**Name: Abhishek Obla Hema**

**Email:** [**oh\_abhishek@yahoo.com**](mailto:oh_abhishek@yahoo.com)

**Spark Assignment-2(Movie Data Analysis) Documentation**

This file details and describes all the attached files for the Spark Assignment-2

**Tools Used:**

1. Python3 – Microsoft VScode
2. Apache Spark (On GCP Cluster)
3. GCP services – DataProc/GCS
4. Jupyter Lab
5. Apache Hive

**Files Attached:**

1. Spark\_Ass2.pdf – This file
2. Spark\_MovieRating.– Pyspark File that details all the analysis

**Process and File Descriptions:**

**Step 1:**

I placed the three csv files in the HDFS location and ingested them into their respective spark dataframes. The three frames being

* Movies
* Ratings
* Tags

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Step 2:**

I then used spark SQL to query the dataframes to get the required outputs. This involved creating tempviews of movies,ratings and tags

A screenshot of a computer error

Description automatically generated

**Step 3:**

I made sure to save the output in a HDFS location. I also checked the files in the HDFS Namenode using the UI.

