Parallel computing with Apache Spark

Andrea Rota, Michele Zanchi
Tenaris 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

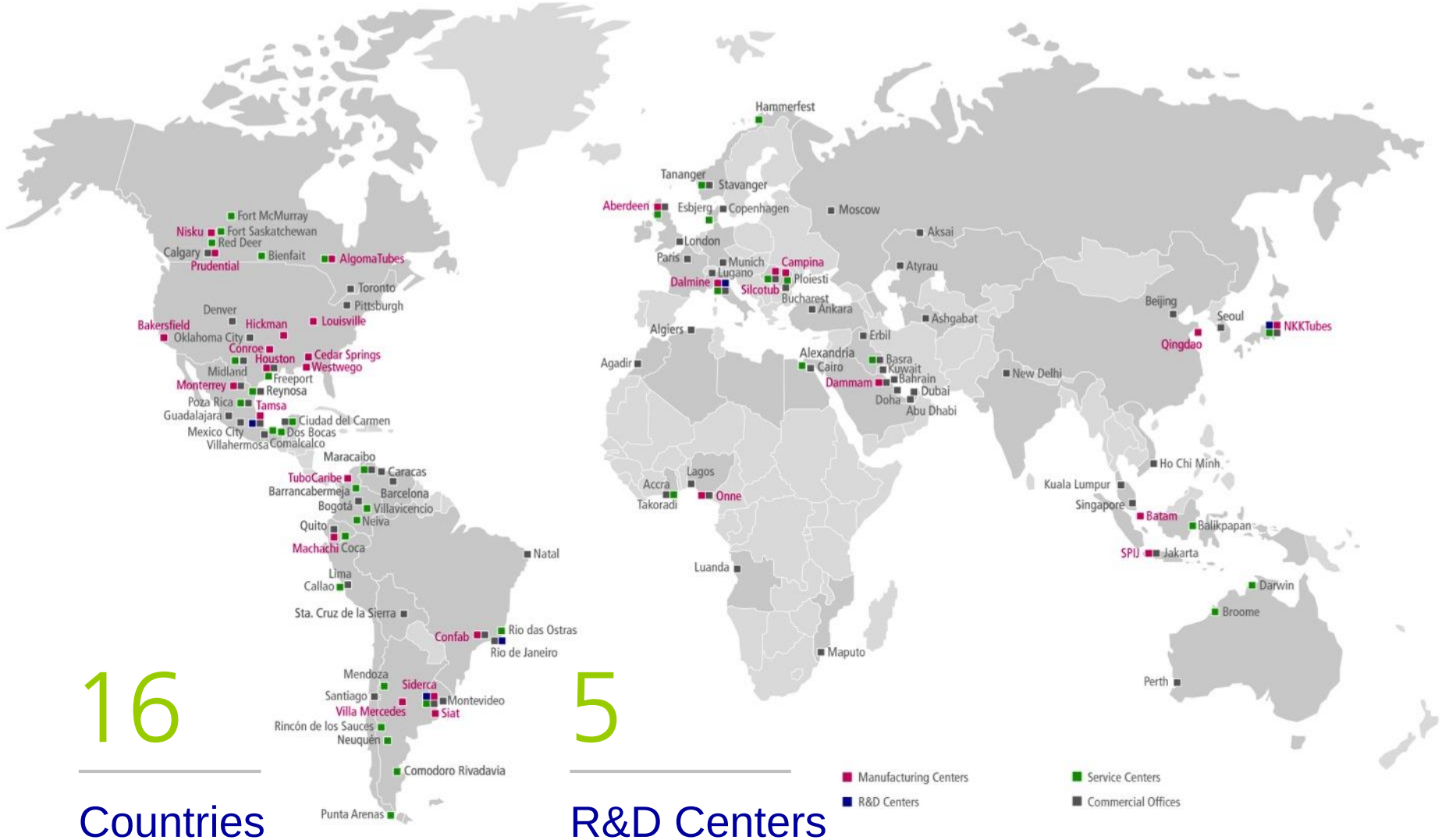
Linepipe per applicazioni onshore e offshore

Trasformazione di idrocarburi

Generazione di energia

Applicazioni industriali e automotive

Tenaris



16

Countries

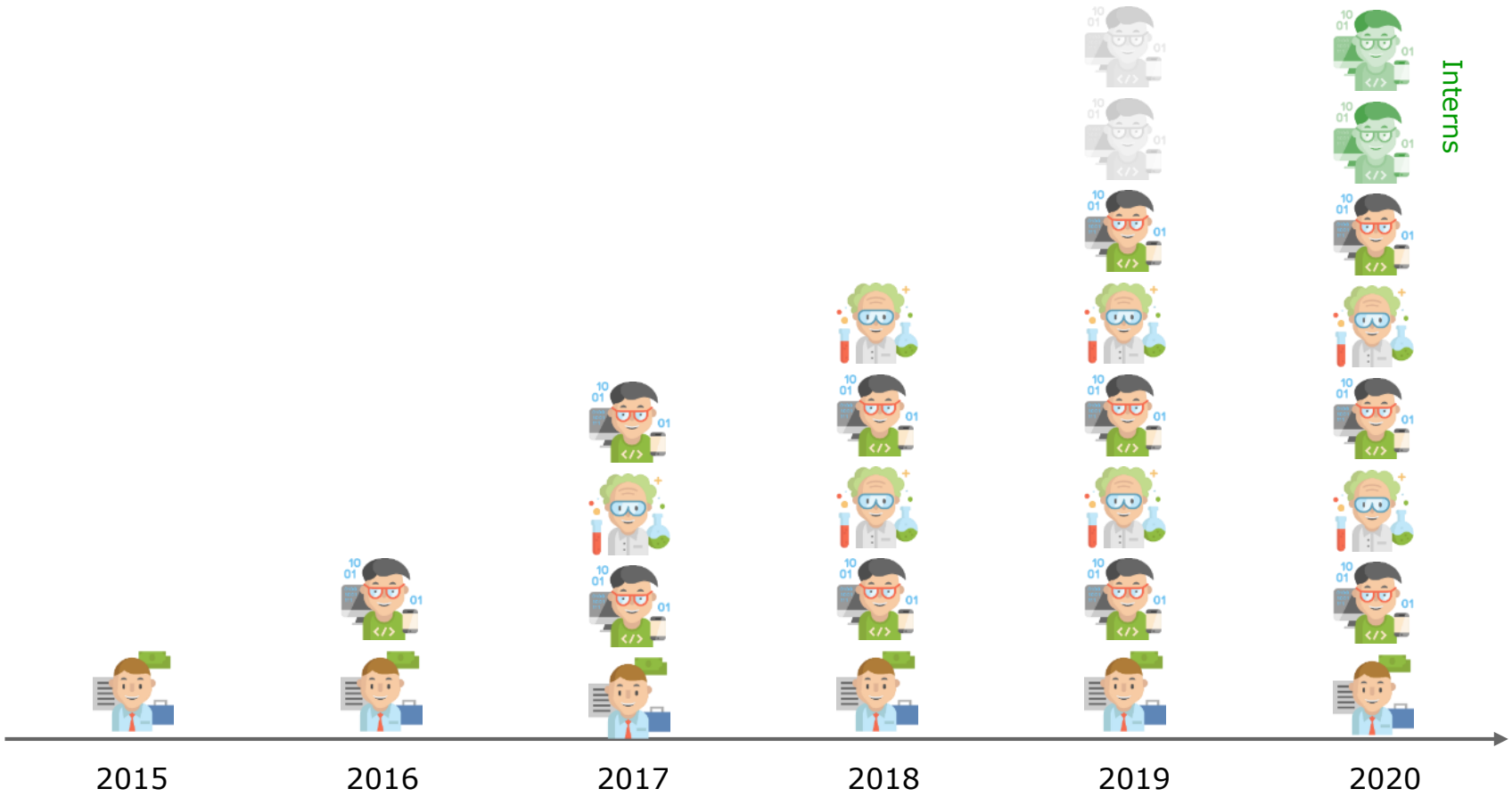
5

R&D Centers

R&D Data Science

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.

Roadmap: 2020/21



Needs addressed by the department

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

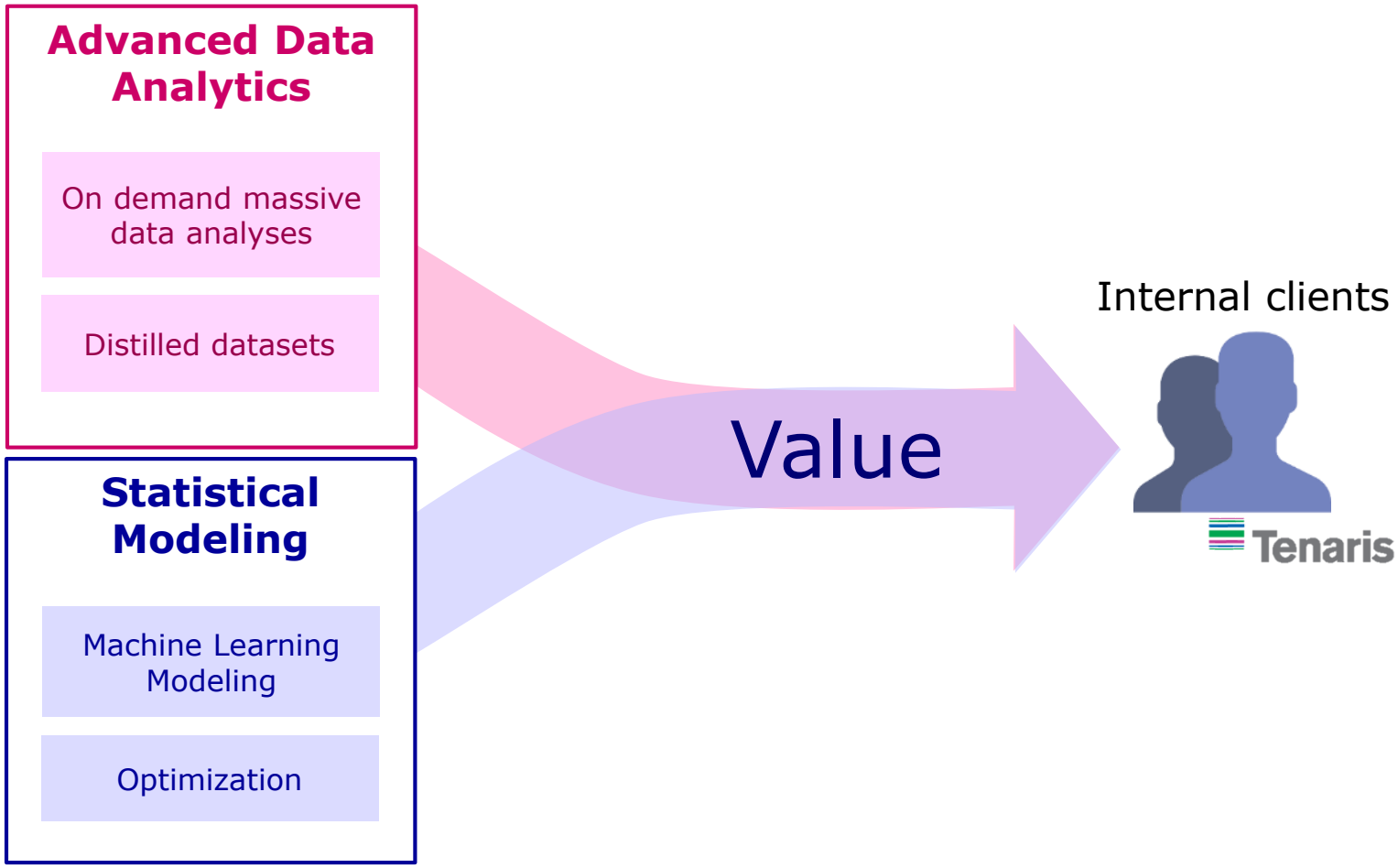
I want to:

- **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

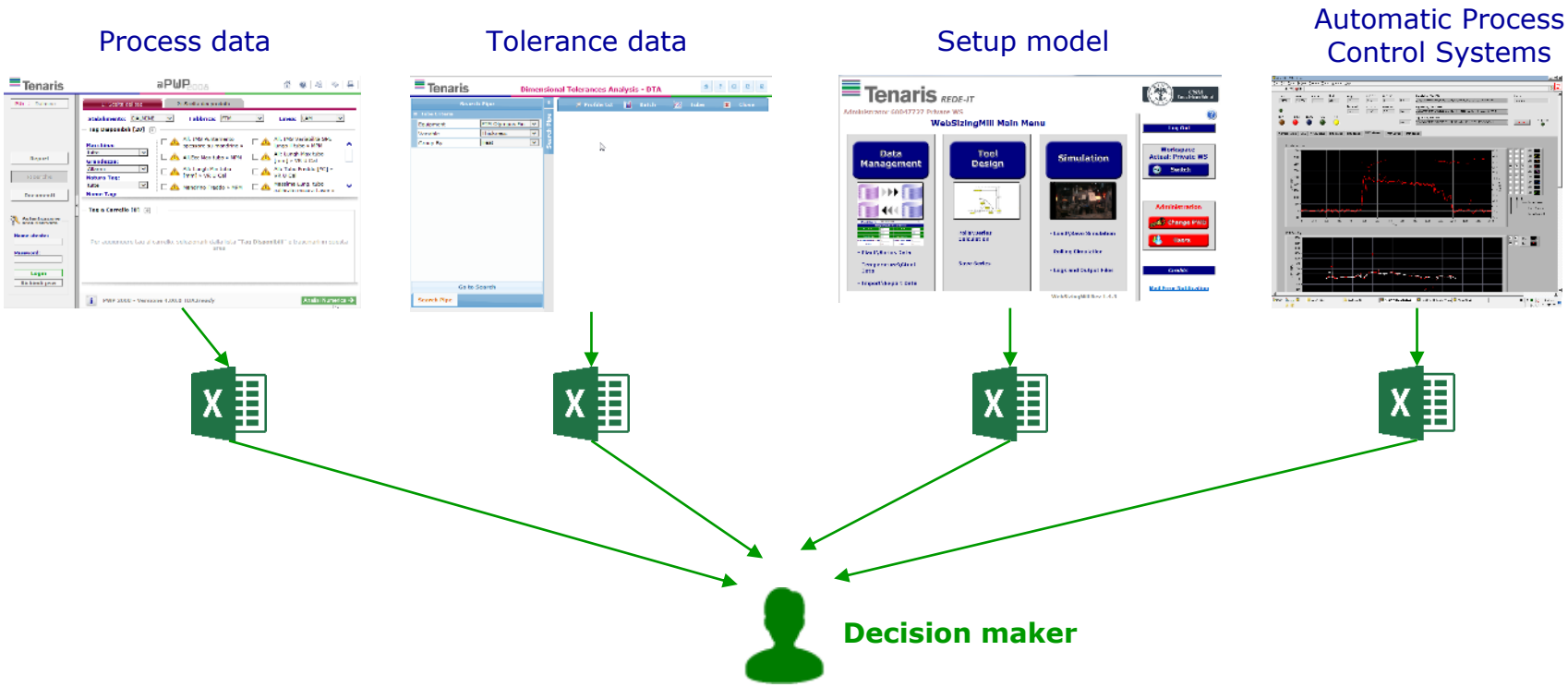
Statistical Modeling

Data Science Products



MASSIVE DATA ANALYSIS

Data Drive Decision Making
Traditional Approach

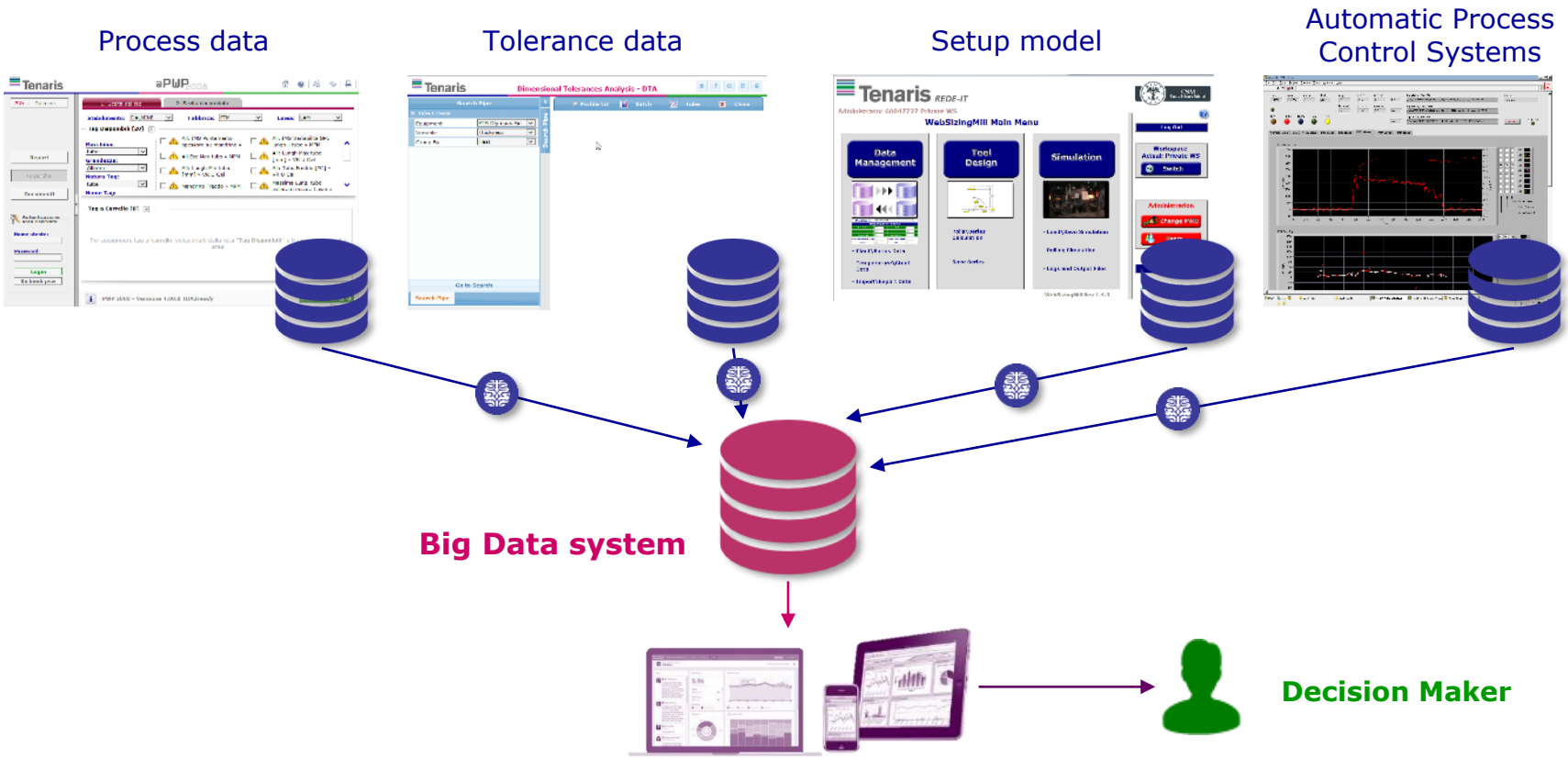


85% of the time goes in data preparation

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)

Data Drive Decision Making
Big Data Approach



Data Analysis and Decision making

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

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 structered, industrial standards, custom solutions)

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

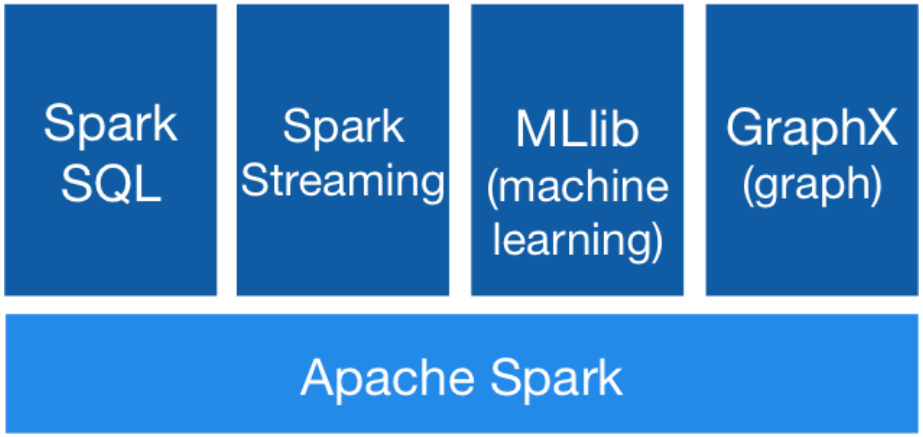
APACHE SPARK

Apache Spark

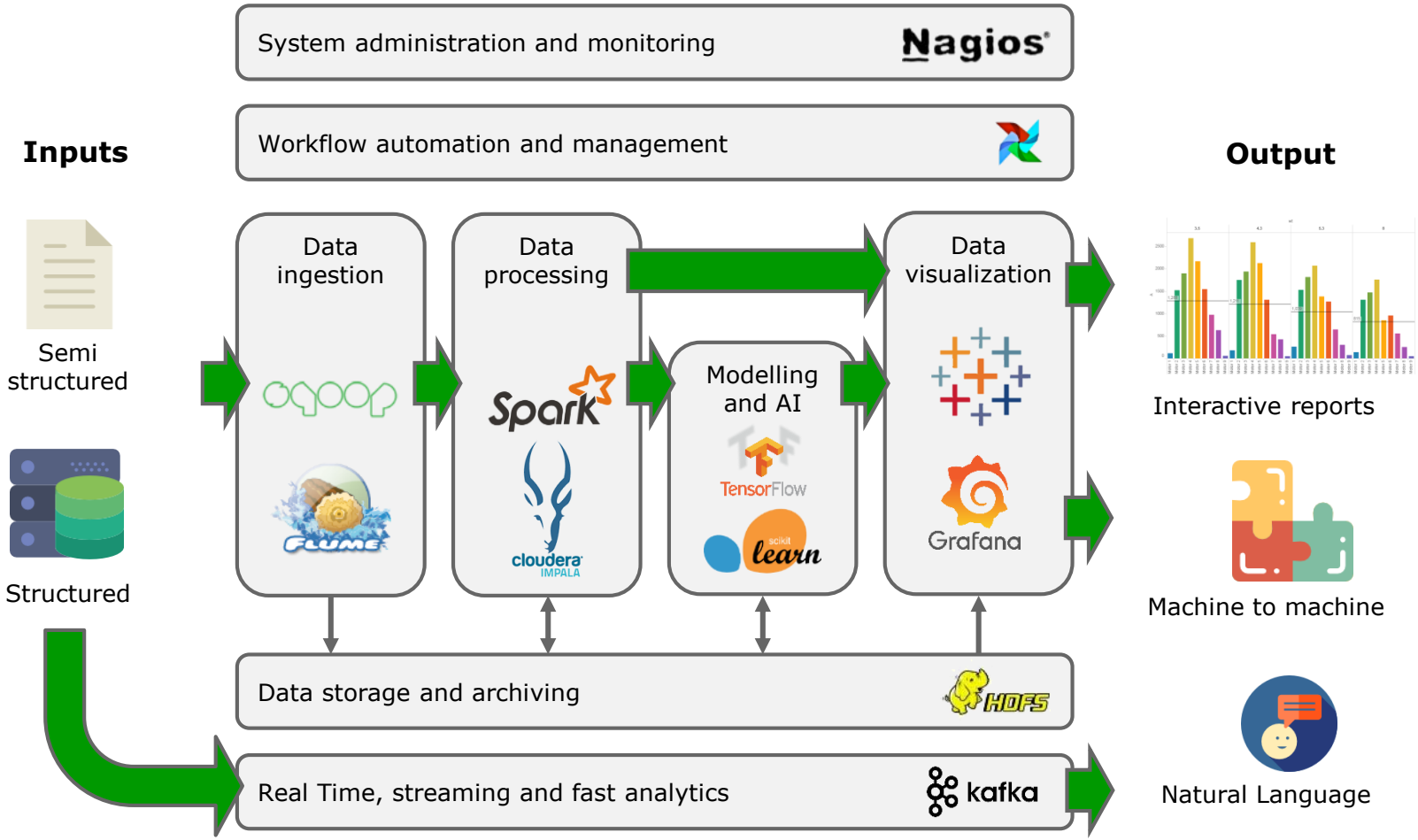
Open-source framework developed by Berkeley's University AMPLab and now mantained by Apache Foundation.

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

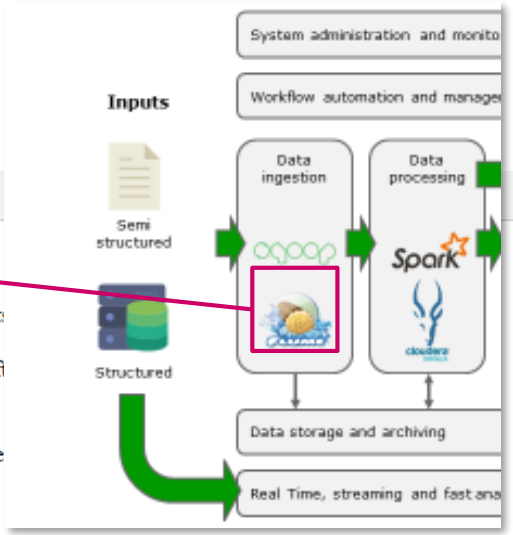
It supports different programming languages: Java, Scala, R and Python.



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 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

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 indipendently 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	Occurences
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)

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?

Word	Occ.
I	3
Am	3
Sam	3

Word	Occ.
Do	2
You	1
Like	2

Word	Occ.
Do	2
You	1
Like	2

Word	Occ.
I	3
Am	3
Sam	3
Do	4
You	2
Like	4
…	…

Intuition: Word Counting on Long Text (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?

Word	Occ.
I	3
Am	3
Sam	3

Word	Occ.
Do	2
You	1
Like	2

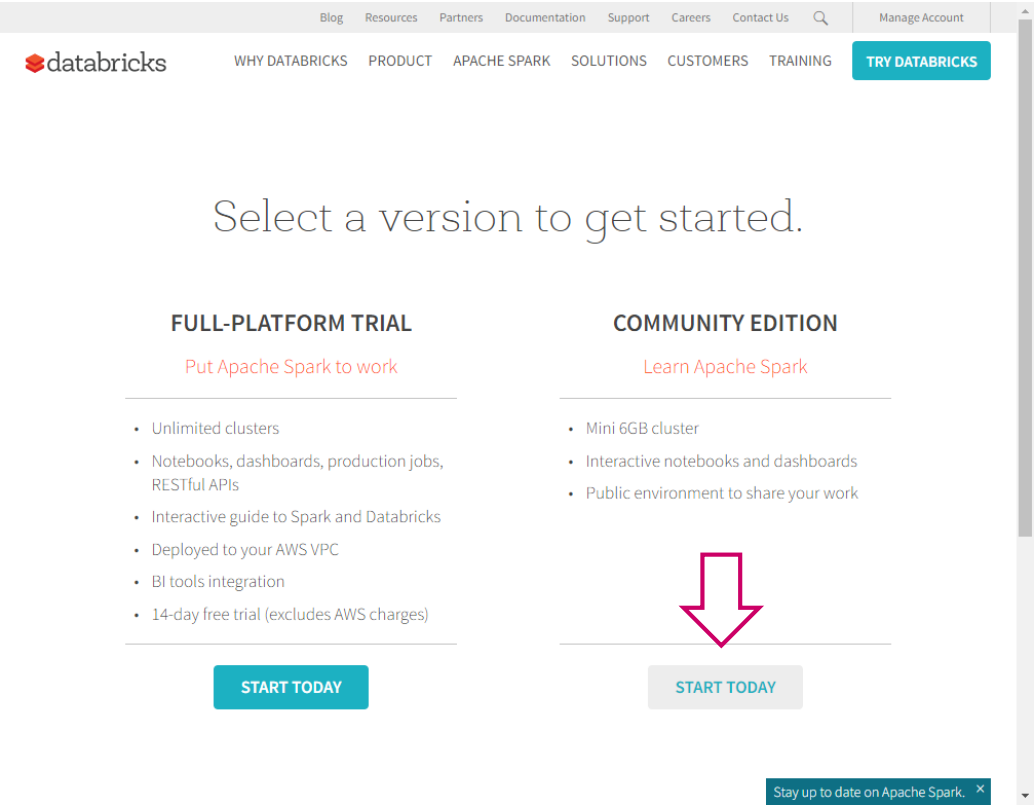
Word	Occ.
Do	2
You	1
Like	2

Word	Occ.
I	3
Am	3
Sam	3
Do	4
You	2
Like	4
...	...

Map **Reduce**

Setup Databricks

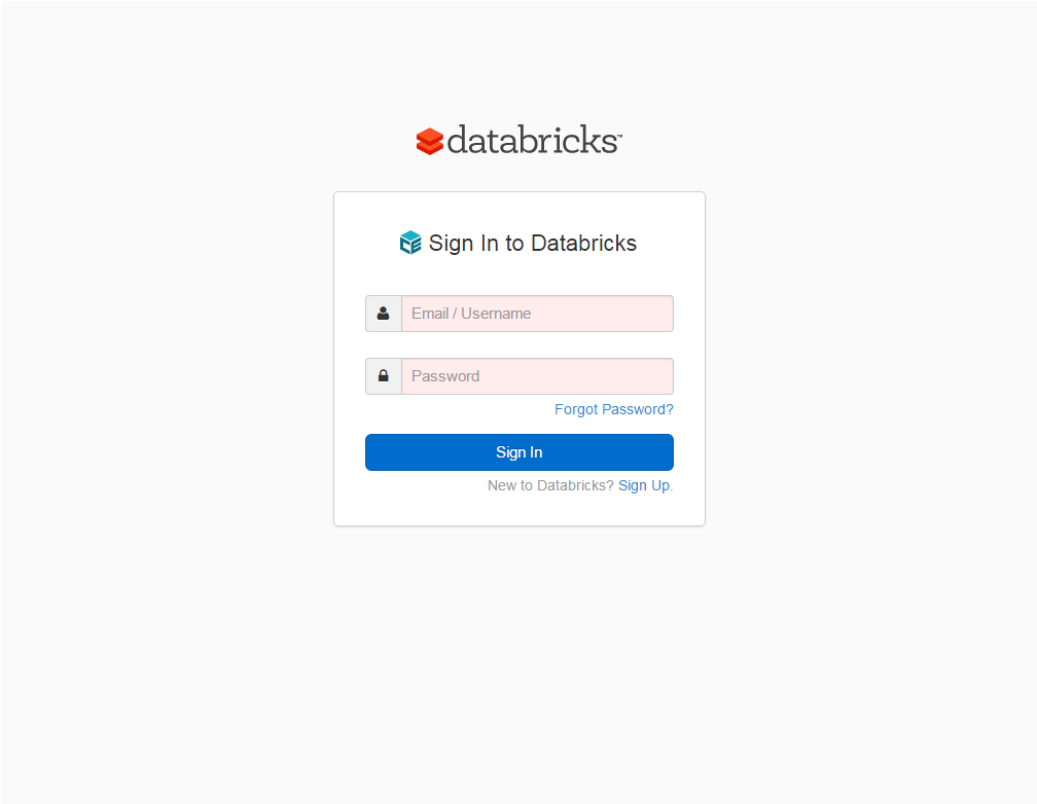
https://databricks.com/try-databricks



GETTING STARTED WITH SPARK

Login databricks

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



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