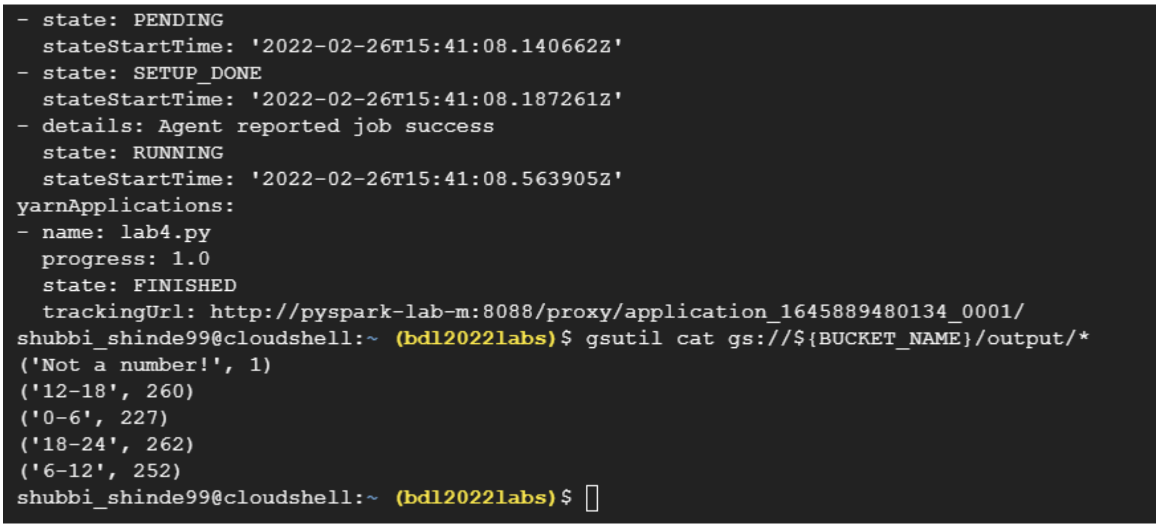
SHINDE SHUBHAM SUNIL **Lab-4 Assignment**
smail: me18b183@smail.iitm.ac.in **Roll No.:** ME18B183
Course: CS4830 - Big Data Laboratory **Semester:** JAN-MAY 2022
Instructor: Prof. Balaraman Ravindran **Due Date:** 26th February, 2022

Problem 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.

Solution: Output of the spark code:



```
- state: PENDING
  stateStartTime: '2022-02-26T15:41:08.140662Z'
- state: SETUP_DONE
  stateStartTime: '2022-02-26T15:41:08.187261Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2022-02-26T15:41:08.563905Z'
yarnApplications:
- name: lab4.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://pyspark-lab-m:8088/proxy/application_1645889480134_0001/
shubbi_shinde99@cloudshell:~ (bdl2022labs)$ gsutil cat gs://${BUCKET_NAME}/output/*
('Not a number!', 1)
('12-18', 260)
('0-6', 227)
('18-24', 262)
('6-12', 252)
shubbi_shinde99@cloudshell:~ (bdl2022labs)$ 
```

Figure 1: Screenshot of the output as displayed in the console

Python code (*lab4.py*):

```python
#!/usr/bin/env python

import pyspark
import sys

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

inputUri=sys.argv[1]
outputUri=sys.argv[2]

def hour_slotting(datetime):
    time = datetime.split(" ")[1]
    hour = time.split(":")[0]
    if hour == 'ID':
        return 'Not a number!'
    if 0 <= int(hour) < 6:
        return '0-6'
    if 6 <= int(hour) < 12:
        return '6-12'
    if 12 <= int(hour) < 18:
        return '12-18'
    if 18 <= int(hour) < 24:
        return '18-24'

sc = pyspark.SparkContext()
userclicks = sc.textFile(sys.argv[1])
count = userclicks.map(hour_slotting)
clicks_keeper = count.map(lambda Time: (Time,1)).reduceByKey(lambda x1, x2: x1 + x2)
clicks_keeper.coalesce(1).saveAsTextFile(sys.argv[2])
```

Problem 2
Provide a brief description of the functionality of the following services:
(a) HDFS; (b) Hive; (c) Pig; and (d) YARN.

Solution:

- **HDFS**
  HDFS is an abbreviation for Hadoop Distributed File System. It supports parallel processing of applications and was created as part of the Apache Nutch online search engine project's infrastructure. It is a highly fault tolerant file system, designed using low cost hardware. It has a large storage capacity and makes data retrieval easy. These files are spread over numerous machines to be able to store huge amounts of data. A duplicate of each file is created to guard against any data loss due to a system failure.

- **Hive**
  Hive is a data warehouse software that facilitates reading, writing, and managing large data-sets residing in distributed storage. Since most data warehousing applications work with SQL-based querying languages, Hive aids portability of SQL-based applications to Hadoop. It was created by Facebook (now Meta) in 2010, currently utilised and developed by other companies like Netflix.

- **Pig**
  Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs and the infrastructure for evaluating them. Pig programs are notable for their structure, which allows for significant parallelization and, as a result, the handling of very large data sets. It has its own platform language known as Pig Latin, that provides us ease of programming, optimization opportunities and extensibilty.

- **YARN**
  YARN is an abbreviation for Yet Another Recourse Navigator. It is a resource management and job scheduling technology that is responsible for assigning system resources to the various Hadoop cluster applications and scheduling tasks to run on different cluster nodes. It provides several data processing engines to operate on and process the data stored in HDFS, such as graph processing, interactive processing, stream processing, and batch processing.

∗ ∗ End of Assignment ∗ ∗