CS 494: Cloud Data Center Systems

Homework 1: Intro to Big Data

PART-3: Pagerank

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

INTRODUCTION—

PageRank is an algorithm which is commonly used by Google search as it ranks the web pages in their search engine. It measures the importance of the website pages. By definition of google PageRank works by counting the number and link quality to a page which gives a rough estimate about the importance of the website.

The basic steps which is followed in the PageRank algorithm are listed below—

1. Set initial rank of each page to be 1.
2. On each iteration, each page contributes to its neighbors by rank(p)/ # of neighbors.
3. Update each page's rank to be 0.15 + 0.85 * (sum of contributions).
4. Go to next iteration.

In this part of homework-1 we were supposed to run PageRank algorithm on two datasets provided in the homework manual, 10 iterations were done on both the datasets.

Below find the code for the PageRank algorithm and the datasets were stored in HDFS—

RESULTS FOR DATASET_1—

1. **Task-1:** Written python code for page rank algorithm and obtained the weights, the elapsed time obtained was 181 seconds
2. **Task-2:** Added custom RDD partioning with 2 ports and obtained an elapsed time of 671 seconds

3. **Task-3:** For persisting the RDD as in memory-objects an elapsed time of 659 seconds was obtained
4. **Task-4:** By killing the worker process an elapsed time of 670 seconds was obtained

RESULTS FOR DATASET_2—

1. **Task-1:** Written python code for page rank algorithm and obtained the weights, the elapsed time obtained was 368 seconds
2. **Task-2:** Added custom RDD partioning with 2 ports and obtained an elapsed time of 1350 seconds
3. **Task-3:** For persisting the RDD as in memory-objects an elapsed time of 1325 seconds was obtained
4. **Task-4:** By killing the worker process an elapsed time of 1345 seconds was obtained

For dataset we were getting errors as there were UTF-8 strings which we were not able to convert to ASCII.

Reasons for difference in the elapsed time are—

In the first task the spark was only dividing/sharding the data into chunks and hence the best time was obtained in that. Whereas in RDD partioning the data was only divided into 2 big sets and hence the time was more. In the third part since persistence was required a copy was supposed to be saved but the partinoing of data was again done in multiple shards thus the elapsed time was less. In the final part worker process was killed and hence the maximum time was taken.