Case Study 4 – Working with Sensor Data

The dataset for the case study is the following csv file.

1. inpatientCharges.csv

This csv file is loaded onto HDFS as follows.

hadoop fs -put inpatientCharges.csv '/user/acadgild/hadoop/Hospital/'

```
[acadgild@localhost Hospital]$ hadoop fs -put inpatientCharges.csv '/user/acadgild/hadoop/Hospital/'
18/07/04 23:29:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Hospital]$ hadoop fs -ls '/user/acadgild/hadoop/Hospital'
18/07/04 23:42:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 1 items
-rw-r--r-- 1 acadgild supergroup 26870105 2018-07-04 23:29
You have new mail in /var/spool/mail/acadgild
```

On listing the contents of the Hospital folder on HDFS, we can see that the csv file is present.

Objective 1:

Load file into spark

```
scala> val session = org.apache.spark.sql.SparkSession.builder.master("local").appName("load CSV").getOrCreate;
18/07/סט סט:נס:נס: אארא sqt.sparkSession;souitder: osing an existing sparkSession; some configuration may not take
session: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@2b625e82
scala> val hospitaldf = session.read.format("com.databricks.spark.csv").option("header","true").option("inferSchema","true")
load("/user/acadgild/hadoop/Hospital/inpatientCharges.csv")
hospitaldf: org.apache.spark.sql.DataFrame = [DRGDefinition: string, ProviderId: int ... 10 more fields]
scala> hospitaldf.show
                                                                                                                                                              Provider Name | Provider Street Address | Provider City | Provider State | Provider Zip Code | Hospit Research Control of the Control of th
                              DRGDefinition|ProviderId|
alReferralRegionDescription|TotalDischarges|AverageCoveredCharges|AverageTotalPayments|ÁverageMedicarePayments|
                                                                                                                                                                                                                                                                                                                                                                                                     AL|
4763.73|
                  - EXTRACRANIA...| 1
AL - Dothan|
                                                                                                         10001|SOUTHEAST ALABAMA...| 1108 ROSS CLARK C...|
                                                                                                                                                                                                                                                                                                                  5777.24|
                                                                                                                                                                                                                              32963.07|
                                                                                                                                                                                                                                                                                                                                                                                                     AL|
4976.71|
                                                                                                         10005|MARSHALL MEDICAL ...| 2505 U S HIGHWAY ...|
 |039 - EXTRACRANIA...|
                                                                                                                                                                                                                                                                                                                                     BOAZ|
                                                                                                                                                                                                                                                                                                                                                                                                                                                         359571
 AL - Birmingham| 14| 15131.85|
|039 - EXTRACRANIA...| 10006|ELIZA COFFEE MEMO...| 205 MARENGO STREET|
                                                                                                                                                                                                                                                                                                                  5787.57
                                                                                                                                                                                                                                                                                                                        FLORENCE|
                                                                                                                                                                                                                                                                                                                                                                                                                                                         35631|
                                                        - Birmingham| 24| 37560.37| 5434.95|
ANIA...| 10011| ST VINCENT'S EAST| 50 MEDICAL PARK E...| BIRMINGHAM|
                                                                                                                                                                                                                                                                                                                                                                                                                  4453.79|
                                                                                                                                                                                                                                                                                                                                                                                                    AL|
4129.16|
                  - EXTRACRANIA...|
                                               AL - Birmingham|
                                                                                                                                                                                                                              13998.281
                                                                                                                                                                                                                                                                                                                     5417.56
                                                                                                                                                                                                                                                                                                                                                                                                    AL|
4851.44|
                                                                                                         10016|SHELBY BAPTIST ME...| 1000 FIRST STREET...|
 |039 - EXTRACRANIA.
                                                                                                                                                                                                                                                                                                                   ALABASTER|
                                                                                                                                                                                                                                                                                                                                                                                                                                                         35007|
                                                                 AL - B1rmingnam;

|039 - EXTRACRANIA...| 10023|BAPTIST MEDICAL ..., 2-16920.79|

AL - Montgomery| 67| 16920.79|

10029|EAST ALABAMA MEDI...| 2000 PEPPERELL PA...|

51| 11977.13|

520 SOUTH 19TH ST...|
                                                                                                                                                                                                                                                                                                                                                                                                    AL|
5374.14|
                                                                                                                                                                                                                                                                                                                                                                                                                                                         361161
                                                                                                                                                                                                                                                                                                                      6653.8|
OPELIKA|
                                                                                                                                                                                                                                                                                                                                                                                                    AL|
4761.41|
| 039 - EATION | AL - BIFMINGHOM, | 10033|UNIVENDED | AL - BIFMINGHOM, | 32| | 32| | 32| | 32| | 32| | 33| | 34| | 35| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 75233.38| | 34| | 752333.38| | 34| | 752333.38| | 34| | 752333.38| | 34| | 7523333.38| | 34| | 75233333| | 34| | 752333
                                                                                                                                                                                                                                                                                                                     5834.74
                                                                                                        | 11977.15| 5034.74| 10033|UNIVERSITY OF ALA...| 619 SOUTH 19TH ST...| BIRMINGHAM| | 32| 35841.09| 8031.12| 10039| HUNTSVILLE HOSPITAL| 101 SIVLEY RD| HUNTSVILLE|
                                                                                                                                                                                                                                                                                                                                                                                                      AL|
                                                                                                                                                                                                                                                                                                                                                                                                                                                         35233 I
                                                                                                                                                                                                                                                                                                                                                                                                                    ₫858.5|
                                                                                                                                                                                                                                                                                                                                                                                                                                                         35801 l
                                                                                                                                                                                                                                                                                                                                                                                                      AL|
                                                                                                                                                                                                                                                                                                                   6113.38|
GADSDEN|
                                                                                                                                                                                                                                                                                                                                                                                                                     5228.4|
                                                                                                                                                                                                                                                                                                                                                                                                      AL|
                                                                                                                                                                                                                                                                                                                                                                                                                   4386.94|
                                                                                                                                                                                                                                                                                                                    5541.05|
                                                                                                                                                                                                                                                                                                                                                                                                    AL|
4493.57|
                                                                                                                                                                                                                                                                                                                            GADSDEN|
                                                                                                                                                                                                                                                                                                                                                                                                                                                         35901|
AL - BIIIMING.... | 1
|039 - EXTRACRANIA...| 1
|AL - Dothan
                                                                                                                                                                                                                                                                                                                   5461.57
                                                                                                                                               FLOWERS HOSPITAL | 4370 WEST MAIN ST...
                                                                                                         10055|
                                                                                                                                                                                                                                                                                                                               DOTHAN|
                                                                                                                                                                                                                                                                                                                                                                                                      AL|
                                                                                                                                                                                                                                                                                                                                                                                                                                                         36305 I
```

We have loaded all the CSV data as a DataFrame into Spark SQL. Here, we have used inferschema as an option so it will automatically infer the data type of the columns.

```
val session =
org.apache.spark.sql.SparkSession.builder.master("local").appName("Spark CSV
Reader").getOrCreate;
val hospitaldf = session.read.format("com.databricks.spark.csv").option("header",
"true").option("inferSchema",
```

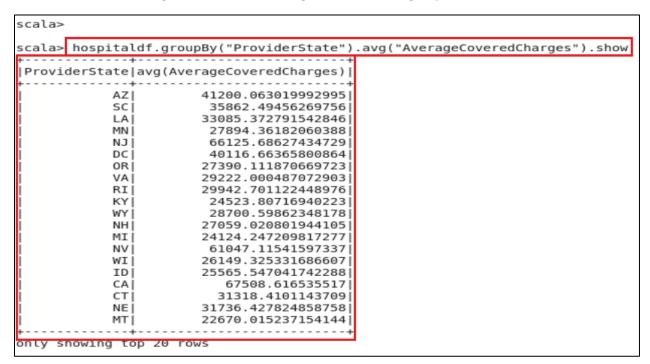
We can see the contents of the dataframe in the screenshot above using the following command.

"true").load("/user/acadgild/hadoop/Hospital/inpatientCharges.csv")

hospitaldf.show

Objective 2:

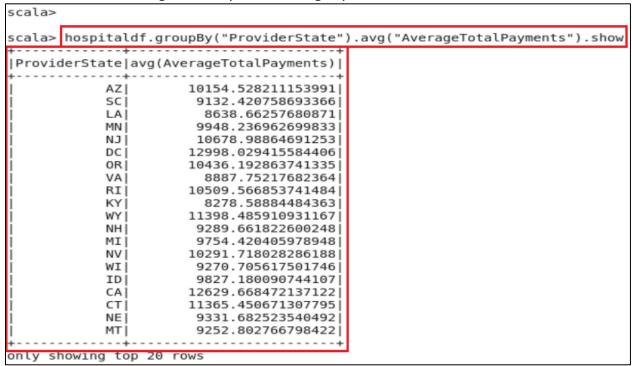
- A. What is the average amount of AverageCoveredCharges per state
- B. Find out the AverageTotalPayments charges per state
- C. Find out the AverageMedicarePayments charges per state.
- A. What is the average amount of AverageCoveredCharges per state



hospital df. group By ("Provider State"). avg ("Average Covered Charges"). show

groupBy and avg functions are used on the dataframe to calculate the average per state. The average amount of AverageCoveredCharges per state is shown as output in the screenshot above.

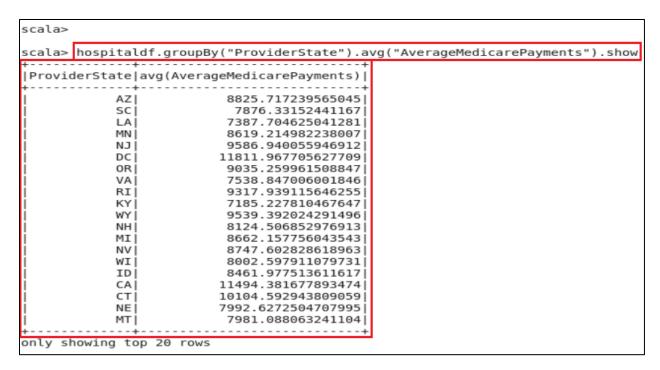
B. Find out the AverageTotalPayments charges per state



hospital df. group By ("Provider State"). avg ("Average Total Payments"). show

groupBy and avg functions are used on the dataframe to calculate the average per state. The average amount of AverageTotalPayments per state is shown as output in the screenshot above.

C. Find out the AverageMedicarePayments charges per state.



hospitaldf.groupBy("ProviderState").avg("AverageMedicarePayments").show

groupBy and avg functions are used on the dataframe to calculate the average per state. The average amount of AverageMedicarePayments per state is shown as output in the screenshot above.

Objective 3:

- A. Find out the total number of Discharges per state and for each disease
- B. Sort the output in descending order of totalDischarges
- A. Find out the total number of Discharges per state and for each disease

```
scala> hospitaldf.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").show
18/07/05 00:35:13 WARN executor.Executor: managed memory
                                                                                                          TID = 511
ProviderState|
                        DRGDefinition|sum(TotalDischarges)|
             KY1065 -
                      INTRACRANTA...
             NY | 101
                       DYSEQUILIBRIUM
             IN|149
                                                           700
                                                           540
             WI | 202
                       BRONCHITIS
                                                           338
                       PERC CARDIO.
ACUTE MYOCA.
                                                           417
413
             WI | 251
             AZ | 292
                       HEART FAILU.
                                                          2643
                                                         13289
             NV 1293
                       HEART FAILU.
                                                           519
                303
                                                           730
             TN | 305
                       HYPERTENSIO.
                308
                       CARDIAC ARR
             NV | 372
                       MAJOR GASTR.
                                                           126
                392
                       ESOPHAGITIS
                                                          3148
             WI | 439
                       DISORDERS O.
                                                           215
            MN|536
DC|563
                       FRACTURES 0
                       FX, SPRN,
                                                            43
             CO | 602
                       CELLULITIS
only showing top 20 rows
```

hospital df. group By (("Provider State"), ("DRGD efinition")). sum ("Total Discharges"). sum

Groupby is applied on a combination of state and DRG definition and then the sum is calculated for each group. The sum of totalDischarges per state and for each disease is shown in the screenshot above.

B. Sort the output in descending order of totalDischarges

```
cala> hospitaldf.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").orderBy(desc(sum("TotalDischarges").toS
              CA|871 -
                         SEPTICEMIA
                         MAJOR JOINT..
MAJOR JOINT..
                                                                29985
              CA 1470
                         MAJOR JOINT.
                                                                29731
                         SEPTICEMIA
SEPTICEMIA
                                                                23144
                                                                21970
              FL | 392
                         ESOPHAGITIS
                                                                21298
              IL|470
NY|470
                         MAJOR JOINT..
                                                                20095
              FL|871
                         SEPTICEMIA
                                                                18660
              TX|690
NY|392
MI|470
                         KIDNEY & UR
ESOPHAGITIS
                                                                17384
                                                                16847
                         MAJOR JOINT.
              PA 1470
                         MAJOR JOINT
                                                                16712
                         HEART FAILU..
KIDNEY & UR..
MAJOR JOINT..
                                                                16639
16405
              OH | 470
                                                                16062
                                                                15820
15610
              NC | 470
                         MAJOR JOINT
              MI|871
                         SEPTICEMIA
                                                                15548
```

hospitaldf.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").s how .orderBy(desc(sum("TotalDischarges").toString)).show

On the above query, orderBy(desc) is applied on the TotalDischarges column to sort the results in a descending order.