Data Lake pour la génomique

Ludovic Legrand

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

lipme

cnrs
dépasser les frontières

Data Lake

Définition

Un Data Lake est un entrepôt centralisé qui permet de stocker toutes vos données structurées ou non et quelque soit le volume

Volume

- Grande capacité de stockage
- Passage à l'échelle

Variété

- Pas de limite sur les formats ou la taille des fichiers
- Pas de contraintes sur la structure pour ne rien perdre

Vélocité

- Ingestion des données en batch et/ou en stream performante
- Pas de contraintes sur la structure pour être performant

Data Lake

Coté bioinformatique

Spécificités

- Vélocité faible à moyenne en génomique
  - Vélocité en augmentation coté phénotypage
- Métadonnées difficile à automatiser
- Données souvent semi-structurées

Fonctionnement/organisation

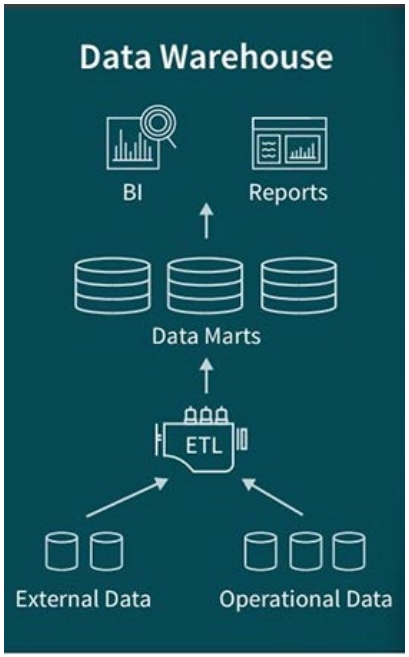
- Equipe restreinte
- Beaucoup de compétence à acquérir
- Moyens financiers limités et intermittents

Besoins

- Identifier les données à mettre dans un Data Lake
- Identifier les workflow intéressant en Big Data
- Identifier les questions pouvant bénéficier du Big Data

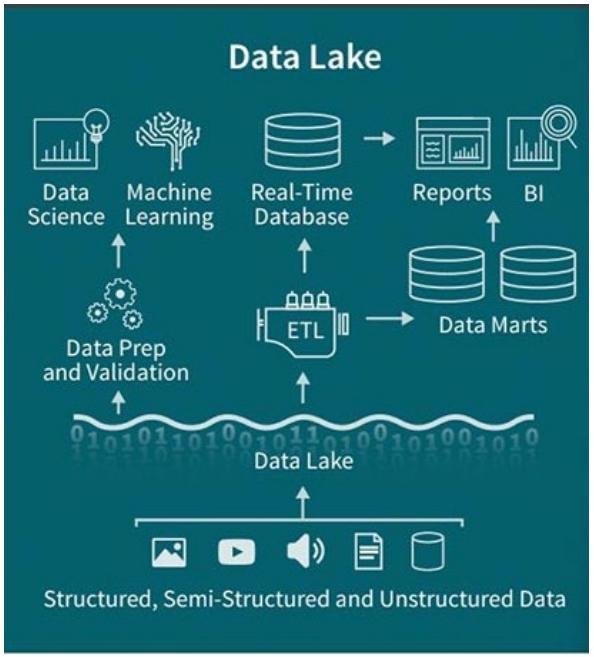
Data Lake

Histoire



Schema-on-write

- Coût élevé et passage à l'échelle difficile
- Format des données limité



Schema-on-read

- Coût raisonnable et passage à l'échelle
- Pas de limitations sur le format des données

Data lake

Pour quoi faire ?

Centralisation

- Standardiser les métadonnées et les données
- Faciliter la recherche de données
- Automatisation des analyses

Exhaustivité

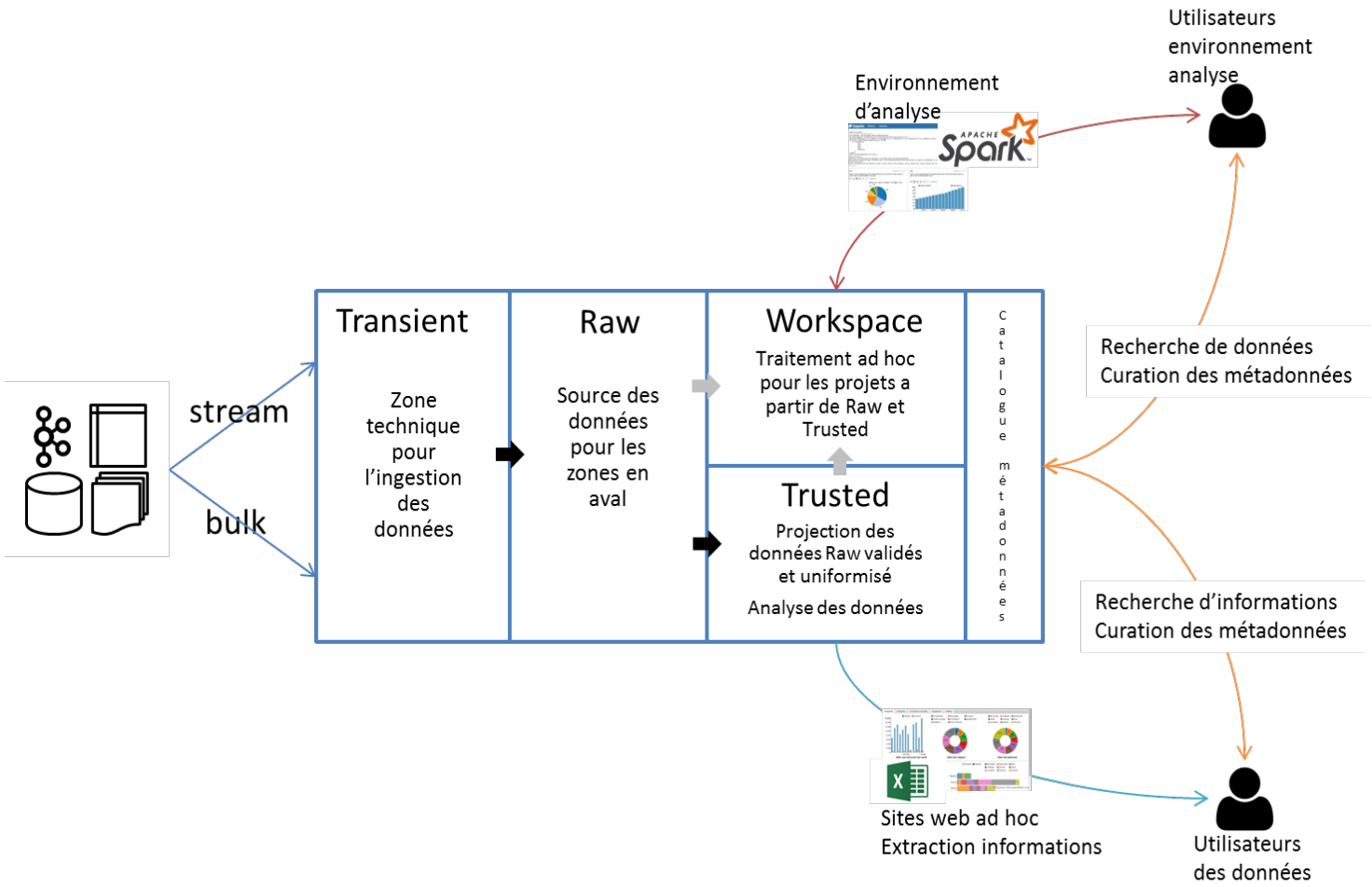
- Faire de meilleurs modèles
- (Ré)analyser des données
  - Nouvelles techniques/données
  - Nouvelles questions

Intégration de données

- Faciliter l'intégration et l'agrégation des données
  - Données formatées
  - Quantité moins limitante
- Machine Learning

Data lake pour la génomique

Structure visée



Organisation des données

En zones

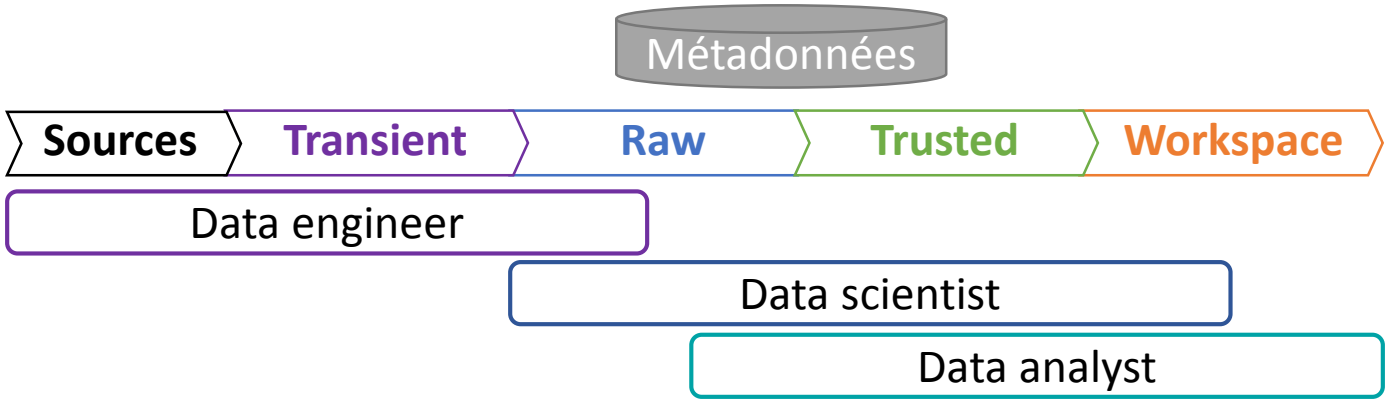
- Format plus ou moins structuré
- Données plus ou moins de confiance
- Droits d'accès
- 4 zones retenues pour cette première version

En répertoire

- Ajouter de l'information
  - Type de données, Espèces …
  - Date, heure, timestamp
- Faciliter le traitement des données

Structuration des données en zones

Zones et rôles



Data engineer

- Gestion des sources et des traitements sur les données Raw
- Compétences en développement/administration

Data scientist

- Fouille des données brutes et préparation des données Trusted
- Compétences en développement/statistiques

Data analyst

- Exploitation des données Trusted avec ses compétences métiers
- Compétences métiers + développement/statistiques

Sources de données

Définition

Connecteur permettant d'ingérer les données et les métadonnées dans le Data lake

1. Interrogation d'une source de données
2. Récupération des données
   - Contrôle d'intégrité
   - Agrégation des données stream
3. Récupération des métadonnées
   - Parsing et validation
   - Métadonnées minimales

- Développement *ad hoc*
- Maintient du code en production

Sources de données

Sources actuelles

	Séquences	Résultats	Phénotypage	Connaissances
Format	Texte et binaire	Texte	Excel/Texte	Excel/Texte
Données	Structurées	Semi-structurées	Semi-structurées	Semi-structurées
Métadonnées	Structurées	Semi-structurées	Semi-structurées	Semi-structurées
Volumétrie	Forte	Moyenne	Faible	Faible
Ingestion	Batch	Batch et Stream	Batch	Batch
Nombre de sources	<10	~10	<10	<10

Sources de données

Sources futures

	Images	IoT
Format	binaire	Texte
Données	Non-structurées	Structurées
Métadonnées	Structurées ?	Structurées ?
Volumétrie	Forte	Faible
Ingestion	Batch	Batch et Stream
Nombre de sources	<10	>10

Structuration des données en zones

Transient

Zone technique fortement couplée aux sources de données et servant de sas avant l'entrée dans le Data Lake

Métadonnées

- Qualité minimale
- Extraction vers un format standardisé

Données

- Agrégation par unité de temps pour les données Stream
- Contrôle d'intégrité pour les fichiers en batch

Défis

- Equilibre entre performance et qualité
  - Trop de contrôle pourrait générer trop d'erreur et freiner l'ingestion des données
  - Trop de laxisme entrainerait l'ajout de données inexploitables dans le Data lake
- Gestion des erreurs

Structuration des données en zones

Raw

Zone centrale du Data lake conservant l'ensemble des données sous un format brute

Métadonnées

- Indexation des métadonnées dans le catalogue

Données

- Formatage léger possible pour simplifier l'utilisation de Spark
- Analyse des données pour valider la qualité et enrichir les métadonnées
  - Recherche de contaminants
- Identifier et protéger les données sensibles
  - Droits limités sur des données confidentielles

Défis

- Garantir que la donnée est correctement décrite et référencée
- Forte volumétrie

Structuration des données en zones

Zone des données validées

Dans cette zone les données brutes auront été normalisées et enrichies dans des fichiers parquet

Métadonnées

- Enrichissement avec les connaissances métiers
  - Synonymes,
- Normalisation
  - Ontologie
- Traçabilité des opérations

Données

- Projection sous le format parquet
- Normalisation des données
- Versionning des données

Défis

- Automatiser les traitements

Structuration des données en zones

Workspace

Zone d'exploitation des données par les utilisateurs.

Métadonnées

- Traçabilité des opérations
- Curation *ad hoc* pour une projet

Données

- Traitements ad hoc depuis différents outils
  - Spark-shell, jupyter ….

Défis

- Faciliter l'utilisation et l'accès aux données/métadonnées par des utilisateurs avec des profiles divers

Structurer de l'arborescence

Intérêts

Performance

- Eviter de scanner des fichiers inutiles
- Certains outils peuvent utiliser le chemin pour restreindre des requêtes
- Découpage des gros fichiers en partie plus petites (partitionning)

Automatisation

- 1 type de données/fichier par répertoire permet de simplifier les scripts de traitement
- Une structure cohérence facilite la reconstruction des chemins

Sécurité

- Facilite la gestion des droits par répertoires

Structure de l'arborescence

Exemple

Structure « statique »

- raw (Zone)
- genome (Type de données)

Structure « dynamique »

- species, genotype, date
  - Dépendant du type de données
- uuid
  - Présentent partout
  - Identifiant unique
- Spark peut automatiquement ajouter en colonne ces informations

```
raw
|-- genome
|   `-- species=Helianthus annuus
|       `-- genotype=XRQ
|           `-- date=2018-05-05
|               `-- <uuid>
|                   `-- HanXRQr2.0-SUNRISE-2.1.genome.csv
```

Défis

Technologies

- Evolution rapide
- Complexité

Alimentation en données

- Connecter et maintenir les sources de données
- Automatiser au maximum la préparation des données
- Structuration/règles

Organisation

- Structuration et gouvernance des données/métadonnées
- Automatiser et faciliter la gestion des métadonnées et la traçabilité
- Facilité la rechercher les données

Défis

Sécurité

- Protéger les données sensibles
- Gérer les accès au données

Utilisation

- Mettre le Data Lake au centre de l'organisation
- Faciliter l'accès par des interfaces ad hoc au type d'utilisateur
- formation