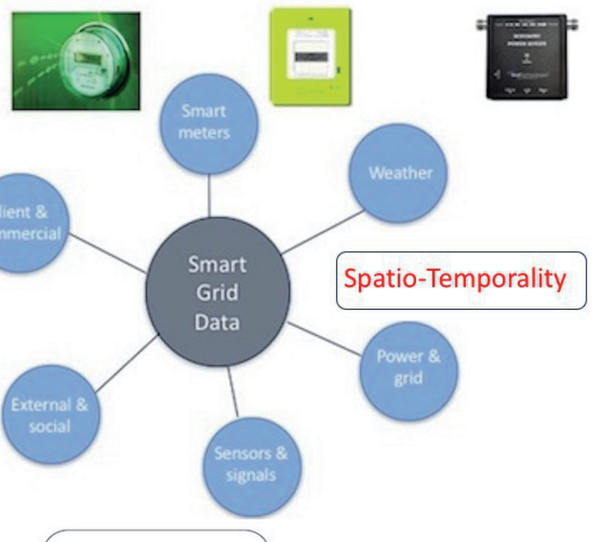
Workload based datasets placement in Big Data Ecosystems

Asma Zgolli,
Christine Collet & Houssem Chihoub

Traitement de Données et de Connaissances à Grande Echelle - HADAS

Smart grid Big data ecosystem

High level of Variety
- Various datasets
- Different formats, structures, constraints, quality, access rules, privacy levels, usage,...



Velocity
- More frequent meausrements
- Data in motion

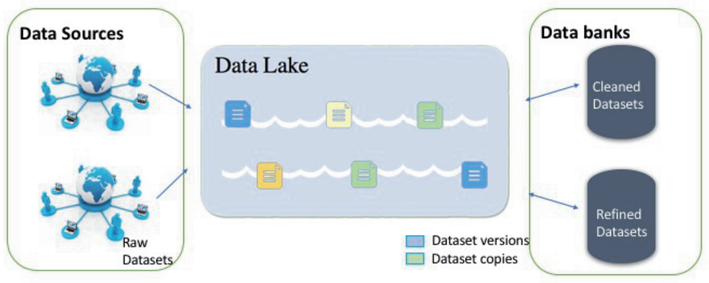
Data Veracity
- Multiplication of sources
- Missing values
- Unreliable measurements

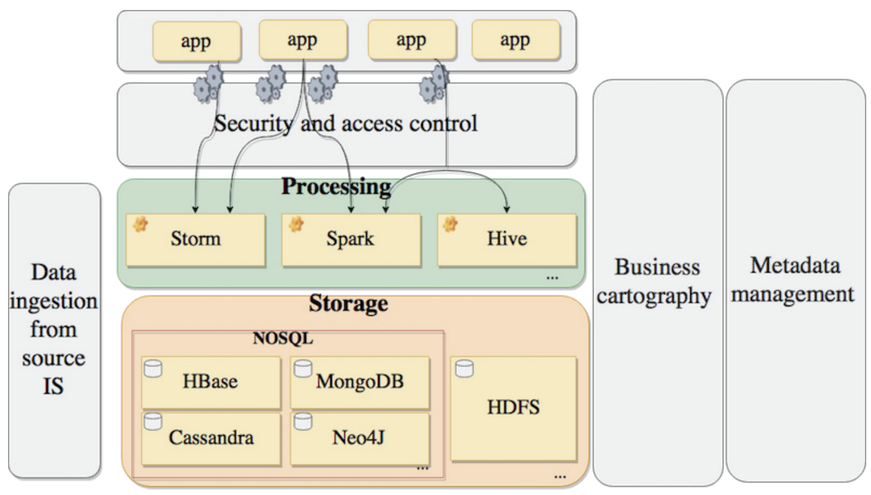
Massive Volumes
- Millions of sensors
- Millions of smart meters
- Historical data

*In one of their white papers, HP show how they manage with their solution Vertica **40+ M** of meter data with measurements every 10 min and a total of **22,5 trillions** measurements*

Smart Grid Data Deluge !

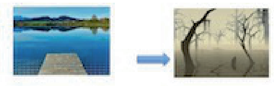
Data @ Rest:
- Ingested from **# sources**
- Coming in **# forms**
- Having **# uses**





data **storage systems**
data **processing systems**
types of **workloads** & **business rules**

***Complexity** of the Data ecosystem!*

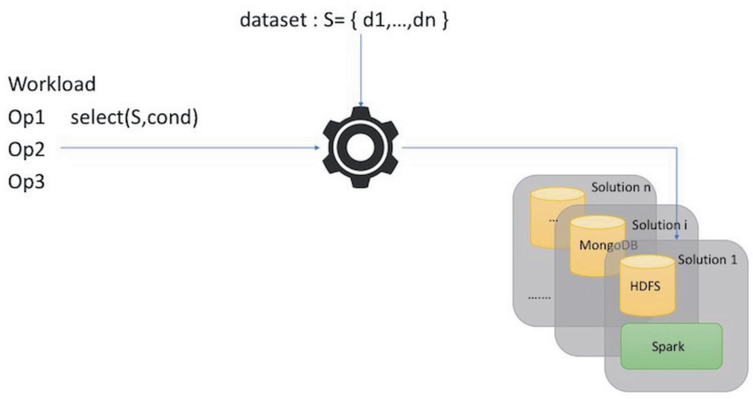


Placement of datasets

Objective :
- Effective **placement** and **processing** of datasets

Constraints :
- Efficient workload processing
- Compliance with business environment & the data ecosystem

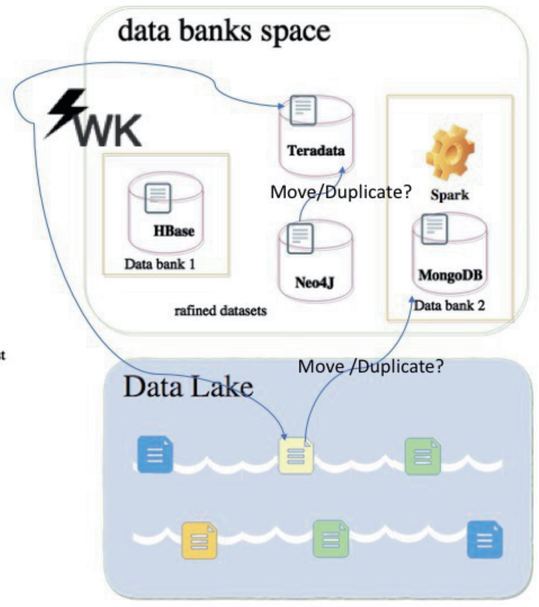
Example :



Where to place datasets ?
- in which databank ?
- in the lake ?

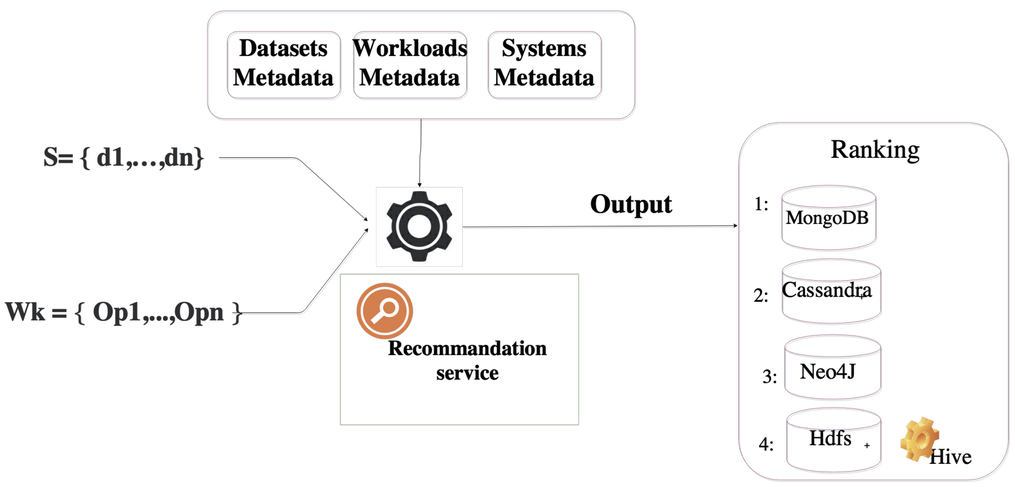
How to place datasets ?
- **duplicate / move** the dataset

When to place datasets ?
- **before processing** (on the input datasets)
- **after processing** (on the result, intermediate datasets)



Workload based placement approach

Recommandation for dataset placement



Recommandations based on metadata about :
- Characteristics of the target systems
- Efficiency of the processing of the <wk, s> in those systems

Metadata

Workloads and applications

- **Workloads** metadata
  - Statistics
  - Plans
  - Patterns
- Business rules
- Annotations

Datasets

- Schema
- Statistics
- Lineage

Data Management Systems

- Data model
- Data partitioning
- Access API & Languages
- Data storage
- Systems configuration & Distribution

Placement creteria

Feasibility

- **Identification of the suitability** of data storage & processing systems properties (data model , query model , partitioning ….) with datasets and workloads metadata

Performance

- **Response time** :
- Time needed for the execution of a workload wk on a data store DS
- Based on cost estimation given workload , data stores and datasets characteristics

- How **scalable** is **performance** ?
- is it crucial to the workload ?

Conformity

- With the **business rules** of Smart Grid Data Ecosystem