In this video We will Cover

- **PySpark Dataframe**
- **Reading The Dataset**
- Checking the Datatypes of the Columns(Schema)
- Selecting Columns And Indexing
- Check Describe option similar to Pandas
- Adding Columns
- Dropping Columns
- Renaming Columns

In [1]:

```
from pyspark.sql import SparkSession
```

In [2]:

```
spark = SparkSession.builder.appName('Dataframes').getOrCreate()
```

In [3]:

```
spark
```

Out[3]: **SparkSession - in-memory**

SparkContext

Spark UI

Version `v3.3.0`
Master `local[*]`
AppName `Dataframes`

In [4]:

```
## read the dataset
spark.read.option('header', 'true').csv('test1.csv')
```

Out[4]:

```
DataFrame[name: string, age: string, experience: string]
```

In [5]:

```
spark.read.option('header', 'true').csv('test1.csv').show()
```

```
+--------+---+----------+
|    name|age|experience|
+--------+---+----------+
| Ajinkya| 32|        10|
|Narendra| 29|         7|
|    Amit| 33|        12|
|  Nikhil| 30|         9|
+--------+---+----------+
```

In [10]:

```
df_pyspark = spark.read.option('header', 'true').csv('test1.csv')
```

In [11]:

```
## check the schema
df_pyspark.printSchema()
```

```
root
 |-- name: string (nullable = true)
 |-- age: string (nullable = true)
 |-- experience: string (nullable = true)
```

```
In [12]: df_pyspark = spark.read.option('header', 'true').csv('test1.csv', inferSchema=True)
```

```
In [13]: df_pyspark.printSchema()
```

```
root
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- experience: integer (nullable = true)
```

```
In [45]: df_pyspark = spark.read.csv('test1.csv', header=True, inferSchema=True)
```

```
In [18]: df_pyspark.printSchema()
```

```
root
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- experience: integer (nullable = true)
```

```
In [19]: df_pyspark.show()
```

```
+--------+---+----------+
|    name|age|experience|
+--------+---+----------+
| Ajinkya| 32|        10|
|Narendra| 29|         7|
|    Amit| 33|        12|
|  Nikhil| 30|         9|
+--------+---+----------+
```

```
In [20]: type(df_pyspark)
```

```
Out[20]: pyspark.sql.dataframe.DataFrame
```

```
In [21]: df_pyspark.columns
```

```
Out[21]: ['name', 'age', 'experience']
```

```
In [22]: df_pyspark.head(3)
```

```
Out[22]: [Row(name='Ajinkya', age=32, experience=10),
          Row(name='Narendra', age=29, experience=7),
          Row(name='Amit', age=33, experience=12)]
```

```
In [23]: df_pyspark.show()
```

```
+--------+---+----------+
|    name|age|experience|
+--------+---+----------+
| Ajinkya| 32|        10|
|Narendra| 29|         7|
|    Amit| 33|        12|
|  Nikhil| 30|         9|
+--------+---+----------+
```

```
In [25]: df_pyspark.select('name', 'age').show()
```

```
+--------+---+
|    name|age|
+--------+---+
| Ajinkya| 32|
|Narendra| 29|
|    Amit| 33|
|  Nikhil| 30|
+--------+---+
```

In [26]:
```
df_pyspark.select('name')
```

Out[26]:
```
DataFrame[name: string]
```

In [27]:
```
df_pyspark.select('name').show()
```

```
+--------+
|    name|
+--------+
| Ajinkya|
|Narendra|
|    Amit|
|  Nikhil|
+--------+
```

In [28]:
```
df_pyspark.select('name', 'experience')
```

Out[28]:
```
DataFrame[name: string, experience: int]
```

In [29]:
```
df_pyspark.select('name', 'experience').show()
```

```
+--------+----------+
|    name|experience|
+--------+----------+
| Ajinkya|        10|
|Narendra|         7|
|    Amit|        12|
|  Nikhil|         9|
+--------+----------+
```

In [32]:
```
df_pyspark.select(['name', 'experience']).show()
```

```
+--------+----------+
|    name|experience|
+--------+----------+
| Ajinkya|        10|
|Narendra|         7|
|    Amit|        12|
|  Nikhil|         9|
+--------+----------+
```

In [33]:
```
df_pyspark['name']
```

Out[33]:
```
Column<'name'>
```

In [34]:
```
df_pyspark.dtypes
```

Out[34]:
```
[('name', 'string'), ('age', 'int'), ('experience', 'int')]
```

In [35]:

```
df_pyspark.describe()
```

Out[35]:

```
DataFrame[summary: string, name: string, age: string, experience: string]
```

In [46]:

```
df_pyspark.describe().show()
```

```
+-------+-------+------------------+------------------+
|summary|   name|               age|        experience|
+-------+-------+------------------+------------------+
|  count|      4|                 4|                 4|
|   mean|   null|              31.0|               9.5|
| stddev|   null|1.8257418583505534|2.0816659994661326|
|    min|Ajinkya|                29|                 7|
|    max| Nikhil|                33|                12|
+-------+-------+------------------+------------------+
```

In [47]:

```
## Adding Columns in dataframe
df_pyspark = df_pyspark.withColumn('experience after 2 yrs', df_pyspark['experience']+2)
```

In [48]:

```
df_pyspark.show()
```

```
+--------+---+----------+----------------------+
|    name|age|experience|experience after 2 yrs|
+--------+---+----------+----------------------+
| Ajinkya| 32|        10|                    12|
|Narendra| 29|         7|                     9|
|    Amit| 33|        12|                    14|
|  Nikhil| 30|         9|                    11|
+--------+---+----------+----------------------+
```

In [ ]:

```
## Drop the Columns
```

In [51]:

```
df_pyspark = df_pyspark.drop('experience after 2 yrs')
```

In [52]:

```
df_pyspark.show()
```

```
+--------+---+----------+
|    name|age|experience|
+--------+---+----------+
| Ajinkya| 32|        10|
|Narendra| 29|         7|
|    Amit| 33|        12|
|  Nikhil| 30|         9|
+--------+---+----------+
```

In [54]:

```
## Rename the Columns
df_pyspark.withColumnRenamed('name', 'new name').show()
```

```
+--------+---+----------+
|new name|age|experience|
+--------+---+----------+
| Ajinkya| 32|        10|
|Narendra| 29|         7|
|    Amit| 33|        12|
|  Nikhil| 30|         9|
+--------+---+----------+
```

In [ ]:

```

```