



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)
 - 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) 

finding HOW to join if
there are no unique IDs

Ingests data streams

P2 objectives

- Integrate all gathered datasets in the Formatted Zone Prepare data for analysis
 - Handle duplicates, reconcile data, clean, etc.
- Implement the calculation of KPIs for descriptive analysis KPI is specific to my use-case. Explain the data we have and calculate aggregates
 - Store them in the Exploitation Zone
- Prepare the input data and train an ML model for predictive analysis
 - Store it in disk Select labels if we do not have labels
- Ingest a data stream
 - Perform predictions applying the model on the data stream elements
 - Describe the data stream using approximate stream analysis algorithms Call the method and use it. Don't go crazy with this, as it's covered in the other class
- Graphically display the results of the analysis Display the results of all of the items we are doing - both meta/architecture calcs and KPIs as well as the results of the predictions/ analysis

Distributed machine learning

Follow the “traditional” process to build a ML model

- 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
 - Descriptive analytics
 - Predictive analytics using distributed ML
 - Stream analytics using approximate algorithms

From these 3, our project must cover 2

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
- 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

Expecting the code is optimal for this part. We are looking to optimize code and apply what we are learning in class

This is essentially a queue. It maintains a stream of data for a period of time. After some period, it drops the data. It's just pushing info into a cue and then starts removing it.

Usually, Spark connects to Kafka to pull data from different places

If we are building a website, we can embed the visualization into the app or website. Website or app is not needed for this class, but it is encouraged for VBP. It is good to have a full demo during the final presentation. Having a nice UI is more assessed as part of the global project.

Delivery

- Document (max 5 pages) Most important ideas and decisions to understand the project
 - 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
 - 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
Data structure, where is it stored, what is being stored
- Code
- Extra material
 - Online access to visualization tool, videos, etc.

Closing

Follow Up Sessions - recommended progress for each session

1 - What do we want to obtain, what is the objective of the data analysis?

2 - following sessions will be on-demand