

Module 4 Cheat Sheet: DataFrames and Spark SQL

Package/Method	Description	Code Example
appName()	A name for your job to display on the cluster web UI.	<div>1. 1</div> <div>2. 2</div> <div>1. from pyspark.sql import SparkSession</div> <div>2. spark = SparkSession.builder.appName("MyApp").getOrCreate()</div> <div>Copied!</div> <div>1. 1</div> <div>2. 2</div> <div>3. 3</div> <div>4. 4</div> <div>1. from pyspark.sql import SparkSession</div> <div>2. spark = SparkSession.builder.appName("MyApp").getOrCreate()</div> <div>3. data = [("Jhon", 30), ("Peter", 25), ("Bob", 35)]</div> <div>4. columns = ["name", "age"]</div>
createDataFrame()	Used to load the data into a Spark DataFrame.	<div>Copied!</div> <div>Creating a DataFrame</div> <div>1. 1</div> <div>1. df = spark.createDataFrame(data, columns)</div> <div>Copied!</div>
createTempView()	Create a temporary view that can later be used to query the data. The only required parameter is the name of the view.	<div>1. 1</div> <div>1. df.createOrReplaceTempView("cust_tbl")</div> <div>Copied!</div> <div>Replace NULL/None values in a DataFrame</div>
fillna()	Used to replace NULL/None values on all or selected multiple DataFrame columns with either zero (0), empty string, space, or any constant literal values.	<div>1. 1</div> <div>1. filled_df = df.fillna(0)</div> <div>Copied!</div> <div>Replace with zero</div>
filter()	Returns an iterator where the items are filtered through a function to test if the item is accepted or not.	<div>1. 1</div> <div>1. filtered_df = df.filter(df['age'] > 30)</div> <div>Copied!</div>
getOrCreate()	Get or instantiate a SparkContext and register it as a singleton object.	<div>1. 1</div> <div>1. spark = SparkSession.builder.getOrCreate()</div> <div>Copied!</div> <div>Grouping data and performing aggregation</div>
groupby()	Used to collect the identical data into groups on DataFrame and perform count, sum, avg, min, max functions on the grouped data.	<div>1. 1</div> <div>1. grouped_df = df.groupBy("age").agg({"age": "count"})</div> <div>Copied!</div> <div>Returning the first 5 rows</div>
head()	Returns the first <i>n</i> rows for the object based on position.	<div>1. 1</div> <div>1. first_5_rows = df.head(5)</div> <div>Copied!</div>
import	Used to make code from one module accessible in another. Python imports are crucial for a successful code structure. You may reuse code and keep your projects manageable by using imports effectively, which can increase your productivity.	<div>1. 1</div> <div>1. from pyspark.sql import SparkSession</div> <div>Copied!</div> <div>1. 1</div> <div>1. import pandas as pd</div> <div>Copied!</div>
pd.read_csv()	Required to access data from the CSV file from Pandas that retrieves data in the form of the data frame.	<div>Reading data from a CSV file into a DataFrame</div> <div>1. 1</div> <div>1. df_from_csv = pd.read_csv("data.csv")</div> <div>Copied!</div>
pip	To ensure that requests will function, the pip program searches for the package in the Python Package Index (PyPI), resolves any dependencies, and installs everything in your current Python environment.	<div>1. 1</div> <div>1. pip list</div> <div>Copied!</div>
pip install	The pip install <package> command looks for the latest version of the package and	<div>1. 1</div> <div>1. pip install pyspark</div>

Package/Method	Description	Code Example
	installs it.	<div>Copied!</div>
printSchema()	Used to print or display the schema of the DataFrame or data set in tree format along with the column name and data type. If you have a DataFrame or data set with a nested structure, it displays the schema in a nested tree format.	<div>1. 1</div> <div>1. df.printSchema()</div> <div>Copied!</div> <div>1. 1</div> <div>1. import pandas as pd</div> <div>Copied!</div> <div>Create a sample DataFrame</div> <div>1. 1</div> <div>2. 2</div> <div>1. data = {'A': [1, 2, 3], 'B': [4, 5, 6]}</div> <div>2. df = pd.DataFrame(data)</div>
rename()	Used to change the row indexes and the column labels.	<div>Copied!</div> <div>Rename columns</div> <div>1. 1</div> <div>1. df = df.rename(columns={'A': 'X', 'B': 'Y'})</div> <div>Copied!</div> <div>The columns 'A' and 'B' are now renamed to 'X' and 'Y'</div> <div>1. 1</div> <div>1. print(df)</div> <div>Copied!</div>
select()	Used to select one or multiple columns, nested columns, column by index, all columns from the list, by regular expression from a DataFrame. select() is a transformation function in Spark and returns a new DataFrame with the selected columns.	<div>1. 1</div> <div>1. selected_df = df.select('name', 'age')</div> <div>Copied!</div>
show()	Spark DataFrame show() is used to display the contents of the DataFrame in a table row and column format. By default, it shows only twenty rows, and the column values are truncated at twenty characters.	<div>1. 1</div> <div>1. df.show()</div> <div>Copied!</div> <div>Sorting DataFrame by a column in ascending order</div> <div>1. 1</div> <div>1. sorted_df = df.sort("age")</div>
sort()	Used to sort DataFrame by ascending or descending order based on single or multiple columns.	<div>Copied!</div> <div>Sorting DataFrame by multiple columns in descending order</div> <div>1. 1</div> <div>1. sorted_df_desc = df.sort(["age", "name"], ascending=[False, True])</div> <div>Copied!</div> <div>1. 1</div> <div>1. from pyspark import SparkContext</div> <div>Copied!</div>
SparkContext()	It is an entry point to Spark and is defined in org.apache.spark package since version 1.x and used to programmatically create Spark RDD, accumulators, and broadcast variables on the cluster.	<div>Creating a SparkContext</div> <div>1. 1</div> <div>1. sc = SparkContext("local", "MyApp")</div> <div>Copied!</div> <div>1. 1</div> <div>1. from pyspark.sql import SparkSession</div> <div>Copied!</div>
SparkSession	It is an entry point to Spark, and creating a SparkSession instance would be the first statement you would write to the program with RDD, DataFrame, and dataset	<div>Creating a SparkSession</div> <div>1. 1</div> <div>1. spark = SparkSession.builder.appName("MyApp").getOrCreate()</div> <div>Copied!</div>

Package/Method	Description	Code Example
spark.read.json()	Spark SQL can automatically infer the schema of a JSON data set and load it as a DataFrame. The read.json() function loads data from a directory of JSON files where each line of the files is a JSON object. Note that the file offered as a JSON file is not a typical JSON file.	<pre>1. 1 1. json_df = spark.read.json("customer.json")</pre> <div>Copied!</div>
spark.sql()	To issue any SQL query, use the sql() method on the SparkSession instance. All spark.sql queries executed in this manner return a DataFrame on which you may perform further Spark operations if required.	<pre>1. 1 2. 2 1. result = spark.sql("SELECT name, age FROM cust_tbl WHERE age > 30") 2. result.show()</pre> <div>Copied!</div>
spark.udf.register()	In PySpark DataFrame, it is used to register a user-defined function (UDF) with Spark, making it accessible for use in Spark SQL queries. This allows you to apply custom logic or operations to DataFrame columns using SQL expressions.	<p>Registering a UDF (User-defined Function)</p> <pre>1. 1 2. 2 3. 3 4. 4 5. 5 1. from pyspark.sql.functions import udf 2. from pyspark.sql.types import StringType 3. def my_udf(value): 4. return value.upper() 5. spark.udf.register("my_udf", my_udf, StringType())</pre> <div>Copied!</div>
where()	Used to filter the rows from DataFrame based on the given condition. Both filter() and where() functions are used for the same purpose.	<p>Filtering rows based on a condition</p> <pre>1. 1 1. filtered_df = df.where(df['age'] > 30)</pre> <div>Copied!</div>
withColumn()	Transformation function of DataFrame used to change the value, convert the data type of an existing column, create a new column, and many more.	<p>Adding a new column and performing transformations</p> <pre>1. 1 2. 2 1. from pyspark.sql.functions import col 2. new_df = df.withColumn("age_squared", col("age") ** 2)</pre> <div>Copied!</div>
withColumnRenamed()	Returns a new DataFrame by renaming an existing column.	<p>Renaming an existing column</p> <pre>1. 1 1. renamed_df = df.withColumnRenamed("age", "years_old")</pre> <div>Copied!</div>



Skills Network