

**BIG DATA EUROPE**

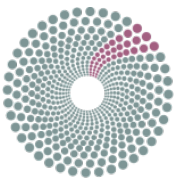
Empowering Communities  
with Data Technologies



## Pilot SC4

BDE Workshop  
Brussels 14.09.2017

BDE Workshop Brussels  
14 Sept. 2017



BIG DATA EUROPE

Empowering Communities  
with Data Technologies

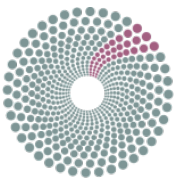
# Objective of the Pilot SC4



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



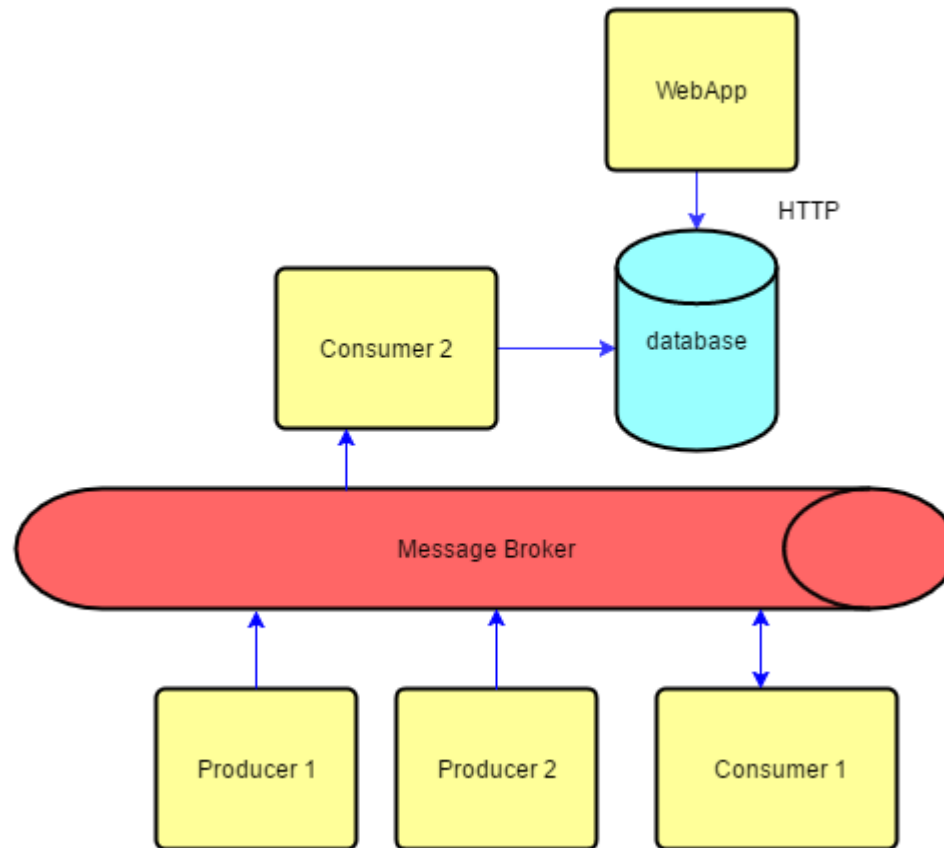
BDE Workshop Brussels  
14 Sept. 2017



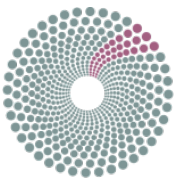
BIG DATA EUROPE

Empowering Communities  
with Data Technologies

# Microservice Architecture



BDE Workshop Brussels  
14 Sept. 2017



**BIG DATA EUROPE**

Empowering Communities  
with Data Technologies

# Message Broker

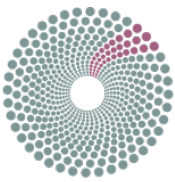
## Apache Kafka



Apache Kafka is a high-throughput distributed durable messaging system



BDE Workshop Brussels  
14 Sept. 2017

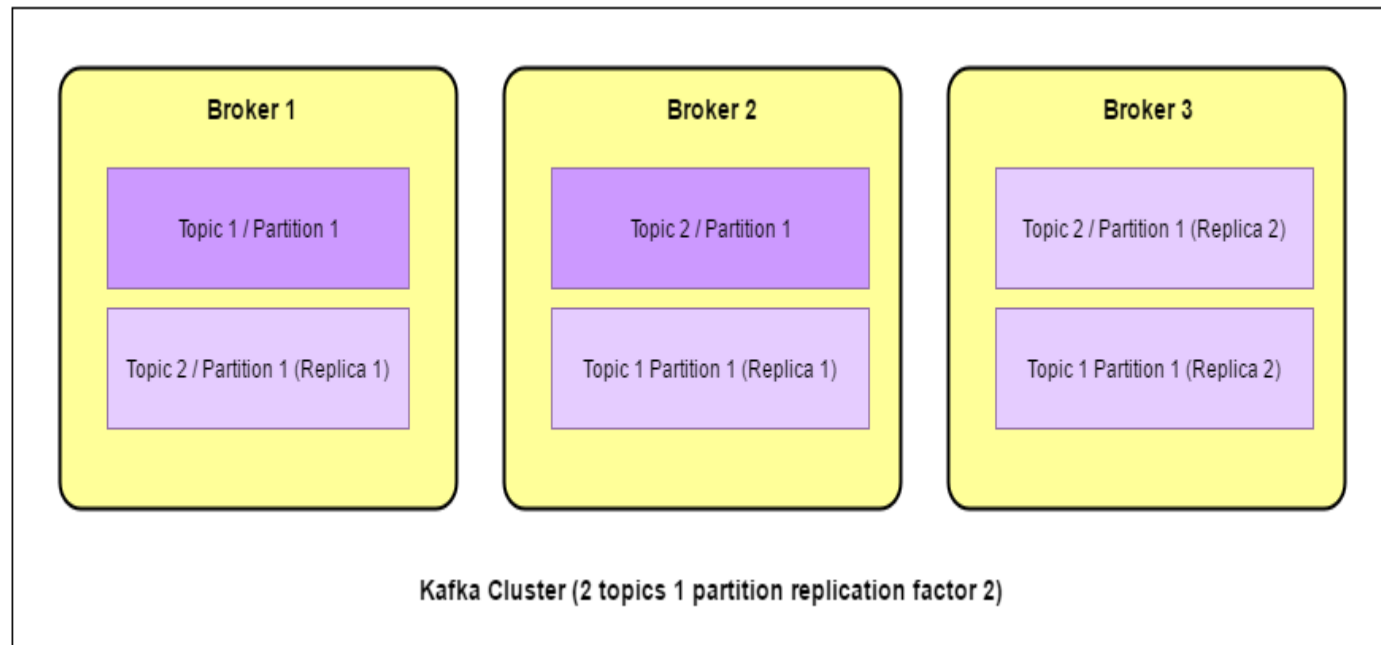


**BIG DATA EUROPE**

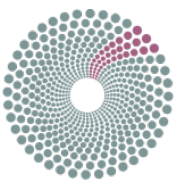
Empowering Communities  
with Data Technologies

# Kafka Cluster

**Apache Kafka**



**BDE Workshop Brussels**  
14 Sept. 2017



**BIG DATA EUROPE**

Empowering Communities  
with Data Technologies

# Stream and Batch Processor

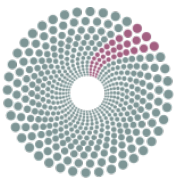
## Apache Flink



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



BDE Workshop Brussels  
14 Sept. 2017

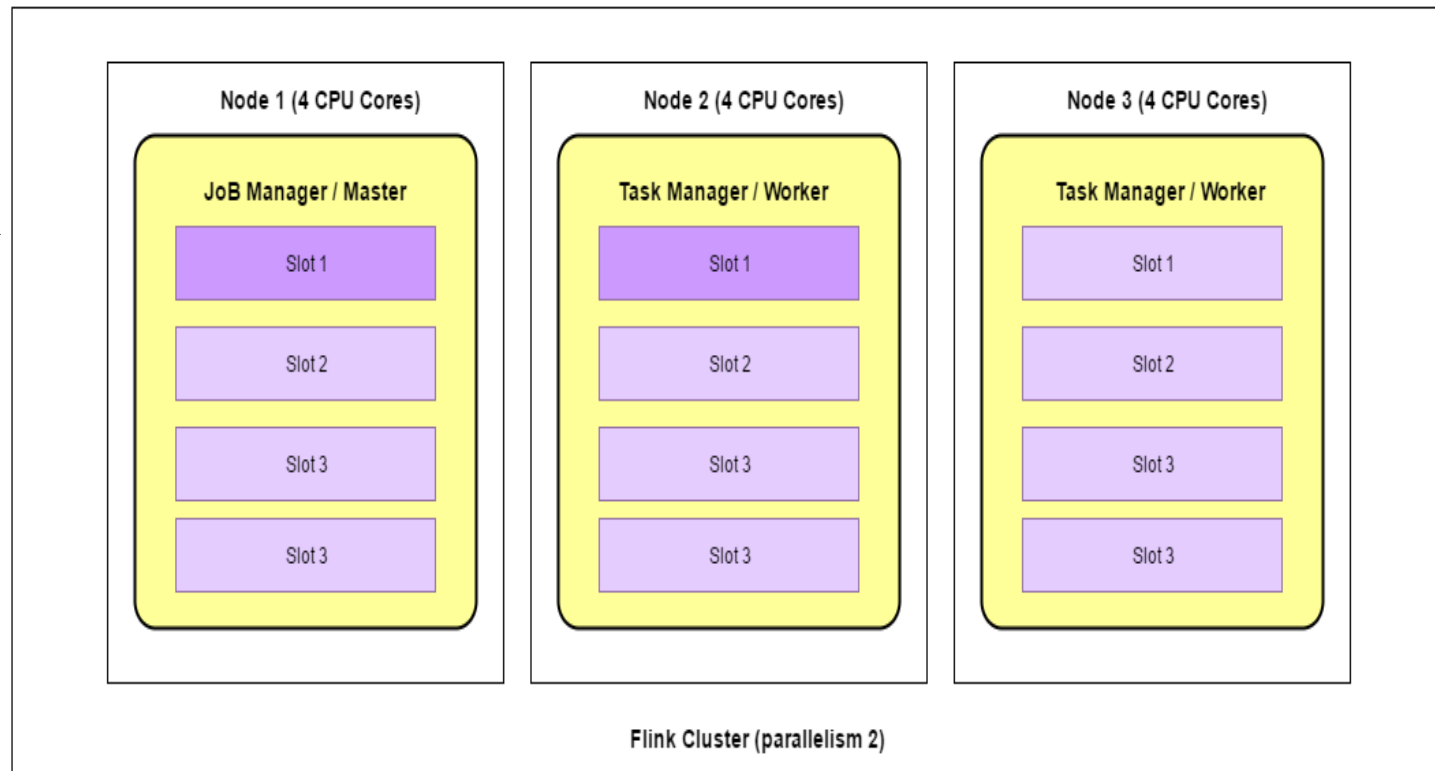


BIG DATA EUROPE

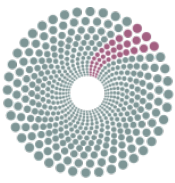
Empowering Communities  
with Data Technologies

# Flink Cluster

Apache Flink



BDE Workshop Brussels  
14 Sept. 2017



BIG DATA EUROPE

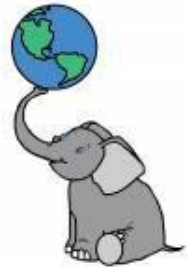
Empowering Communities  
with Data Technologies

# Storage and Indexing



elastic

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

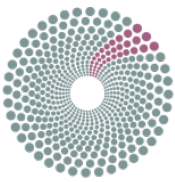


Elasticsearch is a distributed open source document database built on top of Apache Lucene. It stores the result of the workflow.



BDE Workshop Brussels  
14 Sept. 2017



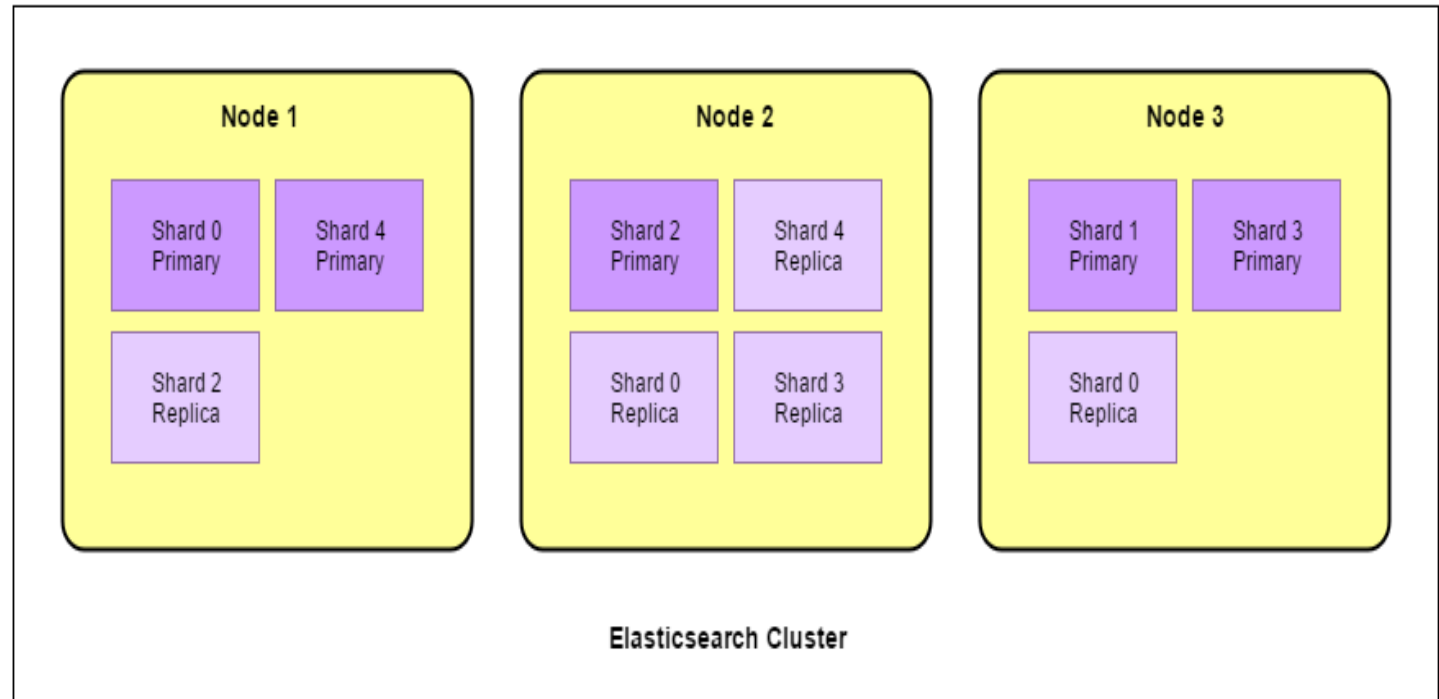


BIG DATA EUROPE

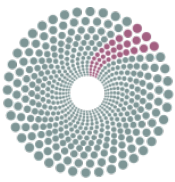
Empowering Communities  
with Data Technologies



# Elasticsearch Cluster



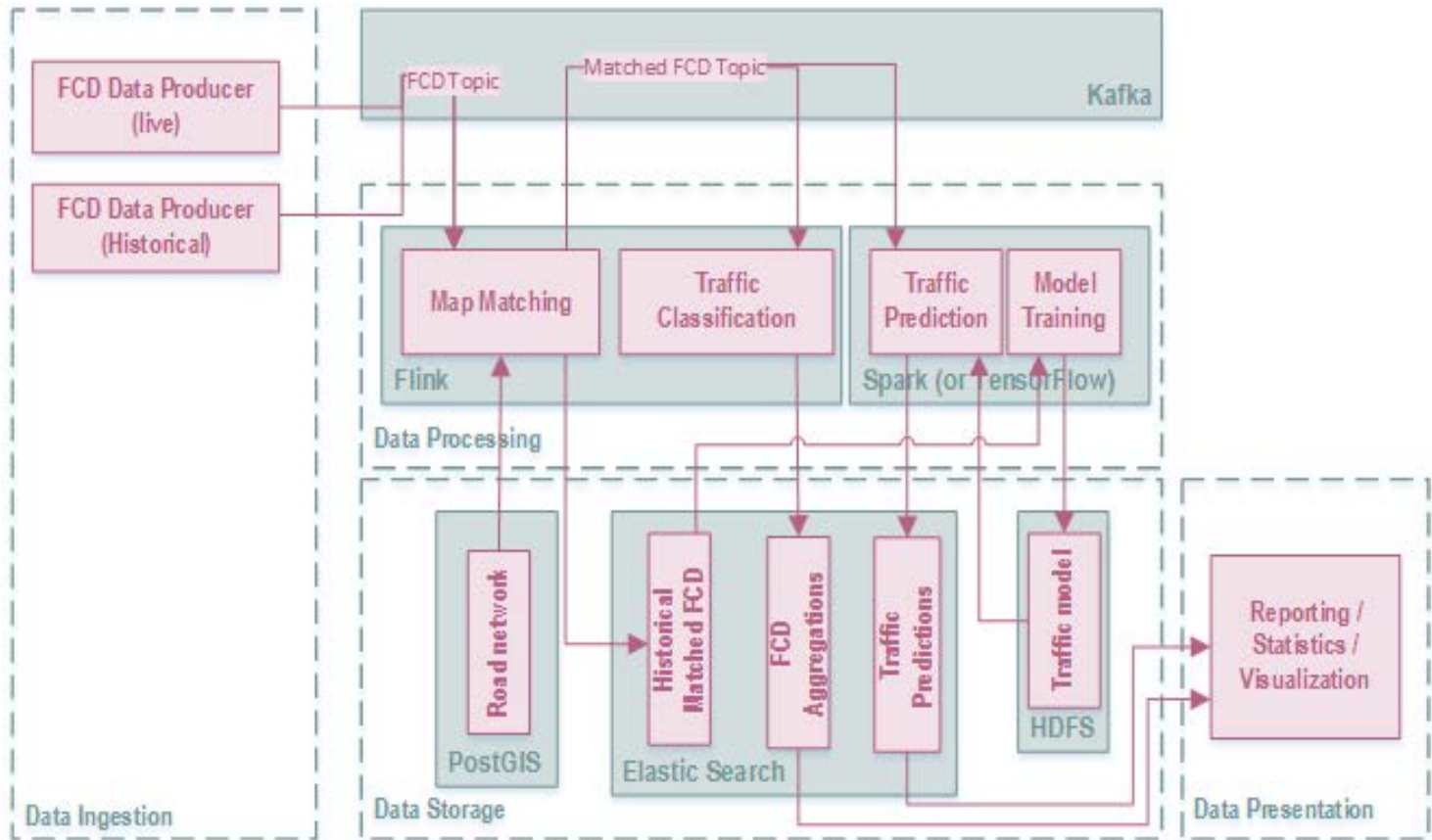
BDE Workshop Brussels  
14 Sept. 2017



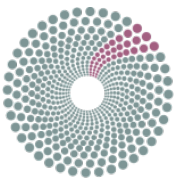
# Pilot Architecture

BIG DATA EUROPE

Empowering Communities  
with Data Technologies

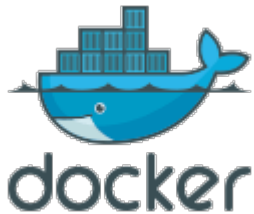


BDE Workshop Brussels  
14 Sept. 2017

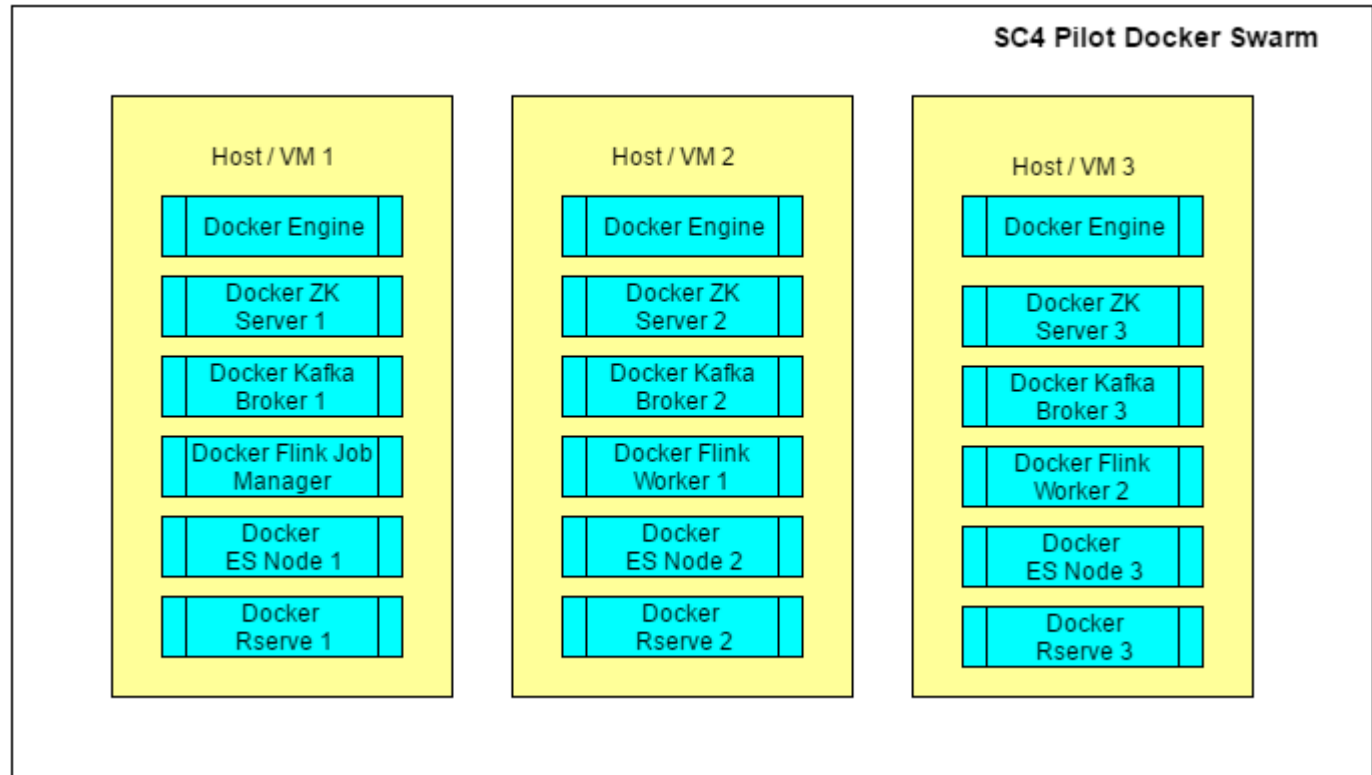


BIG DATA EUROPE

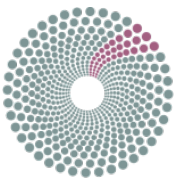
Empowering Communities  
with Data Technologies



# BDE Components



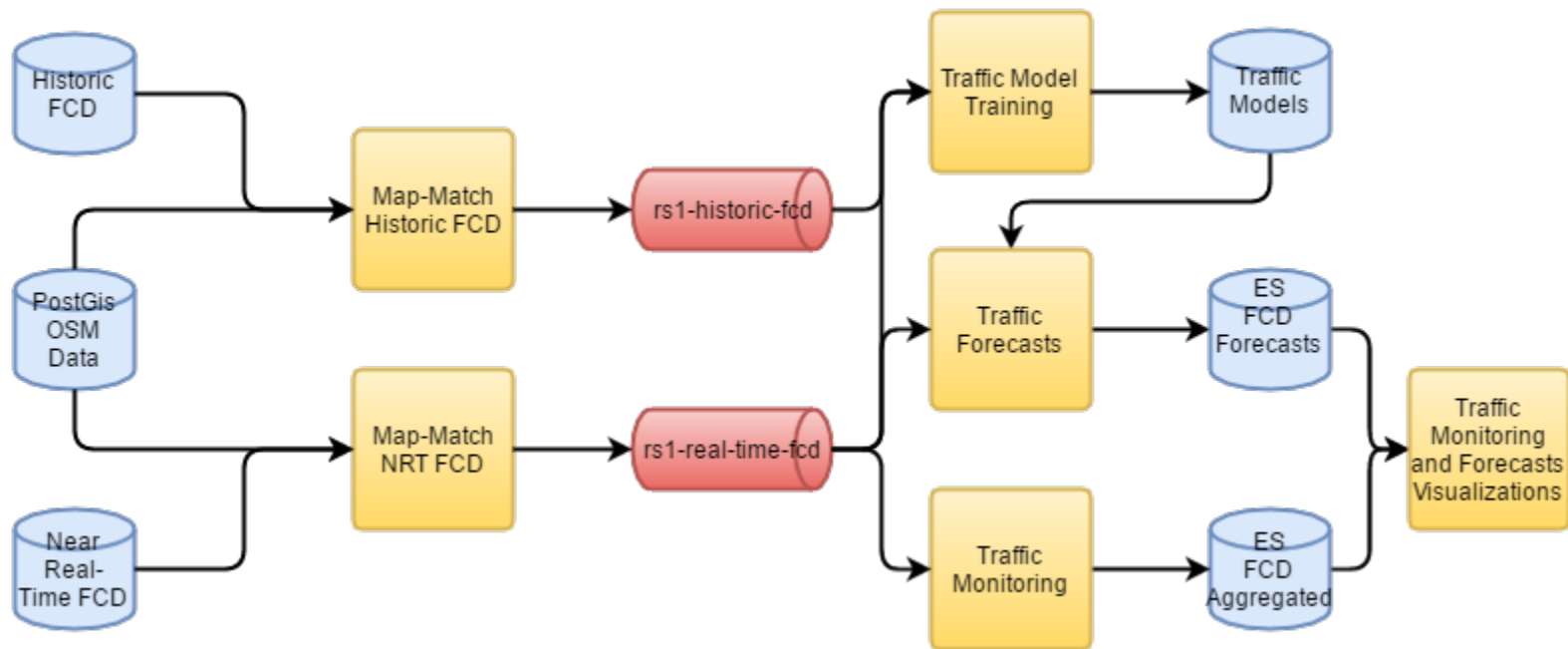
BDE Workshop Brussels  
14 Sept. 2017



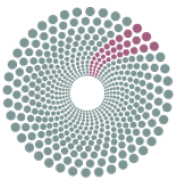
# The FCD Pipeline

BIG DATA EUROPE

Empowering Communities  
with Data Technologies



BDE Workshop Brussels  
14 Sept. 2017



BIG DATA EUROPE

Empowering Communities  
with Data Technologies

## Pilot Cluster

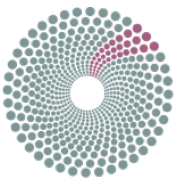
Minimum requirement for fault-tolerance and scalability

- 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

BDE Workshop Brussels  
14 Sept. 2017



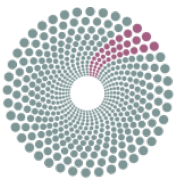


## Parallelization: map-match subtasks

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

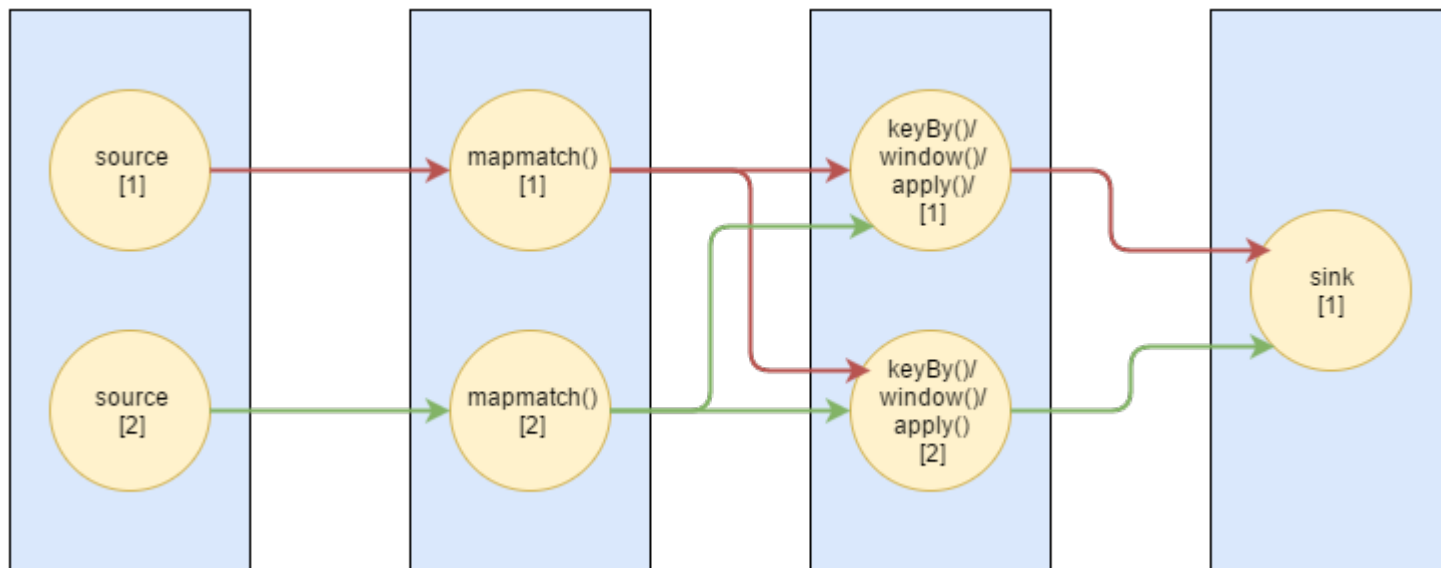
The subtasks can be distributed in slots with different parallelism (e.g. from 1 to 12)

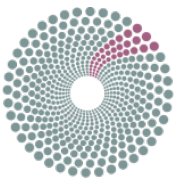




## Parallelization: map-match subtasks

A slot can process all the subtasks in a pipeline





## Parallelization: input and output data

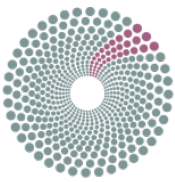
device_id	timestamp	lat	lon	speed	orientation	transit
-----------	-----------	-----	-----	-------	-------------	---------

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




Dashboard Apache Flink Dashboard HDFS Hue Virtuoso Monitor

BDE Pipeline Monitor Pipelines

## Pilot SC4 Start up

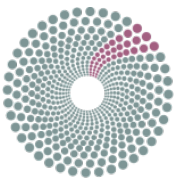
| This pipeline defines the initialization steps for the SC4 pilot

### Steps

 <b>Zookeeper</b> Zookeeper server.	<a href="#">setup_zookeeper</a>
<b>FINISH</b>	
 <b>Kafka</b> Starts a Kafka broker and create a Kafka topic.	<a href="#">setup_kafka</a>
<b>FINISH</b>	
 <b>Setup HDFS</b> Booting of the HDFS cluster.	<a href="#">setup_hdfs</a>
<b>FINISH</b>	



**BDE Workshop Brussels**  
14 Sept. 2017



# Data Upload

BIG DATA EUROPE

Empowering Communities  
with Data Technologies

Dashboard

Apache Flink Dashboard

HDFS

Hue

Virtuoso

Monitor

HUE



File Browser

hue



File Browser

Search for file name

Actions

Move to trash

Upload

New

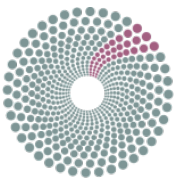
Home / user / hue

History Trash

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		root	supergroup	drwxr-xr-x	March 15, 2017 04:05 AM
<input type="checkbox"/>	.		hue	hue	drwxr-xr-x	April 13, 2017 01:46 AM
<input type="checkbox"/>	.Trash		hue	hue	drwxr-xr-x	April 12, 2017 09:08 AM
<input type="checkbox"/>	taxi_sample_100k.txt	105.5 KB	hue	hue	-rw-r--	April 13, 2017 01:43 AM



BDE Workshop Brussels  
14 Sept. 2017



# Producer and Consumer

BIG DATA EUROPE

Empowering Communities  
with Data Technologies

The screenshot displays the Apache Flink Dashboard interface. At the top, the 'Dashboard' title is visible, along with navigation links for 'Apache Flink Dashboard', 'HDFS', 'Hue', 'Virtuoso', and 'Monitor'. The main navigation bar includes 'Overview', 'Version: 1.2.0', and 'Commit: 1c859c1'. A left sidebar lists navigation options: 'Overview', 'Running Jobs', 'Completed Jobs', 'Task Managers', 'Job Manager', and 'Submit new Job'. The main content area is divided into several sections:

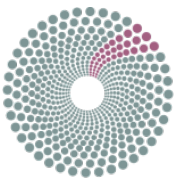
- Task Managers:** 2 Task Managers
- Task Slots:** 2 Task Slots
- Available Task Slots:** 1 Available Task Slots
- Total Jobs Summary:**
  - Running: 1
  - Finished: 1
  - Canceled: 0
  - Failed: 0
- Running Jobs Table:**

Start Time	End Time	Duration	Job Name	Job ID	Tasks	Status
2017-05-03, 11:42:56	2017-05-03, 12:04:18	21m 21s	Read Historic Floating Cars Data from Kalka	196a38a6558ab86f2cb2fd34e7957657	2 0 2 0 0 0 0	<b>RUNNING</b>
- Completed Jobs Table:**

Start Time	End Time	Duration	Job Name	Job ID	Tasks	Status
------------	----------	----------	----------	--------	-------	--------



BDE Workshop Brussels  
14 Sept. 2017



# Visualization

BIG DATA EUROPE

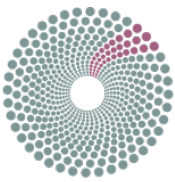
Empowering Communities  
with Data Technologies



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



BDE Workshop Brussels  
14 Sept. 2017

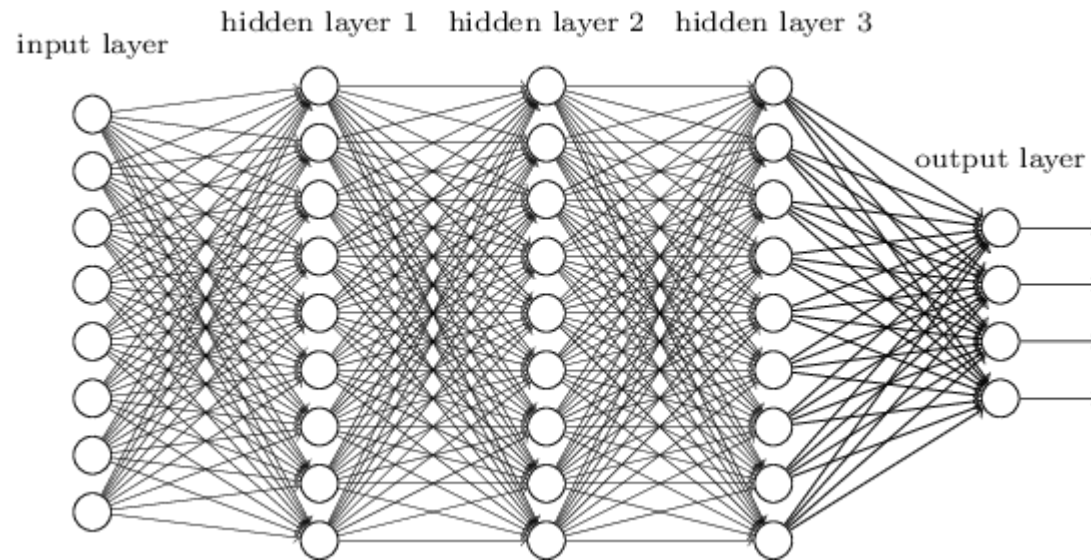


**BIG DATA EUROPE**

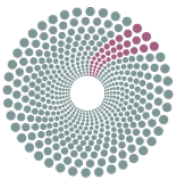
Empowering Communities  
with Data Technologies

# Short-term traffic forecast

## Algorithm: Feedforward ANN



BDE Workshop Brussels  
14 Sept. 2017



## Short-term traffic forecast

Algorithm: Feedforward ANN

Hyperparameters (spatial and temporal correlation):

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

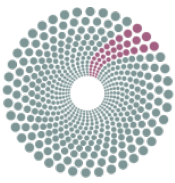
Dd = number of days (e.g. working days, 5 days)

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

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

-> 720 input units





**BIG DATA EUROPE**

Empowering Communities  
with Data Technologies

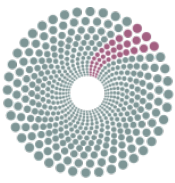
# SANSA-Stack: Big Data + Machine Learning + Semantic Technologies



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



BDE Workshop Brussels  
14 Sept. 2017