Lab 4 - Assignment

MM16B023[1]

[a]*Indian Institute of Technology Madras*

Keyword: pyspark, Hash, Pig, Hive, Yarn, HDFS

Abstract: This paper presents the solutions to second assignment of the Big Data Laboratory course (CS4830) at *IIT Madras*. All the notations used are as according with the textbook Mining of massive data sets by Anand Rajaraman

Problem 1

Write a spark code for executing the Hash example provided in slide 18 on Hashing from Lab 2 Presentation, on the public file: gs://big_data_a2/hash_file.txt.

a. Submit the python file with your code.
b. Also, provide the text file containing your output.

Solution:

```
### Written by H.Vishal MM16B023 at 4:02 pm on 22/02/2020 ###

import pyspark
import sys

if len(sys.argv) != 3:
  raise Exception("Exactly 2 arguments are required: <inputUri> <outputUri>")

# A neat way to store the input and output Uris
inputUri=sys.argv[1]
outputUri=sys.argv[2]

# Extract the lines
sc = pyspark.SparkContext()
lines = sc.textFile(sys.argv[1])

# Transform function definition
def transform_func(data):
        date,time,myid = data.split(" ")
        hr,mins,sec = time.split(":")
        if 0 <= int(hr) < 6:
                return '0-6'
            elif 6 <= int(hr) < 12:
                return '6-12'
            elif 12 <= int(hr) < 18:
                return '12-18'
            elif 18 <= int(hr) < 24:
                return '18-24'

# Input
data = lines.map(transform_func)
```

```
# Save output to file
counts = data.map(lambda frame: (frame,1)).reduceByKey(lambda count1, count2: count1 + count2)
counts.saveAsTextFile(sys.argv[2])
```

```
hvishal512@cloudshell:~ (hopeful-buckeye-266720)$ gsutil cat gs://mm16b023/output/*
('6-12', 13556831)
('18-24', 13592710)
('0-6', 13589270)
('12-18', 13585240)
hvishal512@cloudshell:~ (hopeful-buckeye-266720)$
```

Fig 1: Output

Problem 2

Provide a brief description of the functionality of the following services:
a. HDFS
b. Hive
c. Pig
d. Yarn

Solution:
a) The Hadoop Distributed file system (HDFS) is a high throughput access, high fault-tolerant DFS designated to run on commodity hardware. Some of the features are:

- It is suitable to handle large data sets. A typical file in HDFS is anywhere from GB to TB size.
- A simple coherency structure enables file once created modified, and closed need not be changed except for operations such as appends and truncates.
- Computation required by an application, especially for large data sets is less costlier when executed near the data which is being operated.
- Implementation of HDFS and MapReduce together is the hadoop programming framework

b) The Hive implements a restricted SQL form on top of Hadoop for the purpose of data query and analysis. It supports built-in user defined functions to manipulate the data mining tools such as dates and strings and a special type of indexing, *bitmap index* is used to provide acceleration. Also, all the SQL like queries are implicitly converted in the back-end to MapReduce or Spark jobs.

c) The Pig framework is an implementation of relational algebra on top of Hadoop in order to reduce execution time for data analytics jobs by performing ad-hoc analysis of large data sets. The various features of Pig are as follows:

- Dataflow language - Scope for reordering and optimizing a sequence of operations
- Quick start - User can run Pig queries directly over files from dump section of the search engine logs
- Nested data model - Useful to capture information about positional occurrences in collections of documents
- Allows efficient parallel evaluation since it is designed to deal with web-scale data.

d) Yarn - Yet Another Resource Negotiator is a package management tool to facilitate installation of packages by holding off resources on the cluster. One such for which it is frequently used is for installation of npm (Notoriously Psychic modules)
