FINISHED

```
%sh
wget http://stat-computing.org/dataexpo/2009/2007.csv.bz2 -O /home/scarface/Desktop/sem-3/capstone/Prithvi/Feb2L
wget http://stat-computing.org/dataexpo/2009/2008.csv.bz2 -O /home/scarface/Desktop/sem-3/capstone/Prithvi/Feb2L
wget ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/2007.csv.gz -O /home/scarface/Desktop/sem-3/capstone/Pr
wget ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/2008.csv.gz -O /home/scarface/Desktop/sem-3/capstone/Pr
```

```
--2017-02-02 18:56:32--  http://stat-computing.org/dataexpo/2009/2007.csv.bz2
Resolving stat-computing.org (stat-computing.org)... 54.231.176.191
Connecting to stat-computing.org (stat-computing.org)|54.231.176.191|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 121249243 (116M) [application/x-bzip2]
Saving to: ‘/home/scarface/Desktop/sem-3/capstone/Prithvi/Feb2Lab/flights_2007.csv.bz2’
     0K .......... .......... .......... .......... ..........  0%  210K 9m24s
    50K .......... .......... .......... .......... ..........  0%  321K 7m46s
   100K .......... .......... .......... .......... ..........  0%  580K 6m19s
   150K .......... .......... .......... .......... ..........  0%  318K 6m17s
   200K .......... .......... .......... .......... ..........  0%  607K 5m40s
   250K .......... .......... .......... .......... ..........  0% 8.09M 4m46s
   300K .......... .......... .......... .......... ..........  0%  658K 4m30s
   350K .......... .......... .......... .......... ..........  0%  683K 4m18s
   400K .......... .......... .......... .......... ..........  0% 11.3M 3m50s
   450K .......... .......... .......... .......... ..........  0%  636K 3m46s
   500K .......... .......... .......... .......... ..........  0%  149M 3m25s
   550K                                                         0%  639K 3m24s
```

Took 15 min 34 sec. Last updated by anonymous at February 02 2017, 7:12:05 PM.

FINISHED

```
%spark

import org.apache.spark.rdd._
import scala.collection.JavaConverters._
import au.com.bytecode.opencsv.CSVReader

import java.io._
import org.joda.time._
import org.joda.time.format._
import org.joda.time.format.DateTimeFormat
import org.joda.time.DateTime
import org.joda.time.Days
```

```
import org.apache.spark.rdd._
import scala.collection.JavaConverters._
import au.com.bytecode.opencsv.CSVReader
import java.io._
import org.joda.time._
import org.joda.time.format._
import org.joda.time.format.DateTimeFormat
import org.joda.time.DateTime
import org.joda.time.Days
```

Took 33 sec. Last updated by anonymous at February 02 2017, 8:42:44 PM.

FINISHED

```
case class DelayRec(year: String,
                    month: String,
                    dayOfMonth: String,
                    dayOfWeek: String,
                    crsDepTime: String,
                    depDelay: String,
                    origin: String,
                    distance: String,
                    cancelled: String) {

    val holidays = List("01/01/2007", "01/15/2007", "02/19/2007", "05/28/2007", "06/07/2007", "07/04/2007",
      "09/03/2007", "10/08/2007" ,"11/11/2007", "11/22/2007", "12/25/2007",
      "01/01/2008", "01/21/2008", "02/18/2008", "05/22/2008", "05/26/2008", "07/04/2008",
      "09/01/2008", "10/13/2008" ,"11/11/2008", "11/27/2008", "12/25/2008")

    def gen_features: (String, Array[Double]) = {
      val values = Array(
        depDelay.toDouble,
        month.toDouble,
        dayOfMonth.toDouble,
        dayOfWeek.toDouble,
        get_hour(crsDepTime).toDouble,
        distance.toDouble,
        days_from_nearest_holiday(year.toInt, month.toInt, dayOfMonth.toInt)
      )
      new Tuple2(to_date(year.toInt, month.toInt, dayOfMonth.toInt), values)
    }

    def get_hour(depTime: String) : String = "%04d".format(depTime.toInt).take(2)
    def to_date(year: Int, month: Int, day: Int) = "%04d%02d%02d".format(year, month, day)

    def days_from_nearest_holiday(year:Int, month:Int, day:Int): Int = {
      val sampleDate = new org.joda.time.DateTime(year, month, day, 0, 0)

      holidays.foldLeft(3000) { (r, c) =>
        val holiday = org.joda.time.format.DateTimeFormat.forPattern("MM/dd/yyyy").parseDateTime(c)
        val distance = Math.abs(org.joda.time.Days.daysBetween(holiday, sampleDate).getDays)
        math.min(r, distance)
      }
    }
```

```
warning: Class org.joda.convert.FromString not found - continuing with a stub.
warning: Class org.joda.convert.ToString not found - continuing with a stub.
warning: Class org.joda.convert.ToString not found - continuing with a stub.
warning: Class org.joda.convert.FromString not found - continuing with a stub.
warning: Class org.joda.convert.ToString not found - continuing with a stub.
warning: Class org.joda.convert.FromString not found - continuing with a stub.
warning: Class org.joda.convert.ToString not found - continuing with a stub.
defined class DelayRec
```

Took 4 sec. Last updated by anonymous at February 02 2017, 8:42:51 PM.

FINISHED ▷

```
def prepFlightDelays(infile: String): RDD[DelayRec] = {
    val data = sc.textFile(infile)

    data.map { line =>
      val reader = new CSVReader(new StringReader(line))
      reader.readAll().asScala.toList.map(rec => DelayRec(rec(0),rec(1),rec(2),rec(3),rec(5),rec(15),rec(16),rec
    }.map(list => list(0))
    .filter(rec => rec.year != "Year")
    .filter(rec => rec.cancelled == "0")
    .filter(rec => rec.origin == "ORD")
}
```

```
prepFlightDelays: (infile: String)org.apache.spark.rdd.RDD[DelayRec]
```

Took 2 sec. Last updated by anonymous at February 02 2017, 8:42:59 PM.

FINISHED ▷

```
val data_2007tmp = prepFlightDelays("/home/scarface/Desktop/sem-3/capstone/Prithvi/Feb2Lab/2007.csv")
val data_2007 = data_2007tmp.map(rec => rec.gen_features._2)
val data_2008 = prepFlightDelays("/home/scarface/Desktop/sem-3/capstone/Prithvi/Feb2Lab/2008.csv").map(rec => re
data_2007tmp.toDF().registerTempTable("data_2007tmp")

data_2007.take(5).map(x => x mkString ",").foreach(println)
```

```
data_2007tmp: org.apache.spark.rdd.RDD[DelayRec] = MapPartitionsRDD[6] at filter at <console>:57
data_2007: org.apache.spark.rdd.RDD[Array[Double]] = MapPartitionsRDD[7] at map at <console>:52
data_2008: org.apache.spark.rdd.RDD[Array[Double]] = MapPartitionsRDD[15] at map at <console>:50
warning: there was one deprecation warning; re-run with -deprecation for details
-8.0,1.0,25.0,4.0,11.0,719.0,10.0
41.0,1.0,28.0,7.0,15.0,925.0,13.0
45.0,1.0,29.0,1.0,20.0,316.0,14.0
-9.0,1.0,17.0,3.0,19.0,719.0,2.0
180.0,1.0,12.0,5.0,17.0,316.0,3.0
```

Took 25 sec. Last updated by anonymous at February 02 2017, 8:43:29 PM.

FINISHED

```
%sql

select dayofWeek, case when depDelay > 15 then 'delayed' else 'ok' end , count(1)
from data_2007tmp
group by dayofWeek , case when depDelay > 15 then 'delayed' else 'ok' end
```

| dayofWeek | CASE WHEN (CAST(depDelay AS DOUBLE) > CAST(15 AS DOUBLE)) THEN delaye |
|---|---|
| 1 | delayed |
| 7 | ok |
| 1 | ok |
| 6 | delayed |
| 2 | delayed |
| 3 | ok |
| 4 | delayed |
| 3 | delayed |
| 5 | ok |

Took 42 sec. Last updated by anonymous at February 02 2017, 8:45:26 PM.

FINISHED

```
%sql

select cast( cast(crsDepTime as int) / 100 as int) as hour,  case when depDelay > 15 then 'delayed' else 'ok' er
from  data_2007tmp
group by  cast( cast(crsDepTime as int) / 100 as int),  case when depDelay > 15 then 'delayed' else 'ok' end
```

| hour | delay |
|---|---|
| 12 | ok |
| 13 | ok |

| | |
|---|---|
| 20 | delayed |
| 10 | ok |
| 19 | ok |
| 15 | ok |
| 15 | delayed |

Took 39 sec. Last updated by anonymous at February 02 2017, 8:47:19 PM.

READY

Feb2_Lab Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

**Zeppelin**

Feb2_Lab

default