

Introduzione al calcolo parallelo con Apache Spark



Vincenzo Manzoni, Andrea Rota
Analytics on Industrial Processes

Leader mondiale nella produzione e fornitura di prodotti tubolari e servizi per:

- Trivellazioni, estrazione e produzione di petrolio e gas
- Trasporto di petrolio e gas
- Impianti di trasformazione e centrali elettriche
- Applicazioni specialistiche industriali e automotive



OCTG



**Giunti
Premium**



**Linepipe per
applicazioni
onshore e
offshore**



**Trasformazione
di idrocarburi**

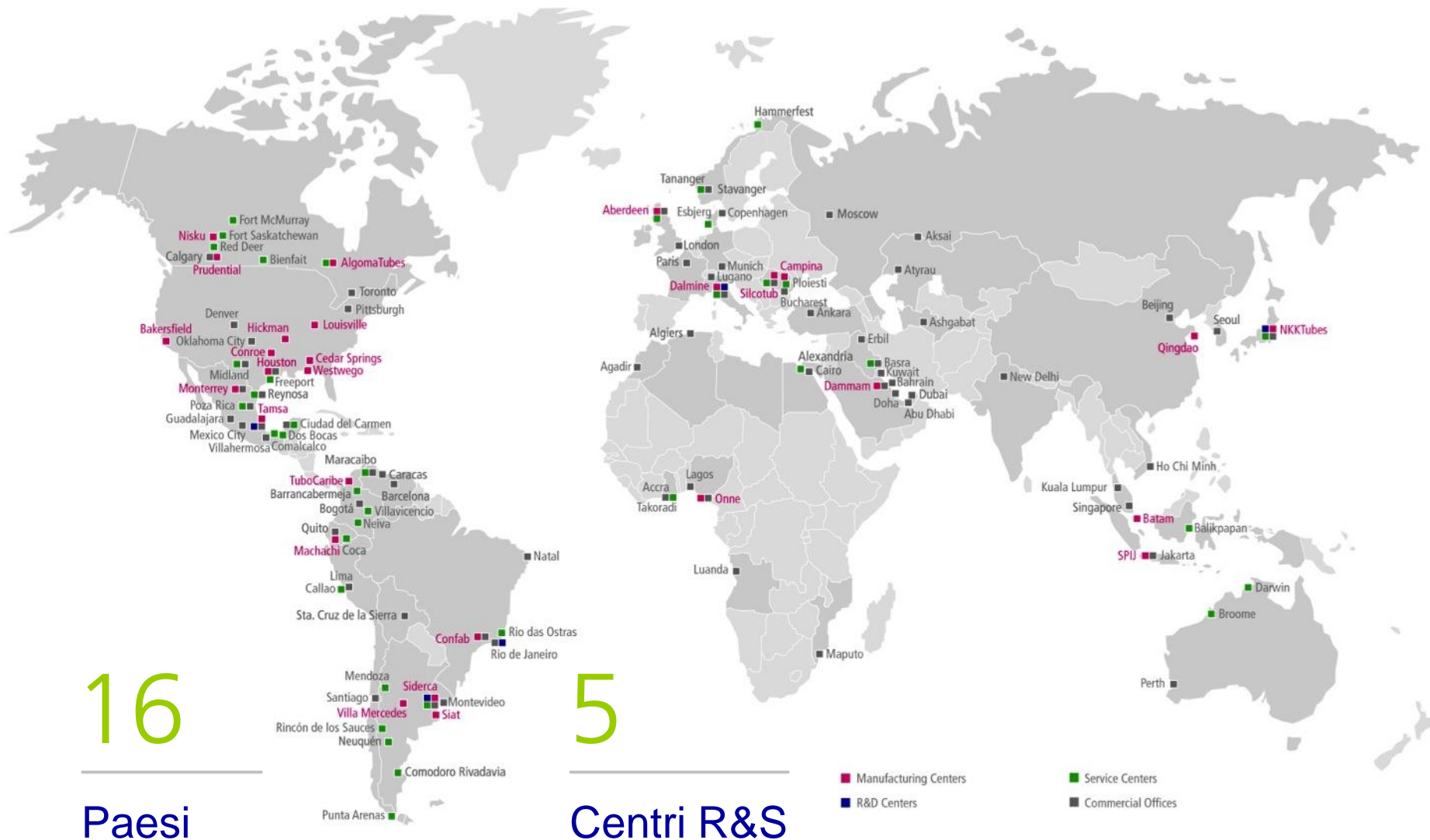


**Generazione di
energia**

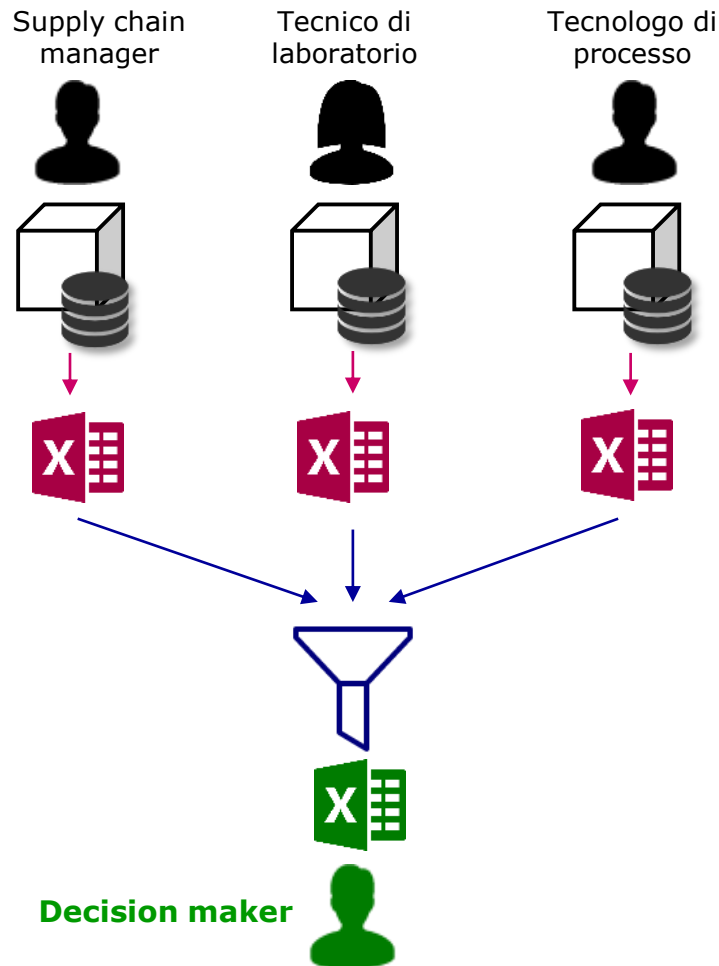


**Applicazioni
industriali e
automotive**

Tenaris nel mondo



Background



Data users

Competenze diverse (tipologia e livello di maturità)
Esigenze diverse

Applicazioni di data analytics

Verticali
Sviluppate custom
Tecnologie eterogenee
Non *mobile-ready*

Export

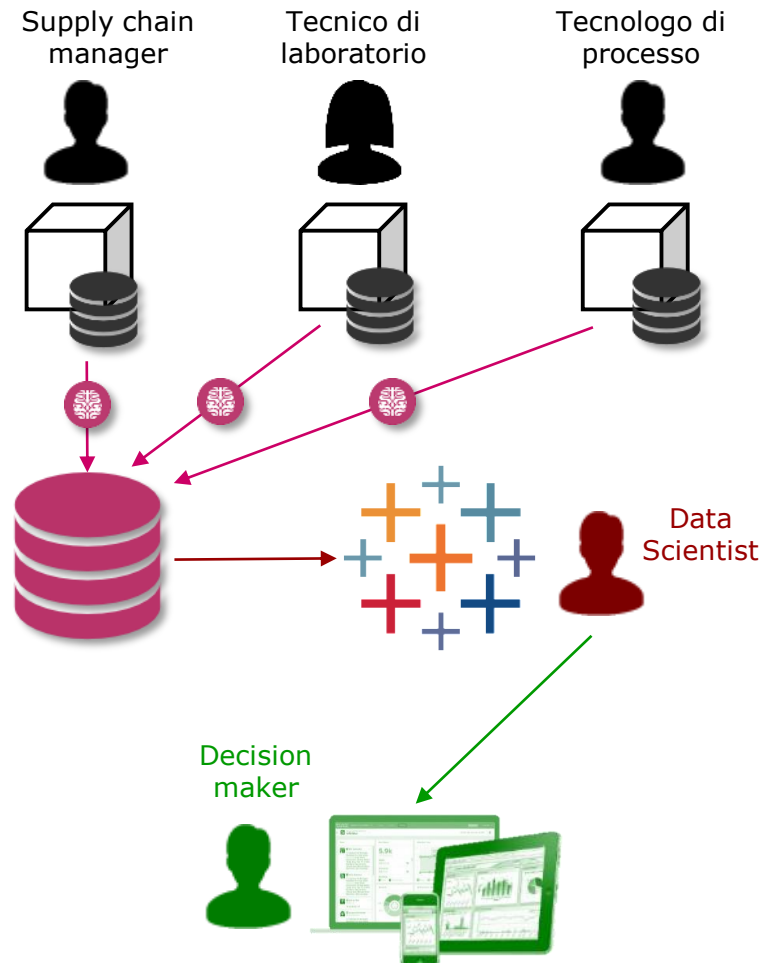
Report strutturati e non strutturati
Diversi formati
Autorizzazione su ogni sistema
Tempo di download

Merge

Operazione manuale → *error prone*
Time consuming
Non real-time
Qualità funzione delle capacità dell'utente



Implementazione Big Data in Tenaris



Perché Big Data



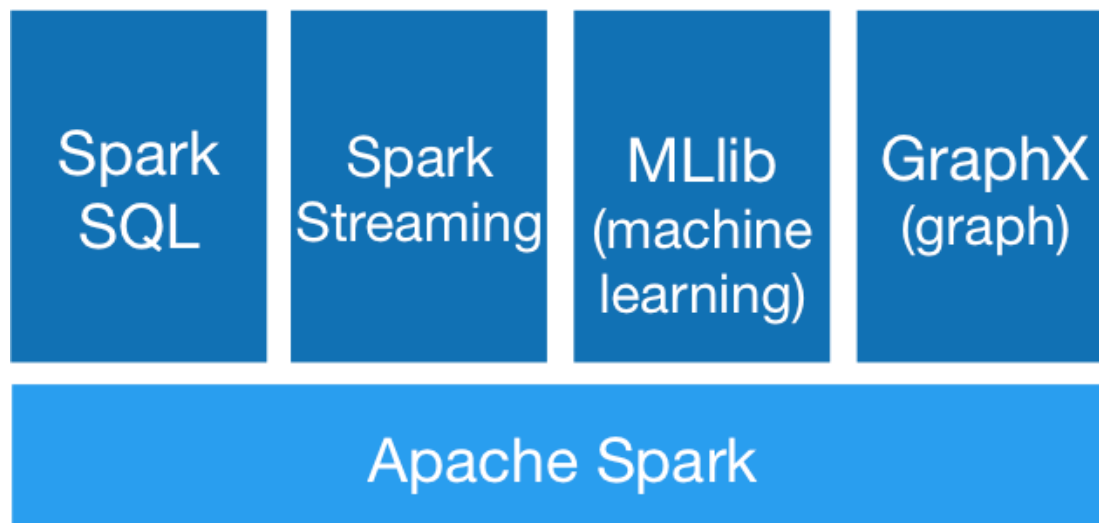
- La **velocità** con cui vengono prodotti i dati cresce più velocemente della capacità di una singola macchina per elaborarli.
- Il **volume** dei dati cresce più velocemente della capacità di storage di una singola macchina.

**Serve una tecnologia per esprimere
ad alto livello analisi in parallelo su
grandi quantità di dati.**

Framework open-source sviluppato dall'AMPLab dell'Università di Berkeley.

A differenza di paradigmi precedenti – tipo Map-Reduce – Spark usa primitive in memoria (fino a 100x in termini di prestazioni).

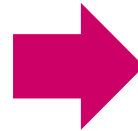
Supporta diversi linguaggi di programmazione, tra cui Java, Scala e Python.



Contare le parole di un testo



«I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham
...»



Parola	Occorrenze
I	3
Am	3
Sam	3
Do	1
You	1
Like	1
...	...

Contare le parole di un testo



«I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham
...»

Parola	Occorrenze
I	1

Contare le parole di un testo



«I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham
...»

Parola	Occorrenze
I	1
Am	1

Contare le parole di un testo



«I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham
...»

Parola	Occorrenze
I	1
Am	1
Sam	1

Contare le parole di un testo



«I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham
...»

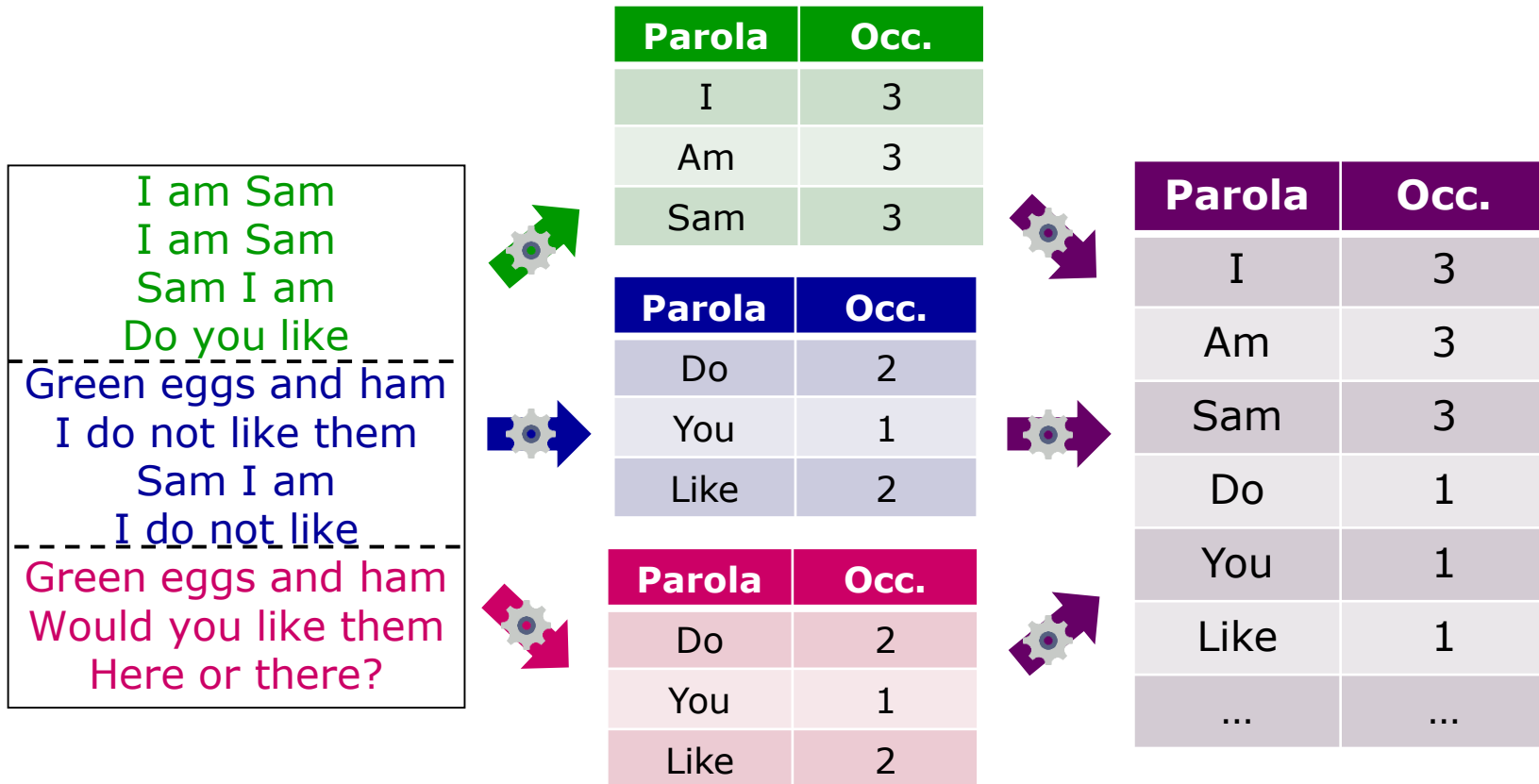
Parola	Occorrenze
I	2
Am	1
Sam	1

Contare le parole di un testo (molto) lungo (1/3)



I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham
I do not like them
Sam I am
I do not like
Green eggs and ham
Would you like them
Here or there?

Contare le parole di un testo (molto) lungo (2/3)



Contare le parole di un testo (molto) lungo (3/3)

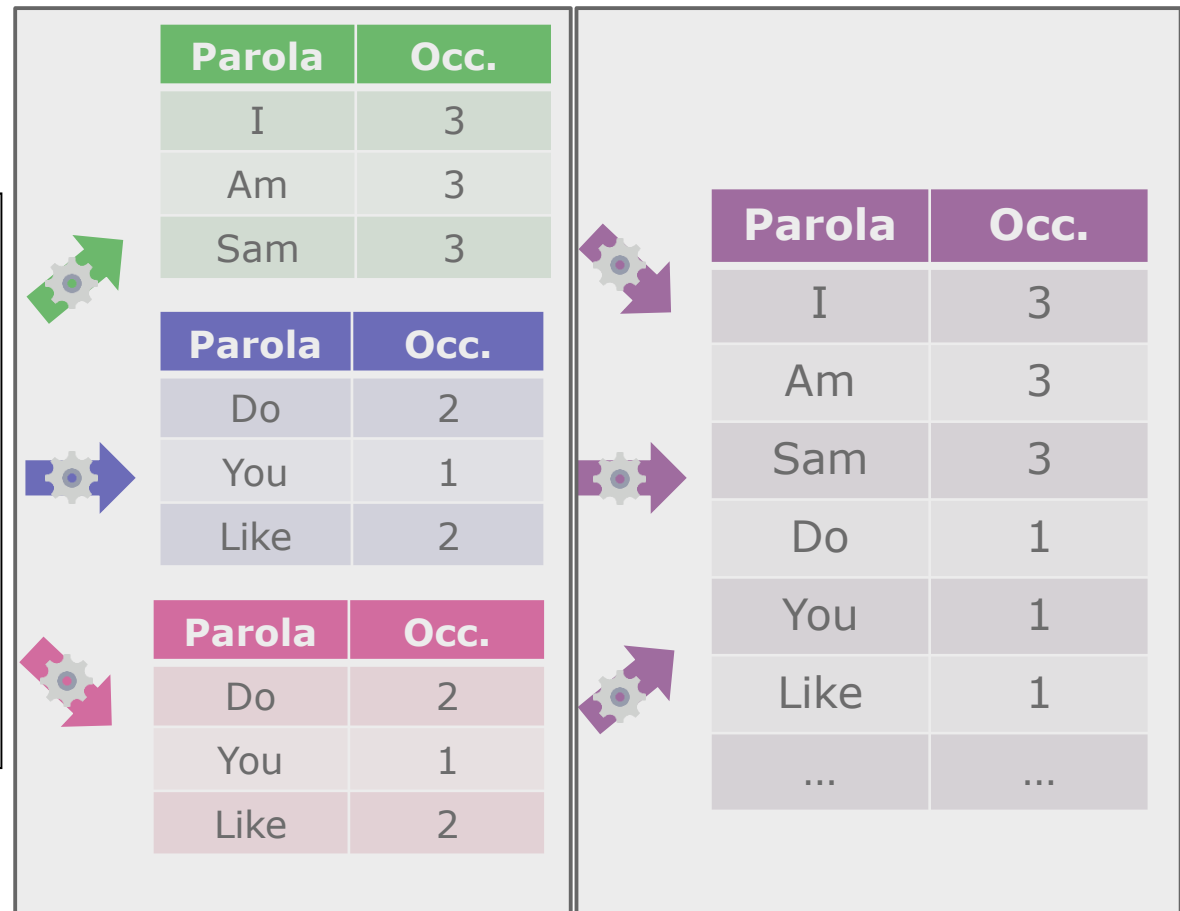


I am Sam
I am Sam
Sam I am
Do you like

Green eggs and ham
I do not like them
Sam I am

I do not like

Green eggs and ham
Would you like them
Here or there?



Map

Reduce

Demo time!



<https://github.com/tenaris/scala-spark-workshop>

Setup Databricks



<https://databricks.com/try-databricks>

The screenshot shows the Databricks website's 'try-databricks' page. At the top, there is a navigation bar with links for Blog, Resources, Partners, Documentation, Support, Careers, Contact Us, and a search icon. Below this is a secondary navigation bar with the Databricks logo, links for WHY DATABRICKS, PRODUCT, APACHE SPARK, SOLUTIONS, CUSTOMERS, TRAINING, and a prominent 'TRY DATABRICKS' button. The main content area features the heading 'Select a version to get started.' followed by two columns. The left column is titled 'FULL-PLATFORM TRIAL' with the subtext 'Put Apache Spark to work'. It lists features: Unlimited clusters, Notebooks, dashboards, production jobs, RESTful APIs, Interactive guide to Spark and Databricks, Deployed to your AWS VPC, BI tools integration, and a 14-day free trial (excluding AWS charges). Below this is a 'START TODAY' button. The right column is titled 'COMMUNITY EDITION' with the subtext 'Learn Apache Spark'. It lists features: Mini 6GB cluster, Interactive notebooks and dashboards, and a Public environment to share your work. Below this is a 'START TODAY' button. A large pink arrow points from the 'COMMUNITY EDITION' section towards the bottom. At the very bottom, there is a small banner that says 'Stay up to date on Apache Spark.' with a close icon.

Blog Resources Partners Documentation Support Careers Contact Us Manage Account

databricks WHY DATABRICKS PRODUCT APACHE SPARK SOLUTIONS CUSTOMERS TRAINING TRY DATABRICKS

Select a version to get started.

FULL-PLATFORM TRIAL
Put Apache Spark to work

- Unlimited clusters
- Notebooks, dashboards, production jobs, RESTful APIs
- Interactive guide to Spark and Databricks
- Deployed to your AWS VPC
- BI tools integration
- 14-day free trial (excludes AWS charges)

START TODAY

COMMUNITY EDITION
Learn Apache Spark

- Mini 6GB cluster
- Interactive notebooks and dashboards
- Public environment to share your work

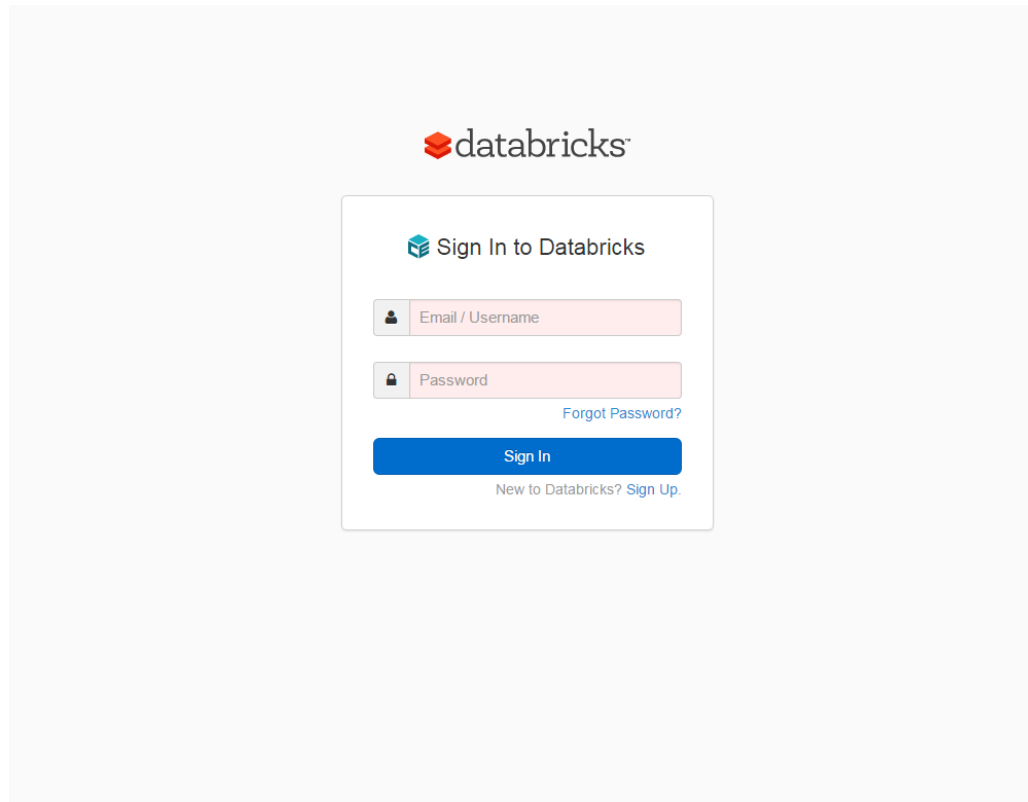
START TODAY

Stay up to date on Apache Spark. X

Login databricks



<https://community.cloud.databricks.com/>



The screenshot shows the Databricks login interface. At the top is the Databricks logo. Below it is a 'Sign In to Databricks' section. This section contains two input fields: 'Email / Username' and 'Password'. Below the password field is a link for 'Forgot Password?'. A blue 'Sign In' button is positioned below the input fields. At the bottom of the sign-in section is a link for 'New to Databricks? Sign Up.'.

Join the dark side!



Siamo alla ricerca di studenti del 5 anno per tesi, tirocini e assegni di ricerca su tematiche **Big Data**, **Machine Learning** e **Artificial Intelligence**.

Per candidarvi, scrivete una mail a:

Vincenzo Manzoni
vmanzoni@tenaris.com

