2018

# BIG DATA HADOOP & SPARK TRAINING

CASE STUDY IV : Case study on Hospital Data Analysis in the United States

- RASHMI KRISHNA

# CASE STUDY IV

## Case study on: Hospital Data Analysis in the United States

**Dataset Description**

*DRG Definition*: The code and description identifying the MS-DRG. MS-DRGs are a classification system that groups similar clinical conditions (diagnoses) and procedures furnished by the hospital during their stay.

*Provider Id*: The CMS Certification Number (CCN) assigned to the Medicare-certified hospital facility.

*Provider Name*: The name of the provider.

*Provider Street Address*: The provider's street address.

*Provider City*: The city where the provider is located.

*Provider State*: The state where the provider is located.

*Provider Zip Code*: The provider's zip code.

*Provider HRR*: The Hospital Referral Region (HRR) where the provider is located.

*Total Discharges*: The number of discharges billed by the provider for inpatient hospital services.

*Average Covered Charges*: The provider's average charge for services covered by Medicare for all discharges in the MS-DRG. These will vary from hospital to hospital because of the differences in hospital charge structures.

*Average Total Payments*: The average total payments to all providers for the MS-DRG including the MSDRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Also included in the average total payments are co-payment and deductible amounts that the patient is responsible for and any additional payments by third parties for coordination of benefits.

*Average Medicare Payments*: The average amount that Medicare pays to the provider for Medicare's share of the MS-DRG. Average Medicare payment amounts include the MS-DRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Medicare payments DO NOT include beneficiary co-payments and deductible amounts nor any additional payments from third parties for coordination of benefits.

You can download the dataset used in this spark SQL use case from below link.4

https://drive.google.com/open?id=13_YDmwENxOQI5asLRa6tOP8FgiqqM9jc

- Let's load the above input file into spark
  - Create a manual schema for the csv file which would provide the schema while loading the CSV file as shown below

```
val Manual_schema = new StructType(Array( new StructField("DRGDefinition",StringType,
true),
  new StructField("ProviderId", LongType, false),
  new StructField("ProviderName", StringType, true),
  new StructField("ProviderStreetAddress", StringType, false),
  new StructField("ProviderCity", StringType, false),
  new StructField("ProviderState", StringType, false),
  new StructField("ProviderZipCode", LongType, false),
  new StructField("HospitalReferralRegionDescription", StringType,true),
  new StructField("TotalDischarges", LongType, false),
  new StructField("AverageCoveredCharges", DoubleType, false),
  new StructField("AverageTotalPayments", DoubleType,false),
  new StructField("AverageMedicarePayments", DoubleType,false)))
```

Note:

StructType is a built-in data type used for Schema definition in Spark SQL, to represent a collection of StructFields that together define a schema or its part.

<schema-name> = new StructType<array_of_columns><Struct_field>(<column_name>,<data_type_of_column>,<nullable_or_not_nullable(true/false)>)

- Now we load the CSV files from local file system to spark as shown below:

```
val Hospital_data = spark.read.format("csv")
  .option("header", "true")
  .schema(Manual_schema)
  .load("E:\\casestudies\\hospitalcasestudy\\inpatient.csv").toDF()
  Hospital_data.show()
```

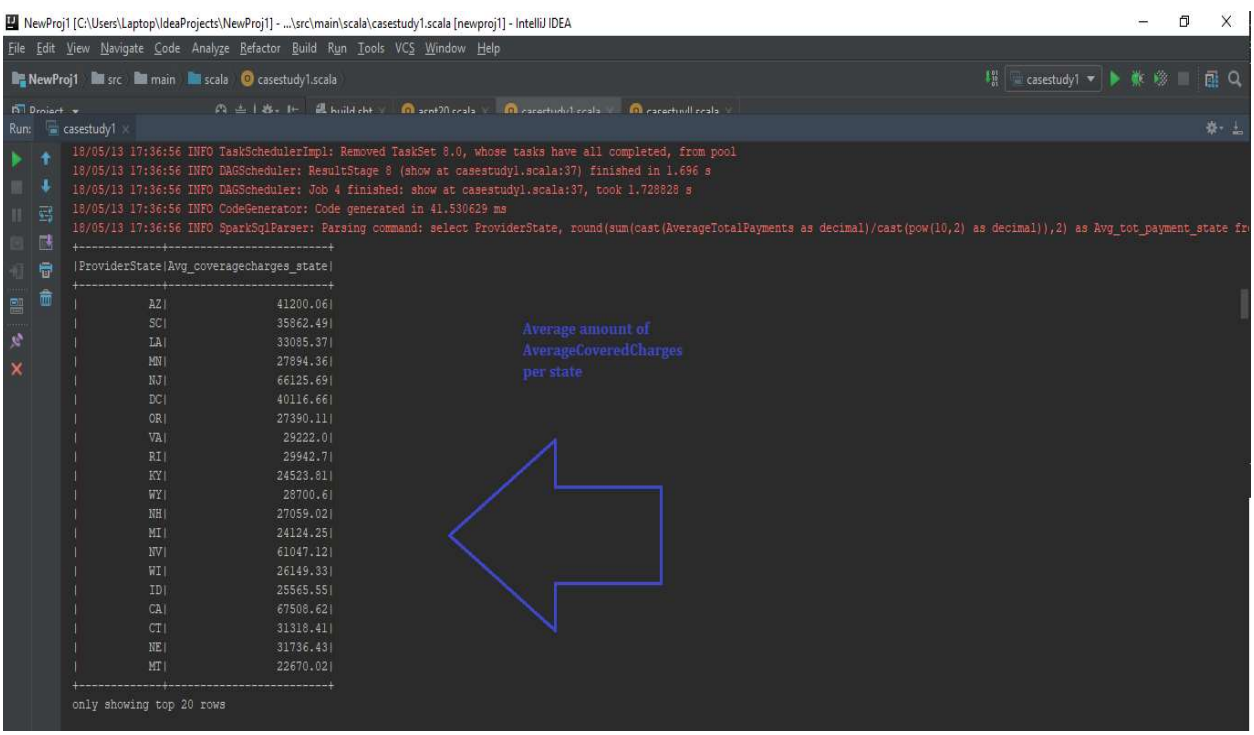
Contents of the input file that is loaded into spark is as shown below:

| DRGDefinition | ProviderId | ProviderName | ProviderStreet&Address | ProviderCity | ProviderState | ProviderZipCode | HospitalReferralRegionDescription | TotalDischarges | AverageCoveredCharges | AverageTotalPayments | AverageMedicarePayments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 039 - EXTRACRANIA... | 10001 | SOUTHEAST ALABAMA... | 1108 ROSS CLARK C... | DOTHAN | AL | 36301 | AL - Dothan | 91 | 32963.07 | 5777.24 | 4763.73 |
| 039 - EXTRACRANIA... | 10005 | MARSHALL MEDICAL ... | 2505 U S HIGHWAY ... | BOAZ | AL | 35957 | AL - Birmingham | 14 | 15131.85 | 5787.57 | 4976.71 |
| 039 - EXTRACRANIA... | 10006 | ELIZA COFFEE MEMO... | 205 MARENGO STREET | FLORENCE | AL | 35631 | AL - Birmingham | 24 | 37560.37 | 5434.95 | 4453.79 |
| 039 - EXTRACRANIA... | 10011 | ST VINCENT'S EAST | 50 MEDICAL PARK E... | BIRMINGHAM | AL | 35235 | AL - Birmingham | 25 | 13998.28 | 5417.56 | 4129.16 |
| 039 - EXTRACRANIA... | 10016 | SHELBY BAPTIST ME... | 1000 FIRST STREET... | ALABASTER | AL | 35007 | AL - Birmingham | 18 | 31633.27 | 5658.33 | 4851.44 |
| 039 - EXTRACRANIA... | 10023 | BAPTIST MEDICAL C... | 2105 EAST SOUTH B... | MONTGOMERY | AL | 36116 | AL - Montgomery | 67 | 16920.79 | 6653.8 | 5374.14 |
| 039 - EXTRACRANIA... | 10029 | EAST ALABAMA MEDI... | 2000 PEPPERELL PA... | OPELIKA | AL | 36801 | AL - Birmingham | 51 | 11977.13 | 5834.74 | 4761.41 |
| 039 - EXTRACRANIA... | 10033 | UNIVERSITY OF ALA... | 619 SOUTH 19TH ST... | BIRMINGHAM | AL | 35233 | AL - Birmingham | 32 | 35841.09 | 8031.12 | 5858.5 |
| 039 - EXTRACRANIA... | 10039 | HUNTSVILLE HOSPITAL | 101 SIVLEY RD | HUNTSVILLE | AL | 35801 | AL - Huntsville | 135 | 28523.39 | 6113.38 | 5228.4 |
| 039 - EXTRACRANIA... | 10040 | GADSDEN REGIONAL ... | 1007 GOODYEAR AVENUE | GADSDEN | AL | 35903 | AL - Birmingham | 34 | 75233.38 | 5541.05 | 4386.94 |
| 039 - EXTRACRANIA... | 10046 | RIVERVIEW REGIONA... | 600 SOUTH THIRD S... | GADSDEN | AL | 35901 | AL - Birmingham | 14 | 67327.92 | 5461.57 | 4493.57 |
| 039 - EXTRACRANIA... | 10055 | FLOWERS HOSPITAL | 4370 WEST MAIN ST... | DOTHAN | AL | 36305 | AL - Dothan | 45 | 39607.28 | 5356.28 | 4408.2 |
| 039 - EXTRACRANIA... | 10056 | ST VINCENT'S BIRM... | 810 ST VINCENT'S ... | BIRMINGHAM | AL | 35205 | AL - Birmingham | 43 | 22962.23 | 5374.65 | 4186.02 |
| 039 - EXTRACRANIA... | 10078 | NORTHEAST ALABAMA... | 400 EAST 10TH STREET | ANNISTON | AL | 36207 | AL - Birmingham | 21 | 31110.85 | 5366.23 | 4376.23 |
| 039 - EXTRACRANIA... | 10083 | SOUTH BALDWIN REG... | 1613 NORTH MCKENZ... | FOLEY | AL | 36535 | AL - Mobile | 15 | 25411.33 | 5282.93 | 4383.73 |
| 039 - EXTRACRANIA... | 10085 | DECATUR GENERAL H... | 1201 7TH STREET SE | DECATUR | AL | 35609 | AL - Huntsville | 27 | 9234.51 | 5676.55 | 4509.11 |
| 039 - EXTRACRANIA... | 10090 | PROVIDENCE HOSPITAL | 6801 AIRPORT BOUL... | MOBILE | AL | 36608 | AL - Mobile | 27 | 15895.85 | 5930.11 | 3972.85 |
| 039 - EXTRACRANIA... | 10092 | D C H REGIONAL ME... | 809 UNIVERSITY BO... | TUSCALOOSA | AL | 35401 | AL - Tuscaloosa | 31 | 19721.16 | 6192.54 | 5179.38 |
| 039 - EXTRACRANIA... | 10100 | THOMAS HOSPITAL | 750 MORPHY AVENUE | FAIRHOPE | AL | 36532 | AL - Mobile | 18 | 10710.88 | 4968.0 | 3898.88 |
| 039 - EXTRACRANIA... | 10103 | BAPTIST MEDICAL C... | 701 PRINCETON AVE... | BIRMINGHAM | AL | 35211 | AL - Birmingham | 33 | 51343.75 | 5996.0 | 4962.45 |

only showing top 20 rows

- What is the average amount of AverageCoveredCharges per state
  - To calculate average, first we shall create a temporary view called "hospital_view"
  - Write sql query on the view previously created to obtain the average amount of AverageCoveredCharges
    - In the query we are rounding the average values to 2 decimal points.

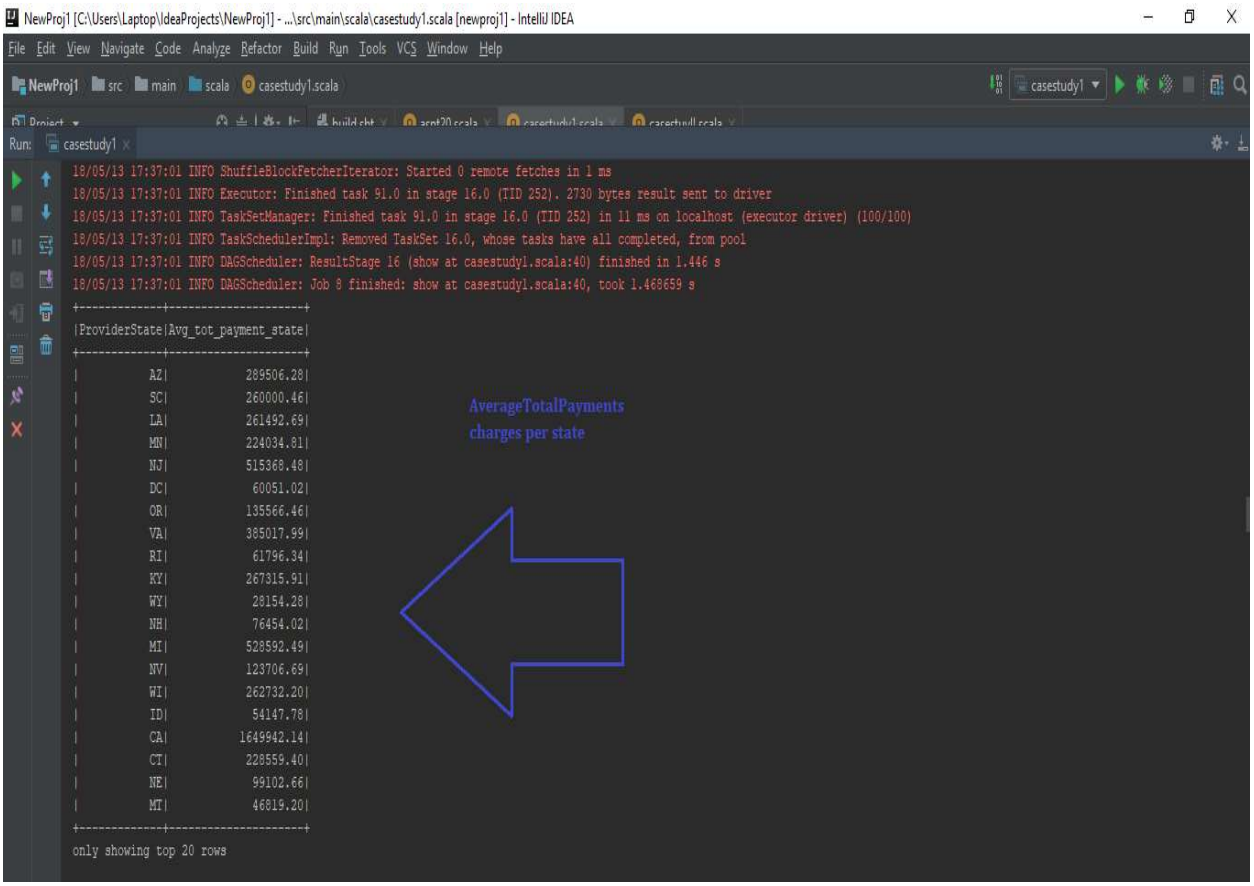
```
Hospital_data.createOrReplaceTempView("Hospital_view")
spark.sql("""select ProviderState, round(avg(AverageCoveredCharges),2) as
Avg_coveragecharges_state from Hospital_view group by ProviderState""").show()
```



- Find out the AverageTotalPayments charges per state

```
//the AverageTotalPayments charges per state
spark.sql("""select ProviderState, round(sum(cast(AverageTotalPayments as
decimal)/cast(pow(10,2) as decimal)),2) as Avg_tot_payment_state from Hospital_view
group by ProviderState""").show()
```

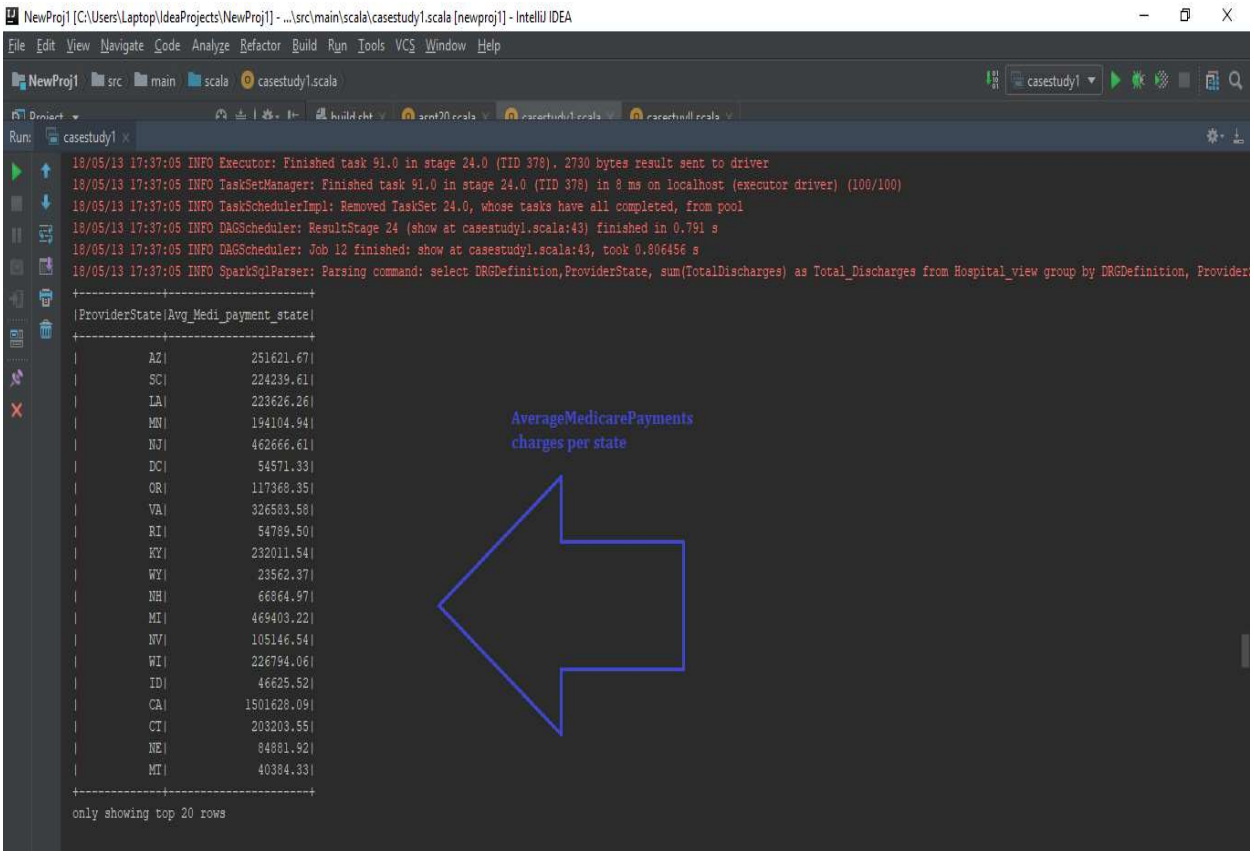
  - To calculate sum, first we shall create a temporary view called "hospital_view"
  - Write sql query on the view previously created to obtain the total amount of AverageTotalPayments per state
    - In the query we are rounding the average values to 2 decimal points and we are casting to decimal data type.



- Find out the AverageMedicarePayments charges per state.

```
//AverageMedicarePayments charges per state
spark.sql("""select ProviderState, round(sum(cast(AverageMedicarePayments as
decimal)/cast(pow(10,2) as decimal)),2) as Avg_Medi_payment_state from Hospital_view
group by ProviderState""").show()
```

  - To calculate sum, first we shall create a temporary view called "hospital_view"
  - Write sql query on the view previously created to obtain the total amount of AverageMedicarePayments per state
    - In the query we are rounding the average values to 2 decimal points and we are casting to decimal data type.



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

```
spark.sql("""select DRGDefinition,ProviderState, sum(TotalDischarges) as
Total_Discharges from Hospital_view group by DRGDefinition, ProviderState order by
Total_Discharges desc """).show()
```

  - To calculate sum, first we shall create a temporary view called "hospital_view"
  - Write sql query on the view previously created to obtain the total amount of TotalDischares per state and per disease.