SPARK BASICS SIMPLY EXPLAINED

PART 3: SET UP YOUR LOCAL SPARK ENVIRONMENT on Windows

- *Cloud providers*
- *Step by Step installation guide*
- *Sample code for testing*

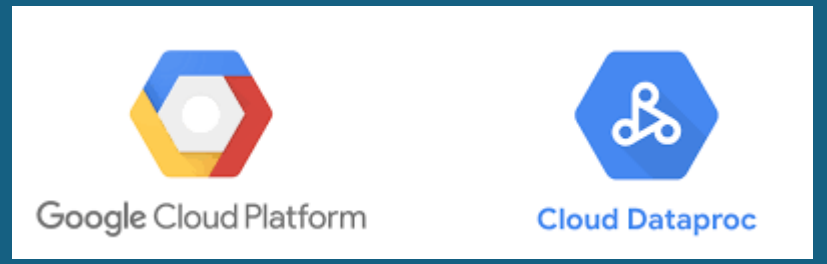
Data with
Nikk the Greek

Cloud providers support Spark

The Data and AI Company — Databricks
Azure Databricks – Open Data Lakehouse in Azure | Microsoft Azure
Azure Databricks documentation | Microsoft Learn

Data Analytics | Microsoft Fabric
Microsoft Fabric documentation - Microsoft Fabric | Microsoft Learn

Big Data Platform - Amazon EMR – AWS
Amazon EMR Documentation



Dataproc | Google Cloud

Local Set-Up – Step 1: Required Installations

- Required Software:
  - Python 3.10.11: Python Release Python 3.10.11 | Python.org
  - Java v8: Latest Releases | Adoptium
  - Git: Git - Downloads (git-scm.com)
  - Miniconde (recommended): Miniconda — Anaconda documentation or Anaconda: Free Download | Anaconda
  - VSCode: Visual Studio Code - Code Editing. Redefined
- Spark and Hadoop Download:
  - Download e.g. Spark 3.5: Downloads | Apache Spark and unzip having the path D:\Spark\spark-3.5.0-bin-hadoop3 and then the e.g. „bin“ folder inside
  - Hadoop 3.3.0: Download winutils and .ddl file from GIT repo to winutils/hadoop-3.3.0/bin at master · kontext-tech/winutils (github.com) to D:\Spark\winutils\bin

Local Set-Up – Step 2: System Variables

- JAVA_HOME = C:\Program Files\Eclipse Adoptium\jdk-8.0.402.6-hotspot
- SPARK_HOME = D:\Spark\spark-3.5.0-bin-hadoop3
- HADOOP_HOME = D:\Spark\winutils
- Path =
  - %JAVA_HOME%\bin
  - %SPARK_HOME%\bin
  - %HADOOP_HOME%\bin

Local Set-Up – Step 3: User Variables

- SPARK_PYTHON = C:\Users\nikol\AppData\Local\Programs\Python\Python310\python.exe
- SPARK_DRIVER_PYTHON = C:\Users\nikol\AppData\Local\Programs\Python\Python310\python.exe
- Path:
  - C:\Users\nikol\AppData\Local\Programs\Python\Python310\
  - C:\Users\nikol\AppData\Local\Programs\Python\Python310\Scripts\

Local Set-Up – Step 4: Test in CMD

Open CMD

Run *pyspark*

Run *sdf = spark.range(10)*

Run *sdf.count()*

Result should be 10

Local Set-Up – Step 5: Conda env

- Open CMD
- Run *conda create –n {myenv} python=3.10* (myenv = pyspark)
- Run conda activate pyspark
- Run *pip install pyspark==3.5.0*
- Run *pip install pyspark-extension==2.11.0.3.5, more info here:* G-Research/spark-extension: A library that provides useful extensions to Apache Spark and PySpark. (github.com)

Local Set-Up – Step 6 VSCode Testing

- Clone my repo or download this script: SparkDeltaDatabricksInternals/Lection1 - Architecture/S01E03_SparkSet-Up-Test.ipynb at main · datanikkthegreek/SparkDeltaDatabricksInternals (github.com)
- Open VSCode
- Choose your conda env
- Execute in VS Code