# BIG DATA PLATFORM FOR RESEARCH AND INNOVATION

# TERALAB

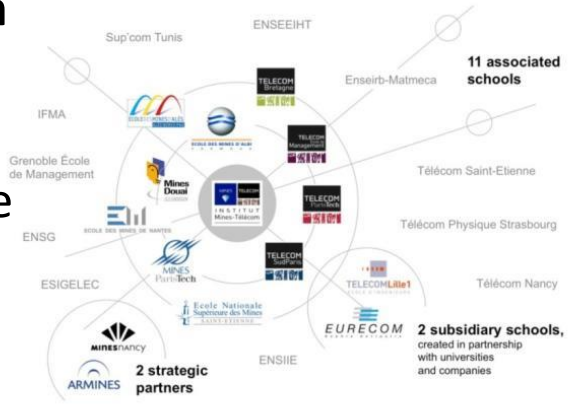
Présentation

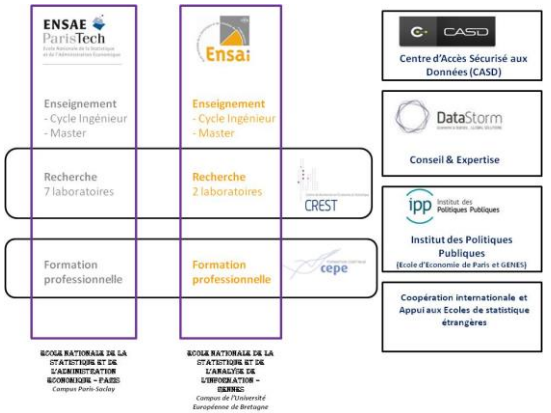
# Teralab

## Vocation : Encourager l'exploitation massive des données

- **Une plateforme dédiée à la recherche / innovation / enseignement**
  - projets de recherche – innovation,
  - Enseignement, dont soutien formation continue

  … en dehors de toute exploitation commerciale.
- **Dan un cadre**
  - De projet collaboratif
  - Ou bilatéral
- **Financée par le PIA (5,7M€)**
  - 2012 : appel à projet PIA remporté par l'**IMT** et le **GENES**, en partenariat avec l'INSEE
  - Fin 2013 : validation du volet financier par la caisses des dépôts et consignation
  - Pour une durée de 5 ans (T0 janvier 2014)
- **La plateforme a été ouverte en avril 2014**



# Teralab

## Un positionnement de catalyseur

- Teralab se positionne en catalyseur / accélérateur de l'exploitation des données par les acteurs économiques

1. **Lève les freins pour les acteurs hésitants à passer à l'action sur leurs données**
2. **Accélère la maturation des usages grâce à l'orientation innovation / R&D comme cadre imposé**
3. **Transfère les projets matures aux acteurs concurrentiels dès lors que le partenaire sait ce dont il à besoin**
4. **Promeut les acteurs innovants de l'écosysteme : installés sur la plateforme pour faciliter la rencontre avec leur marché**

# Teralab / cadre d'intervention

## Synthèse des modalités d'accès

| DISPOSITIFS * | CIBLES | CONDITIONS DE FINANCEMENT |
|---|---|---|
| **PROJETS R&D COLLABORATIFS** | • Des consortiums R&D composés d'industriels, de laboratoire de recherche et de PMEs/startups | • label R&D requis, national ou européen (ANR,FUI, H2020, Carnot…)<br>• Sélection par le comité scientifique TeraLab<br>• Financement variable (jusqu'à 100K€ environ) |
| **CHALLENGES BIG DATA CAP DIGITAL** | • Des grands groupes souhaitant être sponsors de l'initiative<br>• Des PME/Startups big data souhaitant avancer avec un sponsor fournisseur de données | • sélection des challenges par le consortium<br>• Equipe projet est co-financée par l'Etat et le Sponsor à hauteur de 100K€ par an en moyenne (sur une durée de 3 ans) |
| **POCs BILATERAUX** | • Des consortiums commerciaux composés de grands groupes, cabinets de conseils et startups big data | • Sélection par le comité scientifique TeraLab<br>• Pas de financement direct mais possibilité d'inclure d'autres financement (thèses CIFFRE, labos communs) |
| **PROJETS "BOOTSTRAP"** | • Un consortium R&D ou commercial intégrant un chercheur de l'institut Mines Télécom | • Le projet est financé à hauteur de 100K€ maximum par l'intermédiaire du laboratoire de recherche<br>• Sélection par le comité scientifique TeraLab<br>• La décision finale revient à la DGE |
| **ENSEIGNEMENT** | • Des projets type fil rouge académiques ou industriels souhaitant proposer une alternative souveraine à un hébergement classique type aws | • Sélection par le comité scientifique TeraLab |

*document de travail du 06/11/2014

# Teralab / composantes de l'offre

## Panorama global

**Infrastructure technique**
- Hébergement souverain
- compartiments sécurisés et dédiés
- Double environnement : distribués et in memory

**Accompagnement**
- Technique prise en main
- Scientifique
- Ecosystème

**Ecosystème embarqué**
- Les solutions standrds du marché,
- Les solutions de start up innovantes

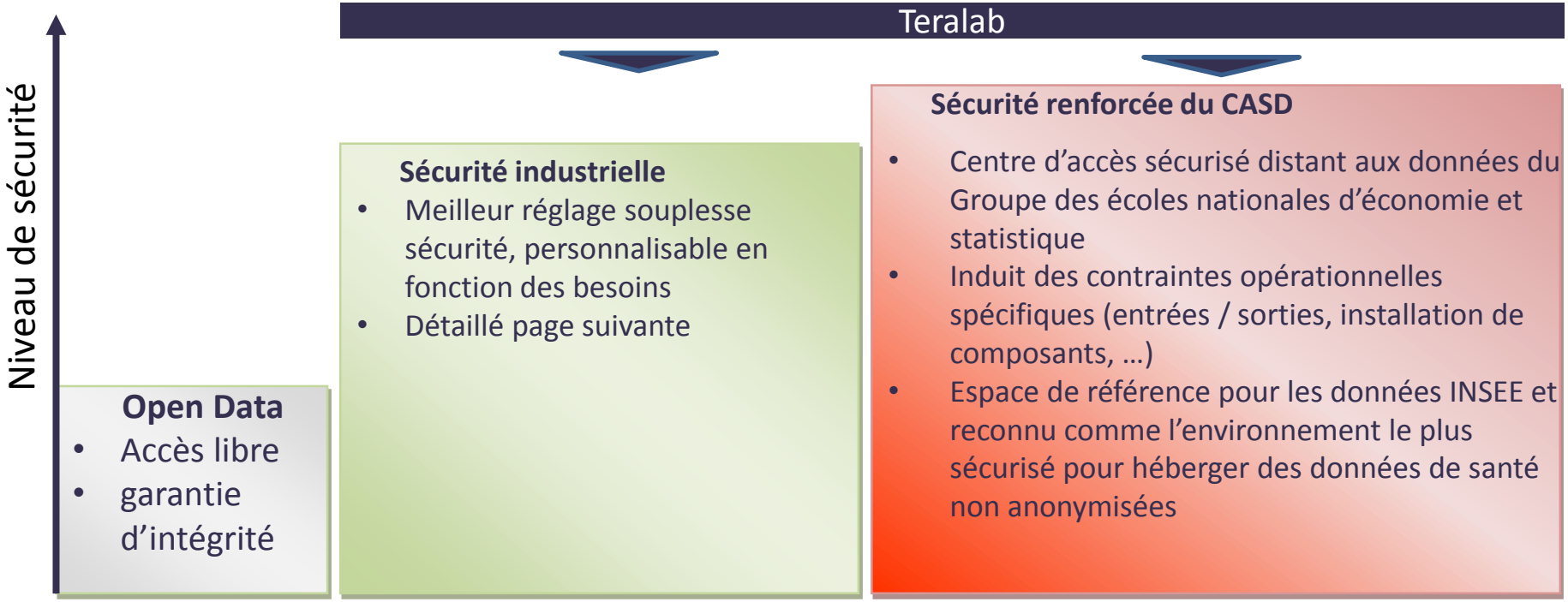
**Corpus de données à disposition**
- Etalab
- Crawling ..
- ....

Cible mi 2015

# La gestion de la sécurité est à la base de notre infrastructure

Offre

- L'activité de recherche / innovation induit une acceptation juridique du risque, cependant Teralab met en œuvre tous les niveaux de sécurité industrielle, déclinée en deux niveaux :

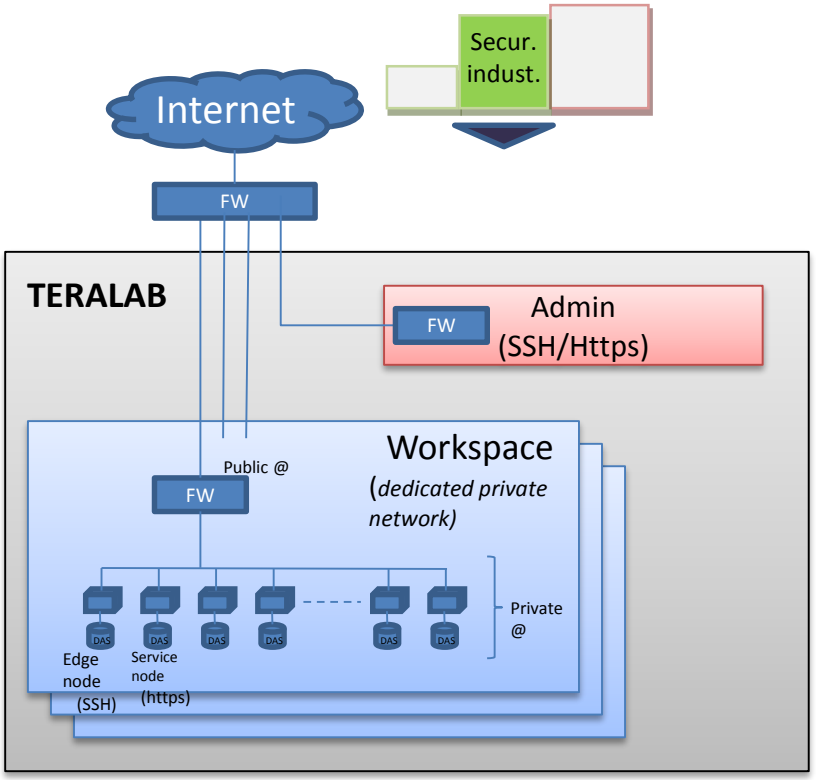



**Une garantie d'hébergement physique en France métropolitaine**

# Teralab / offre - infrastructure

## Le niveau « sécurité industrielle » optimise le compromis accessibilité / sécurité

- Espaces de travail cloisonnés (VLAN)
  - Switch réseau configurés pour constituer des VLAN isolés pour chaque cluster
  - Protège des incursions / perturbations extérieures
- 2 niveaux de firewall
  - En amont de la plateforme : le seul à avoir une IP publique
  - En amont de chaque cluster, personnalisable (sur demande)
- Tout accès nécessite un compte utilisateur (login /mot de passe)
- Droit d'administrateur pour le partenaire lui permettant de personnaliser son dispositif de protection (authentification par clé par exemple)

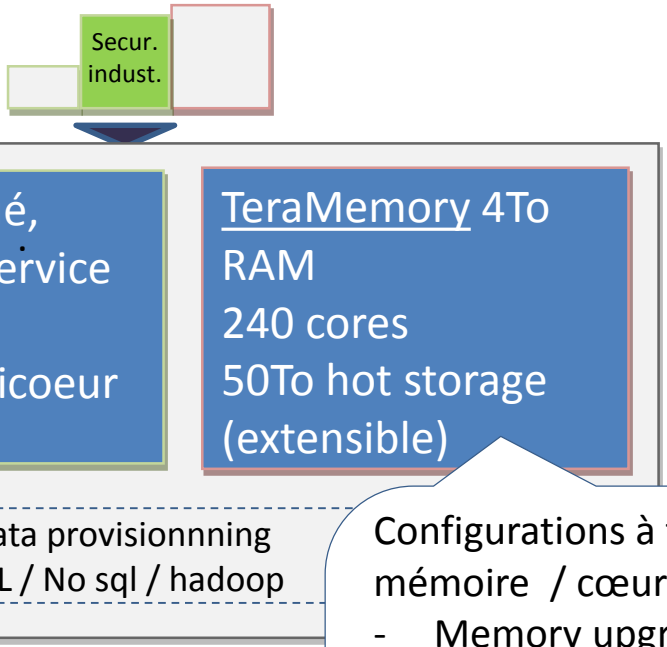



- Le niveau de sécurité peut être ajusté en fonction des besoins d'échange de données du partenaire (objets connectés par exemple)

# Teralab / offre - infrastructure

## 2 infrastructures de calculs complémentaires

Offre




exemples de configuration de cluster en production

| Hadoop | vcpu | RAM (GB) | Storage (GB) |
|---|---|---|---|
| CDH5.0 | 21 | 148 | 3512 |
| CDH5.0 | 13 | 68 | 1944 |
| CDH5.0 | 69 | 404 | 14744 |
| HW | 13 | 68 | 3512 |

Configurations à très fort ratio mémoire / cœur

- Memory upgradable from 4 to 24 TeraByte
- 240 cores (8 Modules x2 Xeon/module x15cores/Xeon)
- 8000 specint
- Net storage (DAS - Raid5):
  - current TeraLab: 50TB / Max : 800TB

# Teralab / offre - écosystème embarqué

## Des ressources instalées / une personnalisation par vos soins

- Les deux compartiments e calcul sont prééquipés de l'infrastructure logicielle au niveau de l'état de l'art

**Linux** : CentOS 6.4

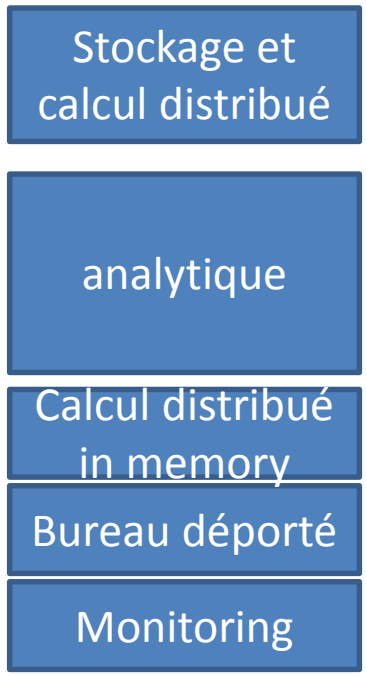
**Hadoop** : Cloudera Full CDH5 ou HW2 (dont web Hue, ClouderaMgr,)

**Python** 2.7 : Canopy (Scikit-learn installable), **R, Dataiku, Pentaho** (partie libre)

**Spark, Flink** (ex stratosphère)

**VNC**

**Ganglia**




Les briques installées sont
- open source
- briques innovantes de start up (ex dataiku) a des fins de promotion

Les gestionnaires de projets hébergés ont un compte administrateur et peuvent **installer toute brique logicielle pour lesquelles ils disposent d'une licence**

# Teralab / offre - écosystème embarqué

## Exemple d'une configuration Teralab

### WORKSPACE (project, team…)

| | |
|---|---|
| **Users** | • Roles (applis, admin. actions)<br>• Groups (files) |
| **Edge nodes** | • Linux server<br>• unix |
| **Cluster** | • Create, start, stop, resize, opera<br>• HDFS, unix |

**Edge node**: The Edge Node is a system intended to be used as a Gateway. It is not a real part of the Cluster but all client library are installed on, thus allowing full access to the cluster from this system.

**Service node**: System hosting most of helper services (Hue, Ganglia, Cloudera Manager, Ambari ...).

**Name nodes**: Two Namenodes acting as active/passive High Availability configuration.

**Data nodes**: The workhorses of the cluster. Provide storage and compute resources.

**Firewall**: A pfsense Firewall controlling all cluster access.

### Environment

| | |
|---|---|
| **Hadoop Environment** | • Hadoop: Cloudera 4.6 + Cloudera 5 |
| **Administration** | • TeraLab mgr (cluster), Nagios (alerts), Ganglia (monitoring) |
| **User operation** | • Hue, ssh |
| **Applications** | • Oozie, Pig, Hive, Impala, Mahout…<br>• Dataiku, Opendatasoft… |
| **Database** | • HBase (columnar) |
| **Parallelisation Engine** | • Yarn, MapReduce, Stratosphere<br>• Spark |
| **File system** | • HDFS |
| **Cluster OS** | • CentOS 6.4 (incl. python 2.6.6)<br>• Others linux distributions |
| **Import/Export** | • Flume, Hive Metastore Mgr, Sqoop |
| **Others** | • Any imported Linux tool (may run on the cluster or on its edge node) |

## Une équipe de 6 personnes dédiée en soutien des utilisateurs de la plateforme

Offre

- **Sur le volet technique**
  - Assistance à la prise en main de l'infrastructure : conseils et aide opérationnelle
  - Hotline avec des tickets 'incidents en ligne
- **Sur le volet scientifique**
  - Une directrice scientifique dédiée au projet, avec une très forte expérience opérationnelle
  - Un lien direct avec les équipes de recherche Mines et Telecom (dont laboratoires de machine learning
- **Sur le volet économique**
  - Une orientation si besoin au sein de l'écosystème économique Big Data : acteurs du marché, financements potentiels, start up sur des activités connexes

# Plateforme référencée dans le cadre du plan Big Data (Plan “nouvelle France Industrielle”)

..

**CHEFS DE PROJET**
**Paul Hermelin est PDG de CapGemini.**
**François Bourdoncle est Président de FB&Cie.**

**LES ÉCOLES ET ORGANISMES DE RECHERCHE**
IMT, ENSAE, ENSAI, CEA Tech, Centrale, Ifsttar, Institut Gustave Roussy, INSERM

**LES ACTEURS PUBLICS**
CGI, CNIL, MERPN, Pôle Emploi

**LES INDUSTRIELS**
Orange, Criteo, AFDEL, Covéa, La Poste, Sicpa, FING, AXA, Solocal Group, Crédit coopératif, Intermarché/Les Mousquetaires, Société Générale, FING, CozyCloud, GDF Suez, Sherpa Engineering, Alstom, CosMo Compagny, Ecogelec, G2 Mobility, MGEN, Integragen, Sogeti Hi tech

# Teralab – écosysteme
## Projets déjà hébergés

| | Project description | Actors involved |
|---|---|---|
| Pure R&D | ▪ Machine Learning project (?)<br>▪ SickitLearn project (?) | ▪ Big Data Researcher: A.Gramfort, S.Clemencon, … |
| Collaborative R&D | ▪ Speedata - Decision-making platform in SaaS mode<br>▪ D4D 'Data for Development Senegal' - innovation challenge open on ICT Big Data for the purposes of societal development in Senegal.<br>▪ Square Predict - Big Data SaaS platform for insurance industry.<br>▪ Normatis - Use of multimodal data with Web information extraction, data integration, and mobility data management. | ▪ Large industrial groups, SMEs, R&D Labs: |
| Teaching | ▪ Yearly project (?)<br>▪ Lab work project (?)<br>▪ Others (?) | ▪ Lecturer and researcher, students of big data masters: |
| Industrial Proof of Concept | ▪ Proof of Concept– Fraud Detection for the french actor of private health insurance<br>▪ Proof of Concept Pôle Emploi (French Employment Office) – Matching of jobs offers/profile/specific training to improve unemployed people market reinsertion | ▪ Large industrial groups with business/big data challenges: |

# 4 chaires dans le périmètre Mines Telecom directement en lien avec le Big Data

| | Pilote | Partenaires |
|---|---|---|
| **Machine Learning for Big Data** | Stéphan Clémençon | BNP Parisbas , Criteo , PSA, Safran |
| **Valeur et politique des informations personnelles** | Claire Levallois Barthes | Groupe Imprimerie Nationale, BNP Paribas et Dassault Système, avec l'aide de la CNIL |
| **réseaux sociaux** | Christine Balagué | la Poste, Pages Jaunes Groupe et Danone |
| **E commerce** | Talel Abdessalem | Partenaires |