Module 5 Cheat Sheet: Development and Runtime Environment Options

Package or method	Description	Code example
SparkSession.builder.getOrCreate()	Gets an existing SparkSession or, if there is no existing one, creates a new one based on the options set in this builder.	1. 1 1. from pyspark.sql import SparkSession Copied! Create a SparkSession or get an existing one: 1. 1 1. spark = SparkSession.builder.appName("myApp").getOrCreate() Copied! Now you can use the spark object to work with Spark
cd	Used to move efficiently from the existing working directory to different directories on your system.	Basic syntax of cd command: 1. 1 1. cd [options]... [directory] Copied! Example 1: Change directory location to folder1: 1. 1 1. cd /usr/local/folder1 Copied! Example 2: Get back to the previous working directory 1. 1 1. cd - Copied! Example 3: Move up one level from the present working directory tree 1. 1 1. cd .. Copied!
docker-compose	Tool for defining and running multicontainer Docker applications. It uses the YAML file to configure the services and enables us to create and start all the services from just one configuration file.	Example (docker-compose.yml) 1. 1 2. 2 3. 3 4. 4 5. 5 6. 6 7. 7 8. 8 1. version: '3' 2. services: 3. web: 4. image: nginx:latest 5. ports: 6. - "80:80" 7. db: 8. image: postgres:latest Copied!
git clone	You can create a copy of a specific repository or branch within a repository.	1. 1 1. git clone REPOSITORY_URL [DESTINATION_DIRECTORY] Copied!
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 to increase productivity.	1. 1 1. from pyspark.sql import SparkSession Copied!
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.	1. 1 1. pip list Copied!

Package or method	Description	Code example
pip install	The pip install <package> command looks for the latest version of the package and installs it.	1. 1 1. pip install package_name Copied!
pip3 install	pip3 is the official package manager and pip command for Python 3. It enables installing and managing third-party software packages with features and functionality not found in the Python standard library. Pip3 installs packages from PyPI (Python Package Index), but it won't resolve dependencies or help you solve dependency conflicts.	1. 1 1. pip3 install package_name Copied!
print()	Prints the specified message to the screen or other standard output device. The message can be a string or any other object; the object will be converted into a string before being written to the screen.	1. 1 1. print("Hello, World!") Copied!
python3	Python 3 is a widely used programming language known for its readability and versatility.	1. 1 2. 2 3. 3 1. sudo apt-get update 2. sudo apt-get install python3 3. python3 --version #to verify the python version Copied!
sc.setloglevel()	Using this method, you can change the log level to the desired level. Valid log levels include ALL, DEBUG, ERROR, FATAL, INFO, OFF, TRACE, and WARN.	Import necessary modules: 1. 1 1. from pyspark import SparkContext Copied! Create a SparkContext: 1. 1 1. sc = SparkContext("local", "LogLevelExample") Copied! Set the log level to a desired level (e.g., INFO, ERROR): 1. 1 1. sc.setLogLevel("INFO") Copied! Now, any logging messages with a severity level equal to or higher than INFO will be displayed
setMaster()	Denotes where to run your Spark application, local or cluster. When you run on a cluster, you need to specify the address of the Spark master or Driver URL for a distributed cluster. We usually assign a local[*] value to setMaster() in Spark while doing internal testing.	1. 1 2. 2 3. 3 1. <pre>from pyspark import SparkContext </pre> 2. Create a SparkContext with a local master: 3. <pre>sc = SparkContext("local[*]", "MyApp")</pre></td> Copied!
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.	1. 1 1. df.show() Copied!
source	Used to execute a script file in the current shell environment, allowing you to modify the current shell environment in the same way you would if you had typed commands manually.	Assuming a Bash shell: 1. 1 1. source myscript.sh Copied!
virtualenv	Primarily a command-line application. It modifies the environment variables in a shell to create an isolated Python environment, so you'll need to have a shell to run it. You can type in virtualenv (name of the application), followed by flags that control its behavior.	Creating a virtual environment named "myenv": 1. 1 1. virtualenv myenv Copied!


Skills Network