Big Data Analytics for extracting mobility patterns in a large urban center

UNINOVA

Summary

- Motivation
- UNINOVA Big Data Architecture
- Processing & Analytics
- Performance
- Visualization
- Conclusions

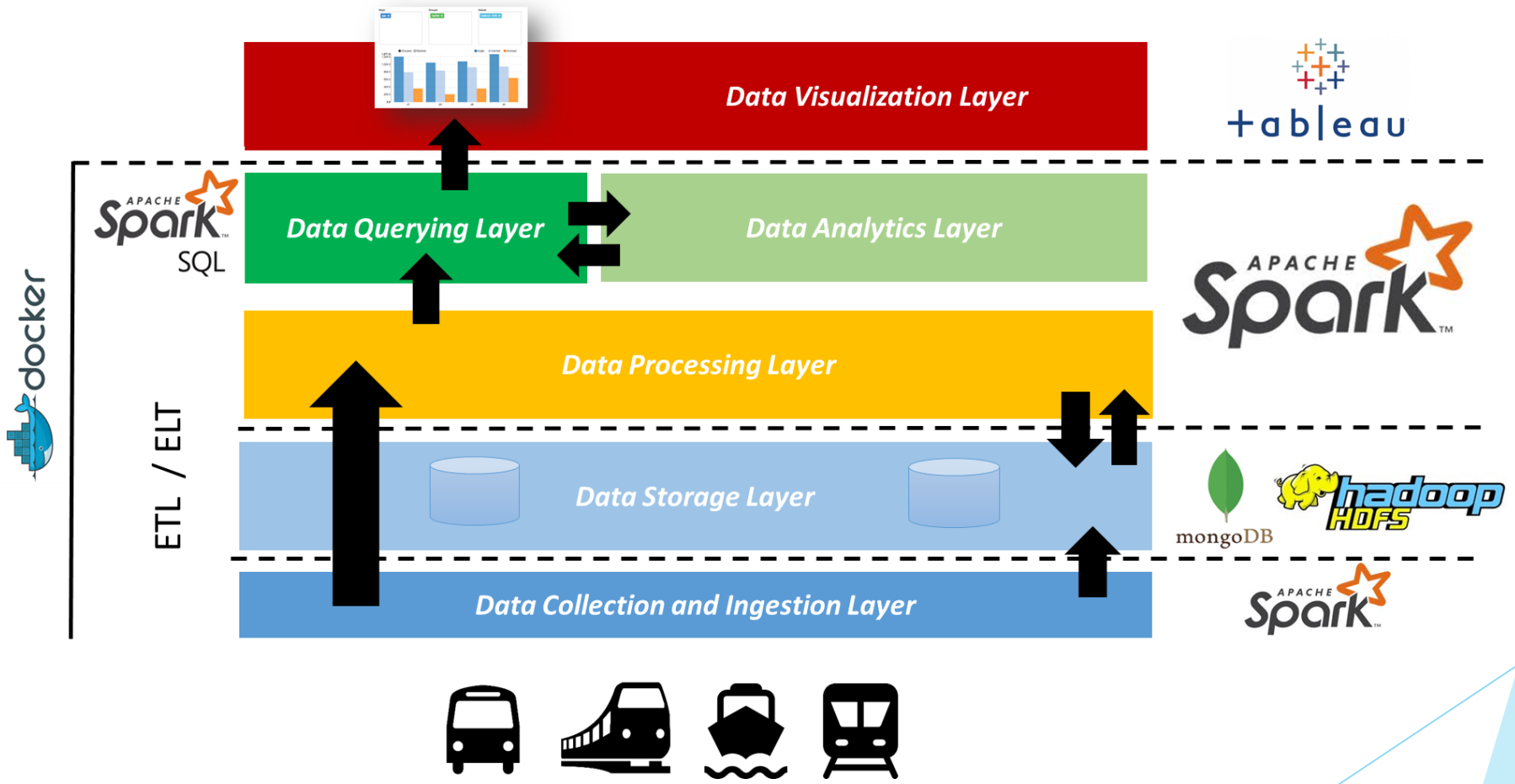
Motivation

- Public Transportation in Lisbon, Portugal
  - Independent public/private operators
  - One association (OTLIS) handles data coming from all operators
    - Ticket validations
    - Stations/stops locations and information
    - Users
  - Data sharing between operators is a challenge
    - Legal/business advantage issues
    - Privacy concerns
  - Analytics performed with traditional techniques
    - Data gathered through questionnaires and human observations
    - Difficulty to get meaningful insights with traditional DW approaches

Research question

- Which technologies can be used in order to provide useful insights about mobility patterns in large urban centers, considering large volumes of ticketing data from different operators?

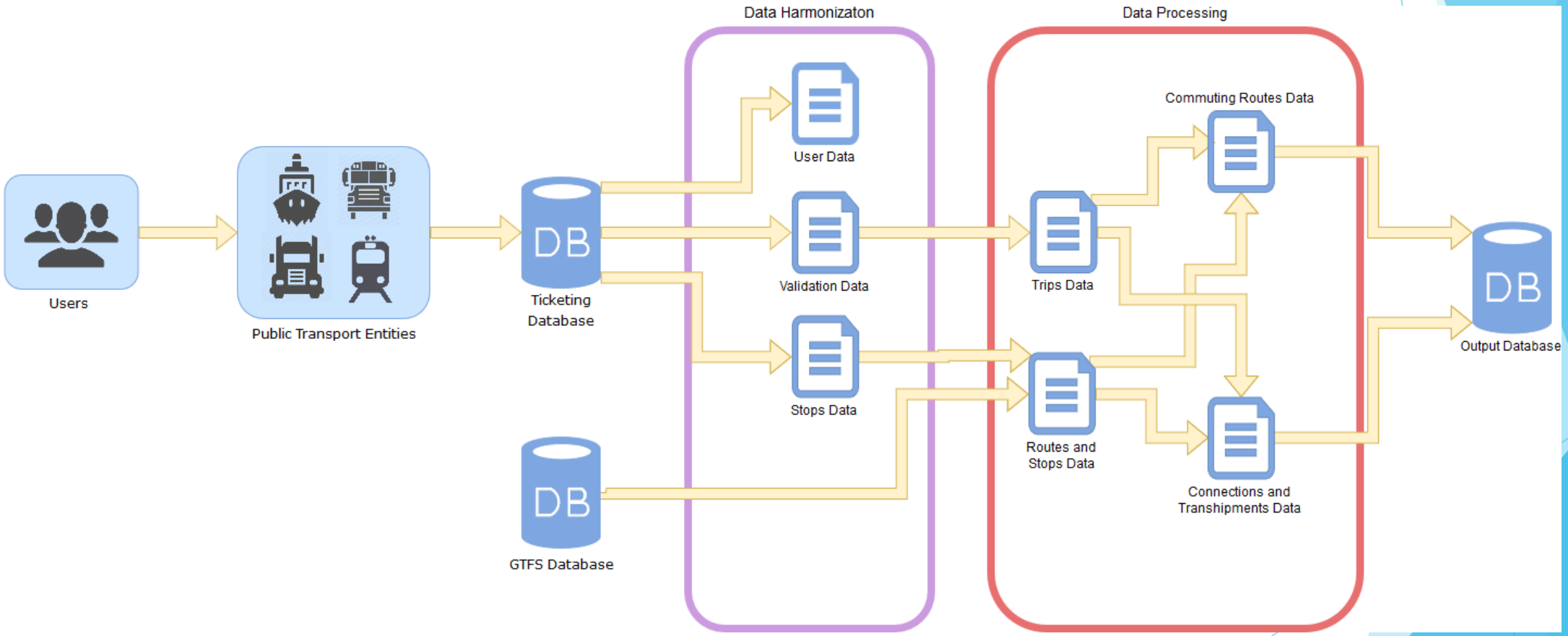
UNINOVA Big Data Architecture



Processing & Analytics

- Clean original data
  - Duplicates
  - Erroneous validations (e.g. consecutive entry validations on the same station)
  - Validations without location information
  - Consecutive entry and/or exit validations with less than 5 minutes between
- Harmonize original data into three distinct formats: Validations, Users, Locations
- Provide semantics via GTFS mappings of locations and routes
- Create new knowledge/insights from collected data (about connections, transhipments and pendular movements)

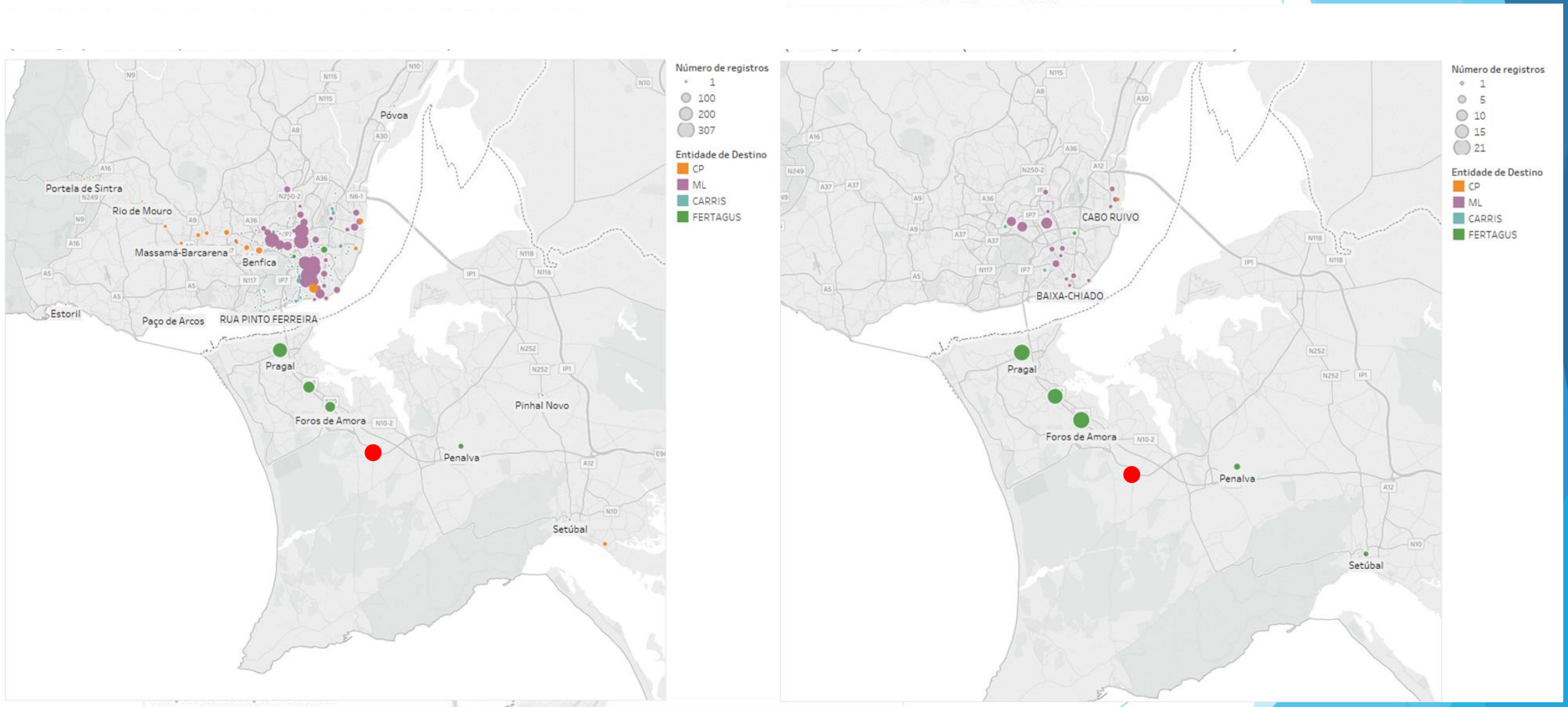
Processing & Analytics



Performance

- Test:
  - One month of data (May 2018): +55 million records
- Before:
  - Oracle Cloud with traditional DW processes
  - Only pre-processing and visualization
  - Time span: Some days – one week
- With UNINOVA Big Data Architecture:
  - Single node (AMD Ryzen 5 1600 - 12CPU's, 32GB RAM (Corsair Vengeance LPX), SSD 120GB + 1TB HDD)
  - Pre-processing + analytics
  - **Time span: 4hours (Reading/writing to MongoDB on each stage, no indexes)**

Visualization



Conclusions

- Novel Big Data architecture for efficiently perform processing and analytics on public transportation data
- The architecture spans the whole life cycle of Big Data
- Development of an unsupervised approach to collect and process data, and to produce meaningful insights
- Comparing with traditional DW processes, the architecture enables much better performances, even on a single machine
  - Less costs (with dedicated Cloud services)
  - Better knowledge and insights
  - Possibility to have an effient in-house solution