Project
Implementation of a (Big) Data Management Backbone

P2 Description

Big Data Management – FIB – UPC

Two parts

- P1 – Data design (Landing Zone)
  - Conceptualization and Data Lake design
  - Technologies: Apache Hadoop (+ file formats), Apache HBase, MongoDB
- P2 – Descriptive and predictive analysis (Formatted and Exploitation Zones)

I modelli saranno costruiti su un pezzo dei dati tale per cui si è pronti a partire. Spark serve a prendere qeul pezzo dalla landing zone

  - Data integration and reconciliation
    - Technologies: Apache Spark (core), a visualization tool (e.g., Tableau)
  - Distributed machine learning and real-time data prediction
    - Technologies: Apache Spark (MLlib, Streaming), Apache Kafka, a visualization tool for streams (e.g., Kibana)

P2 objectives

- Integrate all gathered datasets in the Formatted Zone
  - Handle duplicates, reconciliate data, clean, etc.
- Implement the calculation of KPIs for descriptive analysis
  - Store them in the Exploitation Zone **Build a dashboard**
- Prepare the input data and train an ML model for predictive analysis
  - Store it in disk
- Ingest a data stream
  - Perform predictions applying the model on the data stream elements
  - Describe the data stream using approximate stream analysis algorithms
- Graphically display the results of the analysis

Distributed machine learning

- Create two datasets – perform the necessary transformations, cleaning
  - Training
  - Validation
- Use the training dataset to create a classifier using Spark MLlib (RDD-based)
  - https://spark.apache.org/docs/latest/mllib-guide.html
- You are free to choose the kind of model
  - The objective of the course is not to optimize this part
- Validate the model
  - Compute recall and accuracy
- Store the model
- Ingest and process a data stream to perform predictions using the stored model

Different business ideas

- The focus of your implementation will differ according to the needs of the business idea
- However, everyone must adhere to the zoned Data Management Backbone framework
- Analytical needs might vary **Must choose at leat 2 out of these 3**
  - Descriptive analytics
  - Predictive analytics using distributed ML
  - Stream analytics using approximate algorithms

Your project should cover at least two of such analytical needs

Technologies

- Apache Spark (RDD-based)
  - Integration and reconciliation using lookup tables
  - Calculate KPIs and store them in views
  - Your pipeline must be optimal from the perspective of…
    - Minimizes the number of wide dependencies
    - Caches results when required
    - Exploits parallelism
    - …
- Apache Kafka
  - Endpoint for stream ingestion
- Apache Spark MLlib
  - Classifier and evaluation
- Apache Spark Streaming
- Visualization tool
  - Choose the one you prefer
  - Provide online access or a video of the resulting solution

Delivery

- Document (max 5 pages)
  - Describe the pipelines to integrate and to calculate/store analytical data
    - Sketch the pipelines at a higher abstraction level. Use the notation seen in the lectures to describe the Spark job **Talk also about the disadvantages of the tool!**
    - Elaborate on your assumptions. Refer to any specificity of your solution that should help the lecturer to understand the decisions you made in your code that, otherwise, might look like controversial
  - Describe the extra dataset and new KPIs
  - Describe and justify the data model used in the Formatted and Exploitation Zones
- Code
- Extra material
  - Online access to visualization tool, videos, etc.

Closing