Cloud Infrastructure for Big Data

Description

The aim of this course is to describe the concrete applications offered by this technology and to discover the mechanisms, architectures, and good practices to be implemented.

Learning Objectives and Outcomes

• Understanding distributed computing.

• Know what strategies are used by the cluster manager to deal with large datasets.

• Know how to deploy algorithm on the cluster.

• Know how to setup and administrate the Big Data infrastructure (and tune performance for master, nodes, schedulers, workers).

• Usage of Amazon EMR to deal with spark implementation of ML problems.

• Dealing with different storages to optimize latency and costs.

• Use Cloud-enabled ETL to deal with data flow in the cloud environment.

Course Schedule and Contents

• Session 1 : 4 hours

- Big Data Overview

- Storage solutions

- Big Data Queries

• Session 2: 4 hours

- Setting up storage solutions (S3, Redshift)

- loading data to the cloud, advanced querying and switching strategies.

• Session 3 : 4 hours

- Understanding and setting up Apache Hadoop wit Amazon EMR

- Using Map Reduce with Hadoop

• Session 4 : 4 hours

- Introducing Spark in the cloud with Amazon EMR

- Dealing with Word processing problems with Spark

• Session 5 : 4 hours

- Using ETL to build dataflows between different kinds of storage

• Session 6: 4 hours

- Cost Management for Amazon

- Secure the solutions

- Real world go-live with the cloud

Grading

Quiz : 20 %

Practical Work : 30%

Assignment : 50%

Policies

- I expect you to turn-in your reports on time to receive proper credit/grade.
- Any work submitted must be your own.
- I expect everyone to contribute equally to group assignments
- Attendance in every class is expected and class participation and discussion is strongly encouraged.
- Late work will not be accepted unless prior arrangements have been made directly with me.
- Cases will be decided on an individual basis.

Good Luck!