Spark Project: Analysis and Forecasting of the Urban Pollution

October 9, 2025

Project presentation

The goal of this project is to design an application capable of measuring, analyzing, and predicting pollution in an urban network, using Apache Spark for big data processing and functional programming. This project will highlight how functional programming is particularly well-suited to big data processing, leveraging its capabilities for parallelization and the efficient management of large amounts of data.

Educational objectives

- Mastering the processing of large quantities of data with Spark (RDD, DataFrames, SQL).
- Utilize functional programming with operations such as map, flatMap, filter, reduce, groupBy or aggregate.
- Experiment with batch processing and streaming of data.
- Discover the use of Spark MLlib for simple prediction tasks.

Data sources under consideration

- Pollution measurements: CO2, fine particles, noise, humidity.
- Time data: hours, days, weeks.
- Contextual information: stations, transport lines, weather.

- Additional data: passenger traffic, one-off events, days
  f´eri´es.

Implementation phases

1. Data Ingestion and Preparation

You are invited to build the dataset yourself, either by collecting public data or by generating a simulated dataset adapted to the project's objectives.

- Reading CSV or JSON files with Spark.
- Data cleaning: removal of duplicates and missing values.

2. Functional Transformation and Exploration

- Using map, filter and flatMap to transform data.
- Calculation of statistics per station or per line (average, maximum, minimum)
  mum).
- Extraction of relevant temporal variables (hour, day, month).

3. In-depth analysis

- Identification of stations most exposed to pollution.
- Detection of hourly peaks and critical periods.
- Creation of a global pollution indicator.
- Automatic detection of anomalies in the data.

4. Modeling the relationships between stations

- Representation of stations and their connections in graph form with Spark
  GraphX.
  ´
- Study of the spread of pollution through the network.

5. Prediction

- Construction of a data transformation pipeline to create fea-
  adapted sizes.
- Application of Spark MLlib prediction models (regression, trees of
  decision, random forests).

6. Real-time processing (optional)

- Simulation of data streams from continuous sensors.
- Application of functional transformations to these flows.
- Automatic detection and reporting of anomalies.

Deliverables

- A PDF, in the form of slides, presenting the aspects and concepts of Spark that
  you used in this project.
- Functional Spark code (Scala) using functional programming.
- Report describing the methodology and the results obtained.
- Visualizations of analyses and predictions (optional).

Summary

This project allows for the combination of big data processing, functional programming, statistical analysis and prediction, to build a complete, robust and scalable application.

References

Apache Spark

- Official Apache Spark website: https://spark.apache.org/
  Complete documentation, quick guides, API and tutorials.

- Chambers, B., & Zaharia, M. (2018). Spark: The Definitive Guide. O'Reilly Media.
  A comprehensive book covering RDD, DataFrames, SQL, MLlib and Streaming.
- Damji, JS, et al. (2020). Learning Spark: Lightning-Fast Big Data Analytics-sis. O'Reilly Media.
  Ideal for beginners and intermediate learners with examples in Scala and Python.
- Databricks Spark Documentation: https://docs.databricks.com/ Practical tutorials and concrete notebooks for Spark.

Scala

- Odersky, M., Spoon, L., & Venners, B. (2021). Programming in Scala (4th edition). Artima.
  Book written by the creator of Scala, covering the entire language and functional programming.
- Official Scala documentation: https://docs.scala-lang.org/ Guides, tutorials, API references and best practices.
- Chiusano, P., & Bjarnason, R. (2014). Functional Programming in Scala. Manning Publications.
  To understand functional programming applied to Scala.
- Scala Exercises: https://www.scala-exercises.org/ Interactive platform offering exercises on Scala and functional programming.