1 LINGI20145 Cloud Computing: Project 1

Project 1 has the following objectives and will be graded based on the successful execution of each component.

1.1 Objectives

- Install and operate Apache Spark, a system for large-scale data processing.
- Deploy a cluster computing framework and learn how to launch applications with Mesos.

1.2 Tasks

1. **(Optional) Install Scala.**
   Ensure you have both Scala and sbt (Scala build tool) installed if you wish to write your Spark jobs in Scala.

2. (30%) **Create a basic Spark application.**
   Write a Spark application that executes locally and stores (pageName, pageViewCount) pairs on S3 based using the data set available here:

   `https://aws.amazon.com/datasets/wikipedia-traffic-statistics-v2/`

   This will require packaging the application in a way where `spark-submit` can be used to submit the job for processing.

3. (20%) **Extend application.**
   Modify the application to answer four (4) interesting questions, of your choice, about the data set. For example, you might ask (but don't use these!):

   - Top 5 pages by language (first field of data set.)
   - Top 5 pages viewed overall.

   Ideally, you'll have a single Spark job that will return the results of all of these questions together.

4. (10%) **Deploy application on Elastic Map Reduce with Amazon.**
   Create a script to deploy the job to Amazon Elastic Map Reduce and auto-terminate the cluster when processing is complete.

5. (10%) **Deploy Apache Mesos on Amazon and use Mesos to launch Spark and run the example.**
   You may want to deploy Mesos locally first, to learn how to deploy applications on it before moving to Amazon.

6. (20%) **Deploy a solution for persistence.**
   Rewrite your Spark application to persist the results at the end of the execution in a database of your choice. The database you choose should be deployable on Apache Mesos, and you should have a reason for selecting the database you choose. For example, one database you can choose is Apache Cassandra.

7. (10%) **Periodic scheduling.**
   Identify a way to periodically schedule your Spark job to run in Mesos, and have it periodically analyze the data and store it's results to the database when the cluster is online.

1.3 Deliverables

This project has two deliverables.

1. An application repository that contains the Spark application written in either Java or Scala that is able to be compiled into a JAR and deployed with `spark-submit`.

2. A script that automates deployment on Amazon of the following:

   - Apache Mesos
   - Apache Spark (inside of Mesos as an application)
   - Your Spark application
   - Your database of choice (inside of Mesos)
   - Your scheduling solution (inside of Mesos)