```
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame using spark session variable createDataFrame method
# spark.createDataFrame(data,schema)

df = spark.createDataFrame([1,'ravi','2','Reshwanth',4,'Vikranth'],
["id","name"])
# default minimum partitions 1
# displaying data can be done in three ways.
display(df)  # html table format
df.show() # text table format
df.collect()  # ROW format (raw data in python collection list and tuple)
```

```
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame using spark session variable range method
# spark.range(n)

df = spark.range(10)

# default minimum partitions 1
# displaying data can be done in three ways.
display(df)  # html table format
df.show() # text table format
df.collect()  # ROW format (raw data in python collection list and
tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# converting RDD to DataFrame

rdd = spark.sparkContext.parallelize([(1,"ravi"),(2,"sindhu"),(3,"reshwanth"),
(4,"vikranth")])
df = rdd.toDF(["id","name"])

# default minimum partitions 1
# displaying data can be done in three ways.
display(df)  # html table format
df.show() # text table format
df.collect()  # ROW format (raw data in python collection list and tuple)
```

```
#LearningWithRavi https://youtube.com/@TRRaveendra

# creating dataframe using SPARK READ API. spark.read.csv

df = spark.read.csv("dbfs:/databricks-datasets/asa/airlines/1987.csv",header=True)


# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and
tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# creating dataframe using SPARK READ API. spark.read.json
# Reading json files

df = spark.read.json("dbfs:/databricks-datasets/structured-streaming/events/file-
0.json")
# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# creating dataframe using SPARK READ API. spark.read.text
# Reading text files

df = spark.read.text("dbfs:/databricks-datasets/credit-card-fraud/dbcl-10.txt")

# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and
tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from reading Parquet Files

# creating dataframe using SPARK READ API. spark.read.parquet
# Reading parquet files

df = spark.read.parquet("/databricks-datasets/amazon/data20K")

# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and
tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from reading delta table by path

# creating dataframe using SPARK READ API. spark.read.format("delta")
# Reading delta files

df = spark.read.format("delta").load("dbfs:/databricks-datasets/nyctaxi-with-
zipcodes/subsampled/")
# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from reading avro files

# creating dataframe using SPARK READ API. spark.read.format("avro")
# Reading avro files

df = spark.read.format("avro").load("/avrofiles_path/")

# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and
tuple)
```

```
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from reading orc files

# creating dataframe using SPARK READ API. spark.read.format("orc")
# Reading orc files

df = spark.read.orc("/orcfiles_path/")

# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and
tuple)
```

```
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from reading xml files

# Direct API is not available for xml and excel files.
# install spark-xml library at cluster level from maven central repository
# install spark-xml_2.12  version to avoid issues.

df = spark.read.format("com.databricks.spark.xml")\
               .option("rootTag", "emp")\
               .option("rowTag","row")\
               .load("/FileStore/tables/emp_xml.xml")

# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and
tuple)
```

```
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from reading excel files

# Direct API is not available for excel and excel files.
# install spark-excel library at cluster level from maven central repository
# install this version  - com.crealytics:spark-excel_2.12:0.13.5

df =
spark.read.format("com.crealytics.spark.excel").load("/FileStore/tables/emp.xlsx",header=True)

# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from reading image files

# use format as image for images.
# spark.read.format("image").load("path")

df = spark.read.format("image").load("/databricks-datasets/flower_photos/roses")

# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and
tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from reading binary files like images, videos,audio files..

# spark.read.format("binaryFile").load("path")

df = spark.read.format("binaryFile").load("/databricks-
datasets/cctvVideos/mp4/train")
# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from a table

# spark.table("table_name")

df = spark.table("table_name")

# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and
tuple)
```

```python
#LearningWithRavi https://youtube.com/@TRRaveendra

# Creating DataFrame from a sql query

# spark.sql("select * from table_name")

df = spark.sql("select * from emp")

# default minimum partitions 1
# Depends on data it will create N No of partitions and block size is 128MB
# Get Default Partition block size in bytes.
spark.conf.get("spark.sql.files.maxPartitionBytes")

# displaying data can be done in three ways. use limit function to get few rows.
display(df.limit(100))  # html table format
df.limit(10).show() # text table format
df.limit(10).collect()  # ROW format (raw data in python collection list and
tuple)
```