CASE STUDY 4

Hospital Analysis in US

Assignment 21. 4: Case Study Hospital in US

Problem Statement

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.

Record Size of Data-Set: 1,63,065

Initial Execution:

```
[acadgild@localhost ~]$ jps
3301 Jps
[acadgild@localhost ~]$ sudo service sshd start
[sudo] password for acadgild:
[acadgild@localhost ~]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
18/09/07 00:09:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-secondarynamenode-localhost.localdomain.out
18/09/07 00:10:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-resourcemanager-localhost.localdomain.out
localhost: starting nodemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-nodemanager-localhost.localdomain.out
[acadgild@localhost ~]$ jps
3889 ResourceManager
4740 Jps
3991 NodeManager
3545 DataNode
3449 NameNode
3690 SecondaryNameNode

[acadgild@localhost ~]$ spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
```

```
18/09/07 01:38:12 WARN DataNucleus.Query: Query for candidates of
org.apache.hadoop.hive.metastore.model.MPartitionColumnStatistics and subclasses resulted in no
possible candidates
Error(s) were found while auto-creating/validating the datastore for classes. The errors are printed in
the log, and are attached to this exception.
org.datanucleus.exceptions.NucleusDataStoreException: Error(s) were found while
auto-creating/validating the datastore for classes. The errors are printed in the log, and are attached to
this exception.
        at
org.datanucleus.store.rdbms.RDBMSStoreManager$ClassAdder.verifyErrors(RDBMSStoreManager.
java:3602)
        at
org.datanucleus.store.rdbms.RDBMSStoreManager$ClassAdder.addClassTablesAndValidate(RDBM
SStoreManager.java:3205)
        at
org.datanucleus.store.rdbms.RDBMSStoreManager$ClassAdder.run(RDBMSStoreManager.java:284
1)
        at
org.datanucleus.store.rdbms.AbstractSchemaTransaction.execute(AbstractSchemaTransaction.java:1
22)
        at
org.datanucleus.store.rdbms.RDBMSStoreManager.addClasses(RDBMSStoreManager.java:1605)
        at org.datanucleus.store.AbstractStoreManager.addClass(AbstractStoreManager.java:954)
        at
org.datanucleus.store.rdbms.RDBMSStoreManager.getDatastoreClass(RDBMSStoreManager.java:6
79)
        at
org.datanucleus.store.rdbms.query.RDBMSQueryUtils.getStatementForCandidates(RDBMSQueryUt
ils.java:408)
        at
org.datanucleus.store.rdbms.query.JDOQLQuery.compileQueryFull(JDOQLQuery.java:947)
        at org.datanucleus.store.rdbms.query.JDOQLQuery.compileInternal(JDOQLQuery.java:370)
        at org.datanucleus.store.query.Query.executeQuery(Query.java:1744)
        at org.datanucleus.store.query.Query.executeWithArray(Query.java:1672)
        at org.datanucleus.store.query.Query.execute(Query.java:1654)
        at org.datanucleus.api.jdo.JDOQuery.execute(JDOQuery.java:221)
        at
org.apache.hadoop.hive.metastore.MetaStoreDirectSql.ensureDbInit(MetaStoreDirectSql.java:185)
        at
org.apache.hadoop.hive.metastore.MetaStoreDirectSql.<init>(MetaStoreDirectSql.java:137)
        at org.apache.hadoop.hive.metastore.ObjectStore.initialize(ObjectStore.java:295)
        at org.apache.hadoop.hive.metastore.ObjectStore.setConf(ObjectStore.java:258)
        at org.apache.hadoop.util.ReflectionUtils.setConf(ReflectionUtils.java:73)
        at org.apache.hadoop.util.ReflectionUtils.newInstance(ReflectionUtils.java:133)
        at org.apache.hadoop.hive.metastore.RawStoreProxy.<init>(RawStoreProxy.java:57)
        at org.apache.hadoop.hive.metastore.RawStoreProxy.getProxy(RawStoreProxy.java:66)
        at
```

```
org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.newRawStore(HiveMetaStore.java:
593)
        at
org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.getMS(HiveMetaStore.java:571)
        at
org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.createDefaultDB(HiveMetaStore.ja
va:620)
        at
org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.init(HiveMetaStore.java:461)
        at
org.apache.hadoop.hive.metastore.RetryingHMSHandler.<init>(RetryingHMSHandler.java:66)
        at
org.apache.hadoop.hive.metastore.RetryingHMSHandler.getProxy(RetryingHMSHandler.java:72)
        at
org.apache.hadoop.hive.metastore.HiveMetaStore.newRetryingHMSHandler(HiveMetaStore.java:57
62)
        at
org.apache.hadoop.hive.metastore.HiveMetaStoreClient.<init>(HiveMetaStoreClient.java:199)
        at
org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient.<init>(SessionHiveMetaStoreClien
t.java:74)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at org.apache.hadoop.hive.metastore.MetaStoreUtils.newInstance(MetaStoreUtils.java:1521)
        at
org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.<init>(RetryingMetaStoreClient.java:86)
        at
org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.getProxy(RetryingMetaStoreClient.java:
132)
        at
org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.getProxy(RetryingMetaStoreClient.java:
104)
        at org.apache.hadoop.hive.ql.metadata.Hive.createMetaStoreClient(Hive.java:3005)
        at org.apache.hadoop.hive.ql.metadata.Hive.getMSC(Hive.java:3024)
        at org.apache.hadoop.hive.ql.metadata.Hive.getAllDatabases(Hive.java:1234)
        at org.apache.hadoop.hive.ql.metadata.Hive.reloadFunctions(Hive.java:174)
        at org.apache.hadoop.hive.ql.metadata.Hive.<clinit>(Hive.java:166)
        at org.apache.hadoop.hive.ql.session.SessionState.start(SessionState.java:503)
        at org.apache.spark.sql.hive.client.HiveClientImpl.<init>(HiveClientImpl.scala:192)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at
org.apache.spark.sql.hive.client.IsolatedClientLoader.createClient(IsolatedClientLoader.scala:264)
        at org.apache.spark.sql.hive.HiveUtils$.newClientForMetadata(HiveUtils.scala:366)
        at org.apache.spark.sql.hive.HiveUtils$.newClientForMetadata(HiveUtils.scala:270)
        at org.apache.spark.sql.hive.HiveExternalCatalog.<init>(HiveExternalCatalog.scala:65)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at org.apache.spark.sql.internal.SharedState$.org$apache$spark$sql$internal$SharedState$
$reflect(SharedState.scala:166)
        at org.apache.spark.sql.internal.SharedState.<init>(SharedState.scala:86)
        at org.apache.spark.sql.SparkSession$$anonfun$sharedState$1.apply(SparkSession.scala:101)
        at org.apache.spark.sql.SparkSession$$anonfun$sharedState$1.apply(SparkSession.scala:101)
        at scala.Option.getOrElse(Option.scala:121)
        at org.apache.spark.sql.SparkSession.sharedState$lzycompute(SparkSession.scala:101)
        at org.apache.spark.sql.SparkSession.sharedState(SparkSession.scala:100)
        at org.apache.spark.sql.internal.SessionState.<init>(SessionState.scala:157)
        at org.apache.spark.sql.hive.HiveSessionState.<init>(HiveSessionState.scala:32)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at org.apache.spark.sql.SparkSession$.org$apache$spark$sql$SparkSession$
$reflect(SparkSession.scala:978)
        at org.apache.spark.sql.SparkSession.sessionState$lzycompute(SparkSession.scala:110)
        at org.apache.spark.sql.SparkSession.sessionState(SparkSession.scala:109)
        at org.apache.spark.sql.SparkSession$Builder$
$anonfun$getOrCreate$5.apply(SparkSession.scala:878)
        at org.apache.spark.sql.SparkSession$Builder$
$anonfun$getOrCreate$5.apply(SparkSession.scala:878)
        at scala.collection.mutable.HashMap$$anonfun$foreach$1.apply(HashMap.scala:99)
        at scala.collection.mutable.HashMap$$anonfun$foreach$1.apply(HashMap.scala:99)
        at scala.collection.mutable.HashTable$class.foreachEntry(HashTable.scala:230)
        at scala.collection.mutable.HashMap.foreachEntry(HashMap.scala:40)
        at scala.collection.mutable.HashMap.foreach(HashMap.scala:99)
        at org.apache.spark.sql.SparkSession$Builder.getOrCreate(SparkSession.scala:878)
        at org.apache.spark.repl.Main$.createSparkSession(Main.scala:95)
        at $line3.$read$$iw$$iw.<init>(<console>:15)
        at $line3.$read$$iw.<init>(<console>:42)
        at $line3.$read.<init>(<console>:44)
```

```
	at $line3.$read$.<init>(<console>:48)
	at $line3.$read$.<clinit>(<console>)
	at $line3.$eval$.$print$lzycompute(<console>:7)
	at $line3.$eval$.$print(<console>:6)
	at $line3.$eval.$print(<console>)
	at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
	at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
	at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
	at java.lang.reflect.Method.invoke(Method.java:498)
	at scala.tools.nsc.interpreter.IMain$ReadEvalPrint.call(IMain.scala:786)
	at scala.tools.nsc.interpreter.IMain$Request.loadAndRun(IMain.scala:1047)
	at scala.tools.nsc.interpreter.IMain$WrappedRequest$
$anonfun$loadAndRunReq$1.apply(IMain.scala:638)
	at scala.tools.nsc.interpreter.IMain$WrappedRequest$
$anonfun$loadAndRunReq$1.apply(IMain.scala:637)
	at scala.reflect.internal.util.ScalaClassLoader$class.asContext(ScalaClassLoader.scala:31)
	at
scala.reflect.internal.util.AbstractFileClassLoader.asContext(AbstractFileClassLoader.scala:19)
	at scala.tools.nsc.interpreter.IMain$WrappedRequest.loadAndRunReq(IMain.scala:637)
	at scala.tools.nsc.interpreter.IMain.interpret(IMain.scala:569)
	at scala.tools.nsc.interpreter.IMain.interpret(IMain.scala:565)
	at scala.tools.nsc.interpreter.ILoop.interpretStartingWith(ILoop.scala:807)
	at scala.tools.nsc.interpreter.ILoop.command(ILoop.scala:681)
	at scala.tools.nsc.interpreter.ILoop.processLine(ILoop.scala:395)
	at org.apache.spark.repl.SparkILoop$
$anonfun$initializeSpark$1.apply$mcV$sp(SparkILoop.scala:38)
	at org.apache.spark.repl.SparkILoop$$anonfun$initializeSpark$1.apply(SparkILoop.scala:37)
	at org.apache.spark.repl.SparkILoop$$anonfun$initializeSpark$1.apply(SparkILoop.scala:37)
	at scala.tools.nsc.interpreter.IMain.beQuietDuring(IMain.scala:214)
	at org.apache.spark.repl.SparkILoop.initializeSpark(SparkILoop.scala:37)
	at org.apache.spark.repl.SparkILoop.loadFiles(SparkILoop.scala:105)
	at scala.tools.nsc.interpreter.ILoop$$anonfun$process$1.apply$mcZ$sp(ILoop.scala:920)
	at scala.tools.nsc.interpreter.ILoop$$anonfun$process$1.apply(ILoop.scala:909)
	at scala.tools.nsc.interpreter.ILoop$$anonfun$process$1.apply(ILoop.scala:909)
	at
scala.reflect.internal.util.ScalaClassLoader$.savingContextLoader(ScalaClassLoader.scala:97)
	at scala.tools.nsc.interpreter.ILoop.process(ILoop.scala:909)
	at org.apache.spark.repl.Main$.doMain(Main.scala:68)
	at org.apache.spark.repl.Main$.main(Main.scala:51)
	at org.apache.spark.repl.Main.main(Main.scala)
	at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
	at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
	at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
	at java.lang.reflect.Method.invoke(Method.java:498)
	at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$
$runMain(SparkSubmit.scala:738)
	at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:187)
	at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:212)
	at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:126)
	at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
Caused by: com.mysql.jdbc.exceptions.jdbc4.MySQLSyntaxErrorException: Specified key was too
long; max key length is 3072 bytes
	at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
	at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
	at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
	at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
	at com.mysql.jdbc.Util.handleNewInstance(Util.java:425)
	at com.mysql.jdbc.Util.getInstance(Util.java:408)
	at com.mysql.jdbc.SQLError.createSQLException(SQLError.java:944)
	at com.mysql.jdbc.MysqlIO.checkErrorPacket(MysqlIO.java:3976)
	at com.mysql.jdbc.MysqlIO.checkErrorPacket(MysqlIO.java:3912)
	at com.mysql.jdbc.MysqlIO.sendCommand(MysqlIO.java:2530)
	at com.mysql.jdbc.MysqlIO.sqlQueryDirect(MysqlIO.java:2683)
	at com.mysql.jdbc.ConnectionImpl.execSQL(ConnectionImpl.java:2482)
	at com.mysql.jdbc.ConnectionImpl.execSQL(ConnectionImpl.java:2440)
	at com.mysql.jdbc.StatementImpl.executeInternal(StatementImpl.java:845)
	at com.mysql.jdbc.StatementImpl.execute(StatementImpl.java:745)
	at com.jolbox.bonecp.StatementHandle.execute(StatementHandle.java:254)
	at
org.datanucleus.store.rdbms.table.AbstractTable.executeDdlStatement(AbstractTable.java:760)
	at org.datanucleus.store.rdbms.table.TableImpl.createIndices(TableImpl.java:648)
	at org.datanucleus.store.rdbms.table.TableImpl.createConstraints(TableImpl.java:422)
	at
org.datanucleus.store.rdbms.RDBMSStoreManager$ClassAdder.performTablesValidation(RDBMSSt
oreManager.java:3459)
	at
org.datanucleus.store.rdbms.RDBMSStoreManager$ClassAdder.addClassTablesAndValidate(RDBM
SStoreManager.java:3190)
	... 128 more
Nested Throwables StackTrace:
com.mysql.jdbc.exceptions.jdbc4.MySQLSyntaxErrorException: Specified key was too long; max
key length is 3072 bytes
	at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
	at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
	at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
	at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
	at com.mysql.jdbc.Util.handleNewInstance(Util.java:425)
	at com.mysql.jdbc.Util.getInstance(Util.java:408)
	at com.mysql.jdbc.SQLError.createSQLException(SQLError.java:944)
	at com.mysql.jdbc.MysqlIO.checkErrorPacket(MysqlIO.java:3976)
```

```
        at com.mysql.jdbc.MysqlIO.checkErrorPacket(MysqlIO.java:3912)
        at com.mysql.jdbc.MysqlIO.sendCommand(MysqlIO.java:2530)
        at com.mysql.jdbc.MysqlIO.sqlQueryDirect(MysqlIO.java:2683)
        at com.mysql.jdbc.ConnectionImpl.execSQL(ConnectionImpl.java:2482)
        at com.mysql.jdbc.ConnectionImpl.execSQL(ConnectionImpl.java:2440)
        at com.mysql.jdbc.StatementImpl.executeInternal(StatementImpl.java:845)
        at com.mysql.jdbc.StatementImpl.execute(StatementImpl.java:745)
        at com.jolbox.bonecp.StatementHandle.execute(StatementHandle.java:254)
        at
org.datanucleus.store.rdbms.table.AbstractTable.executeDdlStatement(AbstractTable.java:760)
        at org.datanucleus.store.rdbms.table.TableImpl.createIndices(TableImpl.java:648)
        at org.datanucleus.store.rdbms.table.TableImpl.createConstraints(TableImpl.java:422)
        at
org.datanucleus.store.rdbms.RDBMSStoreManager$ClassAdder.performTablesValidation(RDBMSSt
oreManager.java:3459)
        at
org.datanucleus.store.rdbms.RDBMSStoreManager$ClassAdder.addClassTablesAndValidate(RDBM
SStoreManager.java:3190)
        at
org.datanucleus.store.rdbms.RDBMSStoreManager$ClassAdder.run(RDBMSStoreManager.java:284
1)
        at
org.datanucleus.store.rdbms.AbstractSchemaTransaction.execute(AbstractSchemaTransaction.java:1
22)
        at
org.datanucleus.store.rdbms.RDBMSStoreManager.addClasses(RDBMSStoreManager.java:1605)
        at org.datanucleus.store.AbstractStoreManager.addClass(AbstractStoreManager.java:954)
        at
org.datanucleus.store.rdbms.RDBMSStoreManager.getDatastoreClass(RDBMSStoreManager.java:6
79)
        at
org.datanucleus.store.rdbms.query.RDBMSQueryUtils.getStatementForCandidates(RDBMSQueryUt
ils.java:408)
        at
org.datanucleus.store.rdbms.query.JDOQLQuery.compileQueryFull(JDOQLQuery.java:947)
        at org.datanucleus.store.rdbms.query.JDOQLQuery.compileInternal(JDOQLQuery.java:370)
        at org.datanucleus.store.query.Query.executeQuery(Query.java:1744)
        at org.datanucleus.store.query.Query.executeWithArray(Query.java:1672)
        at org.datanucleus.store.query.Query.execute(Query.java:1654)
        at org.datanucleus.api.jdo.JDOQuery.execute(JDOQuery.java:221)
        at
org.apache.hadoop.hive.metastore.MetaStoreDirectSql.ensureDbInit(MetaStoreDirectSql.java:185)
        at
org.apache.hadoop.hive.metastore.MetaStoreDirectSql.<init>(MetaStoreDirectSql.java:137)
        at org.apache.hadoop.hive.metastore.ObjectStore.initialize(ObjectStore.java:295)
        at org.apache.hadoop.hive.metastore.ObjectStore.setConf(ObjectStore.java:258)
        at org.apache.hadoop.util.ReflectionUtils.setConf(ReflectionUtils.java:73)
        at org.apache.hadoop.util.ReflectionUtils.newInstance(ReflectionUtils.java:133)
        at org.apache.hadoop.hive.metastore.RawStoreProxy.<init>(RawStoreProxy.java:57)
        at org.apache.hadoop.hive.metastore.RawStoreProxy.getProxy(RawStoreProxy.java:66)
        at
org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.newRawStore(HiveMetaStore.java:
593)
        at
org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.getMS(HiveMetaStore.java:571)
        at
org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.createDefaultDB(HiveMetaStore.ja
va:620)
        at
org.apache.hadoop.hive.metastore.HiveMetaStore$HMSHandler.init(HiveMetaStore.java:461)
        at
org.apache.hadoop.hive.metastore.RetryingHMSHandler.<init>(RetryingHMSHandler.java:66)
        at
org.apache.hadoop.hive.metastore.RetryingHMSHandler.getProxy(RetryingHMSHandler.java:72)
        at
org.apache.hadoop.hive.metastore.HiveMetaStore.newRetryingHMSHandler(HiveMetaStore.java:57
62)
        at
org.apache.hadoop.hive.metastore.HiveMetaStoreClient.<init>(HiveMetaStoreClient.java:199)
        at
org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClient.<init>(SessionHiveMetaStoreClien
t.java:74)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at org.apache.hadoop.hive.metastore.MetaStoreUtils.newInstance(MetaStoreUtils.java:1521)
        at
org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.<init>(RetryingMetaStoreClient.java:86)
        at
org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.getProxy(RetryingMetaStoreClient.java:
132)
        at
org.apache.hadoop.hive.metastore.RetryingMetaStoreClient.getProxy(RetryingMetaStoreClient.java:
104)
        at org.apache.hadoop.hive.ql.metadata.Hive.createMetaStoreClient(Hive.java:3005)
        at org.apache.hadoop.hive.ql.metadata.Hive.getMSC(Hive.java:3024)
        at org.apache.hadoop.hive.ql.metadata.Hive.getAllDatabases(Hive.java:1234)
        at org.apache.hadoop.hive.ql.metadata.Hive.reloadFunctions(Hive.java:174)
        at org.apache.hadoop.hive.ql.metadata.Hive.<clinit>(Hive.java:166)
        at org.apache.hadoop.hive.ql.session.SessionState.start(SessionState.java:503)
        at org.apache.spark.sql.hive.client.HiveClientImpl.<init>(HiveClientImpl.scala:192)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
```

```
        at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at
org.apache.spark.sql.hive.client.IsolatedClientLoader.createClient(IsolatedClientLoader.scala:264)
        at org.apache.spark.sql.hive.HiveUtils$.newClientForMetadata(HiveUtils.scala:366)
        at org.apache.spark.sql.hive.HiveUtils$.newClientForMetadata(HiveUtils.scala:270)
        at org.apache.spark.sql.hive.HiveExternalCatalog.<init>(HiveExternalCatalog.scala:65)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at org.apache.spark.sql.internal.SharedState$.org$apache$spark$sql$internal$SharedState$
$reflect(SharedState.scala:166)
        at org.apache.spark.sql.internal.SharedState.<init>(SharedState.scala:86)
        at org.apache.spark.sql.SparkSession$$anonfun$sharedState$1.apply(SparkSession.scala:101)
        at org.apache.spark.sql.SparkSession$$anonfun$sharedState$1.apply(SparkSession.scala:101)
        at scala.Option.getOrElse(Option.scala:121)
        at org.apache.spark.sql.SparkSession.sharedState$lzycompute(SparkSession.scala:101)
        at org.apache.spark.sql.SparkSession.sharedState(SparkSession.scala:100)
        at org.apache.spark.sql.internal.SessionState.<init>(SessionState.scala:157)
        at org.apache.spark.sql.hive.HiveSessionState.<init>(HiveSessionState.scala:32)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
        at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.ja
va:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
        at org.apache.spark.sql.SparkSession$.org$apache$spark$sql$SparkSession$
$reflect(SparkSession.scala:978)
        at org.apache.spark.sql.SparkSession.sessionState$lzycompute(SparkSession.scala:110)
        at org.apache.spark.sql.SparkSession.sessionState(SparkSession.scala:109)
        at org.apache.spark.sql.SparkSession$Builder$
$anonfun$getOrCreate$5.apply(SparkSession.scala:878)
        at org.apache.spark.sql.SparkSession$Builder$
$anonfun$getOrCreate$5.apply(SparkSession.scala:878)
        at scala.collection.mutable.HashMap$$anonfun$foreach$1.apply(HashMap.scala:99)
        at scala.collection.mutable.HashMap$$anonfun$foreach$1.apply(HashMap.scala:99)
        at scala.collection.mutable.HashTable$class.foreachEntry(HashTable.scala:230)
        at scala.collection.mutable.HashMap.foreachEntry(HashMap.scala:40)
        at scala.collection.mutable.HashMap.foreach(HashMap.scala:99)
        at org.apache.spark.sql.SparkSession$Builder.getOrCreate(SparkSession.scala:878)
        at org.apache.spark.repl.Main$.createSparkSession(Main.scala:95)
        at $line3.$read$$iw$$iw.<init>(<console>:15)
```

```
	at $line3.$read$$iw.<init>(<console>:42)
	at $line3.$read.<init>(<console>:44)
	at $line3.$read$.<init>(<console>:48)
	at $line3.$read$.<clinit>(<console>)
	at $line3.$eval$.$print$lzycompute(<console>:7)
	at $line3.$eval$.$print(<console>:6)
	at $line3.$eval.$print(<console>)
	at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
	at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
	at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
	at java.lang.reflect.Method.invoke(Method.java:498)
	at scala.tools.nsc.interpreter.IMain$ReadEvalPrint.call(IMain.scala:786)
	at scala.tools.nsc.interpreter.IMain$Request.loadAndRun(IMain.scala:1047)
	at scala.tools.nsc.interpreter.IMain$WrappedRequest$
$anonfun$loadAndRunReq$1.apply(IMain.scala:638)
	at scala.tools.nsc.interpreter.IMain$WrappedRequest$
$anonfun$loadAndRunReq$1.apply(IMain.scala:637)
	at scala.reflect.internal.util.ScalaClassLoader$class.asContext(ScalaClassLoader.scala:31)
	at
scala.reflect.internal.util.AbstractFileClassLoader.asContext(AbstractFileClassLoader.scala:19)
	at scala.tools.nsc.interpreter.IMain$WrappedRequest.loadAndRunReq(IMain.scala:637)
	at scala.tools.nsc.interpreter.IMain.interpret(IMain.scala:569)
	at scala.tools.nsc.interpreter.IMain.interpret(IMain.scala:565)
	at scala.tools.nsc.interpreter.ILoop.interpretStartingWith(ILoop.scala:807)
	at scala.tools.nsc.interpreter.ILoop.command(ILoop.scala:681)
	at scala.tools.nsc.interpreter.ILoop.processLine(ILoop.scala:395)
	at org.apache.spark.repl.SparkILoop$
$anonfun$initializeSpark$1.apply$mcV$sp(SparkILoop.scala:38)
	at org.apache.spark.repl.SparkILoop$$anonfun$initializeSpark$1.apply(SparkILoop.scala:37)
	at org.apache.spark.repl.SparkILoop$$anonfun$initializeSpark$1.apply(SparkILoop.scala:37)
	at scala.tools.nsc.interpreter.IMain.beQuietDuring(IMain.scala:214)
	at org.apache.spark.repl.SparkILoop.initializeSpark(SparkILoop.scala:37)
	at org.apache.spark.repl.SparkILoop.loadFiles(SparkILoop.scala:105)
	at scala.tools.nsc.interpreter.ILoop$$anonfun$process$1.apply$mcZ$sp(ILoop.scala:920)
	at scala.tools.nsc.interpreter.ILoop$$anonfun$process$1.apply(ILoop.scala:909)
	at scala.tools.nsc.interpreter.ILoop$$anonfun$process$1.apply(ILoop.scala:909)
	at
scala.reflect.internal.util.ScalaClassLoader$.savingContextLoader(ScalaClassLoader.scala:97)
	at scala.tools.nsc.interpreter.ILoop.process(ILoop.scala:909)
	at org.apache.spark.repl.Main$.doMain(Main.scala:68)
	at org.apache.spark.repl.Main$.main(Main.scala:51)
	at org.apache.spark.repl.Main.main(Main.scala)
	at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
	at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
	at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
	at java.lang.reflect.Method.invoke(Method.java:498)
	at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$
```

```
$runMain(SparkSubmit.scala:738)
	at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:187)
	at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:212)
	at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:126)
	at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
18/09/07 01:38:18 WARN metastore.ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1536264475926).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.1.0
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) Client VM, Java 1.8.0_171)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Solution:

Objective 1 :

Load file into spark

Query :

In Spark-2.0, we can load a CSV file directly into the Spark SQL context as follows:

1. **val session =** org.apache.spark.sql.SparkSession.builder.master("local").appName("Spark CSV Reader").getOrCreate;

2. val df = session.read.format("com.databricks.spark.csv").option("header","true").option("inferSchema","true").load("file:///home/acadgild/Desktop/inpatientCharges.csv")

With this, 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.

- To see the contents inside the DataFrame, type the following:

  df.show

- Next, we will save the data in a table by registering it in a temp table as shown below.

  df.registerTempTable(“hospital_charges”)

Terminal Execution:

```
scala> val session = org.apache.spark.sql.SparkSession.builder.master("local").appName("Spark
CSV Reader").getOrCreate;
18/09/07 02:11:47 WARN sql.SparkSession$Builder: Using an existing SparkSession; some
configuration may not take effect.
session: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@b55a0e

scala> val df  =
session.read.format("com.databricks.spark.csv").option("header","true").option("inferSchema","true"
).load("file:///home/acadgild/Desktop/inpatientCharges.csv")
18/09/07 02:20:30 WARN util.SizeEstimator: Failed to check whether UseCompressedOops is set;
assuming yes
df: org.apache.spark.sql.DataFrame = [DRGDefinition: string, ProviderId: int ... 10 more fields]


scala> df.show
+--------------------+----------+--------------------+--------------------+------------+-------------+--------------
-+--------------------------------+---------------+--------------------+--------------------+-----------------------
+
|       DRGDefinition|ProviderId|
ProviderName|ProviderStreetAddress|ProviderCity|ProviderState|ProviderZipCode|HospitalReferral
RegionDescription|TotalDischarges|AverageCoveredCharges|AverageTotalPayments|AverageMedicar
ePayments|
+--------------------+----------+--------------------+--------------------+------------+-------------+--------------
-+--------------------------------+---------------+--------------------+--------------------+-----------------------
+
|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|            22862.23|
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


scala> df.registerTempTable("hospital_charges")
warning: there was one deprecation warning; re-run with -deprecation for details
```

Objective 2 :

1. **What is the average amount of AverageCoveredCharges per state**

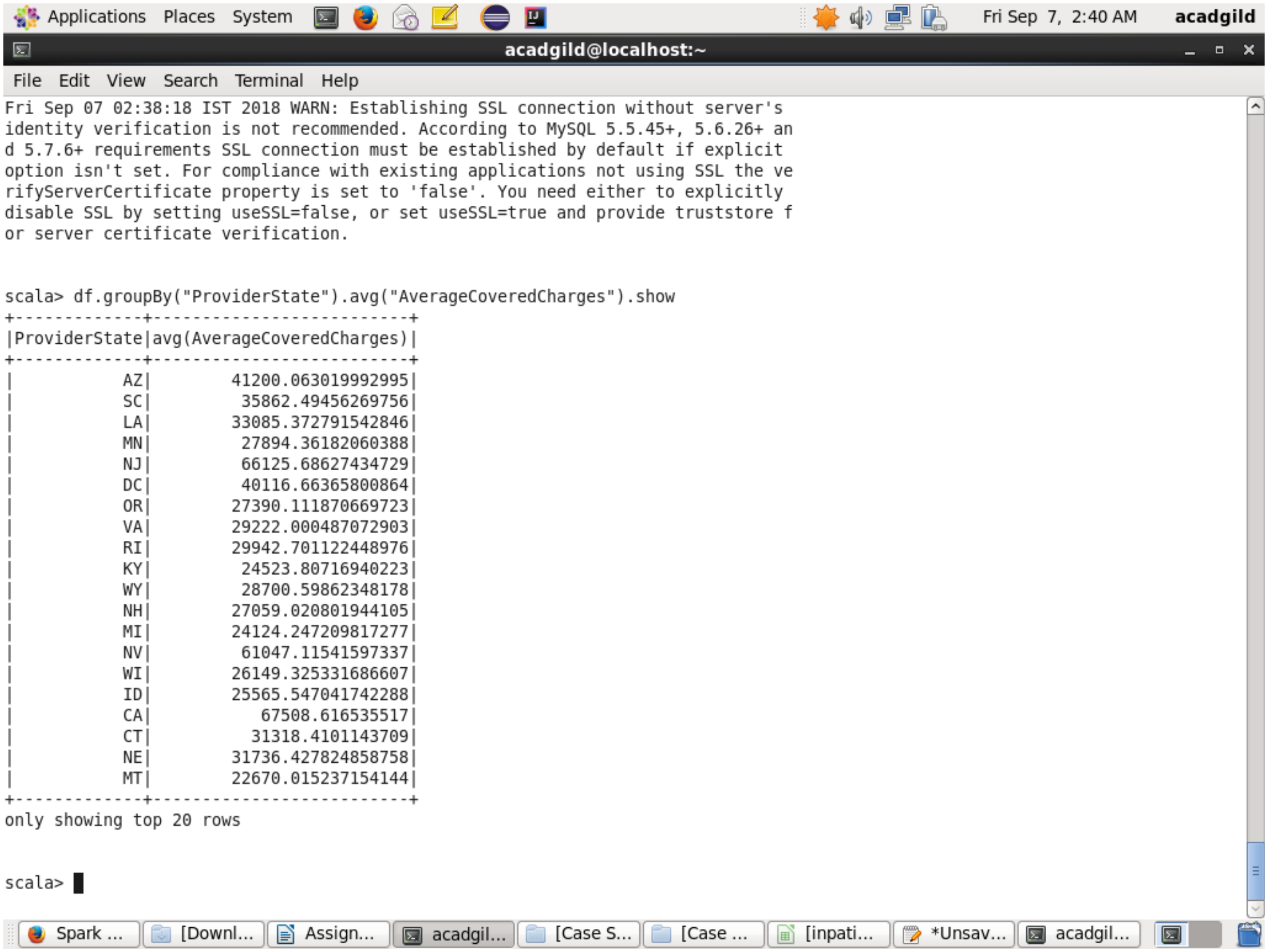
Query :

```
df.groupBy("ProviderState").avg("AverageCoveredCharges").show
```

Terminal Execution:

```
scala> df.groupBy("ProviderState").avg("AverageCoveredCharges").show
+-------------+--------------------------+
|ProviderState|avg(AverageCoveredCharges)|
+-------------+--------------------------+
|           AZ|        41200.063019992995|
|           SC|         35862.49456269756|
|           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
```

Output :



```
Fri Sep 07 02:38:18 IST 2018 WARN: Establishing SSL connection without server's
identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ an
d 5.7.6+ requirements SSL connection must be established by default if explicit
option isn't set. For compliance with existing applications not using SSL the ve
rifyServerCertificate property is set to 'false'. You need either to explicitly
disable SSL by setting useSSL=false, or set useSSL=true and provide truststore f
or server certificate verification.


scala> df.groupBy("ProviderState").avg("AverageCoveredCharges").show
+-------------+--------------------------+
|ProviderState|avg(AverageCoveredCharges)|
+-------------+--------------------------+
|           AZ|        41200.063019992995|
|           SC|         35862.49456269756|
|           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


scala> █
```

2. Find out the AverageTotalPayments charges per state

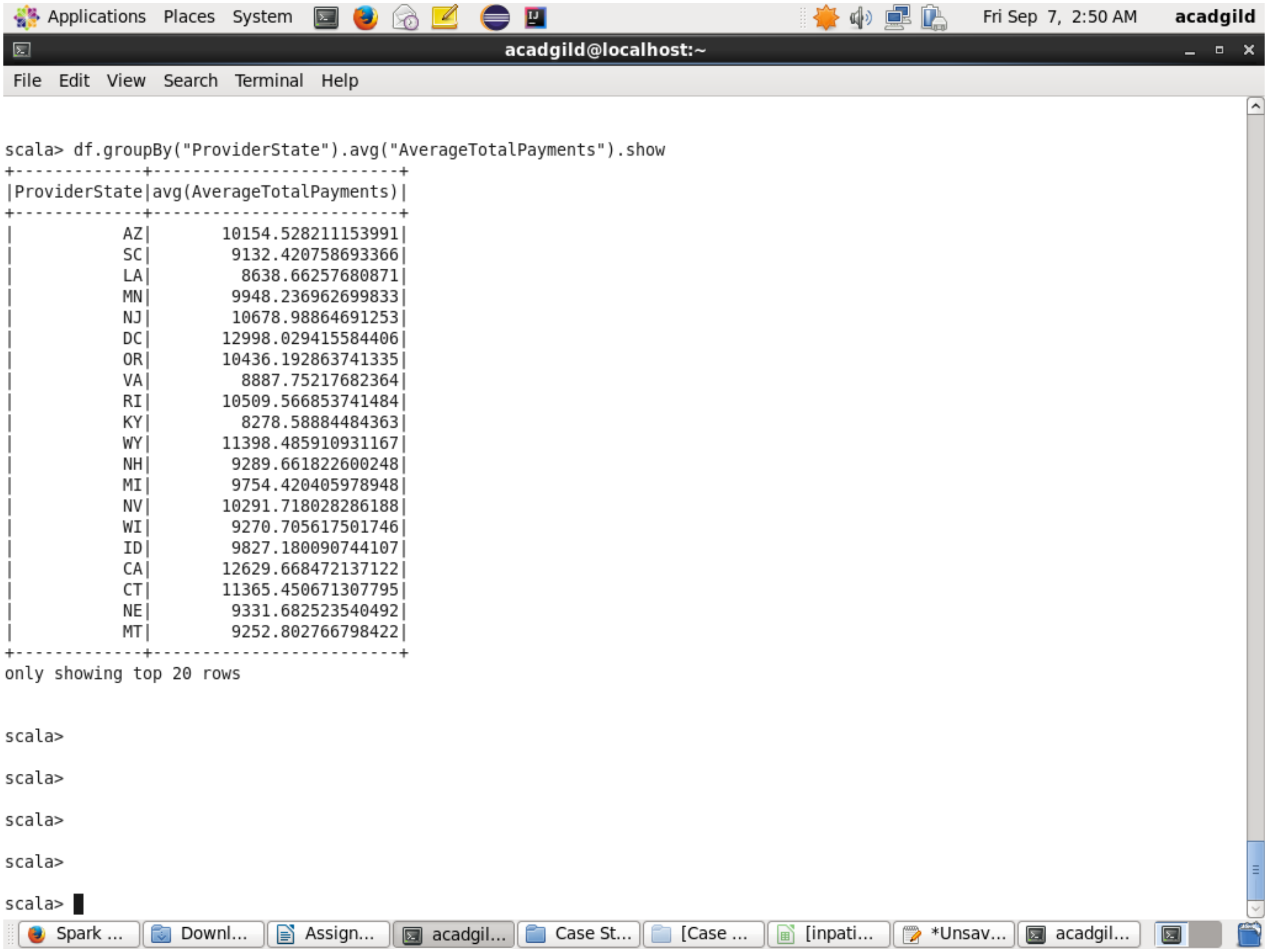
Query :

```
df.groupBy("ProviderState").avg("AverageTotalPayments").show
```

Terminal Execution:

```
scala> df.groupBy("ProviderState").avg("AverageTotalPayments").show
+-------------+-------------------------+
|ProviderState|avg(AverageTotalPayments)|
+-------------+-------------------------+
|           AZ|       10154.528211153991|
|           SC|        9132.420758693366|
|           LA|         8638.66257680871|
|           MN|        9948.236962699833|
|           NJ|        10678.98864691253|
|           DC|       12998.029415584406|
|           OR|       10436.192863741335|
|           VA|         8887.75217682364|
|           RI|       10509.566853741484|
|           KY|         8278.58884484363|
|           WY|       11398.485910931167|
|           NH|        9289.661822600248|
|           MI|        9754.420405978948|
|           NV|       10291.718028286188|
|           WI|        9270.705617501746|
|           ID|        9827.180090744107|
|           CA|       12629.668472137122|
|           CT|       11365.450671307795|
|           NE|        9331.682523540492|
|           MT|        9252.802766798422|
+-------------+-------------------------+
only showing top 20 rows
```

Output :



```
scala> df.groupBy("ProviderState").avg("AverageTotalPayments").show
+-------------+------------------------+
|ProviderState|avg(AverageTotalPayments)|
+-------------+------------------------+
|           AZ|       10154.528211153991|
|           SC|        9132.420758693366|
|           LA|         8638.66257680871|
|           MN|        9948.236962699833|
|           NJ|        10678.98864691253|
|           DC|       12998.029415584406|
|           OR|       10436.192863741335|
|           VA|         8887.75217682364|
|           RI|       10509.566853741484|
|           KY|         8278.58884484363|
|           WY|       11398.485910931167|
|           NH|        9289.661822600248|
|           MI|        9754.420405978948|
|           NV|       10291.718028286188|
|           WI|        9270.705617501746|
|           ID|        9827.180090744107|
|           CA|       12629.668472137122|
|           CT|       11365.450671307795|
|           NE|        9331.682523540492|
|           MT|        9252.802766798422|
+-------------+------------------------+
only showing top 20 rows


scala>

scala>

scala>

scala>

scala> 
```

3. Find out the AverageMedicarePayments charges per state.

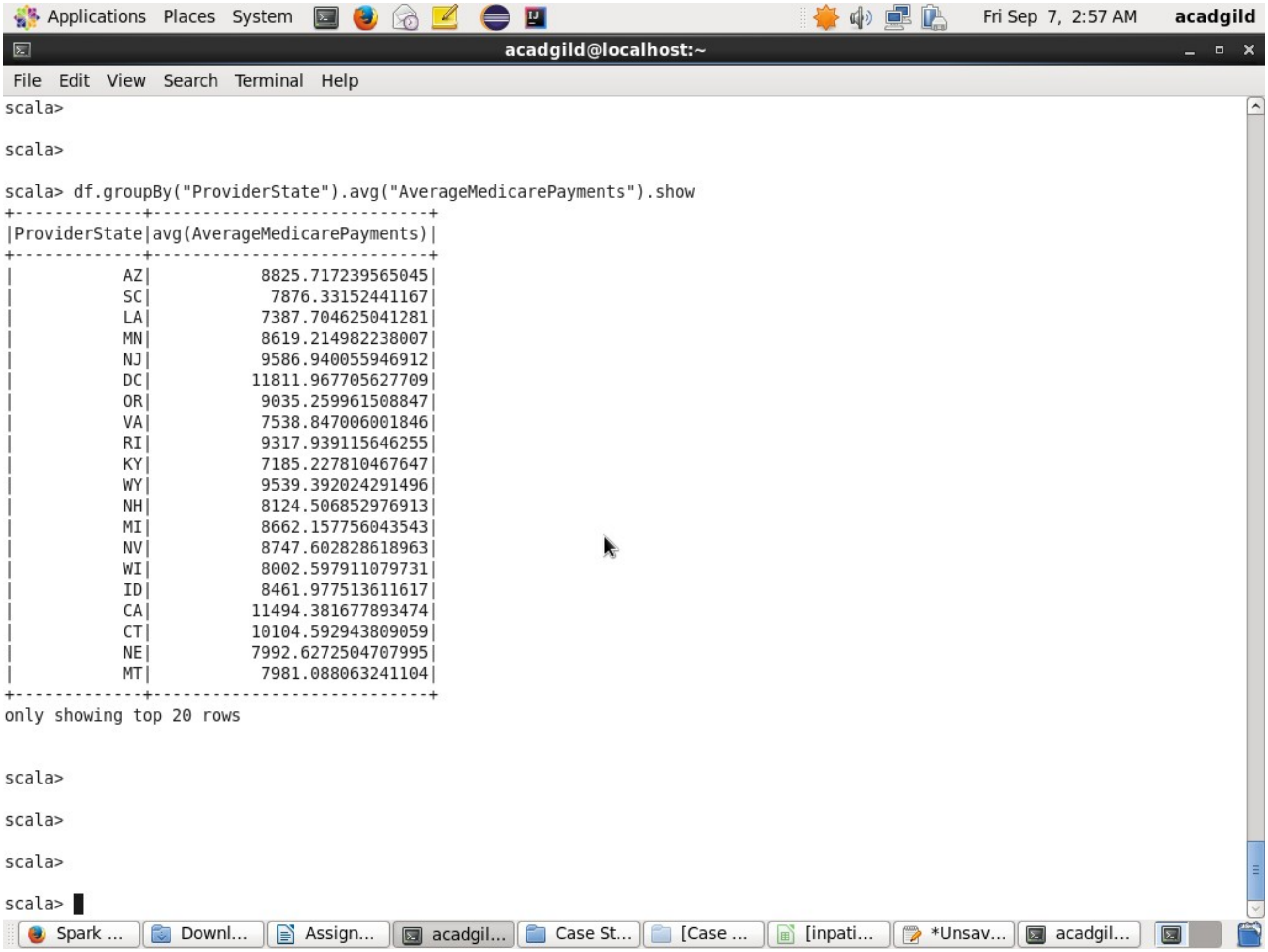
Query :

```
df.groupBy("ProviderState").avg("AverageMedicarePayments").show
```

- **Terminal Execution:**

```
scala> df.groupBy("ProviderState").avg("AverageMedicarePayments").show
+-------------+----------------------------+
|ProviderState|avg(AverageMedicarePayments)|
+-------------+----------------------------+
|           AZ|           8825.717239565045|
|           SC|            7876.33152441167|
|           LA|           7387.704625041281|
|           MN|           8619.214982238007|
|           NJ|           9586.940055946912|
|           DC|          11811.967705627709|
|           OR|           9035.259961508847|
|           VA|           7538.847006001846|
|           RI|           9317.939115646255|
|           KY|           7185.227810467647|
|           WY|           9539.392024291496|
|           NH|           8124.506852976913|
|           MI|           8662.157756043543|
|           NV|           8747.602828618963|
|           WI|           8002.597911079731|
|           ID|           8461.977513611617|
|           CA|          11494.381677893474|
|           CT|          10104.592943809059|
|           NE|          7992.6272504707995|
|           MT|           7981.088063241104|
+-------------+----------------------------+
only showing top 20 rows
```

Output :



```
scala>

scala>

scala> df.groupBy("ProviderState").avg("AverageMedicarePayments").show
+-------------+----------------------------+
|ProviderState|avg(AverageMedicarePayments)|
+-------------+----------------------------+
|           AZ|           8825.717239565045|
|           SC|            7876.33152441167|
|           LA|           7387.704625041281|
|           MN|           8619.214982238007|
|           NJ|           9586.940055946912|
|           DC|          11811.967705627709|
|           OR|           9035.259961508847|
|           VA|           7538.847006001846|
|           RI|           9317.939115646255|
|           KY|           7185.227810467647|
|           WY|           9539.392024291496|
|           NH|           8124.506852976913|
|           MI|           8662.157756043543|
|           NV|           8747.602828618963|
|           WI|           8002.597911079731|
|           ID|           8461.977513611617|
|           CA|          11494.381677893474|
|           CT|          10104.592943809059|
|           NE|          7992.6272504707995|
|           MT|           7981.088063241104|
+-------------+----------------------------+
only showing top 20 rows


scala>

scala>

scala>

scala> █
```

Objective 3 :

1. **Find out the total number of Discharges per state and for each disease**

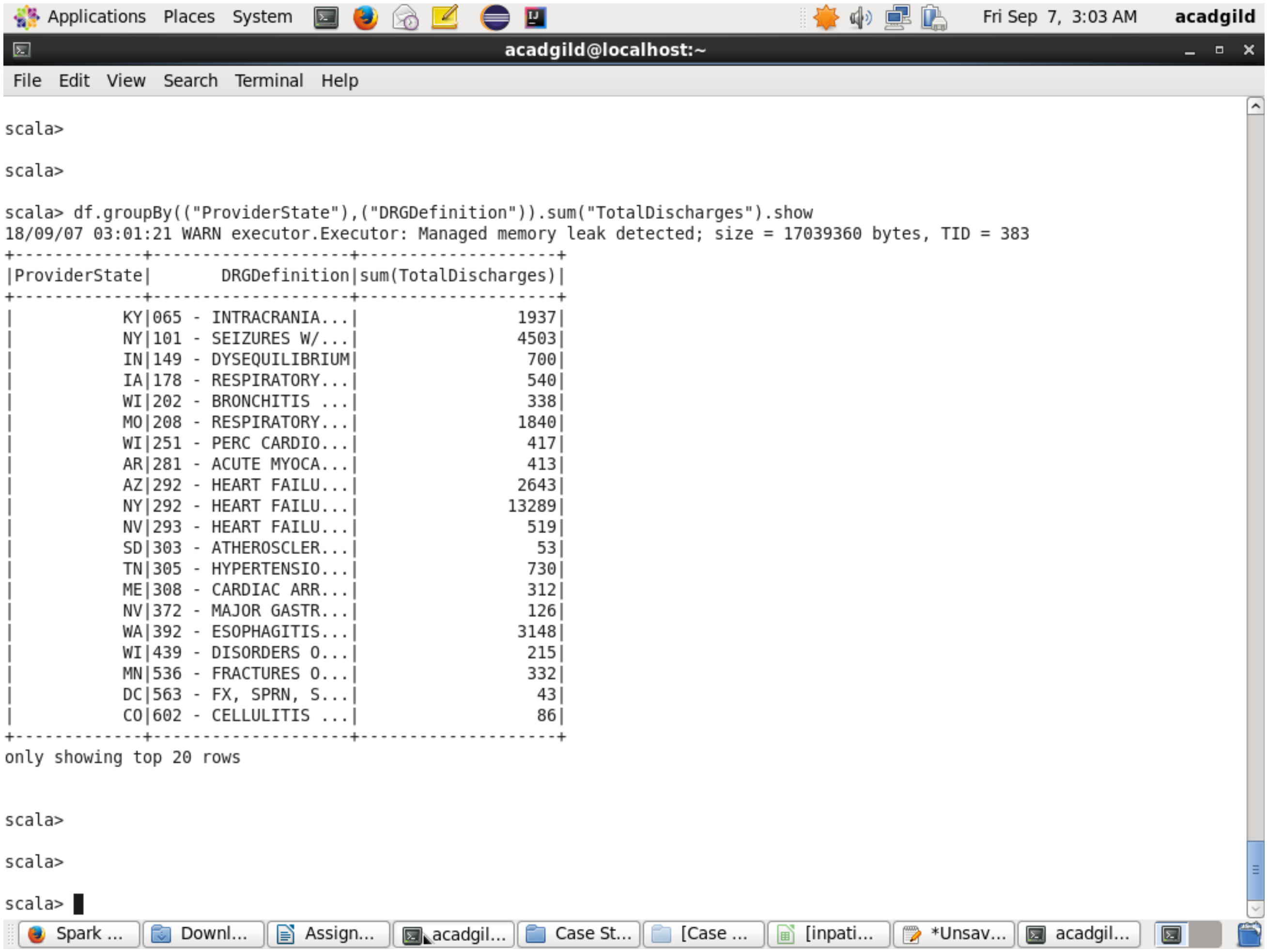
   Query :

   df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").show

Terminal Execution:

```
 scala> df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").show
18/09/07 03:01:21 WARN executor.Executor: Managed memory leak detected; size = 17039360
bytes, TID = 383
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           KY|065 - INTRACRANIA...|                1937|
|           NY|101 - SEIZURES W/...|                4503|
|           IN|149 - DYSEQUILIBRIUM|                 700|
|           IA|178 - RESPIRATORY...|                 540|
|           WI|202 - BRONCHITIS ...|                 338|
|           MO|208 - RESPIRATORY...|                1840|
|           WI|251 - PERC CARDIO...|                 417|
|           AR|281 - ACUTE MYOCA...|                 413|
|           AZ|292 - HEART FAILU...|                2643|
|           NY|292 - HEART FAILU...|               13289|
|           NV|293 - HEART FAILU...|                 519|
|           SD|303 - ATHEROSCLER...|                  53|
|           TN|305 - HYPERTENSIO...|                 730|
|           ME|308 - CARDIAC ARR...|                 312|
|           NV|372 - MAJOR GASTR...|                 126|
|           WA|392 - ESOPHAGITIS...|                3148|
|           WI|439 - DISORDERS O...|                 215|
|           MN|536 - FRACTURES O...|                 332|
|           DC|563 - FX, SPRN, S...|                  43|
|           CO|602 - CELLULITIS ...|                  86|
+-------------+--------------------+--------------------+
only showing top 20 rows
```

Output :



```
scala>

scala>

scala> df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").show
18/09/07 03:01:21 WARN executor.Executor: Managed memory leak detected; size = 17039360 bytes, TID = 383
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           KY|065 - INTRACRANIA...|                1937|
|           NY|101 - SEIZURES W/...|                4503|
|           IN|149 - DYSEQUILIBRIUM|                 700|
|           IA|178 - RESPIRATORY...|                 540|
|           WI|202 - BRONCHITIS ...|                 338|
|           MO|208 - RESPIRATORY...|                1840|
|           WI|251 - PERC CARDIO...|                 417|
|           AR|281 - ACUTE MYOCA...|                 413|
|           AZ|292 - HEART FAILU...|                2643|
|           NY|292 - HEART FAILU...|               13289|
|           NV|293 - HEART FAILU...|                 519|
|           SD|303 - ATHEROSCLER...|                  53|
|           TN|305 - HYPERTENSIO...|                 730|
|           ME|308 - CARDIAC ARR...|                 312|
|           NV|372 - MAJOR GASTR...|                 126|
|           WA|392 - ESOPHAGITIS...|                3148|
|           WI|439 - DISORDERS O...|                 215|
|           MN|536 - FRACTURES O...|                 332|
|           DC|563 - FX, SPRN, S...|                  43|
|           CO|602 - CELLULITIS ...|                  86|
+-------------+--------------------+--------------------+
only showing top 20 rows


scala>

scala>

scala> 
```

2. **Sort the output in descending order of totalDischarges**

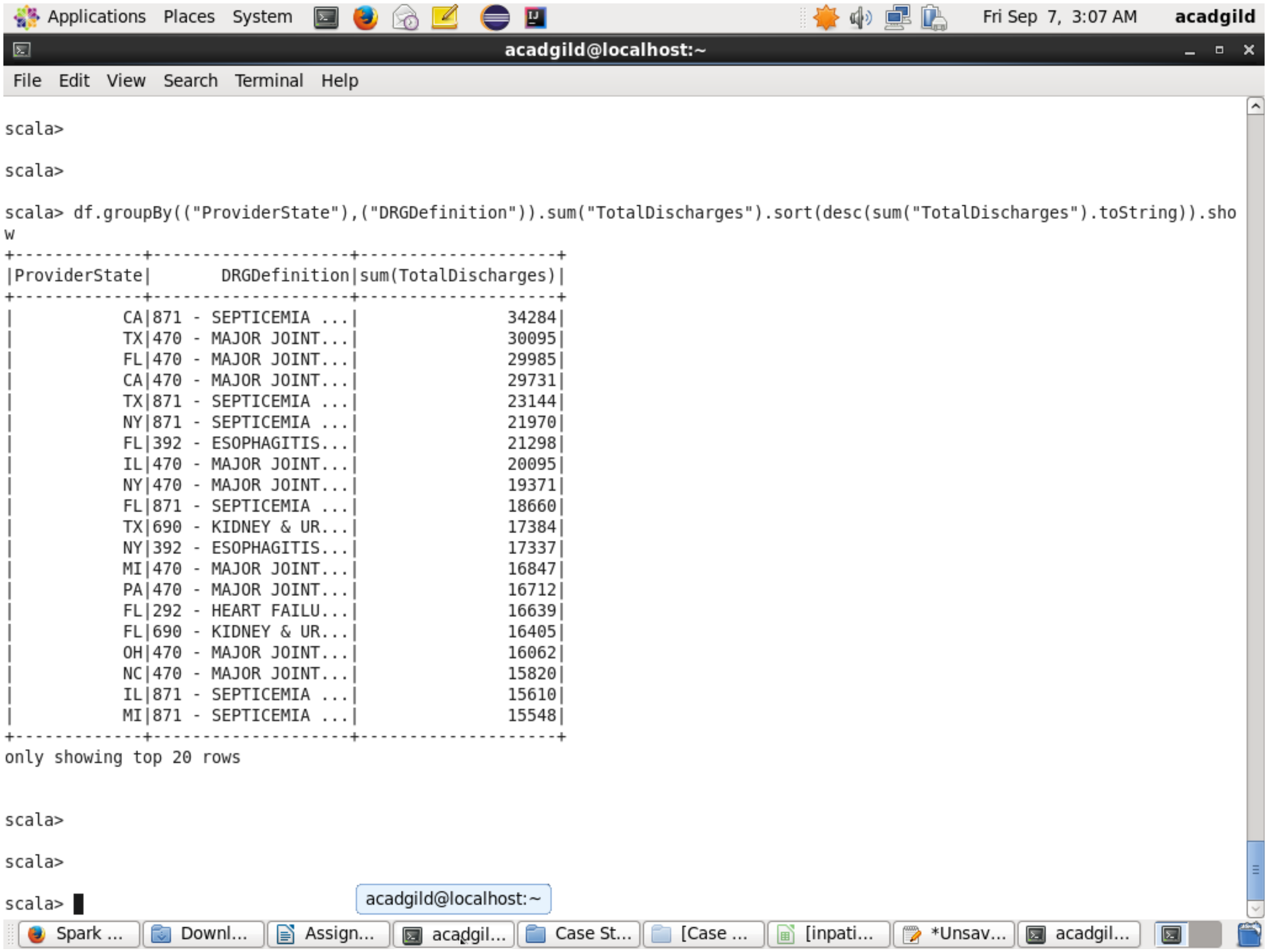
Query :

```
df.groupBy(("ProviderState"),
("DRGDefinition")).sum("TotalDischarges").sort(desc(sum("TotalDischarges").toString)).show
```

Terminal Execution:

```
scala> df.groupBy(("ProviderState"),
("DRGDefinition")).sum("TotalDischarges").sort(desc(sum("TotalDischarges").toString)).show
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           CA|871 - SEPTICEMIA ...|               34284|
|           TX|470 - MAJOR JOINT...|               30095|
|           FL|470 - MAJOR JOINT...|               29985|
|           CA|470 - MAJOR JOINT...|               29731|
|           TX|871 - SEPTICEMIA ...|               23144|
|           NY|871 - SEPTICEMIA ...|               21970|
|           FL|392 - ESOPHAGITIS...|               21298|
|           IL|470 - MAJOR JOINT...|               20095|
|           NY|470 - MAJOR JOINT...|               19371|
|           FL|871 - SEPTICEMIA ...|               18660|
|           TX|690 - KIDNEY & UR...|               17384|
|           NY|392 - ESOPHAGITIS...|               17337|
|           MI|470 - MAJOR JOINT...|               16847|
|           PA|470 - MAJOR JOINT...|               16712|
|           FL|292 - HEART FAILU...|               16639|
|           FL|690 - KIDNEY & UR...|               16405|
|           OH|470 - MAJOR JOINT...|               16062|
|           NC|470 - MAJOR JOINT...|               15820|
|           IL|871 - SEPTICEMIA ...|               15610|
|           MI|871 - SEPTICEMIA ...|               15548|
+-------------+--------------------+--------------------+
only showing top 20 rows
```

Output :



```
scala>

scala>

scala> df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").sort(desc(sum("TotalDischarges").toString)).sho
w
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           CA|871 - SEPTICEMIA ...|               34284|
|           TX|470 - MAJOR JOINT...|               30095|
|           FL|470 - MAJOR JOINT...|               29985|
|           CA|470 - MAJOR JOINT...|               29731|
|           TX|871 - SEPTICEMIA ...|               23144|
|           NY|871 - SEPTICEMIA ...|               21970|
|           FL|392 - ESOPHAGITIS...|               21298|
|           IL|470 - MAJOR JOINT...|               20095|
|           NY|470 - MAJOR JOINT...|               19371|
|           FL|871 - SEPTICEMIA ...|               18660|
|           TX|690 - KIDNEY & UR...|               17384|
|           NY|392 - ESOPHAGITIS...|               17337|
|           MI|470 - MAJOR JOINT...|               16847|
|           PA|470 - MAJOR JOINT...|               16712|
|           FL|292 - HEART FAILU...|               16639|
|           FL|690 - KIDNEY & UR...|               16405|
|           OH|470 - MAJOR JOINT...|               16062|
|           NC|470 - MAJOR JOINT...|               15820|
|           IL|871 - SEPTICEMIA ...|               15610|
|           MI|871 - SEPTICEMIA ...|               15548|
+-------------+--------------------+--------------------+
only showing top 20 rows


scala>

scala>

scala> 
```