COS80023 Big Data – Lab 3

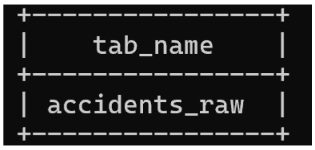
Lab 3: Pass Task 3 – Extraction, Transformation and Loading

Student Name: Arun Ragavendhar Arunachalam Palaniyappan

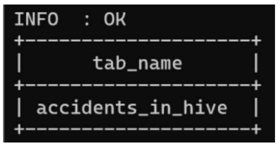
Student ID: 104837257

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.

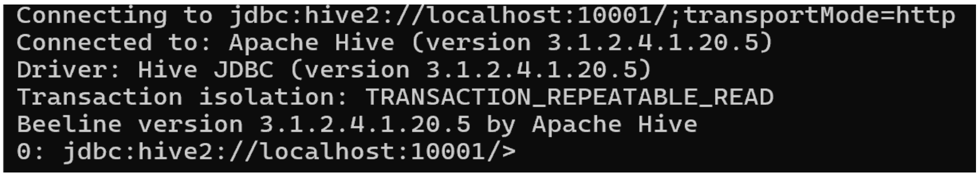
Finally, the results of this query were written into a new output directory (/accidents/output) in HDFS, stored as tab-separated values. This meant Hive had successfully taken raw CSV data, structured it, applied cleaning and grouping, and then output a processed dataset that is easier to analyse.



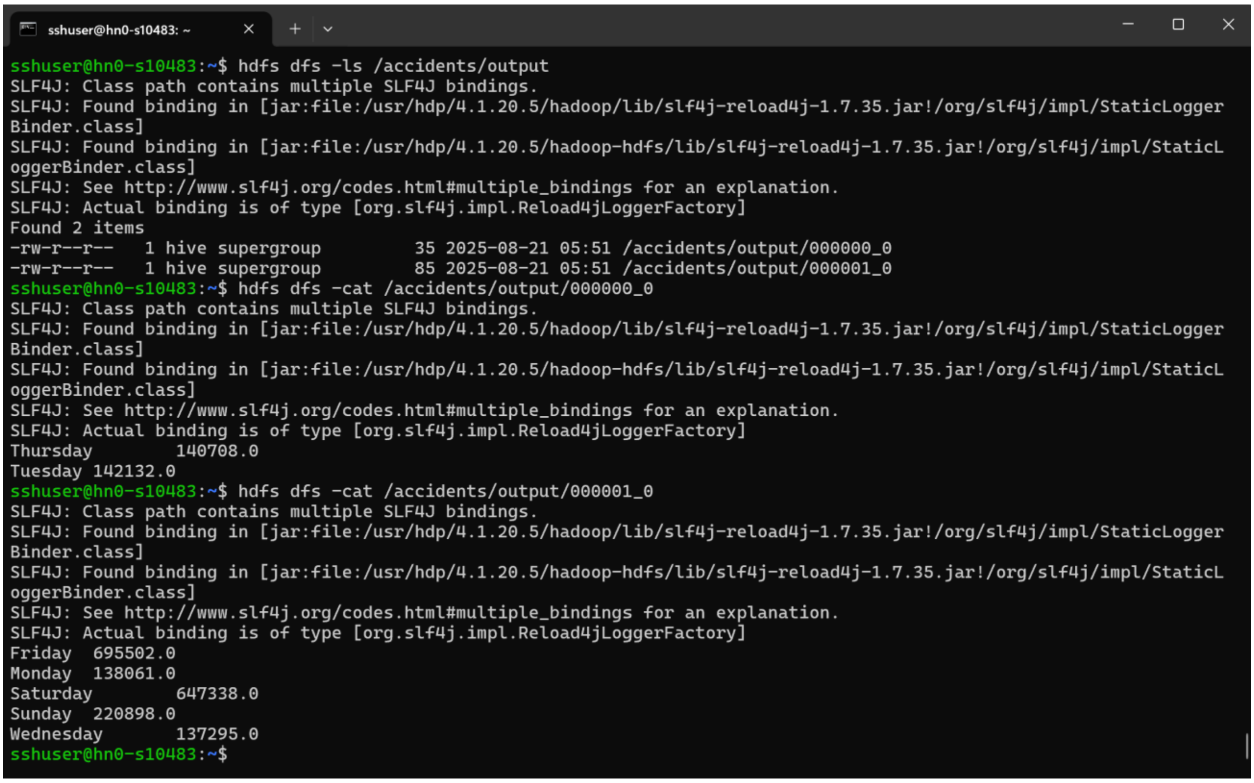
```
+-----------------+
|    tab_name     |
+-----------------+
| accidents_raw   |
+-----------------+
```



```
INFO  : OK
+--------------------+
|      tab_name      |
+--------------------+
| accidents_in_hive  |
+--------------------+
```



```
Connecting to jdbc:hive2://localhost:10001/;transportMode=http
Connected to: Apache Hive (version 3.1.2.4.1.20.5)
Driver: Hive JDBC (version 3.1.2.4.1.20.5)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.2.4.1.20.5 by Apache Hive
0: jdbc:hive2://localhost:10001/>
```



```
sshuser@hn0-s10483:~$ hdfs dfs -ls /accidents/output
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/4.1.20.5/hadoop/lib/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticLogger
Binder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/4.1.20.5/hadoop-hdfs/lib/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticL
oggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
Found 2 items
-rw-r--r--   1 hive supergroup         35 2025-08-21 05:51 /accidents/output/000000_0
-rw-r--r--   1 hive supergroup         85 2025-08-21 05:51 /accidents/output/000001_0
sshuser@hn0-s10483:~$ hdfs dfs -cat /accidents/output/000000_0
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/4.1.20.5/hadoop/lib/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticLogger
Binder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/4.1.20.5/hadoop-hdfs/lib/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticL
oggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
Thursday        140708.0
Tuesday 142132.0
sshuser@hn0-s10483:~$ hdfs dfs -cat /accidents/output/000001_0
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/4.1.20.5/hadoop/lib/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticLogger
Binder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/4.1.20.5/hadoop-hdfs/lib/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticL
oggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
Friday  695502.0
Monday  138061.0
Saturday        647338.0
Sunday  220898.0
Wednesday       137295.0
sshuser@hn0-s10483:~$
```