Semantic COMPSs:
Distributing data lakes and queries

•••

Ramon Amela Milian
Pol Alvarez Vecino

Overview

- Goal
- Data
- Architecture
- Queries
- Demo

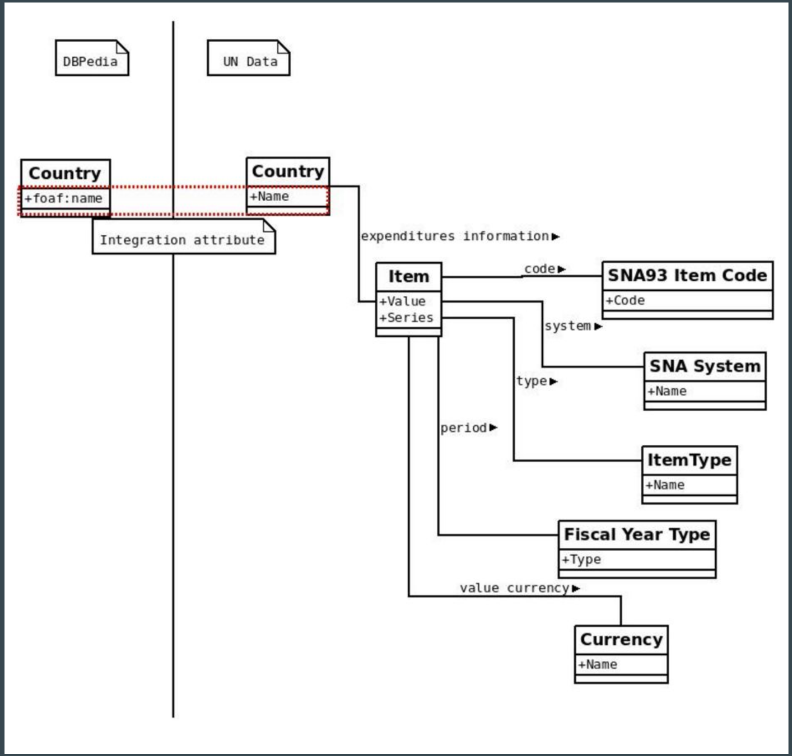
Goal

- Implements a distributed RDF data lake
- Data integration automatization handling semantic enrichment and updates
- Distributed queries over the whole system with minimal requirements

Data

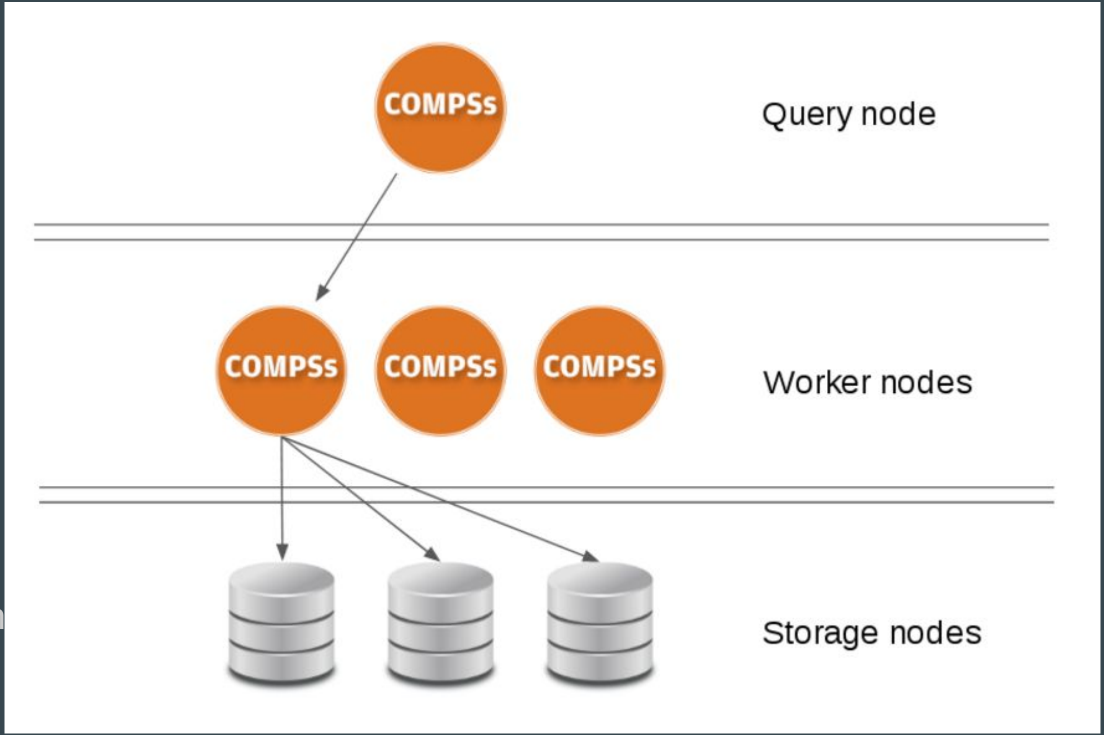
- United Nations data
  - Tabular data
  - Format: XML
- DBpedia
  - Graph data
  - RDF format

- Integration
  - Country name



Architecture

- Query Node
  - Processes user queries
  - Distributes subqueries
  - Tracks already available data
- Worker Nodes
  - Contain local graph
  - Perform user query in local graph
- Storage Nodes
  - Contain whole data
  - Receive subqueries required to gath user-query data



Queries

- Requirements
  - All objects and subjects type (class) must be present in the WHERE clause

- Dataflow
  - Each type and the relation is retrieved (together with its literals) from all endpoints (storage nodes)
  - Data is consolidated and entity resolution is done
  - Data is inserted into compute nodes local graph
  - Query is run on the compute node data graph
  - Result is returned to the Query node and presented to the user

Thanks

- Questions
- Demo