Hospital Data

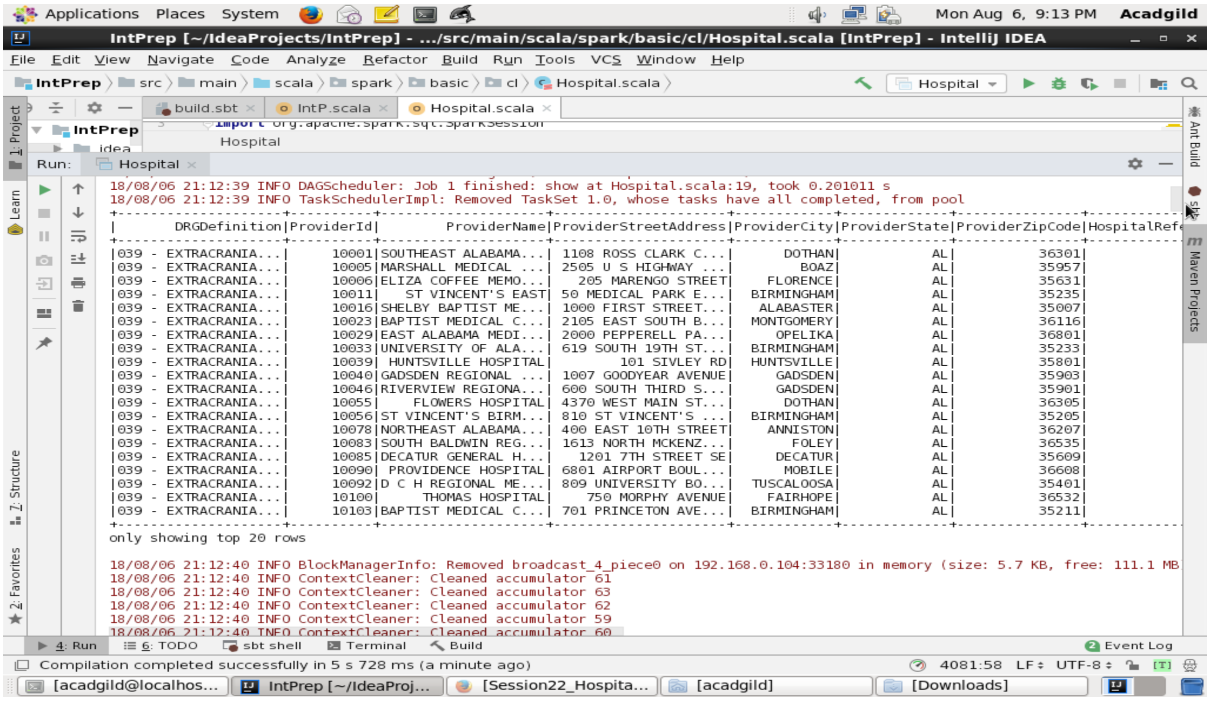
Code:

```
package spark.basic.cl

import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.types.IntegerType
import org.apache.spark.sql.functions._

object Hospital extends App{

  val sparkSession = SparkSession.builder.master("local")
    .appName("spark").getOrCreate()

  val sparkcontext = sparkSession.sparkContext

  //OBJECTIVE 1

  val hp = sparkSession.read.format("csv").option("header","true")
    .load("/home/acadgild/Downloads/inpatientCharges.csv")

  hp.show()

  import sparkSession.implicits._

  //Objective 2

  val avgCC = hp.groupBy($"ProviderState").agg(avg("AverageCoveredCharges")).show()

  val sumATP = hp.groupBy($"ProviderState").agg(sum("AverageTotalPayments")).show()

  val sumAMP =
hp.groupBy($"ProviderState").agg(sum("AverageMedicarePayments")).show()

  //OBJECTIVE 3

  val PSD = hp.groupBy($"ProviderState",$"DRGDefinition").agg(sum("TotalDischarges"))
  PSD.show()

  val DEOR = PSD.sort(desc("sum(TotalDischarges)")).show()


}
```

OBJECTIVE 1

Load file into spark

```
  val hp = sparkSession.read.format("csv").option("header","true")
    .load("/home/acadgild/Downloads/inpatientCharges.csv")

  hp.show()
```

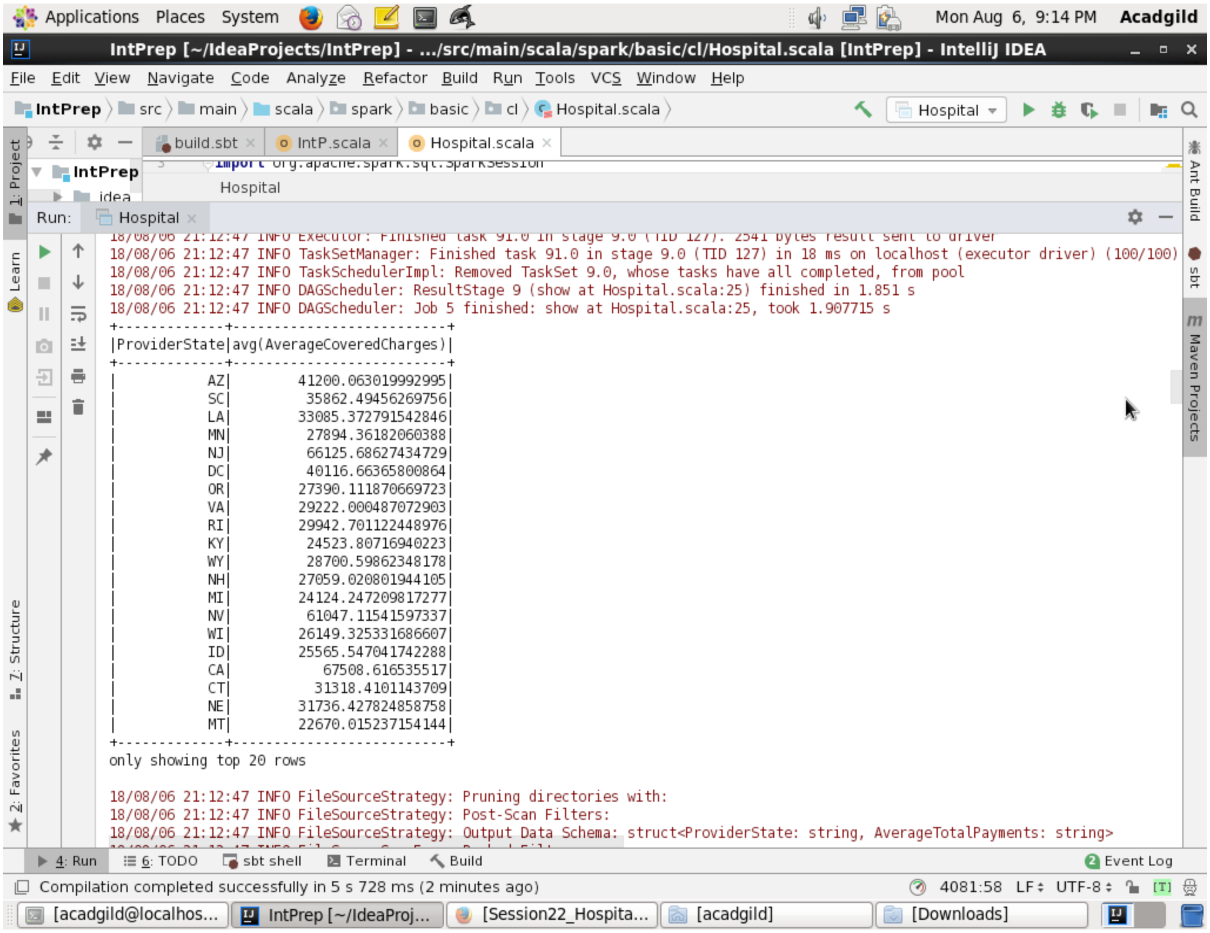

Applications Places System
Mon Aug 6, 9:13 PM Acadgild
IntPrep [~/IdeaProjects/IntPrep] - .../src/main/scala/spark/basic/cl/Hospital.scala [IntPrep] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IntPrep src main scala spark basic cl Hospital.scala
build.sbt IntP.scala Hospital.scala
Run: Hospital
4: Run 6: TODO sbt shell Terminal Build Event Log
Compilation completed successfully in 5 s 728 ms (a minute ago)
4081:58 LF UTF-8

```
18/08/06 21:12:39 INFO DAGScheduler: Job 1 finished: show at Hospital.scala:19, took 0.201011 s
18/08/06 21:12:39 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
+--------------------+----------+--------------------+---------------------+------------+-------------+---------------+------------
|       DRGDefinition|ProviderId|        ProviderName|ProviderStreetAddress|ProviderCity|ProviderState|ProviderZipCode|HospitalRefe
+--------------------+----------+--------------------+---------------------+------------+-------------+---------------+------------
|039 - EXTRACRANIA...|     10001|SOUTHEAST ALABAMA...| 1108 ROSS CLARK C...|      DOTHAN|           AL|          36301|
|039 - EXTRACRANIA...|     10005|MARSHALL MEDICAL ...| 2505 U S HIGHWAY ...|        BOAZ|           AL|          35957|
|039 - EXTRACRANIA...|     10006|ELIZA COFFEE MEMO...|   205 MARENGO STREET|    FLORENCE|           AL|          35631|
|039 - EXTRACRANIA...|     10011|   ST VINCENT'S EAST| 50 MEDICAL PARK E...|  BIRMINGHAM|           AL|          35235|
|039 - EXTRACRANIA...|     10016|SHELBY BAPTIST ME...| 1000 FIRST STREET...|   ALABASTER|           AL|          35007|
|039 - EXTRACRANIA...|     10023|BAPTIST MEDICAL C...| 2105 EAST SOUTH B...|  MONTGOMERY|           AL|          36116|
|039 - EXTRACRANIA...|     10029|EAST ALABAMA MEDI...| 2000 PEPPERELL PA...|     OPELIKA|           AL|          36801|
|039 - EXTRACRANIA...|     10033|UNIVERSITY OF ALA...| 619 SOUTH 19TH ST...|  BIRMINGHAM|           AL|          35233|
|039 - EXTRACRANIA...|     10039| HUNTSVILLE HOSPITAL|        101 SIVLEY RD|  HUNTSVILLE|           AL|          35801|
|039 - EXTRACRANIA...|     10040|GADSDEN REGIONAL ...| 1007 GOODYEAR AVENUE|     GADSDEN|           AL|          35903|
|039 - EXTRACRANIA...|     10046|RIVERVIEW REGIONA...| 600 SOUTH THIRD S...|     GADSDEN|           AL|          35901|
|039 - EXTRACRANIA...|     10055|    FLOWERS HOSPITAL| 4370 WEST MAIN ST...|      DOTHAN|           AL|          36305|
|039 - EXTRACRANIA...|     10056|ST VINCENT'S BIRM...| 810 ST VINCENT'S ...|  BIRMINGHAM|           AL|          35205|
|039 - EXTRACRANIA...|     10078|NORTHEAST ALABAMA...| 400 EAST 10TH STREET|    ANNISTON|           AL|          36207|
|039 - EXTRACRANIA...|     10083|SOUTH BALDWIN REG...| 1613 NORTH MCKENZ...|       FOLEY|           AL|          36535|
|039 - EXTRACRANIA...|     10085|DECATUR GENERAL H...|   1201 7TH STREET SE|     DECATUR|           AL|          35609|
|039 - EXTRACRANIA...|     10090| PROVIDENCE HOSPITAL| 6801 AIRPORT BOUL...|      MOBILE|           AL|          36608|
|039 - EXTRACRANIA...|     10092|D C H REGIONAL ME...| 809 UNIVERSITY BO...|  TUSCALOOSA|           AL|          35401|
|039 - EXTRACRANIA...|     10100|     THOMAS HOSPITAL|    750 MORPHY AVENUE|    FAIRHOPE|           AL|          36532|
|039 - EXTRACRANIA...|     10103|BAPTIST MEDICAL C...| 701 PRINCETON AVE...|  BIRMINGHAM|           AL|          35211|
+--------------------+----------+--------------------+---------------------+------------+-------------+---------------+------------
only showing top 20 rows

18/08/06 21:12:40 INFO BlockManagerInfo: Removed broadcast_4_piece0 on 192.168.0.104:33180 in memory (size: 5.7 KB, free: 111.1 MB
18/08/06 21:12:40 INFO ContextCleaner: Cleaned accumulator 61
18/08/06 21:12:40 INFO ContextCleaner: Cleaned accumulator 63
18/08/06 21:12:40 INFO ContextCleaner: Cleaned accumulator 62
18/08/06 21:12:40 INFO ContextCleaner: Cleaned accumulator 59
18/08/06 21:12:40 INFO ContextCleaner: Cleaned accumulator 60
```

OBJECTIVE 2

➢ What is the average amount of AverageCoveredCharges per state

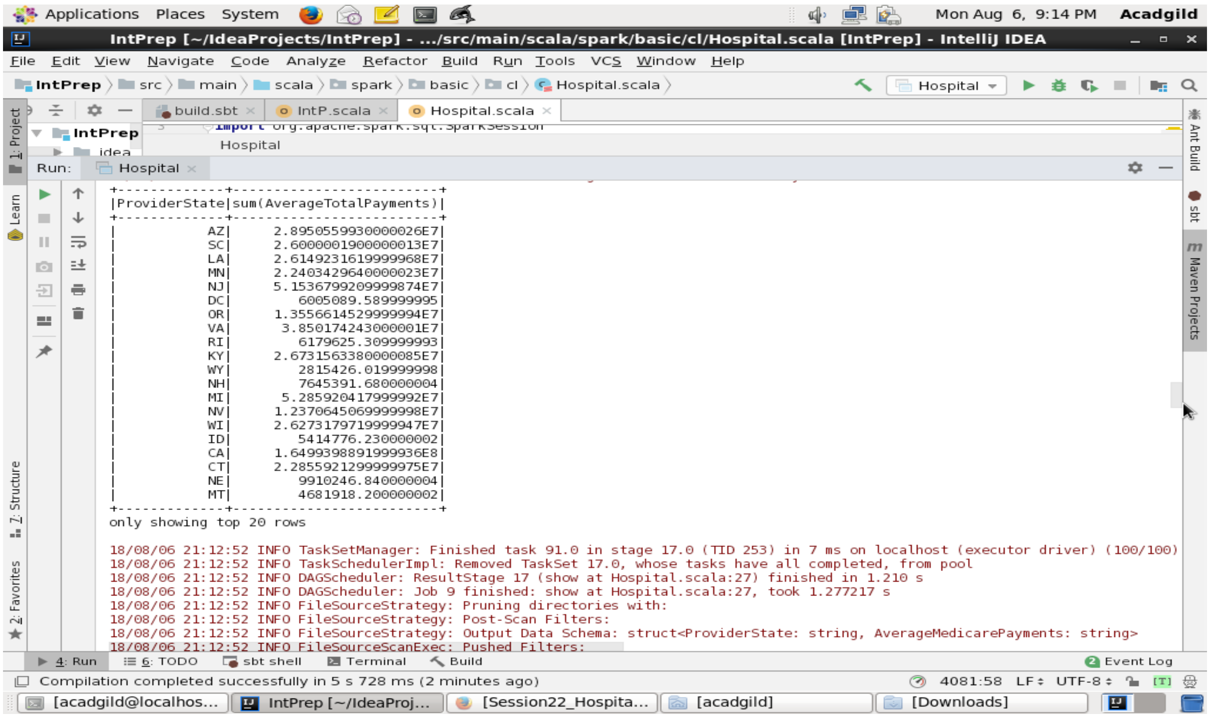
```
val avgCC = hp.groupBy($"ProviderState").agg(avg("AverageCoveredCharges")).show()
```


Applications Places System
Mon Aug 6, 9:14 PM Acadgild
IntPrep [~/IdeaProjects/IntPrep] - .../src/main/scala/spark/basic/cl/Hospital.scala [IntPrep] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IntPrep src main scala spark basic cl Hospital.scala
build.sbt IntP.scala Hospital.scala
Run: Hospital
4: Run 6: TODO sbt shell Terminal Build Event Log
Compilation completed successfully in 5 s 728 ms (2 minutes ago)
4081:58 LF UTF-8

```
18/08/06 21:12:47 INFO Executor: Finished task 91.0 in stage 9.0 (TID 127). 2541 bytes result sent to driver
18/08/06 21:12:47 INFO TaskSetManager: Finished task 91.0 in stage 9.0 (TID 127) in 18 ms on localhost (executor driver) (100/100)
18/08/06 21:12:47 INFO TaskSchedulerImpl: Removed TaskSet 9.0, whose tasks have all completed, from pool
18/08/06 21:12:47 INFO DAGScheduler: ResultStage 9 (show at Hospital.scala:25) finished in 1.851 s
18/08/06 21:12:47 INFO DAGScheduler: Job 5 finished: show at Hospital.scala:25, took 1.907715 s
+-------------+------------------------+
|ProviderState|avg(AverageCoveredCharges)|
+-------------+------------------------+
|           AZ|      41200.063019992995|
|           SC|        35862.4945629756|
|           LA|      33085.372791542846|
|           MN|       27894.36182060388|
|           NJ|       66125.68627434729|
|           DC|       40116.66365800864|
|           OR|      27390.111870669723|
|           VA|      29222.000487072903|
|           RI|      29942.701122448976|
|           KY|       24523.80716940223|
|           WY|       28700.59862348178|
|           NH|      27059.020801944105|
|           MI|      24124.247209817277|
|           NV|       61047.11541597337|
|           WI|      26149.325331686607|
|           ID|      25565.547041742288|
|           CA|         67508.616535517|
|           CT|        31318.4101143709|
|           NE|      31736.427824858758|
|           MT|      22670.015237154144|
+-------------+------------------------+
only showing top 20 rows

18/08/06 21:12:47 INFO FileSourceStrategy: Pruning directories with:
18/08/06 21:12:47 INFO FileSourceStrategy: Post-Scan Filters:
18/08/06 21:12:47 INFO FileSourceStrategy: Output Data Schema: struct<ProviderState: string, AverageTotalPayments: string>
```

➢ find out the AverageTotalPayments charges per state

```
val sumATP = hp.groupBy($"ProviderState").agg(sum("AverageTotalPayments")).show()
```



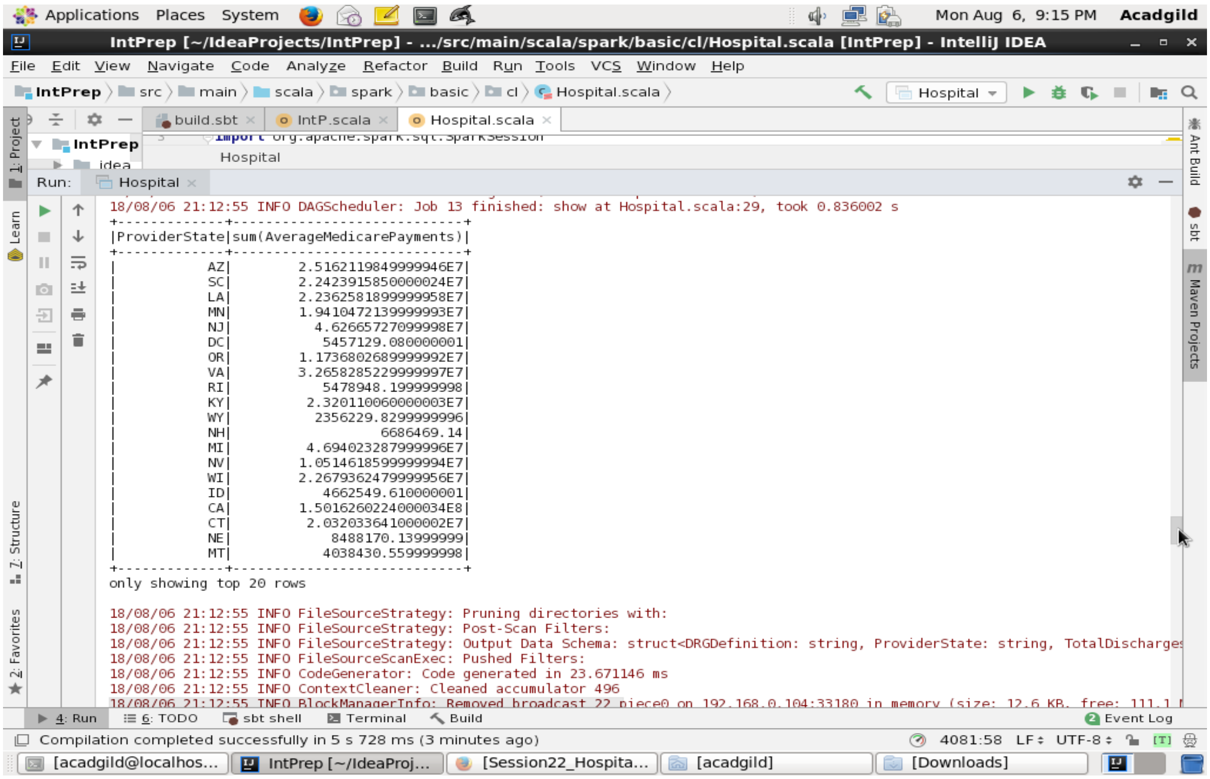
```
+-------------+-------------------------+
|ProviderState|sum(AverageTotalPayments)|
+-------------+-------------------------+
|           AZ|     2.8950559930000026E7|
|           SC|     2.6000001900000013E7|
|           LA|     2.6149231619999968E7|
|           MN|     2.2403429640000023E7|
|           NJ|     5.1536799209999874E7|
|           DC|        6005089.589999995|
|           OR|     1.3556614529999994E7|
|           VA|      3.850174243000001E7|
|           RI|        6179625.309999993|
|           KY|     2.6731563380000085E7|
|           WY|        2815426.019999998|
|           NH|        7645391.680000004|
|           MI|      5.285920417999992E7|
|           NV|     1.2370645069999998E7|
|           WI|     2.6273179719999947E7|
|           ID|        5414776.230000002|
|           CA|     1.6499398891999936E8|
|           CT|     2.2855921299999975E7|
|           NE|        9910246.840000004|
|           MT|        4681918.200000002|
+-------------+-------------------------+
only showing top 20 rows

18/08/06 21:12:52 INFO TaskSetManager: Finished task 91.0 in stage 17.0 (TID 253) in 7 ms on localhost (executor driver) (100/100)
18/08/06 21:12:52 INFO TaskSchedulerImpl: Removed TaskSet 17.0, whose tasks have all completed, from pool
18/08/06 21:12:52 INFO DAGScheduler: ResultStage 17 (show at Hospital.scala:27) finished in 1.210 s
18/08/06 21:12:52 INFO DAGScheduler: Job 9 finished: show at Hospital.scala:27, took 1.277217 s
18/08/06 21:12:52 INFO FileSourceStrategy: Pruning directories with:
18/08/06 21:12:52 INFO FileSourceStrategy: Post-Scan Filters:
18/08/06 21:12:52 INFO FileSourceStrategy: Output Data Schema: struct<ProviderState: string, AverageMedicarePayments: string>
18/08/06 21:12:52 INFO FileSourceScanExec: Pushed Filters:
```

Compilation completed successfully in 5 s 728 ms (2 minutes ago)

Applications Places System
Mon Aug 6, 9:14 PM
Acadgild
IntPrep [~/IdeaProjects/IntPrep] - .../src/main/scala/spark/basic/cl/Hospital.scala [IntPrep] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IntPrep › src › main › scala › spark › basic › cl › Hospital.scala
build.sbt IntP.scala Hospital.scala
Run: Hospital
4: Run 6: TODO sbt shell Terminal Build
Event Log
4081:58 LF UTF-8
[acadgild@localhos... IntPrep [~/IdeaProj... [Session22_Hospita... [acadgild] [Downloads]

➢ find out the AverageMedicarePayments charges per state.

```
val sumAMP =
hp.groupBy($"ProviderState").agg(sum("AverageMedicarePayments")).show()
```



```
18/08/06 21:12:55 INFO DAGScheduler: Job 13 finished: show at Hospital.scala:29, took 0.836002 s
+-------------+----------------------------+
|ProviderState|sum(AverageMedicarePayments)|
+-------------+----------------------------+
|           AZ|        2.5162119849999946E7|
|           SC|        2.2423915850000024E7|
|           LA|        2.2362581899999958E7|
|           MN|        1.9410472139999993E7|
|           NJ|          4.62665727099998E7|
|           DC|           5457129.080000001|
|           OR|        1.1736802689999992E7|
|           VA|        3.2658285229999997E7|
|           RI|           5478948.199999998|
|           KY|         2.320110060000003E7|
|           WY|          2356229.8299999996|
|           NH|                  6686469.14|
|           MI|         4.694023287999996E7|
|           NV|        1.0514618599999994E7|
|           WI|        2.2679362479999956E7|
|           ID|           4662549.610000001|
|           CA|        1.5016260224000034E8|
|           CT|         2.032033641000002E7|
|           NE|            8488170.13999999|
|           MT|           4038430.559999998|
+-------------+----------------------------+
only showing top 20 rows

18/08/06 21:12:55 INFO FileSourceStrategy: Pruning directories with:
18/08/06 21:12:55 INFO FileSourceStrategy: Post-Scan Filters:
18/08/06 21:12:55 INFO FileSourceStrategy: Output Data Schema: struct<DRGDefinition: string, ProviderState: string, TotalDischarges
18/08/06 21:12:55 INFO FileSourceScanExec: Pushed Filters:
18/08/06 21:12:55 INFO CodeGenerator: Code generated in 23.671146 ms
18/08/06 21:12:55 INFO ContextCleaner: Cleaned accumulator 496
18/08/06 21:12:55 INFO BlockManagerInfo: Removed broadcast 22 piece0 on 192.168.0.104:33180 in memory (size: 12.6 KB, free: 111.1 M
```

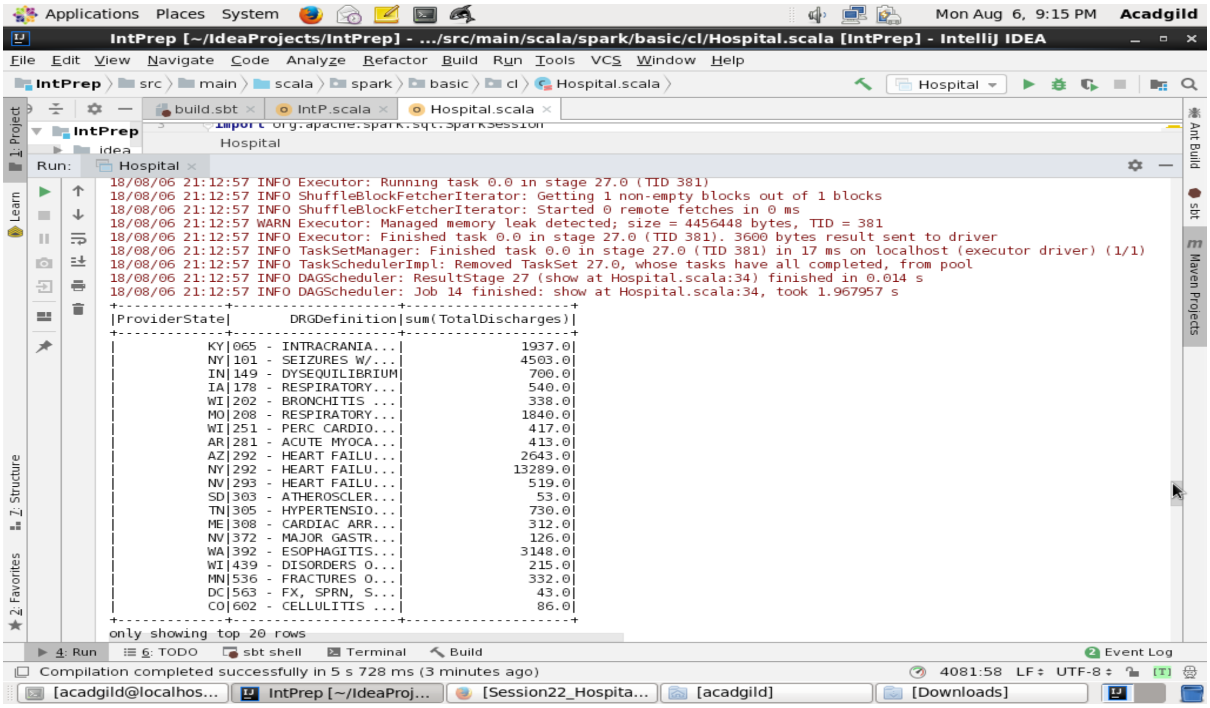
Compilation completed successfully in 5 s 728 ms (3 minutes ago)

Applications Places System
Mon Aug 6, 9:15 PM
Acadgild
IntPrep [~/IdeaProjects/IntPrep] - .../src/main/scala/spark/basic/cl/Hospital.scala [IntPrep] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IntPrep › src › main › scala › spark › basic › cl › Hospital.scala
build.sbt IntP.scala Hospital.scala
Run: Hospital
4: Run 6: TODO sbt shell Terminal Build
Event Log
4081:58 LF UTF-8
[acadgild@localhos... IntPrep [~/IdeaProj... [Session22_Hospita... [acadgild] [Downloads]

OBJECTIVE 3

➢ Find out the total number of Discharges per state and for each disease

```
val PSD = hp.groupBy($"ProviderState",$"DRGDefinition").agg(sum("TotalDischarges"))
 PSD.show()
```

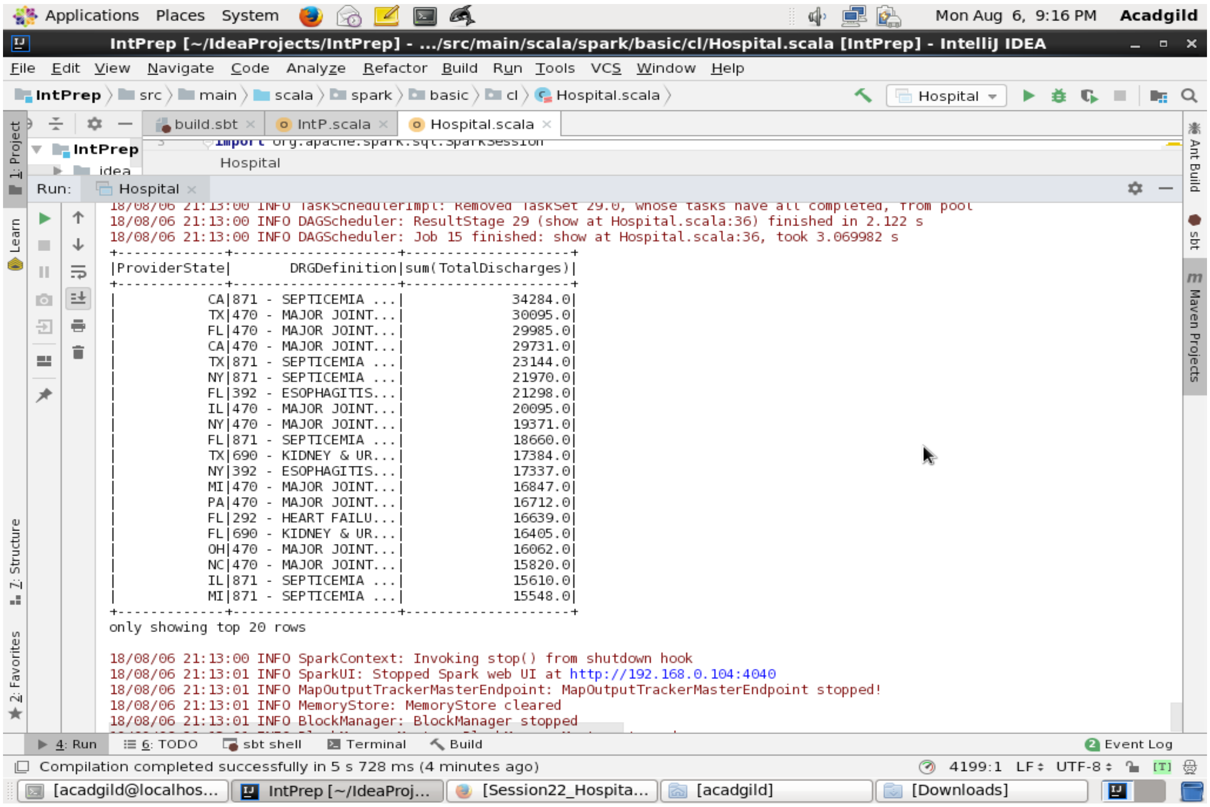

Applications Places System
Mon Aug 6, 9:15 PM Acadgild
IntPrep [~/IdeaProjects/IntPrep] - .../src/main/scala/spark/basic/cl/Hospital.scala [IntPrep] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IntPrep src main scala spark basic cl Hospital.scala
Hospital
build.sbt IntP.scala Hospital.scala
Run: Hospital
4: Run 6: TODO sbt shell Terminal Build Event Log
4081:58 LF UTF-8
[acadgild@localhos... IntPrep [~/IdeaProj... [Session22_Hospita... [acadgild] [Downloads]

```
18/08/06 21:12:57 INFO Executor: Running task 0.0 in stage 27.0 (TID 381)
18/08/06 21:12:57 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/08/06 21:12:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/08/06 21:12:57 WARN Executor: Managed memory leak detected; size = 4456448 bytes, TID = 381
18/08/06 21:12:57 INFO Executor: Finished task 0.0 in stage 27.0 (TID 381). 3600 bytes result sent to driver
18/08/06 21:12:57 INFO TaskSetManager: Finished task 0.0 in stage 27.0 (TID 381) in 17 ms on localhost (executor driver) (1/1)
18/08/06 21:12:57 INFO TaskSchedulerImpl: Removed TaskSet 27.0, whose tasks have all completed, from pool
18/08/06 21:12:57 INFO DAGScheduler: ResultStage 27 (show at Hospital.scala:34) finished in 0.014 s
18/08/06 21:12:57 INFO DAGScheduler: Job 14 finished: show at Hospital.scala:34, took 1.967957 s
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           KY|065 - INTRACRANIA...|              1937.0|
|           NY|101 - SEIZURES W/...|              4503.0|
|           IN|149 - DYSEQUILIBRIUM|               700.0|
|           IA|178 - RESPIRATORY...|               540.0|
|           WI|202 - BRONCHITIS ...|               338.0|
|           MO|208 - RESPIRATORY...|              1840.0|
|           WI|251 - PERC CARDIO...|               417.0|
|           AR|281 - ACUTE MYOCA...|               413.0|
|           AZ|292 - HEART FAILU...|              2643.0|
|           NY|292 - HEART FAILU...|             13289.0|
|           NV|293 - HEART FAILU...|               519.0|
|           SD|303 - ATHEROSCLER...|                53.0|
|           TN|305 - HYPERTENSIO...|               730.0|
|           ME|308 - CARDIAC ARR...|               312.0|
|           NV|372 - MAJOR GASTR...|               126.0|
|           WA|392 - ESOPHAGITIS...|              3148.0|
|           WI|439 - DISORDERS O...|               215.0|
|           MN|536 - FRACTURES O...|               332.0|
|           DC|563 - FX, SPRN, S...|                43.0|
|           CO|602 - CELLULITIS ...|                86.0|
+-------------+--------------------+--------------------+
only showing top 20 rows
```

Compilation completed successfully in 5 s 728 ms (3 minutes ago)

➢ Sort the output in descending order of totalDischarges

```
val DEOR = PSD.sort(desc("sum(TotalDischarges)")).show()
```


Applications Places System
Mon Aug 6, 9:16 PM Acadgild
IntPrep [~/IdeaProjects/IntPrep] - .../src/main/scala/spark/basic/cl/Hospital.scala [IntPrep] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
IntPrep src main scala spark basic cl Hospital.scala
Hospital
build.sbt IntP.scala Hospital.scala
Run: Hospital
4: Run 6: TODO sbt shell Terminal Build Event Log
4199:1 LF UTF-8
[acadgild@localhos... IntPrep [~/IdeaProj... [Session22_Hospita... [acadgild] [Downloads]

```
18/08/06 21:13:00 INFO TaskSchedulerImpl: Removed TaskSet 29.0, whose tasks have all completed, from pool
18/08/06 21:13:00 INFO DAGScheduler: ResultStage 29 (show at Hospital.scala:36) finished in 2.122 s
18/08/06 21:13:00 INFO DAGScheduler: Job 15 finished: show at Hospital.scala:36, took 3.069982 s
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           CA|871 - SEPTICEMIA ...|             34284.0|
|           TX|470 - MAJOR JOINT...|             30095.0|
|           FL|470 - MAJOR JOINT...|             29985.0|
|           CA|470 - MAJOR JOINT...|             29731.0|
|           TX|871 - SEPTICEMIA ...|             23144.0|
|           NY|871 - SEPTICEMIA ...|             21970.0|
|           FL|392 - ESOPHAGITIS...|             21298.0|
|           IL|470 - MAJOR JOINT...|             20095.0|
|           NY|470 - MAJOR JOINT...|             19371.0|
|           FL|871 - SEPTICEMIA ...|             18660.0|
|           TX|690 - KIDNEY & UR...|             17384.0|
|           NY|392 - ESOPHAGITIS...|             17337.0|
|           MI|470 - MAJOR JOINT...|             16847.0|
|           PA|470 - MAJOR JOINT...|             16712.0|
|           FL|292 - HEART FAILU...|             16639.0|
|           FL|690 - KIDNEY & UR...|             16405.0|
|           OH|470 - MAJOR JOINT...|             16062.0|
|           NC|470 - MAJOR JOINT...|             15820.0|
|           IL|871 - SEPTICEMIA ...|             15610.0|
|           MI|871 - SEPTICEMIA ...|             15548.0|
+-------------+--------------------+--------------------+
only showing top 20 rows

18/08/06 21:13:00 INFO SparkContext: Invoking stop() from shutdown hook
18/08/06 21:13:01 INFO SparkUI: Stopped Spark web UI at http://192.168.0.104:4040
18/08/06 21:13:01 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/08/06 21:13:01 INFO MemoryStore: MemoryStore cleared
18/08/06 21:13:01 INFO BlockManager: BlockManager stopped
```

Compilation completed successfully in 5 s 728 ms (4 minutes ago)