**School of Computing and Engineering: University of Missouri – Kansas City**

**Course Name: CS5525 – Cloud Computing**

**Assignment 5 (Spark and Dataproc)**

**Name: Rakesh Rao Thakkalapelly**

**Email:** rthf8@umsystem.edu

**Student Id**: 16338585

**1. Installing Cloud SDK:**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Google Cloud Command Line Interface(CLI):**

Text

Description automatically generated

Text

Description automatically generated

**2. Dataproc:**

**Enabling Dataproc API in gcloud console:**

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

**Creating a Cluster using command line interface & running a Spark Job for Calculating Pi Value:**

Text

Description automatically generated

Text

Description automatically generated

A screenshot of a computer

Description automatically generated

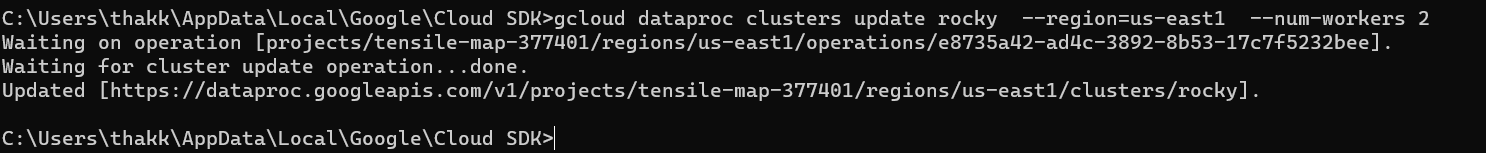
Cluster “rocky” is created successfully with 3 nodes (1 master + 2 workers).

Graphical user interface, application

Description automatically generated

**Pi value is** : 3.1415108714151088

**Updating a cluster by increasing or decreasing number of workers:**



Graphical user interface, text, application, email

Description automatically generated

**Deleting a cluster:**

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**3. BigQuery:**

Graphical user interface, application, Word

Description automatically generated

**Querying a public dataset:**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Preview the data from the public dataset:**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Query to list 500 results:**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Creating a dataset “babynames” in the project:**

**Graphical user interface, application

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Creating a custom table “names\_2020” and loading the data into the dataset:**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

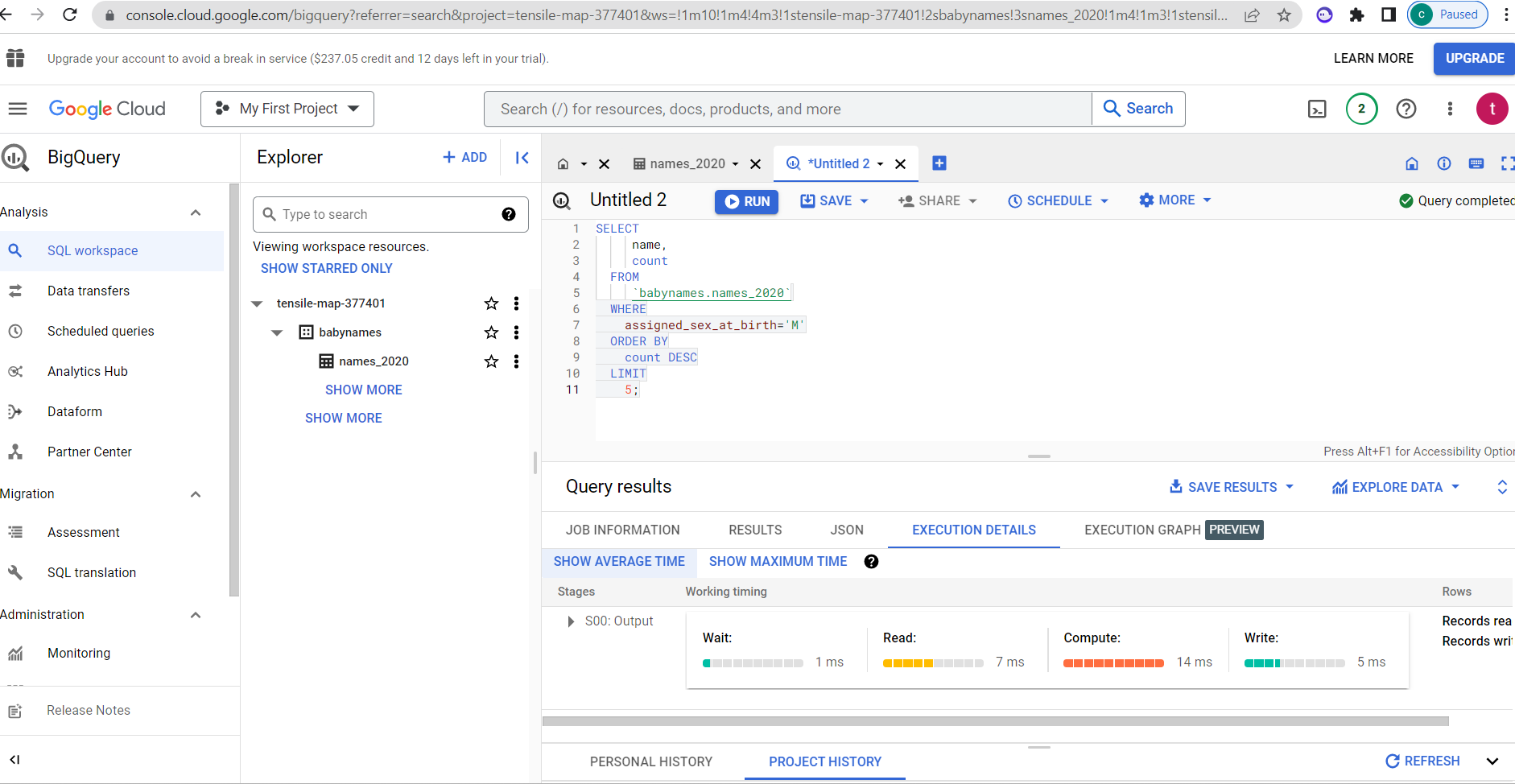
Graphical user interface, application

Description automatically generated

**Querying a table for a limit of 5 results:**

Graphical user interface, text, application, email

Description automatically generated



**Querying a table for a limit of 20 results:**

**Graphical user interface, text, application, email

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**4. Briefly write any challenges (or comments) while doing this assignment.**

**Ans:** Creating a cluster can be a complex and challenging task that requires expertise in multiple areas. One of the main challenges in creating a cluster is configuring the different components, including hardware, software, networking, and security, to work together seamlessly. Ensuring compatibility between these components and designing a cluster that can scale efficiently to meet changing demands are also significant challenges. Allocating resources among the different nodes of the cluster, configuring the network topology, and ensuring the security of the cluster are other key challenges.