**COS80023 Big Data – Lab 3**

**Lab 3: Pass Task 3 – Extraction, Transformation and Loading**

**Student Name: Arun Ragavendhar Arunachalam Palaniyappan**

**Student ID: 104837257**

In this task, I used Hive to transform and process accident data stored in Azure storage through a Hadoop cluster. First, I uploaded the CSV file and the Hive script into the Blob storage. Then, using the Hive script (staging.hql), I created two tables: accidents\_raw and accidents\_in\_hive. These tables allowed the raw data to be structured in a relational format inside Hive so it could be queried more easily. After the tables were created, I ran a Hive query to extract useful information. The query grouped the data by day\_week\_description and calculated the total number of vehicles involved in accidents for each day of the week. It also cleaned the text by removing unwanted quotation marks from the day names.

Finally, the results of this query were written into a new output directory (/accidents/output) in HDFS, stored as tab-separated values. This meant Hive had successfully taken raw CSV data, structured it, applied cleaning and grouping, and then output a processed dataset that is easier to analyse.

 



