

Assignment 28

Delayed_Flights.csv Datasets (Downloaded from https://drive.google.com/file/d/0B_Qjau8wv1KoWTVDUVF0dzIJNWM/view)

There are 29 columns in this dataset. Some of them have been mentioned below:

- Year: 1987 – 2008
- Month: 1 – 12
- FlightNum: Flight number
- Canceled: Was the flight canceled?
- CancellationCode: The reason for cancellation.

Problem Statement 1

- Find out the top 5 most visited destinations.

Problem Statement 2

- Which month has seen the most number of cancellations due to bad weather?

Problem Statement 3

- Which route (origin & destination) has seen the maximum diversion?

Complete Code

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._
object Assignment_28 {
  case class Flight(Month:Int,Origin: String,Dest:String,CancellationCode:String,Diverted:Int)
  def main(args: Array[String]): Unit = {
    println("hey scala")
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Assignment 28")
      .config("spark.some.config.option", "some-value")
```

```

    .getOrCreate()

    println("Spark Session Object created")

    val data =
spark.sparkContext.textFile("E:\\Avani\\Acadgild\\Datasets\\DelayedFlights.csv");

    val header = data.first()

    val data1 = data.filter(x => x != header)

    val num = println("aviation data->>" + data1.count())

    println("removed header")

import spark.implicits._

    val s = data1.map(x => x.split(",")).map(x =>
Flight(x(2).toInt,x(17),x(18),x(23),x(24).toInt)).toDF

    println("Aviation data")

    s.registerTempTable("fly")

    println("temp table created")

    val Destinations = spark.sql("SELECT Dest,count(Dest) c from fly group by Dest order by c
desc").show(5)

    println("Task1 output")

    val cancel = spark.sql("SELECT month,count(CancellationCode) c from fly where
CancellationCode = 'B' group by month order by c desc").show(1)

    println("Task2 output")

    val diversion = spark.sql("SELECT Origin, Dest, count(Diverted)d from fly WHERE diverted =
1 GROUP BY origin,dest ORDER BY d DESC").show(1)

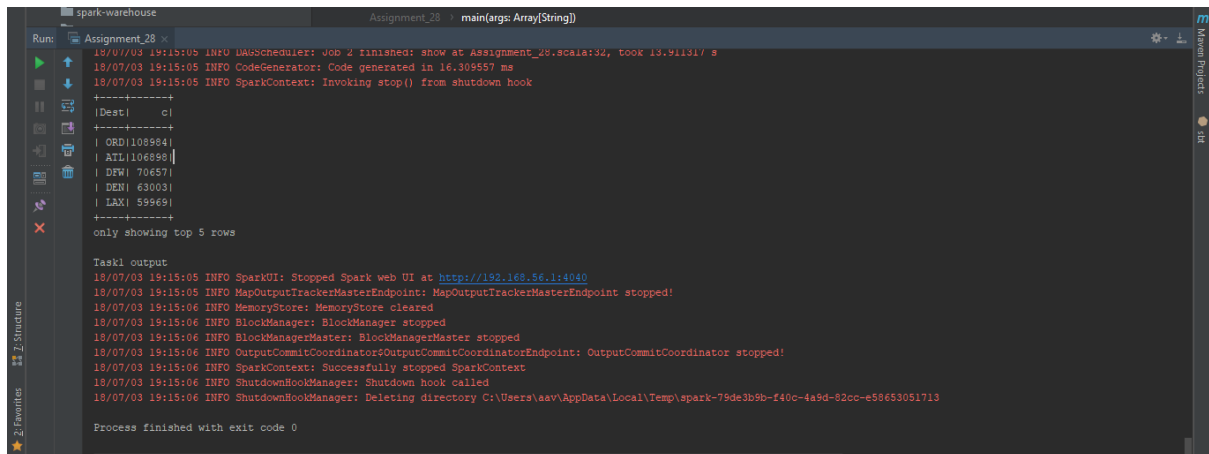
    println("Task3 output")

}

}

```

Screenshots

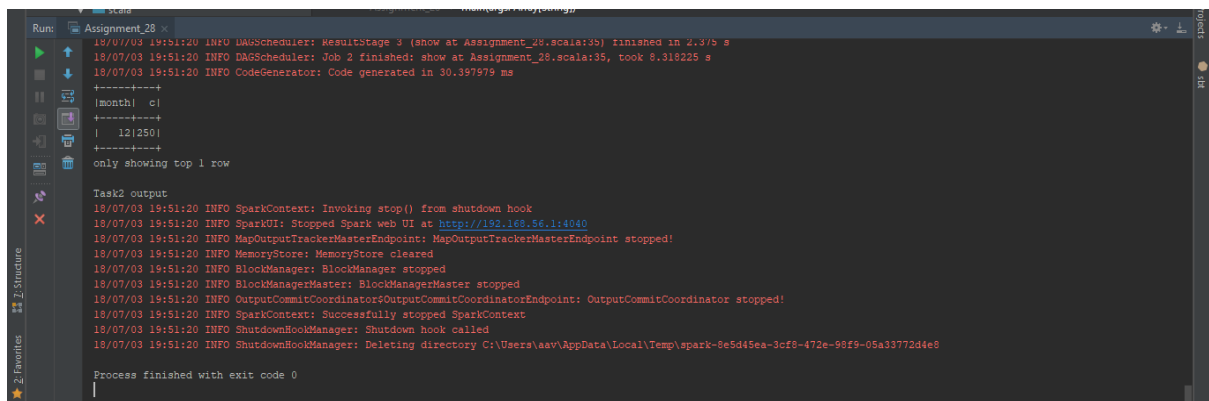


The screenshot shows a Spark job run in Databricks. The job is named "Assignment_28" and is running on a cluster named "main(args: Array[String])". The output of Task1 is displayed, showing a table with 5 rows and 2 columns: "Dest" and "c". The table contains the following data:

Dest	c
ORD	108984
ATL	106898
DFW	70657
DEN	63003
LAX	59969

The output also shows the Spark web UI at <http://192.168.56.1:4040> and the SparkContext successfully stopped.

Task1

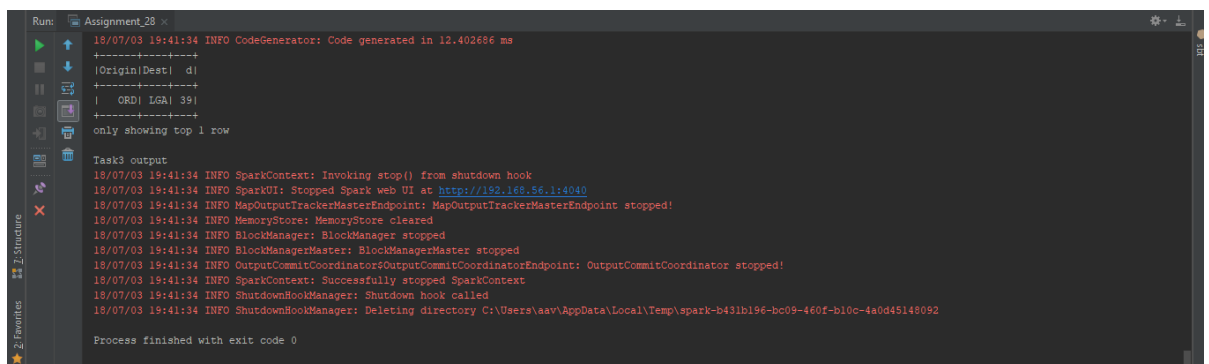


The screenshot shows a Spark job run in Databricks. The job is named "Assignment_28" and is running on a cluster named "main(args: Array[String])". The output of Task2 is displayed, showing a table with 1 row and 2 columns: "month" and "c". The table contains the following data:

month	c
12	2501

The output also shows the Spark web UI at <http://192.168.56.1:4040> and the SparkContext successfully stopped.

Task2



The screenshot shows a Spark job run in Databricks. The job is named "Assignment_28" and is running on a cluster named "main(args: Array[String])". The output of Task3 is displayed, showing a table with 1 row and 3 columns: "Original", "Dest", and "d". The table contains the following data:

Original	Dest	d
ORD	LGA	391

The output also shows the Spark web UI at <http://192.168.56.1:4040> and the SparkContext successfully stopped.

Task3