CS 494: Cloud Data Center Systems

Homework 1: Intro to Big Data

PART-2: A simple Spark application

Group-2
Suboor Jamal(673151084)
Bhuvan Jain(667704103)
Sidhant Pani(652453025)

INTRODUCTION

Apache Spark is an open-source distributed general-purpose cluster-computing framework. It provides an interface for programming entire clusters with various benefits such as data parallelism and fault tolerance. It has an architectural foundation which is a resilient distributed dataset (RDD), also it is just a read only multiset of items which are distributed over the clusters.

In the part of homework-1 we are running a sorting program using spark application. There are two parameters on which the data has to be sorted, firstly it gets sorted by country code alphabetically and then that sorted data is filtered according to their timestamp. The dataset was provided in the homework manual which is as follows - https://www2.cs.uic.edu/~brents/cs494-cdcs/data/export

STEPS INVOLVED IN RUNNING THE SPARK APPLICATION

Before starting to set up spark, there were various steps in order to set up the clusters(master and slave) according to the given configuration in the HW manual. We were supposed to use 4 node cluster from which one represented the master node and the other three represented the slave nodes. The first and the foremost process was to copy the public/private key pair of the master node to all the other three nodes. All the configurations were done in ~/.shh/ directory and the folder was authorized_keys where the key of master node was copied to all the other slave nodes. All these configurations were performed in the home directory only. Parallel ssh was required to copy any changes in the configuration to all the other three nodes automatically. Hadoop was downloaded on the master node and using parallel ssh it was replicated to the other nodes also. Apache Hadoop was used to save the dataset in order to run the sorting algorithm in part-2 and PageRank algorithm in part-3. Few configuration changes were done in the hadoop's core-site.xml and hdfs-site.xml files which were provided in the HW manual. The same configuration was also replicated to the other nodes. In order to run the spark application to perform sorting, Python code was written which performs two sorting algorithm first was to sort the country code and the second was to sort the sorted data according to the timestamp. In order to perform this the data was uploaded to HDFS and then using the proper configuration we downloaded the Spark files using the given commands and added the IP addresses of the slave in spark-2.2.0-bin-hadoop2.7/conf/slaves file. We then ran sudo spark-2.2.0-bin-hadoop2.7/sbin/start-all.sh.

After configuring hdfs, We then wrote the spark application for Sorting algorithm (using Python for Spark) for the given data set.ort the data firstly by the country code alphabetically (the third column) then by the timestamp (the last column). The output of the code was written in a csv file. There are two arguments, where the first is the path to the input of the file and the second path will be the output to the file.