1/09/2022-pyspark task

1.Difference between filter and where?

A. Both ‘filter’ and ‘where’ in Spark SQL gives same result. **There is no difference between the two**. It’s just filter is simply the standard Scala name for such a function, and where is for people who prefer SQL.

2. how select is used in pyspark?

In PySpark, select() function is used to select single, multiple, column by index, all columns from the list and the nested columns from a DataFrame, Pyspark select() is a transformation function hence it returns a new DataFrame with the selected columns.

3. what is different types of filter in Pyspark?

The startsWith() and endsWith() methods in Pyspark belong to the Column class and are used to search DataFrame rows by checking if the column value starts with some value or ends with some value. Both are used for filtering data in applications.

* **startsWith() method:** This method is used to return a Boolean value. It shows TRUE when the column's value starts with the specified string and FALSE when the match is not satisfied in that column value.
* **endsWith() method:** This method is used to return a Boolean value. It shows TRUE when the column's value ends with the specified string and FALSE when the match is not satisfied in that column value. Both methods are case-sensitive.

4. ways to read a file in Pyspark and what are the file extension that Pyspark support to read?

df = spark.read.csv("file path.csv") (for single file reading)

df = spark.read.csv("path1,path2,path3") (for multiple file reading)

df = spark.read.csv("Folder path") (for all csv file reading in that folder)

5.ways to write a file in Pyspark and what are the file extension that Pyspark support to write?

df.write.option("header",True) \

.csv("/file path")

df2.write.options(header='True', delimiter=',') \

.csv("/file path")

df2.write.mode('overwrite').csv("/file path")

df2.write.format("csv").mode('overwrite').save("/file path")

The following three file systems are supported by Spark:

1. Hadoop Distributed File System (HDFS).
2. Local File system.
3. Amazon S3