

Parallel computing with Apache Spark



Andrea Rota, Michele Zanchi
Tenaris Data Science

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**

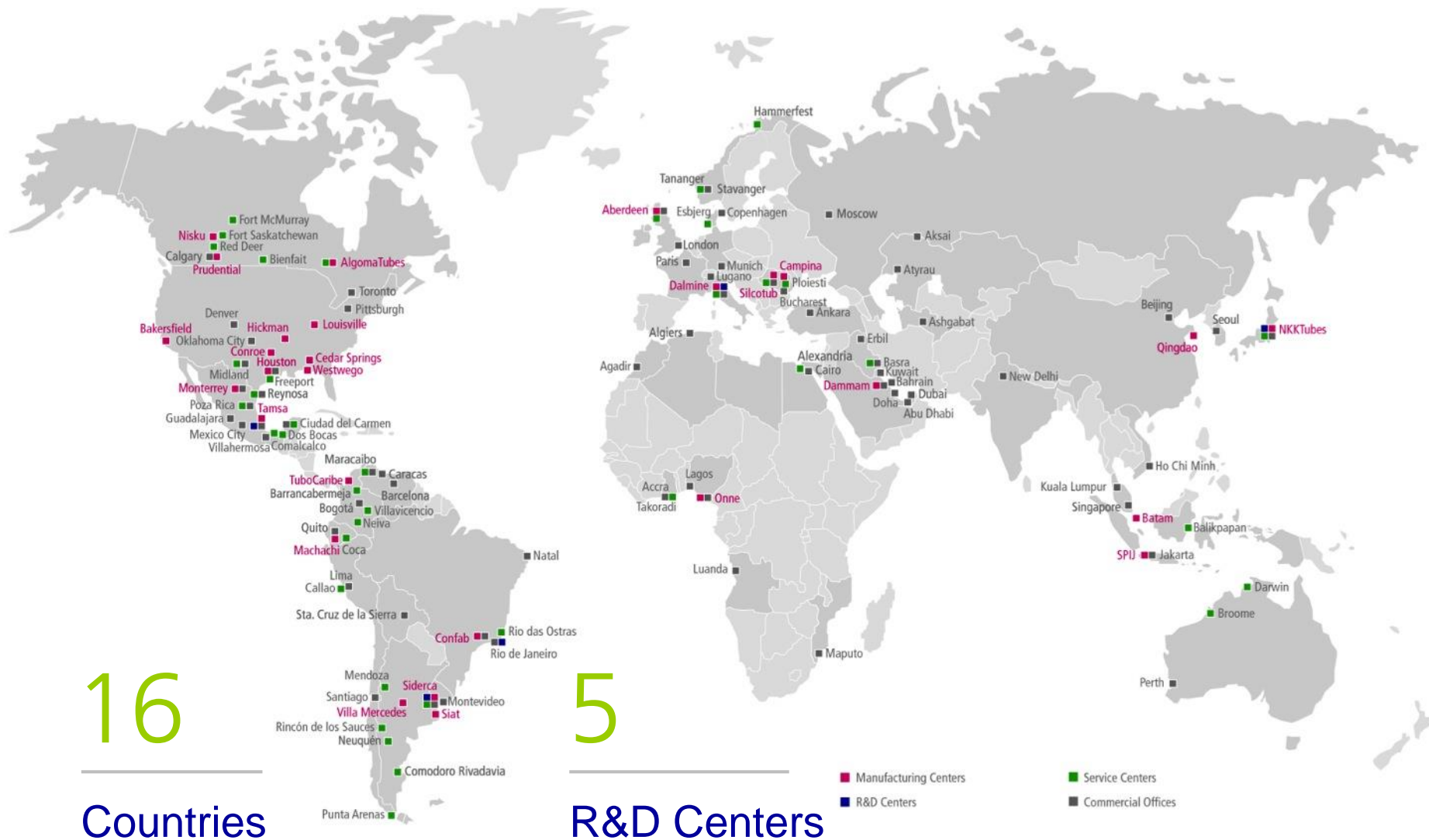


**Generazione di
energia**



**Applicazioni
industriali e
automotive**

Tenaris





Tenaris Data Science Department aims to extract value from company's data, using a scientific approach, also known as Data Science, and to apply Big Data technologies to the industrial field.

[illegible]

Needs addressed by the department



"I have a big amount of data spread across different databases.

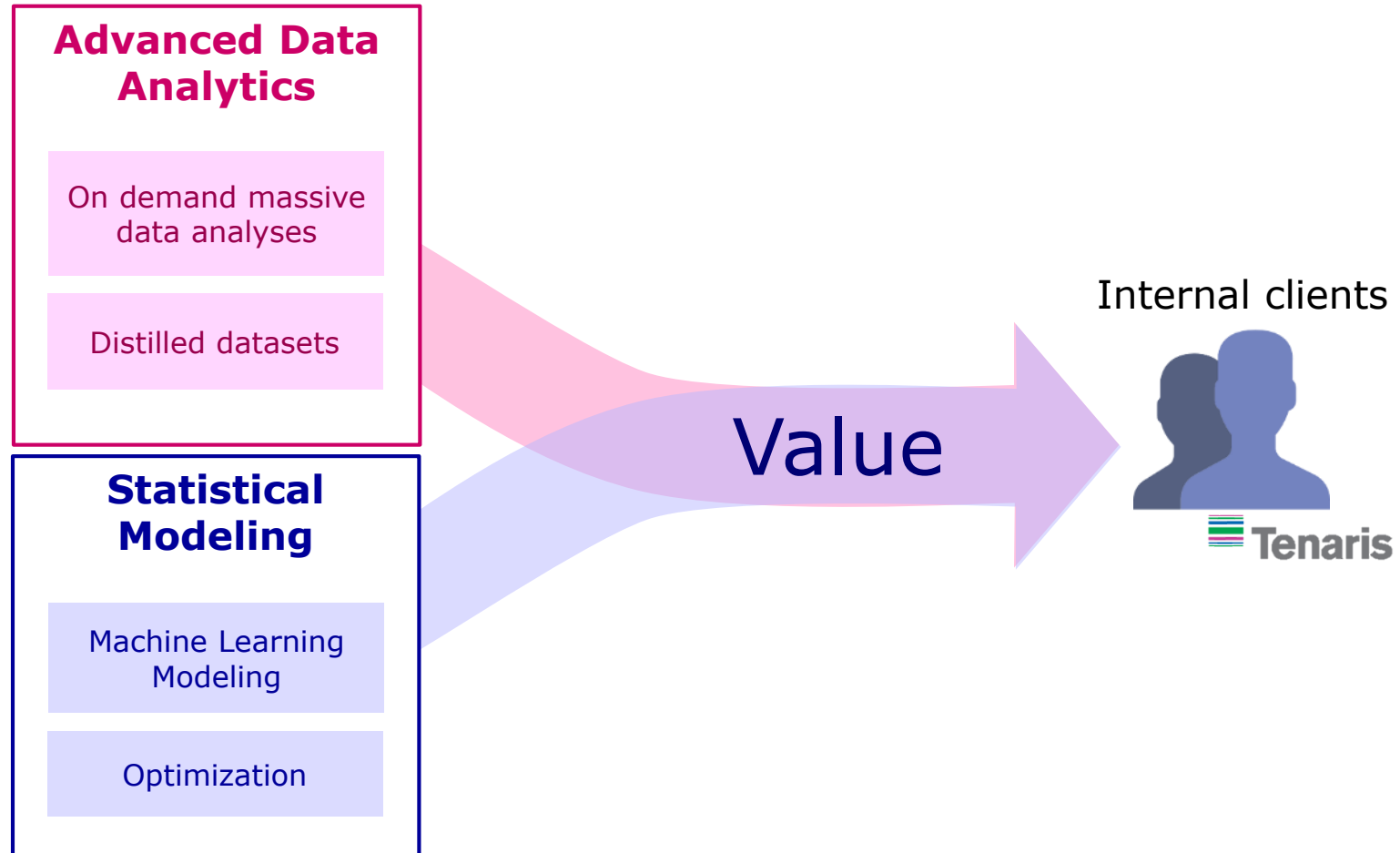
I want to:

- **Extract new metrics** from complex data (e.g. timeseries, images, ...).
- **Visualize** different **KPIs and their relations** to check a set of hypotheses.
- Perform **what-if analyses**.
- Create **mathematical models** to predict the process behavior.
- **Identify patterns** and **recognize anomaly** behavior.
- **Optimize** allocation of resources."

Advanced (Big)
Data Analytics

Statistical
Modeling

Data Science Products

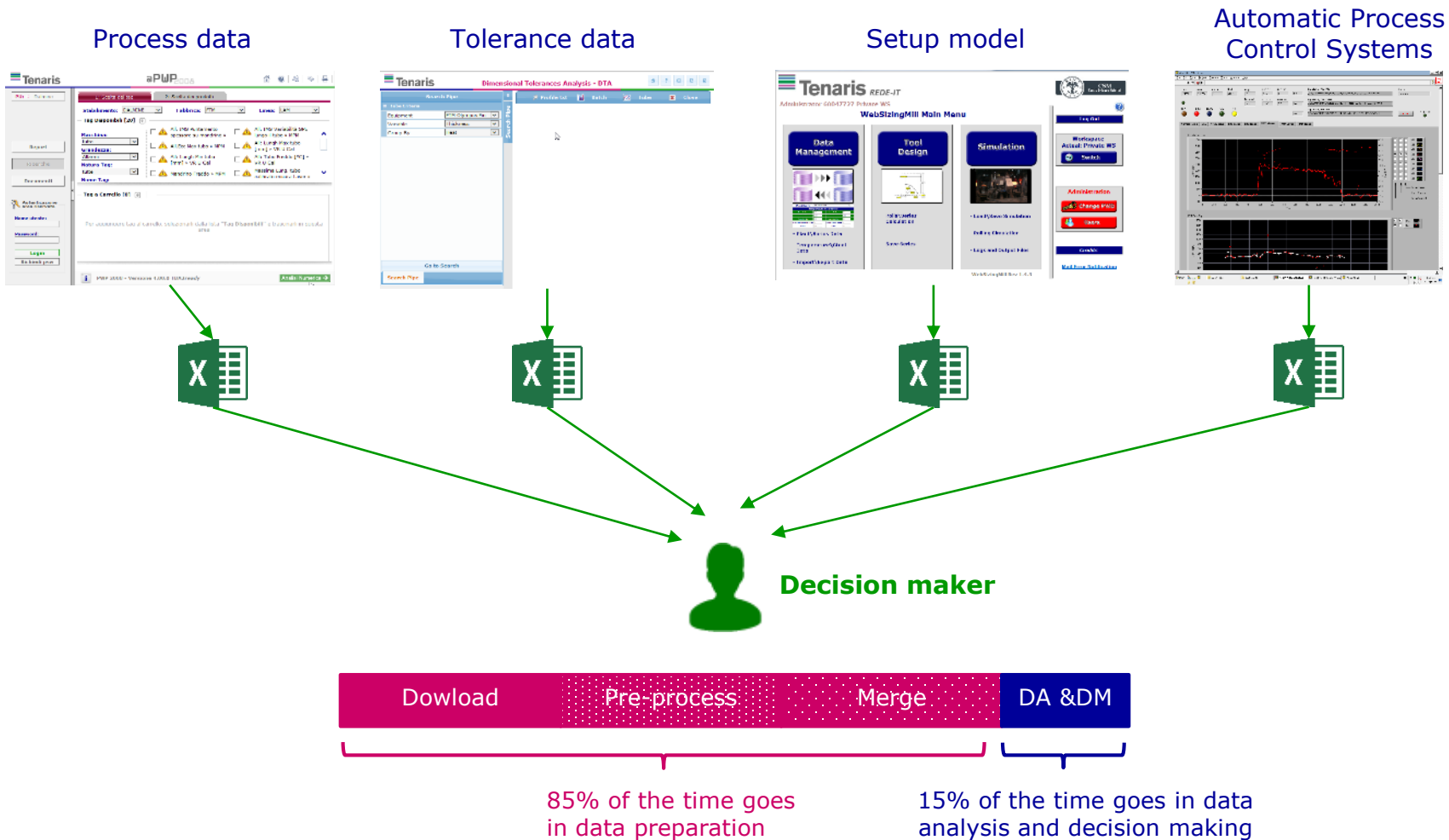




MASSIVE DATA ANALYSIS

Data Drive Decision Making

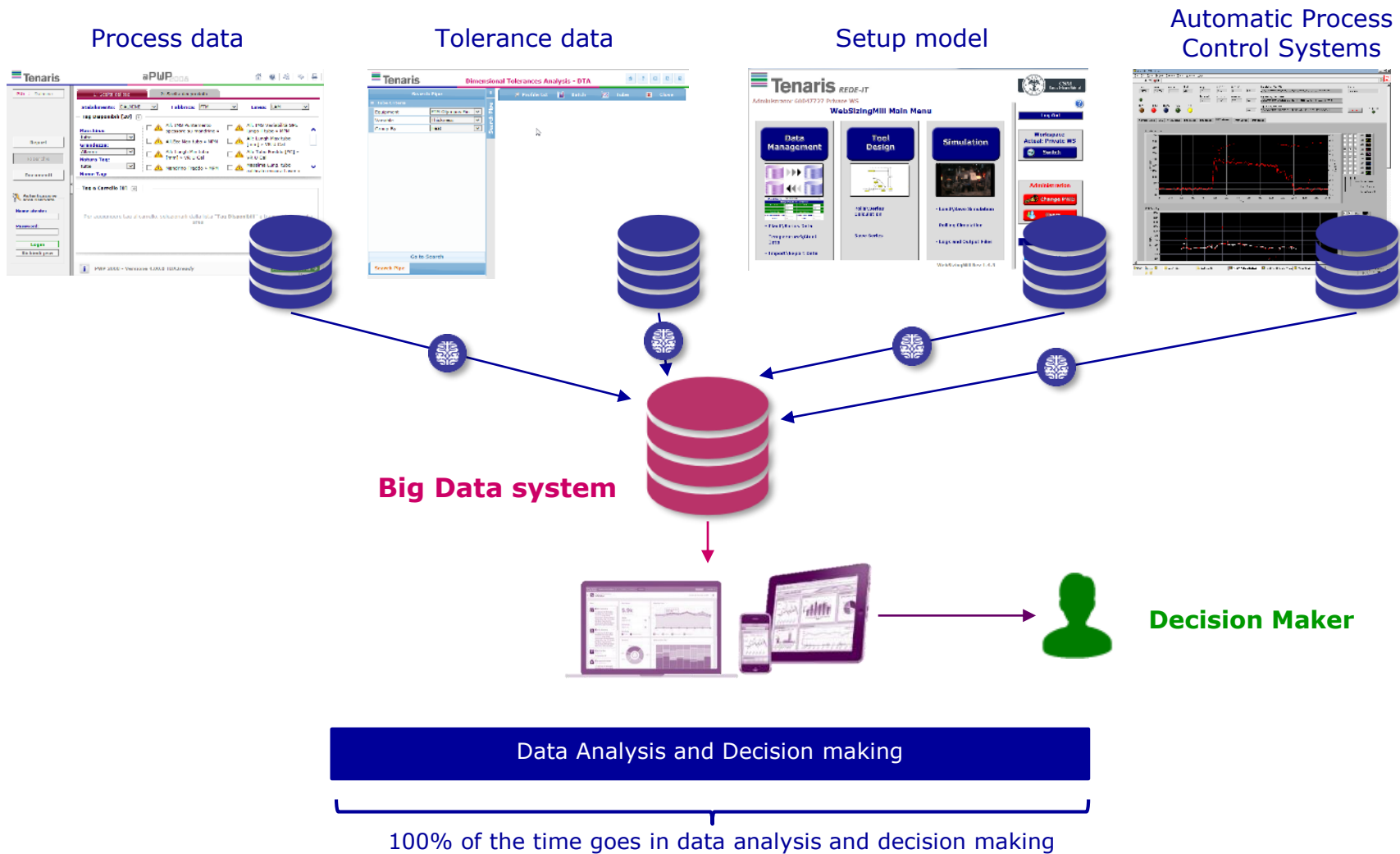
Traditional Approach



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

Data Drive Decision Making

Big Data Approach



Why Big Data



Velocity, data is produced faster than a single machine can process it.

Volume of data is growing faster than the storage capacity of a single machine.

Variety of data, as input data can be in several format (structured, not structured, industrial standards, custom solutions)

**We need a technology to perform
high-level, parallel and flexible
analysis on large amounts of data.**

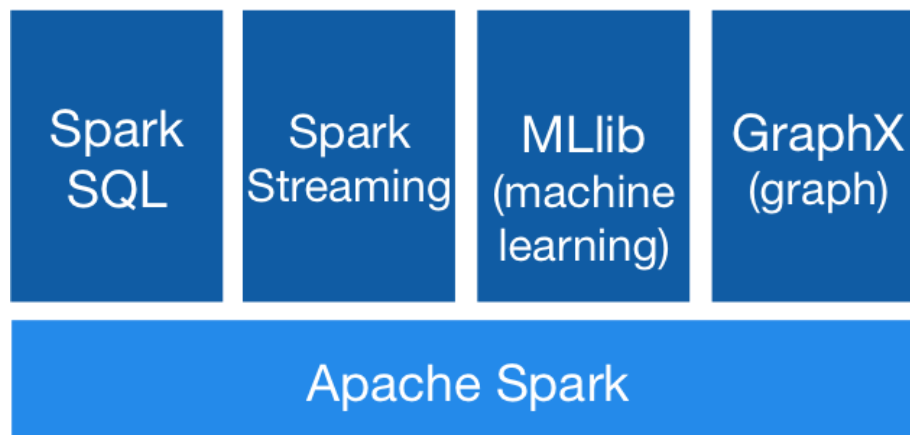


APACHE SPARK

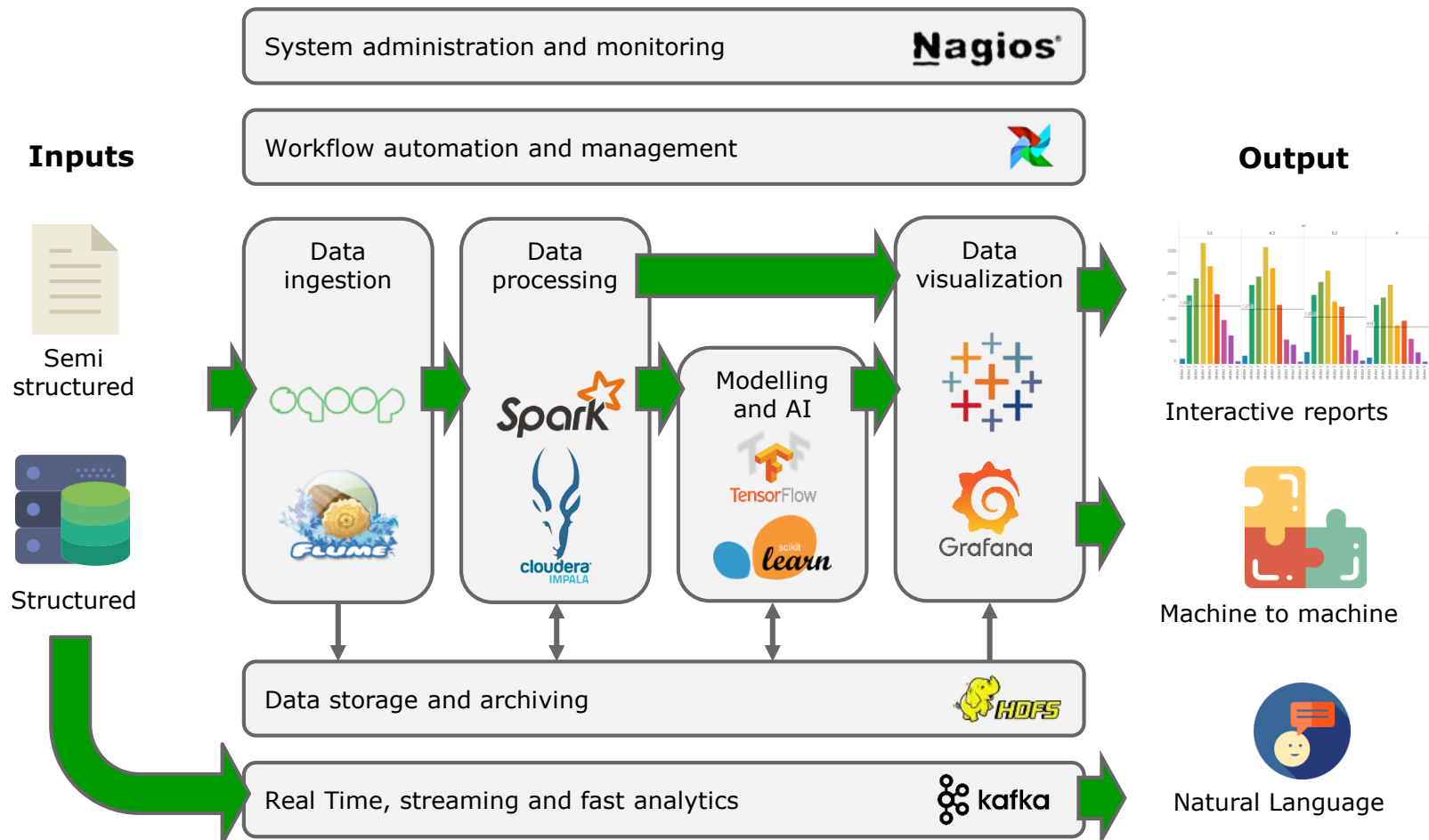
Open-source framework developed by Berkeley's University AMPLab and now maintained by Apache Foundation.

Unlike other paradigms – such as Map-Reduce – Spark uses in-memory functions (with performances up to 100x).

It supports different programming languages: Java, Scala, R and Python.



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 soft versions of the Flume 1.x codeline.

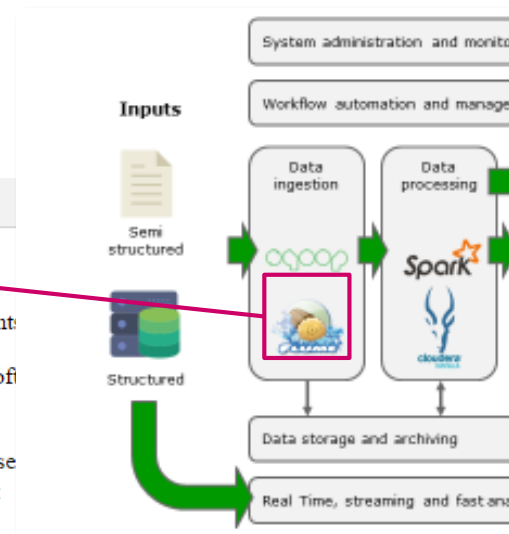
Several months of active development went into this release: almost 100 patches were committed since 1.6.0, represe While the full change log can be found on the 1.7.0 release page (link below), here are a few new feature highlights:

- 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

**Now included in
all the major Big
Data releases!**



Spark Main Intuitions



- Spark allows developer to decompose their algorithms in tasks that can be run across several machines (i.e. cluster)
- Each task is run independently and Spark consolidates the results
- All the complexity of handling, optimizing and running code in a distributed environment is managed transparently by Spark

**What could go wrong when running
in a distributed environment?**

Intuition: Word Counting



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



Word	Occurences
I	3
Am	3
Sam	3
Do	1
You	1
Like	1
...	...

Intuition: Word Counting



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

Word	Occurences
I	1

Intuition: Word Counting



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

Word	Occurences
I	1
Am	1

Intuition: Word Counting



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

Word	Occurrences
I	1
Am	1
Sam	1

Intuition: Word Counting



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

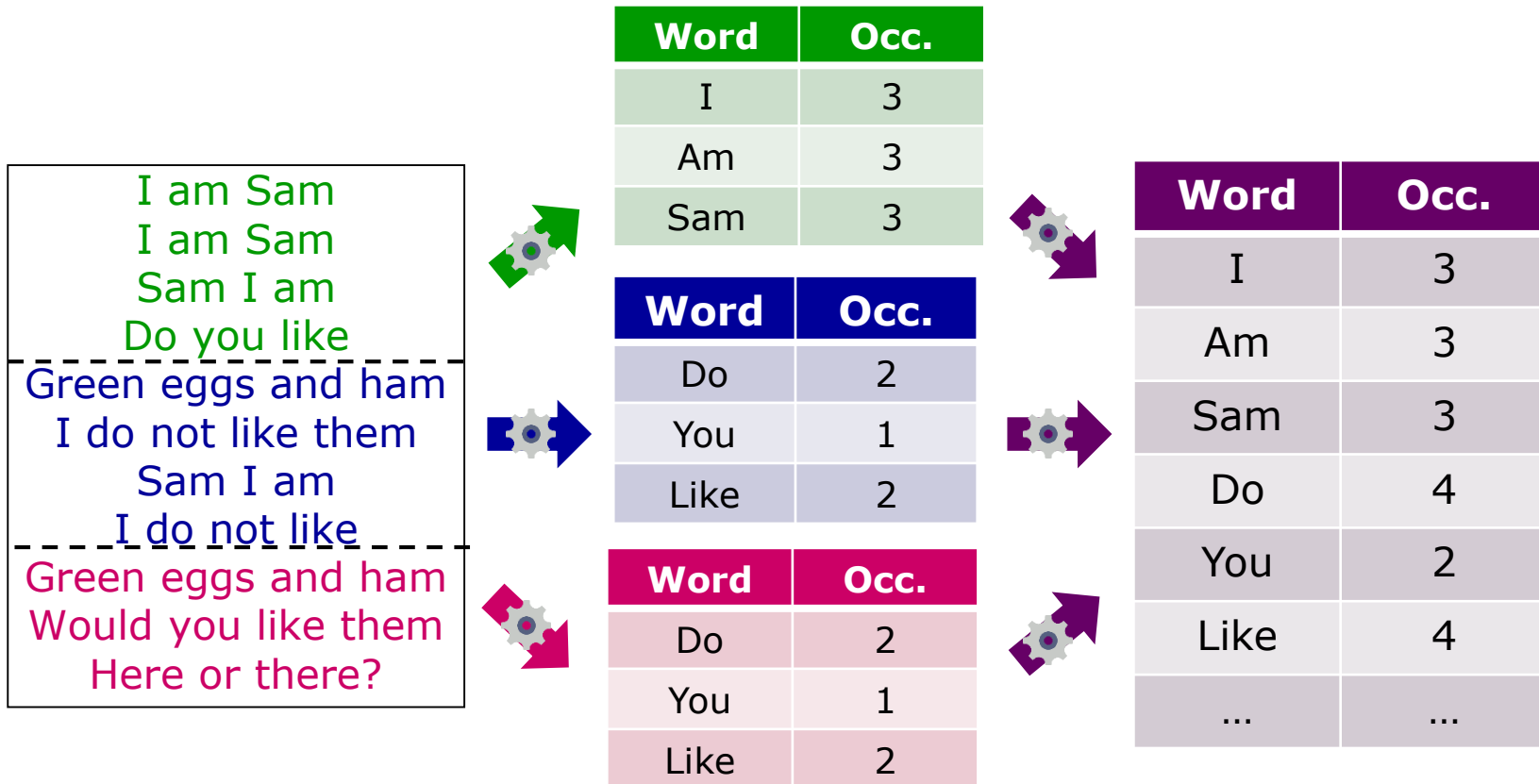
Word	Occurences
I	2
Am	1
Sam	1

Intuition: Word Counting on Long Text (1/3)

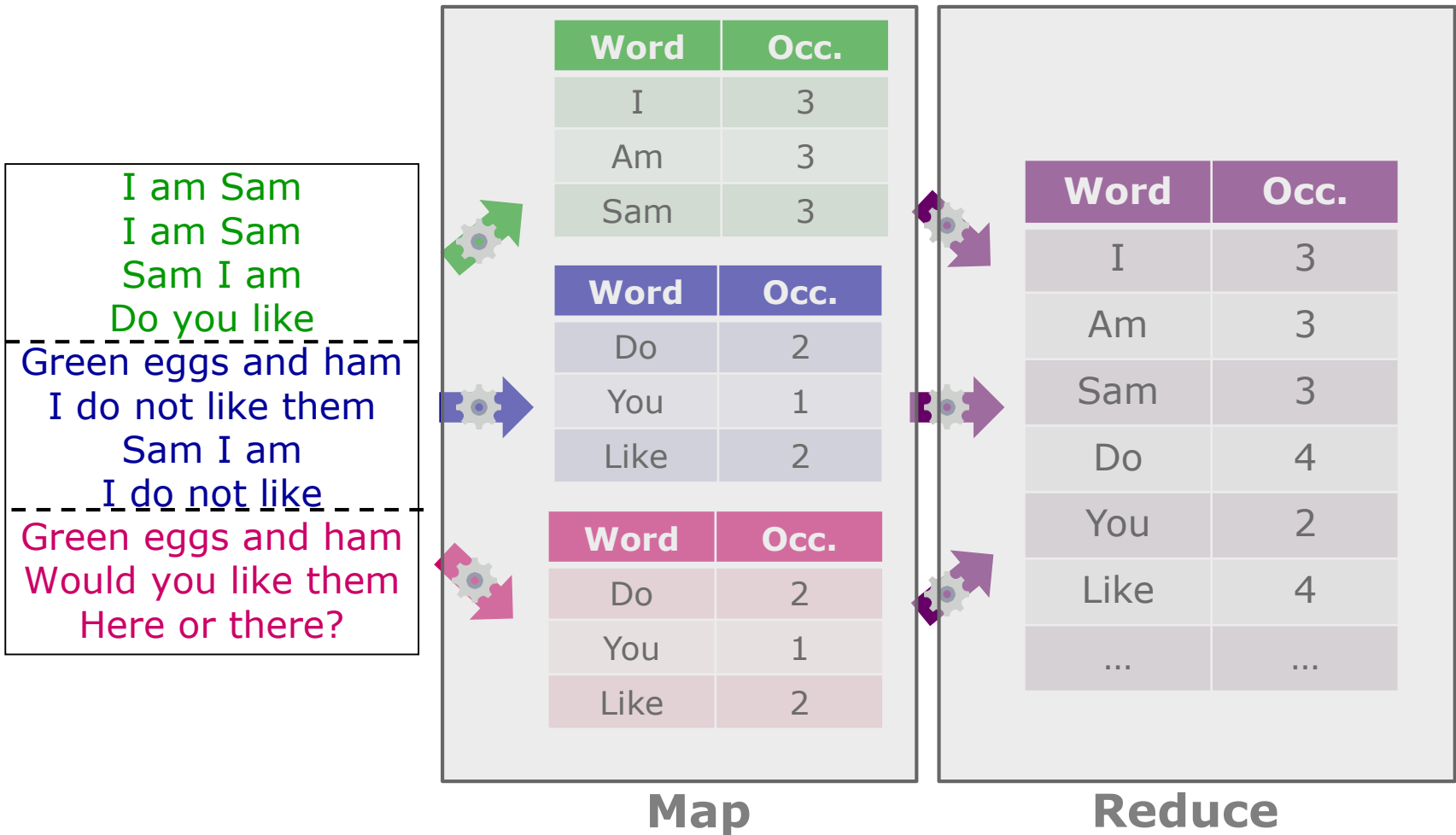


I am Sam
I am Sam
Sam I am
Do you like
Green eggs and ham
I do not like them
Sam I am
I do not like
Green eggs and ham
Would you like them
Here or there?

Intuition: Word Counting on Long Text (2/3)



Intuition: Word Counting on Long Text (3/3)



Setup Databricks



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

The screenshot shows the Databricks website's 'try-databricks' page. At the top 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 for the 'FULL-PLATFORM TRIAL', described as 'Put Apache Spark to work', and lists features like unlimited clusters, notebooks, and a 14-day free trial. The right column is for the 'COMMUNITY EDITION', described as 'Learn Apache Spark', and lists features like a mini 6GB cluster and interactive notebooks. A large pink arrow points from the Community Edition section towards a 'START TODAY' button. At the bottom right, 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

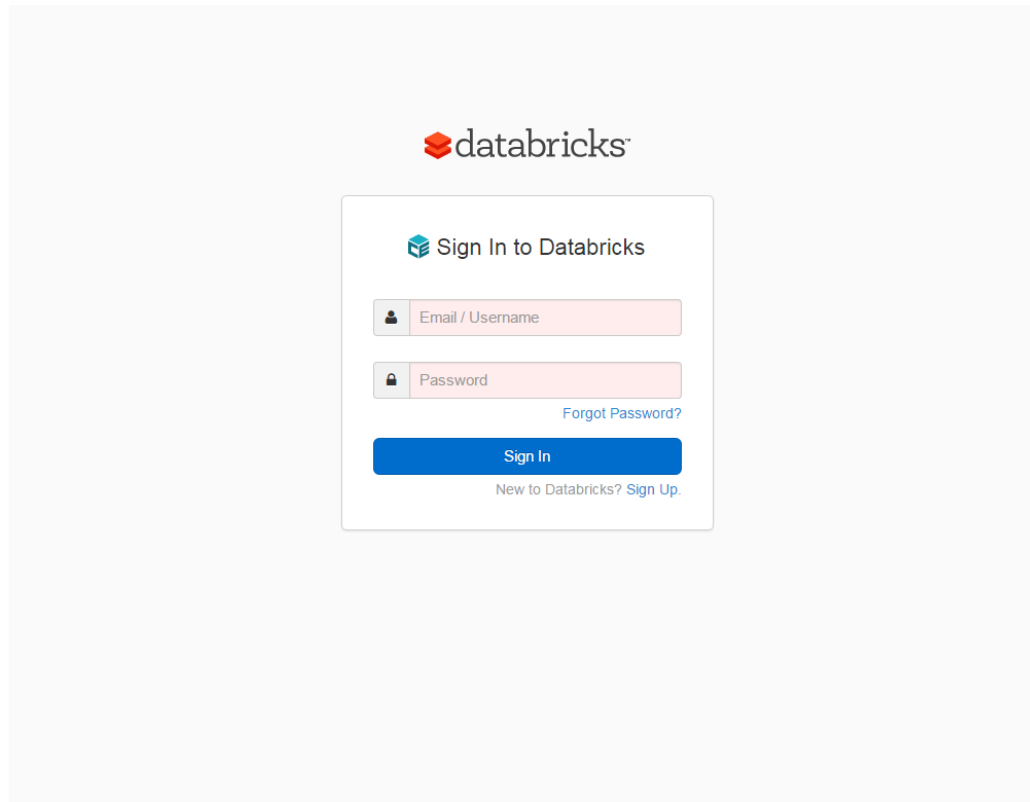


GETTING STARTED WITH SPARK

Login databricks



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



The screenshot shows the Databricks login interface. At the top, the Databricks logo is displayed. Below it, a central box contains the text "Sign In to Databricks" with a small icon. There are 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 box, there is a link for "New to Databricks? Sign Up."

GitHub repository



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



Internships



Are you looking for an internship on **Big Data**, **Machine Learning** or **Artificial Intelligence**?

Send us an email!

Andrea Rota
arota@tenaris.com

Michele Zanchi
mzanchi@tenaris.com

