Università di Catania

Dipartimento di Matematica e Informatica
Corso di Big Data
a.a. 2022/2023

Introduzione pratica a Spark

Dott. Roberto Grasso

Informazioni utili

E-mail: roberto.grasso@phd.unict.it

Repository del corso:
https://github.com/RobertoGrasso96/LaboratorioBigDataUniCT

Overview

- Da MapReduce a Spark
- Resilient Distributed Dataset (RDD)
- Persistenza e Caching
- DataFrame
- Metodi di partizionamento
- Machine Learning Library (MLlib)

Ambiente di sviluppo

- Nel corso di queste lezioni useremo Python e la libreria PySpark. Ovviamente siete liberi di utilizzare il linguaggio che preferite (Java, Scala o R).
- Non è necessario installare Spark sui vostri computer! Nel materiale del corso troverete tutto il necessario per configurare l'ambiente di lavoro in maniera totalmente containerizzata.
  Per chi volesse installare Spark: https://spark.apache.org/downloads.html
- Tutto ciò che ci servirà sarà Docker e un editor. Il mio suggerimento è di usare Visual Studio Code. Il motivo? Visual Studio Code offre un buon supporto per Python, i container Docker e i Jupyter Notebook.

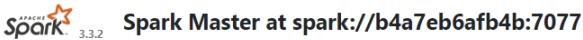
Ambiente di sviluppo

Docker:
https://www.docker.com/

Visual Studio Code:
https://code.visualstudio.com/

- Per avviare i container basta lanciare (dalla root del progetto) il seguente comando:
  `docker compose up -d`
- Per arrestarli:
  `docker compose stop`
- È possibile aumentare il numero di worker, ad esempio a 3, avviando i container con il seguente comando:
  `docker compose up -d --scale spark-worker=3`
- Spark Web UI:
  `http://localhost:8080/`

Spark Web UI



URL: spark://b4a7eb6afb4b:7077
Alive Workers: 3
Cores in use: 3 Total, 0 Used
Memory in use: 3.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (3)

Worker Id	Address	State	Cores	Memory	Resources
worker-20230407144904-172.18.0.2-32985	172.18.0.2:32985	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20230407144905-172.18.0.5-36985	172.18.0.5:36985	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20230407144905-172.18.0.6-43741	172.18.0.6:43741	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	

Running Applications (0)

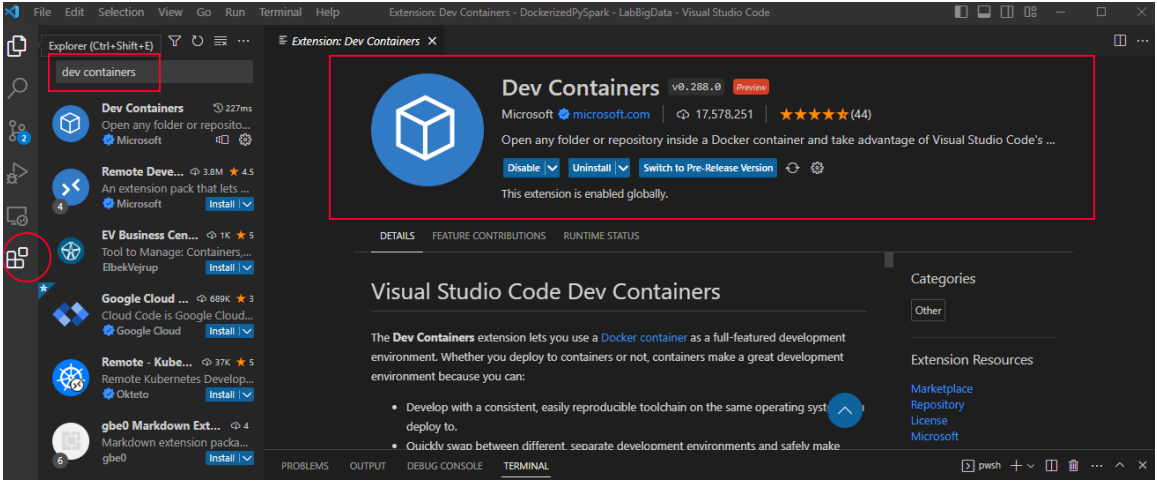
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

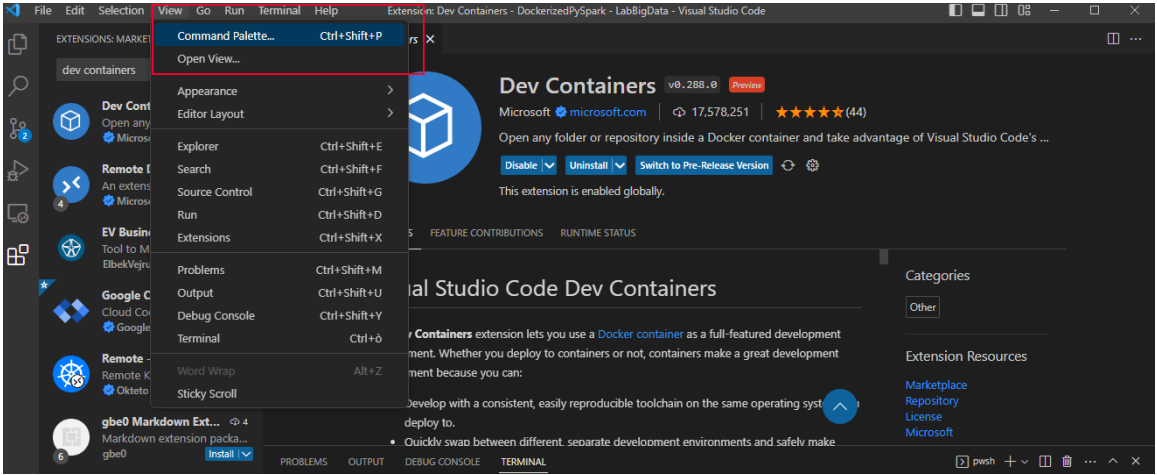
Ambiente di sviluppo

Dopo aver avviato Spark, occupiamoci del container per lo sviluppo.
Aprite il progetto su Visual Studio Code e installate la seguente estensione:



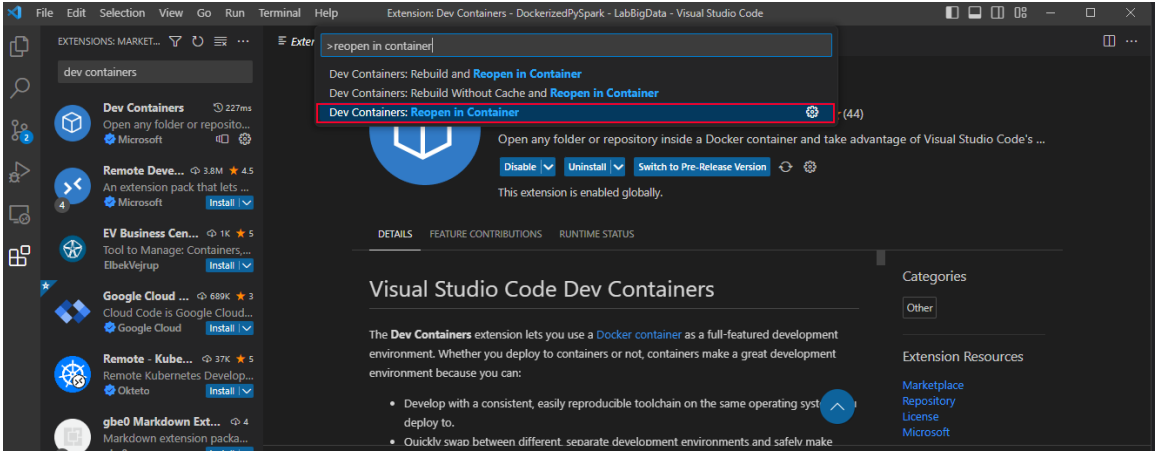
Ambiente di sviluppo

Aprite la command palette:



Ambiente di sviluppo

Digitate *reopen in container*



Ambiente di sviluppo

Alla prima apertura verrano installate e configurate in automatico tutte le estesioni che ci serviranno.
A questo punto abbiamo tutti gli strumenti necessari per poter lavorare!

Link Utili

Documentazione PySpark

URL: https://spark.apache.org/docs/latest/api/python/

QuickStart: DataFrame

URL: https://spark.apache.org/docs/latest/api/python/getting_started/quickstart_df.html

Machine Learning Library in PySpark (DataFrame-based)

URL: https://spark.apache.org/docs/latest/api/python/reference/pyspark.ml.html

Guida MLlib

URL: https://spark.apache.org/docs/latest/ml-guide.html