european data forum

INSTITUT Mines-Télécom

GROUPE-GENES

INVESTISSEMENTS D'AVENIR

TERALAB

TeraLab, A Secure Big Data Platform Description And Use Cases

Franck Cotton – INSEE
Kamel Gadouche – GENES/CASD

TERALAB: DESCRIPTION

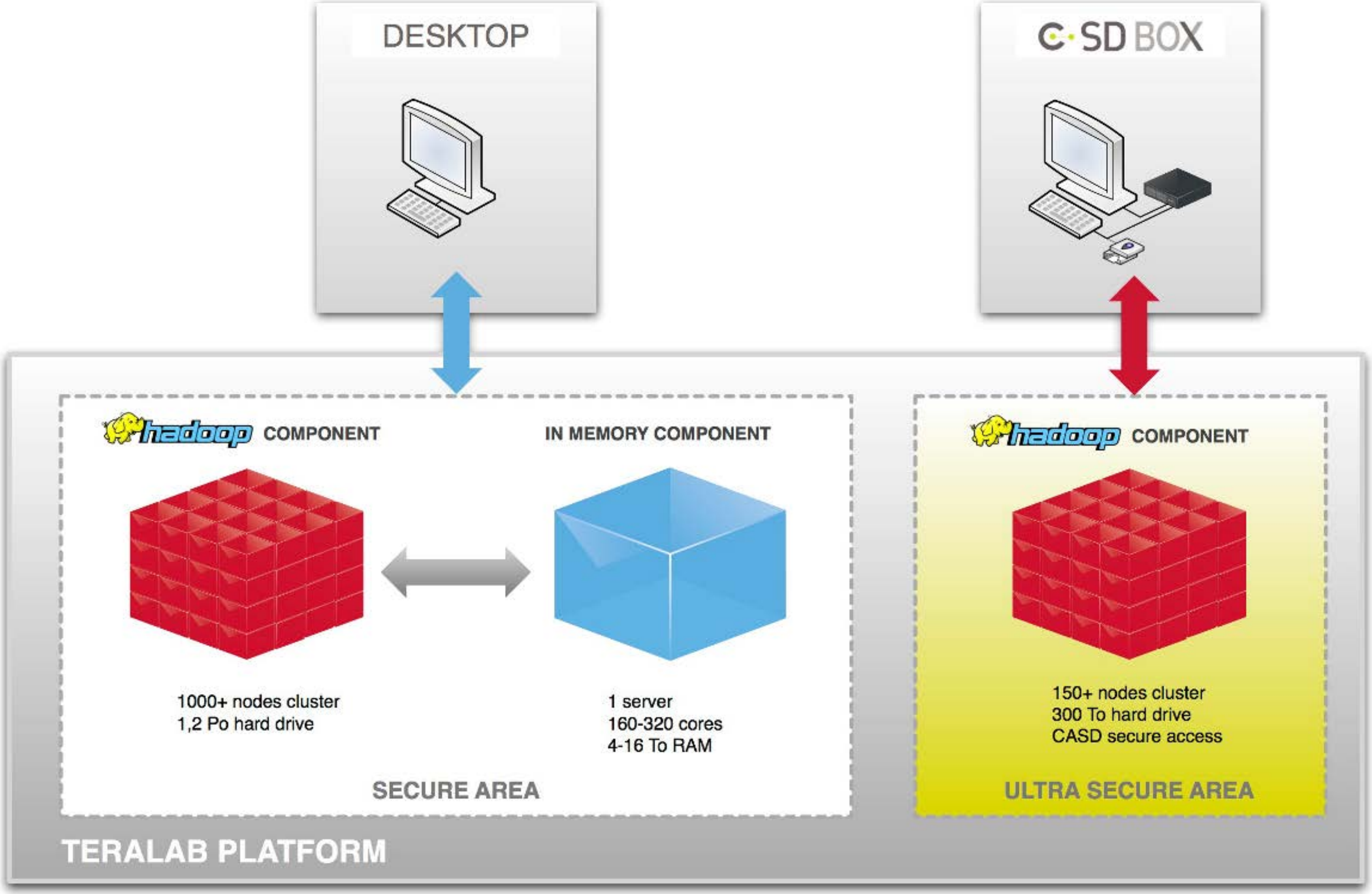
Birth of the TeraLab project

- 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 TeraLab platform

- A state-of-the-art technical infrastructure
  - Elastic distributed system + tera-memory server
  - With unique security features
- A rich catalogue of software tools
  - Data storage (MPP, 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

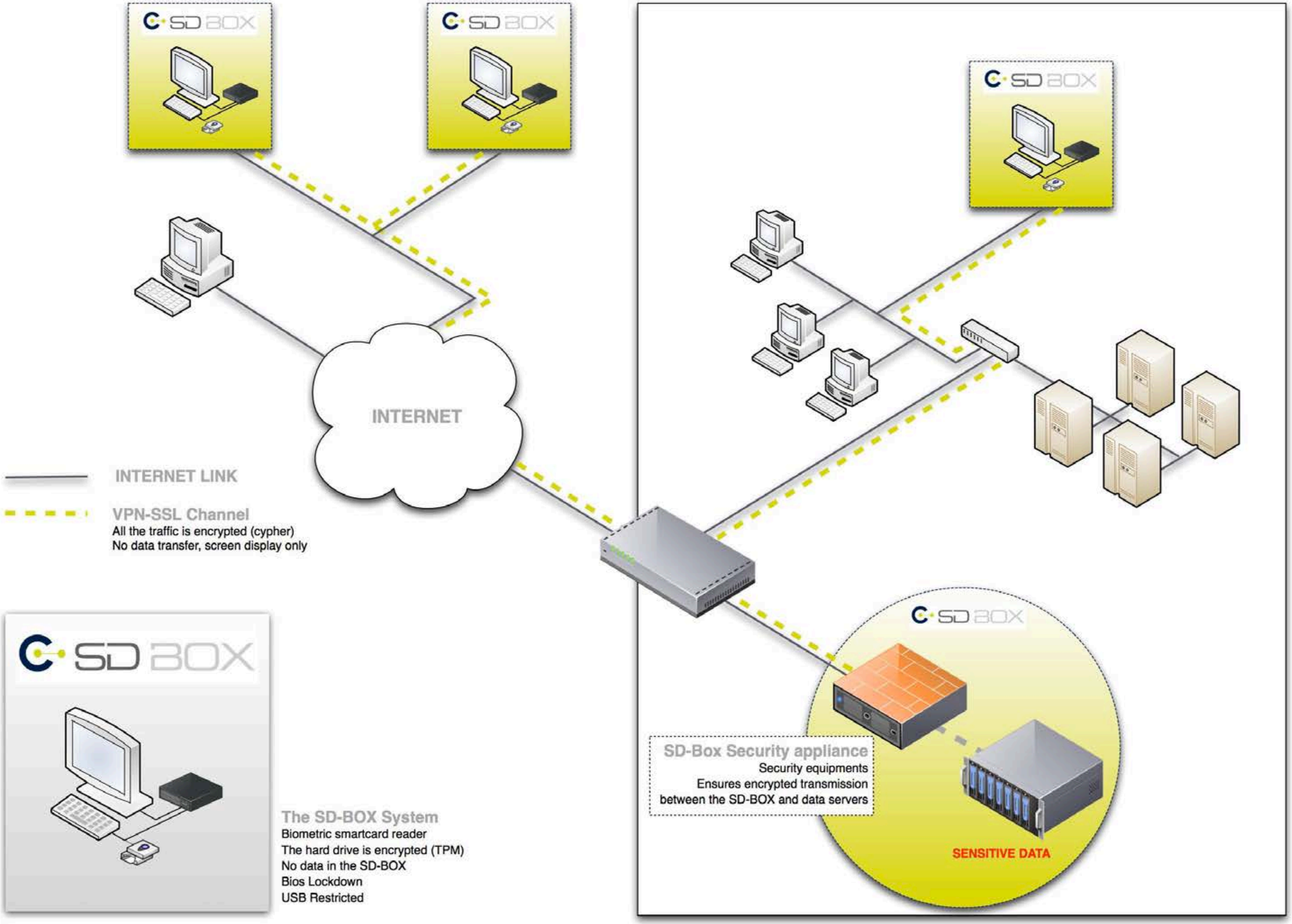
Platform organization



The CASD

- The CASD is a facility including
  - A central secure computing infrastructure (IICE): “the bubble”
  - **Specific access devices (SD-Box™)**, guarantying imperviousness as the sole means for accessing the IICE.
- With the SD-Box, researchers may
  - Work remotely on confidential data
    - With 64-bit statistics sofware: SAS, Stata, R, Gauss, Matlab, Latex, Excel...
    - Soon with Big Data software: Hive, Pig, Mahout, Revolution Analytics, Python…
  - Request inputs or outputs
    - Scripts or data
    - Inputs and outputs are monitored
- With the SD-Box, data owners are sure that
  - The authorized researcher is the one behind the SD-Box (smartcard and biometry)
  - No data can be retrieved by researchers (no copy or paste, printing, USB keys...)

CASD
.NET v2+ Bio
testAuthentification testAuthentification
GENES
INSEE
CASD
SD-BOX
SD-BOX V2 GENES
PRESS ENTER...

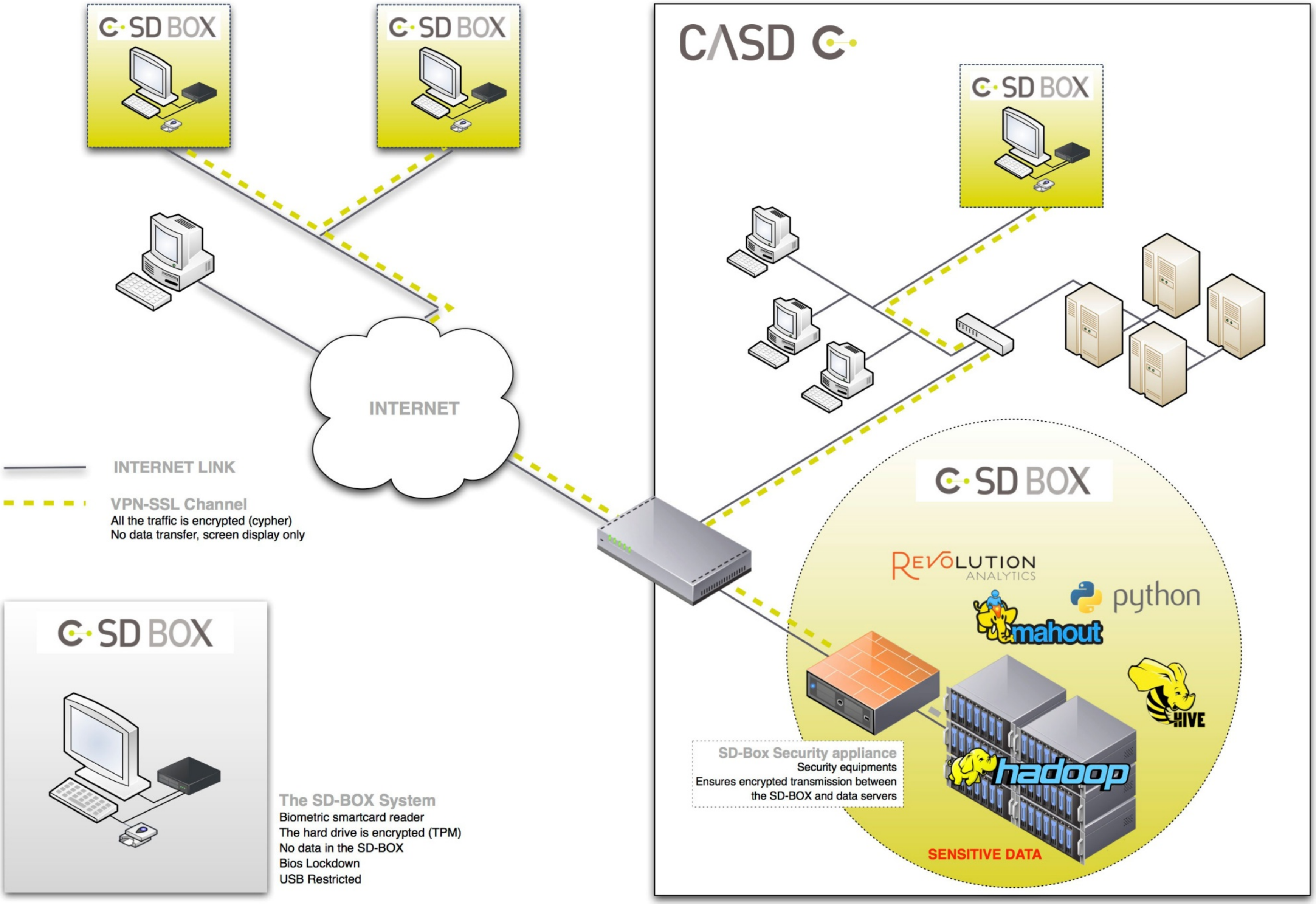


INTERNET LINK

VPN-SSL Channel
All the traffic is encrypted (cypher)
No data transfer, screen display only

SD-Box Security appliance
Security equipments
Ensures encrypted transmission
between the SD-BOX and data servers

The SD-BOX System
Biometric smartcard reader
The hard drive is encrypted (TPM)
No data in the SD-BOX
Bios Lockdown
USB Restricted



INTERNET LINK

VPN-SSL Channel
All the traffic is encrypted (cypher)
No data transfer, screen display only

SD-Box Security appliance
Security equipments
Ensures encrypted transmission between
the SD-BOX and data servers

The SD-BOX System
Biometric smartcard reader
The hard drive is encrypted (TPM)
No data in the SD-BOX
Bios Lockdown
USB Restricted

The TeraLab platform – planning

- 2014 - 2015
  - Incremental platform construction
  - Pilot projects
  - No cost
- 2016 - 2018
  - Professionalization (business model, methodology, client support, etc.)
  - Operating expenses recovery
- 2019 and beyond
  - Target service offer
  - Commercial mode

TERALAB: SOME USE CASES

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 To relational database
  - Other ideas expressed
    - Telco data for tourism statistics
    - Web site log analysis
    - Next-generation social declarations

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

Use case for data challenges

- The DataScience web site (http://datascience.net)
  - Allows data owners to issue public or private challenges based on their data
  - Allows data scientists to analyze the data, to submit models and their results and to get evaluation scores (ranking). The winner gets a prize in euros.
- The goals are to improve the knowledge
  - On methodological aspects
  - On data management aspects
- How TeraLab fits in
  - Allow to organize challenges on Big Data
    - hosted by TeraLab – standard
    - hosted by TeraLab – bubble
  - Help disseminate Big Data technologies

CONCLUSION

Where are we now?

- The story has just begun
  - The planning is tight
    - A ß-version of the distributed service will open in April 2014
    - The “tera-memory” server will open in summer 2014
    - The ultra-secure compartment will open in September 2014
  - The team is currently being set up
  - Several pilot projects have been identified
  - The methodology for projects management is being defined
- Contact us if you have a Big Data project
- Visit us at http://www.teralab-datascience.fr/

Thank you for your attention

franck.cotton@insee.fr

kamel.gadouche@casd.eu