BTS | Barcelona Technology School

Real-time Data Analysis Curriculum

Master in Big Data Solutions

1 Subject description

Subject: Real Time Data Analysis
Year: 2020-2021
Quarter: 2nd & 3rd
Degree: Master in Big Data Solutions
Number of credits: 5 **/ Hours of class:** 48 **/ Hours of homework:** 48

Teaching Staff:

Name: Joan Gasull Jolis
Email: joan.gasull@bts.tech
Linkedin: https://www.linkedin.com/in/joan-gasull-jolis-592239123/

Schedule:

- 4-hour sessions from 9:00 – 13:00
- Weekly home work: 4 hours

2 Subject introduction and goals

There are several applications to process an enormous volume of data and to process it as quick as possible.

Natural phenomena predictions (hurricanes, tsunamis, etc), with delays, can cost human's lives. Delays in processing new content via video, audio, twitter and news report can provoke wrong trading decisions. Delays in traffic jam prediction cost extra time. Credit card fraud detection needs accurate results in fractions of seconds.

The two operatives in the previous examples are velocity and latency, and that's where classical distributed big data batch processing systems (like Hadoop map reduce) fall short. They are designed to deliver in batch mode and can't operate at a latency of nanoseconds/milliseconds.

This course studies in depth the spark framework as a modern approach to deal with the complexity of processing real-time big data taking advantages of both batch and streaming processing methods.

The goal of the course is to enable students to characterize the real time data analysis processing problems arising on big data and solve them using the techniques, algorithms and tools implemented on Spark framework. Also by following a test driven development approach, to get good practices on software development for big data systems.

3 Teaching methodology

This course will have focus in both theory and practice, putting a special effort on the real-word problem solving.

The teaching methodologies will depend on the subject being studied and include the following:

- Teaching with Discussions (Lectures, Lecture-demonstration)
- Student-centered (Analysis and discussion of case studies, Presentations by student panels, Student-group reports, Individual reading and research with oral presentation)

Everyday classes are complemented with assignments and topics to review, oriented to be developed both individually or as a team. In most of the cases the students will participate in the evaluation and co-evaluation of their learning.

4 Contents

1. Parallel programming on large scale data processing.
2. Parallel programming using Spark framework.
3. Batch processing on Spark:
   - Structured data processing on pyspark, Spark SQL.
4. Machine learning pipelines on spark, MLIB.
5. Stream processing:
   - Outlook
   - Structure Streaming with Twitter
   - Web Scraping stocks and newspapers
6. Lambda Structure

2 Schedule of contents and activities

Session	Activity at class (4 hours)	Activity at home (4 hours)
Session 1 8-mar	**Content** Computer science foundations on parallel and distributed programming. Introduction to Spark Framework. Spark context, *resilient distributed data-set* (RDD). **Activity** Setup programming environment. First overview of pyspark	Quiz about basics of parallel and distributed programming and RDDs

Session 2 15-mar	**Content** Transformations, actions and basic operations in pyspark **Activity** To solve basic problems using pyspark Explore new transformations and actions	**Individual Home Work** To solve basic problems using pyspark **Bibliography:** https://spark.apache.org/docs/latest/rdd-programming-guide.html
Session 3 12-apr	**Content** More transformations, actions and operations in pyspark. First contact with Dataframes **Activity** Basic spark programs. RDD Programming with Transformations and actions.	**Individual Home Work** To build basic data-processing programs on spark.
Session 4 19-apr	**Content** Advanced RDD operations on spark: Joins, Shuffle, Shared variables. **Activity** To build batch processing spark task. **Working on a database**	**Individual Assignment (GRADED)** To solve data analysis problem on batch processing spark task. SQL-Datasets
Session 5 26-apr	**Content** **Working on a database.** Spark SQL. Using data sources. Basic Statistics **Activity** To solve data analysis problem on batch processing spark.	**Individual Home Work** To solve data analysis problem on batch processing spark task using data sources.
Session 6 03-may	**Content** Machine learning on spark (MLIB). ML Pipelines. Extracting, transforming and selecting features. **Activity** Implementing machine learning pipelines on spark.	**Individual Home Work** Implementing machine learning pipelines on spark.
Session 7 10-may	**Content** MLIB: Machine Learning algorithms & Clustering. **Activity** Build application to solve classical clustering problems using "MLIB Clustering"	**Individual Home Work** Implementing ML solutions on pyspark

Session 8 17-may	**Content** More Machine Learning algorithms. Functional programming **Activity** Build application to solve Regression/Classification problems.	**Individual Assignment (Graded)** Machine learning implementation in pyspark
Session 9 31-may	**Content** Fundamentals of streaming analysis. Case 1: Implementing an automatic emailing solution. **Activity** To build basic email solution based on dataset	**Individual Home Work** Self study: To build basic stream processing pipelines using TDD. https://spark.apache.org/docs/latest/streaming-programming-guide.html
Session 10 07-jun	**Content** Coding Hackathon. Case 2: Scraping the web in real time. **Activity** Coding Hackathon	**Individual Home Work** Improve the solution presented in class
Session 11 14-jun	**Content** Lambda Architecture Case 3: Spider to find articles in a newspaper **Activity** Build the solution	**Individual Home Work** Implement a similar solution to explore the web
Session 12 21-jun	**Content** Applying pyspark/real time processing to the final project Case 4: Studying twitter profiles	**Group/Individual Assignment (Graded)** Practical implementation of full Lambda architecture in a real world example // Integrate it in the final project.

The Schedule of activities can (and most likely, will) be modified according to the program needs. Assignments evaluation mode will be specified on Virtual Campus.

3 Qualification system

Participation: 10%
Active participation at class is expected during the subject.

Assignments: 90%
Continuous assessment by delivery of individual exercises and group exercises. All the assignments will have several deliver item different levels of complexity that define the final grade of it.
There will be a total of 3 assignments. Each of them weights a 30%.

Conditions to recover the subject
If the final mark is below 5, students will be able to deliver a final assignment.

***The 85% of attendance to each subject is required to pass the Master.**

4 Bibliography

Basic:

Big Data Principles and best practices of scalable realtime data systems	https://www.manning.com/books/big-data
PySpark	https://spark.apache.org/docs/latest/api/python/index.html
Streaming	https://spark.apache.org/docs/latest/streaming-programming-guide.html
Beautiful Soup	https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Complementary:

Real-Time Big Data Analytics	https://www.amazon.co.uk/Real-Time-Data-Analytics-Sumit-Gupta/dp/1784391409
Fast Data Processing with Spark	https://www.packtpub.com/big-data-and-business-intelligence/fast-data-processing-spark

Other academic resources:

Spark web site.	https://spark.apache.org/docs/latest/index.html
Applying the Lambda Architecture with Spark	https://databricks.com/session/applying-the-lambda-architecture-with-spark
Predictions on streaming	https://www.youtube.com/watch?v=fPlgoTLJh38
Lambda Architecture with spark	https://blog.knoldus.com/2017/01/31/twitters-tweets-analysis-using-lambda-architecture/ https://github.com/knoldus/Lambda-Arch-Spark https://www.youtube.com/watch?v=fPlgoTLJh38

Bibliography and other academic resources will be detailed and updated in the Campus.