Big Data Systems and techniques (6 units)

Dr. Dinos Arkoumanis: arkoumanis.dinos@gmail.com

Overview

Techniques and best practices for the development of production Big Data systems using Parquet and ORC columnar storage files in Hadoop and the Apache Spark data processing framework with SQL Query Engines (Spark SQL, Presto). Integration with latest parallel Machine Learning Frameworks. Cloud service technologies like Amazon EMR. Streaming and real time processing with Apache Storm + Kafka.

Key Outcomes

After completing the course, the students will be able to:

- Set up a Hadoop cluster from scratch
- Set up an Apache Spark cluster from scratch
- Import JSON/CSV data on an Apache Spark cluster and save it to files in HDFS with columnar formats Parquet and ORC with Java
- Map HDFS files as HIVE tables and query them with Spark SQL
- Set up a Presto SQL engine
- Set up an Apache Spark Cluster in Amazon AWS with EMR technology
- Process data with Apache Spark feed them to Spark ML machine learning algorithms and save trained models
- Use Apache Kafka to save datastream feeds to multiple storage systems
- Use Apache Storm for real time processing of datastream feeds

Requirements and Prerequisites

This course is hand-on and students will be evaluated by a final hands-on project.

The course does not assume any prior experience in Apache Spark, Hadoop or any other software that will be presented. However, basic knowledge of programming and computer science concepts is required. Basic knowledge of Java and SQL is necessary.

Students will need to bring their laptops in class in order to try out interactively the material being presented. Laptops should by strong enough to run at the same time 3 linux VMs with 2 GBs memory each.

Required Course Materials

There is no required textbook. All course materials will be provided in class and available for downloading.

Books

There are many books on the subject; the following selection provides a good foundation for those students who wish to delve deeper on the topics discussed in class:

- Fast Data Processing with Spark, 2nd Edition. Krishna Sankar, Holden Karau. Packt publishing
- Learning Spark Lightning-Fast Big Data Analysis.Holden Karau, Andy Konwinski, Patrick Wendell. Matei Zaharia. O'Reilly

Software/Computing requirements

A laptop strong enough to run at the same time 3 linux VMs with 2 GBs memory each.
Before each lecture a detailed list of the software packages needed to be installed will be emailed to the students.

Grading

Students will be graded on a final project. Each student will work alone on the final project. The project will have five requirements and each requirement will contribute 10% to 30% of the final score if it is successfully delivered.

Copied work from other students will be rejected automatically.

The course does not have exams.

Participation

All lectures will require the use of your laptop.

Attendance Requirements

This is a hands-on course. There is no point getting it if someone does not plan to attend it. Students are responsible for keeping up with the course material, including lectures, from the first day of this class, forward. It is the student's obligation to bring oneself up to date on any missed coursework.

Code of Ethics

Students will generally work alone but can cooperate and exchange ideas except on the final project. Copied code from another student in the final project will be penalized as discussed above.

Course Syllabus

The course comprises of ten units of three hours each and one unit for final project examination.

Unit 1: Introduction – Setup a hadoop cluster

Introduction about production Big Data systems.
Setup a linux VM for the cluster.
Hadoop basic concepts.
Setup a hadoop cluster from scratch.

Unit 2: Setup a Spark cluster over Hadoop

Spark basic concepts.
Setup a spark claster cluster from scratch on hadoop cluster from Unit 2.
Read JSON/CSV files with spark.
Columnar format basic concepts.
Write JSON/parquer/ORC files.

Units 3-4: SQL engines: Spark-SQL / Beeline / HIVE. Setup Presto SQL engine and performance comparison

SQL engines over HDFS files basic concepts.
Spark-SQL/beeline/HIVE.
Map JSON and columnar format files as HIVE tables in Spark SQL.
Execute basic queries in Spark-SQL.
Presto SQL engine setup.
SQL queries in Presto.

Performance comparison between Spark-SQL and Presto.

Units 5-6: Process data for spark ML algorithms / retrieve and save the results

Basic concepts and featuring algorithms of SparkML machine learning framework.
Create a JAVA project linked with Spark Libraries and use IntelliJ community edition as IDE.
Basic Spark RDD/Dataframes concepts.
Load data from and columnar format files using JAVA.
Transform data and create vectors.
Train algorithms machine language algorithms of SparkML with JAVA.
Store trained models with JAVA.
Load models and incremental training.
Validate new data with trained models.

Unit 7: Setup a cluster on AWS using EMR

Basic concepts of AWS.
Setting up a Spark/Hadoop cluster on AWS instances.
EMR basic concepts and differences from a typical Spark/Hadoop cluster.
Setting up a Spark/Hadoop cluster on AWS instances using Amazon EMR.
Connection of a Spark cluster with Tableau for visualizations using Tableau Spark SQL connection and Hive Thriftserver.

Units 8-9: Data streams. Use Apache Kafka to have data sterams stored to multiple systems / Apache storm to basic real-time processing

Setup a stream of data to the cluster from Unit 6.
Setup Apache Kafka to store stream to more than one system.
Setup Apache Storm to perform basic real time processing on the data streams.

Units 10: DevOps - Ansible - Final project discussion

DevOps basic concepts and Ansible scripts.
Discussion and assignment of the final project.

Unit 11: Final presentation of projects

Each student presents his course project to instructor. Instructor will ask details to determine that the project demonstrated the requested functionality and how well the student understood various topics discussed in the lesson.

The final project will involve:

- setting up a spark cluster on a different OS than Centos used on class
- data file manipulation
- machine learning algorithm execution
- DevOps and automation
- Bitbucket project and wiki