
Brokering IG
Community Capability Model IG
Data Citation WG
Data Description Registry Interoperability

Commission Européenne

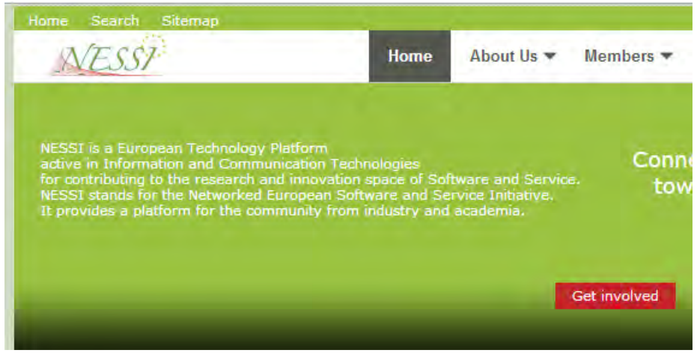
eurostat Votre clé d'accès à la Statistique européen

Commission européenne > Eurostat > Home

Accueil
Statistiques
Pu
Base de données
Calendriers des parutions
Tables préférées

Communiqués de presse

13.06.2014	L'er l'UE
13.06.2014	Exc bie



NESSI is a European Technology Platform
active in Information and Communication Technologies
for contributing to the research and innovation space of Software and Service.
NESSI stands for the Networked European Software and Service Initiative.
It provides a platform for the community from industry and academia.

KIT Karlsruhe Institute of Technology

Smart Data Innovation Lab

Smart Data Innovation Lab (SDIL)

Accelerating Data driven Innovation

NESSI Summit May 27, 2014

Prof. Dr.-Ing. Michael Beigl – Department of Informatics

KIT – University of the State of Baden-Wuerttemberg and National Research Center of the Helmholtz Association

www.kit.edu

european data forum

Driving European leadership in ICT innovation for economic growth and quality of life

Empowering ICT Top Talents for the Future
Bringing ICT Innovations to Life
News & Events
eit EIT ICT Labs
About EIT ICT Labs

Commissioner Vassiliou met students at EIT ICT Labs Helsinki

INSTITUT Mines-Télécom

GROUPE-GENES

TERALAB DATA SCIENCE FOR EUROPE

INVESTISSEMENTS D'AVENIR

cap·digital

HPC & Big Data Architectures

Richard SIJBRANDIJ

Big Data Appliances Product Management, BULL

INSTITUT Mines-Télécom · GROUPE-GENES · TERALAB DATA SCIENCE FOR EUROPE · INVESTISSEMENTS D'AVENIR · cap·digital

Both Big Data and HPC have similarities

- Processing large volumes of data
- Parallel processing needs
- Store large volumes of data
- Large data centres:
  - –max perf with min power/cooling
- Similar underlying architectures / constrains
  - –Latency and processor performance (vs in-memory)
  - –Head Node ~ Name Node (vs Hadoop)
  - –Compute Node ~ Data Node (vs Hadoop)

HPC generally applies

©P. Stroppa/CEA

- Circumscribed to 1 area of knowledge:
  - Physics, Chemistry, …
  - Solves a known equation
  - Homogenous data sets
- Few programming languages:
  - C, C++, CUDA/OpenCL, Fortran
- Additional nodes: Storage, Login, …
- Workflow:
  - Move data from Storage to Compute Nodes
  - Communication between Compute Nodes
  - Data can be partitioned (MPI)
- Storage:
  - Huge volumes of data
  - High bandwidth
  - Parallel
  - /scratch: large & parallel accessed temp files
  - Data «belongs» to the company
- MPI and OpenMP paradigms

Big Data takes advantage of

Integrates several areas of knowledge

- Business, statistics, psychology, sociology, marketing, ...
- Analyze trends
- Heterogeneous data

Vast amount of programming languages

- Though mainly based on Java
- Other: R, Pig Latin, SQL, ...

Workflow:

- Don’t move data, or don’t interrupt data flow
- No communication between Data Nodes

Storage:

- Often Integrated Storage into Data/compute Node (in-memory of database)
- Parallel storage: Data can be partitioned (Hadoop)
- Huge volumes , high-speed and vast variety of data
- Local /scratch to each node
- Data is not only the company’s data

What Big-Data adds to HPC

- HPC
  - Race to EXA-flop
- Centralized data
- Architecture criteria
  - Peta-flop/Watt
  - Peta-Flop/M2
  - Latency “nodes”
- Surprised by HW failures
  - Re-do calculation on different node

- Big-Data
  - Value streaming/real-time data
- Analyses of Structured and /or none-structured data
  - Multiple data streams
  - Huge amount of transactional data
- Architecture criteria
  - Max RAM capacity
  - CPU core/RAM
  - Storage latency/capacity
  - Network bandwith
- Robust HW architectures with resiliency and redundancy
  - Transactional data lives in RAM
  - Used for transactions and analytics

Zoom on
Tera-memory Compartment of Teralab

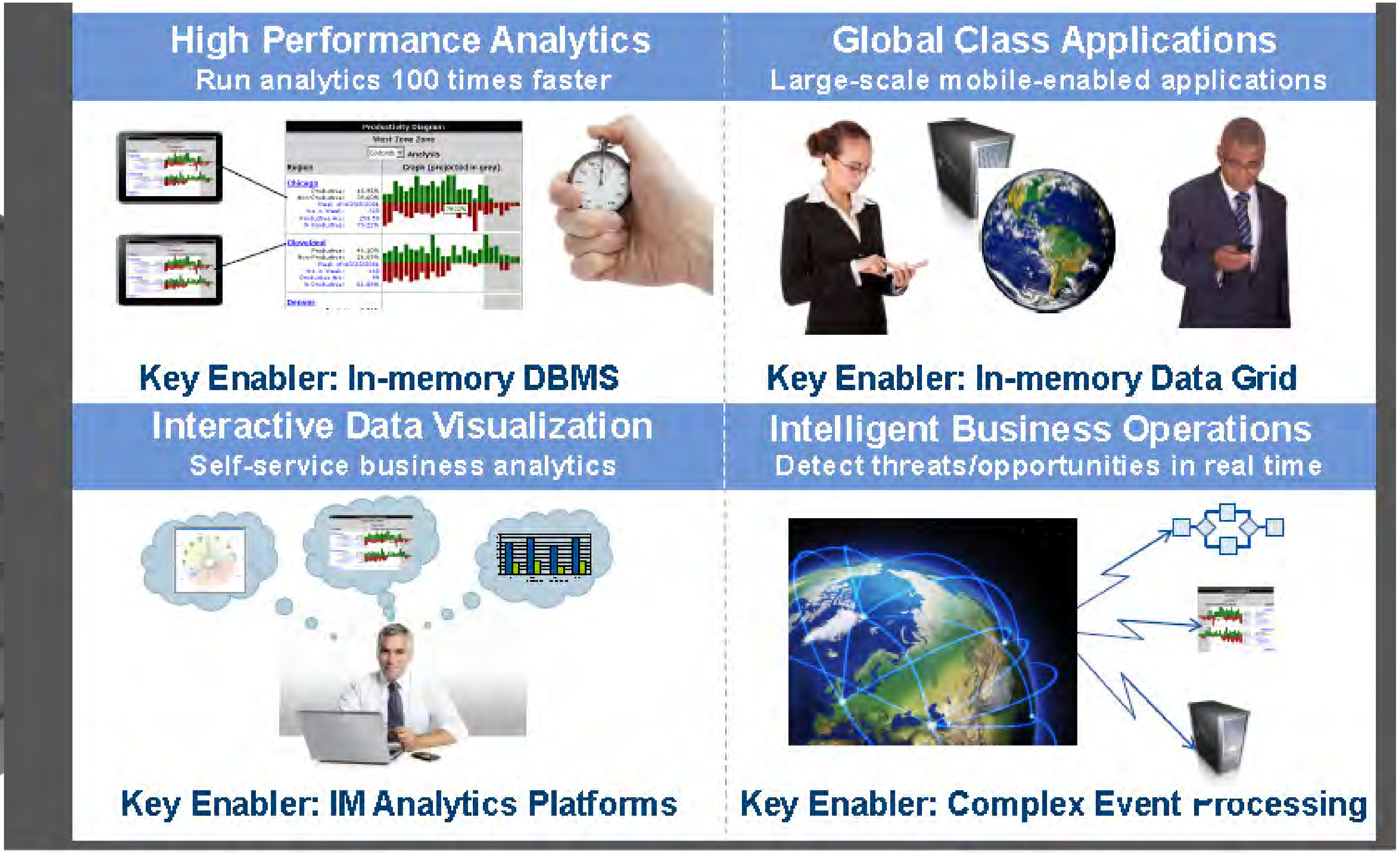
Technical differentiation

- Scales per physical server:
  - Memory upgradable from 4 to 24 TeraByte
  - 240 cores (8 Modulesx2Xeon/modulex15cores/Xeon)
  - > 8000 specint
  - 120To Cold storage
- No proprietary network requirements
- Application communication inside same server
  - Run database, analytics and visualization on a single server

User benefits

- Runs “any” Big-Data requirement
- Very low latency

Bullion S for big-data in-memory computing



High Performance Analytics

Run analytics 100 times faster

Key Enabler: In-memory DBMS

Global Class Applications

Large-scale mobile-enabled applications

Key Enabler: In-memory Data Grid

Interactive Data Visualization

Self-service business analytics

Key Enabler: IM Analytics Platforms

Intelligent Business Operations

Detect threats/opportunities in real time

Key Enabler: Complex Event Processing

Why bullion technology for " in Memory"

- RAM is the new "disk"
  - Adding RAM is as easy as adding disks
  - On the fly without application interruption
- IO scalability, latency and performance improvements
  - Adding IO adapters (RDMA based 10Gb/s, FC) on the fly
- Memory per physical server
  - Up to 24TB of memory per server
- Add dynamically more Compute Modules containing CPU, memory + IO blades
  - With up to 16x sockets and 240x cores

Bullion the most advanced working place for your fast data

- Fast-data = in-memory computing

- RAM = the new disk
  - Adding RAM is as easy as adding a disk with bullion and its memory blades

- RAS for RAM to run production database in memory
  - RAID 5 on HDD = Rank sparing with RAM
  - Bull memory blade migration and extraction = RAID 5 with Hot-Swap disks

- Bullion
  - Only server with 24TB RAM
  - Only server with Hot-swap RAM blades