



# Università di Catania

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

---

*Introduzione pratica a Spark*

---

**Dott. Roberto Grasso**



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



**Repository del corso:**

<https://github.com/RobertoGrasso96/LaboratorioBigDataUniCT>

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



- 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.





## 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 Master at spark://b4a7eb6afb4b:7077

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
<a href="#">worker-20230407144904-172.18.0.2-32985</a>	172.18.0.2:32985	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
<a href="#">worker-20230407144905-172.18.0.5-36985</a>	172.18.0.5:36985	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
<a href="#">worker-20230407144905-172.18.0.6-43741</a>	172.18.0.6:43741	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	

### ▼ Running Applications (0)

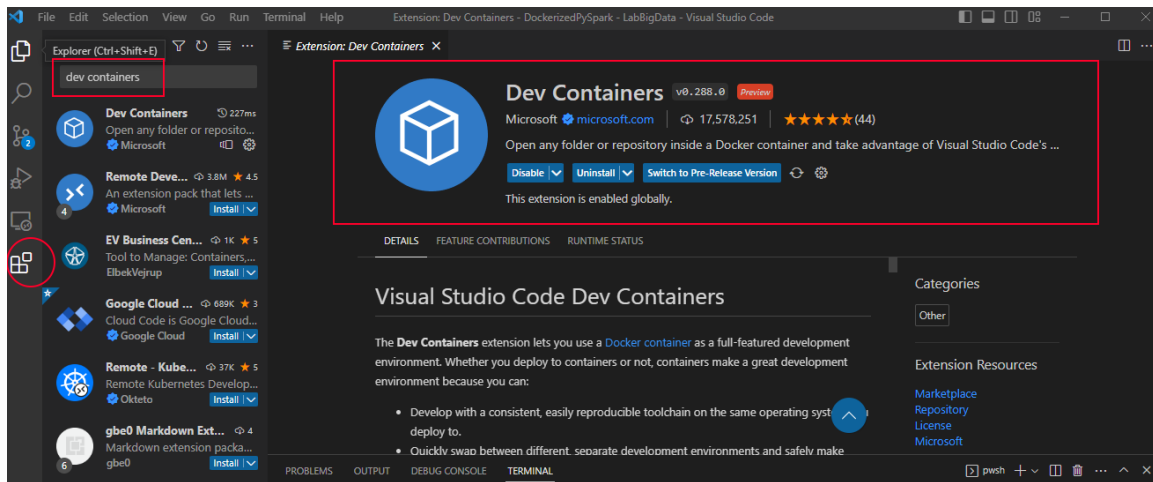
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

### ▼ Completed Applications (0)

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

# Ambiente di sviluppo

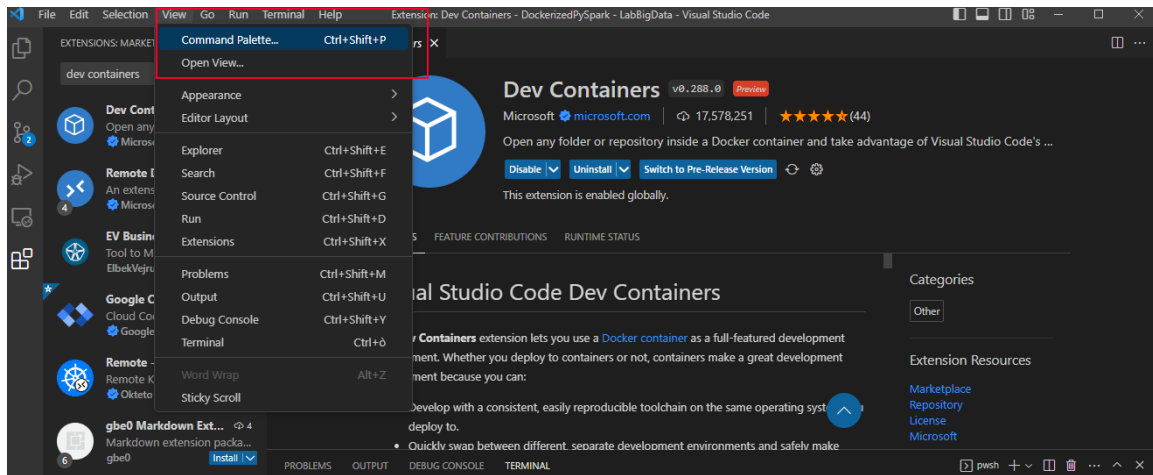
Dopo aver avviato Spark, occupiamoci del container per lo sviluppo.  
Aprirete il progetto su Visual Studio Code e installate la seguente estensione:





# Ambiente di sviluppo

Aprire la command palette:



# Ambiente di sviluppo

## Digitate *reopen in container*

The screenshot shows the Visual Studio Code interface with the 'EXTENSIONS: MARKET...' view on the left. A search for 'dev containers' has been performed. The 'Dev Containers' extension by Microsoft is highlighted. The search results list several options, with 'Dev Containers: Reopen in Container' selected. The details panel for this extension is visible, showing its description and installation status. The extension is described as a tool to open any folder or repository inside a Docker container. It is currently installed and enabled globally. The details panel also includes links to 'Disable', 'Uninstall', and 'Switch to Pre-Release Version'. The 'Visual Studio Code Dev Containers' section provides a brief overview of the extension's purpose and benefits.

File Edit Selection View Go Run Terminal Help Extension: Dev Containers - DockerizedPySpark - LabBigData - Visual Studio Code

EXTENSIONS: MARKET... dev containers

**Dev Containers** 227ms  
Open any folder or repository inside a Docker container and take advantage of Visual Studio Code's ...  
Microsoft

**Remote Development** 3.8M 4.5  
An extension pack that lets you develop in a remote environment.  
Microsoft

**EV Business Center** 1K 5  
Tool to Manage: Containers, Docker, Kubernetes, and more.  
ElbekVejrup

**Google Cloud SDK** 689K 3  
Cloud Code is Google Cloud's integrated development environment for Google Cloud.  
Google Cloud

**Remote - Kubernetes** 37K 5  
Remote Kubernetes Development Environment  
Okteto

**gbe0 Markdown Extension** 4  
Markdown extension pack for Visual Studio Code  
gbe0

**Dev Containers: Reopen in Container** (44)  
Open any folder or repository inside a Docker container and take advantage of Visual Studio Code's ...  
Disable Uninstall Switch to Pre-Release Version  
This extension is enabled globally.

DETAILS FEATURE CONTRIBUTIONS RUNTIME STATUS

### Visual Studio Code Dev Containers

The **Dev Containers** extension lets you use a **Docker container** as a full-featured development environment. Whether you deploy to containers or not, containers make a great development environment because you can:

- Develop with a consistent, easily reproducible toolchain on the same operating system and deploy to.
- Quickly swap between different, separate development environments and safely make

Categories  
Other

Extension Resources  
Marketplace  
Repository  
License  
Microsoft

Alla prima apertura verranno installate e configurate in automatico tutte le estensioni che ci serviranno.

A questo punto abbiamo tutti gli strumenti necessari per poter lavorare!





## 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](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>