Paradigms of Big Data

Master's in data science

2022/2023

Goal

The project of Paradigms of Big Data course aims to explore:

- Big Data technologies, mainly, Spark, Kafka, and Cassandra
- Stream processing

Description

The company XPTO needs a system which determines in real-time failures in a machine.

In Paradigms of Big Data, the students should develop a system for anomaly detection or predictive maintenance using Kafka, Spark, and Cassandra to support the XPTO request.

The project involves collecting sensor data from various sources using Kafka, processing and analysing the data using Spark Streaming, and storing the results in Cassandra.

The practical work will be developed using Python language and it is composed of two different components:

1. The first part encompasses Kafka (50 %) as well as an exploratory analysis;
2. In the second component the student must apply *Machine Learning* models, using the Spark Streaming (50 %) storing data in Cassandra.

As datasets the student can choose between:

- NASA Turbofan Jet Engine Data Set
- NASA Bearing Dataset

Component I (50 %)

In the first part of the work, the student should:

- Choose a dataset
- Do an exploratory analysis.
- Install Spark, Kafka and Cassandra
- Develop a producer and consumer for the dataset variables
- Elaborate a paper with abstract + introduction + related-work

Component II (50 %)

In the second part, the student should:

- Complete Spark Installation
- Connect Kafka, Cassandra and Spark
- Store in Cassandra data which will be required by Machine Learning Model Analyse the accuracy of the model
- Apply Machine Learning models using Spark Streaming for anomaly detection or predictive maintenance
- Analyse the performance in terms of running time
- Finish the paper with Proposed Method + Results + Conclusions

Submission

The work must be carried out in group. Each work should be submitted in Moodle using a ZIP archive with the PDF report and all Python scripts. In addition, the student should include the declaration of authorship also available in Moodle.

Each paper has a limit of 10 pages. The deadlines will be announced in Moodle. The oral presentation is always in the next class.

References and Resources

The student should consider the bibliography of this subject as well as the material provided by professor. The development environment is free choice. However, we recommend using Pycharm or Jupyter

Good work!

Fátima Leal