



BIG DATA PLATFORM FOR RESEARCH AND INNOVATION

TERALAB

Denis Oblin

denis.oblin@mines-telecom.fr



TERALAB

DATA SCIENCE FOR EUROPE



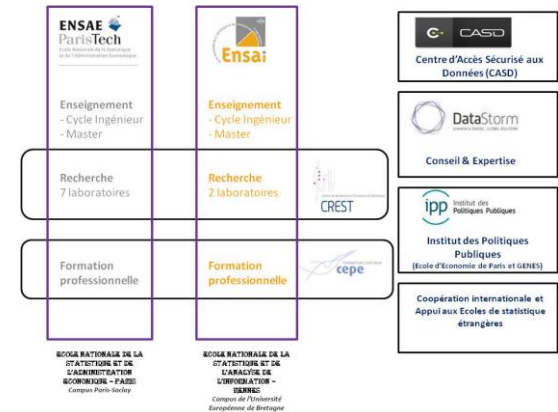
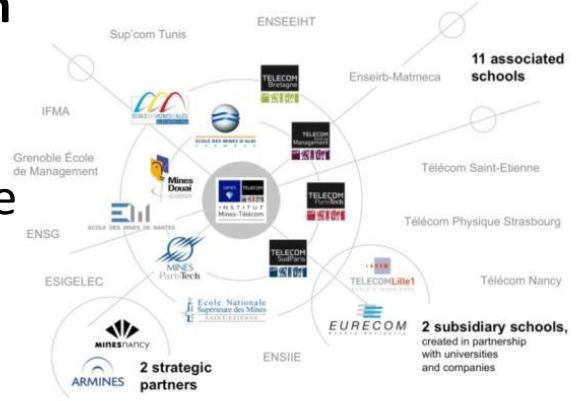
cap·digital



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 un cadre de partenariat avec un laboratoire

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'écosystème : installés sur la plateforme pour faciliter la rencontre avec leur marché



TERALAB

DATA SCIENCE FOR EUROPE



cap·digital
Paris Region



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 standards du marché,
- Les solutions de start up innovantes

Corpus de données à disposition

- Etalab
- Crawling ..
-

Cible mi 2015



TERALAB

DATA SCIENCE FOR EUROPE



cap.digital

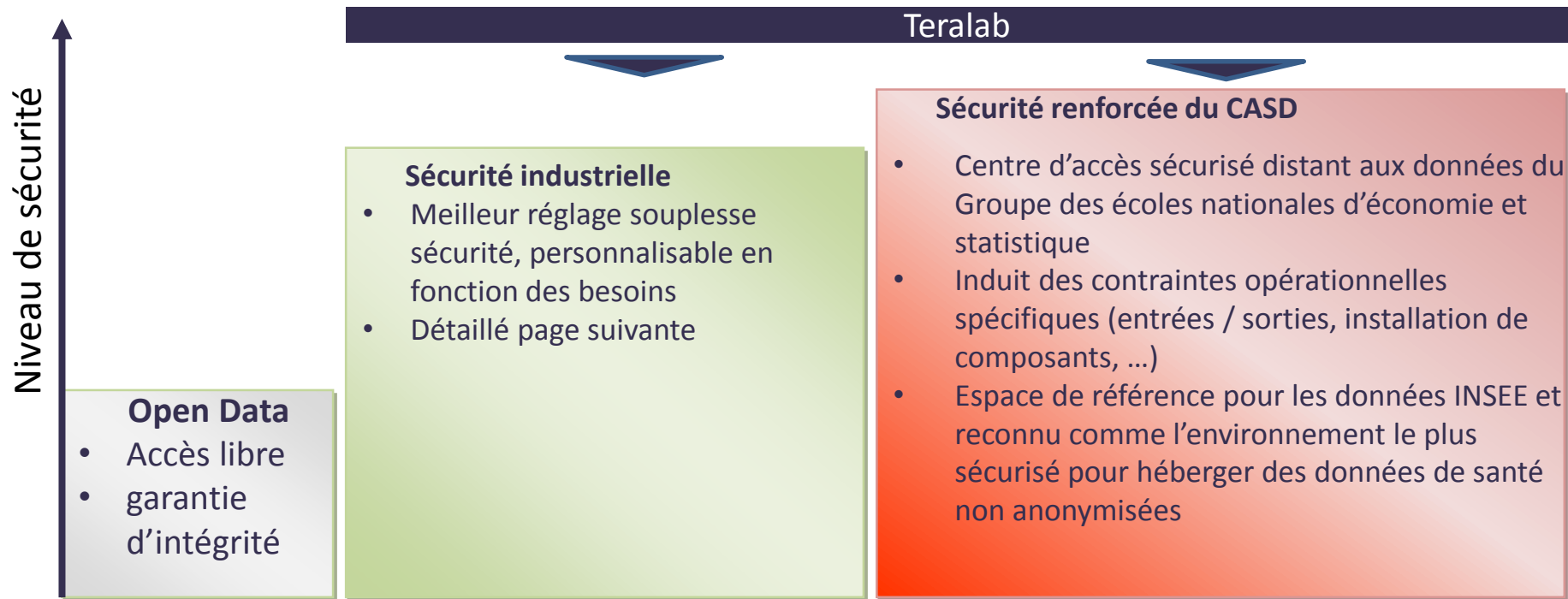


Teralab / offre - infrastructure

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

2 infrastructures de calcul complémentaires



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

Data provisionning
SQL / No sql / hadoop

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 installées / 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

Stockage et
calcul distribué

analytique

Calcul distribué
in memory

Bureau déporté

Monitoring

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

DATA SCIENCE FOR EUROPE



cap.digital



Teralab / offre - Accompagnement

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



- **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”)

..





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

	Pilote	Partenaires
Machine Learning for Big Data	Stéphan Cléménçon	BNP Paribas , 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



TERALAB

DATA SCIENCE FOR EUROPE



cap·digital