Policy Cloud
Cloud for Data Driven Policy Management

02/07/2020

PolicyCLOUD Technical Overview

Pavlos Kranas (LeanXcale S.L)

Policy Cloud
Cloud for Data Driven Policy Management

PolicyCloud has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870675.

Objective

- PolicyCLOUD: Analytics-as-a -Service facilitating efficient data-driven public policy management

Background

- Facts
  - Increasing use of devices and networks leading to the generation of vast quantities of data
  - Data linking is becoming the norm (e.g. linking new data sources with established data sources)
  - Current approaches in policy making are not evidence-based
  - Mature approaches to analyse and understand the “environment”

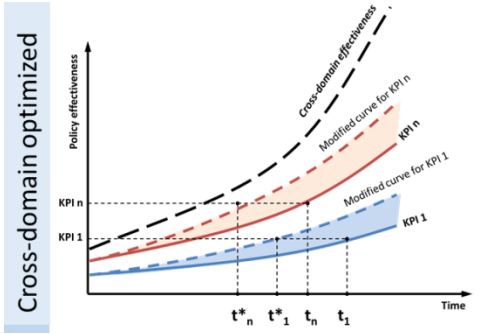
- Goal
  - Creation of efficient and effective policies through data-driven policy management
  - Decision support to authorities for policy modelling, implementation and simulation through identified populations, as well as for policy enforcement and adaptation

Main challenges (1/5)

- A data-driven approach for effective policies management
  - Across the complete data path, including data modelling, representation and interoperability, cleaning, heterogeneous datasets linking, analytics for knowledge extraction
  - Exploit the collective knowledge out of policy “collections” combined with the data from several sources (e.g. sensor readings, online platforms, etc.)

Main challenges (2/5)

- Compilation, assessment and optimization of multi-domain policies
  - Holistic policy modelling, making and implementation in different sectors (e.g. environment, migration, goods and services, etc.), through the analysis and linking of KPIs of different policies that may be interdependent and inter-correlated (e.g. environment)
  - Analysis of (unexpected) patterns and policies relationships
    - Identification of effective KPIs to be re-used and non-effective ones (including the causes for not being effective) towards their improvement



Main challenges (3/5)

- Data management techniques across the complete data path
  - Meta-interpretation layer for the semantic and syntactic capturing of data properties and their representation
  - Data cleaning to ensure data quality and coherence including the adaptive selection of information sources based on evolving volatility levels (i.e. changing availability or engagement level of information sources)

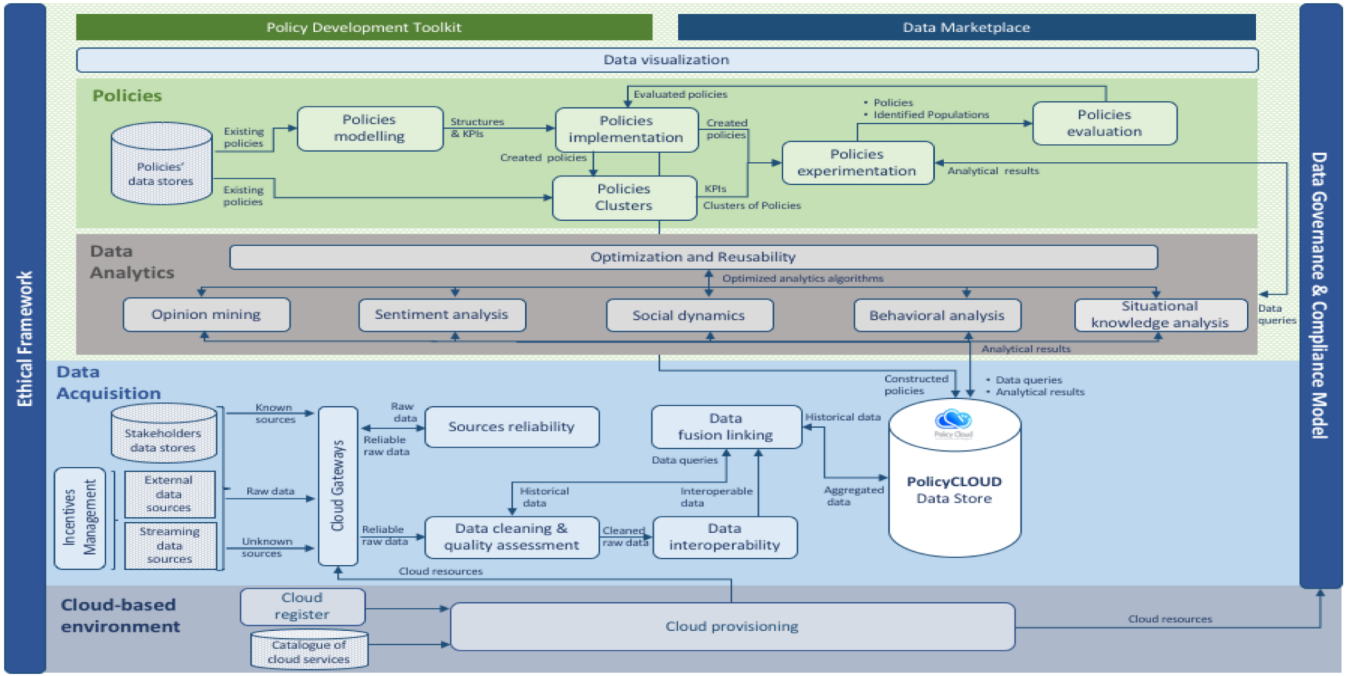
Main challenges (4/5)

- Analytics as a service reusable on top of different datasets
  - Machine and deep learning techniques (e.g. classification, regression, clustering and frequent pattern mining) to infer new data and knowledge
    - Opinion mining, sentiment analysis, social dynamics and behavioral data analytics
  - Technologies that allow analytics tasks to be decoupled from specific datasets and thus be triggered as services and applies to various cases and datasets

Main challenges (5/5)

- Unique endpoint to exploit analytics in different cases
  - Execution of different models / analytical tools on data (e.g. to identify trends, to mine opinion artefacts, to explore situational and context awareness information, to identify sensitives, etc.)
  - Modelled policies (through their KPIs) realized / implemented and monitored against these KPIs
  - Adaptive and incremental visualization enabling the policy lifecycle to be visualized in different ways, while the visualization can be modified on the fly and can enable the specification of the assets to be visualized (e.g. data sources or meta-processed information)

Conceptual architecture



Seamless Analytical Framework

- Baseline Technology firstly introduced and implemented as Proof-Of-Concept in H2020 BigDataStack
- Collaborative work between IBM and LeanXcale
- First prototype already delivered! Its functionality is planned to be extended in PolicyCLOUD

Seamless Analytical Framework

- Modern enterprises use
  - Operational databases for OLTP load
  - Key-Value for IoT data
  - Data warehouses for data analytics
  - Datalakes
  - etc ...

- Need polyglot capabilities

Seamless Analytical Framework

- Nowadays: Data Federation using Spark

Seamless Analytical Framework

- Nowadays: Data Federation using Spark

- BUT:
  - Can be very resource consuming
  - Cannot exploit the specific capabilities of each different datastore

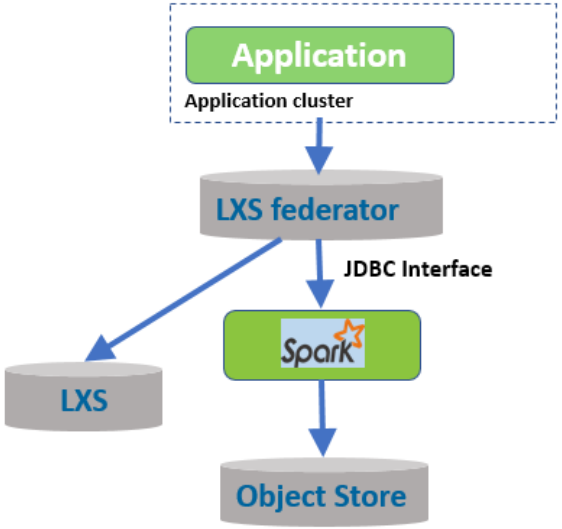
Seamless Analytical Framework – User Story

- Data ingestion in operational datastore (LeanXcale)
- Old data becomes historical, with no modifications
- Data Warehouse to perform analytics on big data volumes
- Distribution of datasets is problematic
  - Data to be retrieved from both stores
  - To be merged in the application level
  - Data consistency considerations when moving datasets

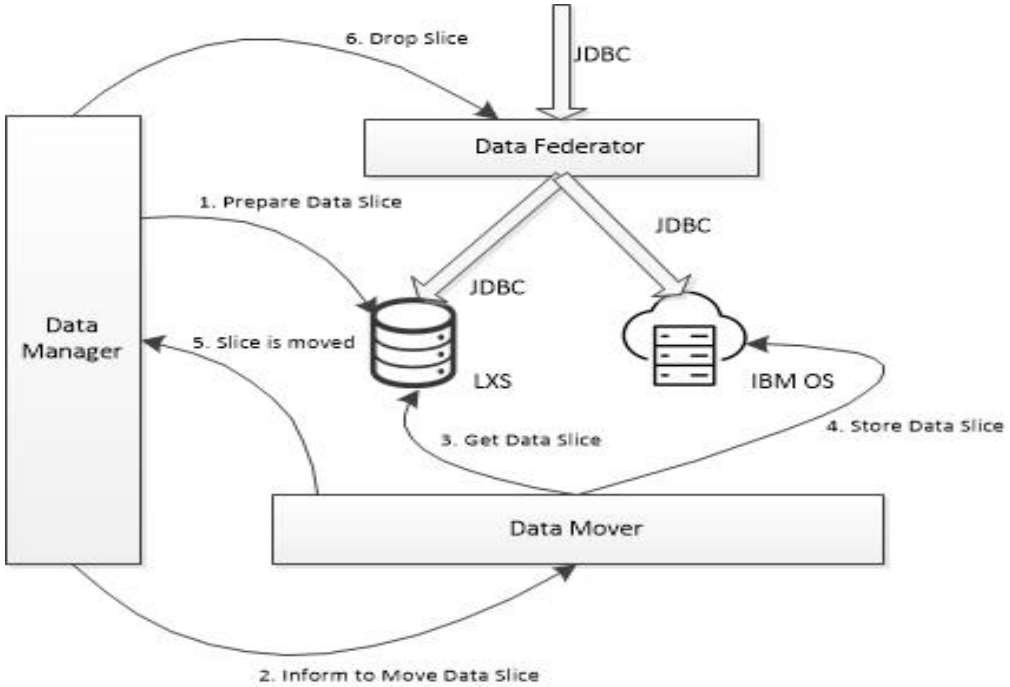
Seamless Analytical Framework – Solution

- Seamless Analytical Framework
  - Federate data coming from two different datastores:
    - HTAP Relational LXS Datastore
    - IBM Object store
  - sharing the **SAME** dataset
  - Single (black box) component that
    - consists of two datastores
    - exploits unique characteristics of each one
    - transparently from the user
    - does not compromise some requirements for the benefits of others

Query Federation



Data Movement High Level

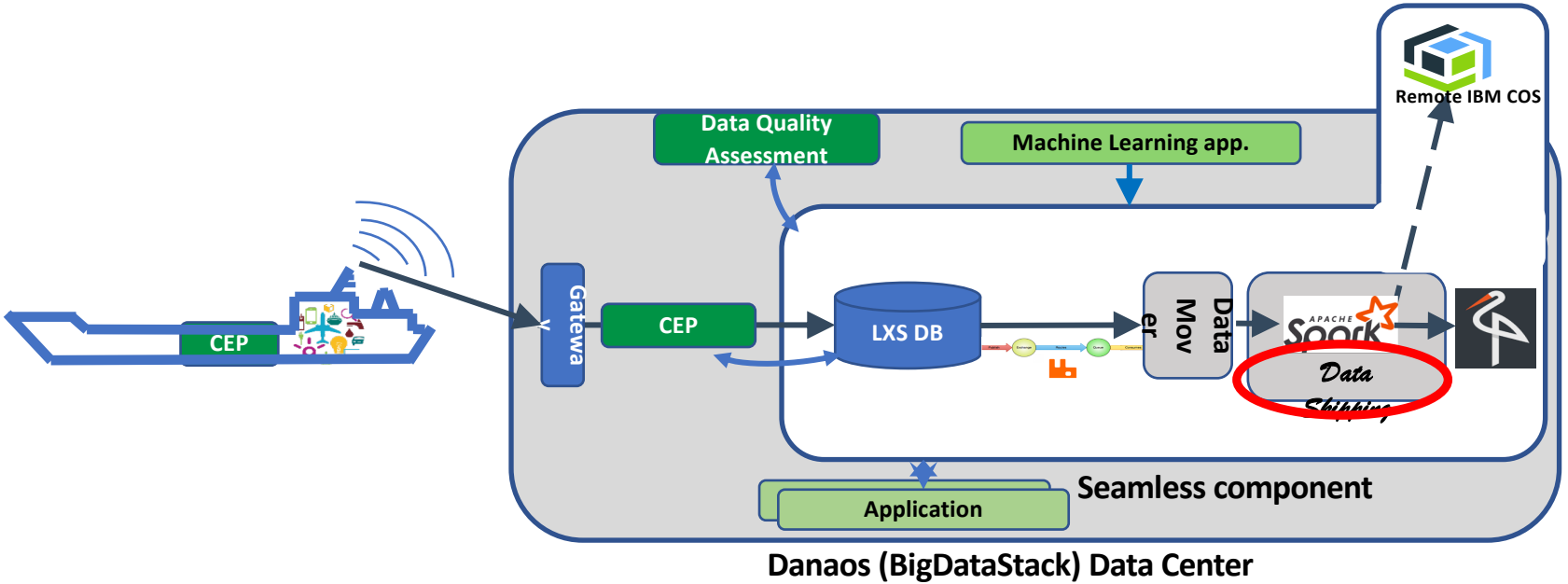


Supported Operations

- Currently supports
  - Full Scan
  - Ordered Scan
  - LIMIT
  - Aggregations
  - Group By Aggregations
  - Ordered Group By Aggregations

- Does not yet support
  - JOIN on fragmented data tables

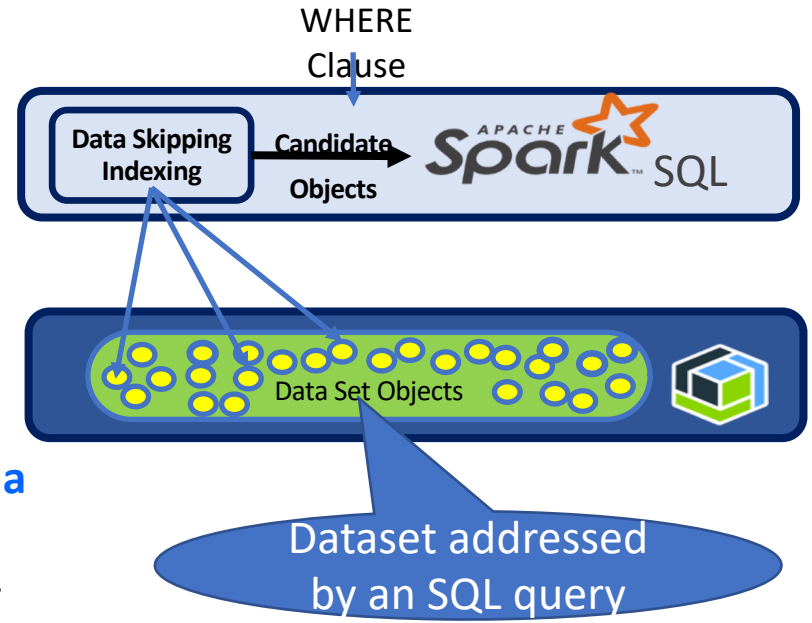
Data Skipping



Data Skipping

- Relevant for SQL queries
- Implemented for Apache Spark SQL
- Up to latest Apache Spark version 3.0
- Standalone technology but also nicely complements the seamless component
- Determine which objects are NOT relevant to a SQL query using a *data skipping index*

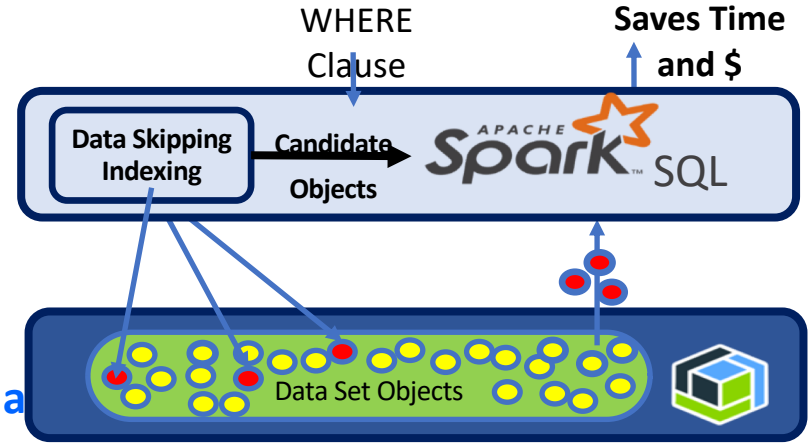
  Stores and indexes tiny summary metadata for each object.
- Skipping over irrelevant objects reduc
  bytes scanned



Query example: retrieve data of violent storms

```
SELECT vessel_code, datetime, longitude, latitude, wind_speed
FROM cos://us-south/…/danaos stored  as parquet
WHERE wind_speed > 30
```

Data Skipping



- **Relevant for SQL queries**
- **Implemented for Apache Spark SQL**
- **Standalone technology but also nicely complements the seamless component**
- **Determine which objects are NOT relevant to a SQL query using a *data skipping index***

  Stores and indexes tiny summary metadata for each object.
- **Skipping over irrelevant objects reduces the bytes scanned**

  Saves time and $

Example: Look for data in violent storm conditions

```
SELECT vessel_code, datetime, longitude, latitude, wind_speed
FROM cos://us-south/…/danaos stored  as parquet
WHERE wind_speed > 30
```

When could we try Data Skipping ….

Joint demo IBM/Danaos at the high visibility in IBM THINK ’19 conference

When could we try Data Skipping …

Joint demo IBM/Danaos at the high visibility in IBM THINK ’19 conference

Data Skipping technology integrated as open beta into IBM Cloud SQL Query

Policy Cloud

Cloud for Data-Driven Policy Management

GET IN TOUCH

www.policycloud.eu

@PolicyCloudEU

PolicyCloud EU

PolicyCloud has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870675.