Lab1 |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled

Zeppelin

Lab1

default

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 4 sec. Last updated by anonymous at February 03 2017, 1:56:49 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",
        "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)
```

Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

Zeppelin

Lab1

default

```
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").parseDateT
    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 03 2017, 1:57:02 PM.

ERROR

```
%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
```

```
Table or view not found: data_2007tmp; line 2 pos 5
set zeppelin.spark.sql.stacktrace = true to see full stacktrace
```

Took 13 sec. Last updated by anonymous at February 03 2017, 1:32:03 AM.

FINISHED

```
%sh
wget http://stat-computing.org/dataexpo/2009/2008.csv.bz2 -O /Users/jyothi/Downloads/flights_
```

```
--2017-02-03 14:04:36--  http://stat-computing.org/dataexpo/2009/2008.csv.bz2
Resolving stat-computing.org... 52.218.160.39
Connecting to stat-computing.org|52.218.160.39|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 113753229 (108M) [application/x-bzip2]
Saving to: ‘/Users/jyothi/Downloads/flights_2008.csv.bz2’
     0K .......... .......... .......... .......... ..........  0%  322K 5m45s
    50K .......... .......... .......... .......... ..........  0%  625K 4m21s
   100K .......... .......... .......... .......... ..........  0%  627K 3m53s
   150K .......... .......... .......... .......... ..........  0%  678K 3m36s
   200K .......... .......... .......... .......... ..........  0% 5.55M 2m56s
   250K .......... .......... .......... .......... ..........  0%  818K 2m49s
   300K .......... .......... .......... .......... ..........  0% 2.33M 2m32s
   350K .......... .......... .......... .......... ..........  0% 9.04M 2m14s
   400K .......... .......... .......... .......... ..........  0% 12.2M 2m0s
   450K .......... .......... .......... .......... ..........  0%  885K 2m1s
   500K .......... .......... .......... .......... ..........  0% 4.34M 1m52s
```

Lab1 Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

1 sec. Last updated by anonymous at February 03 2017, 2:04:48 PM.

Zeppelin

Lab1

ERROR default

```
%sh
#remove existing copies of dataset from HDFS
hadoop fs -rm -r -f /tmp/airflightsdelays
```

```
17/02/03 13:48:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
rm: Call From Jyothi.local/172.17.204.216 to localhost:9000 failed on connection exception: ja
va.net.ConnectException: Connection refused; For more details see:  http://wiki.apache.org/had
oop/ConnectionRefused
17/02/03 13:48:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
rm: Call From Jyothi.local/172.17.204.216 to localhost:9000 failed on connection exception: ja
va.net.ConnectException: Connection refused; For more details see:  http://wiki.apache.org/had
oop/ConnectionRefused
ExitValue: 1
```

Took 8 sec. Last updated by anonymous at February 03 2017, 1:48:37 PM.

FINISHED

```
// function to do a preprocessing step for a given file
def prepFlightDelays(infile: String): RDD[DelayRec] = {
    val data = sc.textFile(infile)

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

val data_2007tmp = prepFlightDelays("/Users/jyothi/Downloads/flights_2007.csv")
val data_2007 = data_2007tmp.map(rec => rec.gen_features._2)
val data_2008 = prepFlightDelays("/Users/jyothi/Downloads/airflightsdelays/flights_2008.csv")

data_2007tmp.toDF().registerTempTable("data_2007tmp")

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

```
prepFlightDelays: (infile: String)org.apache.spark.rdd.RDD[DelayRec]
data_2007tmp: org.apache.spark.rdd.RDD[DelayRec] = MapPartitionsRDD[64] at filter at <console
>:58
data_2007: org.apache.spark.rdd.RDD[Array[Double]] = MapPartitionsRDD[65] at map at <console>:
52
data_2008: org.apache.spark.rdd.RDD[Array[Double]] = MapPartitionsRDD[73] 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 5 sec. Last updated by anonymous at February 03 2017, 2:07:09 PM.

Lab1 |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled

Zeppelin

Lab1 default

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 delayed EL
1	delayed
7	ok
1	ok
6	delayed
2	delayed
3	ok
4	delayed
3	delayed

Took 26 sec. Last updated by anonymous at February 03 2017, 2:07:43 PM.

FINISHED

```
%sql

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

hour	delay	cc
12	ok	13
13	ok	20
20	delayed	10
10	ok	17
19	ok	12
15	ok	14
15	delayed	7,
21	ok	8,

Took 27 sec. Last updated by anonymous at February 03 2017, 2:09:43 PM.

Zeppelin

READY

Lab1

default