Introduzione al calcolo parallelo con Apache Spark

Vincenzo Manzoni, Andrea Rota
Data Science

Tenaris

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

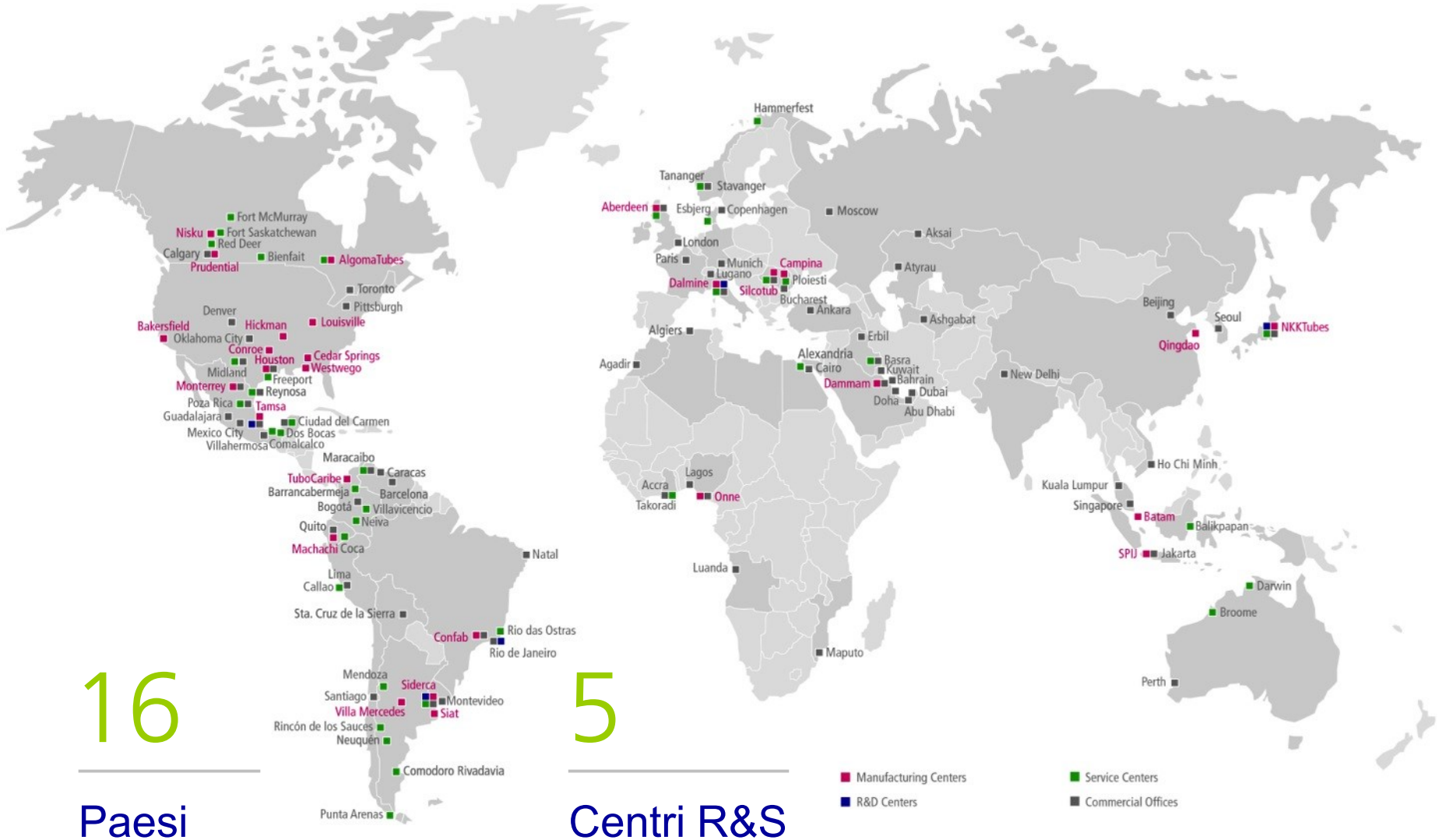
Linepipe per applicazioni onshore e offshore

Trasformazione di idrocarburi

Generazione di energia

Applicazioni industriali e automotive

Tenaris nel mondo



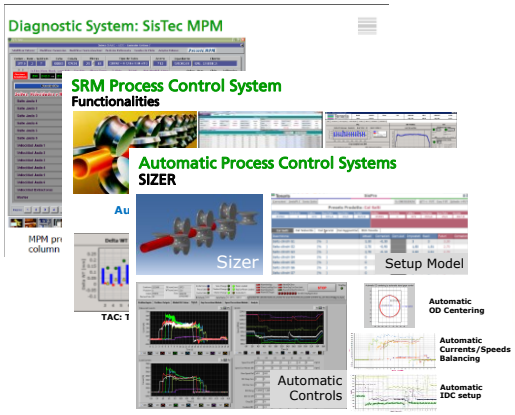
16

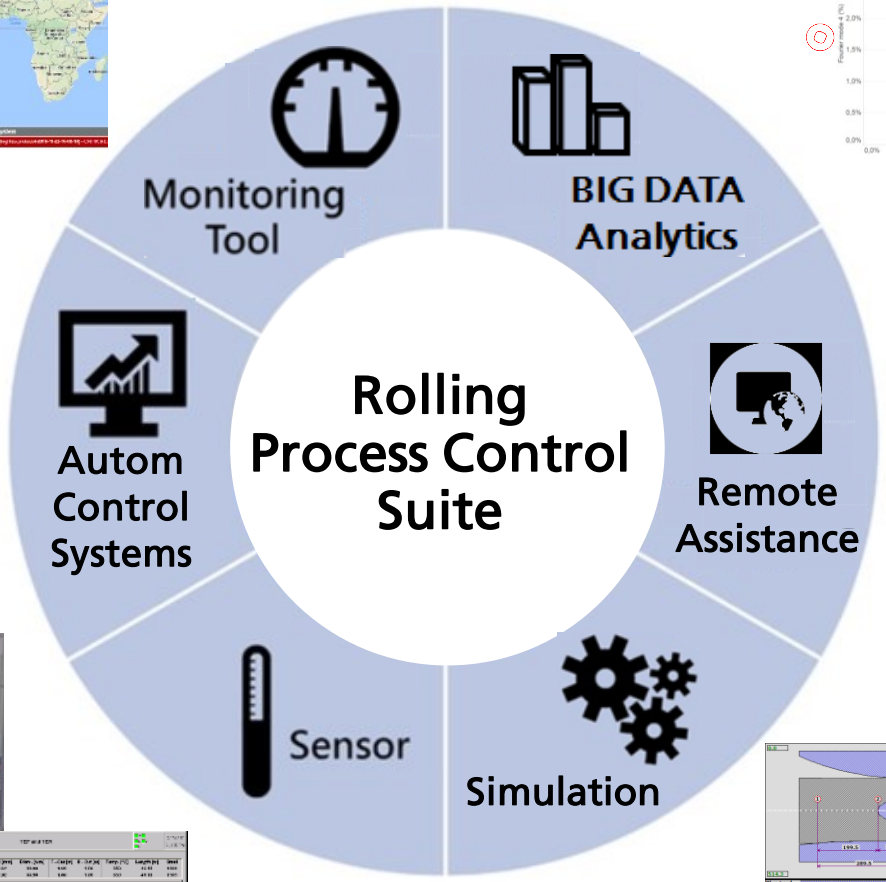
Paesi

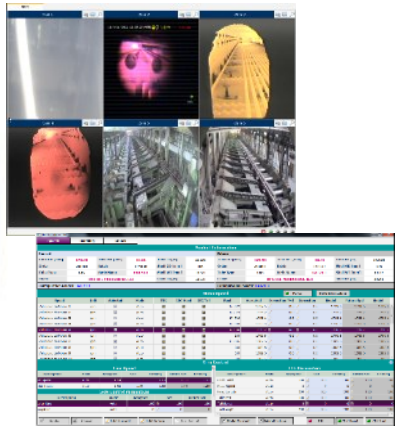
5

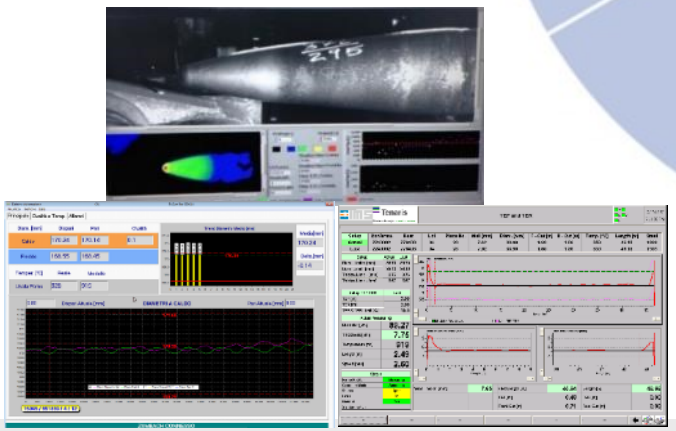
Centri R&S

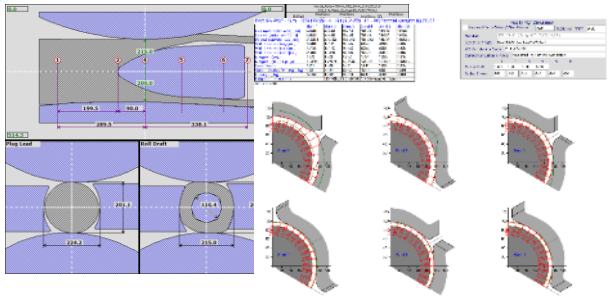
Industry 4.0 topics in Tenaris



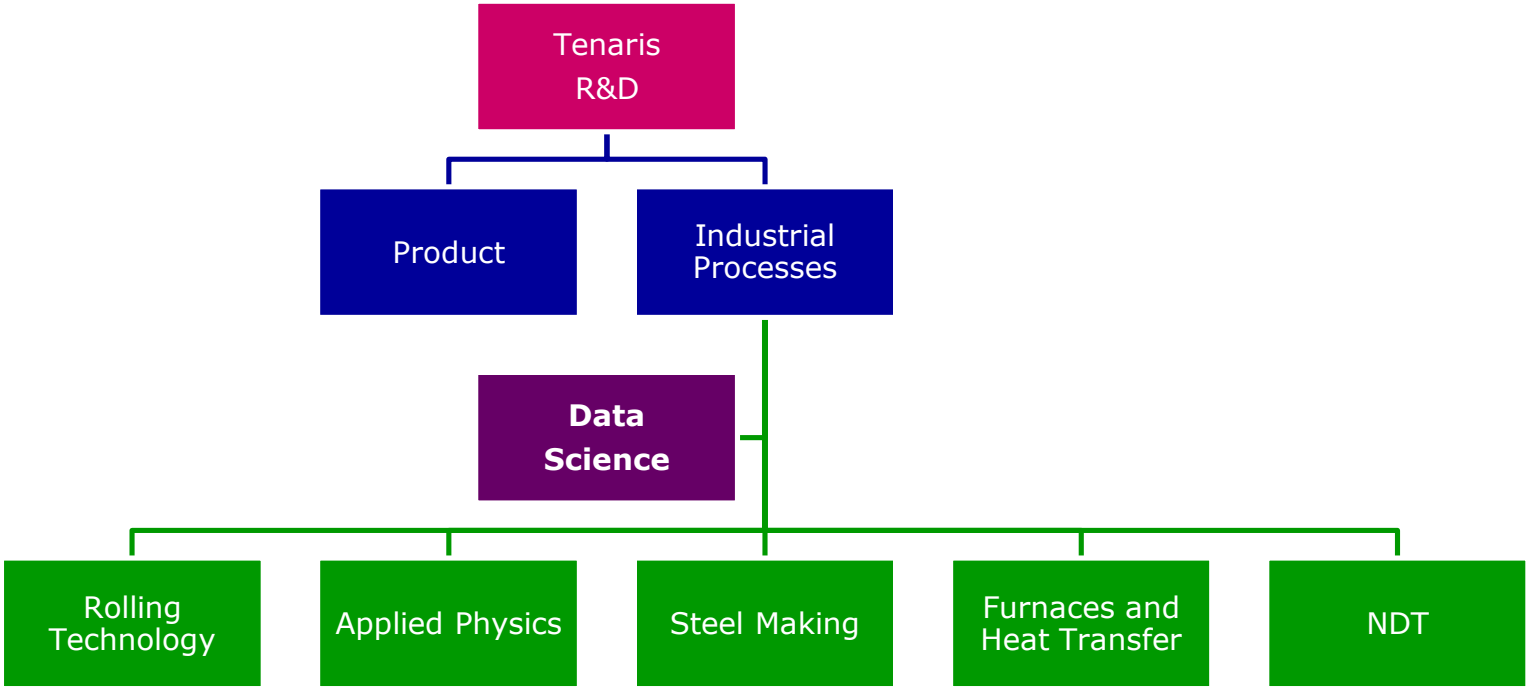








Chi siamo


Tenaris
R&D
Product
Industrial
Processes
Data
Science
Rolling
Technology
Applied Physics
Steel Making
Furnaces and
Heat Transfer
NDT

Big Data and AI examples

Massive Data Analysis

Offline applications for **batch** decision making.

Process Monitoring

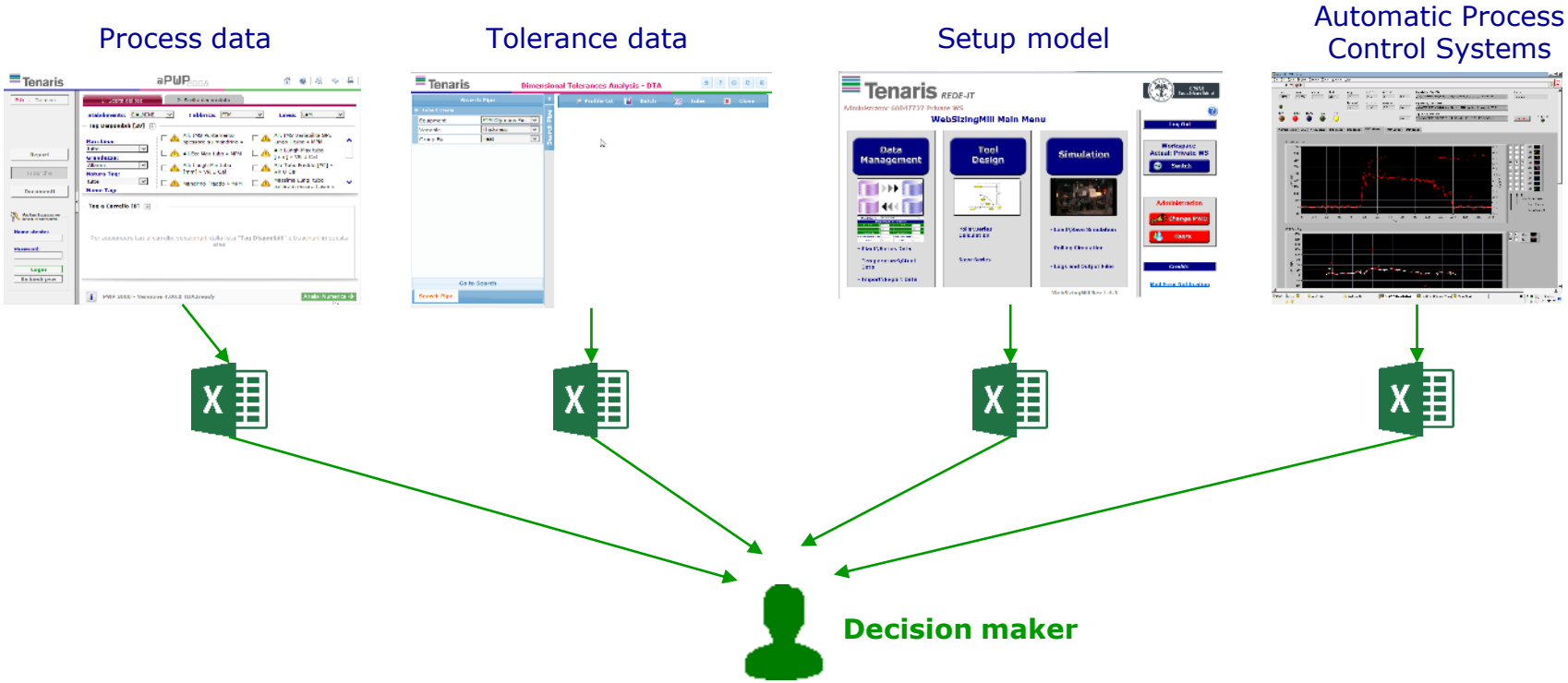
Streaming application for **real-time** decision making.

Artificial Intelligence and Machine Learning Modeling

Model Industrial Processes behavior from data through Artificial Intelligence and Machine Learning algorithms.

MASSIVE DATA ANALYSIS

Processo decisionale tradizionale

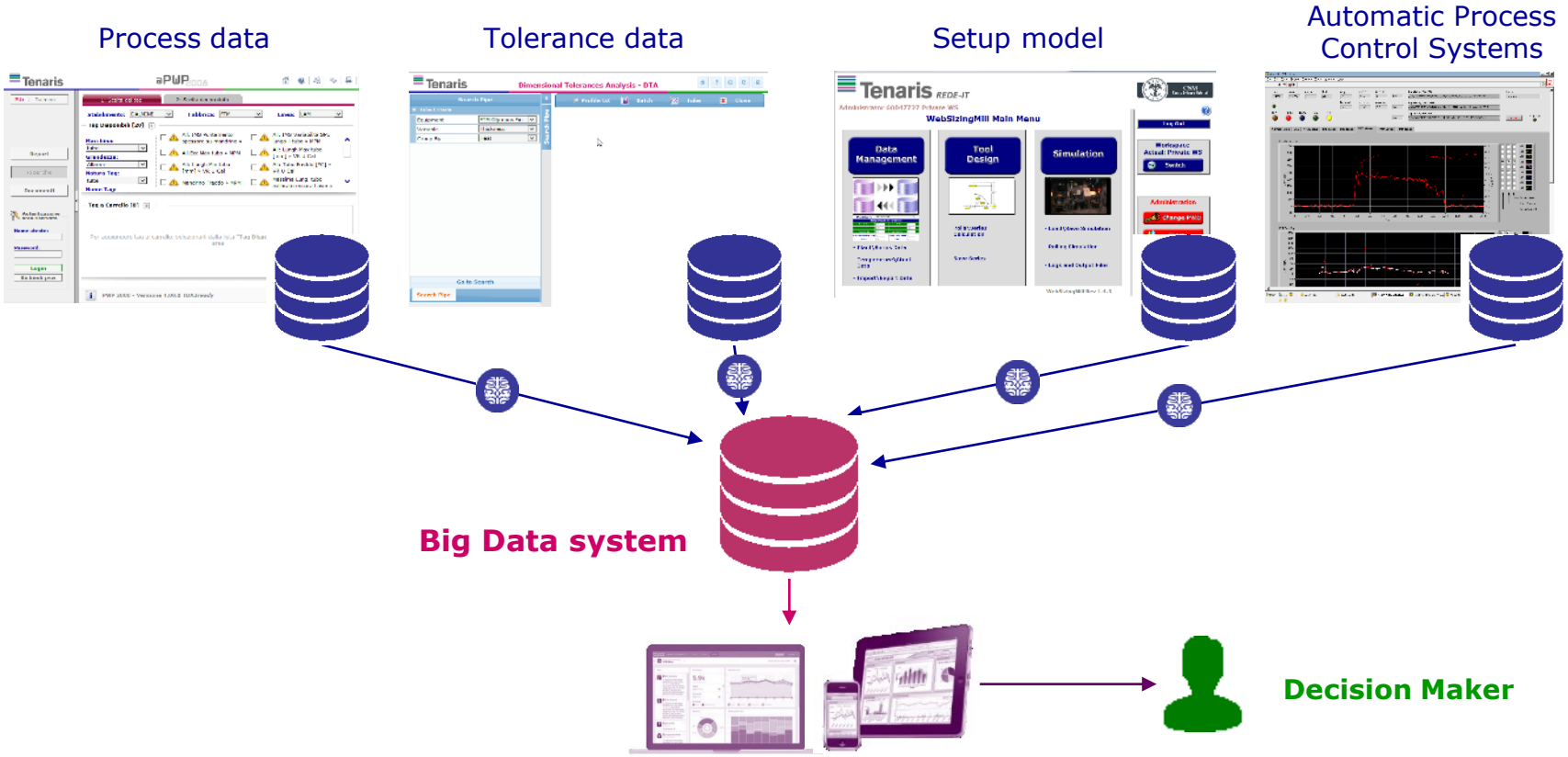


85% of the time goes in data preparation

15% of the time goes in data analysis and decision making

Munson, *"A Study on the Importance of and Time Spent on Different Modeling Steps"* (link)

Processo decisionala supportato da tecnologie Big Data



Data Analysis and Decision making

100% of the time goes in data analysis and decision making

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.

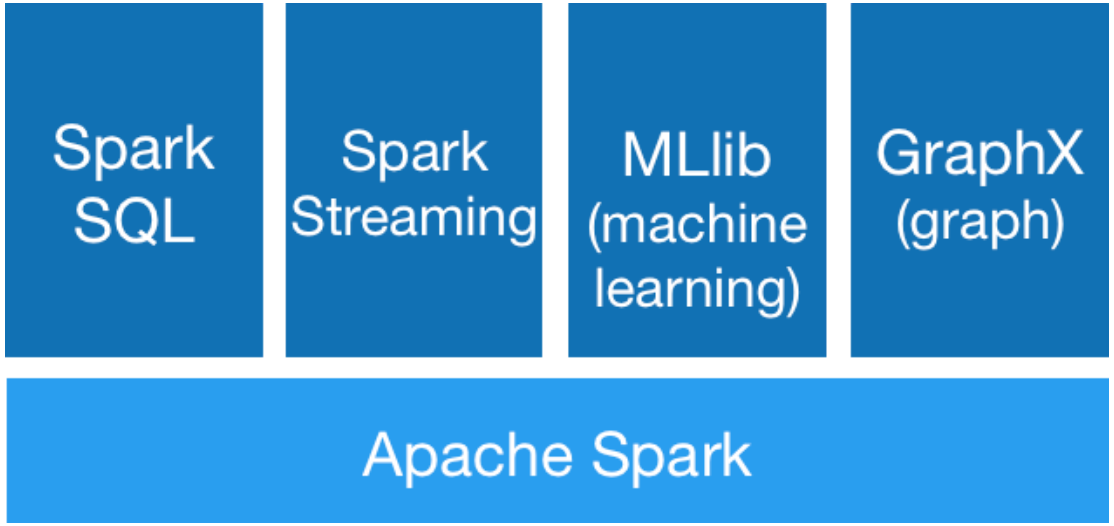
Serve una tecnologia per esprimere ad alto livello analisi in parallelo su grandi quantità di dati.

Apach Spark

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)

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?

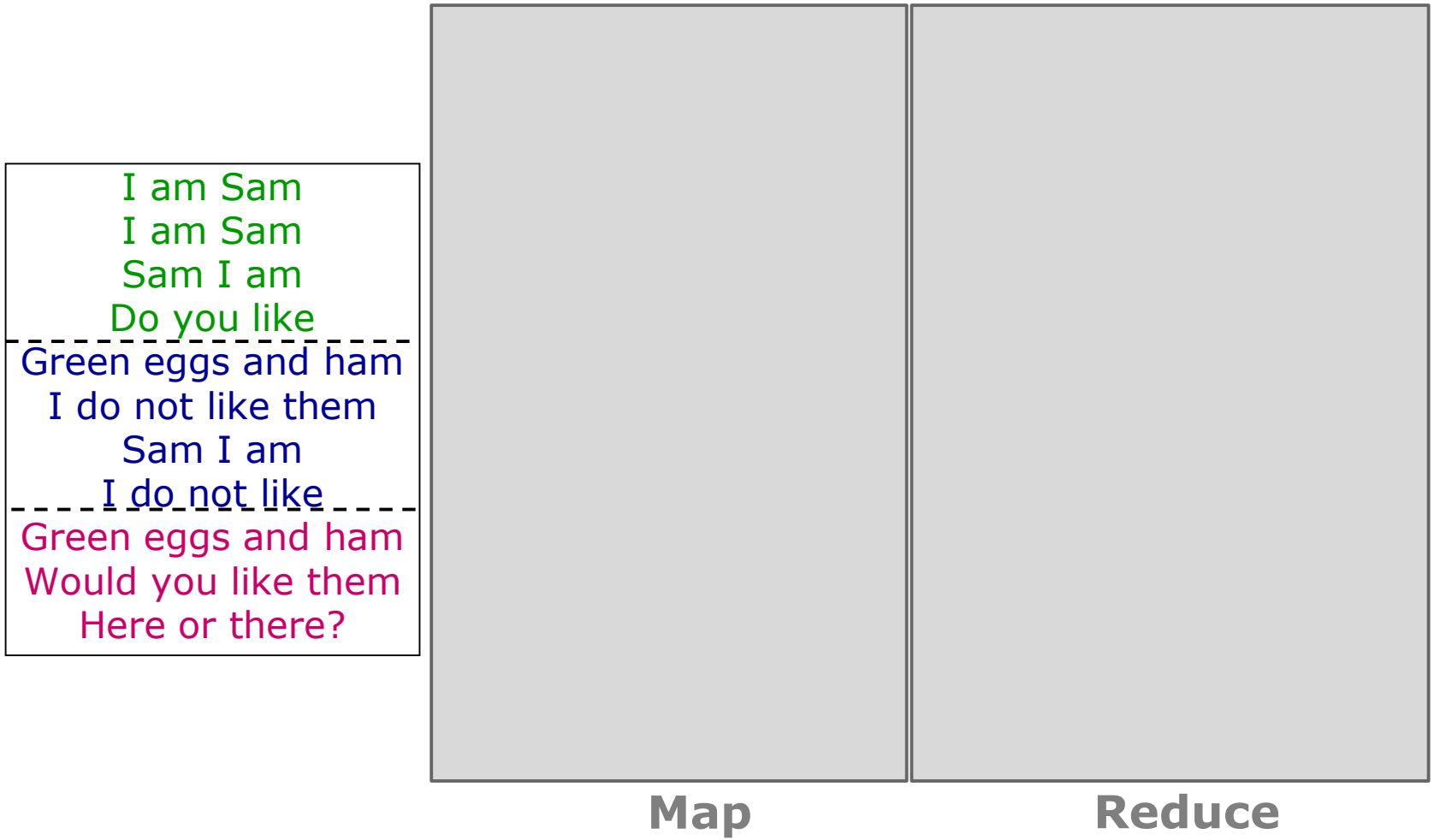
Parola	Occ.
I	3
Am	3
Sam	3

Parola	Occ.
Do	2
You	1
Like	2

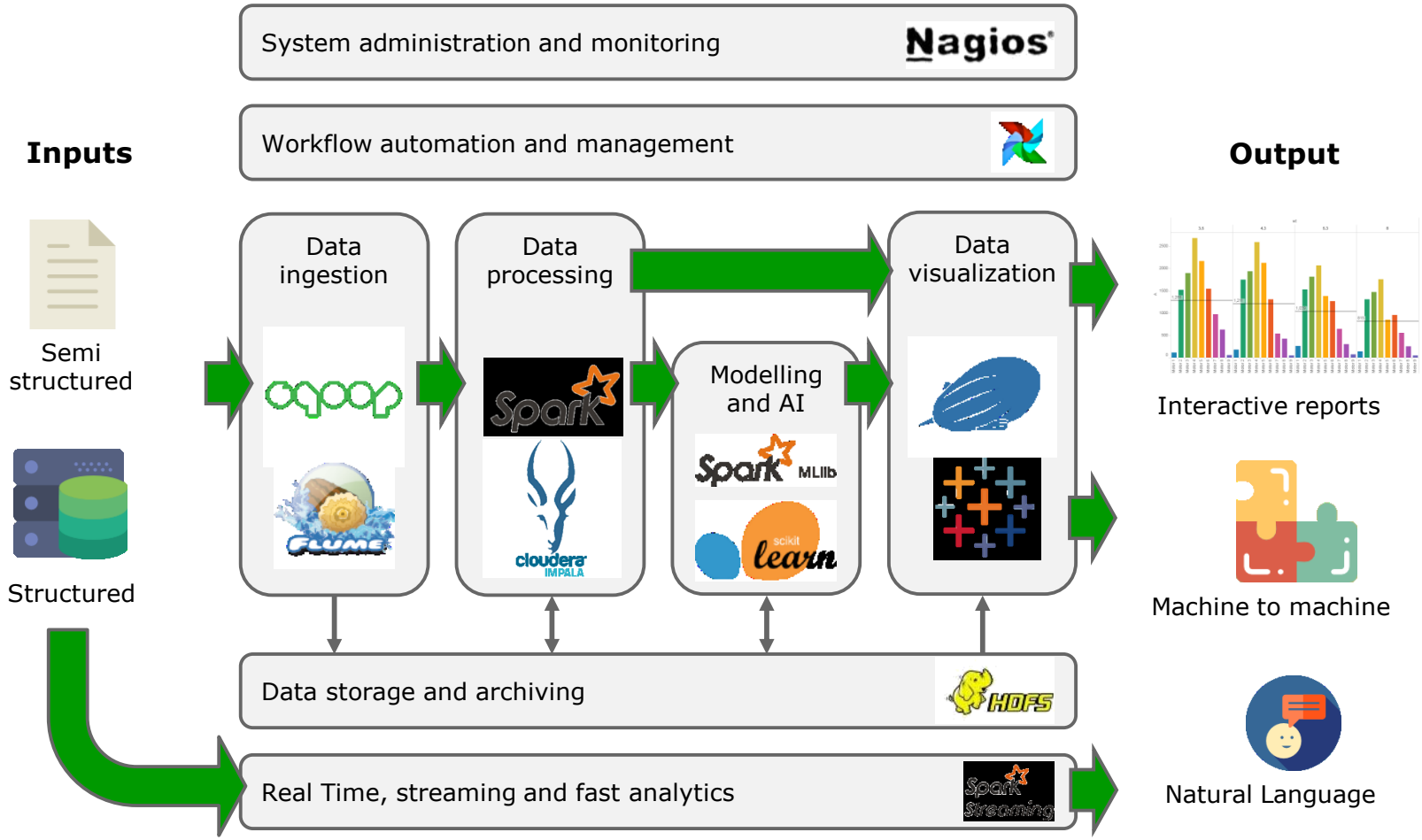
Parola	Occ.
Do	2
You	1
Like	2

Parola	Occ.
I	3
Am	3
Sam	3
Do	1
You	1
Like	1
…	…

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



Tenaris Big Data technology stack



Tenaris Big Data technologies

October 17, 2016 - Apache Flume 1.7.0 Released

The Apache Flume team is pleased to announce the release of Flume 1.7.0.

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amount

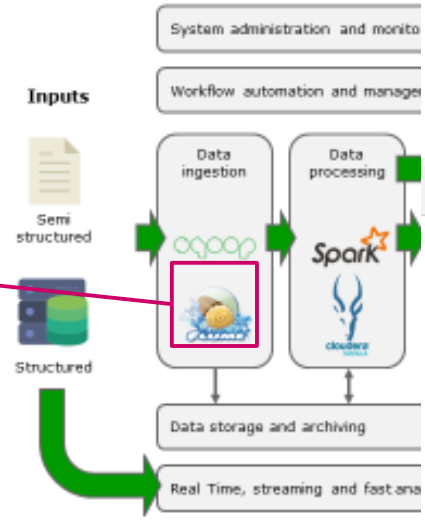
Version 1.7.0 is the tenth Flume release as an Apache top-level project. Flume 1.7.0 is stable, production-ready sofi versions of the Flume 1.x codeline.

Several months of active development went into this release: almost 100 patches were committed since 1.6.0, represe While the full change log can be found on the 1.7.0 release page (link below), here are a few new feature highlights:

- Taildir source
- Kafka integration improvements (eg. security)

Below is the list of people (from Git/SVN logs) who submitted and/or reviewed improvements to Flume during the 1.7.0 development cycle:

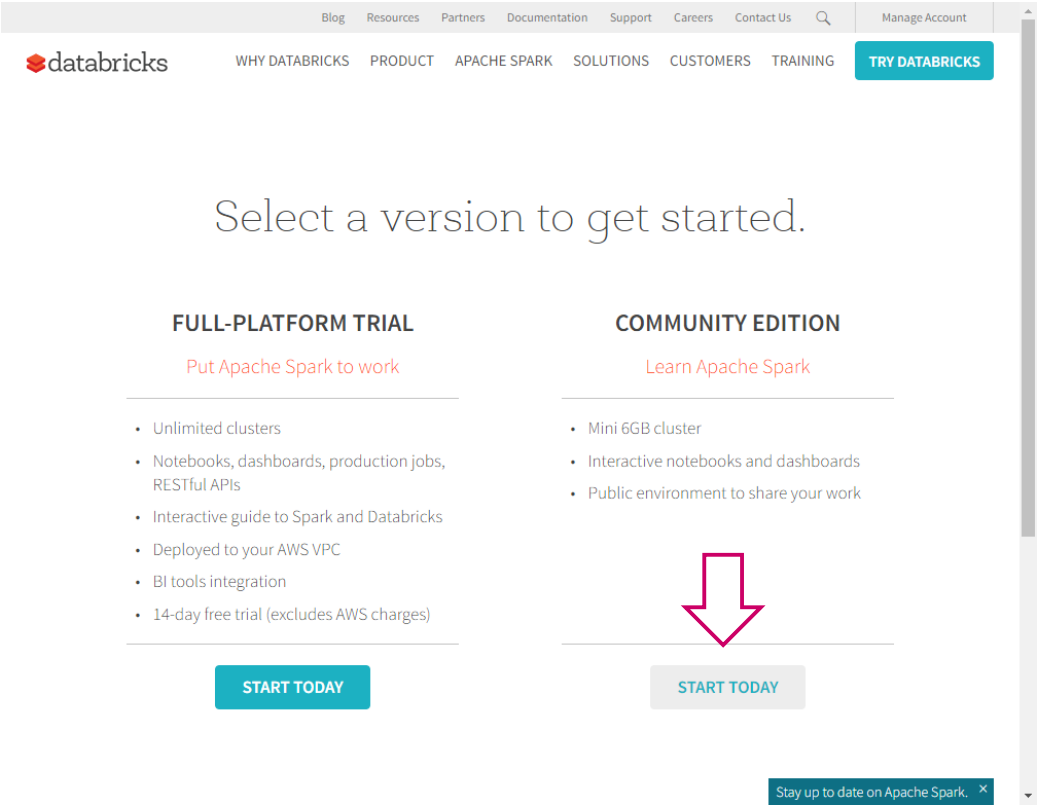
- Abraham Fine
- Alexandre Dutra
- Andrea Rota
- Ashish Paliwal
- Attila Simon
- Bessenyei Balázs Donát
- Daniel Templeton
- Deepesh Khandelwal



DEMO TIME!

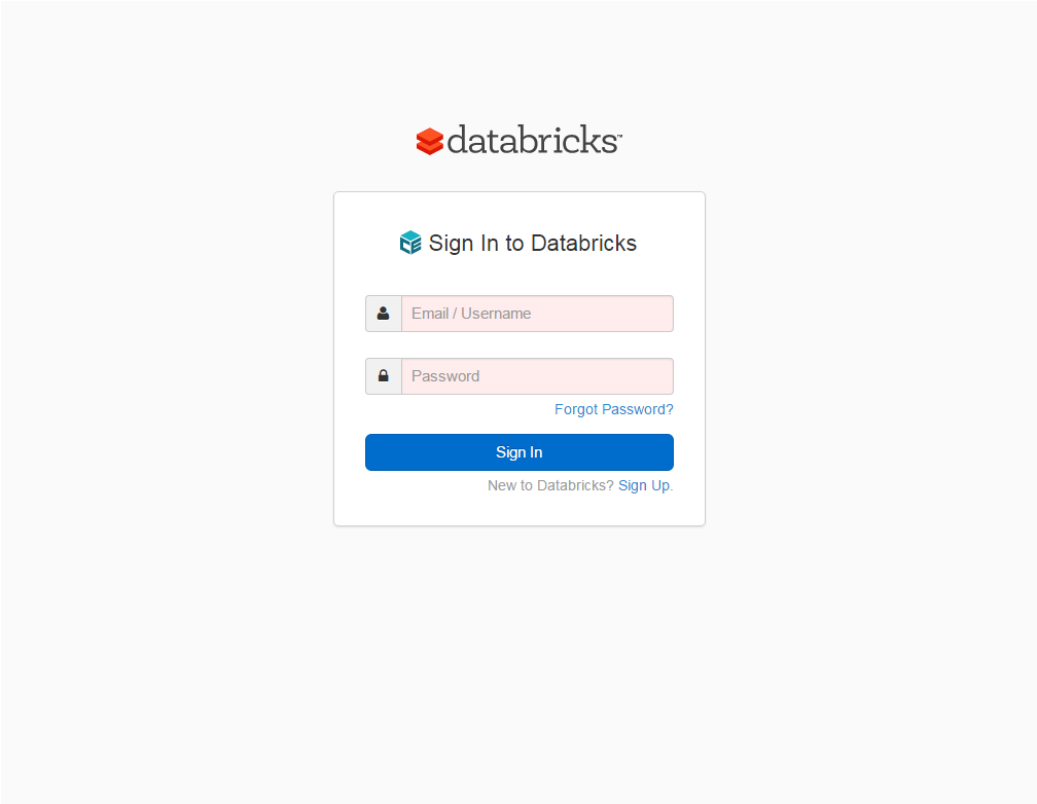
Setup Databricks

https://databricks.com/try-databricks



Login databricks

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



GitHub repository

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

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

Andrea Rota
arota@tenaris.com