

1. Copy delayed_flights.csv to HDFS using

```
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
cloudera@quickstart:~$ hadoop dfs -ls /tmp
DEPRECATED: Use of this script to execute dfs command is deprecated.
Instead use the hdfs command for it.

Found 14 items
-rw-r--r-- 1 cloudera supergroup 602 2017-10-17 06:02 /tmp/17.2.Dataset.txt
-rw-r--r-- 1 cloudera supergroup 2793 2017-11-19 03:58 /tmp/Delayedflights.description.csv
-rw-r--r-- 1 cloudera supergroup 899 2017-11-11 05:50 /tmp/S18.Dataset.Holidays.txt
-rw-r--r-- 1 cloudera supergroup 40 2017-11-11 05:50 /tmp/S18.Dataset.Transport.txt
-rw-r--r-- 1 cloudera supergroup 108 2017-11-11 05:51 /tmp/S18.Dataset.User.Details.txt
-rw-r--r-- 1 cloudera supergroup 2025764 2017-11-19 06:21 /tmp/demonetization_tweets_filter.csv
drwxr-xrwt - mapred mapred 0 2017-07-19 05:34 /tmp/hadoop-yarn
drwxr-xrwx - hive supergroup 0 2017-10-06 03:59 /tmp/hive
drwxr-xrwt - mapred hadoop 0 2017-07-19 05:36 /tmp/logs
-rw-r--r-- 1 cloudera supergroup 114 2017-10-30 03:49 /tmp/session17.assignment1.txt
drwxr-xr-- 1 cloudera supergroup 0 2017-11-15 01:10 /tmp/session19task3.parquet
-rw-r--r-- 1 cloudera supergroup 1017 2017-11-12 21:16 /tmp/sports.data.txt
drwxr-xr-x - cloudera supergroup 0 2017-10-26 06:51 /tmp/temp-1468848784
-rw-r--r-- 1 cloudera supergroup 2543 2017-11-19 00:29 /tmp/tweets.json
[cloudera@quickstart Downloads]$ hadoop dfs -put DelayedFlights.csv /tmp/
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
[cloudera@quickstart Downloads]$
```

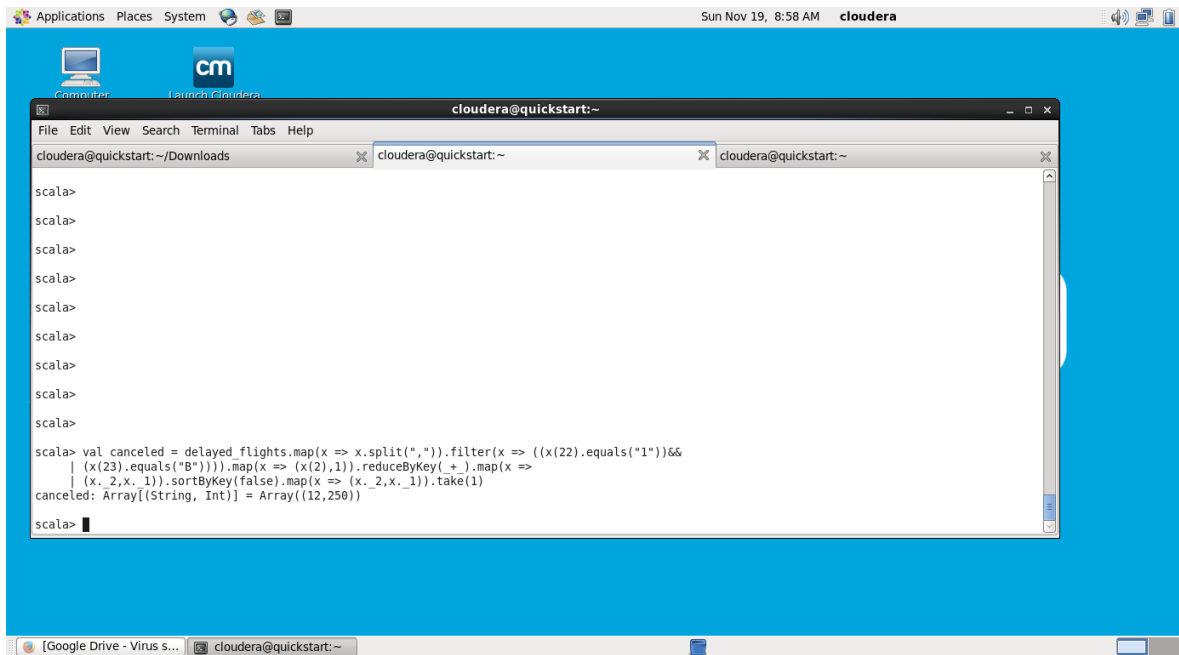
Problem statement 1:

1. Find out the top 5 most visited destinations.

```
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~/Downloads
scala>
scala>
scala>
scala>
scala>
scala>
scala>
scala> val delayedFlights = sc.textFile("hdfs://quickstart.cloudera:8020/tmp/DelayedFlights.csv")
delayedFlights: org.apache.spark.rdd.RDD[String] = hdfs://quickstart.cloudera:8020/tmp/DelayedFlights.csv MapPartitionsRDD[189] at textFile at <console>:43
scala> val mapping = delayedFlights.map(x => x.split(",")).map(x => (x(18),1)).filter(x =>
  | x._1!=null).reduceByKey(_+_).map(x => (x._2,x._1)).sortByKey(false).map(x => (x._2,x._1)).take(5)
mapping: Array[(String, Int)] = Array((ORD,108984), (ATL,106898), (DFW,70657), (DEN,63003), (LAX,59969))
scala>
```

Problem statement 2

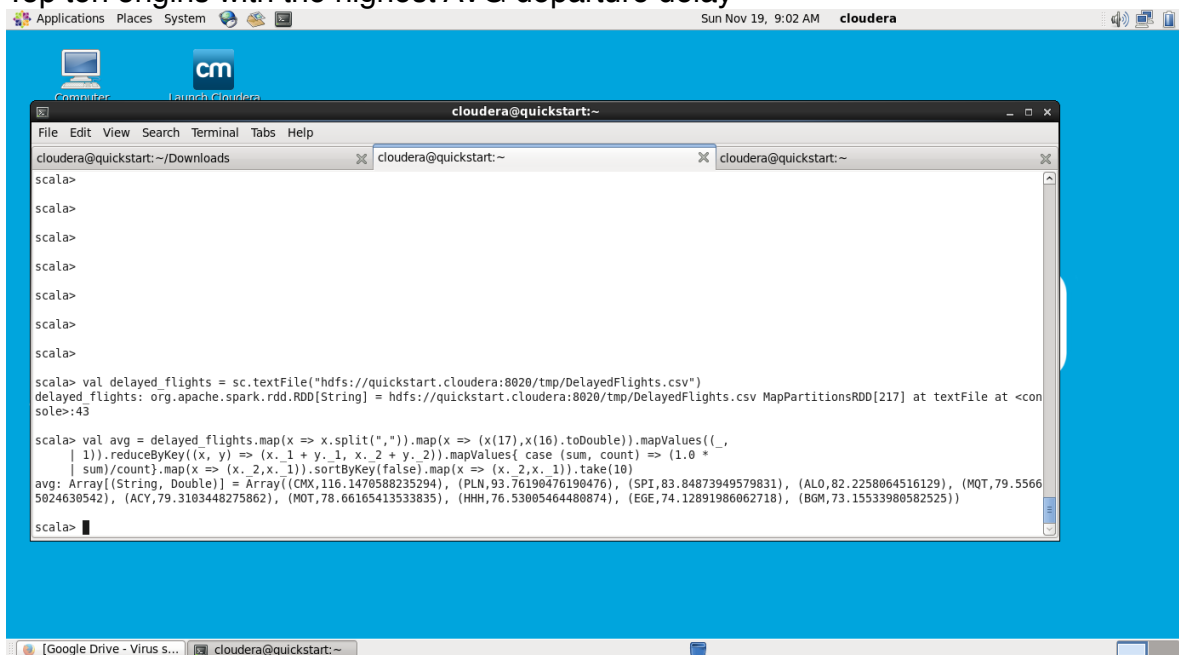
Which month has seen the most number of cancellations due to bad weather?



The screenshot shows a Cloudera Quickstart desktop environment with a terminal window open. The terminal displays the following Scala code and its output:

```
scala>
scala>
scala>
scala>
scala>
scala>
scala>
scala>
scala>
scala> val canceled = delayed_flights.map(x => x.split(",")).filter(x => ((x(22).equals("1")) &&
  | (x(23).equals("B")))).map(x => (x(2),1)).reduceByKey( + ).map(x =>
  | (x._2,x._1)).sortByKey(false).map(x => (x._2,x._1)).take(1)
canceled: Array[(String, Int)] = Array((12,250))
scala>
```

Problem statement 3: Top ten origins with the highest AVG departure delay



The screenshot shows a Cloudera Quickstart desktop environment with a terminal window open. The terminal displays the following Scala code and its output:

```
scala>
scala>
scala>
scala>
scala>
scala>
scala>
scala> val delayed_flights = sc.textFile("hdfs://quickstart.cloudera:8020/tmp/DelayedFlights.csv")
delayed_flights: org.apache.spark.rdd.RDD[String] = hdfs://quickstart.cloudera:8020/tmp/DelayedFlights.csv MapPartitionsRDD[217] at textFile at <console>:43
scala> val avg = delayed_flights.map(x => x.split(",")).map(x => (x(17),x(16).toDouble)).mapValues(_ =>
  | 1)).reduceByKey((x,y) => (x._1 + y._1, x._2 + y._2)).mapValues{ case (sum, count) => (1.0 *
  | sum)/count}.map(x => (x._2,x._1)).sortByKey(false).map(x => (x._2,x._1)).take(10)
avg: Array[(String, Double)] = Array((CMX,116.1470580235294), (PLN,93.76190476190476), (SPI,83.84873949579831), (ALO,82.2258064516129), (MOT,79.5566
5824630542), (ACY,79.3103448275862), (MOT,78.66165413533635), (HHH,76.53005464480874), (EGE,74.1289196062718), (BGM,73.15533980582525))
scala>
```

Problem statement 4 Which route (origin & destination) has seen the maximum diversion?

