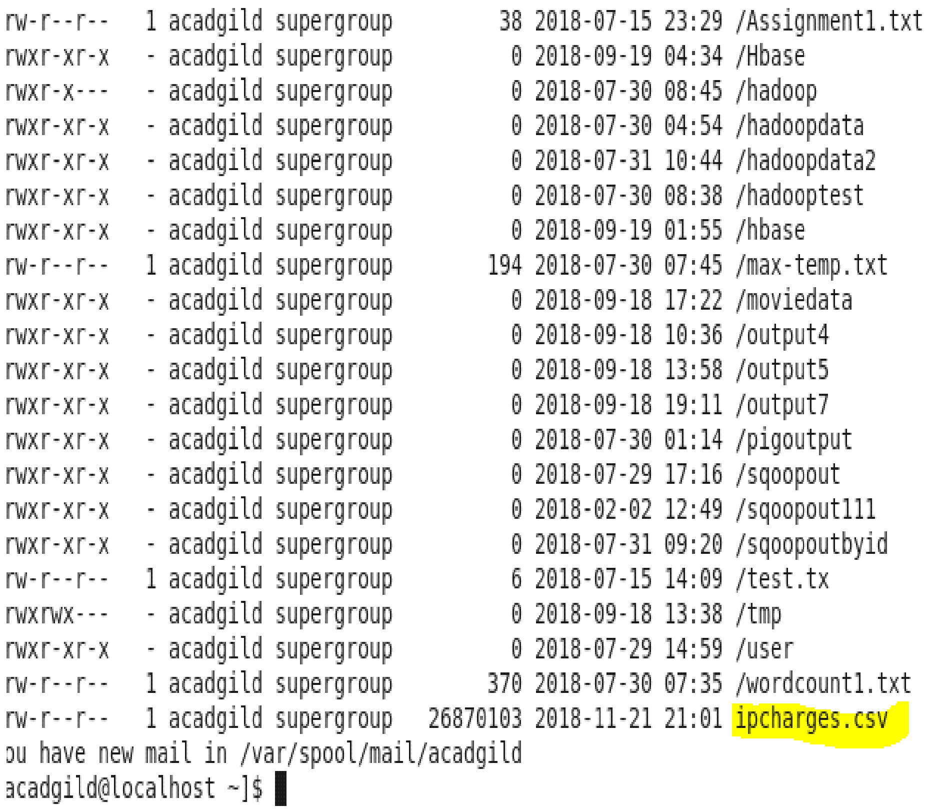
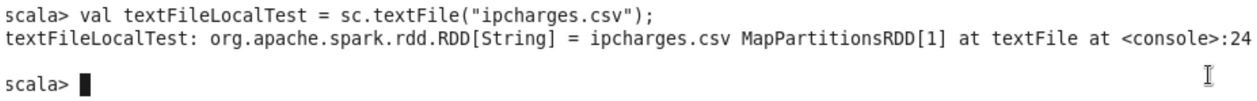
Case study 5

Moving the data set to HDFS



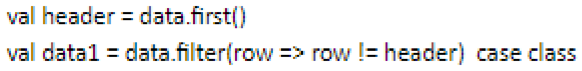
```
rw-r--r--   1 acadgild supergroup        38 2018-07-15 23:29 /Assignment1.txt
rwxr-xr-x   - acadgild supergroup         0 2018-09-19 04:34 /Hbase
rwxr-x---   - acadgild supergroup         0 2018-07-30 08:45 /hadoop
rwxr-xr-x   - acadgild supergroup         0 2018-07-30 04:54 /hadoopdata
rwxr-xr-x   - acadgild supergroup         0 2018-07-31 10:44 /hadoopdata2
rwxr-xr-x   - acadgild supergroup         0 2018-07-30 08:38 /hadooptest
rwxr-xr-x   - acadgild supergroup         0 2018-09-19 01:55 /hbase
rw-r--r--   1 acadgild supergroup       194 2018-07-30 07:45 /max-temp.txt
rwxr-xr-x   - acadgild supergroup         0 2018-09-18 17:22 /moviedata
rwxr-xr-x   - acadgild supergroup         0 2018-09-18 10:36 /output4
rwxr-xr-x   - acadgild supergroup         0 2018-09-18 13:58 /output5
rwxr-xr-x   - acadgild supergroup         0 2018-09-18 19:11 /output7
rwxr-xr-x   - acadgild supergroup         0 2018-07-30 01:14 /pigoutput
rwxr-xr-x   - acadgild supergroup         0 2018-07-29 17:16 /sqoopout
rwxr-xr-x   - acadgild supergroup         0 2018-02-02 12:49 /sqoopout111
rwxr-xr-x   - acadgild supergroup         0 2018-07-31 09:20 /sqoopoutbyid
rw-r--r--   1 acadgild supergroup         6 2018-07-15 14:09 /test.tx
rwxrwx---   - acadgild supergroup         0 2018-09-18 13:38 /tmp
rwxr-xr-x   - acadgild supergroup         0 2018-07-29 14:59 /user
rw-r--r--   1 acadgild supergroup       370 2018-07-30 07:35 /wordcount1.txt
rw-r--r--   1 acadgild supergroup  26870103 2018-11-21 21:01 ipcharges.csv
ou have new mail in /var/spool/mail/acadgild
acadgild@localhost ~]$
```

Loading the dataset on to spark



```
scala> val textFileLocalTest = sc.textFile("ipcharges.csv");
textFileLocalTest: org.apache.spark.rdd.RDD[String] = ipcharges.csv MapPartitionsRDD[1] at textFile at <console>:24

scala>
```



```
val header = data.first()
val data1 = data.filter(row => row != header)  case class
```

```
Charges(DRG:int,pid:int,pname:string,padd:string,pcity:string,pstate:string,pcode:int,referal:string,discharges:int,avgcovered:int,avgtotal:int,avgcare:int);


val Charges = sc.parallelize(1 to 10)


val avgperstate = sc.parallelize(Array(("pstate"), ("avgtotal")));
val join = avgperstate. join (avgtotal).


join.collect();
```