Projet Xdata

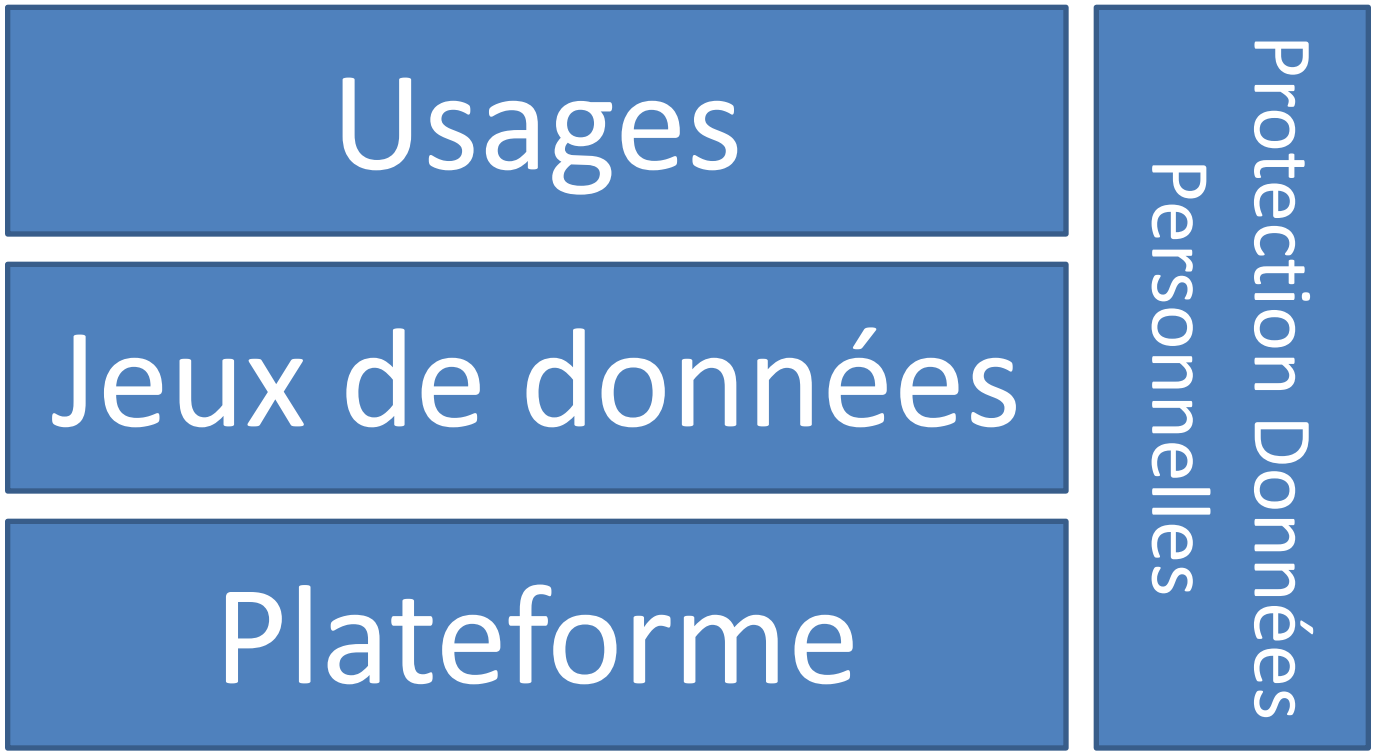
Cinequant, Data Publica, EDF, ESRI,
Hurence, INRIA, Institut Mines Telecom,
La Poste, Orange, Veolia

Mutualisation des données

XDATA

- XData = Cross Data
- En croisant des données d'origine diverses, on peut générer de nouveaux usages et de nouvelles opportunités
- Faire la démonstration de cette idée
  - Jeux de données
  - Usages
  - Plateforme
  - Méthodes

Eléments


Usages
Jeux de données
Plateforme
Protection Données Personnelles

Thèmes généraux

- Mouvement des individus
  - Densité
  - Déplacements
- Consommation de services par les individus et les entreprises
- Echelle de temps
  - Heure/Jour/Semaine/Mois/Année
- Echelle d’espace
  - Pays/Région/Departement/Commune/Quartier

Démarche

- Chainage avant
  - Identifier des jeux de données, puis trouver des usages
- Chainage arrière
  - Identifier des usages et en déduire les jeux de données

- Profiter de cette approche double pour enrichir notre liste d'usage et notre liste de jeux de données

Type d'usages et d'applications

- Catégories d'usages
  - Usages internes pour les participants
  - Usages externes (Clients à identifier)
- Catégories d'usages
  - Applications
  - Analyses
  - Dataviz

Plateforme technique

- Une plateforme pour
  - Partager les données
  - Partager des composants
- Plateforme matérielle
  - Parallélisme
- Plateforme logicielle
  - Environnement Hadoop, bases de données
  - Dataviz, analyse, prédiction

Plateforme commerciale

- La plateforme XData peut être la préfiguration d'une place marché de mutualisation de données et de technologies associées
- Partage de données
- Partage d'outils de traitement de données
- Modèles économiques de mutualisation

Protection des données personnelles

- Un élément clé du projet
  - Voir défis et difficultés
- Conflit classique open data vs. protection vie privée
- Prise en compte à l'initiation du projet
- Méthodologie et stratégie d'anonymisation
- Consultation de la CNIL

Présentation des partenaires

XDATA

- Cinequant
- Data Publica
- EDF R&D
- ESRI
- Hurence
- INRIA (Indes, Privatics & Zenith)
- Telecom ParisTech WebDB
- La Poste, Dides
- Orange
- Veolia, Veri

XDATA

Ouverture

- XData est un projet ouvert
- Nous invitons des partenaires à nous rejoindre pour apporter expertise, technologie et données

Infrastructure

XDATA

Applications actuelles

- **UI Analyse des données en OLAP**
- **UI Analyse des migrations entre départements**
- **Etc.**

- un système d'exploitation CentOS version 64 b
- 12 à 16 cœurs CPU
- 72 à 96 Gigas de RAM par serveur
- 2 x 2 To de disque SATA par serveur en DAS
- Un accès Internet montant et descendant important

Idéalement une connectique réseau de 10 gigabits / seconde

Distribution Hadoop + Storm + Elastic Search

- Outils de l'écosystème
- **PIG, Mahout** et les outils disponibles de la distribution (Hive/Impala, HBase)…
- HDFS
- Map Reduce
- Installation Administration
- Monitoring
- Access Rights + Kerberos

Outils de restitution

- Apache Servers applicatifs
- Kibana
- Elastic Search UI

Outils de sécurité

- Serveurs Kerberos Authentification Gestion des clés de sessions cryptées)

Hadoop + ES cluster + Storm: 4 nœuds de gestion + 20 workers Hadoop / Storm / ES + 4 machines edge (4 cœurs, 32G RAM, 2 teras SATA)

Jeux de données

- Grands groupes
  - EDF
  - La Poste
  - Orange
  - Veolia
- Réseaux sociaux (Telecom ParisTech)
- Open Data (Data Publica)

Applications témoin

- Objectif :
  - Aider à sortir des données
  - Valider le processus avec la CNIL
  - Produire des candidats d'usage
- Applications
  - Zones de chalandise
  - Consommation quartier
  - Tourisme
  - Fragilité des quartiers
  - Pression immobilière

Le Défi de la Protection des données Personnelles

- Certains principes peuvent paraitre contradictoires:
  - ***Recommandation CNIL***: la minimisation des données collectées (seules les données nécessaires au fonctionnement du service doivent être recueillies)
  - ***L'idée de l'Open Data est à l'opposé*** : recueillir le maximum de données, puis voir éventuellement ce qu'on peut en faire après.

Premiers résultats

- Premiers jeux de données
- Maitrise du processus protection des données privées
- Première version de la plateforme
- Premières applications
- Début d'ouverture à de nouveaux partenaires

XDATA

Premier Bilan

- Bonnes surprises
  - Un gros consortium assez agile
  - On arrive rapidement à faire des premiers démonstrateurs
- Mauvaises surprises
  - Faire sortir des données est un exercice long et compliqué
  - Le problème protection des données personnelles est un problème complexe

Objectifs deuxième phase

- Plus de données
- Démonstration anonymisation
- Nouvelle version de la plateforme
  - Technique et business
- Nouveaux usages
  - Analyses, Dataviz, Applications
- Plateforme ouverte