Spark DataFrames

#Creating DataFrames

```
 data=[
... {'id':1, 'name':'Ankit'},
... {'id':2, 'name':'Ravi'},
... {'id':3, 'name':'Pankaj'}
... ]


df=sc.parralelize(data)
df=spark.createDataFrame(df)
```

#Displaying the Data

```
df.show()
```

#Printing the Schema

```
df.printSchema()
```

#Projection(Displaying Selected Columns)

```
df.select("name").show()
```

#Selection(Displaying Selected Rows)

```
df.filter(df["id"]=2).show()
```

#Summarizing the Dataframe

```
df.describe().show()
```

#Loading .csv into Dataframe

```
product = spark.read.csv('product.csv', inferSchema=True, header=True)
product.printSchema()
```

#Ordering Records

```
product.orderBy("PRICE").show()
product.orderBy(product["PRICE"]).show()
product.orderBy(product["PRICE"].desc()).show()
```

#Grouping Records

```
product.groupBy("CID")
```

1. agg
2. avg
3. count
4. max
5. min
6. mean
7. sum

```
product.groupBy("CID").sum().show()
product.groupBy("BID").sum().show()


category_wise_data = product.groupBy("CID")
category_wise_data.sum().show()
category_wise_data.agg({'PRICE':'sum'}).show()
category_wise_data.agg({'PRICE':'avg'}).show()

product.agg({'PRICE':'sum'}).show()
```