Big Data et données confidentielles

LE CASD ET LA PLATEFORME TERALAB

Alexandre Marty [alexandre.marty@casd.eu]

CASD

Plan

- Le CASD
  - Démo
- La plateforme TeraLab
- Le projet Données de Caisse
  - Démo

Le CASD

Le Centre d'Accès Sécurisé aux Données

- Accès et traitement de données confidentielles
- Activités :
  - Premier hébergeur français de données administratives pour la recherche
  - Fournisseur de services pour le public et le privé
- Plusieurs aspects :
  - Hébergement, data center
  - Terminaux, technologie d'accès
  - Conseil, support, maintenance

CASD

Quelques chiffres

- Equipe d'une vingtaine de personnes
- 80 sources de données
- 200 projets
- 800 chercheurs
- Projets d'entreprises privées

- Croissance rapide

CASD

Technologie

- Equipement permettant d'assurer la sécurité de données très sensibles
  - Authentification forte par biométrie
  - Prévient physiquement et logiquement toute évasion de fichier de données
- Equipex
- Des travaux de valorisation de la technologie

CASD

La SD-Box


Un boitier autonome et dédié à l'accès
Lecteur biométrique
Lecteur de carte à puce
Connecteur RJ45
Connecteur VGA
Connecteurs USB
Ecran et boutons de paramétrage sur site
CASD
CENTRE D'ACCES
SECURISE AUX DONNEES
SDBOX

Un boîtier sécurisé conçu par nos soins

Hautement sécurisée et dédiée à l'accès sécurisé aux données sensibles

Authentification forte et incontournable

Verrouillée physiquement et logiquement, pour limiter la récupération de fichiers

Pas de données sensibles stockées sur la SD-Box

Des boîtiers standardisés

- Fiabilité (configuration unique, stable et validée)
- Déploiement simple et économique
- Besoin très limité d'assistance

Facile à déployer

- Nécessite un écran, un clavier & une souris
- Nécessite une connexion à internet
- Simple à configurer
- Pas d'incidence sur le reste du SI

CASD

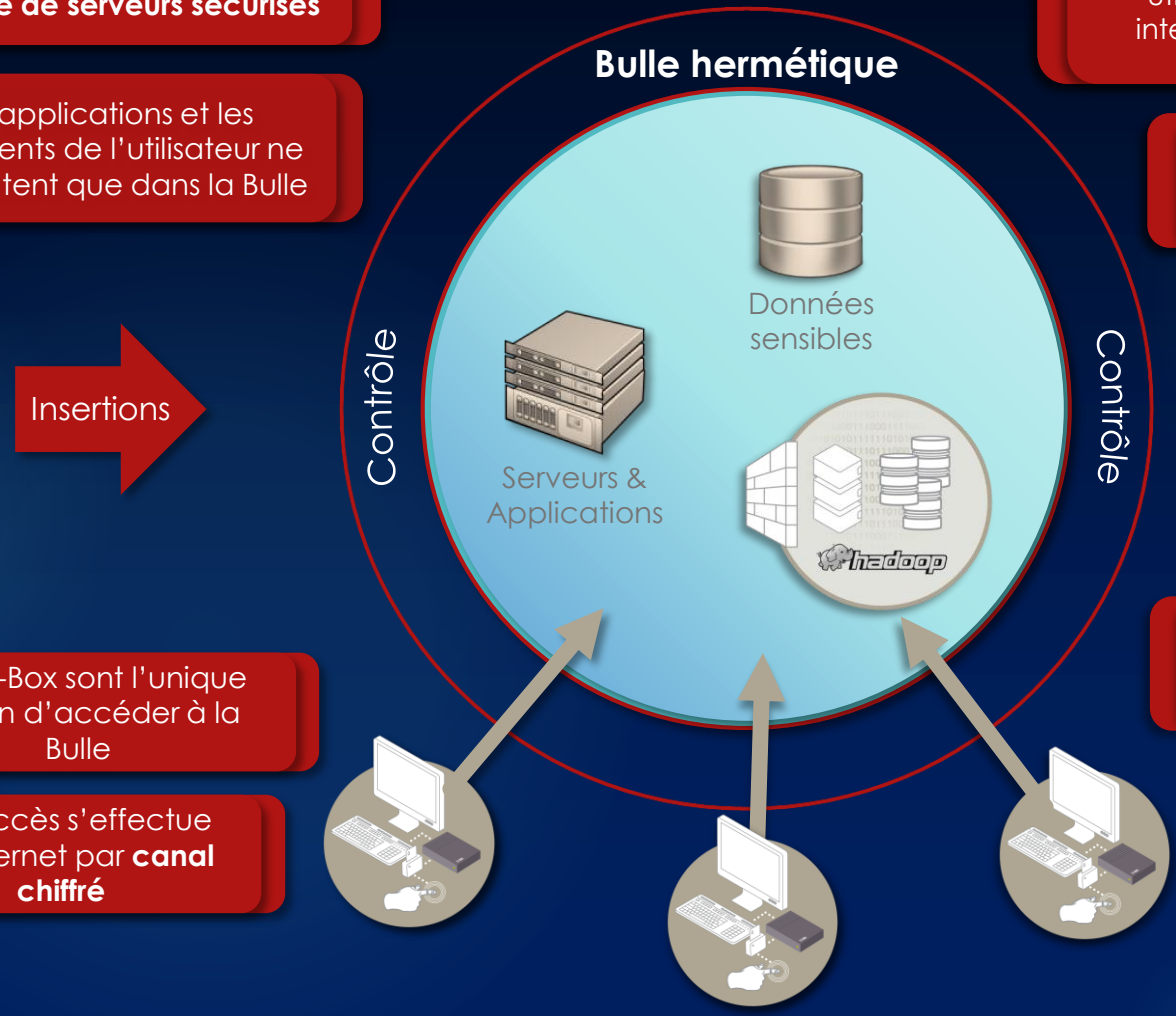
La bulle : notre infrastructure étanche

La Bulle est **un ensemble étanche de serveurs sécurisés**

Les applications et les traitements de l'utilisateur ne s'exécutent que dans la Bulle

Les insertions / extractions de données sont **contrôlées**. Les utilisateurs n'ont pas accès à internet depuis leur espace de travail.

Les données sensibles sont hébergées uniquement dans la Bulle



Les SD-Box sont l'unique moyen d'accéder à la Bulle

Cet accès s'effectue via internet par **canal chiffré**

Un cluster Hadoop est à disposition pour les traitements **BigData**

CASD

La SD-Box

La bulle

La SD-Box

Les données

Les outils

CASD

L'offre classique CASD

- Tous les outils dont les chercheurs ont besoin
  - Environnement Windows
  - SAS, R, Stata, SPSS, Python…
  - Ajout de tout logiciel sous réserve de licence compatible
- Plug & Play
  - Il suffit de brancher la SD-Box et ça marche
  - Support technique

Démo

CASD

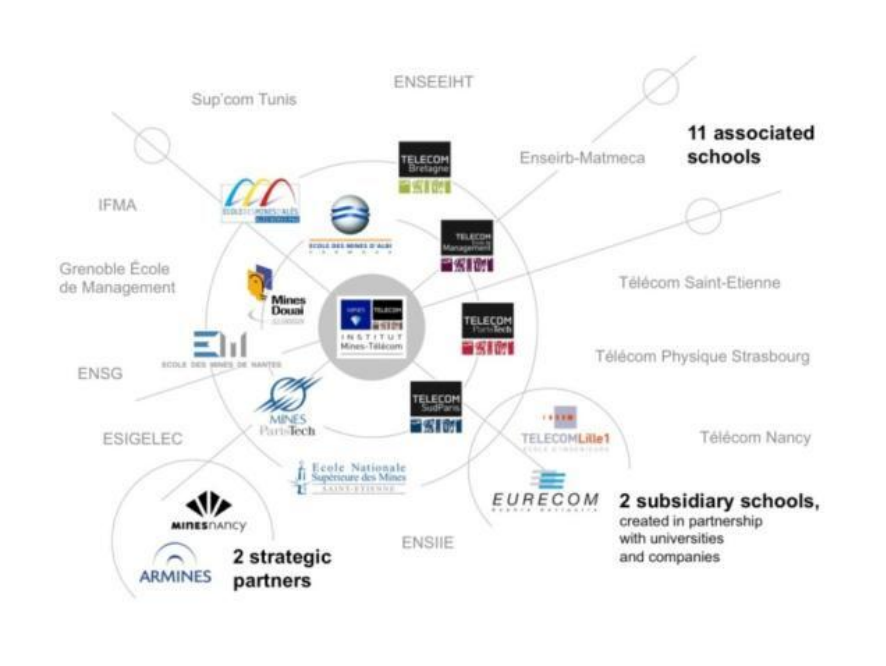
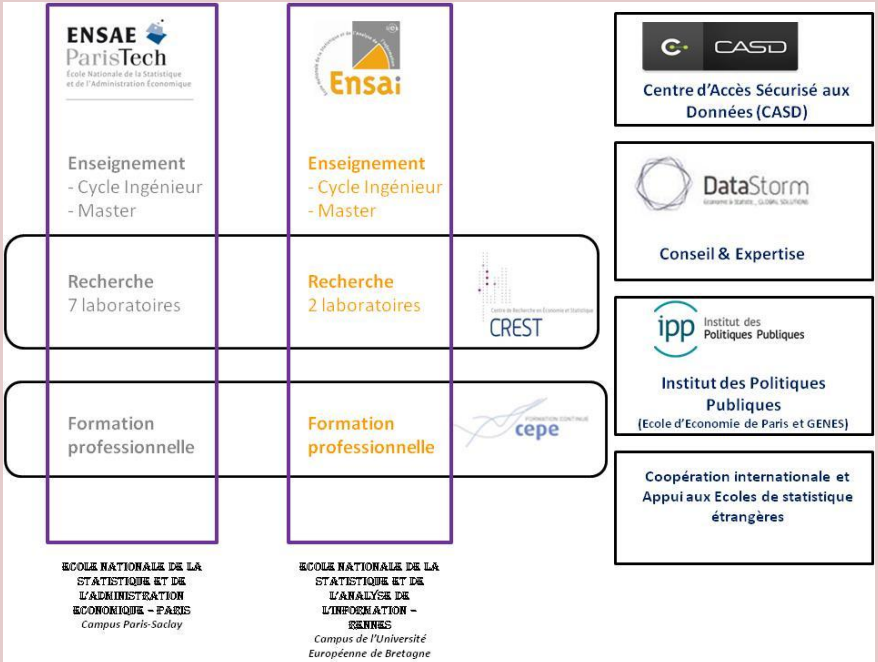
Plateforme TERALAB

CASD

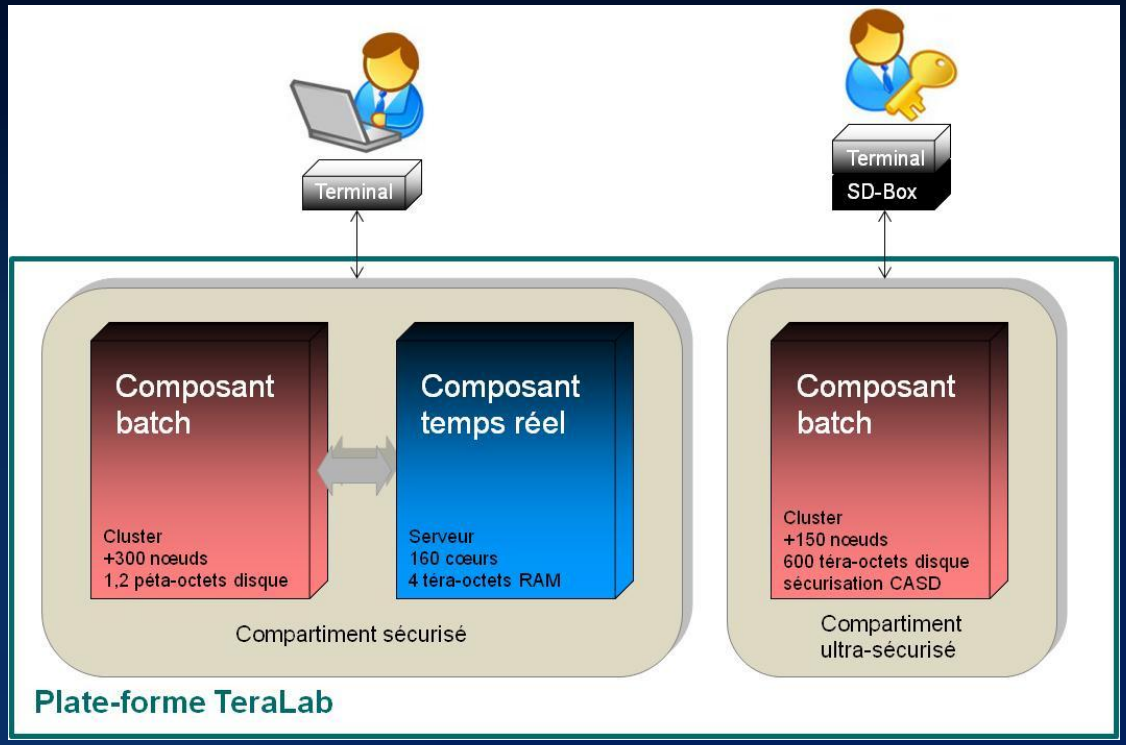
Le projet TERALAB

- Appel à projets "Cloud computing/Big Data" dans le cadre du PIA (Programme d'Investissements d'Avenir)
- Construction et exploitation d'une plateforme Big Data
  - Pour la recherche, l'innovation et l'enseignement
  - Soumise par un consortium comprenant
    - L'IMT (Institut Mines-Télécom)
    - Le GENES, et particulièrement le CASD
    - En partenariat avec l'INSEE
- Projet sélectionné et lancé
  - Budget de 5,7 M€
  - Durée de 5 ans
  - Signature en Décembre 2013

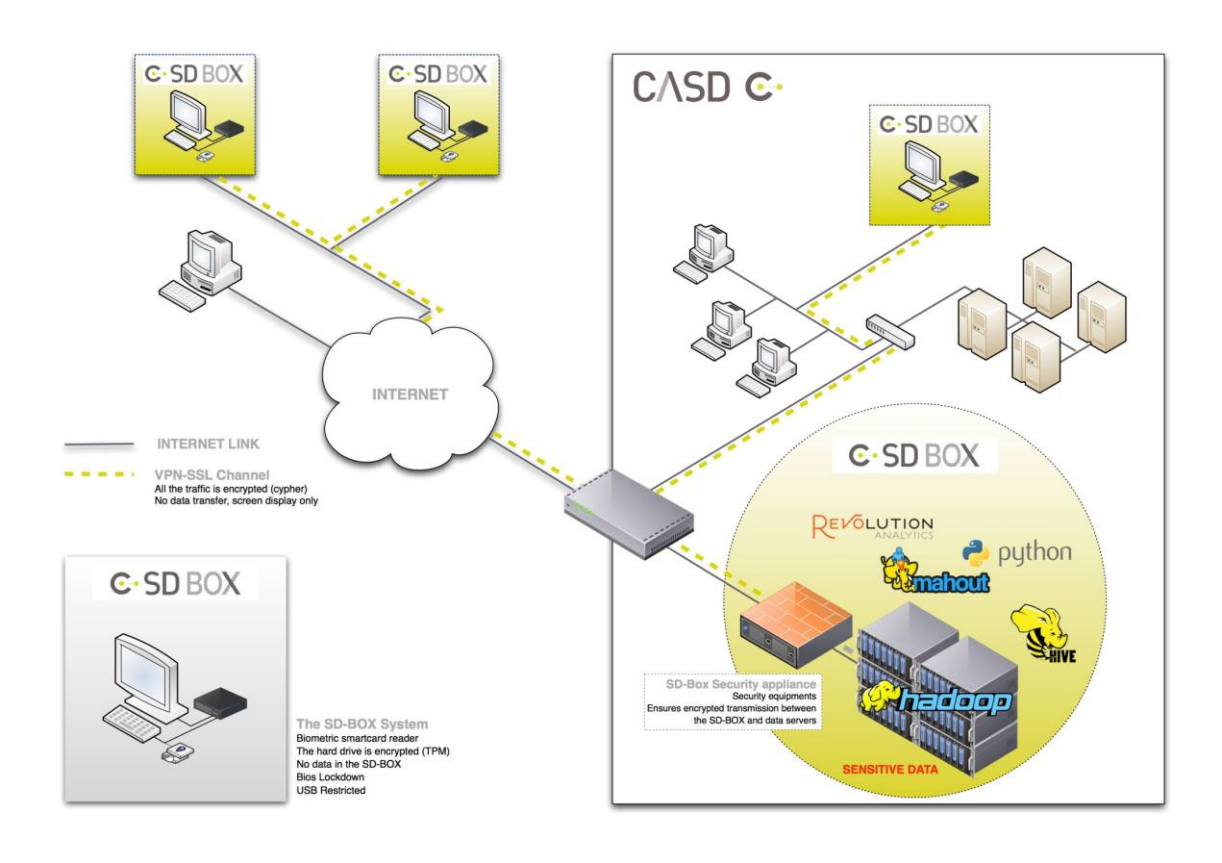
Acteurs

	Institut Mines-Télécom	GENES
		
Domaines majeurs	Numérique, énergie, environnement Matériaux, économie/ entreprise/ société	Economie, Sciences humaines et sociales mathématiques appliquées, Statistiques, économétrie
Tutelles	Ministère du redressement Productif, de l'Enseignement Supérieur et de la recherche	Ministère de l'Economie et des finances (INSEE)

Compartiments de la plateforme



Architecture TERALAB-CASD



CASD

Technologies Big Data

- Infrastructure puissante et extensible
- Environnement Hadoop
- Nombreux outils pour les data scientists :
  - Python, R
  - Machine learning
  - Serveur Open Street Map

CASD

Projets

- Données de caisse
- RTE (Réseau de Transport d'Electricité)
  - Problématiques d'optimisation du réseau (prévision, maintenance et consommation)
  - Grande diversité de sources
  - Développement d'applications innovantes
- En discussion : Données de santé
  - Sujet en pleine actualité
  - Grande confidentialité
- Implication dans des projets européens : DwB, Eurostat Big Data Task Force

Projet Données de Caisse

CASD

Contexte

- Données de caisse de grande taille
- Pertinence des technologies Big Data
- Cluster Hadoop ultra sécurisé du CASD
- Outils scientifiques puissants

CASD

Objectif

- Amélioration de l'IPC :
  - capture des prix réels des ventes
  - exhaustivité
- Nouvelles productions, en gardant les concepts actuels de l'IPC :
  - prix moyens
  - indices régionaux
  - comparaisons spatiales

Données

- Fichier par semaine et par enseigne avec les mesures suivantes :
  - Prix de vente
  - Quantité
  - Point de vente
  - Code barre (EAN)
  - Date de vente
  - Famille de l'article
  - Description de l'article
- Prix et quantités simulés
- Référentiels des points de vente + Référentiel des articles : volume faible

Cluster actuel

- 6 machines : 2 nœuds de gestion , 4 nœuds de travail
- 48 cœurs, 512 GB de RAM
- 33 TB de stockage

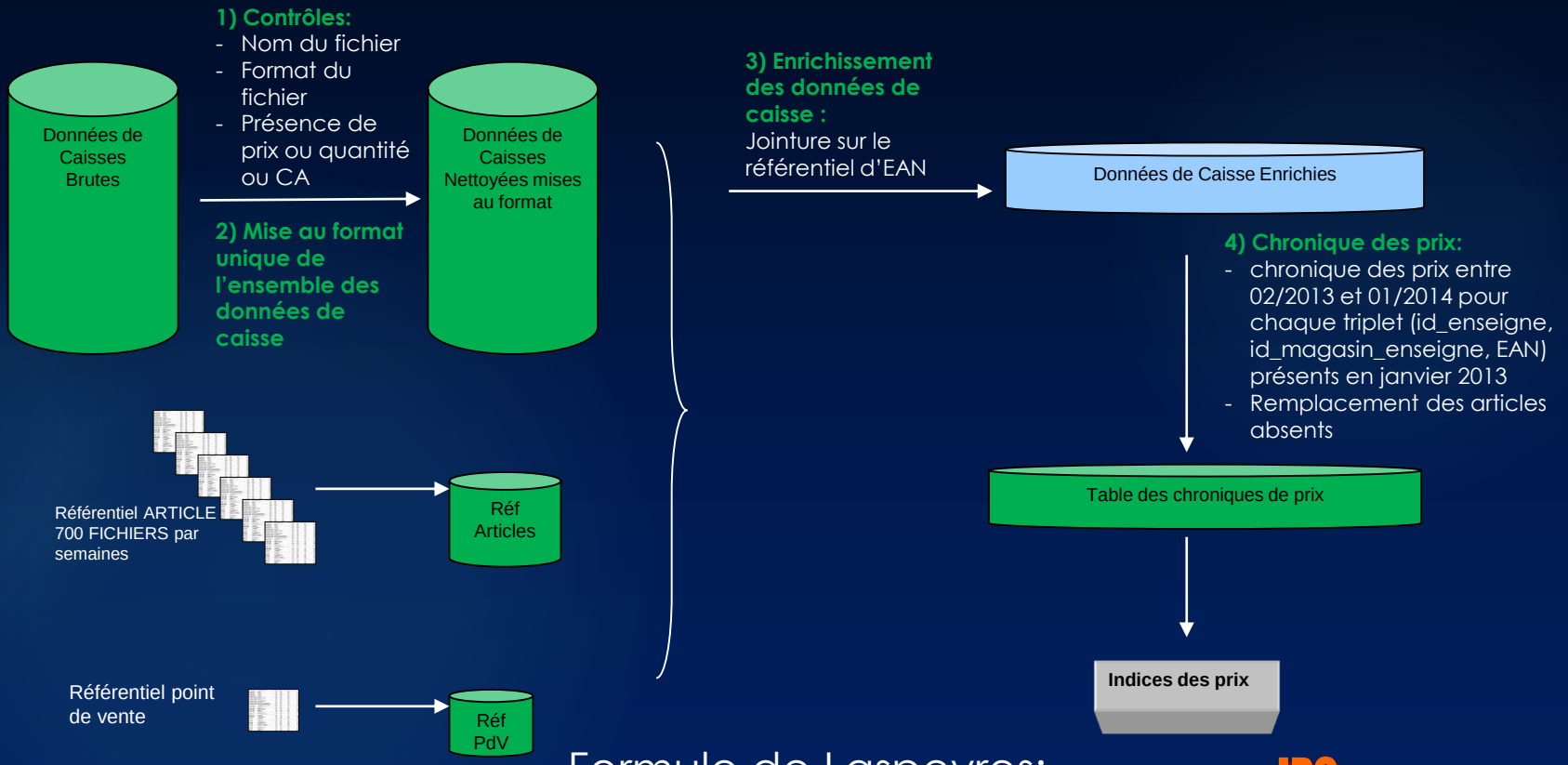
- Extensible à la volée

CASD

Volumétrie

- Situation réelle actuelle :
  - 5 enseignes livrent 10 Go de données par semaine
  - 900 Go de données à traiter
  - 5,7 milliards de lignes
- Situation de test :
  - Situation réelle x5
  - 28,5 milliards de lignes
- Situation de long terme : 10x situation actuelle

Traitements



Formule de Laspeyres:
IPC par famille de produits =

$$\prod_{i=1}^{n}\left(\frac{P_i^t}{P_i^0}\right)^{ca/\sum ca}$$

→ IPC

Outils

- Pig, Hive, Impala :
  - Chargement et traitements type base de données SQL
- Python :
  - Orchestration des tâches

Résultats

- Temps de traitement :
  - Chargement de 70 semaines de données : 5h
  - Constitution du panier : 2h40
  - Jointure référentiel produit : 3h05
- Par rapport à solution Oracle actuelle :
  - Complexité et temps de développement très réduits
  - Temps de traitement très inférieurs

Merci !

Questions ?

Alexandre Marty [alexandre.marty@casd.eu]

CASD