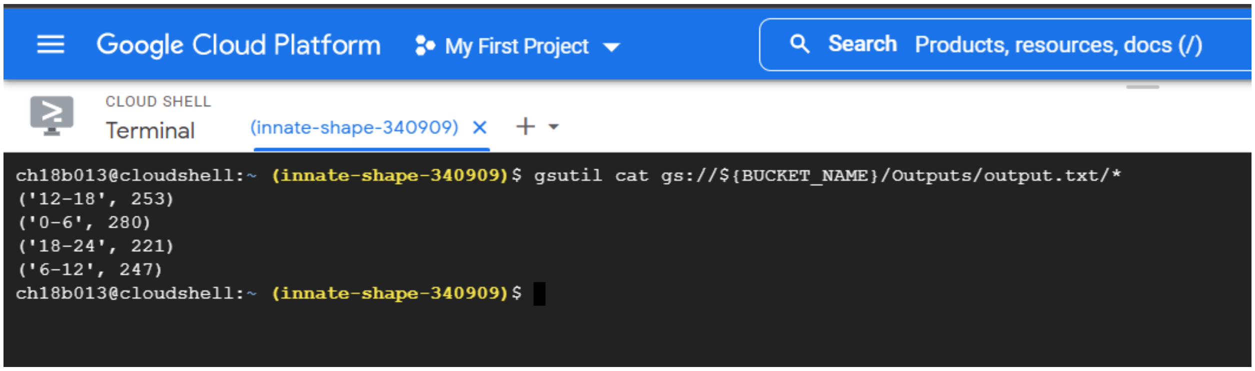
CS4830 - BIG DATA LABORATORY
VISHAL RISHI MK - CH18B013
LAB 4

1. Write a spark code for executing the Hash example provided in slide 14 on Hashing from Lab 1 Presentation, on the public file: ‘gs://bdl2022/lab4_dataset.csv’. You would have to find the number of user clicks between 0-6, 6-12, 12-18, and 18-24, as was discussed in the first class.

The python file ***‘dataproc.py’*** contains the code. The output text file **‘outputs.txt’** contains the number of user clicks between 0-6, 6-12, 12-18, and 18-24.



```
ch18b013@cloudshell:~ (innate-shape-340909)$ gsutil cat gs://${BUCKET_NAME}/Outputs/output.txt/*
('12-18', 253)
('0-6', 280)
('18-24', 221)
('6-12', 247)
ch18b013@cloudshell:~ (innate-shape-340909)$
```

Figure 1: The output obtained after running the data proc job

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

a. HDFS (Hadoop File System):
The Hadoop File System is developed using distributed file system design, run on commodity hardware, is highly fault-tolerant, and is designed using low-cost hardware. HDFS holds a very large amount of data and provides easier access. This is done through storage across multiple machines. Files are stored in a redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available for parallel processing.

Features of HDFS:

- HDFS is suitable for distributed storage and processing.
- Hadoop provides a command interface to interact with HDFS.
- The built-in servers of the name node and data node help users to easily check the status of the cluster.
- Provides streaming access to file system data.
- Provides file permissions and authentication.

b. Hive:

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data and makes querying and analyzing easy.

Features of Hive:

- It stores schema in a database and processes data into HDFS.
- It is designed for OLAP.
- It provides SQL-type language for querying called HiveQL or HQL.
- It is familiar, fast, scalable, and extensible.

c. Pig:

Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows. Pig is generally used with Hadoop; we can perform all the data manipulation operations in Hadoop using Apache Pig.

Features of Pig:

- Rich set of operators − It provides many operators to perform operations like join, sort, filer, etc.
- Ease of programming − Pig Latin is similar to SQL and it is easy to write a Pig script if you are good at SQL.
- Optimization opportunities − The tasks in Apache Pig optimize their execution automatically, so the programmers need to focus only on the semantics of the language.
- Extensibility − Using the existing operators, users can develop their own functions to read, process, and write data.
- UDF's − Pig provides the facility to create User-defined Functions in other programming languages such as Java and invoke or embed them in Pig Scripts.
- Handles all kinds of data − Apache Pig analyses all kinds of data, both structured as well as unstructured. It stores the results in HDFS.

d. YARN:

The Apache Hadoop YARN stands for Yet Another Resource Negotiator. It is a very efficient technology to manage the Hadoop cluster. YARN is a completely new way of processing data and is now rightly at the center of The Hadoop architecture. Using this revolutionary technology, it is possible to stream real-time, use interactive SQL, process data using multiple engines, manage data using batch processing on a single platform, and so on.

Features of Yarn:

- High degree of compatibility - The applications that are created using the MapReduce framework can easily run YARN in a seamless manner
- Better cluster utilization - YARN allocates the cluster resources in an efficient and dynamic manner and due to this the utilization is much better compared to the previous version of Hadoop.

- Utmost scalability - As and when the number of nodes in the Hadoop cluster expands, the YARN Resource Manager ensures that the requirements are met and the processing power of the data center does not face any hurdles.
- Multi-tenancy - The various engines that access data on the Hadoop cluster can seamlessly work thanks to YARN being a highly versatile technology.