Name : Omkar B Auti



“VIDYALANKAR SCHOOL OF INFORMATION TECHNOLOGY, WADALA”

AFFILIATED TO

UNIVERSITY OF MUMBAI

INSTITUTE OF DISTANCE AND OPEN LEARNING (IDOL)

CERTIFICATE

This is to certify that, ' **Omkar Auti** ' of SYMCA. Semester - III with Application ID **129457** has completed the practical of ‘Big Data Analytics and Visualization’ in this college during the academic year 2022 — 2023.

Subject In-Charge Coordinator -In-Charge Prof. Geeta Sahu

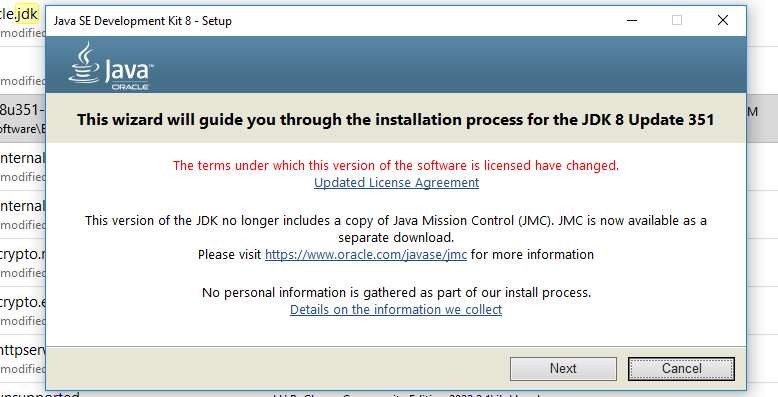
Examined By:

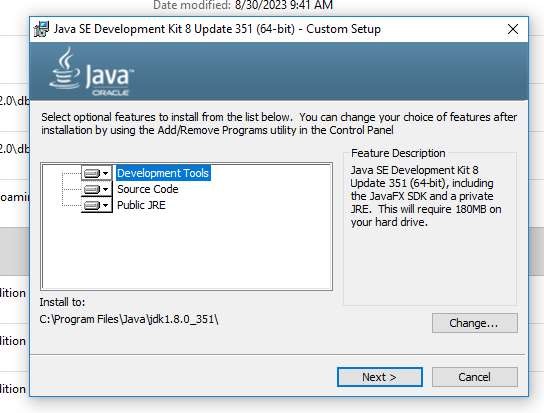
INDEX

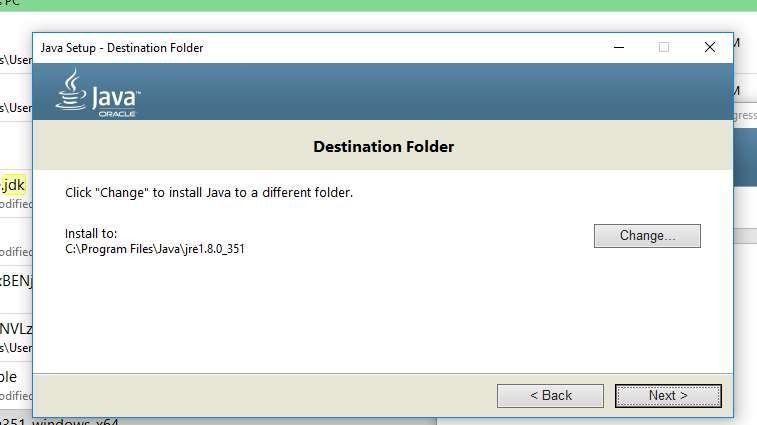
|  |  |  |  |
| --- | --- | --- | --- |
| **Practical No** | **Title** | **Date** | **Signature** |
| 1 | Set up and Configuration Hadoop |  |  |
| 2 | Map Reduce Programming |  |  |
| 3 | Mongo DB: Installation and CRUD Operations |  |  |
| 4 | Hive Database Operations |  |  |
| 5 | Pig Shell Operations |  |  |
| 6 | Visualization with Excel |  |  |
| 7 | Visualization with Tableau |  |  |

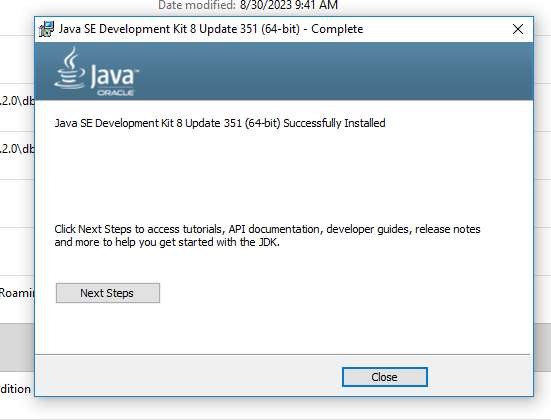
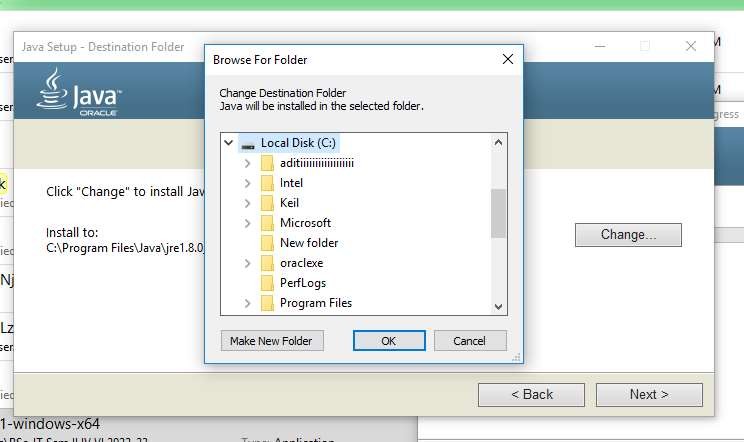
### Hadoop Installation

Step 1: First install the JDK

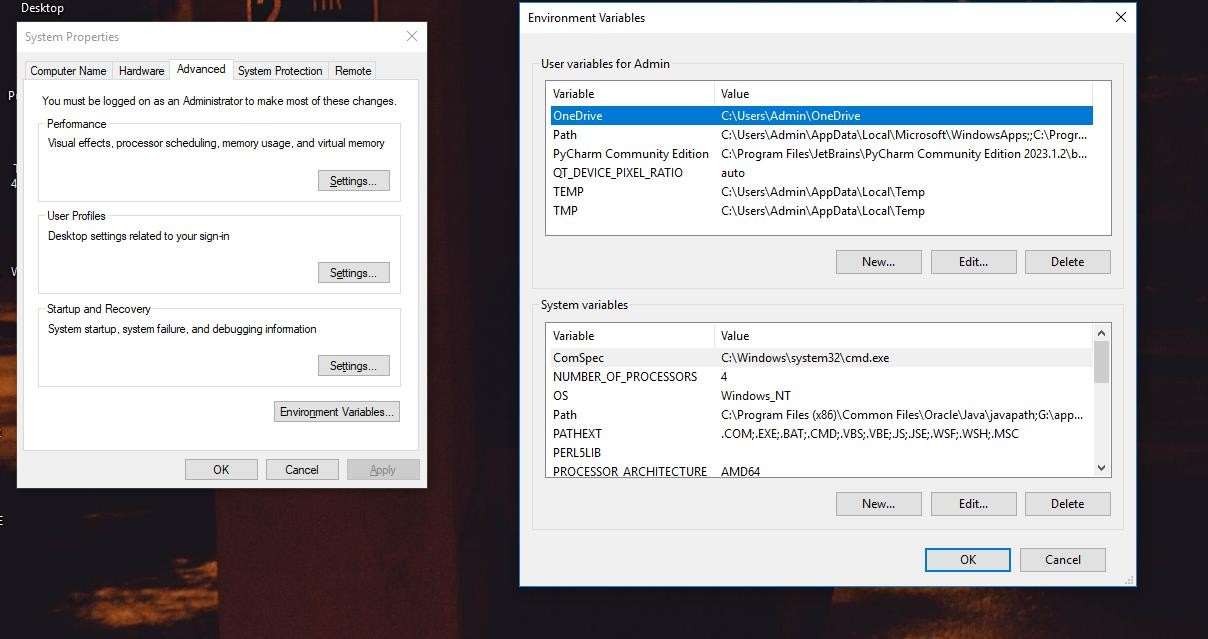


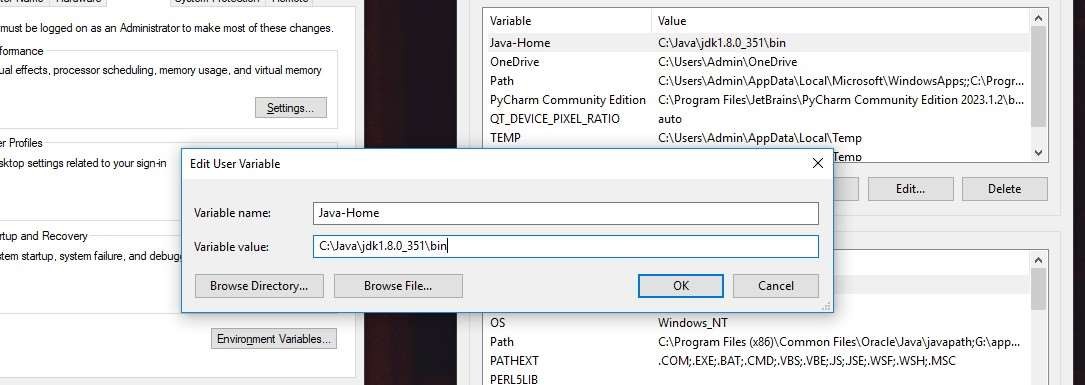






Step 2: Set Environment variable for java





Step 3: Create a path in system variable



## Step 4: Hadoop installation process

Download the Hadoop file form browser using internet After download extract it and paste the file in c drive

Now go to Hadoop file

Inside Hadoop file go to the etc folder Inside the etc folder edit five file Which will be edit only in note++ Following are the files –

## core-site.xml

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

## mapred-site.xml

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>

## yarn-site.xml

<configuration>

<!-- Site specific YARN configuration properties -->

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.auxservices.mapreduce.shuffle.clas s</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

## </configuration> hdfs-site.xml

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>C:\Users\Rasika\Desktop\MCA\SEM 3\big data lab\Software\hadoop-3.1.0\hadoop- 3.1.0\data\datanode</value>

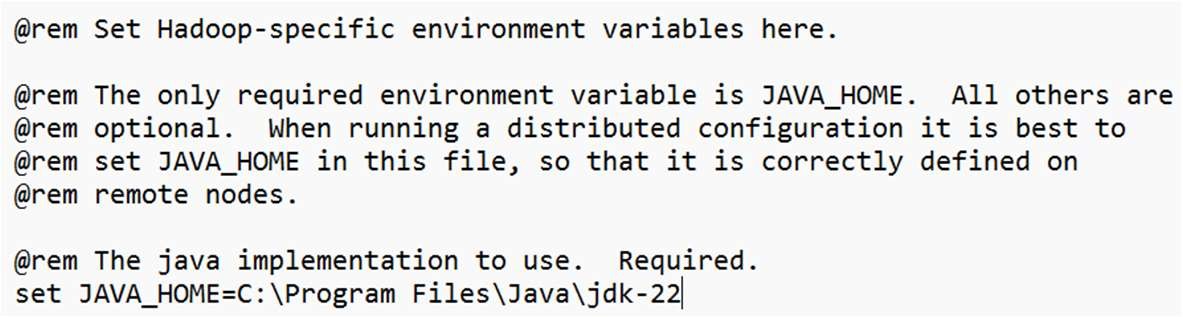
</property> <property> <name>dfs.datanode.data.dir</name>

<value>C:\Users\Rasika\Desktop\MCA\SEM 3\big data lab\Software\hadoop-3.1.0\hadoop- 3.1.0\data\namenode</value>

</property>

</configuration>

## hadoop-env.cmd

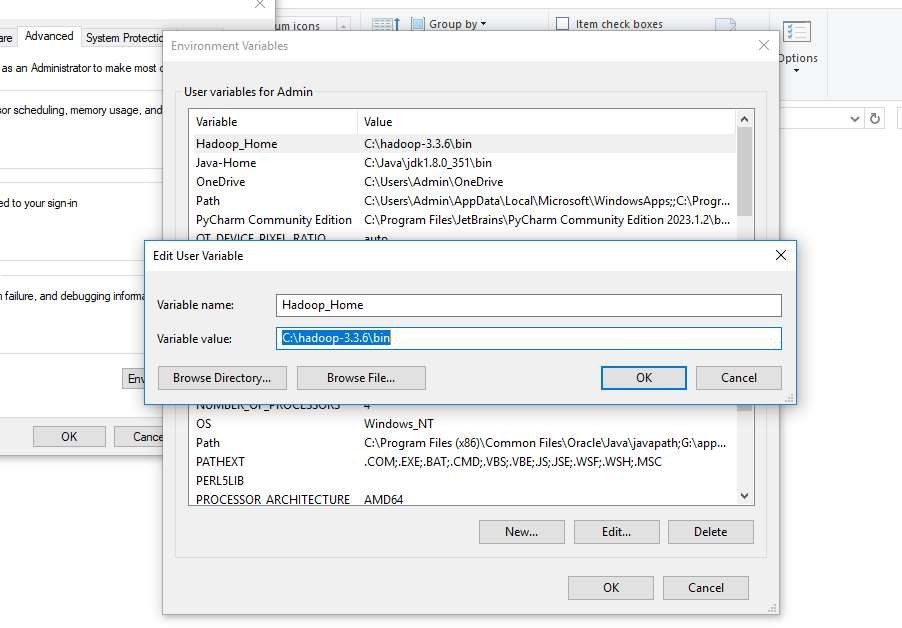


All above file opened

Before that we have create folder “Data” inside the Hadoop folder

Inside the data folder we have to create two folder Namenode and datanode

After that we have to create a Hadoop new environment variable and then give the path here as well



After all configuration just check format using command prompt Use this Hadoop command “ hdfs namenode –format”



Hadoop successfully installed

To check whether the Hadoop properly function or not run the following step by step First go to the Hadoop folder in c drive in cmd

Inside the Hadoop folder go to the sbin folder after the excute the hdfs commands 1)start-dfs 2)start-yarn.cmd 3)jps

### MapReduce in Hadoop – Word count program

public class WordCount {

public static void main(String [] args) throws Exception

{

Configuration c=new Configuration(); String[] files=new GenericOptionsParser(c,args).getRemainingArgs();Path input=new Path(files[0]); Path output=new Path(files[1]); Job j=new Job(c,"wordcount"); j.setJarByClass(WordCount.class);

j.setMapperClass(MapForWordCount.class); j.setReducerClass(ReduceForWordCount.class); j.setOutputKeyClass(Text.class); j.setOutputValueClass(IntWritable.class); FileInputFormat.addInputPath(j, input); FileOutputFormat.setOutputPath(j, output); System.exit(j.waitForCompletion(true)?0:1);

}

public static class MapForWordCount extends Mapper<LongWritable, Text, Text, IntWritable>{

public void map(LongWritable key, Text value, Context con) throws IOException, InterruptedException

{

String line = value.toString(); String[] words=line.split(",");

for(String word: words )

{

Text outputKey = new Text(word.toUpperCase().trim());IntWritable outputValue = new IntWritable(1); con.write(outputKey, outputValue);

}

}

}

public static class ReduceForWordCount extends Reducer<Text,IntWritable, Text, IntWritabl{ void reduce(Text word, Iterable<IntWritable> values,Context con) throws IOException, InterruptedException

{

int sum = 0;

for(IntWritable value : values)

{

sum += value.get();

}

con.write(word, new IntWritable(sum));

}

}

}

The above program consists of three classes:

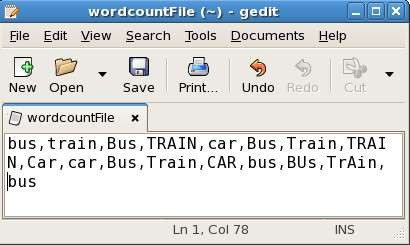
Driver class (Public, void, static, or main; this is the entrypoint). The Map class which **extends** the public class

Mapper<KEYIN,VALUEIN,KEYOUT,VALUEOUT> andimplements the Map function. The Reduce class which extends the public class Reducer<KEYIN,VALUEIN,KEYOUT,VALUEOUT> andimplements the Reduce function. Make a jar file

Right Click on Project> Export> Select export destination as **JarFile** > next> Finish



Take a text file and move it into HDFS format:

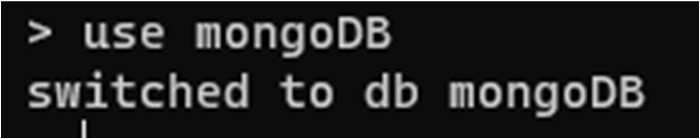


To move this into Hadoop directly, open the terminal and enter the following commands:

[training@localhost ~]$ hadoop fs -ls MRDir1

### MongoDB Installation and CRUD Operations

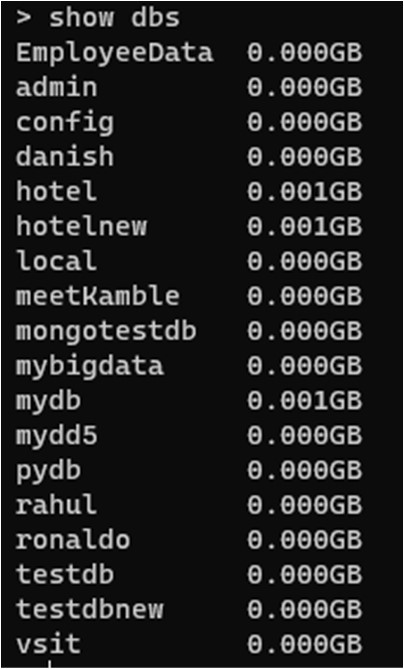
To create database:



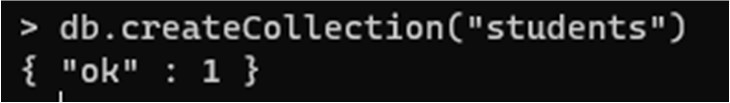
To check your currently selected database, use the command

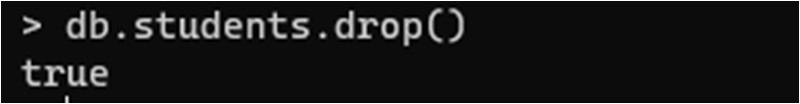
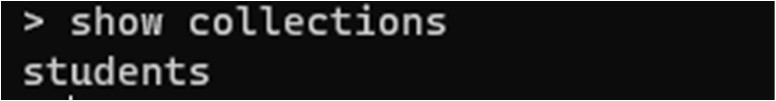


If you want to check your databases list, use the command

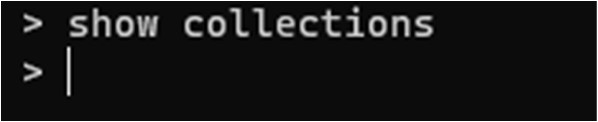


To create collection



You can check the created collection by using the command show collections To drop a collection from the database.

To verify collection is dropped



To drop a database



### CRUD:

In MongoDB, you don't need to create collection. MongoDB creates collection automatically, when you insert some document.

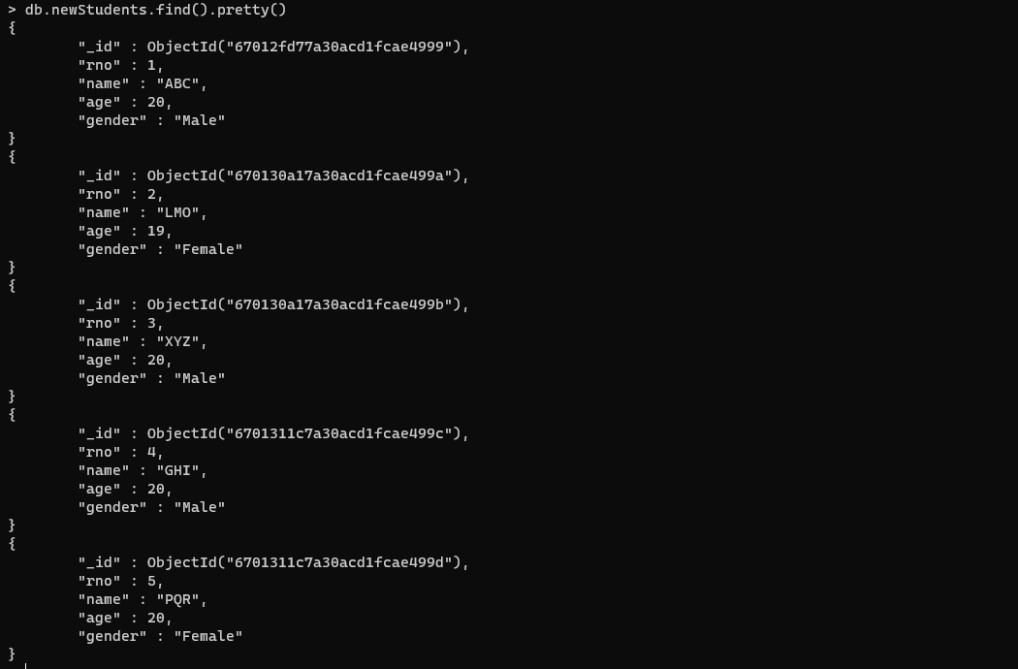
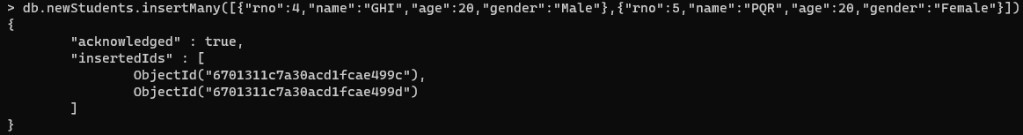
To insert & show element in db



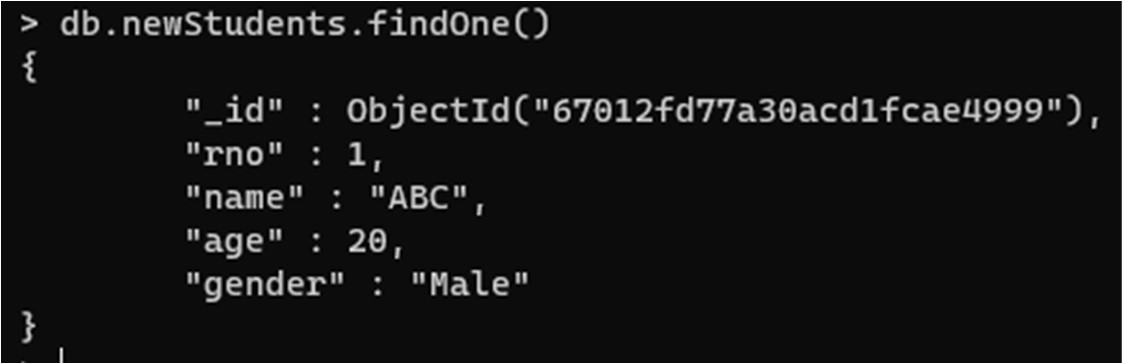
To insert multiple elements at a same time



Or



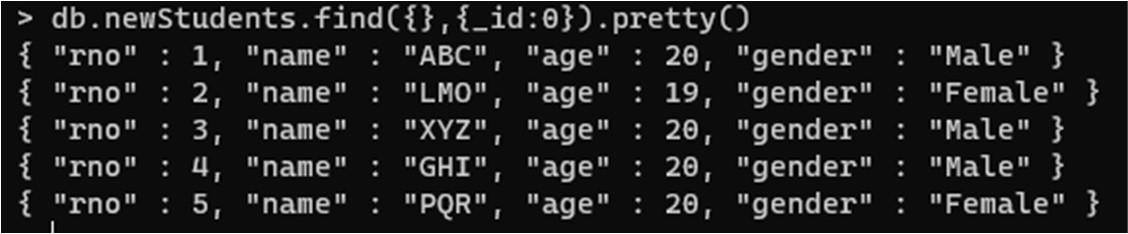
To get 1st data



To get particular data

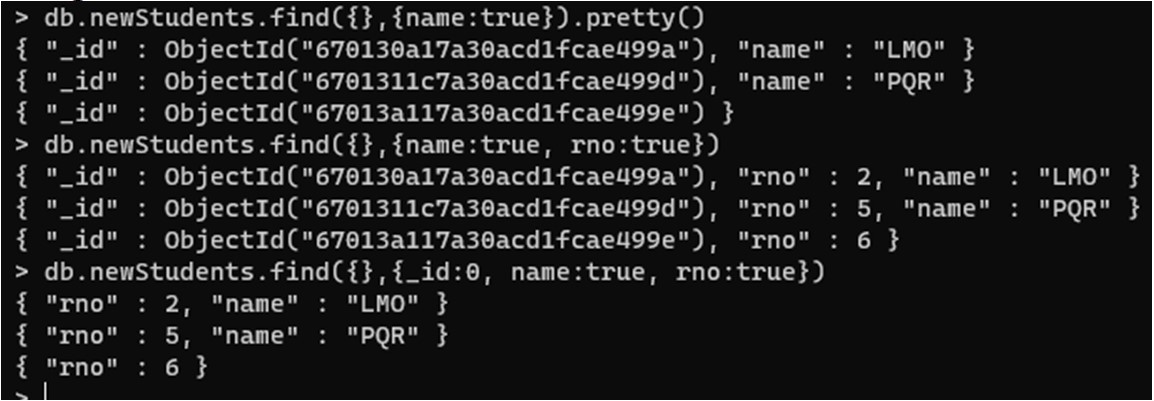


To get data based on condition

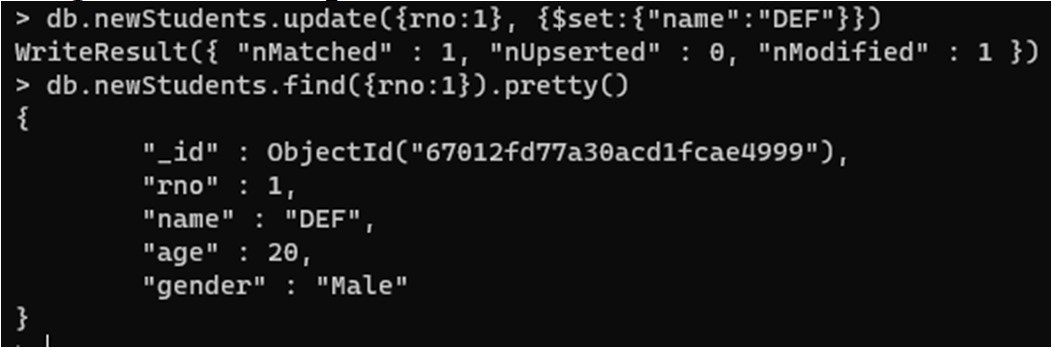


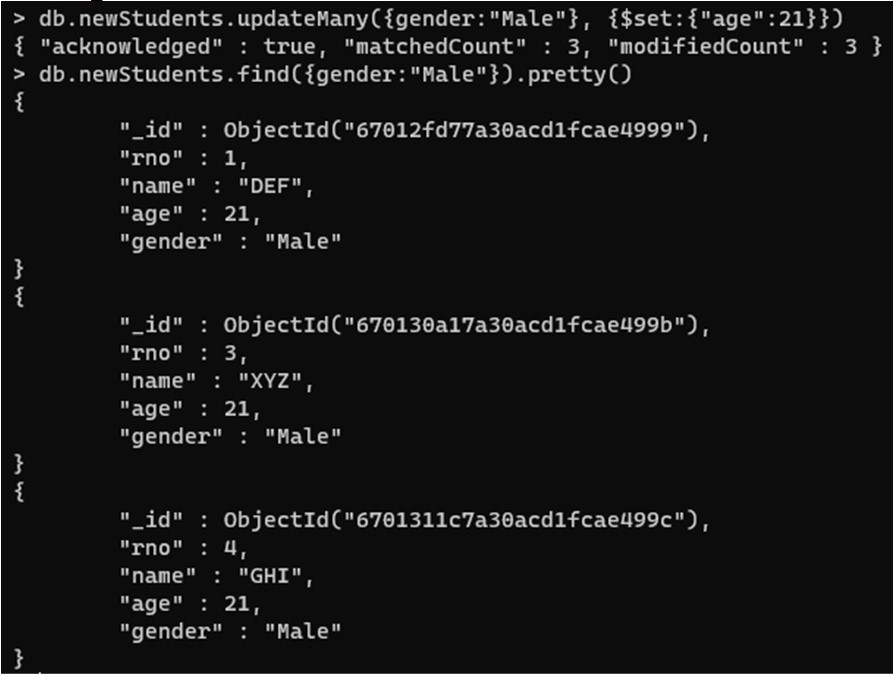
\

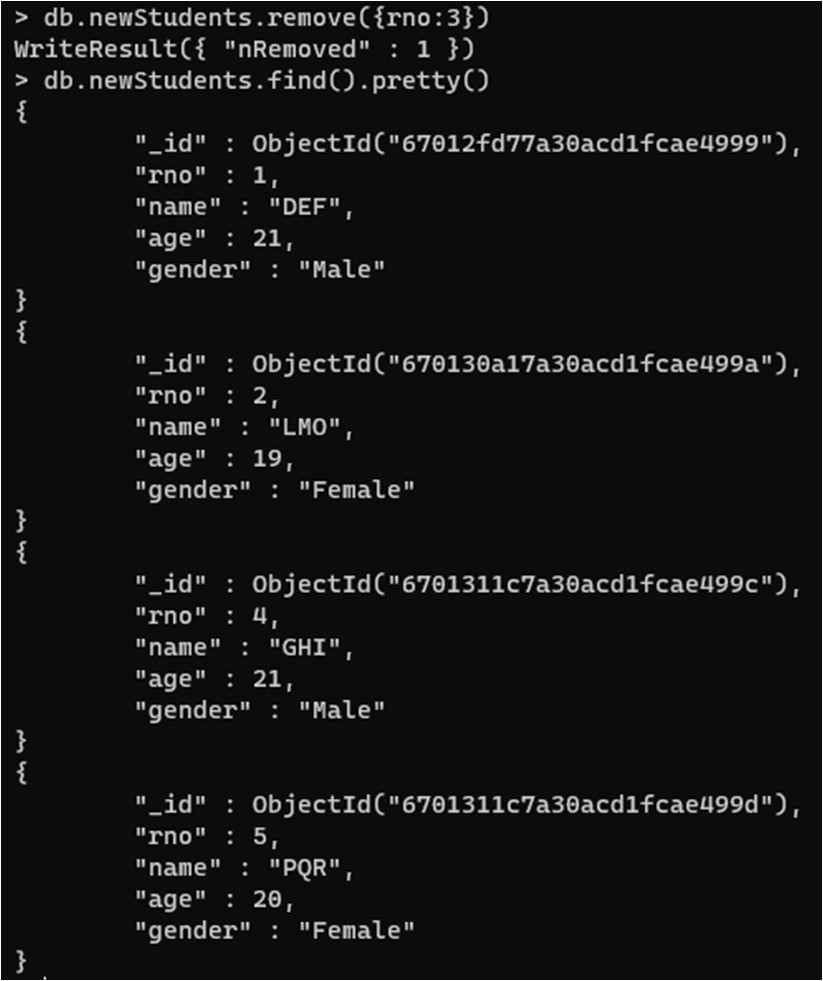
To get all data without id



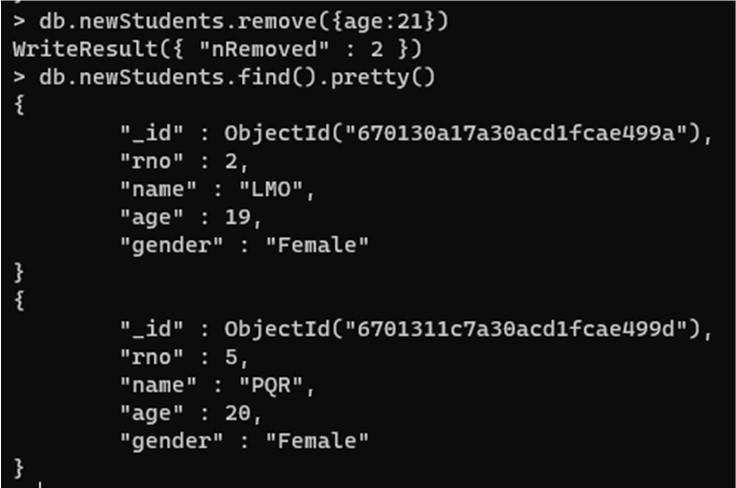
To get data with particular columns



To update record from db



Remove data from database

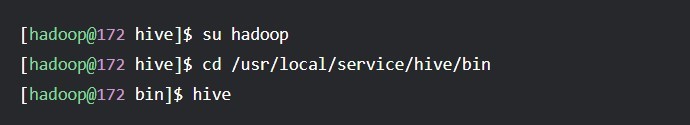


# Hive operations

Here we are explaining the operators available in Hive. There are types of operators in Hive: Relational Operators Arithmetic

Operators Logical Operators Complex Operators

Log in to a master node of the EMR cluster, switch to the Hadoop user, go to the Hive directory, and connect to Hive:



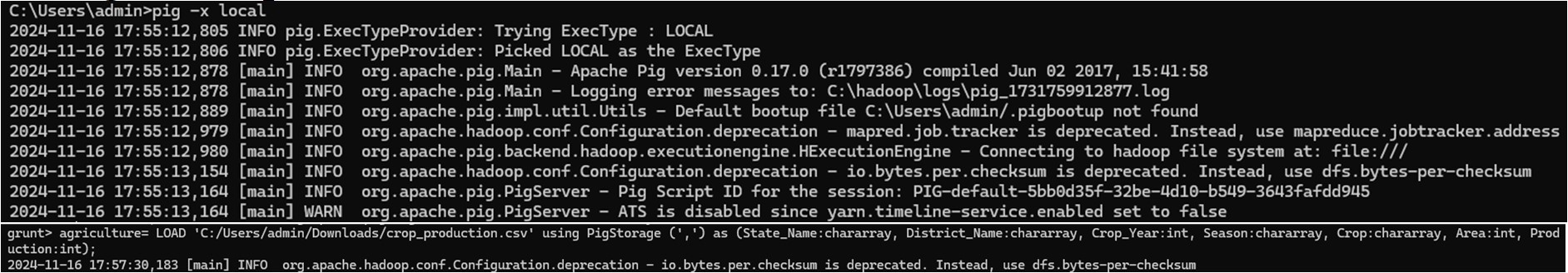
You can use the -h parameter to get basic information on Hive commands. You can also use the Beeline mode to connect to a database. To do so, you also need to log in to a master node in EMR, switch to the Hadoop user, and go to the Hive directory. In the conf/hive-site.xml configuration file, get the connection port $port and host address $host of Hive server 2

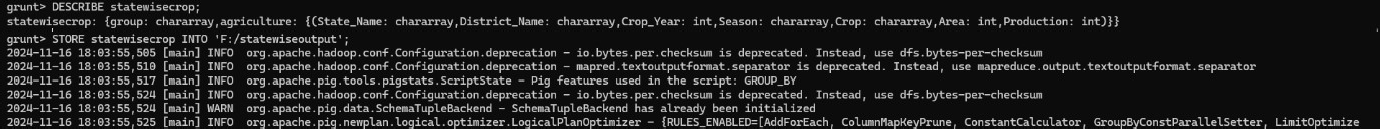
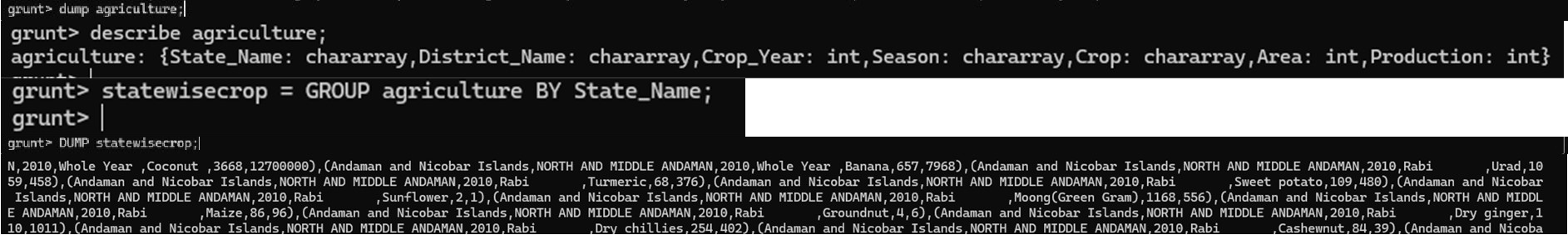


In the bin directory, run the following statement to connect to Hive



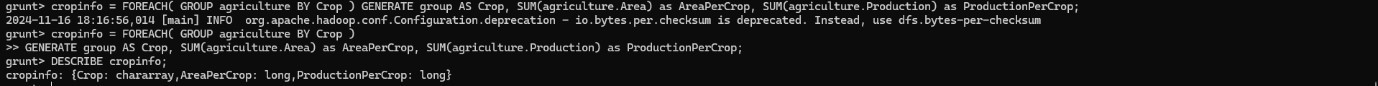
# Pig Shell Operations





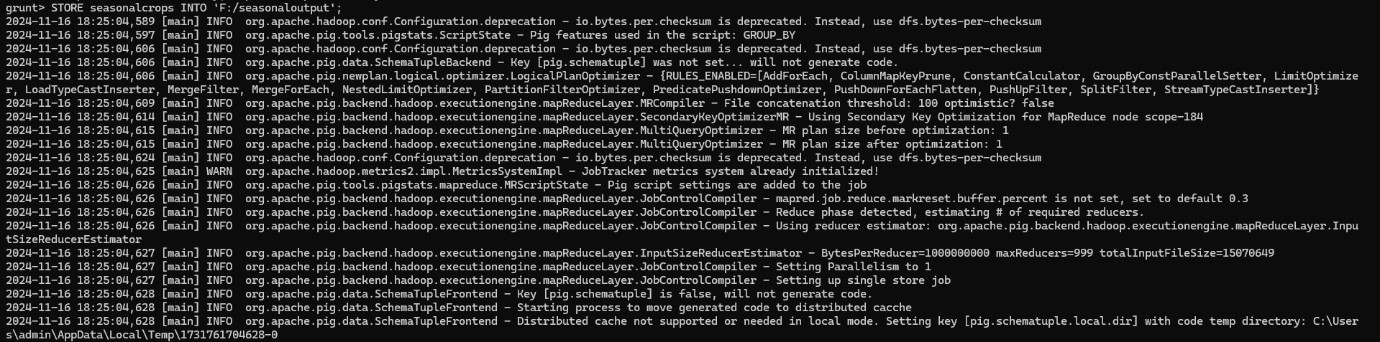
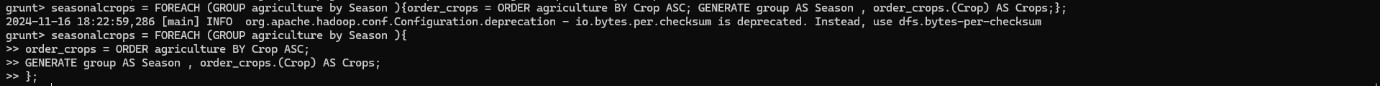
The output will be in a file named ‘part-r-00000’ which needs to be renamed as ‘part-r-00000.csv’ to be opened in the Excel format and to make it readable.

# Query 2: Generate Total Crop wise Production and Area



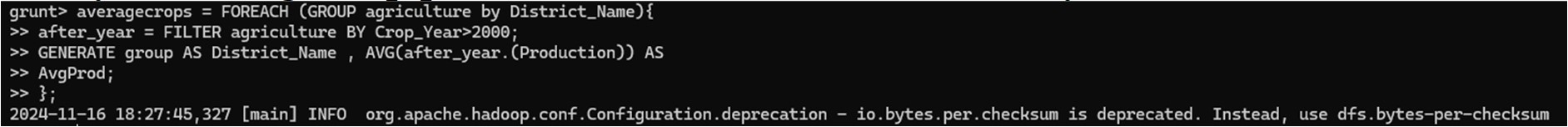
The output will be in a file named ‘part-r-00000’ which needs to be renamed as ‘part-r-00000.csv’ to be opened in the Excel format and to make it readable.

# Query 3: The majority of crops are grown in a Season and in which year.



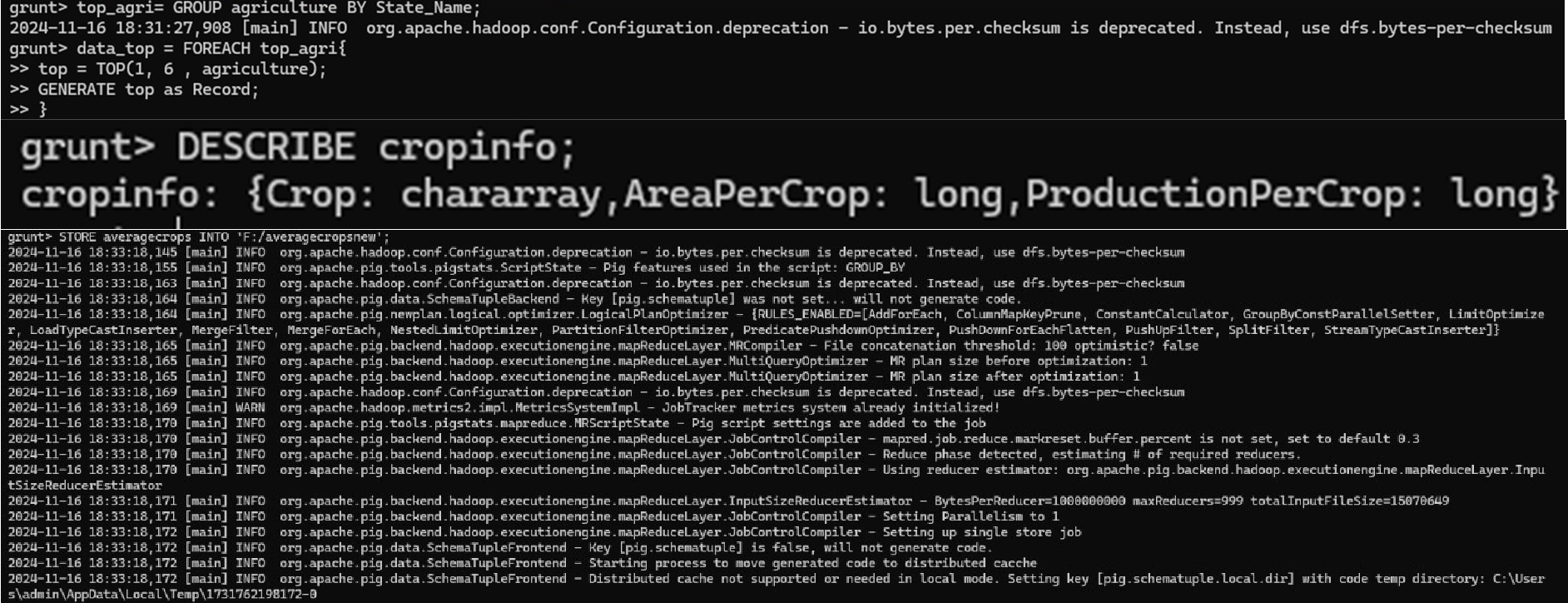
The output will be in a file named ‘part-r-00000’ which needs to be renamed as ‘part-r-00000.csv’ to be opened in the Excel format and to make it readable.

# Query 4: Average crop production in each district after the year 2000.





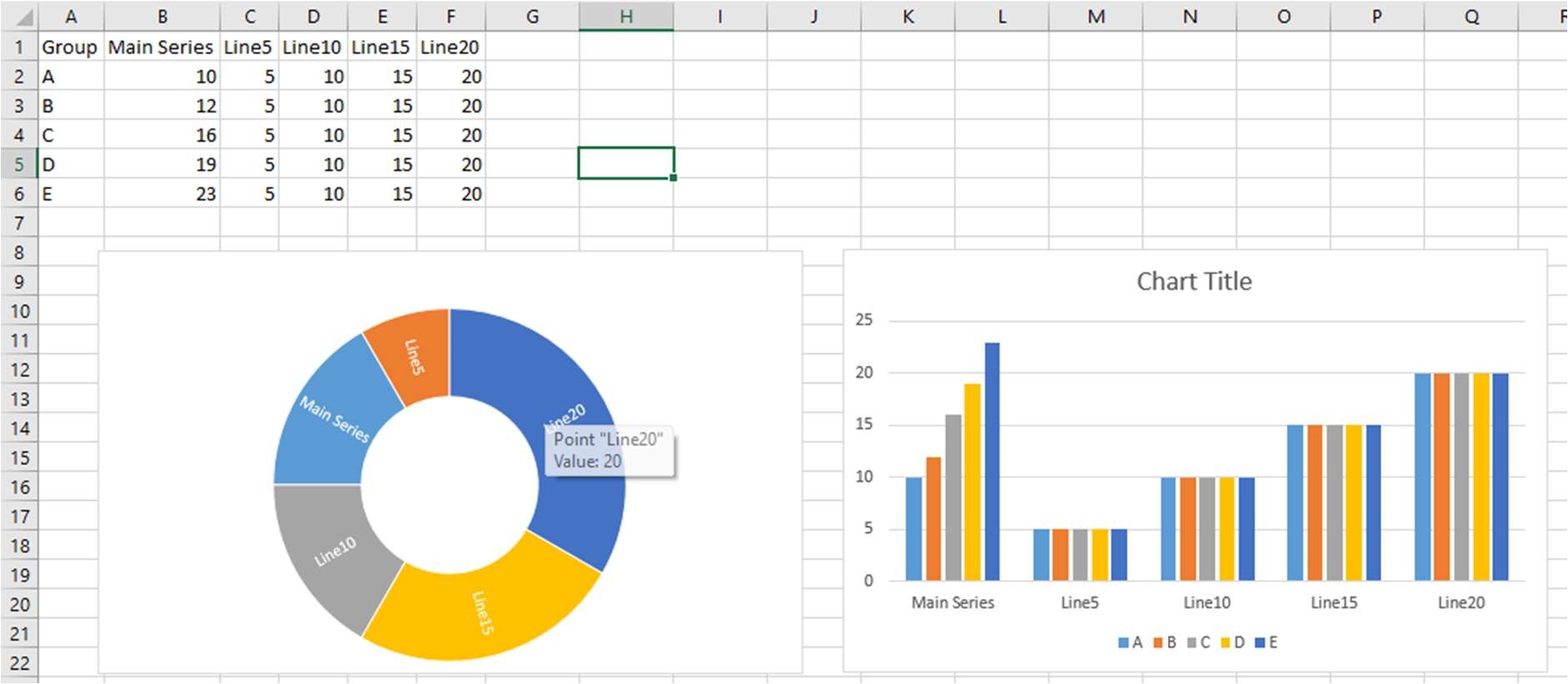
**Query 5: Highest produced crops and details from each State.**



# Data visualization using Excel

## Create some data -> Select data - >Insert -> Recommended Chart -> All charts ->Select any chart -

>Done



# Data visualization using Tableau

Step 1: Set Up Your Tableau Environment

Download Tableau: If you haven't already, download and install Tableau Desktop. You can use the Tableau Public version if you're just starting out (free but with limitations on privacy).

Open Tableau: Launch Tableau Desktop and start a new project by clicking on File > New. Step 2: Load Your Data into Tableau

Connect to Data:

On the left sidebar, under Connections, click the data source type (e.g., Microsoft Excel, Text File, Web Data Connector, SQL Server).

Select your file (Excel, CSV, etc.) and click Open. Tableau will load the data into the workspace. Preview the Data:

Tableau will display a preview of the data. It automatically identifies the dimensions (qualitative variables like Region, Category) and measures (quantitative variables like Sales, Profit).

Drag and Drop to Workspace:

Drag the data table (sheet) you want to analyze to the canvas. Create Basic Visualizations

In Tableau, visualizations are built by dragging dimensions (e.g., Category, Region) and measures (e.g., Sales, Profit) onto rows, columns, and the "marks" area.

Bar Chart (Sales by Category)

Drag ‘Category’ to Columns: This will create categories along the x-axis.

Drag ‘Sales’ to Rows: Tableau will generate a bar chart showing sales across different categories.

Sort Bars: To sort bars by sales, click the Sort icon on the chart or right-click on the axis and choose “Sort”. Line Chart (Sales Over Time)

Drag ‘Date’ to Columns: Tableau will automatically recognize that this is a time-based field and will create a time axis. Drag ‘Sales’ to Rows: A line chart will be generated showing sales trends over time.

Adjust Time Granularity: You can adjust the time scale (e.g., from year to month or day) by right-clicking on the Date field and selecting the level of granularity.

Pie Chart (Sales by Region)

Drag ‘Region’ to Columns: This will set the regions along the pie chart.

Drag ‘Sales’ to Rows: Tableau will create a pie chart where each slice represents sales by region. Show Labels: Click on the pie chart and enable labels to show the sales values on each segment. Scatter Plot (Profit vs. Sales) Drag ‘Profit’ to Columns: This will be one axis. Drag ‘Sales’

to Rows: This will be the other axis.

Adjust Size and Color: Use the Size and Color shelves to adjust the size of the scatter points based on a measure (like Quantity) and color based on another (like Region).

Step 4: Refine Your Visualizations Filters:

Drag a dimension (e.g., Region) to the Filters shelf to filter data by a specific region. You can also filter by measures (e.g., only show sales greater than $100,000).

Color:

Drag a dimension (like Category) to the Color shelf to differentiate segments using color.

Adjust the color scheme by clicking on the Color legend and selecting “Edit Colors” to choose a color palette that fits the data.

Tooltips:

Customize tooltips to provide more detailed information when hovering over data points. Click on Tooltip in the Marks card to modify what is displayed.

Add Annotations:

Right-click on a data point or part of the visualization and choose Annotate to add notes or explanations. Step 5: Create a Dashboard

To combine multiple visualizations into a single dashboard:

Create a New Dashboard:

Click on the Dashboard tab at the bottom of Tableau and then select New Dashboard. Drag Worksheets to Dashboard:

Drag each worksheet (e.g., bar chart, line chart, pie chart) from the left sidebar into the dashboard area. Size Your

Dashboard: Adjust the size of the dashboard by choosing Fixed Size or Automatic under the Size section on the left. Add Interactivity:

To add interactivity between visualizations, use Actions. For example, you can set up a filter action where clicking on a bar in the bar chart filters data in the line chart.

Go to Dashboard > Actions > Add Action > Filter to set up this feature. Step 6: Final Touches and Sharing

Customize Titles and Labels:

Edit titles and labels to make your visualizations clearer.

Right-click the title area and choose Edit to change the text, fonts, and alignment. Publish the Dashboard:

Once satisfied with your visualizations, you can publish the dashboard.

Click File > Save As to save the workbook locally or Server to upload it to Tableau Online or Tableau Server. Export:

Export the dashboard as a PDF, image, or PowerPoint by going to File > Export. Practical Example Scenario: Sales and Profit Analysis

Let’s assume you have a dataset with the following fields: Region, Category, Product, Sales, Profit, and Date. Here’s how you would visualize this data:

Bar Chart: Show total sales for each product category. Drag Category to Columns.

Drag Sales to Rows.

Sort the chart to display categories with the highest sales first. Line Chart: Show sales trends over time (e.g., monthly sales). Drag Date to Columns (ensure it’s set to a monthly level).

Drag Sales to Rows to create a line chart showing sales over the months. Scatter Plot: Show the relationship between sales and profit.

Drag Sales to Columns. Drag Profit to Rows.

Color the scatter points by Region to see the geographical distribution of sales and profit. Pie Chart: Show the sales distribution by region.

Drag Region to Columns.

Drag Sales to Rows to create a pie chart.

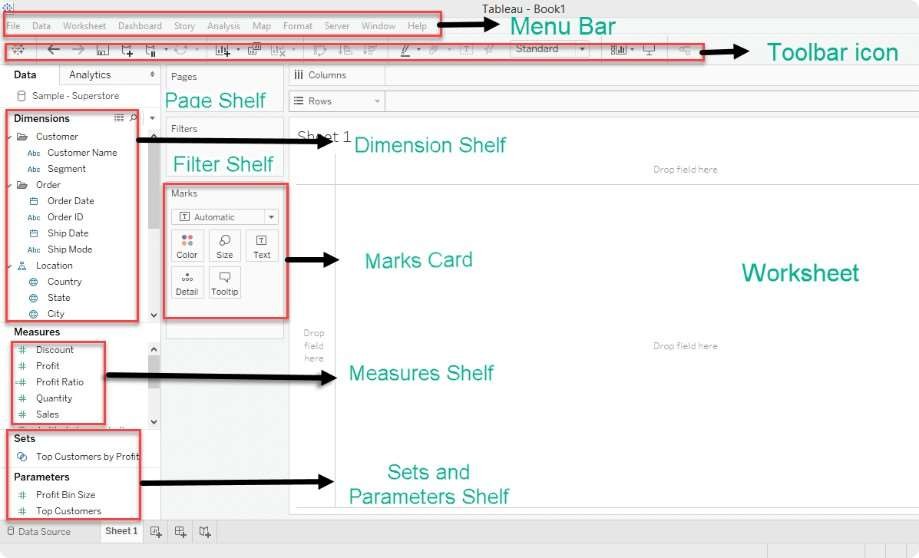
Label the slices with the sales values for better clarity. Dashboard:

Combine the bar chart, line chart, scatter plot, and pie chart into a single dashboard.

Add filter actions, e.g., clicking on a region in the pie chart filters the data in the bar and line charts. Conclusion By following this process, you will be able to create interactive, insightful data visualizations using

Tableau. From importing data to refining visualizations and creating dashboards, Tableau enables you to tell a compelling story with your data. With these basic skills, you can analyze trends, identify patterns, and make informed decisions based on your data.

The downloaded file is a zip file that contains an excel that looks like the given picture below: will demonstrate to you how to load data and perform some preprocessing.

Tableau work-page consist of different section. Let’s understand them first before plotting our graphs.