**SESSION 22 ASSIGNMENT 1**

* **Importing APIs**

scala> import org.apache.spark.sql.Row

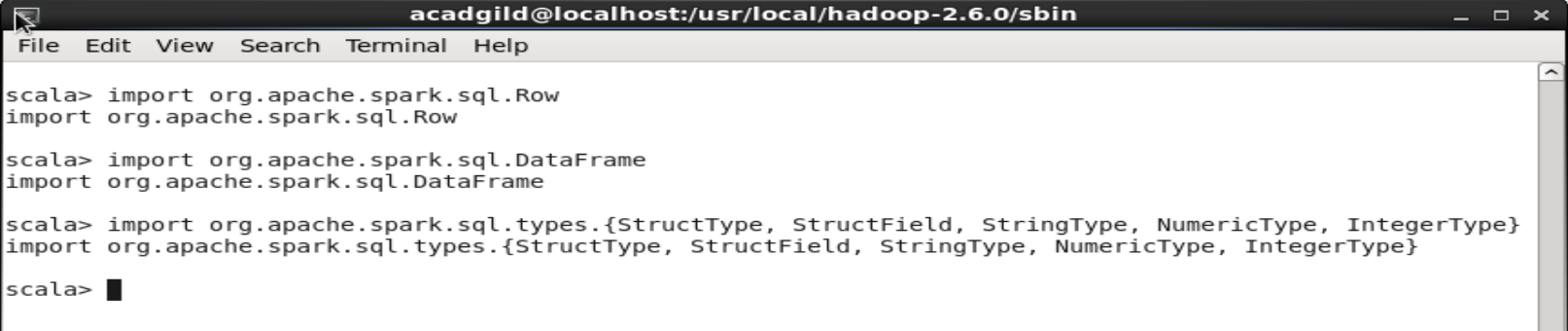
import org.apache.spark.sql.Row

scala> import org.apache.spark.sql.DataFrame

import org.apache.spark.sql.DataFrame

scala> import org.apache.spark.sql.types.{StructType, StructField, StringType, NumericType, IntegerType}

import org.apache.spark.sql.types.{StructType, StructField, StringType, NumericType, IntegerType}

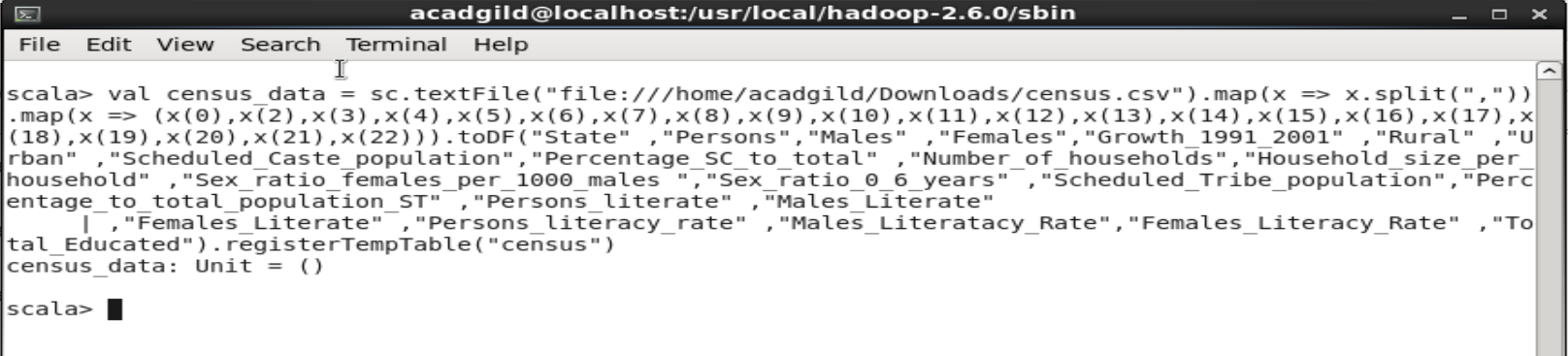


* **Loading census data by splitting with comma and creating data frame**

scala> val census\_data = sc.textFile("file:///home/acadgild/Downloads/census.csv").map(x => x.split(",")).map(x => (x(0),x(2),x(3),x(4),x(5),x(6),x(7),x(8),x(9),x(10),x(11),x(12),x(13),x(14),x(15),x(16),x(17),x(18),x(19),x(20),x(21),x(22))).toDF("State" ,"Persons","Males" ,"Females","Growth\_1991\_2001" ,"Rural" ,"Urban" ,"Scheduled\_Caste\_population","Percentage\_SC\_to\_total" ,"Number\_of\_households","Household\_size\_per\_household" ,"Sex\_ratio\_females\_per\_1000\_males ","Sex\_ratio\_0\_6\_years" ,"Scheduled\_Tribe\_population","Percentage\_to\_total\_population\_ST" ,"Persons\_literate" ,"Males\_Literate"

| ,"Females\_Literate" ,"Persons\_literacy\_rate" ,"Males\_Literatacy\_Rate","Females\_Literacy\_Rate" ,"Total\_Educated").registerTempTable("census")

census\_data: Unit = ()



* **State wise population and order by state**

**// spark sql to selecting state and sum of persons and ordering it by population in a descending order**

scala> val population = sqlContext.sql("select state,sum(persons) as total\_population from census group by state order by total\_population desc").show

+-----------+----------------+

| state|total\_population|

+-----------+----------------+

| UP| 1.66197921E8|

|Maharashtra| 9.6878627E7|

| Bihar| 8.2998509E7|

| WB| 8.0176197E7|

| Andhra| 7.1308587E7|

| TN| 6.2405679E7|

| MP| 6.0348023E7|

| Rajasthan| 5.6507188E7|

| Karnataka| 5.2850562E7|

| Gujarat| 5.0671017E7|

| Orrisa| 3.5664657E7|

| Kerala| 3.1841374E7|

| Jharkhand| 2.6945829E7|

| Assam| 2.6655528E7|

| Punjab| 2.4358999E7|

| Haryana| 2.1144564E7|

| CG| 2.0833803E7|

| Delhi| 1.3850507E7|

| JK| 1.01437E7|

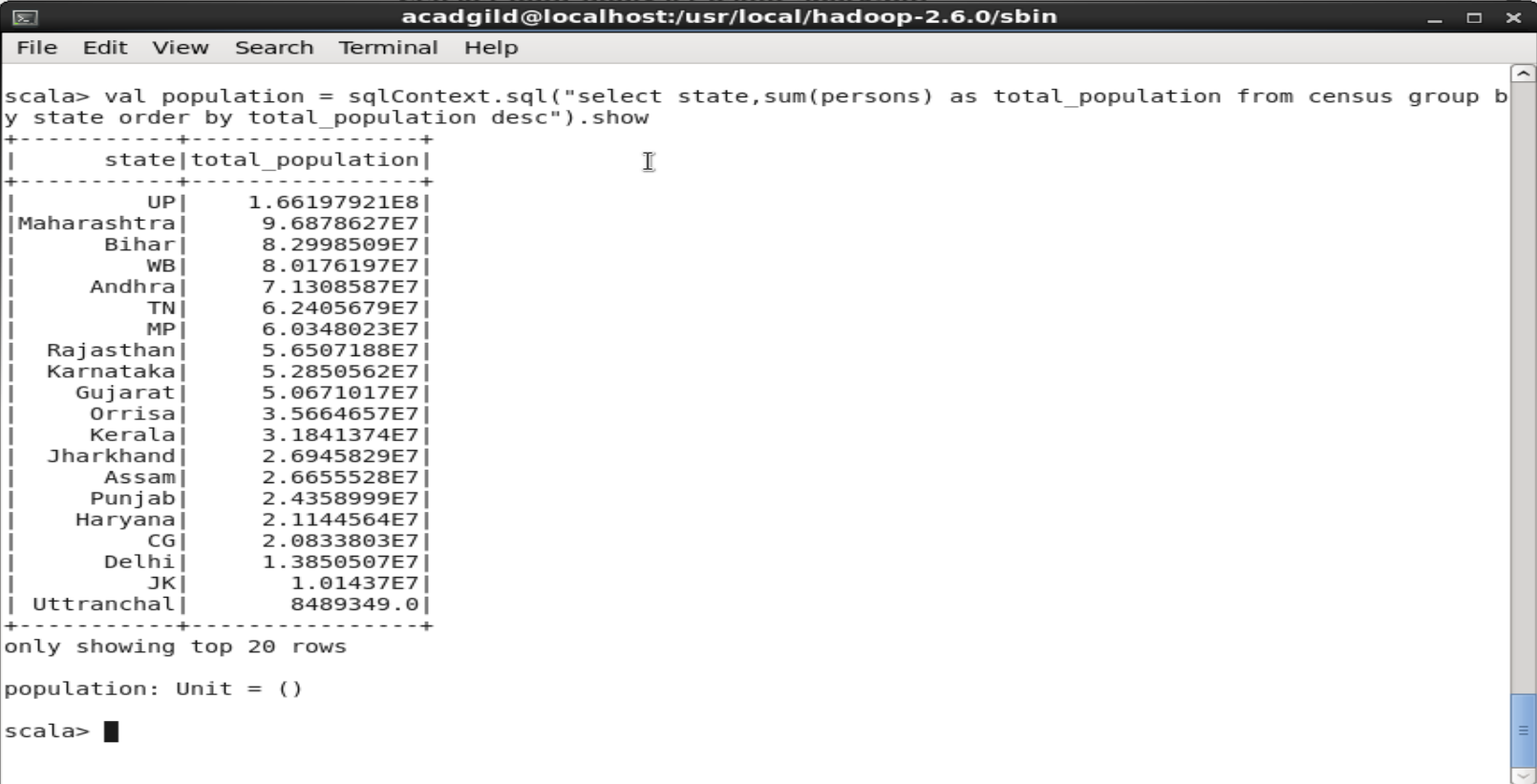
| Uttranchal| 8489349.0|

+-----------+----------------+

only showing top 20 rows

population: Unit = ()

scala>



* **Growth Rate of Each State Between 1991-2001**

**// spark sql to selecting state and average of growth rate between 1991 to 2001 year and grouping it by state**

scala> val growth\_rate = sqlContext.sql("select state,avg(Growth\_1991\_2001) as total\_growth from census group by state").show

+----------------+------------------+

| state| total\_growth|

+----------------+------------------+

| Maharashtra|19.607142857142865|

| TN|10.127666666666668|

| Gujarat| 20.8248|

| Orrisa|15.551379310344826|

| Sikkim|31.834999999999997|

| AN| 18.665|

| Chandigarh| 40.33|

| Bihar|28.605945945945955|

| HP| 17.53083333333333|

| UP| 25.70228571428572|

|ArunachalPradesh| 25.46999999999999|

| Tripura|15.405000000000001|

| D\_N\_H| 59.2|

| Uttranchal|17.092307692307692|

| Haryana|27.816842105263152|

| CG|17.506249999999998|

| WB|18.424999999999997|

| Manipur|29.240000000000002|

| JK|28.785714285714285|

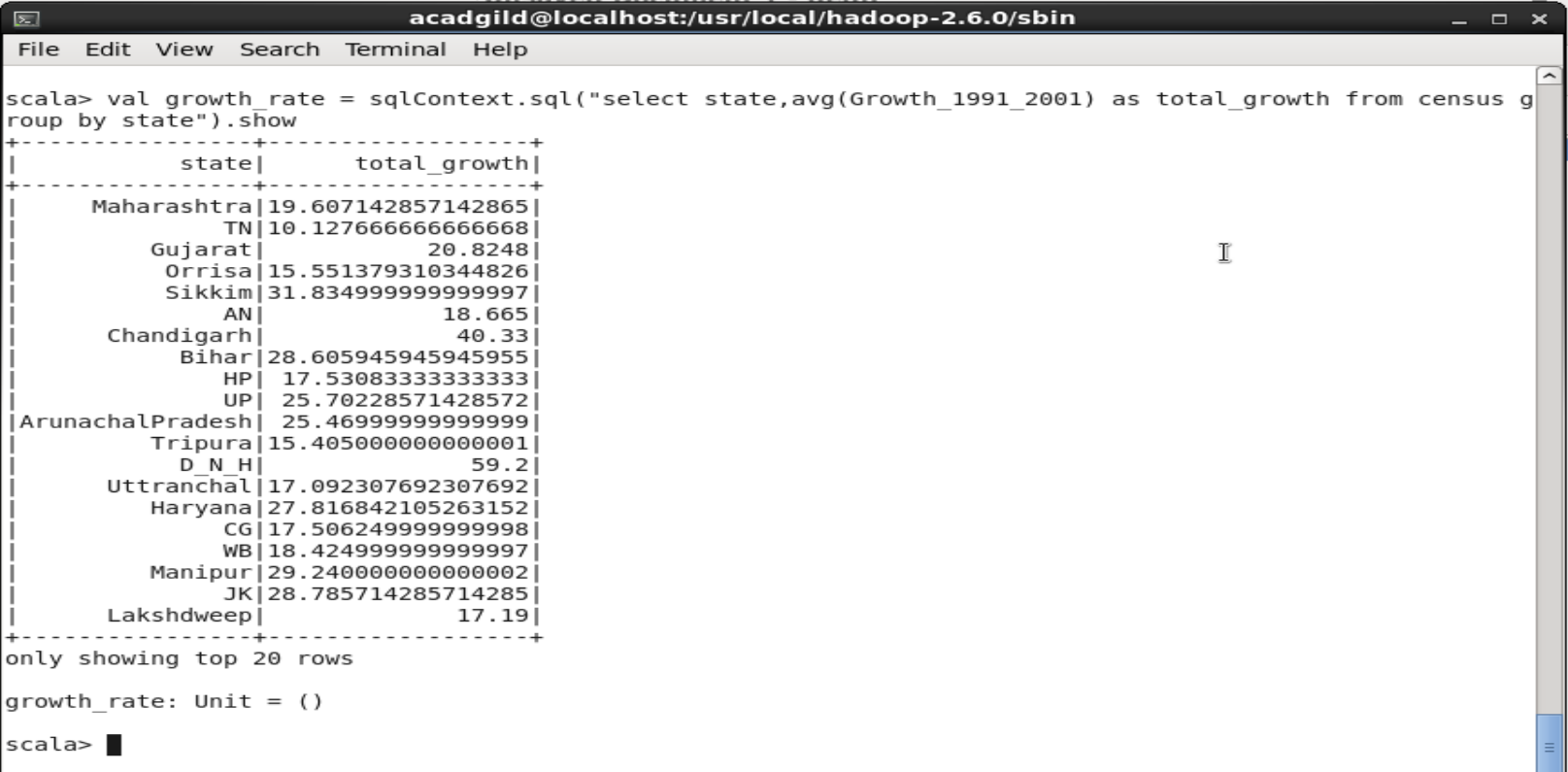
| Lakshdweep| 17.19|

+----------------+------------------+

only showing top 20 rows

growth\_rate: Unit = ()

scala>



* **Literacy rate of each state**

**// spark sql to selecting state and average of people Literacy rate and grouping it by state**

scala> val literacy = sqlContext.sql("select state,avg(Persons\_literacy\_rate) from census group by state").show

+----------------+------------------+

| state| \_c1|

+----------------+------------------+

| Maharashtra| 74.55342857142857|

| TN| 72.94266666666665|

| Gujarat| 67.07480000000001|

| Orrisa| 59.97965517241381|

| Sikkim| 66.9975|

| AN| 77.41999999999999|

| Chandigarh| 81.94|

| Bihar| 46.42135135135135|

| HP| 75.50833333333333|

| UP| 56.01057142857144|

|ArunachalPradesh|53.166923076923084|

| Tripura| 70.27000000000001|

| D\_N\_H| 57.63|

| Uttranchal| 72.01769230769231|

| Haryana| 68.24473684210527|

| CG| 63.02312499999999|

| WB| 66.07|

| Manipur| 68.6125|

| JK|54.867142857142845|

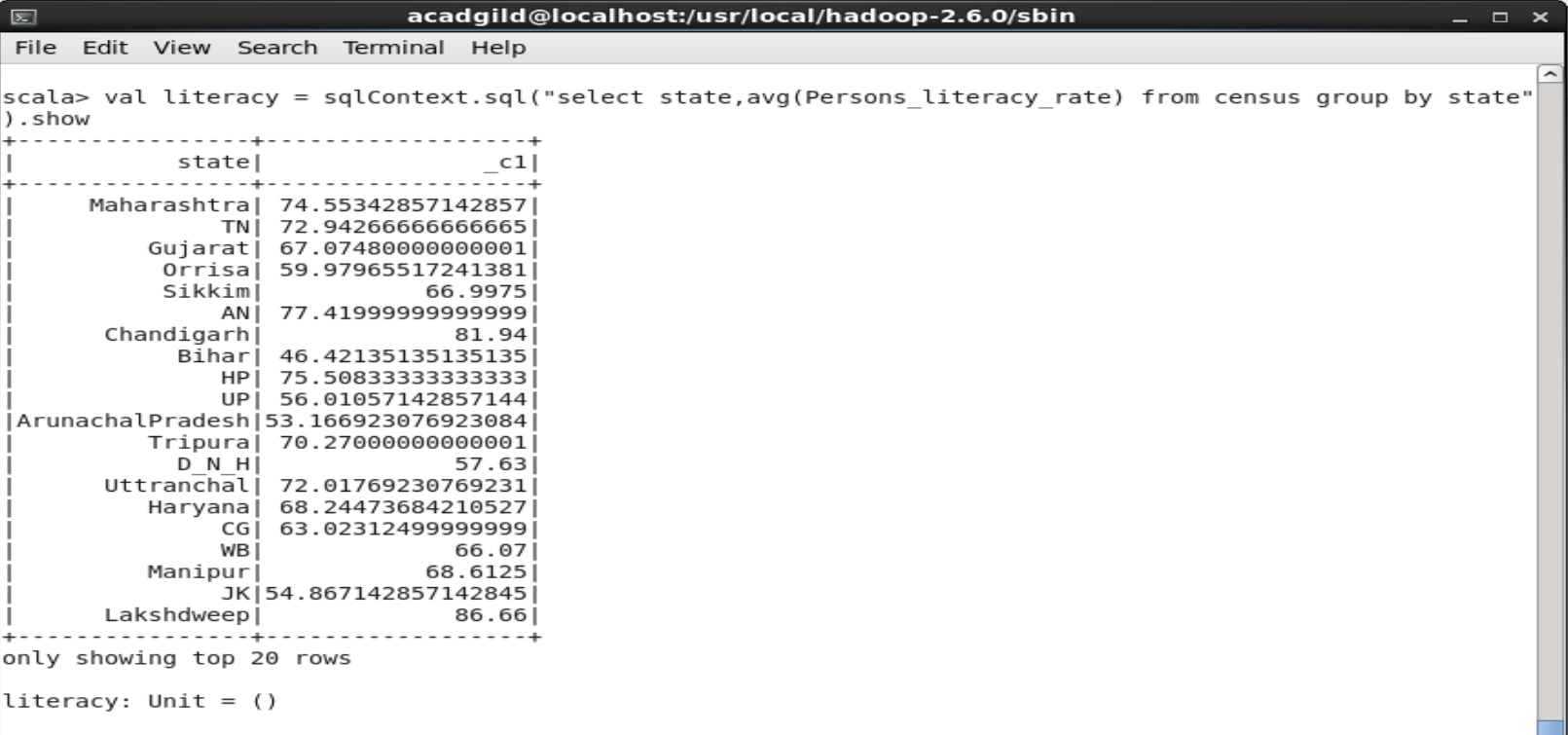
| Lakshdweep| 86.66|

+----------------+------------------+

only showing top 20 rows

literacy: Unit = ()

scala>



* **States with More Female Population**

**// spark sql to selecting state and sum of Female and grouping it by state**

scala> val female\_pop = sqlContext.sql("select state, sum(Males)-sum(Females) from census group by state").show

+----------------+---------+

| state| \_c1|

+----------------+---------+

| Maharashtra|3922565.0|

| TN| 396139.0|

| Gujarat|2100137.0|

| Orrisa| 482015.0|

| Sikkim| 36117.0|

| AN| 29792.0|

| Chandigarh| 113241.0|

| Bihar|3489081.0|

| HP| 97980.0|

| UP|8932817.0|

|ArunachalPradesh| 61914.0|

| Tripura| 85247.0|

| D\_N\_H| 22842.0|

| Uttranchal| 162499.0|

| Haryana|1583342.0|

| CG| 114633.0|

| WB|2755773.0|

| Manipur| 20533.0|

| JK| 578152.0|

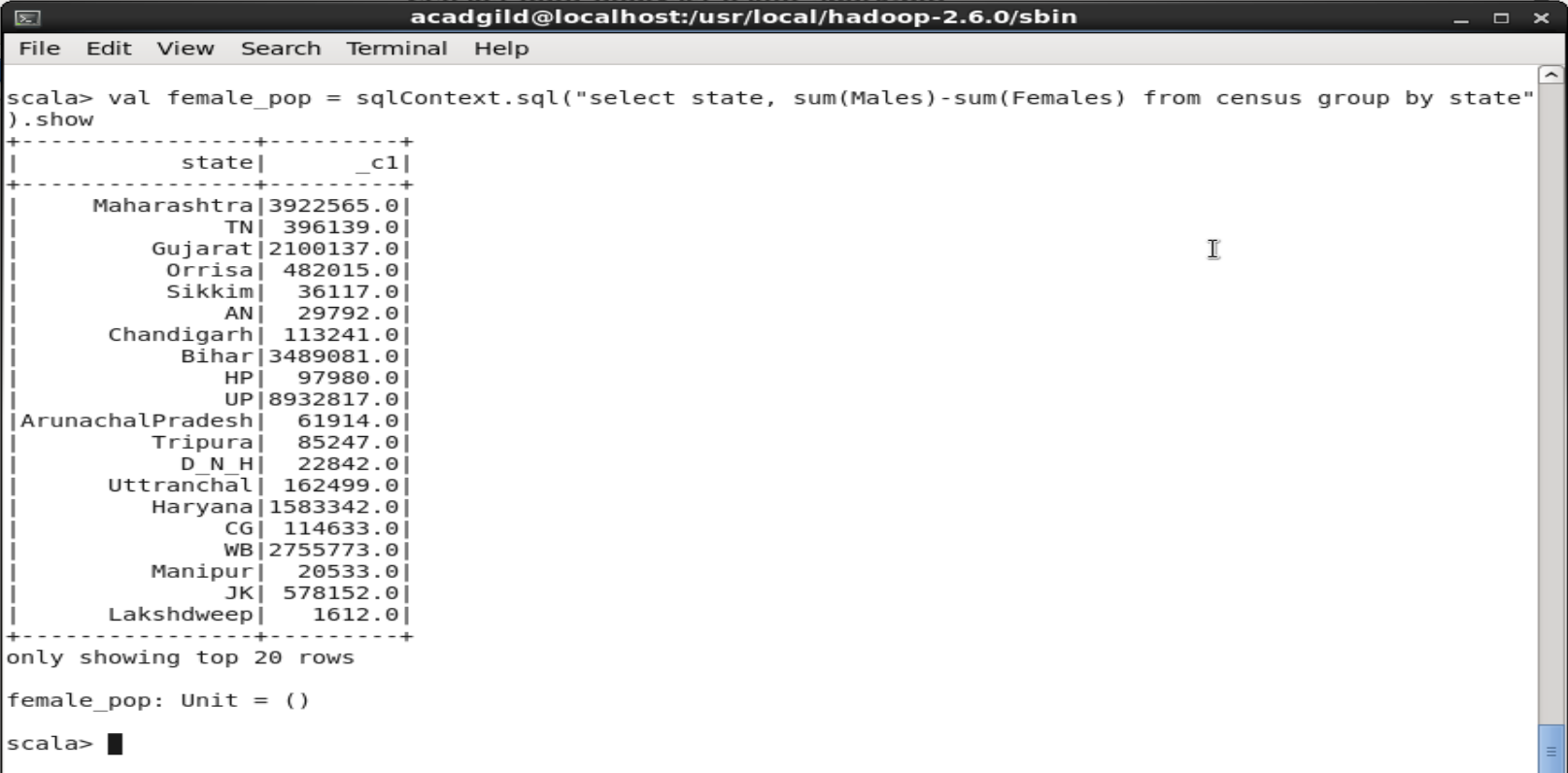
| Lakshdweep| 1612.0|

+----------------+---------+

only showing top 20 rows

female\_pop: Unit = ()

scala>



* **Percentage of Population in Every State**

**// spark sql to selecting state and percentage of population and grouping it by state**

scala> val percenet\_pop = sqlContext.sql("select state, (sum(persons) \* 100.0) / SUM(sum(persons)) over() as percent\_pop\_by\_state from census group by state").show

18/03/11 11:15:29 WARN Window: No Partition Defined for Window operation! Moving all data to a single partition, this can cause serious performance degradation.

+----------------+--------------------+

| state|percent\_pop\_by\_state|

+----------------+--------------------+

| Maharashtra| 9.475494209385522|

| TN| 6.103767861999858|

| Gujarat| 4.956025317815201|

| Orrisa| 3.488284891601744|

| Sikkim| 0.05289949576432755|

| AN| 0.03483447606726582|

| Chandigarh| 0.08808921009243792|

| Bihar| 8.117909138174843|

| HP| 0.5944665819347776|

| UP| 16.25546817511578|

|ArunachalPradesh| 0.10738993468694186|

| Tripura| 0.31290729895613395|

| D\_N\_H| 0.02156566193106157|

| Uttranchal| 0.8303253233652121|

| Haryana| 2.0681052152192616|

| CG| 2.0377103371415317|

| WB| 7.841864753141607|

| Manipur| 0.19662075848548596|

| JK| 0.9921339059826262|

| Lakshdweep|0.005932048601382...|

+----------------+--------------------+

only showing top 20 rows

percenet\_pop: Unit = ()

scala>

