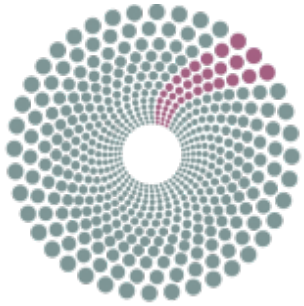
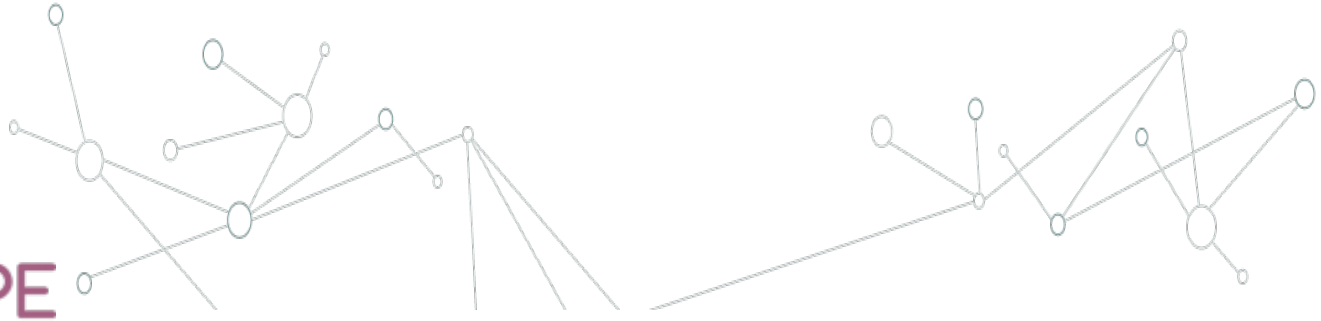


BIG DATA EUROPE

Empowering Communities
with Data Technologies



Pilot SC4

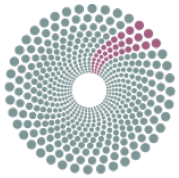
BDE Workshop Brussels
14 Sept. 2017

BIG DATA EUROPE

Empowering Communities with Data Technologies

Objective of the Pilot SC4

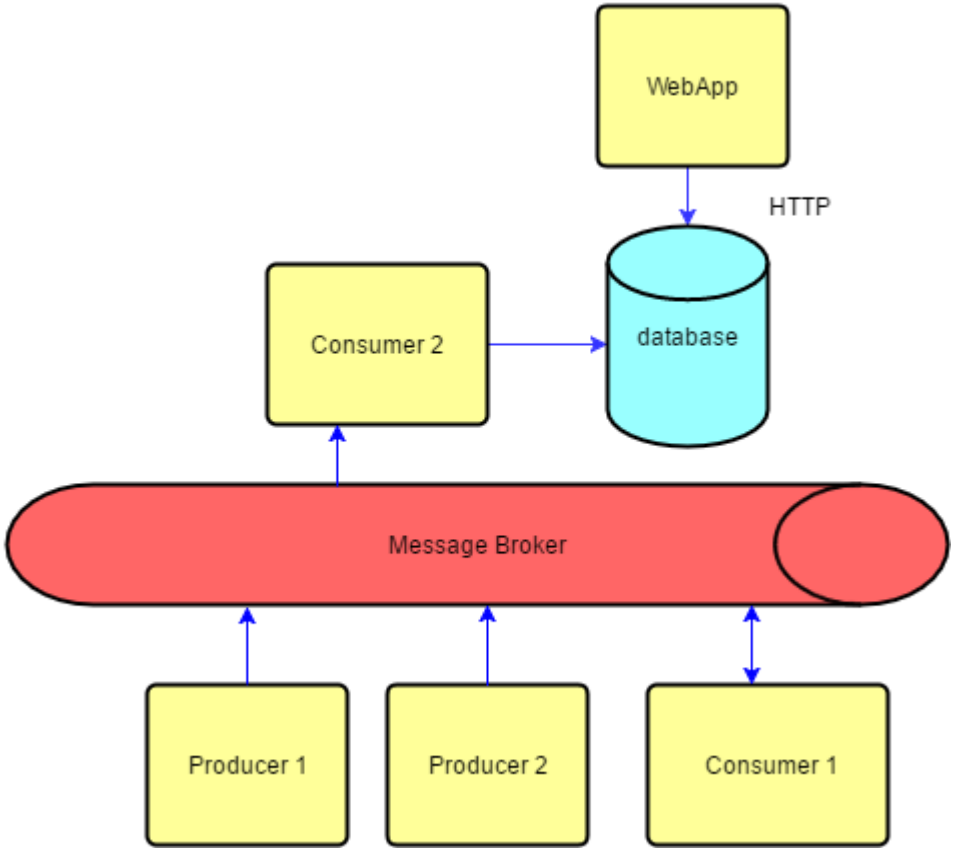
A scalable, fault-tolerant and flexible platform based on open source frameworks that can process unbounded data sets.

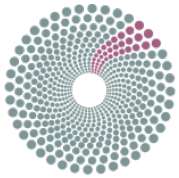


BIG DATA EUROPE

Empowering Communities
with Data Technologies

Microservice Architecture





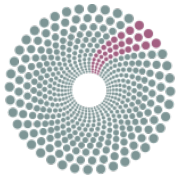
BIG DATA EUROPE

Empowering Communities with Data Technologies

Message Broker

Apache Kafka

Apache Kafka is a high-throughput distributed durable messaging system

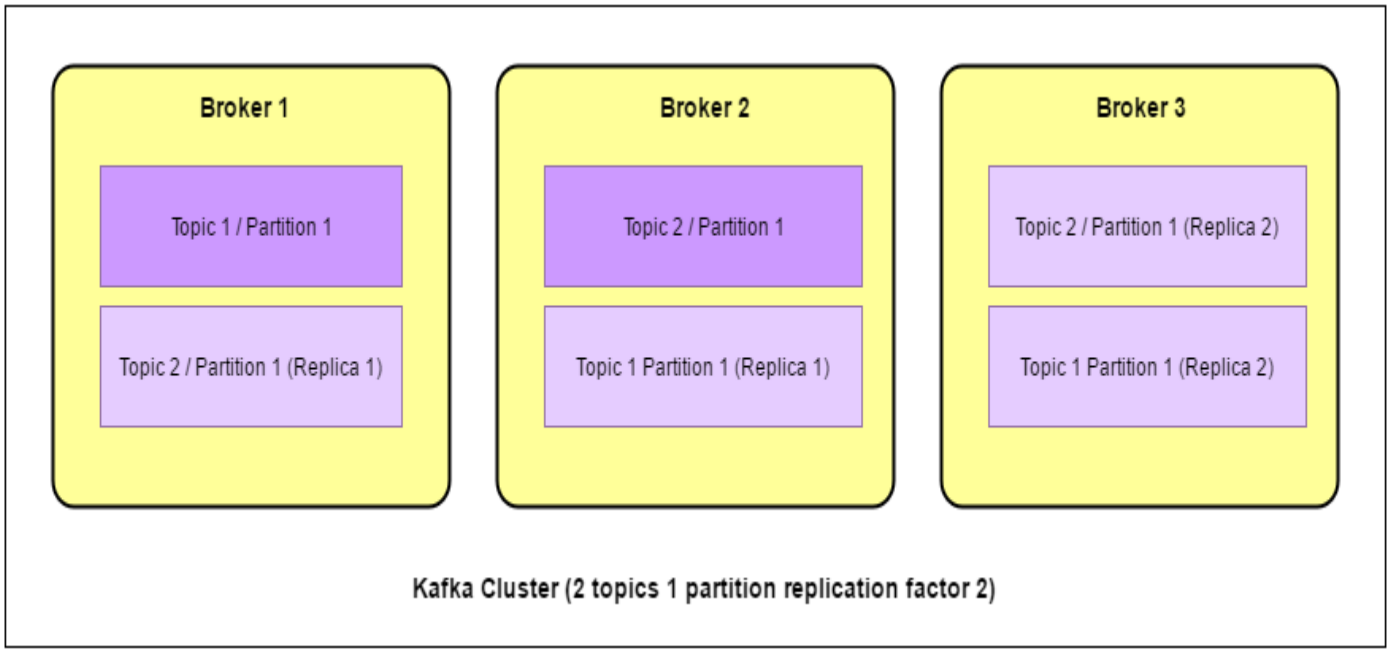


BIG DATA EUROPE

Empowering Communities with Data Technologies

Kafka Cluster

Apache Kafka



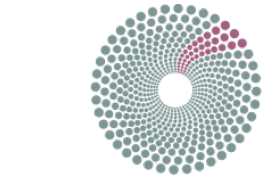
BIG DATA EUROPE

Empowering Communities with Data Technologies

Stream and Batch Processor

Apache Flink

Apache Flink is an open source platform for distributed stream and batch data processing.

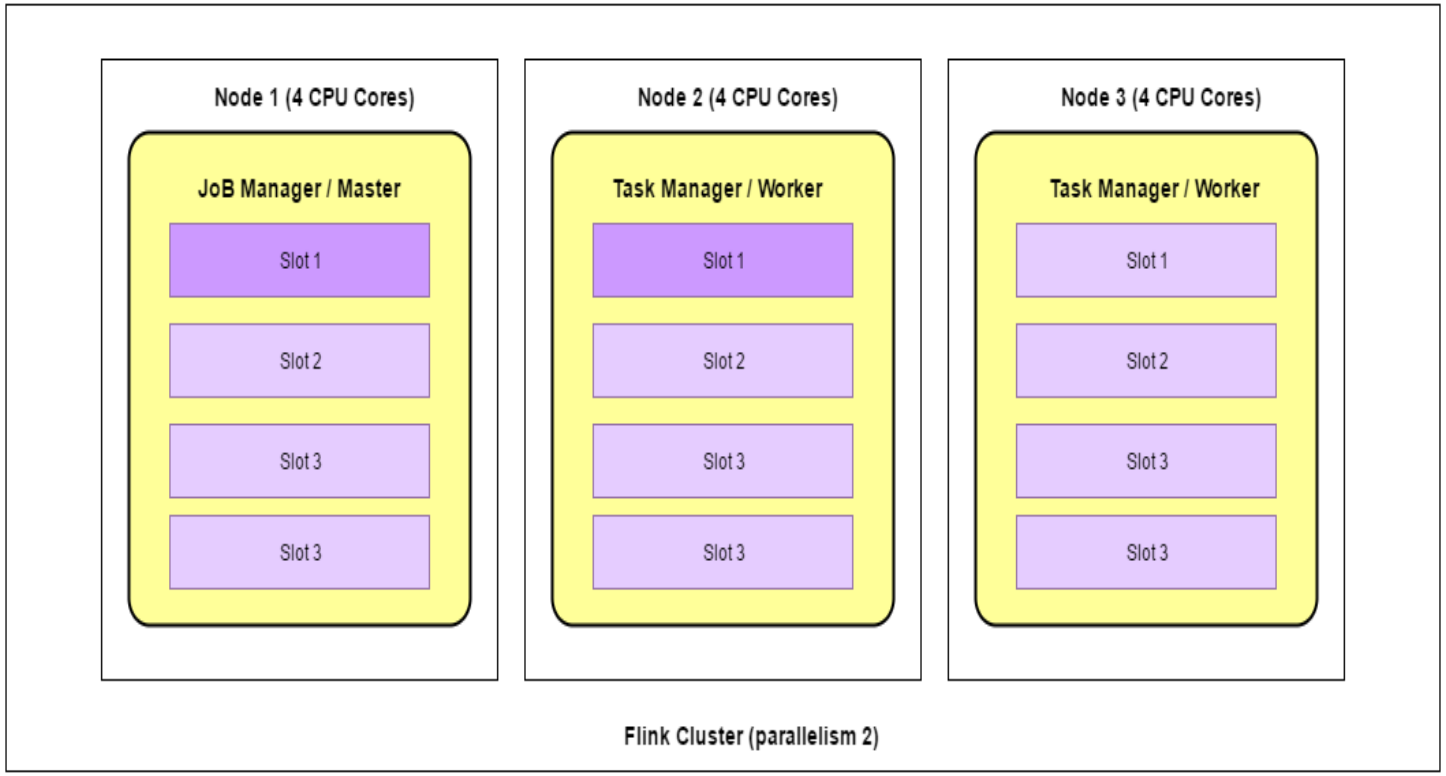


BIG DATA EUROPE

Empowering Communities with Data Technologies

Flink Cluster

Apache Flink

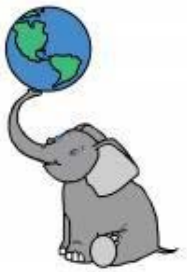


BIG DATA EUROPE

Empowering Communities with Data Technologies

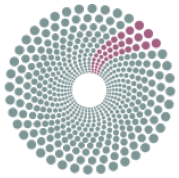
Storage and Indexing

elastic



PostGis is a spatial database that stores the road network data.

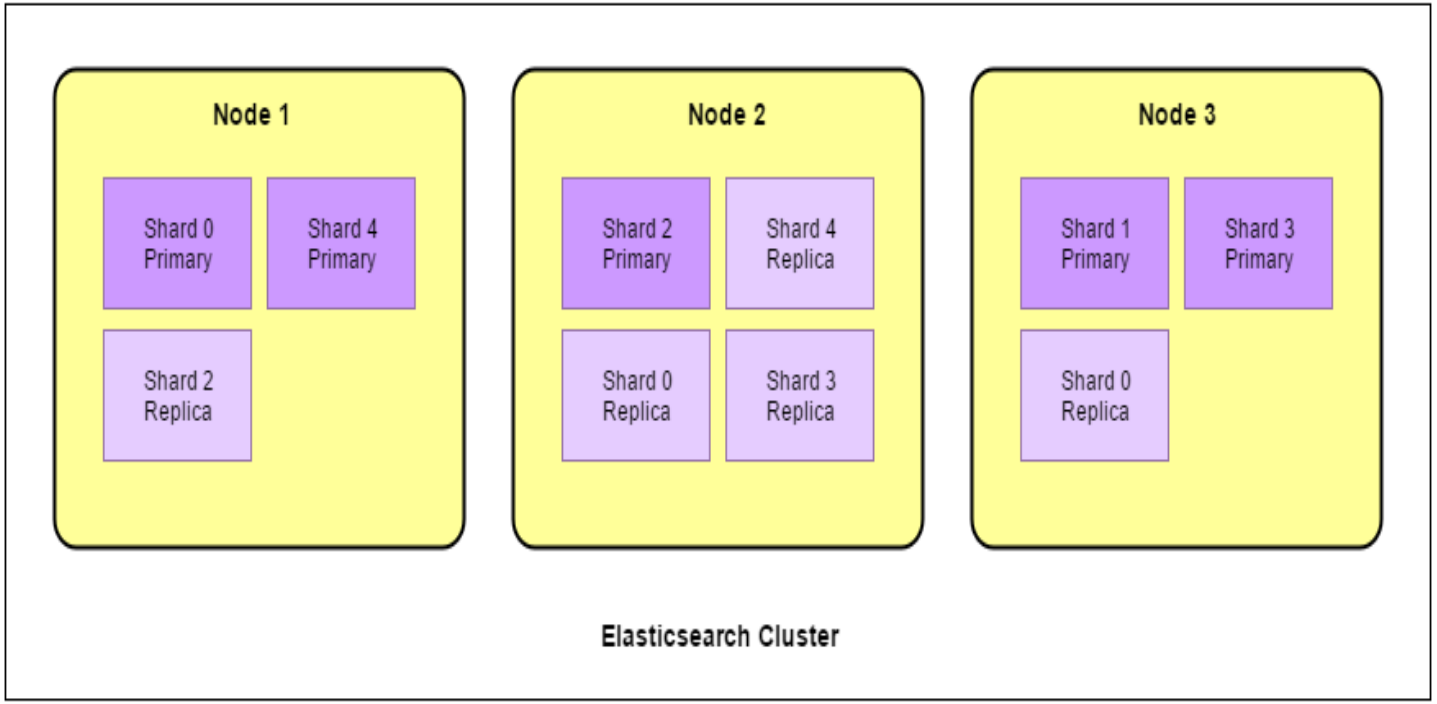
Elasticsearch is a distributed open source document database built on top of Apache Lucene. It stores the result of the workflow.



BIG DATA EUROPE

Empowering Communities with Data Technologies

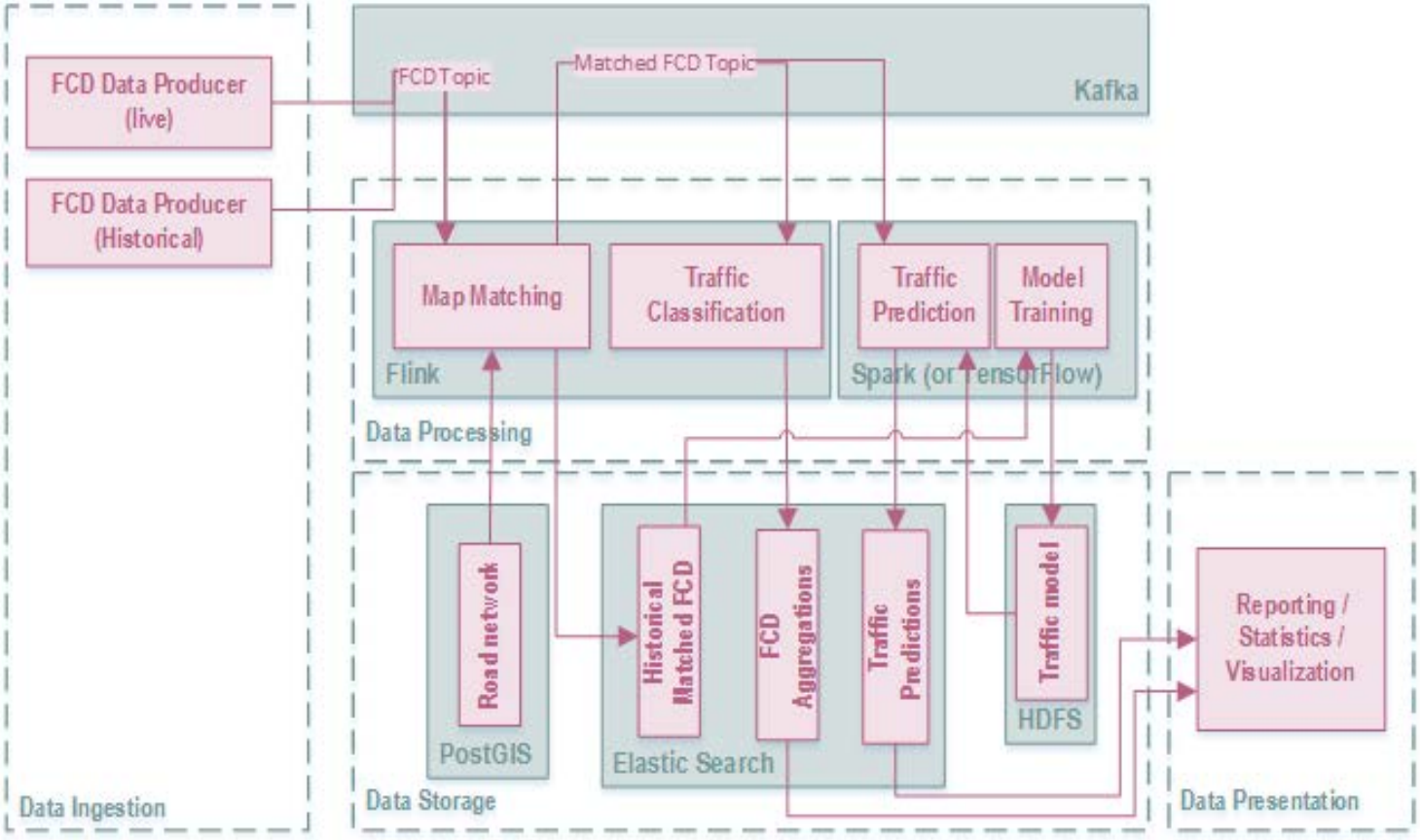
Elasticsearch Cluster



BIG DATA EUROPE

Empowering Communities with Data Technologies

Pilot Architecture

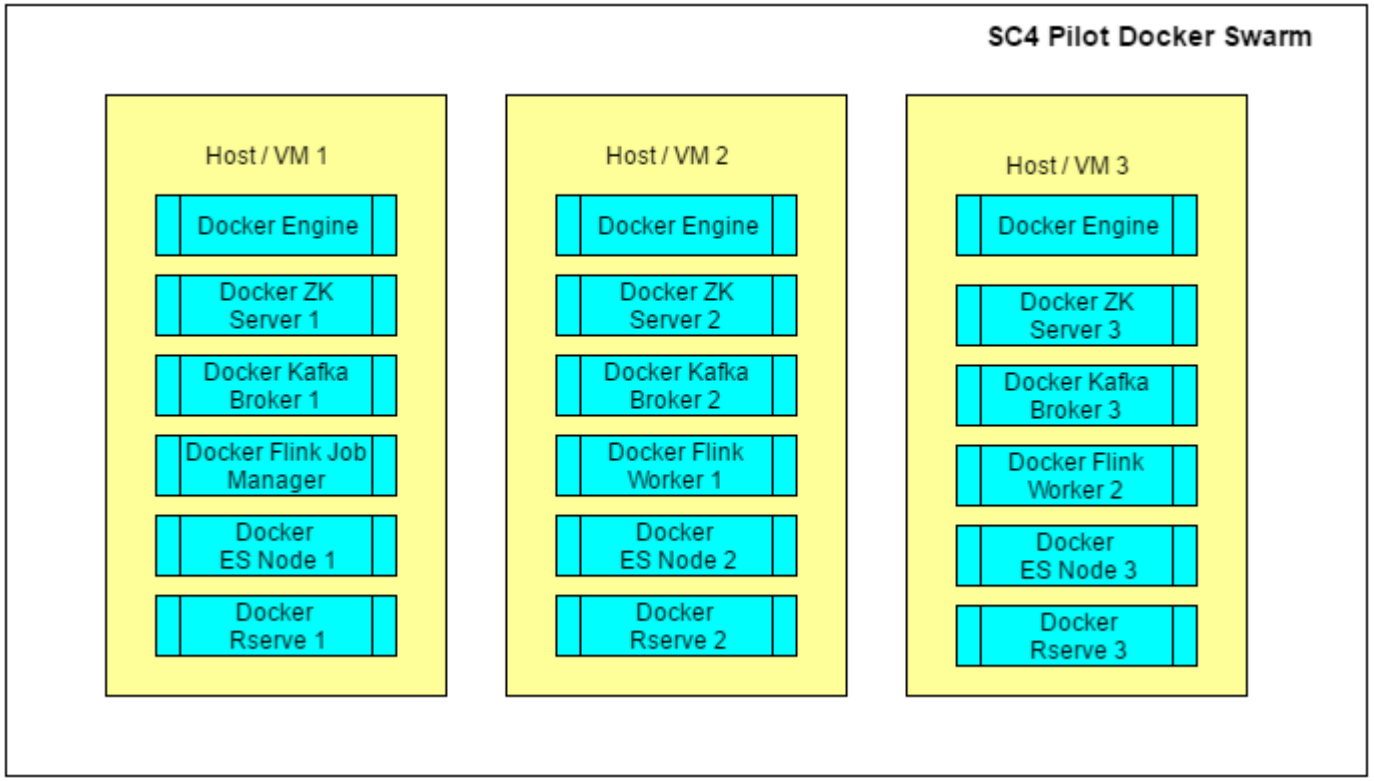


BDE Workshop Brussels
14 Sept. 2017

BIG DATA EUROPE

Empowering Communities with Data Technologies

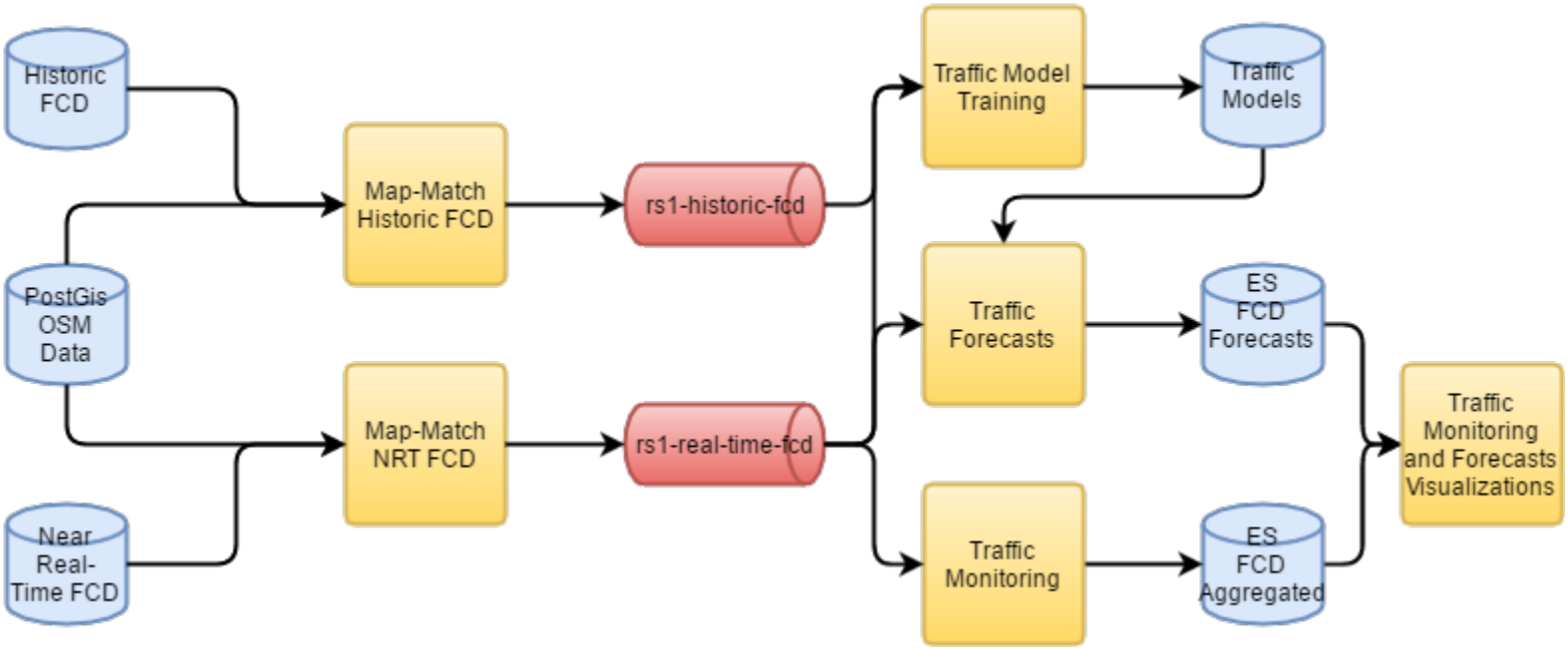
BDE Components

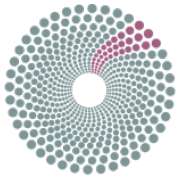


BIG DATA EUROPE

Empowering Communities with Data Technologies

The FCD Pipeline





BIG DATA EUROPE

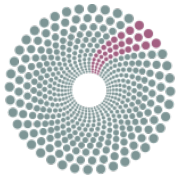
Empowering Communities with Data Technologies

Pilot Cluster

Minimum requirement for fault-tolerance and scalability

- Cluster of 3 nodes (Docker swarm)
- 4 CPU cores x node
- 1 (Flink) worker x node
- 1 (Flink) slot x CPU core

Max parallelism = 12



BIG DATA EUROPE

Empowering Communities with Data Technologies

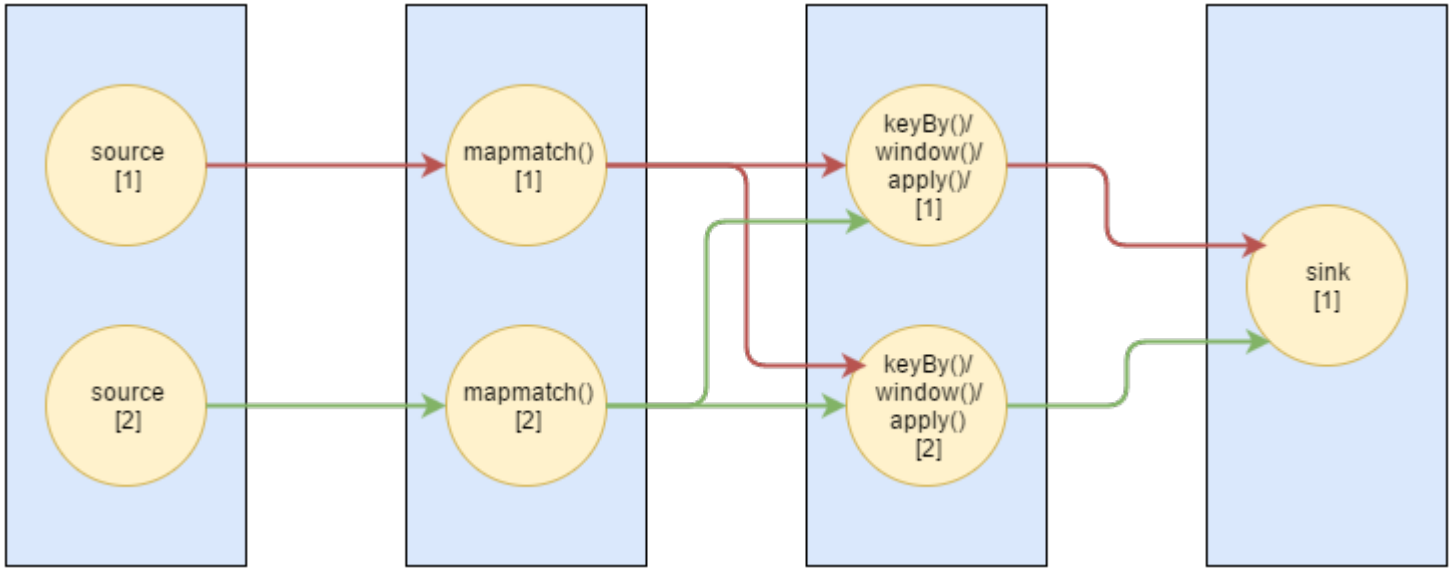
Parallelization: map-match subtasks

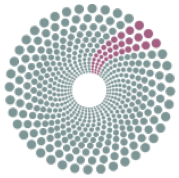
1. source()
2. mapMatch()
3. keyBy()/window()/apply()
4. sink()

The subtasks can be distributed in slots with different parallelism (e.g. from 1 to 12)

Parallelization: map-match subtasks

A slot can process all the subtasks in a pipeline





BIG DATA EUROPE

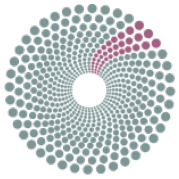
Empowering Communities with Data Technologies

Parallelization: input and output data

device_id	timestamp	lat	lon	speed	orientation	transit

The mapMatch subtask keeps the time order so that the next task keyBy(road_seg)/window(15')/apply(average_speed) will return the correct result within the time window for each road segment.

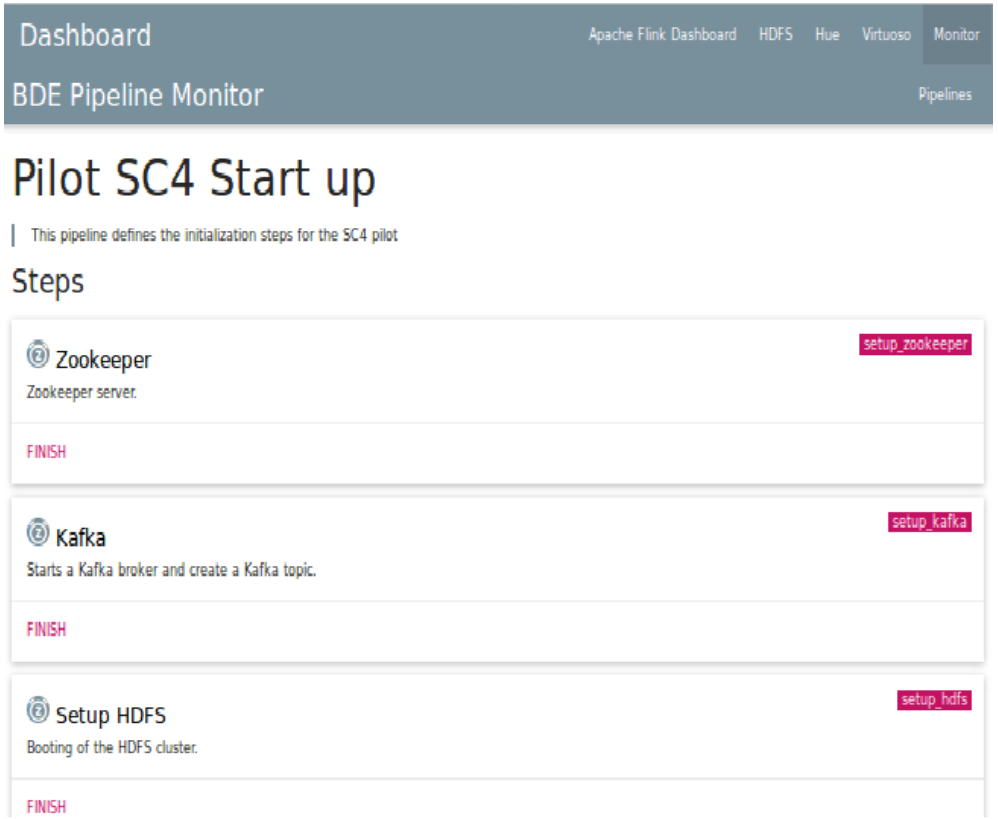
road_seg_id	start_date	num_vehicles	avg_speed



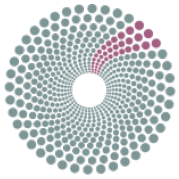
BIG DATA EUROPE

Empowering Communities with Data Technologies

SC4 Pilot Pipeline



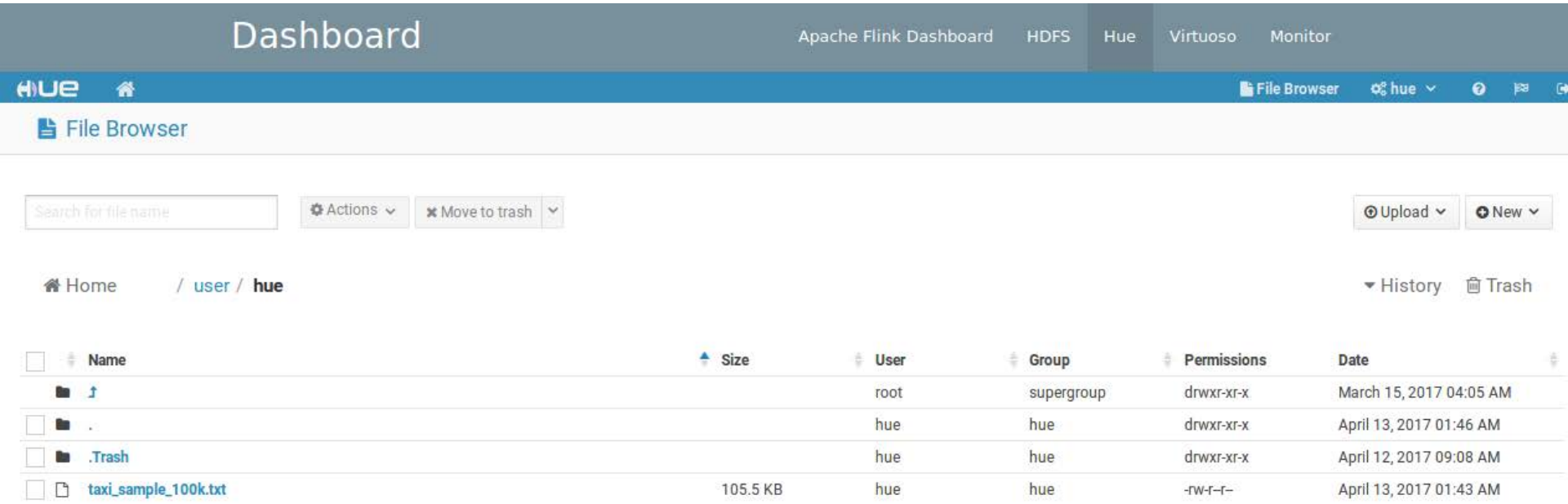
BDE Workshop Brussels
14 Sept. 2017



BIG DATA EUROPE

Empowering Communities with Data Technologies

Data Upload

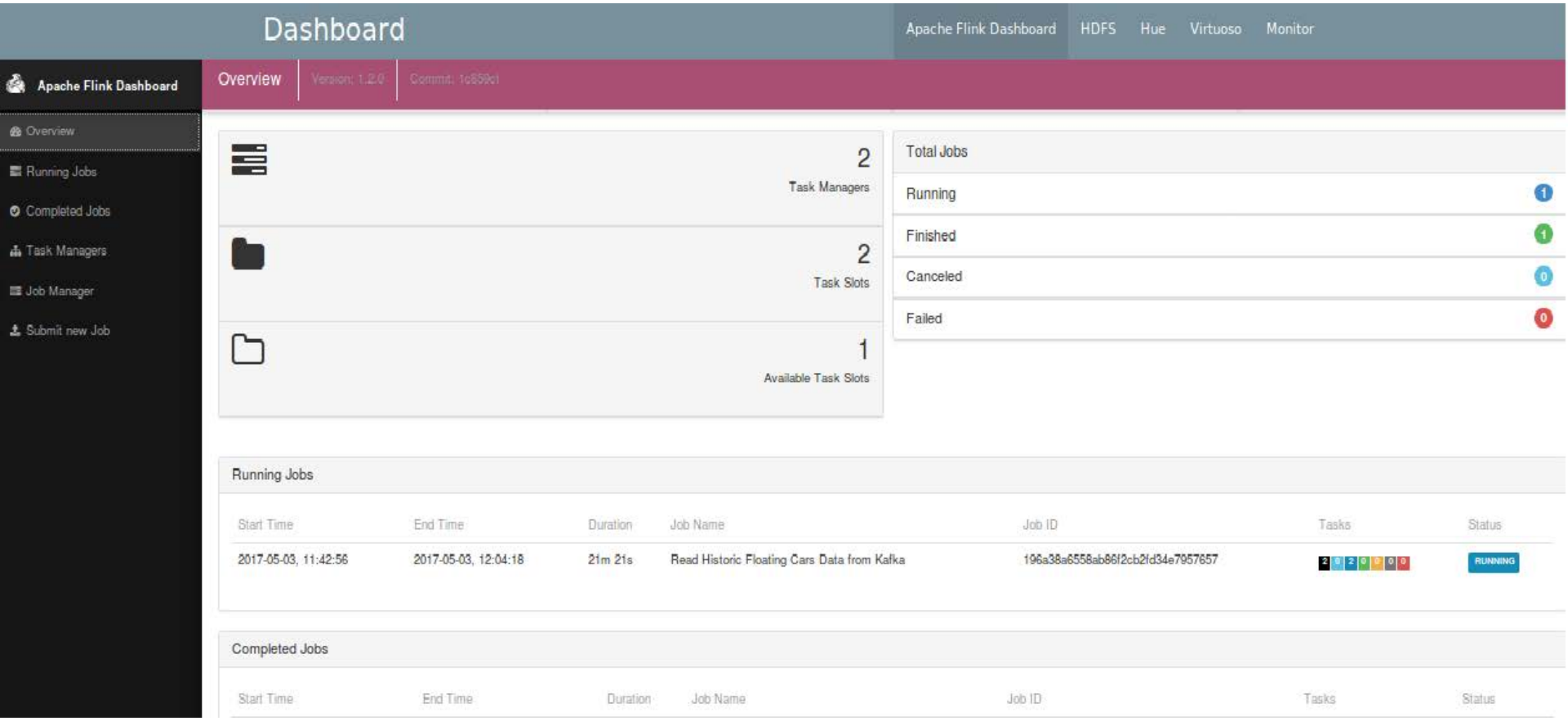


BDE Workshop Brussels
14 Sept. 2017

BIG DATA EUROPE

Empowering Communities with Data Technologies

Producer and Consumer



BDE Workshop Brussels
14 Sept. 2017

BIG DATA EUROPE

Empowering Communities with Data Technologies

Visualization

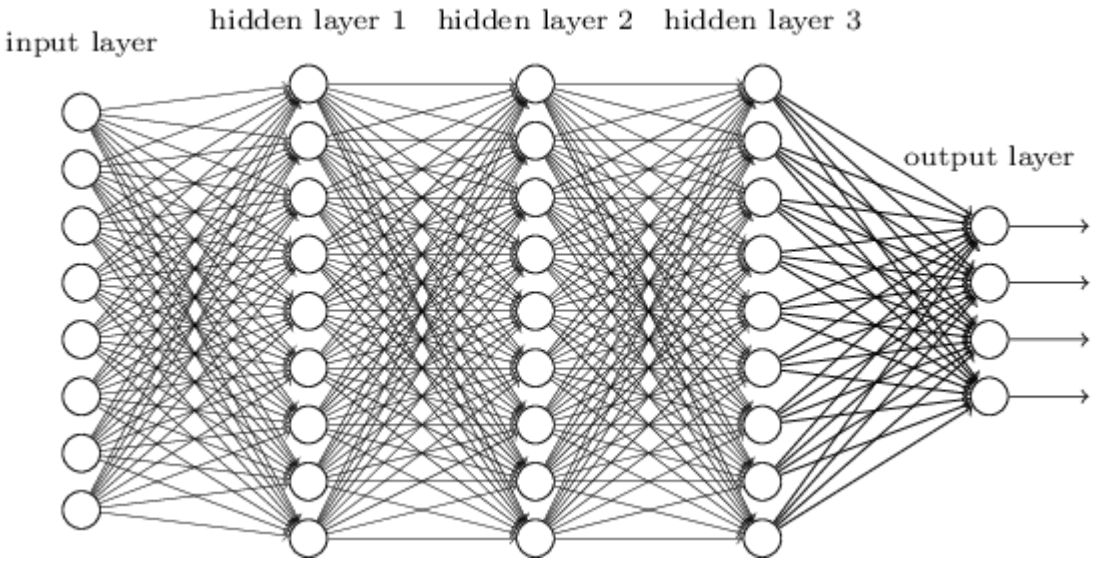
The pilot SC4 can process real-time FCD data for map-matching and classify a road segment according to the traffic level.

BIG DATA EUROPE

Empowering Communities with Data Technologies

Short-term traffic forecast

Algorithm: Feedforward ANN



BIG DATA EUROPE

Empowering Communities with Data Technologies

Short-term traffic forecast

Algorithm: Feedforward ANN

Hyperparameters (spatial and temporal correlation):

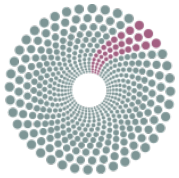
Input layer units: (Dd*24*60*Cr)/Tw

Dd = number of days (e.g. working days, 5 days)

Tw = time window (e.g. 30')

Cr = connected road segments (e.g. 3)

-> 720 input units



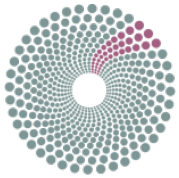
BIG DATA EUROPE

Empowering Communities with Data Technologies

SANSA-Stack: Big Data + Machine Learning + Semantic Technologies

SANSA-Stack, part of the BDE project, and RDF data sets based on semantic technologies such as LinkedGeoData, will enable more use cases related to SC4

BDE Workshop Brussels
14 Sept. 2017



BIG DATA EUROPE

Empowering Communities with Data Technologies

Thanks

BDE project website:

https://www.big-data-europe.eu/

Code repository:

https://github.com/big-data-europe

Contact:

luigi.selmi@iais.fraunhofer.de