Zeppelin

FINISHED

```
%sh

wget http://stat-computing.org/dataexpo/2009/2007.csv.bz2 -O /tmp/flights_2007.csv.bz2
wget http://stat-computing.org/dataexpo/2009/2008.csv.bz2 -O /tmp/flights_2008.csv.bz2
wget ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/2007.csv.gz -O /tmp/weather_2007.c
wget ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/2008.csv.gz -O /tmp/weather_2008.c
echo "downloaded"
```

```
--2017-02-02 19:46:31--  http://stat-computing.org/dataexpo/2009/2007.csv.bz2
Resolving stat-computing.org... 54.231.177.47
Connecting to stat-computing.org|54.231.177.47|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 121249243 (116M) [application/x-bzip2]
Saving to: ‘/tmp/flights_2007.csv.bz2’
     0K .......... .......... .......... .......... ..........  0%  321K 6m8s
    50K .......... .......... .......... .......... ..........  0%  625K 4m39s
   100K .......... .......... .......... .......... ..........  0%  638K 4m8s
   150K .......... .......... .......... .......... ..........  0% 9.06M 3m9s
   200K .......... .......... .......... .......... ..........  0%  685K 3m5s
   250K .......... .......... .......... .......... ..........  0% 4.97M 2m38s
   300K .......... .......... .......... .......... ..........  0%  752K 2m38s
   350K .......... .......... .......... .......... ..........  0% 8.34M 2m20s
   400K .......... .......... .......... .......... ..........  0% 4.93M 2m7s
   450K .......... .......... .......... .......... ..........  0% 9.48M 1m55s
   500K .......... .......... .......... .......... ..........  0%  889K 1m57s
   550K                                                        0% 5.73M 1m49s
```

FINISHED

```
 1 %sh
 2
 3 #remove existing copies of dataset from HDFS
 4 hadoop fs -rm -r -f /tmp/airflightsdelays
 5 hadoop fs -mkdir /tmp/airflightsdelays
 6
 7 #put data into HDFS
 8 hadoop fs -put /tmp/flights_200*.bz2 /tmp/airflightsdelays/
 9 hadoop fs -put /tmp/weather_200*.gz /tmp/airflightsdelays/
10 hadoop fs -ls -h /tmp/airflightsdelays/
```

```
17/02/02 19:47:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
17/02/02 19:47:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
17/02/02 19:47:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
17/02/02 19:47:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
put: `/tmp/weather_200*.gz': No such file or directory
17/02/02 19:47:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 Rhon wheel    115.6 M 2017-02-02 19:47 /tmp/airflightsdelays/flights_2007.cs
v.bz2
-rw-r--r--   1 Rhon wheel    108.5 M 2017-02-02 19:47 /tmp/airflightsdelays/flights_2008.cs
v.bz2
```

FINISHED

```
%dep

z.reset()
z.load("joda-time:joda-time:2.9.1")
```

```
DepInterpreter(%dep) deprecated. Remove dependencies and repositories through GUI interpret
er menu instead.
DepInterpreter(%dep) deprecated. Load dependency through GUI interpreter menu instead.
res0: org.apache.zeppelin.dep.Dependency = org.apache.zeppelin.dep.Dependency@143e9a9f
```

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
```

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", "06/0
      "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", "07/04/20
      "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,

    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").pa
        val distance = Math.abs(org.joda.time.Days.daysBetween(holiday, sampleDate).g
        math.min(r, distance)
      }
    }
  }
```

```
defined class DelayRec
```

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
```

```
null
set zeppelin.spark.sql.stacktrace = true to see full stacktrace
```

READY

```
%sql

select cast( cast(crsDepTime as int) / 100 as int) as hour,  case when depDelay > 15 then
delay, count(1) as count
from  data_2007tmp
```