From a data perspective we see endless streams of data collected by sensors and devices immersed in a graph of infrastructures.
Objective of the SC4 Pilots

A scalable, fault-tolerant and flexible platform based on open source frameworks that can process unbounded data sets and graphs.
Apache Kafka is a high-throughput distributed durable messaging system.
Kafka Cluster

Apache Kafka

Kafka Cluster (2 topics 1 partition replication factor 2)
Apache Flink is an open source platform for distributed stream and batch data processing.
Flink Cluster

Apache Flink

Node 1 (4 CPU Cores)
- Job Manager / Master
  - Slot 1
  - Slot 2
  - Slot 3
  - Slot 3

Node 2 (4 CPU Cores)
- Task Manager / Worker
  - Slot 1
  - Slot 2
  - Slot 3
  - Slot 3

Node 3 (4 CPU Cores)
- Task Manager / Worker
  - Slot 1
  - Slot 2
  - Slot 3
  - Slot 3

Flink Cluster (parallelism 2)

L. Selmi - BDE - Societal Challenge on Transport
Storage and Indexing

Elasticsearch is a distributed open source document database built on top of Apache Lucene
Elasticsearch Cluster

Node 1
- Shard 0 Primary
- Shard 2 Replica

Node 2
- Shard 2 Primary
- Shard 0 Replica

Node 3
- Shard 1 Primary
- Shard 0 Replica

Elasticsearch Cluster

L. Selmi - BDE - Societal Challenge on Transport
The R Project

R is a free software environment for statistical computing. It is used in the pilot to run the map-matching and the prediction algorithms.
Pilot 1 Architecture

BIG DATA EUROPE
Empowering Communities with Data Technologies

L. Selmi - BDE - Societal Challenge on Transport
BDE Components

SC4 Pilot Docker Swarm

Host / VM 1
- Docker Engine
- Docker ZK Server 1
- Docker Kafka Broker 1
- Docker Flink Job Manager
- Docker ES Node 1
- Docker Rserve 1

Host / VM 2
- Docker Engine
- Docker ZK Server 2
- Docker Kafka Broker 2
- Docker Flink Worker 1
- Docker ES Node 2
- Docker Rserve 2

Host / VM 3
- Docker Engine
- Docker ZK Server 3
- Docker Kafka Broker 3
- Docker Flink Worker 2
- Docker ES Node 3
- Docker Rserve 3

L. Selmi - BDE - Societal Challenge on Transport
A Pipeline

BIG DATA EUROPE
Empowering Communities with Data Technologies

L. Selmi - BDE - Societal Challenge on Transport
SC4 Pilot 1 can process real-time FCD data for map-matching and aggregations (e.g. average speed)
Future Work

- Pilot 2: integration of the FCD historical data and prediction algorithm
- Pilot 3: more data sets (BT sensors, NYC taxi data, OpenStreetMap)