Pyspark basics

1. File reading

Spark.read.format(‘’) \

.options(header=True, inferSchema=True….)

.load(file\_path)

1. User defined schema

Method 1: StructType

StructType([

StructField(column\_name, dtype, nullable)

])

dtype [StringType(), IntegerType()…]

nullable [True, False]

Method 2: DDL

Schema = ‘’’ Name String, Age Integer ‘’’

Spark.read.format(‘’) \

.options(header=True, schema=schema\_var….)

.load(file\_path)

1. Select
2. Selecting few columns:

df\_new = df.select(‘column1’, ‘column2’….)

or

df\_new = df.select(col(column\_name1, col(column\_name2)…)

1. Alias
2. df.select(col('Name').alias('Student Name'), col('Age')).display()



1. Filter()
2. df.filter(col('Name') == 'Rahul').display()
3. df.filter((col('Name') == 'Rahul') & (col('Age') > 22))
4. Isin()

Just like SQL in condition

1. df.filter(col('Name').isin('Rahul','Deepak'))
2. isNull() and isNotNull()
3. df.filter(col('Name').isNull())
4. withColumnRenamed():

df.withColumnRenamed('Age','Student Age').display()

1. sort()
2. df.sort(col(column\_name).desc())