**Session 22: DEPLOYING A SPARK APPLICATION**

**Assignment 22.1**

**Problem Statement**

Implement the below blog at your end and send the complete documentation.

[https://docs.google.com/document/d/1csLBlMiEXs\_hXWV2Z8VpBlrj\_R6RoDQLlZUnA0uBTCk/edit](https://docs.google.com/document/d/1csLBlMiEXs_hXWV2Z8VpBlrj_R6RoDQLlZUnA0uBTCk/edit%20)

**Census data analysis**

You can download the dataset from the below link,

<https://drive.google.com/open?id=0ByJLBTmJojjzWllGZFJFaXFVbU0>

Due to the limitation of **22** elements for a map function, we are taking only **22** columns from the data set. Here is the total dataset description

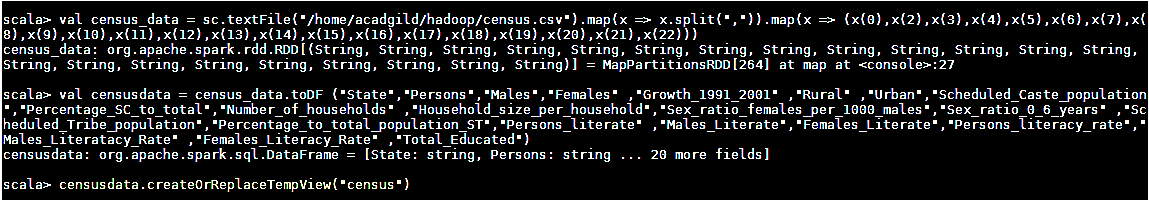
**State String,DistrictString,PersonsString,Malesint,Females int,Growth\_1991\_2001 int,Ruralint,Urbanint,Scheduled\_Caste\_populationint,Percentage\_SC\_to\_totalint,Number\_of\_householdsint,Household\_size\_per\_household int,Sex\_ratio\_females\_per\_1000\_males int ,Sex\_ratio\_0\_6\_years int,Scheduled\_Tribe\_populationint,Percentage\_to\_total\_population\_STint,Persons\_literateint,Males\_Literateint,Females\_Literateint,Persons\_literacy\_rateint,Males\_Literatacy\_Rateint,Females\_Literacy\_Rateint,Total\_Educatedint,Data\_without\_levelint,Below\_Primaryint,Primaryint,Middleint,Matric\_Higher\_Secondary\_Diplomaint,Graduate\_and\_Above int,X0\_4\_years int,X5\_14\_years int,X15\_59\_years int,X60\_years\_and\_above\_Incl\_ANS int,Total\_workersint,Main\_workersint,Marginal\_workersint,Non\_workers int,SC\_1\_Name String,SC\_1\_Population int,SC\_2\_Name String,SC\_2\_Population int,SC\_3\_Name String,SC\_3\_Population int,Religeon\_1\_Name String,Religeon\_1\_Population int,Religeon\_2\_Name String,Religeon\_2\_Population int,Religeon\_3\_Name String,Religeon\_3\_Population int,ST\_1\_Name String,ST\_1\_Population int,ST\_2\_Name String,ST\_2\_Population int,ST\_3\_Name String,ST\_3\_Population int,Imp\_Town\_1\_Name String,Imp\_Town\_1\_Population int,Imp\_Town\_2\_Name String,Imp\_Town\_2\_Population int,Imp\_Town\_3\_Name String,Imp\_Town\_3\_Population int,Total\_Inhabited\_Villagesint,Drinking\_water\_facilitiesint,Safe\_Drinking\_waterint,Electricity\_Power\_Supplyint,Electricity\_domesticint,Electricity\_Agricultureint,Primary\_schoolint,Middle\_schoolsint,Secondary\_Sr\_Secondary\_schoolsint,Collegeint,Medical\_facilityint,Primary\_Health\_Centreint,Primary\_Health\_Sub\_Centreint,Post\_telegraph\_and\_telephone\_facilityint,Bus\_servicesint,Paved\_approach\_roadint,Mud\_approach\_roadint,Permanent\_Houseint,Semi\_permanent\_Houseint,Temporary\_Houseint**

**Here is what we are taking**

**"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***"*

Creating a RDD,

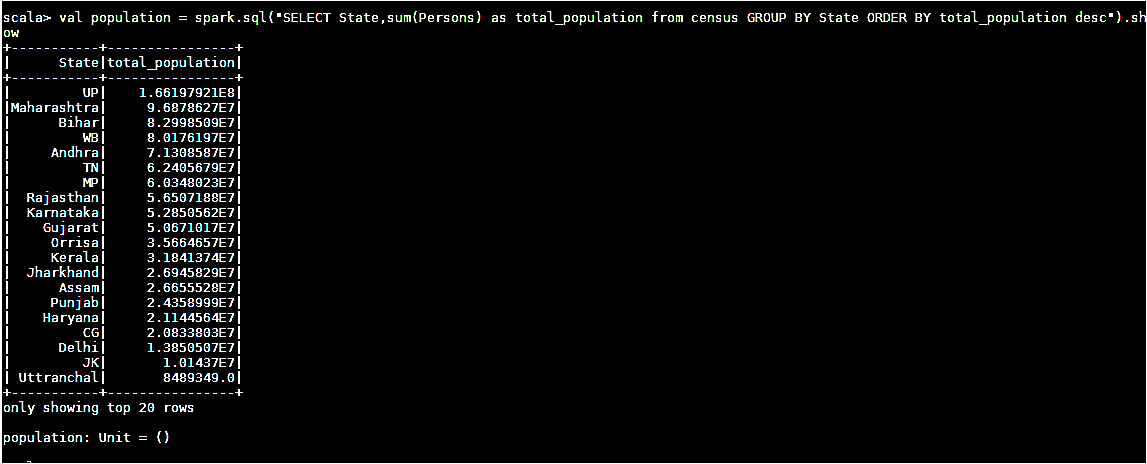
* ***valcensus\_data = sc.textFile("/home/acadgild/hadoop/census.csv").map(x =>x.split(",")).map(x => (x(0),x(2),x(3),x(4),x(5),x(6),x(7),x(8),x(9),x(10),x(11),x(12),x(13),x(14),x(15),x(16),x(17),x(18),x(19),x(20),x(21),x(22)))***
* ***valcensusdata = census\_data.toDF ("State","Persons","Males","Females" ,"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")***
* ***censusdata.createOrReplaceTempView("census")***



**1. Find out the state wise population and order by state**

***val population = spark.sql("SELECT State,sum(Persons) as total\_population from census GROUP BY State ORDER BY total\_populationdesc").show***

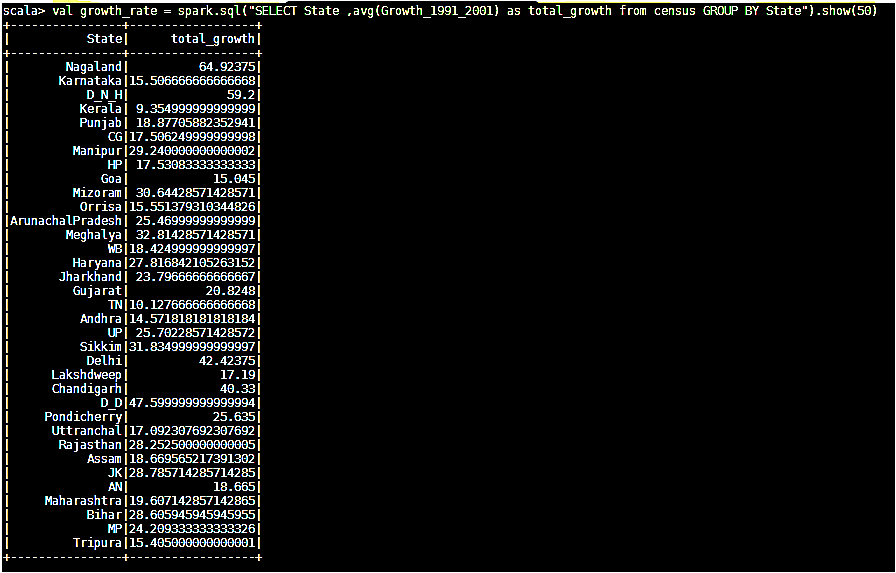
**Output**



**2. Find out the Growth Rate of Each State Between 1991-2001**

***valgrowth\_rate = spark.sql("SELECT State ,avg(Growth\_1991\_2001) as total\_growth from census GROUP BY State ASC").show(50)***

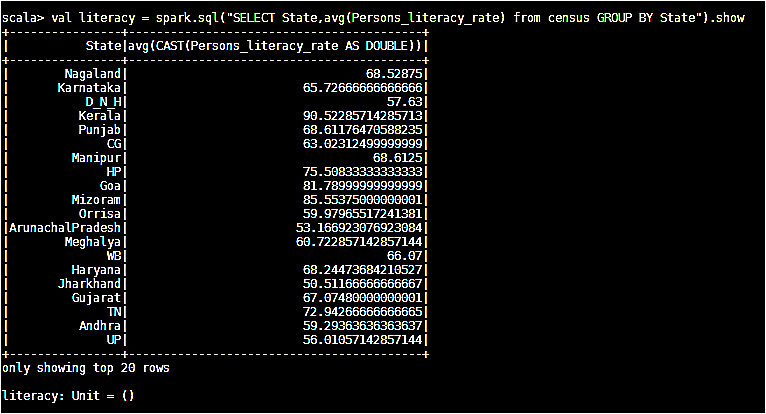
**Output**



**3. Find the literacy rate of each state**

***val literacy = spark.sql("SELECT State,avg(Persons\_literacy\_rate) from census GROUP BY State").show***

**Output**



**4. Find out the States with More Female Population**

***valfemale\_pop = spark.sql("SELECT State, sum(Males)-sum(Females) from census GROUP BY State").show***

**Output**



**5. Find out the Percentage of Population in Every State**

***valpercenet\_pop = spark.sql("SELECT State, (sum(Persons) \* 100.0) / SUM(sum(Persons)) over() as percent\_pop\_by\_state from census GROUP BY State").show(35)***

**Output**

