

# SOMACHINE



INSTITUTO DE  
ASTROFÍSICA DE  
ANDALUCÍA



CSIC



Andalusian  
Research Institute in  
Data Science and  
Computational Intelligence



UNIVERSIDAD  
DE GRANADA



Instituto Andaluz Interuniversitario en  
Data Science and Computational Intelligence

## Big Data: Foundations and Frameworks

A. Fernández. Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence. **Universidad de Granada.**



# Astronomy questions require astronomical data

How the first stars and galaxies were formed? What is the nature of the dark matter?

Is there any life in other planets?

Simulations to reproduce the observable universe are complex, and generate Big Data

# Outline



1

- Big Data. Big Data Science

2

- Why Big Data? Google and the MapReduce programming model

3

- Big Data technologies: Hadoop / Spark ecosystem

4

- Big Data Analytics: Libraries for Data Analytics in Big Data. Case studies

5

- Final Comments

# Outline



1

- **Big Data. Big Data Science**

2

- Why Big Data? Google and the MapReduce programming model

3

- Big Data technologies: Hadoop / Spark ecosystem

4

- Big Data Analytics: Libraries for Data Analytics in Big Data. Case studies

5

- Final Comments



# What is Big Data

Big Data and Data Science





SCALE OF DATA  
**VOLUME**



FORMS OF DATA  
**VARIETY**

# BIG DATA

**VELOCITY**

ANALYSIS OF DATA-FLOW



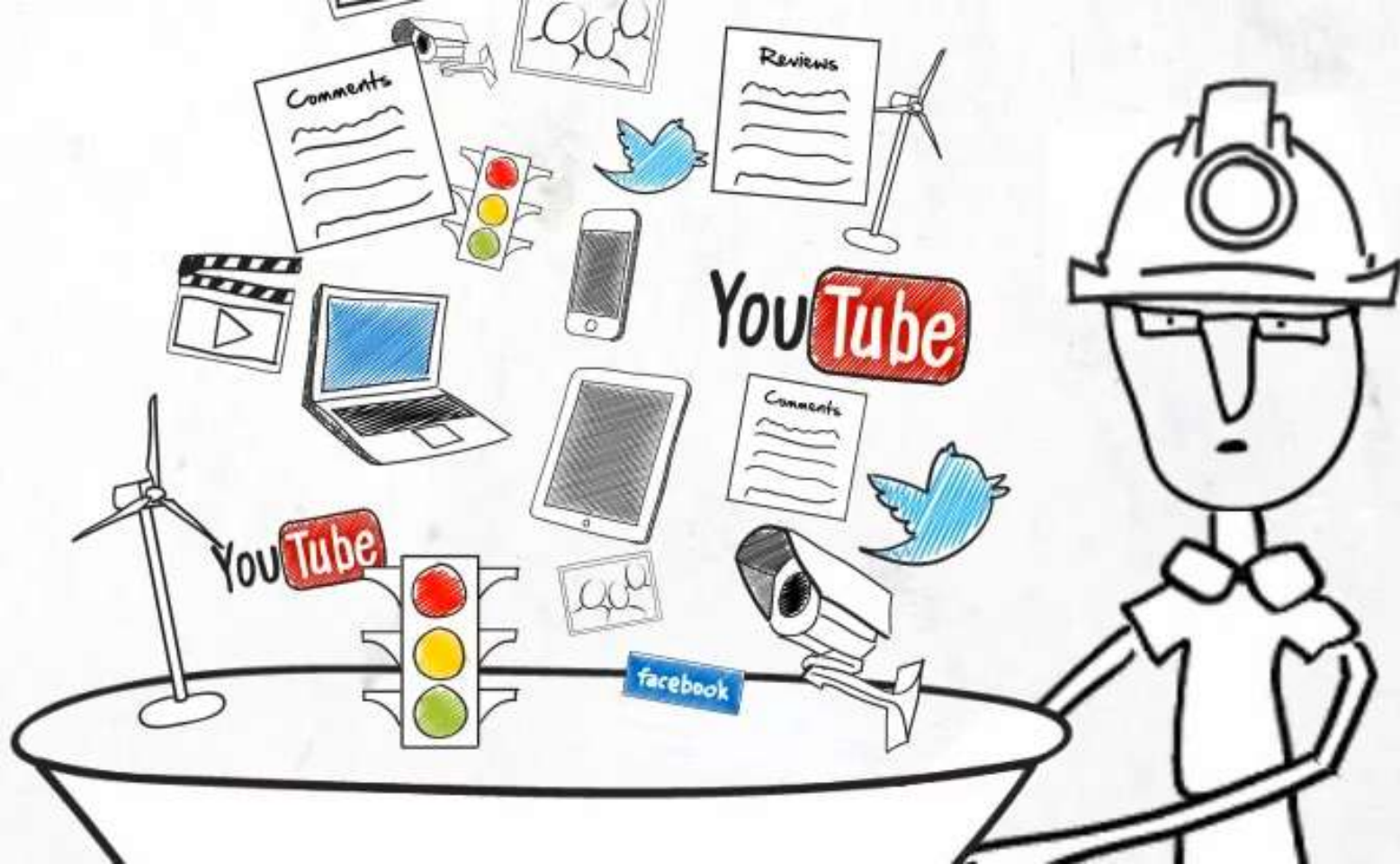
**VERACITY**

UNCERTAINTY OF DATA



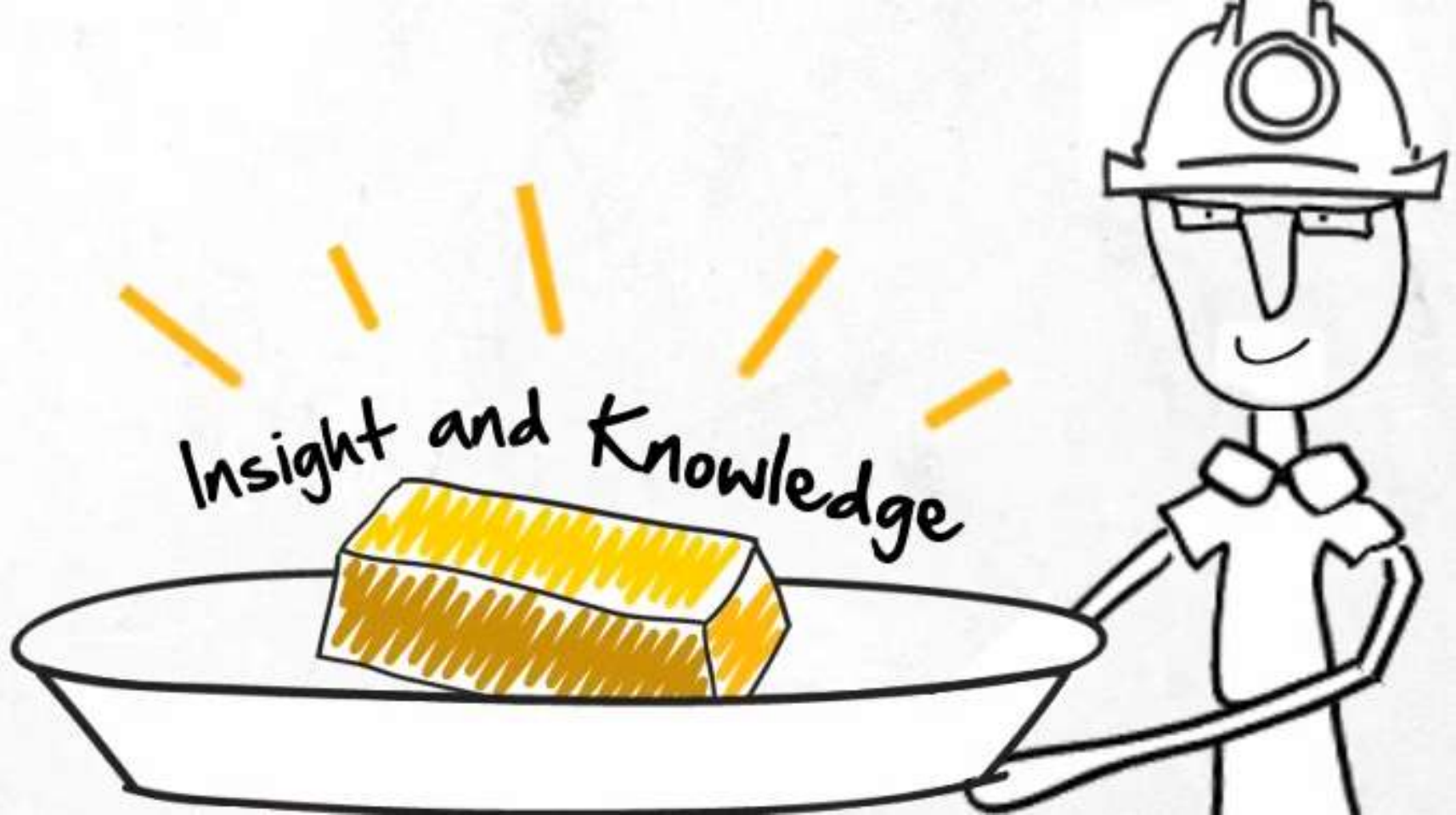
## What is Big Data?

The “4” Vs determining the sides of the story



## The “Internet of Things”

Different sources generate a large amount of data



## What is Big Data?

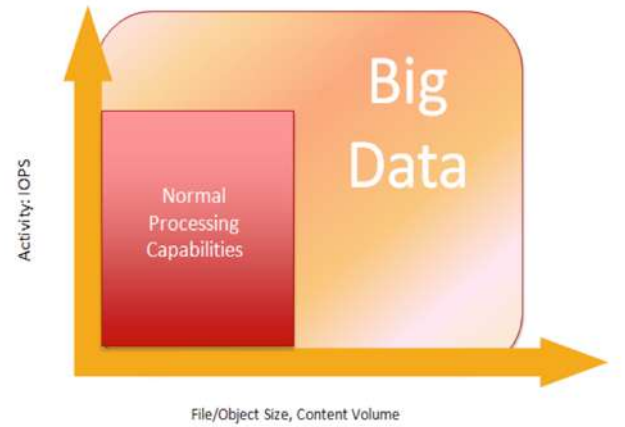
Methods to transform raw data to actionable knowledge



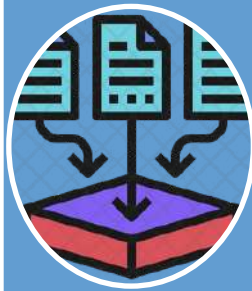
# What is Big Data? Change the way we design solutions and solve problems

There is not a standard definition!

“*Big Data*” involves data whose **Volume**, **Variety** (diversity), **Velocity**, & complexity requires new techniques, algorithms and analyses to extract **valuable knowledge** (hidden).



# Data Science LifeCycle (MLOps)



## Data Collection

- Source of information
- Get as many samples as possible
- Data in different formats
- Define the problem



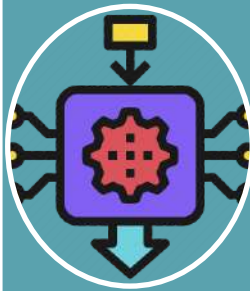
## Exploratory Data Analysis

- How data looks like
- Check issues
- Preliminary estimation of performance



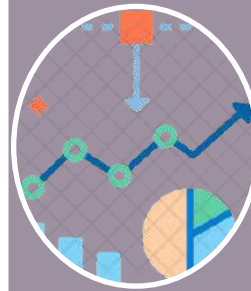
## Data Cleaning

- Data format
- Aggregation
- Normalisation
- Missing information
- Data reduction



## Model Building

- Apply learning methods
- Knowledge extraction
- Predictive or descriptive models



## Visualisation interpretation

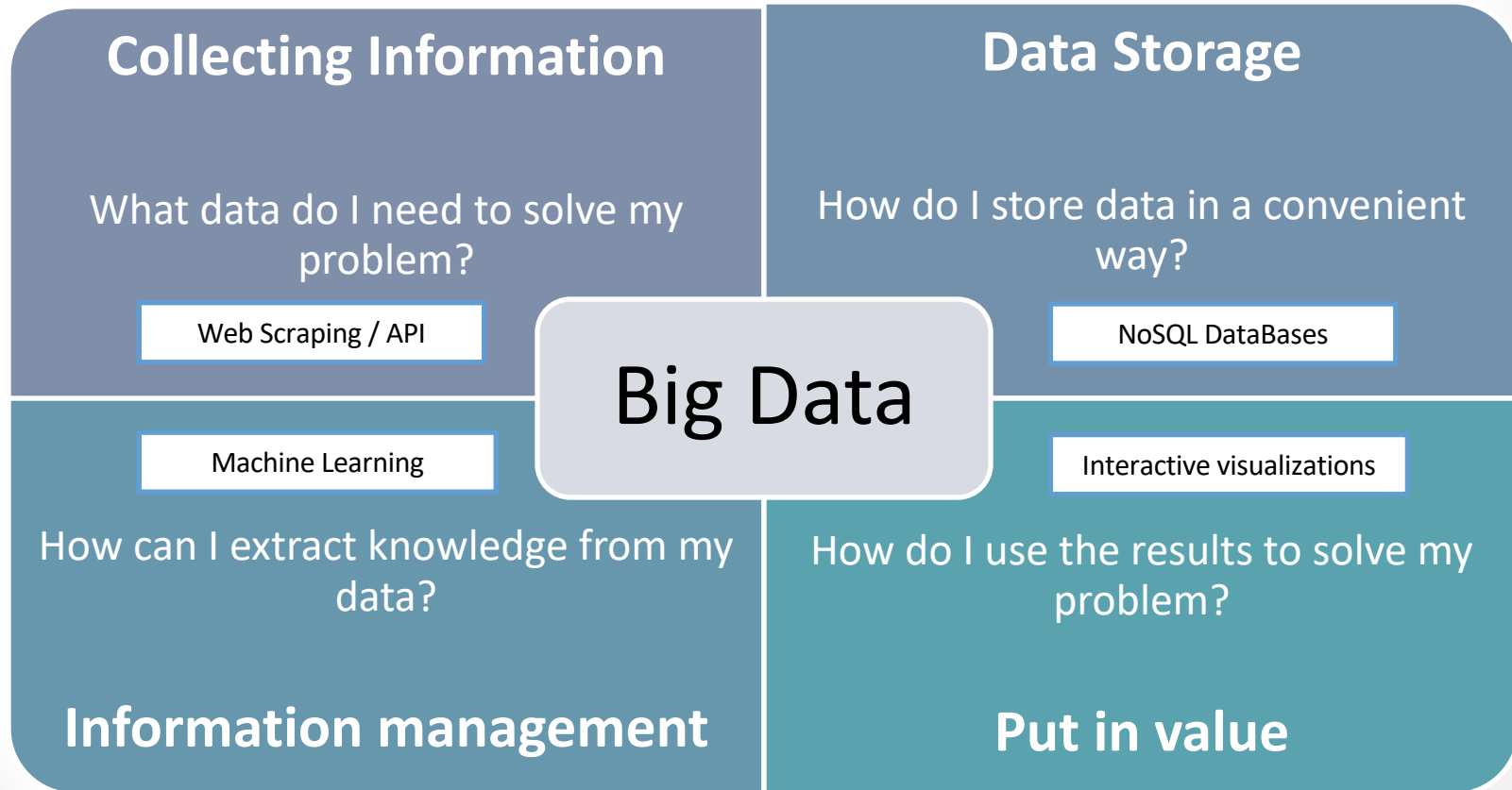
- Explain the behavior
- Interpret the results
- Observe interesting patterns
- Trust and support decisions

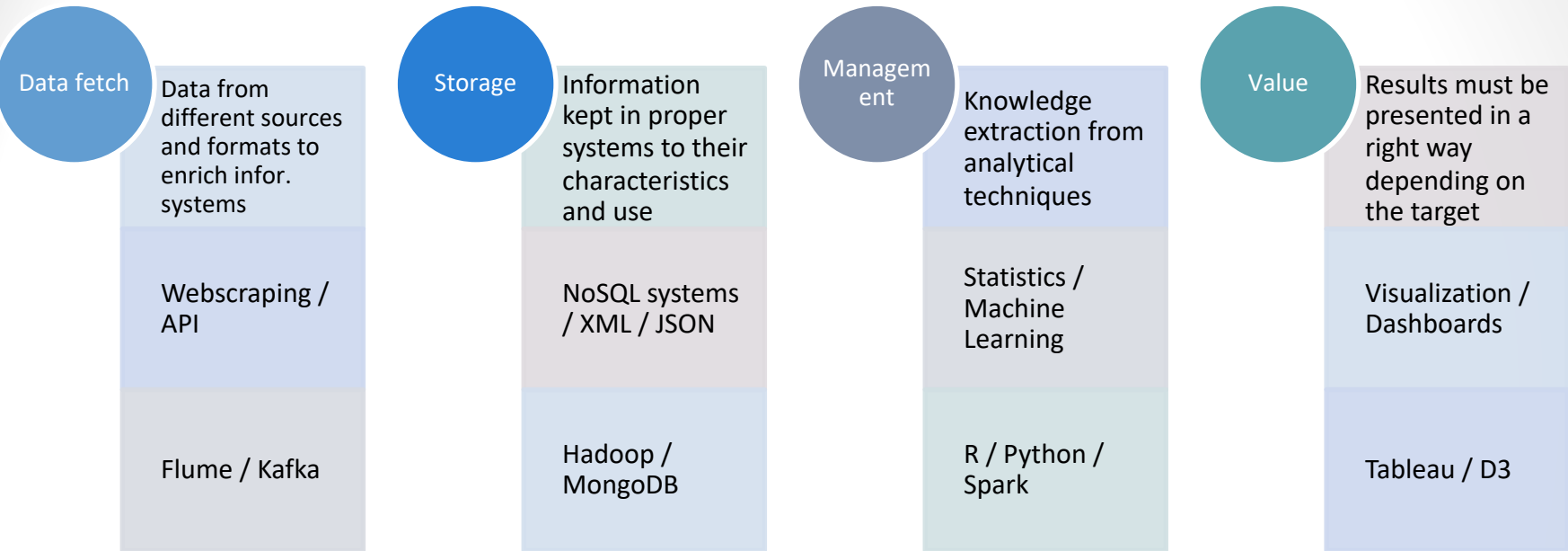


## Model Deployment

- Integrate the system
- Make the application
- Maintaining and update the models

# Big Data simplifies the translation from data to actionable knowledge





**All management information procedures must consider transversal aspects such as information quality, traceability, security, privacy, among others**

Examples of Tools and Technologies – A multidisciplinary approach

# Issues when working in Big Data

Cheap, abundant storage.

Faster processors.

Affordable open source, distributed big data platforms (Hadoop).

Parallel processing, MPP, virtualization, grid environments, high throughputs.

Cloud computing and other flexible resource allocation arrangements.





## DATA &amp; AI LANDSCAPE 2020

## INFRASTRUCTURE

The collage features logos for various cloud services, organized into several categories:

- STORAGE:** Amazon S3, Google Cloud Storage, Microsoft Azure, AWS, IBM, Oracle, SAP, and others.
- HADOOP:** Cloudera, Databricks, and others.
- DATA LAKES:** Snowflake, Databricks, and others.
- DATA WAREHOUSES:** Amazon Redshift, Google BigQuery, Microsoft Azure, and others.
- STREAMING / IN-MEMORY:** Apache Kafka, Databricks, and others.

NoSQL DATABASES	SQL DATABASES	GRAPH DBs	MPP DBs	SERVER-LESS	CLUSTER SVCS
Google Cloud Bigtable Amazon DynamoDB Microsoft Azure Cosmos DB ORACLE MongoDB MarkLogic Couchbase CockroachDB ScyllaDB Aerospike	Microsoft SQL Server Amazon RDS Google Cloud SQL Microsoft Azure SQL Database IBM Db2 ORACLE Microsoft Access Microsoft Excel Microsoft Word Microsoft PowerPoint Microsoft Outlook Microsoft OneDrive Microsoft Teams Microsoft SharePoint Microsoft Dynamics 365 Microsoft Power BI Microsoft Azure IoT Hub Microsoft Azure DevOps Microsoft Azure Active Directory Microsoft Azure Key Vault Microsoft Azure Storage Microsoft Azure Blob Storage Microsoft Azure File Storage Microsoft Azure Data Lake Storage Microsoft Azure Cosmos DB Microsoft Azure Databricks Microsoft Azure Synapse Analytics Microsoft Azure Machine Learning Microsoft Azure Cognitive Services Microsoft Azure Bot Service Microsoft Azure Logic Apps Microsoft Azure Functions Microsoft Azure App Service Microsoft Azure Container Instances Microsoft Azure Kubernetes Service Microsoft Azure Service Fabric Microsoft Azure Virtual Machines Microsoft Azure Virtual Network Microsoft Azure Firewall Microsoft Azure Front Door Microsoft Azure CDN Microsoft Azure Media Services Microsoft Azure Video Analytics Microsoft Azure Speech Services Microsoft Azure Text Analytics Microsoft Azure Computer Vision Microsoft Azure Face API Microsoft Azure Bot Service Microsoft Azure Logic Apps Microsoft Azure Functions Microsoft Azure App Service Microsoft Azure Container Instances Microsoft Azure Kubernetes Service Microsoft Azure Service Fabric Microsoft Azure Virtual Machines Microsoft Azure Virtual Network Microsoft Azure Firewall Microsoft Azure Front Door Microsoft Azure CDN Microsoft Azure Media Services Microsoft Azure Video Analytics Microsoft Azure Speech Services Microsoft Azure Text Analytics Microsoft Azure Computer Vision Microsoft Azure Face API	Neo4j Amazon Neptune IBM Graph ORACLE Microsoft Access Microsoft Excel Microsoft Word Microsoft PowerPoint Microsoft Outlook Microsoft OneDrive Microsoft Teams Microsoft SharePoint Microsoft Dynamics 365 Microsoft Power BI Microsoft Azure IoT Hub Microsoft Azure DevOps Microsoft Azure Active Directory Microsoft Azure Key Vault Microsoft Azure Storage Microsoft Azure Blob Storage Microsoft Azure File Storage Microsoft Azure Data Lake Storage Microsoft Azure Cosmos DB Microsoft Azure Databricks Microsoft Azure Synapse Analytics Microsoft Azure Machine Learning Microsoft Azure Cognitive Services Microsoft Azure Bot Service Microsoft Azure Logic Apps Microsoft Azure Functions Microsoft Azure App Service Microsoft Azure Container Instances Microsoft Azure Kubernetes Service Microsoft Azure Service Fabric Microsoft Azure Virtual Machines Microsoft Azure Virtual Network Microsoft Azure Firewall Microsoft Azure Front Door Microsoft Azure CDN Microsoft Azure Media Services Microsoft Azure Video Analytics Microsoft Azure Speech Services Microsoft Azure Text Analytics Microsoft Azure Computer Vision Microsoft Azure Face API	Teradata Amazon EMR IBM ORACLE Microsoft Access Microsoft Excel Microsoft Word Microsoft PowerPoint Microsoft Outlook Microsoft OneDrive Microsoft Teams Microsoft SharePoint Microsoft Dynamics 365 Microsoft Power BI Microsoft Azure IoT Hub Microsoft Azure DevOps Microsoft Azure Active Directory Microsoft Azure Key Vault Microsoft Azure Storage Microsoft Azure Blob Storage Microsoft Azure File Storage Microsoft Azure Data Lake Storage Microsoft Azure Cosmos DB Microsoft Azure Databricks Microsoft Azure Synapse Analytics Microsoft Azure Machine Learning Microsoft Azure Cognitive Services Microsoft Azure Bot Service Microsoft Azure Logic Apps Microsoft Azure Functions Microsoft Azure App Service Microsoft Azure Container Instances Microsoft Azure Kubernetes Service Microsoft Azure Service Fabric Microsoft Azure Virtual Machines Microsoft Azure Virtual Network Microsoft Azure Firewall Microsoft Azure Front Door Microsoft Azure CDN Microsoft Azure Media Services Microsoft Azure Video Analytics Microsoft Azure Speech Services Microsoft Azure Text Analytics Microsoft Azure Computer Vision Microsoft Azure Face API	Amazon S3 IBM ORACLE Microsoft Access Microsoft Excel Microsoft Word Microsoft PowerPoint Microsoft Outlook Microsoft OneDrive Microsoft Teams Microsoft SharePoint Microsoft Dynamics 365 Microsoft Power BI Microsoft Azure IoT Hub Microsoft Azure DevOps Microsoft Azure Active Directory Microsoft Azure Key Vault Microsoft Azure Storage Microsoft Azure Blob Storage Microsoft Azure File Storage Microsoft Azure Data Lake Storage Microsoft Azure Cosmos DB Microsoft Azure Databricks Microsoft Azure Synapse Analytics Microsoft Azure Machine Learning Microsoft Azure Cognitive Services Microsoft Azure Bot Service Microsoft Azure Logic Apps Microsoft Azure Functions Microsoft Azure App Service Microsoft Azure Container Instances Microsoft Azure Kubernetes Service Microsoft Azure Service Fabric Microsoft Azure Virtual Machines Microsoft Azure Virtual Network Microsoft Azure Firewall Microsoft Azure Front Door Microsoft Azure CDN Microsoft Azure Media Services Microsoft Azure Video Analytics Microsoft Azure Speech Services Microsoft Azure Text Analytics Microsoft Azure Computer Vision Microsoft Azure Face API	Amazon S3 IBM ORACLE Microsoft Access Microsoft Excel Microsoft Word Microsoft PowerPoint Microsoft Outlook Microsoft OneDrive Microsoft Teams Microsoft SharePoint Microsoft Dynamics 365 Microsoft Power BI Microsoft Azure IoT Hub Microsoft Azure DevOps Microsoft Azure Active Directory Microsoft Azure Key Vault Microsoft Azure Storage Microsoft Azure Blob Storage Microsoft Azure File Storage Microsoft Azure Data Lake Storage Microsoft Azure Cosmos DB Microsoft Azure Databricks Microsoft Azure Synapse Analytics Microsoft Azure Machine Learning Microsoft Azure Cognitive Services Microsoft Azure Bot Service Microsoft Azure Logic Apps Microsoft Azure Functions Microsoft Azure App Service Microsoft Azure Container Instances Microsoft Azure Kubernetes Service Microsoft Azure Service Fabric Microsoft Azure Virtual Machines Microsoft Azure Virtual Network Microsoft Azure Firewall Microsoft Azure Front Door Microsoft Azure CDN Microsoft Azure Media Services Microsoft Azure Video Analytics Microsoft Azure Speech Services Microsoft Azure Text Analytics Microsoft Azure Computer Vision Microsoft Azure Face API

ETL / DATA TRANSFORMATION	DATA INTEGRATION	DATA GOVERNANCE	DATA QUALITY
 <b>Talend</b>  <b>pentaho</b>  <b>alteryx</b>  <b>informatica</b>  <b>DataStage</b>  <b>IBM DataStage</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>	 <b>Informatica</b>  <b>Talend</b>  <b>alteryx</b>  <b>informatica</b>  <b>DataStage</b>  <b>IBM DataStage</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>	 <b>Informatica</b>  <b>Talend</b>  <b>alteryx</b>  <b>informatica</b>  <b>DataStage</b>  <b>IBM DataStage</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>  <b>pentaho Data Integration</b>  <b>informatica Data Integration</b>  <b>IBM Data Integration</b>	 <b>Talend</b>  <b>pentaho</b>  <b>alteryx</b>  <b>informatica</b> 

[illegible]

## ANALYTICS &amp; MACHINE INTELLIGENCE

[illegible][illegible][illegible]

The collage displays logos for the following companies:

- SEARCH:** Elasticsearch, Amazon Services, Oracle, Microsoft, Algolia, Coveo, Baidu, Lucidworks, ATTIVO, Swiftly, Experio, Acquisense, Amni, PIRANA.
- LOG ANALYTICS:** Splunk, Google Cloud Logging, Sumologic, Loggly, Logz.io, Loggly, Loggly, Loggly.
- SOCIAL ANALYTICS:** Hootsuite, Sprinklr, Netbase, Synthesio, SimpleReach, Bitly, Sprinklr, SimilarWeb.
- WEB / MOBILE / COMMERCE ANALYTICS:** Google Analytics, Mixpanel, Sigopt, Airtable, Resc, Granify, Amplitude.

## – APPLICATIONS – ENTERPRISE

SALES	MARKETING - B2B	MARKETING - B2C	CUSTOMER EXPERIENCE / SERVICE	HUMAN CAPITAL
 Pipedrive  Chorus.io  Envo  Tactile  Hustleflow	 App Annie  Latent  Sense  Tubular  Jango.io	 iStock  Silicon  Seed&Spark  Amperity  Bluecore  Invoca  Braze	 Galileo  Medallia  Gainight  Interlogix  ASAP  Kallid  Jahia  Mogenta  Gigamonks  Send	 iStock  Workday  Myle  Wade & Mandy

[illegible]

## – APPLICATIONS – INDUSTRY

The collage displays logos for fintech companies across seven sectors:

- ADVERTISING:** ApolloTech, MediaMath, critico!, IAS, ORACLE, calvert, gumgum, Apper, theRadioClick, Tapad.
- EDUCATION:** 猿辅导, 猿题库, KNEWTON, Declara, monocrow, anuradha, SPACEHACKS, GEOPIXY, KOREST.
- REAL ESTATE:** REDFIN, 贝壳找房, 房多多, Orchard, money.
- GOVT & INTELLIGENCE:** Palantir, SPENGO, Datagen, MARK33, ANDRILL, FiscalNote, 10101 PRIMER.
- COMMERCE:** FAIRRE, STITCH FIX, FitGood, STANDARD.
- FINANCE - LENDING:** affirm, Lendavo, ZEST, CASH, ZEPHYRUS, BLISSBANK, LULU, Upstart, 100Credit.
- INSURANCE:** ROOT, Genworth, Amdocs, SHR Technology, CAPE, FORTIS, ZEPHYRUS, ZEPHYRUS, ZEPHYRUS.

[illegible]

## - OPEN SOURCE

<b>FRAMEWORKS</b> 	<b>QUERY / DATA FLOW</b> 	<b>DATA ACCESS &amp; DATABASES</b> 	<b>ADMINISTRATION &amp; PIPELINES</b> 	<b>STREAMING &amp; MESSAGING</b> 	<b>STAT TOOLS &amp; LANGUAGES</b> 	<b>AI OPS &amp; INFRA</b> 	<b>AI / MACHINE LEARNING / DEEP LEARNING</b> 	<b>SEARCH</b> 	<b>LOGGING &amp; MONITORING</b> 	<b>VISUALIZATION</b> 	<b>COLLABORATION</b> 	<b>SECURITY</b> 
-----------------------	------------------------------	--	---	--------------------------------------	---------------------------------------	-------------------------------	--	-------------------	-------------------------------------	--------------------------	--------------------------	---------------------

## DATA SOURCES & APIs:

The banner displays logos for various data and analytics companies, organized into six categories:

- DATA MARKETPLACES & DISCOVERY:** AWS Data Exchange, snowflake, DAWEX, EXPLOREM, data.world, narrative.
- FINANCIAL & ECONOMIC DATA:** Bloomberg, THOMSON REUTERS, DOW JONES, Quandl, CAPITALIS, CBRE, FLUID, STOCKHOLM, StockTelligence, xignite, predata, MOC, thk, SmallNews.
- AIR / SPACE / SEA:** Global Insight, AIRBOTICS, spire, WINDWARD, EXOLABS, Synapse.
- PEOPLE / ENTITIES:** zoominfo, CxI, Experian, InsideView, Quantcast, BASIS, Demyst, melissa, SAFEGRAPH.
- LOCATION INTELLIGENCE:** FOURSQUARE, mapbox, esri, PlaceIQ, KAYAK, Radar, Mapillary, cuebic, OpenStreetMap.
- OTHER:** DATA GOV, IMAGENET, Lab4Life, CRUX.

## DATA RESOURCES

**DATA SERVICES**

- QuantumBlack
- IBM
- Booz | Allen | Hamilton
- Kaggle
- ElectrifAI
- fractalx
- EXL
- DataKind
- innopLUS

**INCUBATORS & SCHOOLS**

- Playbook
- General Assembly
- DataCamp
- Data360
- galvanize
- metric
- The Data Incubator
- 100XVC

**RESEARCH**

- OpenAI
- facebook research
- MIT
- MIR
- VECTOR INSTITUTE
- AT&T

# Outline



1

- Big Data. Big Data Science

2

- **Why Big Data? Google and the MapReduce programming model**

3

- Big Data technologies: Hadoop / Spark ecosystem

4

- Big Data Analytics: Libraries for Data Analytics in Big Data. Case studies

5

- Final Comments



# Why Big Data?

The MapReduce programming model

# Pulsars detection from radiofrequency by Imaging stacking

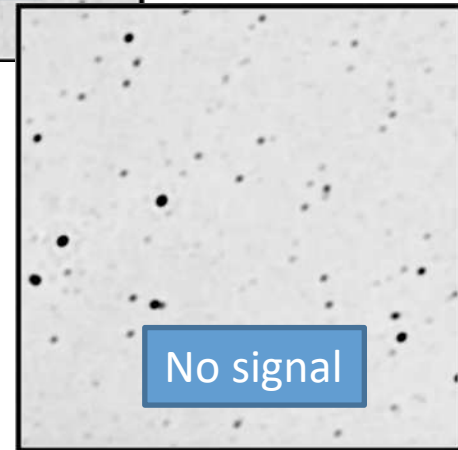
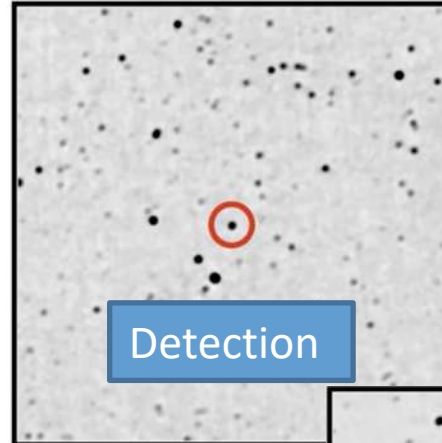
Use flux density as a measure of spectral power



Positive detection when signal is 5 times noise in the local radius



Aggregate regions of an image to amplify signal to noise ratio.



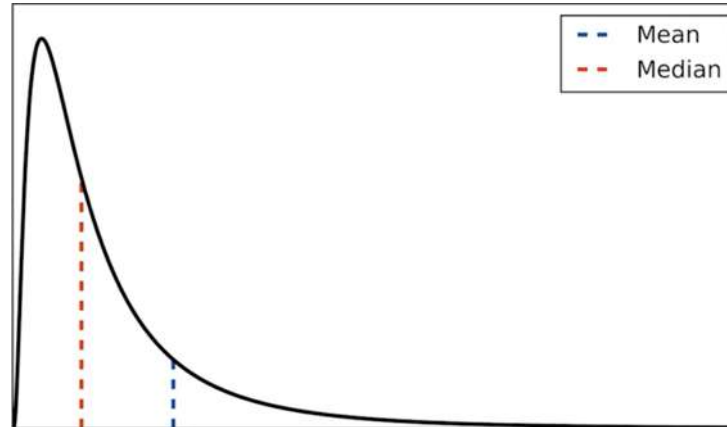
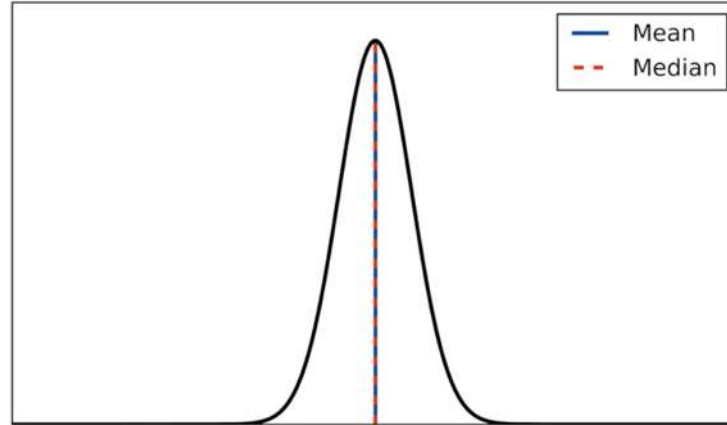
# Procedure for pulsar detection: How to aggregate values?

Find images containing positions of known pulsars

Crop and shift each pulsar to the centre of each image

Aggregate brightness at each pixel

Create new image





# Challenge: A example on scalability.

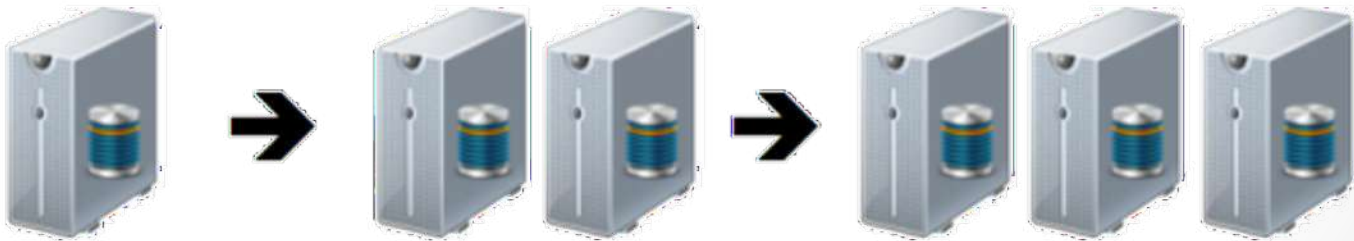
- Compute the mean is simpler, but does not provide accurate results
- Compute the median leads to a significant increase of computational resources. Why?
  - Let's say we start from 600,000 images
  - 1 image has a 200x200 pixel resolution --> 40,000 pixels
  - 1 pixel needs 8 bytes in memory
  - The amount of memory needed: 192 GB!!
- All data cannot be in memory at the same time. Available solutions:
  - Invest money on better equipment
  - Reformulation of the problem
  - Smart design of an scalable solution -> **Go for linear and parallel solutions**

# Dealing data intensive applications: Scale-up vs. Scale-out

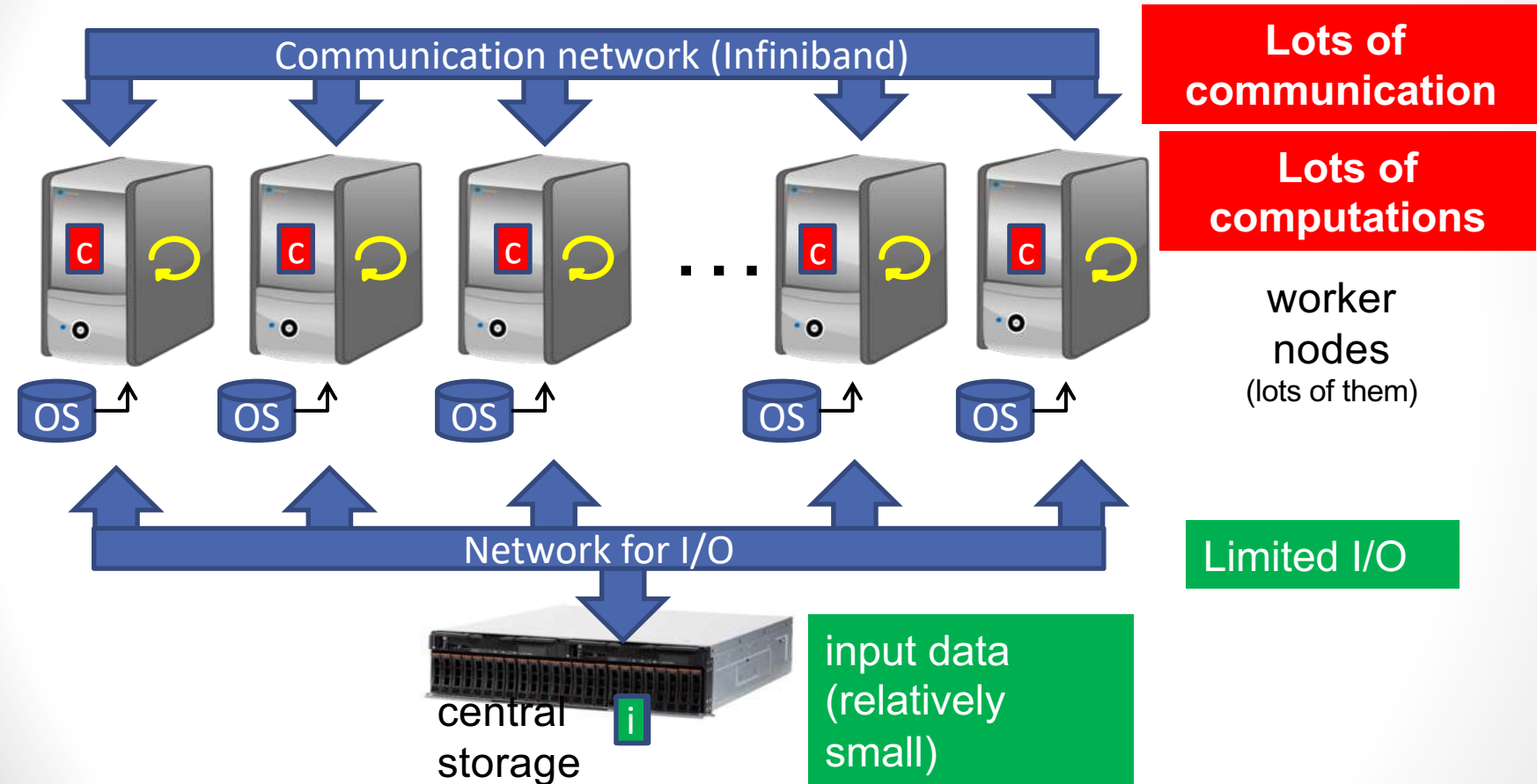
Scale-Up



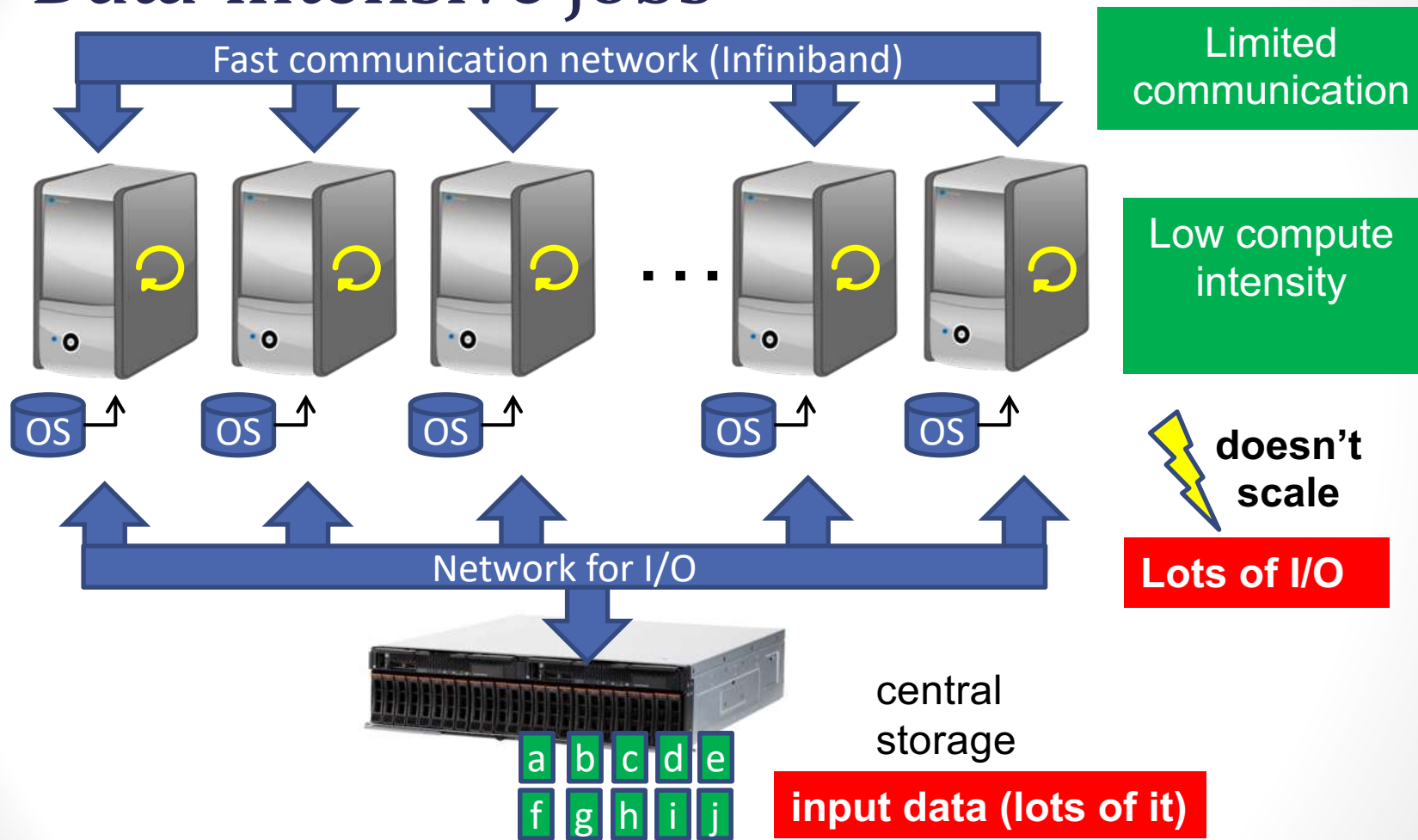
Scale-Out



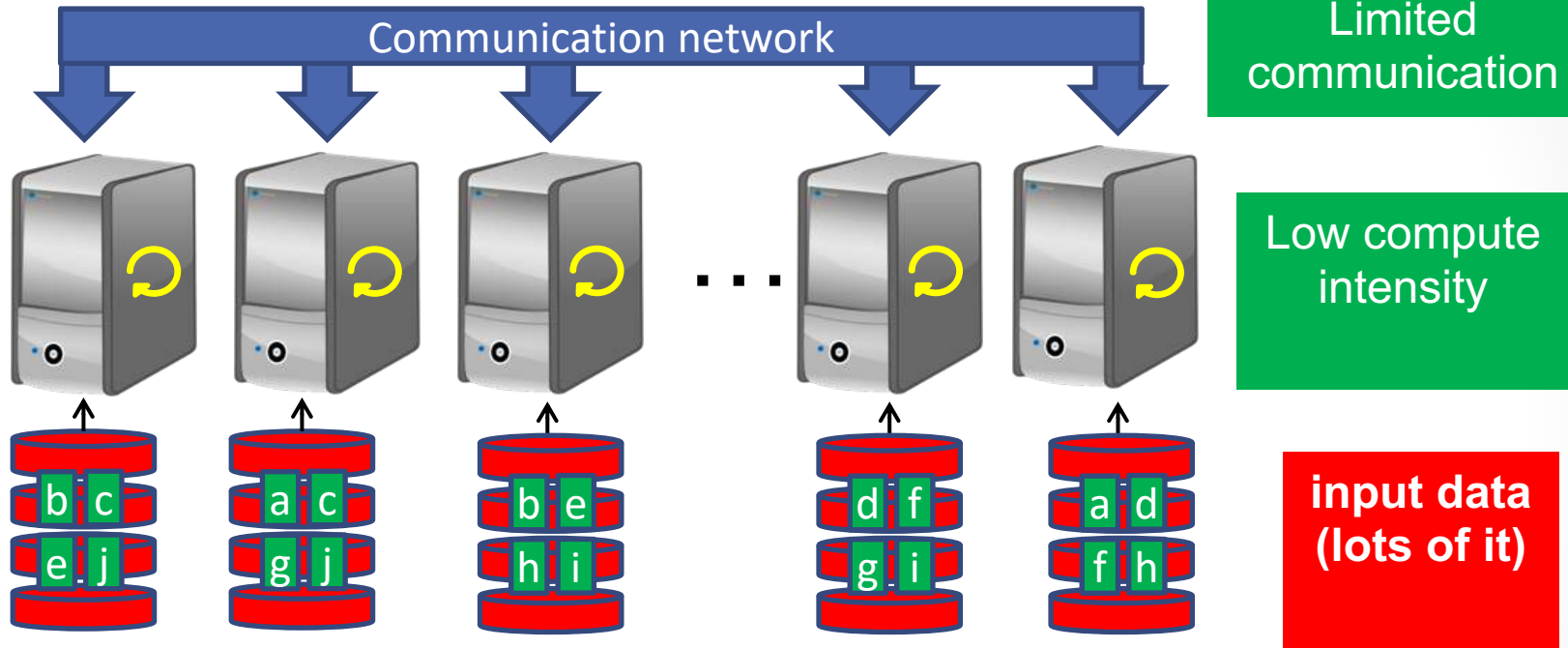
# Traditional HPC way of doing things



# Data-intensive jobs



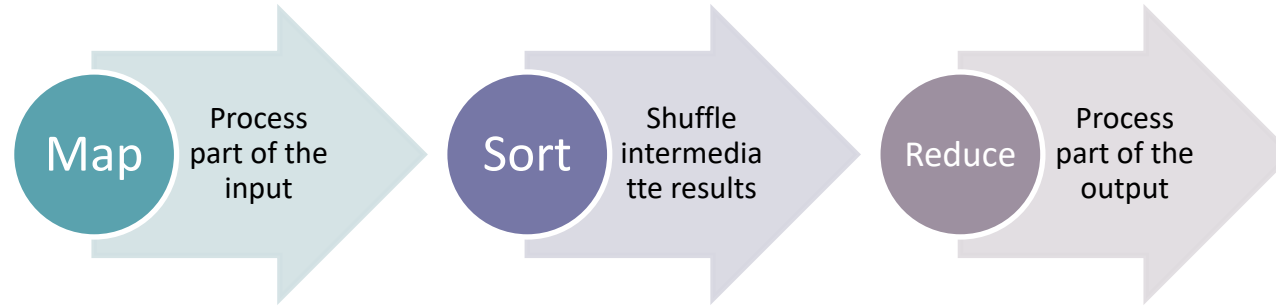
# Data-intensive jobs



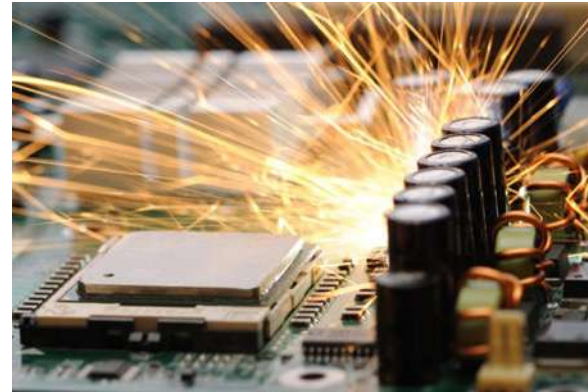
**Solution:** store data on local disks of the nodes that perform computations on that data (“**data locality**”)



# We need new parallel programming paradigm



Example: vote counting



With fault tolerance

# Distributed systems in Big Data.

Aim: to apply an operation to all data

- One machine cannot process or store all data
  - Data is distributed in a cluster of computing nodes
  - It does not matter which machine executes the operation
  - It does not matter if it is run twice in different nodes (due to failures or straggler nodes)
  - We look for an abstraction of the complexity behind distributed systems
- **DATA LOCALITY is crucial**
  - Avoid data transfers between machines as much as possible

# Distributed systems in Big Data

## New programming model: MapReduce

- *“Moving computation is cheaper than moving computation and data at the same time”*
- **Idea**
  - Data is distributed among nodes (**distributed file system**)
  - Functions/operations to process data are distributed to all the computing nodes
  - Each computing node works with the data stored in it
  - Only the necessary data is moved across the network

# Hadoop Distributed File System: HDFS

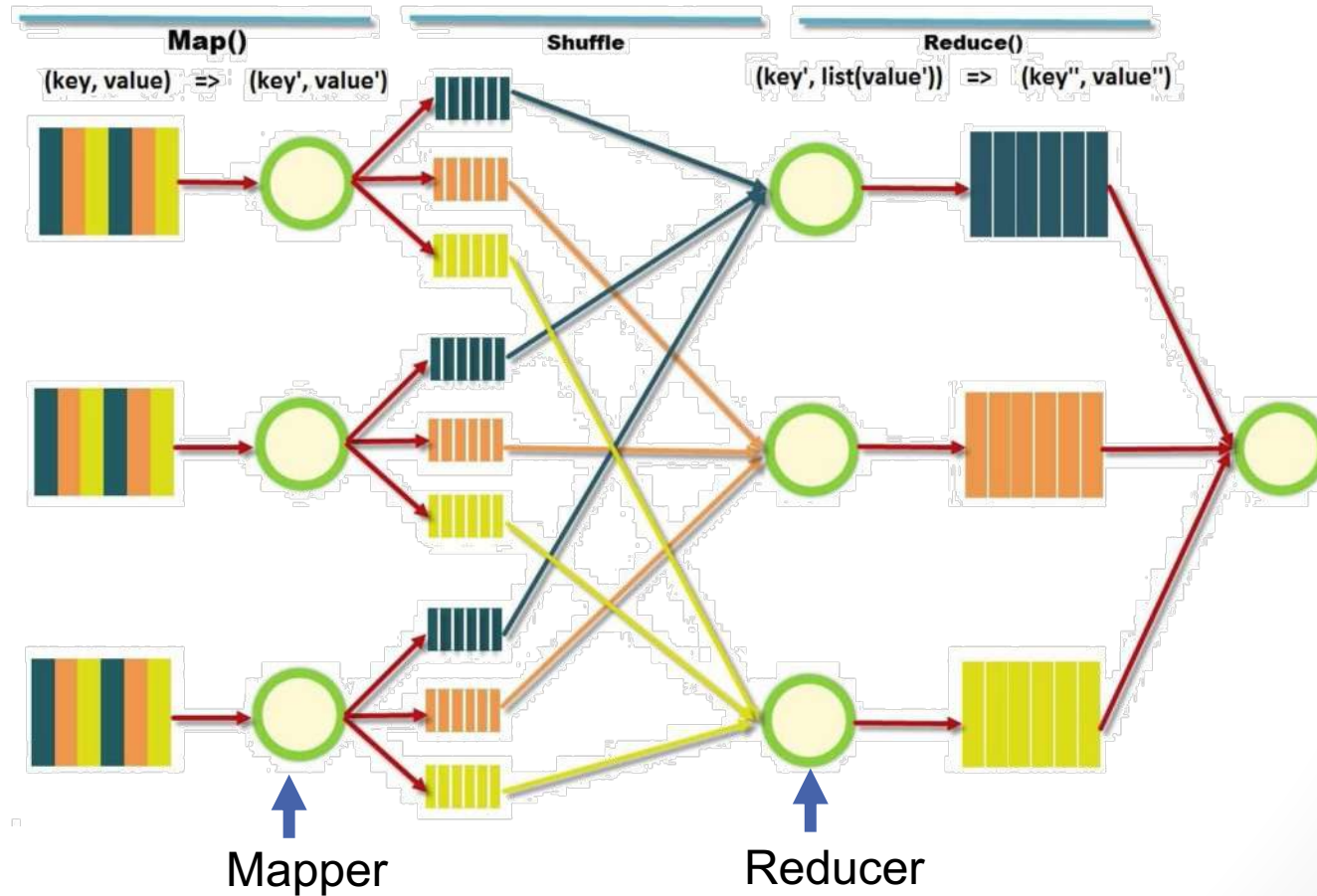
- Distributed File System written in Java
- Scales to clusters with **thousands of computing nodes**
  - Each node stores part of the data in the system
- **Fault tolerant** due to data replication
- Designed for big files and low-cost hardware
  - GBs, TBs, PBs..
- **Efficient for read and append operations** (random updates are rare). **Overhead** when writing into disk.
- High throughput (for bulk data) more important than low latency

# MapReduce

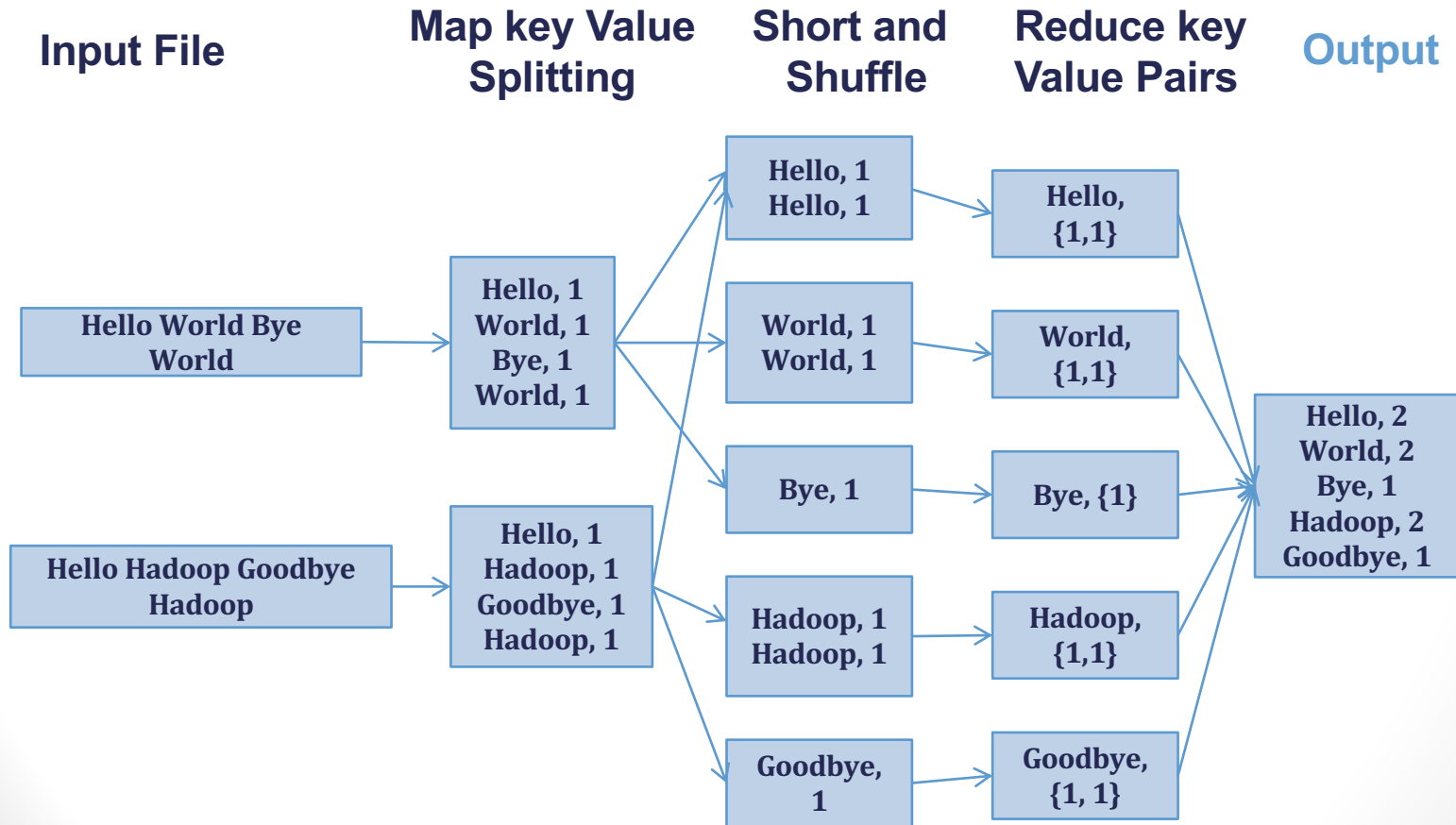
- Parallel Programming model
- **Divide & conquer strategy**
  - **divide**: partition dataset into smaller, independent chunks to be processed in parallel (map)
  - **conquer**: combine, merge or otherwise aggregate the results from the previous step (reduce)
- Based on **simplicity** and **transparency** to the programmers, and assumes **data locality**
- Becomes popular thanks to the open-source project Hadoop! (Used by [Google](#), [Facebook](#), [Amazon](#), ...)



# MapReduce: How it works



# MapReduce: WordCount Example



# MapReduce: Features

	<b>Automatic parallelization:</b>	<ul style="list-style-type: none"><li>• Size of the INPUT DATA → multiple MAP tasks</li><li>• Number of &lt;key, value&gt; intermediate partitions → several REDUCE tasks</li></ul>
	<b>Scalability:</b>	<ul style="list-style-type: none"><li>• It works over any cluster of nodes /processor</li><li>• Can work from 2 to 10,000 machines</li></ul>
	<b>Programming transparency</b>	<ul style="list-style-type: none"><li>• Management of the failures of the machine</li><li>• Management of the communication among machines</li></ul>

# MapReduce

## Main Features

- Scalable architectures
- Optimized planning
- Elasticity and availability
- Flexibility
- Security and authentication

## Limitations

- Machine Learning: Iterative computation
- Graph processing
- Real time processing (streams)
- Functionality of process communication
- Difficultiy in the MR-like implementation

# Outline



1

- Big Data. Big Data Science

2

- Why Big Data? Google and the MapReduce programming model

3

- **Big Data technologies: Hadoop / Spark ecosystem**

4

- Big Data Analytics: Libraries for Data Analytics in Big Data. Case studies

5

- Final Comments



# Big Data Technologies

Hadoop Ecosystem

# Hadoop Ecosystem:

## Apache Hadoop Modules



<http://hadoop.apache.org/>

### Hadoop Common:

Utilities  
supporting  
other  
Hadoop  
modules.

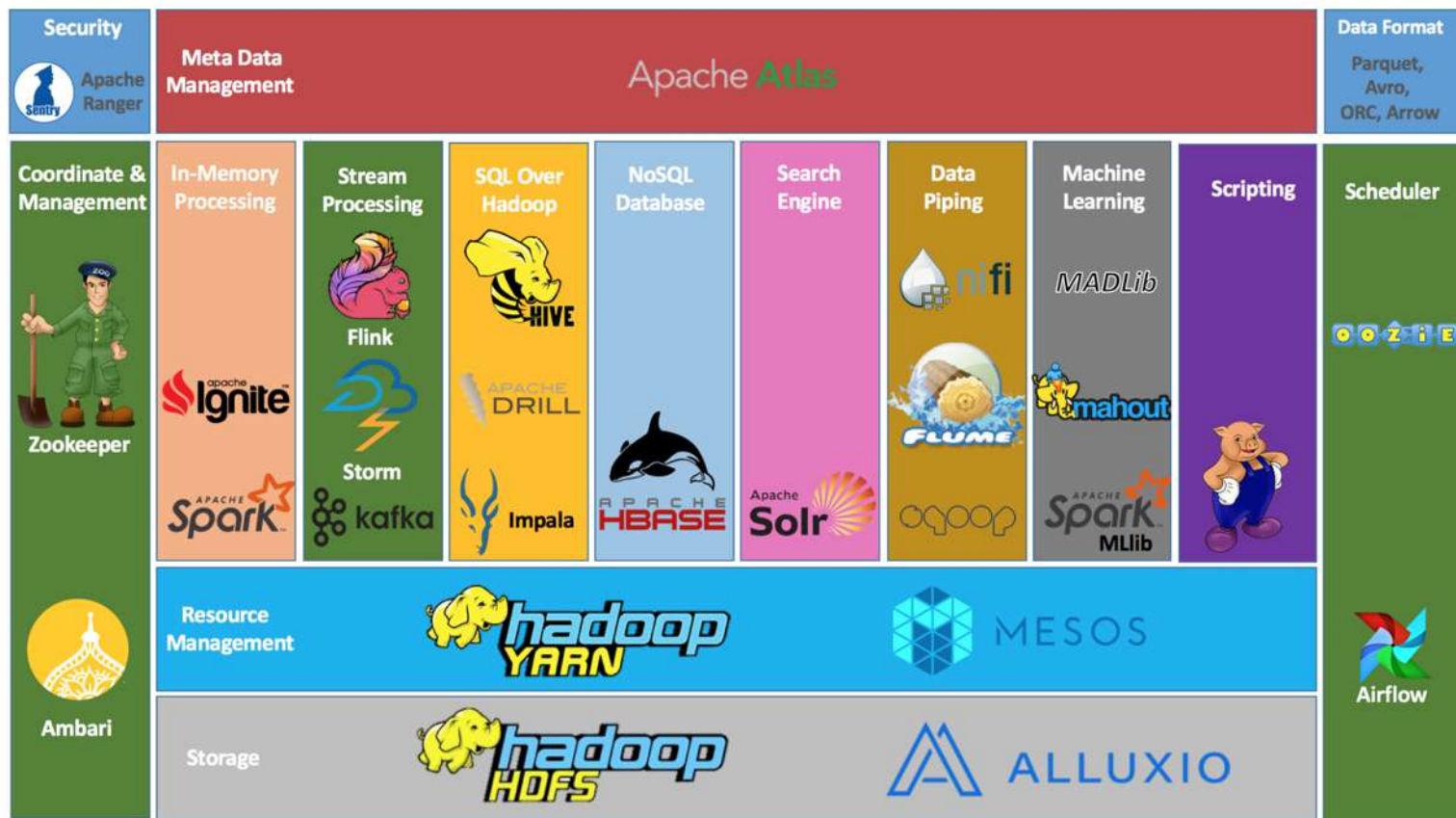
### Hadoop Distributed File System (HDFS):

The file  
system that  
grant  
access

### Hadoop YARN:

Framework for management the programming  
resources.



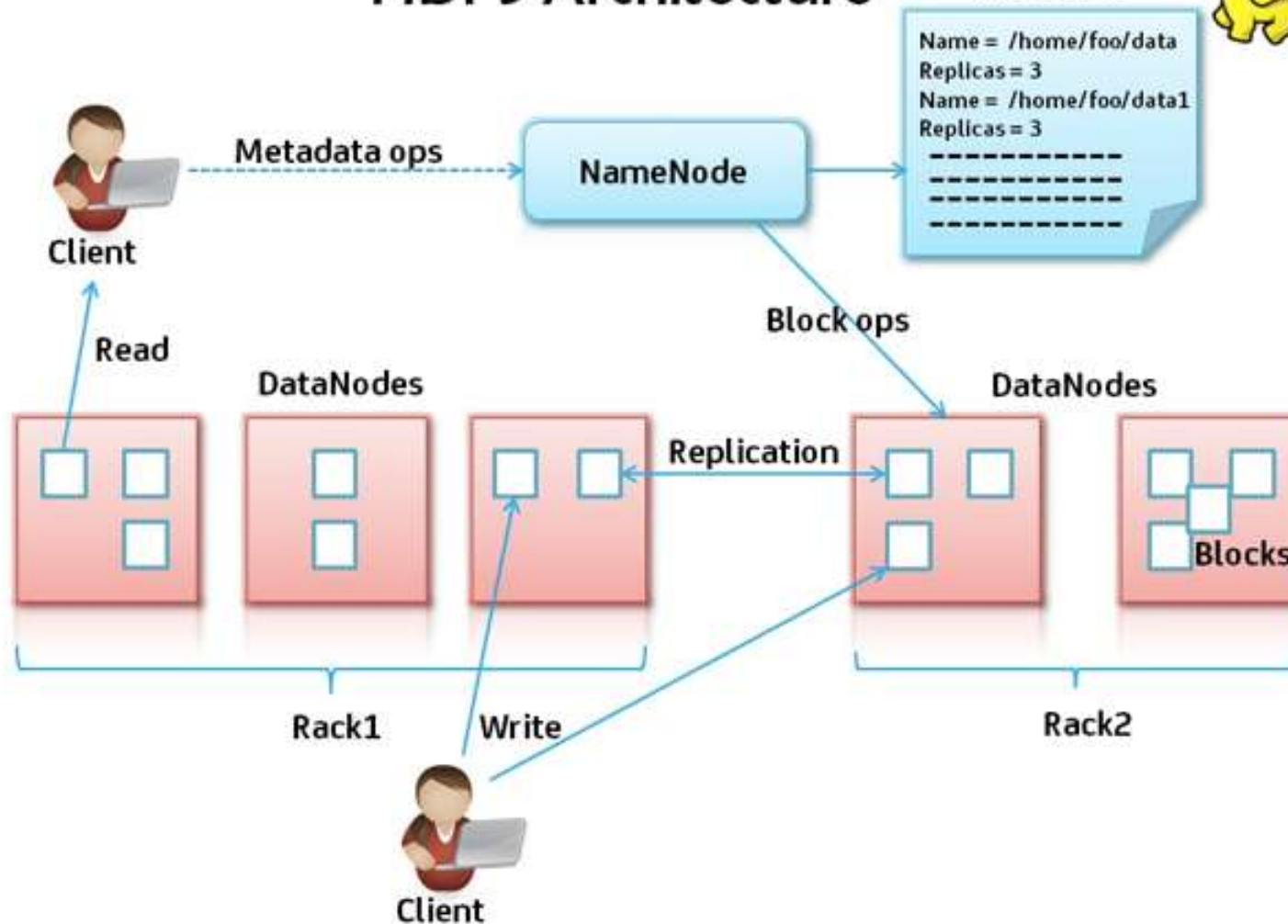


## Hadoop Ecosystem

Complete technologies to address Big Data Analytics

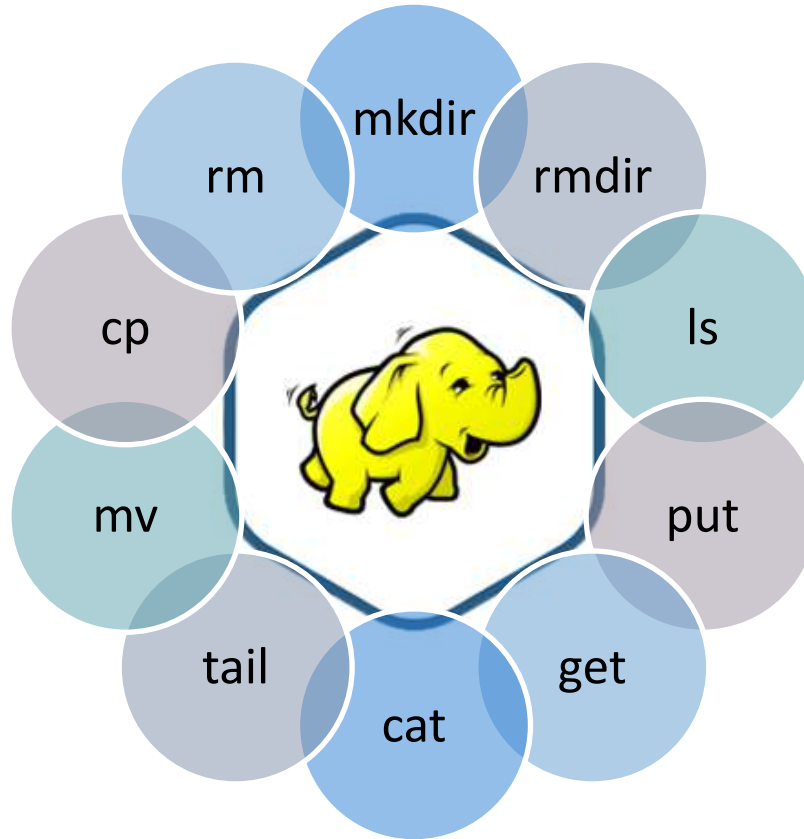
# HDFS Architecture

Metadata



# File management with HDFS.

Command “hdfs dfs -<option>”



# Apache Spark (Birth 2009-2010)



Fast and Expressive Cluster Computing  
Engine Compatible with Apache Hadoop

Up to **10x** faster on disk,  
**100x** in memory

**Efficient**

- General execution graphs
- In-memory storage

**2-5x** less code

**Usable**

- Rich APIs in Java, **Scala**, Python
- Interactive shell

# What is Spark?



Data processing engine (only)

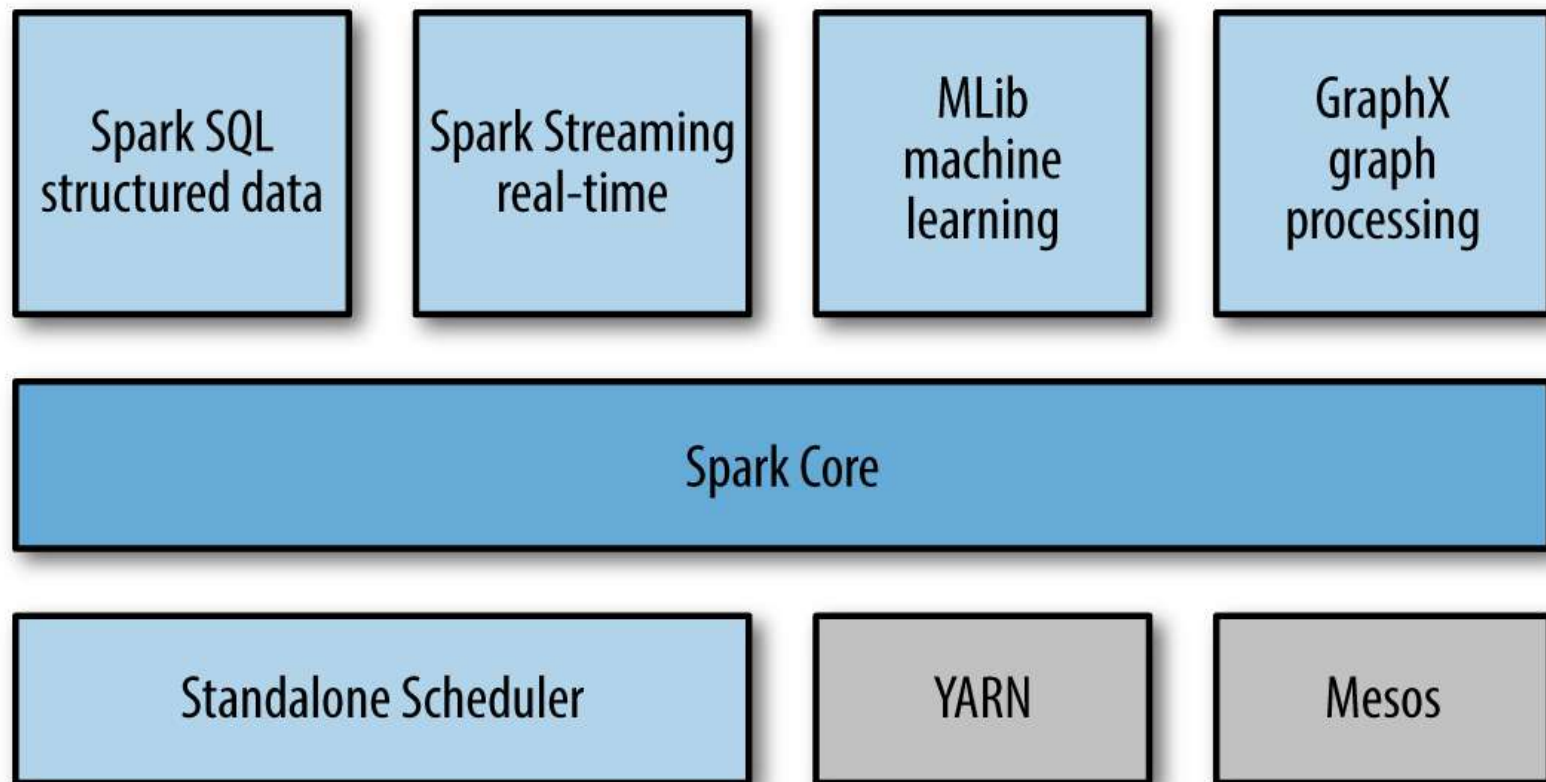
Without a distributed file system

- Uses other existing DFS
  - HDFS, NoSQL...
  - Hadoop is not a prerequisite

Works with different cluster management tools

- Hadoop (YARN)
- Mesos
- Standalone mode (included in Spark)

# What is Spark?





# Spark Goal

Provide distributed memory abstractions for clusters to support apps with working sets

## **Retain the attractive properties of MapReduce:**

- Fault tolerance (for crashes & stragglers)
- Data locality
- Scalability

Initial Solution: augment data flow model with “resilient distributed datasets” (RDDs)

# Apache Spark

- KEY Concept: RDD (Resilient Distributed Datasets)
  - Fault-tolerant collection of elements that can be operated on in parallel.
- There are two ways to create RDDs:
  - Parallelizing an existing collection in your driver program
  - Referencing a dataset in an external storage system, such as a shared filesystem, HDFS, Hbase.
- Objects spread across a cluster, stored in RAM or on Disk
- *Can be cached for future reuse*

# Apache Spark: Operations

## Transformations

Create a new RDD / dataset from an existing one

Lazy in nature: executed only when some action is performed

Example

- Map(func)
- Filter(func)
- Distinct()

## Actions

Returns a value or exports data after performing a computation

Example

- Count()
- Reduce(func)
- Collect()

## Persistence

Caching RDD / dataset in-memory for future operations

Store on disk or RAM or mixed

Example

- Persist()
- Cache

# RDDs operations. Word Count example

```
>>> lines = sc.textFile("README.md") # Creates an RDD
>>> lines.count() # Counts the number of elements in the
RDD
```

```
127
```

```
>>> lines.first() # First element of the RDD -> 1st line
of README.md
```

**u'# Apache Spark`**

```
text_file = sc.textFile("hdfs://...")
counts = text_file.flatMap(lambda line: line.split(" "))
                    .map(lambda word: (word, 1))
                    .reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://...")
```

# Current Spark Environment

## Spark has released version 3.0

- June 2020: 10 year anniversary!
- Spark 3.0 is said to be 2 times faster than Spark 2.4
- SparkSQL is the main component for the update: APIs for DataFrames.
- PySpark has become the widest used language:
  - Improvement of functionality and usability
  - Pandas UDF redesign and types
  - Error handling

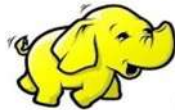
## Feature highlights

- Adaptive query execution;
- Dynamic partition pruning;
- ANSI SQL compliance;
- Significant improvements in pandas APIs;
- New UI for structured streaming;
- Up to 40x speedups for calling R user-defined functions;
- Accelerator-aware scheduler;
- SQL reference documentation.

<https://spark.apache.org/releases/spark-release-3-0-0.html>

# Which Language Should I Use?

- Standalone programs can be written in any, but console is only Python & Scala
- Python developers: can stay with Python for both
- Java developers: consider using Scala for console (to learn the API)
- **Performance:** Java / Scala will be faster (statically typed), but Python can do well for numerical work with NumPy
- **Effectiveness:** Scala and python code is less lengthy than java. Scala has faster performance.
- **Comfort:** Moving from one language to another?
- **Future Prospects:** Many supported languages, Scala keeps being the main one.



# How do I get access to a Big Data platform?

Cloud platforms with  
complete IaaS / PaaS



Google Cloud Platform

Pseudo-distributed installation

Pre-compiled Hadoop and  
Spark stand-alone

Installation in a cluster

cloudera



kubernetes



MESOS



# Outline



1

- Big Data. Big Data Science

2

- Why Big Data? Google and the MapReduce programming model

3

- Big Data technologies: Hadoop / Spark ecosystem

4

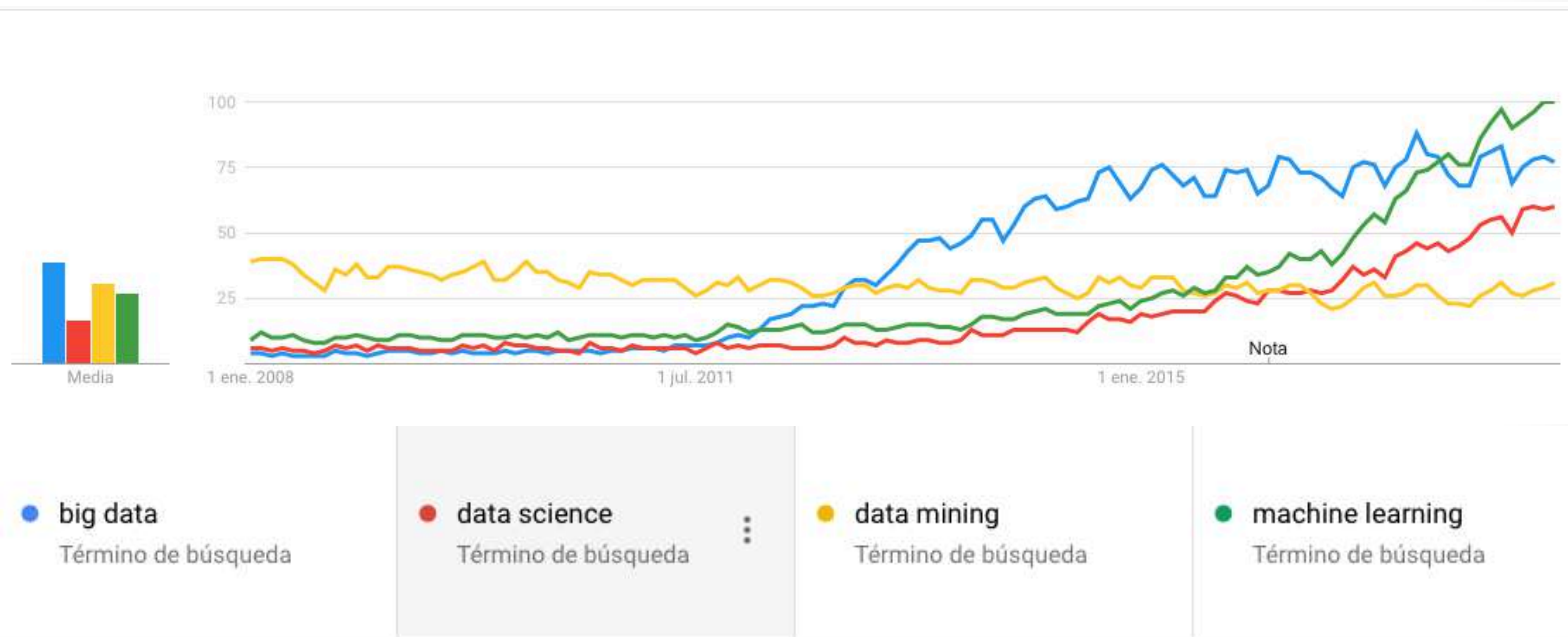
- **Big Data Analytics: Libraries for Data Analytics in Big Data. Case studies**

5

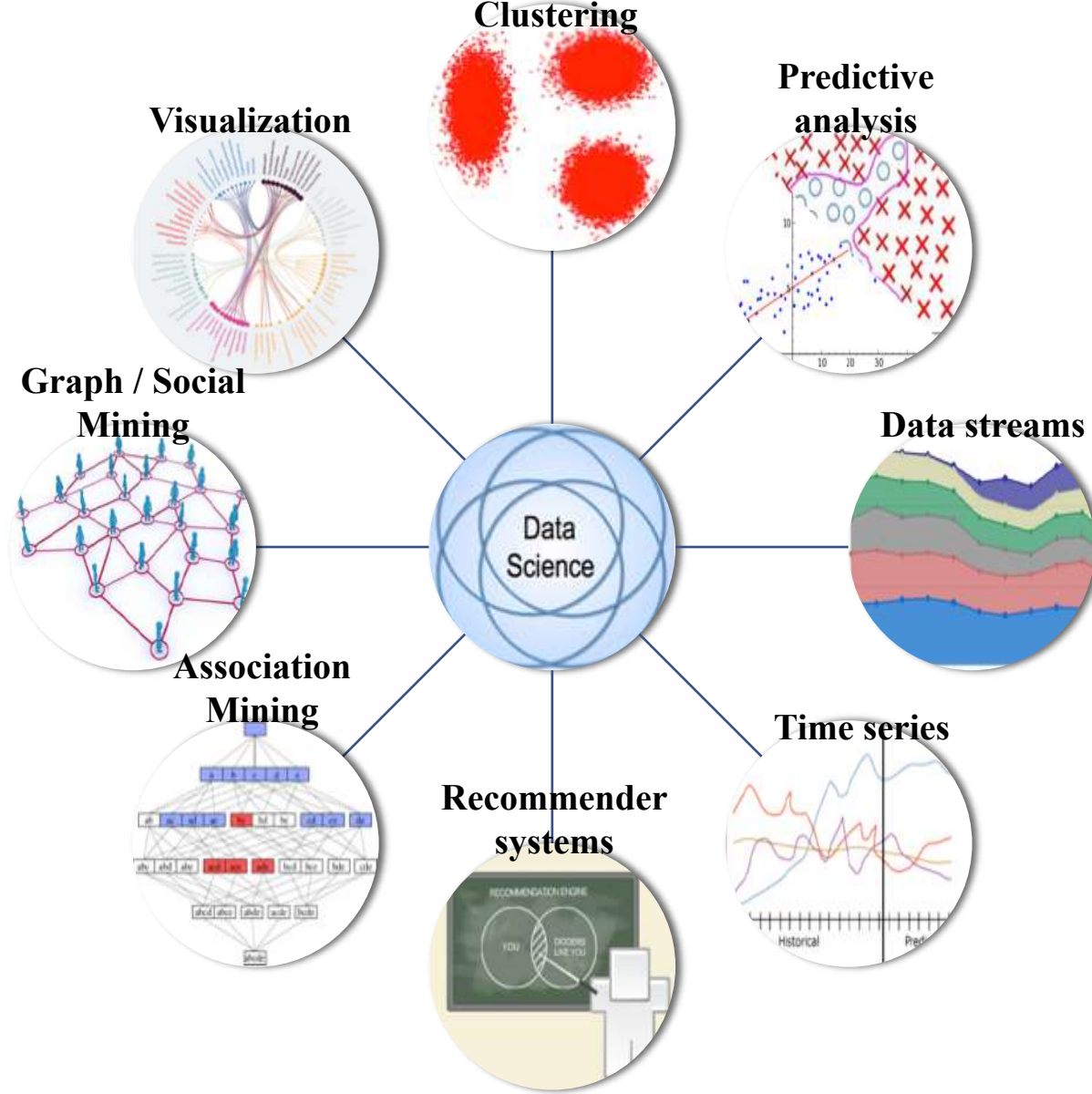
- Final Comments



# Evolution of the popularity of Big Data and related concepts



[Google Trends](#)



# Machine Learning for Big Data

- Data mining techniques have demonstrated to be very useful tools to extract new valuable knowledge from data.
- Knowledge extraction process from big data has become a very difficult task for most of the data mining tools.
- The main challenges are to deal with:
  - The increasing scale of data
    - at the level of instances
    - at the level of features
  - The complexity of the problem.
  - ...and many other points (unstructured data, redundancy, ...)

# Big Data Analytics Tools: Spark MLlib

- Goal: make practical Machine Learning scalable and easy.
- At a high level, it provides tools such as:
  - **ML Algorithms:** common learning algorithms such as classification, regression, clustering, and collaborative filtering
  - **Featurization:** feature extraction, transformation, dimensionality reduction, and selection
  - **Pipelines:** tools for constructing, evaluating, and tuning ML Pipelines
  - **Persistence:** saving and load algorithms, models, and Pipelines
  - **Utilities:** linear algebra, statistics, data handling, etc.





## MLlib: Main Guide

- [Basic statistics](#)
- [Data sources](#)
- [Pipelines](#)
- [Extracting, transforming and loading data](#)
- [Classification and Regression](#)
- [Clustering](#)
- [Collaborative filtering](#)
- [Frequent Pattern Mining](#)
- [Model selection and tuning](#)
- [Advanced topics](#)

## MLlib: RDD-based API

- [Data types](#)
- [Basic statistics](#)
- [Classification and regression](#)
- [Collaborative filtering](#)
- [Clustering](#)
- [Dimensionality reduction](#)
- [Feature extraction and transformation](#)
- [Frequent pattern mining](#)
- [Evaluation metrics](#)
- [PMML model export](#)
- [Optimization \(developer\)](#)

# Machine Learning Library (MLlib) Guide

MLlib is Spark's machine learning (ML) library. Its goal is to make practical machine learning scalable and easy. At a high level, it provides tools such as:

- **ML Algorithms:** common learning algorithms such as classification, regression, clustering, and collaborative filtering
- **Featurization:** feature extraction, transformation, dimensionality reduction, and selection
- **Pipelines:** tools for constructing, evaluating, and tuning ML Pipelines
- **Persistence:** saving and load algorithms, models, and Pipelines
- **Utilities:** linear algebra, statistics, data handling, etc.

## Announcement: DataFrame-based API is primary API

**The MLlib RDD-based API is now in maintenance mode.**

As of Spark 2.0, the [RDD-based](#) APIs in the `spark.mllib` package have entered maintenance mode. The primary Machine Learning API for Spark is now the [DataFrame-based](#) API in the `spark.ml` package.

*What are the implications?*

- MLlib will still support the RDD-based API in `spark.mllib` with bug fixes.
- MLlib will not add new features to the RDD-based API.
- In the Spark 2.x releases, MLlib will add features to the DataFrames-based API to reach feature parity with the RDD-based API.

*Why is MLlib switching to the DataFrame-based API?*

- DataFrames provide a more user-friendly API than RDDs. The many benefits of DataFrames include Spark Datasources, SQL/DataFrame queries, Tungsten and Catalyst optimizations, and uniform APIs across languages.
- The DataFrame-based API for MLlib provides a uniform API across ML algorithms and across multiple languages.
- DataFrames facilitate practical ML Pipelines, particularly feature transformations. See the [Pipelines guide](#) for details.

*What is "Spark ML"?*

- "Spark ML" is not an official name but occasionally used to refer to the MLlib DataFrame-based API. This is majorly due to the `org.apache.spark.ml` Scala package name used by the DataFrame-based API, and the "Spark ML Pipelines" term we used initially to emphasize the pipeline concept.



# Spark ML (Mllib): Basic Statistics

- Calculation of descriptive statistics: mean, variance, max, min, etc.
- Computation of the degree of correlation among variables.
- Stratified sampling, through two methods `sampleByKey` and `sampleByKeyExact`.
- Hypothesis contrast, for example through the Chi-square test.
- Generation of random data following a certain distribution, Normal or Poisson, for example.

# Spark ML (Mllib): Classification and Regression

## Classification models

- Simple, binomial, and multinomial logistic regression
- Random forest classifier
- Gradient-boosted tree classifier
- Multilayer perceptron classifier
- Linear Support Vector Machine
- One-vs-Rest classifier (a.k.a. One-vs-All)
- Naive Bayes
- Factorization machines classifier

## Regression models

- Simple and Generalised Linear regression
- Decision tree regression
- Random forest regression
- Gradient-boosted tree regression
- Survival regression
- Isotonic regression
- Factorization machines regressor

# Spark ML (Mllib): Unsupervised

## Clustering algorithms

- K-means
- Latent Dirichlet allocation (LDA)
- Bisecting k-means
- Gaussian Mixture Model (GMM)
- Power Iteration Clustering (PIC)

## Frequent pattern mining

- FP-Growth
- PrefixSpan

## spark-packages.org

- External, community-managed list of third-party libraries, add-ons, and applications that work with Apache Spark.

## Infrastructure Projects

- Zeppelin - Multi-purpose notebook supports 20+ language backends, including Apache Spark
- MLflow - Open source platform to manage the ML lifecycle, including deploying models from diverse machine learning libraries on Apache Spark.

## Applications Using Spark

- Apache Mahout - Previously on Hadoop MapReduce, now switched using Spark as backend
- Natural Language Processing for Apache Spark - A library to provide simple, performant, and accurate NLP annotations for machine learning pipelines

## Performance, Monitoring, and Debugging Tools for Spark

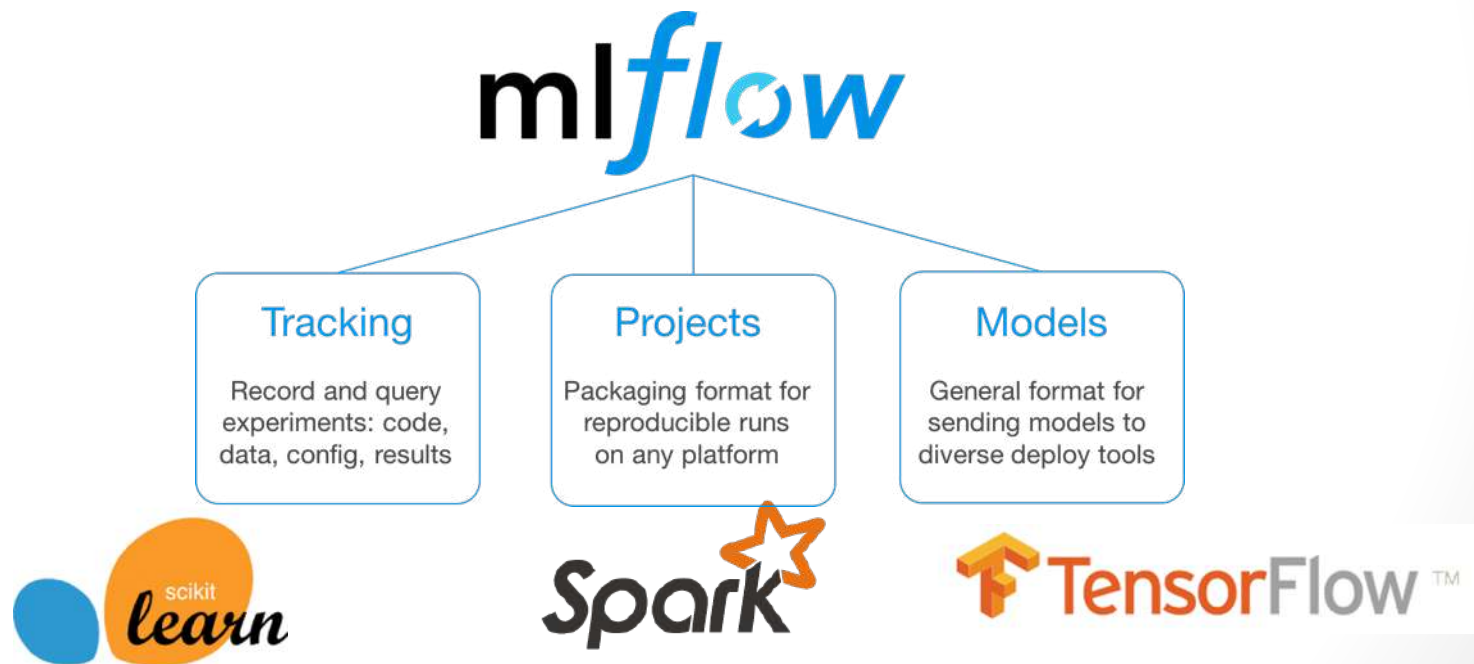
- Performance and debugging library - A library to analyze Spark and PySpark applications for improving performance and finding the cause of failures
- Data Mechanics Delight - Delight is a free, hosted, cross-platform Spark UI alternative backed by an open-source Spark agent. It features new metrics and visualizations to simplify Spark monitoring and performance tuning.

## Additional Language Bindings

- C# / .NET: Mobius: C# and F# language binding and extensions to Apache Spark
- Julia: Spark.jl
- Kotlin: Kotlin for Apache Spark

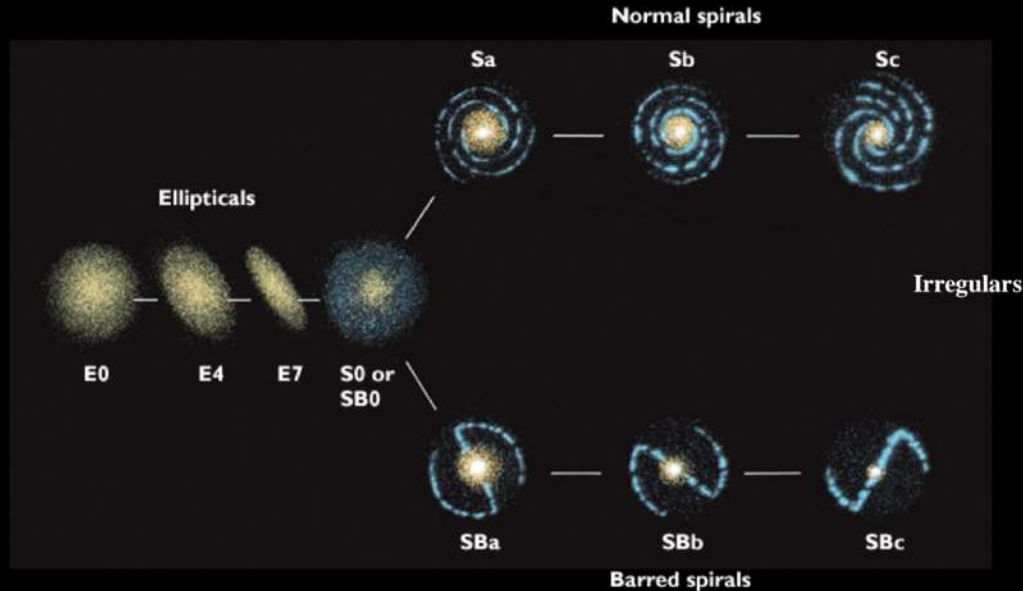
# A 4th generation of ML libraries?

A cross-cloud open source framework for the complete Machine Learning lifecycle.



# Case Study: Classification of Galaxies

## Hubble Tuning Fork Classification



# Collecting Data: Sloan Digital Sky Survey (SDSS) + Galaxy zoo project

- Manual labelling (high confidence galaxies)
- Three major classes: elliptical, spiral and merger
- Descriptive features:
  - Sloan filter colors
  - Color of the Galaxy: Age
  - Ellipticity of the Galaxy: Shape by minor / major axis
  - Luminosity profile



*“The degree of success attained by the method depends largely upon the significance of the features selected as the basis of classification”*

Avoid raw pixels: high dimension (only for DL)

Edwin Hubble – The Realm of the Nebulae



# Classification Results for Galaxy dataset (10fcv, default parameters)

## Spark Decision Tree

- Overall accuracy: 79.1% (4/5 of the galaxies well-classified)
- Possible drawbacks:
  - Overfitting risk
  - High bias on noisy data

## Spark Random Forest

- Overall accuracy: 83.1% (better approach)
- Benefits of RF:
  - More stable approach
  - Feature importance interpretation

Source name	u-g colour	g-r colour	r-i colour	r-z colour	Ellipticity 1	Ellipticity 2	Luminosity (u)	Luminosity (r)	Luminosity (z)	Class
Galaxy 1	1.86	0.67	0.42	0.31	0.59	0.57	0.60	0.46	0.33	merger
Galaxy 2	2.11	0.97	0.57	0.34	0.63	0.63	0.34	0.43	0.31	merger
Galaxy 3	1.99	1.01	0.43	0.35	0.84	0.83	0.46	0.31	0.31	elliptical
Galaxy 4	1.98	0.93	0.45	0.30	0.86	0.87	0.33	0.30	0.30	elliptical
Galaxy 5	1.44	0.66	0.35	0.27	0.71	0.72	0.57	0.44	0.43	spiral
Galaxy 6	1.70	0.74	0.40	0.29	0.57	0.56	0.47	0.45	0.46	spiral
...	...	...	...	...	...	...	...	...	...	...

Classification features

# Outline



1

- Big Data. Big Data Science

2

- Why Big Data? Google and the MapReduce programming model

3

- Big Data technologies: Hadoop / Spark ecosystem

4

- Big Data Analytics: Libraries for Data Analytics in Big Data. Case studies

5

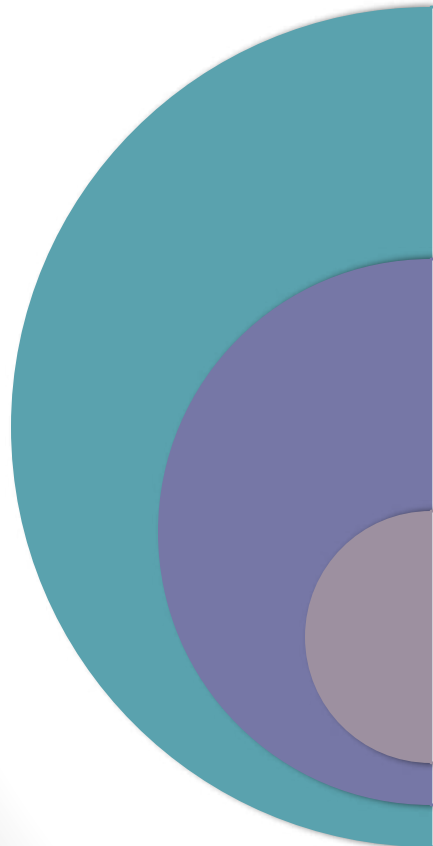
- **Final Comments**



# Final Comments

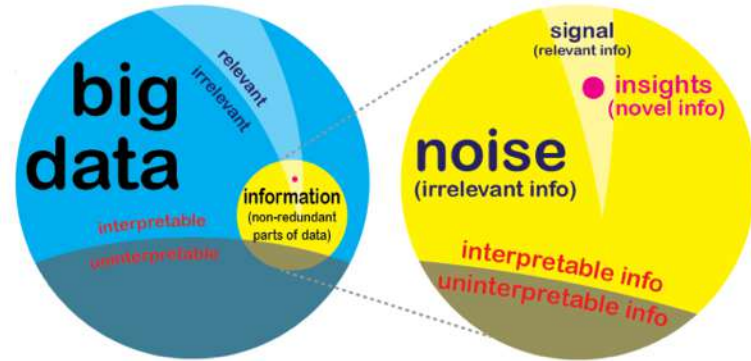
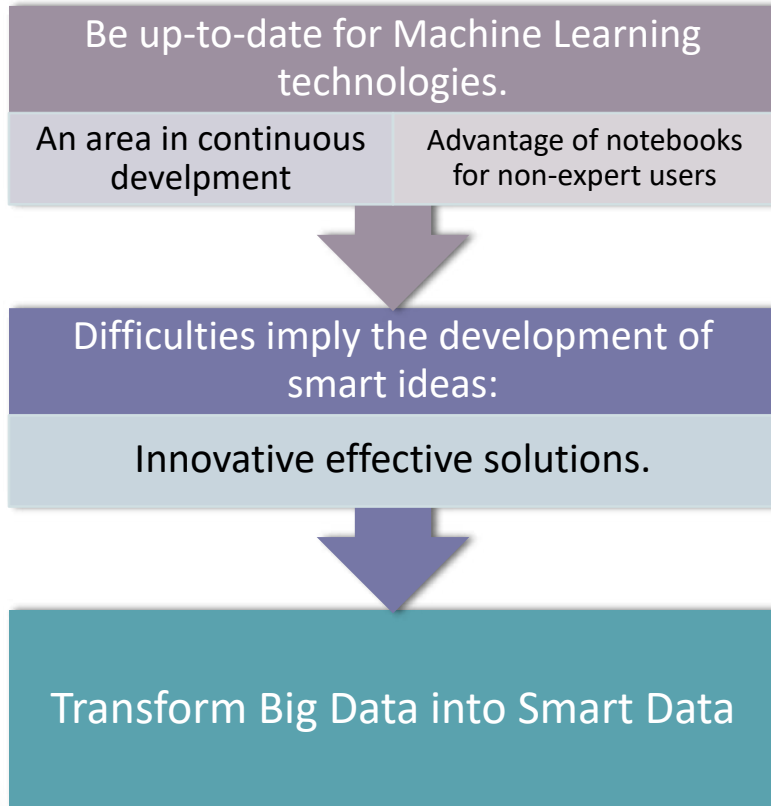
...and now what!?

# Concluding remarks



Big data is a new computational paradigm	<ul style="list-style-type: none"><li>• Big data is not HPC</li><li>• More than just large datasets</li></ul>
We need a new strategies to perform ML in those datasets	<ul style="list-style-type: none"><li>• Choosing the right technology is like choosing the right data structure in a program.</li></ul>
Frequently confused terms:	<ul style="list-style-type: none"><li>• Hadoop and Spark are frameworks, MapReduce and RDDs are programming paradigms.</li></ul>

# What should be done from here?



© 2020

## Big Data Preprocessing

Enabling Smart Data

Authors: Luengo, J., García-Gil, D., Ramírez-Gallego, S., García López, S., Herrera, F.

# THANKS!

# QUESTIONS?





# SOMACHINE

# BIG DATA



UNIVERSIDAD  
DE GRANADA



Instituto Andaluz Interuniversitario en  
Data Science and Computational Intelligence

## Big Data: Foundations and Frameworks

A. Fernández. Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence. **Universidad de Granada.**