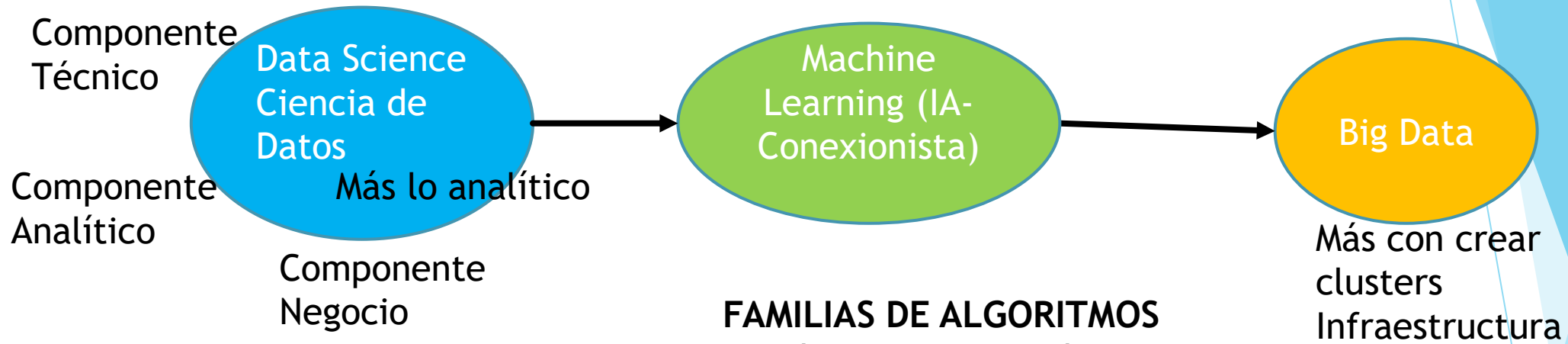


# Introducción a Big Data -Parte 2

Dra. Elvia Ruiz Beltrán  
Tecnológico Nacional de México  
Instituto Tecnológico de Aguascalientes



### ETAPAS

1. Recolectar datos(Data collection)
2. Almacenar datos(Data storage)
3. Limpieza de datos (Data cleaning)
4. Análisis de datos(Data Analysis)
5. Visualización (Visualization)
6. Toma de decisiones(Decision)

### FAMILIAS DE ALGORITMOS

1. Aprendizaje supervisado
2. Aprendizaje No-Supervisado
3. Algoritmos de Reforzamiento
4. Rede Neuronales→Deep Learning

Bases de datos  
relacionales  
Ficheros



Tradicional  
Data  
Science

R  
Python  
SAS  
IBM



Almacenamiento

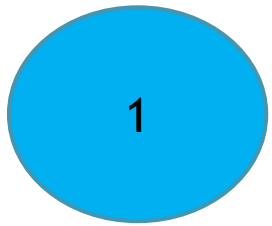
Procesamiento

HDFS  
NoSQL

Hadoop

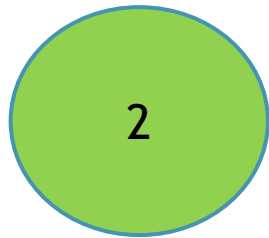
Hive  
Pig  
MapReduced  
Spark

Big Data



Data Science

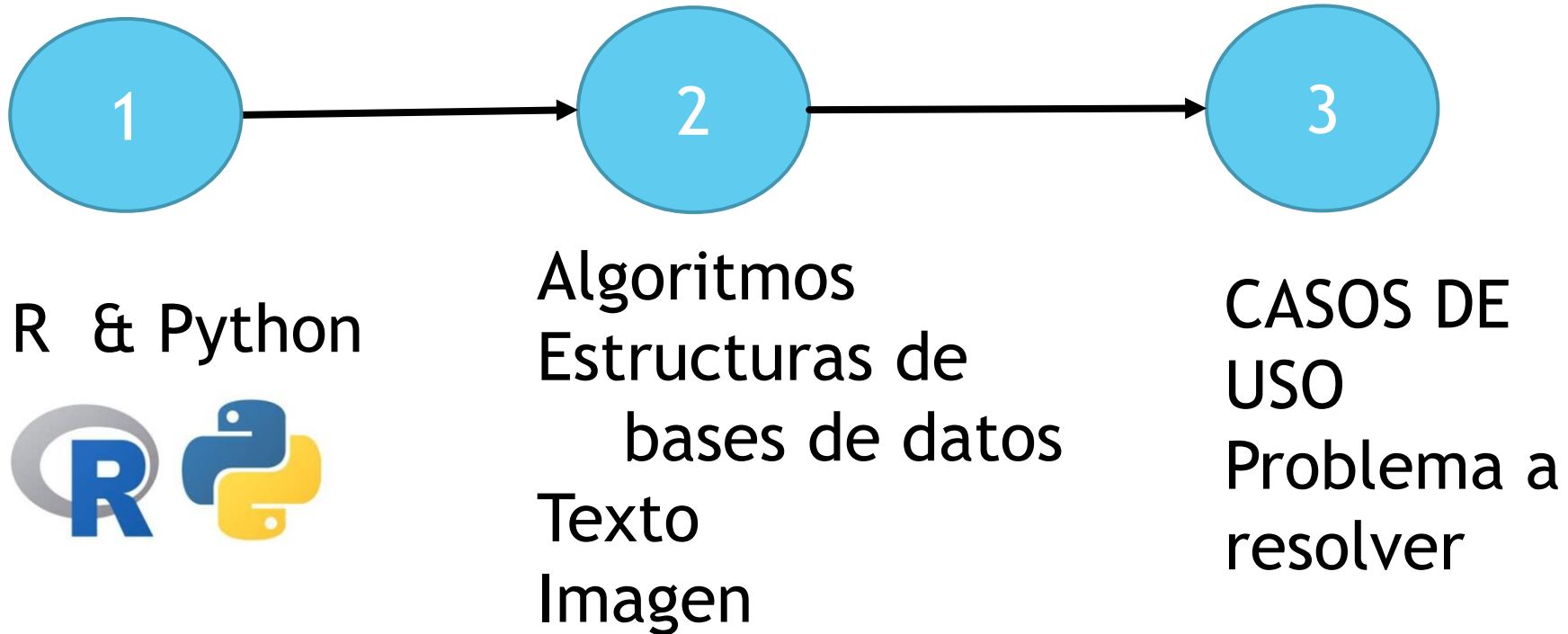
&



Business Intelligence

- **Proyectos de Discovery-** Analizar información para encontrar patrones->Conclusiones (Insights)  
Como se comportan los clientes, análisis de proveedores para ver cómo reducir costos con ellos, Análisis para descubrir nichos de mercados, etc.
- **Proyectos de Automation**
  - Crear producto de datos o algún sistema que se ejecute de manera más automática y nos mejore cierto proceso (Modelo implantado dentro de un sistema y mejore un proceso de negocio.)
  - Identificar qué acciones comerciales van a funcionar mejor con los clientes, los que muestran mayor interés.
  - Localizar transacciones fraudulenta en tiempo real y bloquear antes de que se produzcan.
  - Calcular el score de un cliente para conceder créditos.

# ¿Cómo empezar Data Science?



# Big Data + Analytics



## 6. Big Data Analytics

# Big Data

# Data mining (Minería de datos)

# Big data Analytics



ETL(Extract, transform and load)  
Extraer (Extract)  
Transformar (Transform)  
Cargar (Load)

Top ETL Tools 2022 | IT Business Edge

Mathematical operations can be applied (statical learning, Data Science, machine learning, Data Mining, etc).



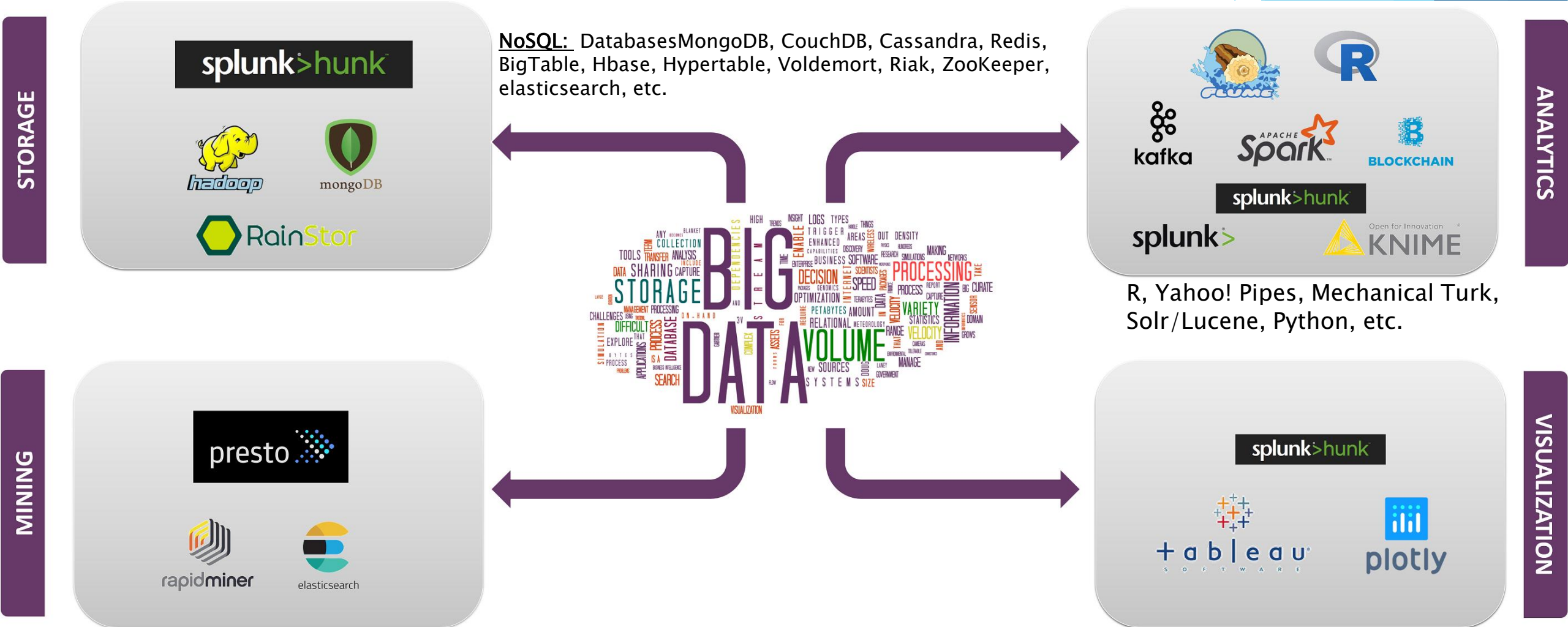
**Servers:** EC2(Amazon Elastic Compute Cloud ), Google App Engine, Elastic, Beanstalk, Heroku, etc.

Hadoop, Hive, Pig, Cascading  
Cascalog, mrjob, Caffeine,  
S4, MapR, Acunu, Flume,  
Kafka, Azkaban, Oozie,  
Greenplum, etc.

**Storage:** S3, Hadoop Distributed File System, Cloudera, Snow, etc.

**NoSQL:** Databases MongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper, elasticsearch, etc.

R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, Python, etc.







Google <https://www.youtube.com/watch?v=XZmGGAhHqa0>

<https://www.youtube.com/watch?v=k d33UVZhnAA>



<https://www.youtube.com/watch?v=rhQ5L86100Q>

Infotec



Facebook

<https://www.youtube.com/watch?v=frzVtaNrHU0>



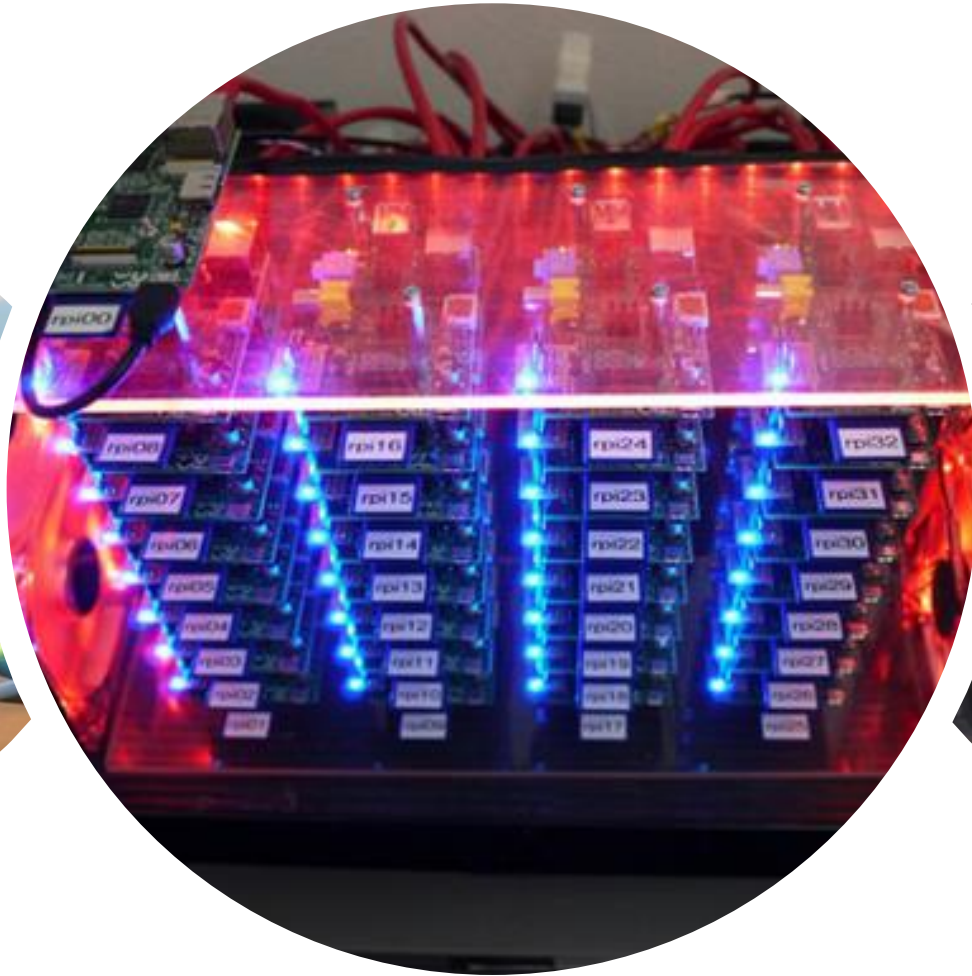
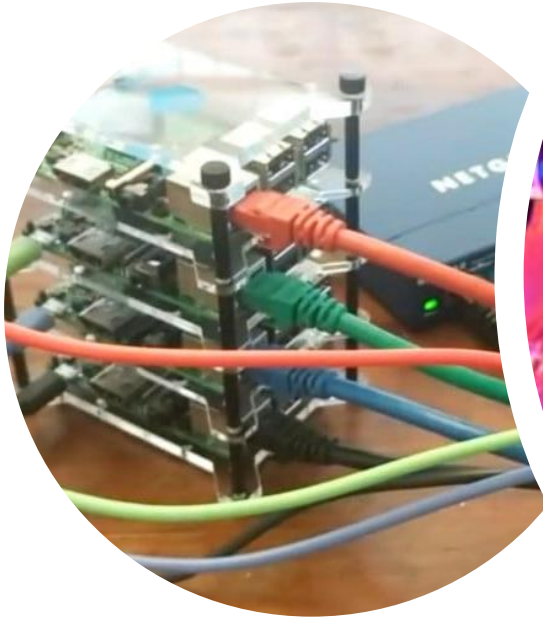


# Ubicación de los centros de datos de AWS

## AWS Regions









## Big Data Integration

All Components of the big data ecosystem from Hadoop to NoSQL Databases, each one of them have its own approach for extracting, transforming and loading data.

Leaflet



Google Cloud Platform



hadoop



amazon web services™ EC2



Studio®



Microsoft  
Excel 2013

IBM



rCharts



SQLite



HIVE



Hortonworks

Spark

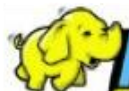


STATA

APACHE  
HBASE



STORM



hadoop  
MapReduce



cloudera®



mongoDB



cassandra

MySQL®

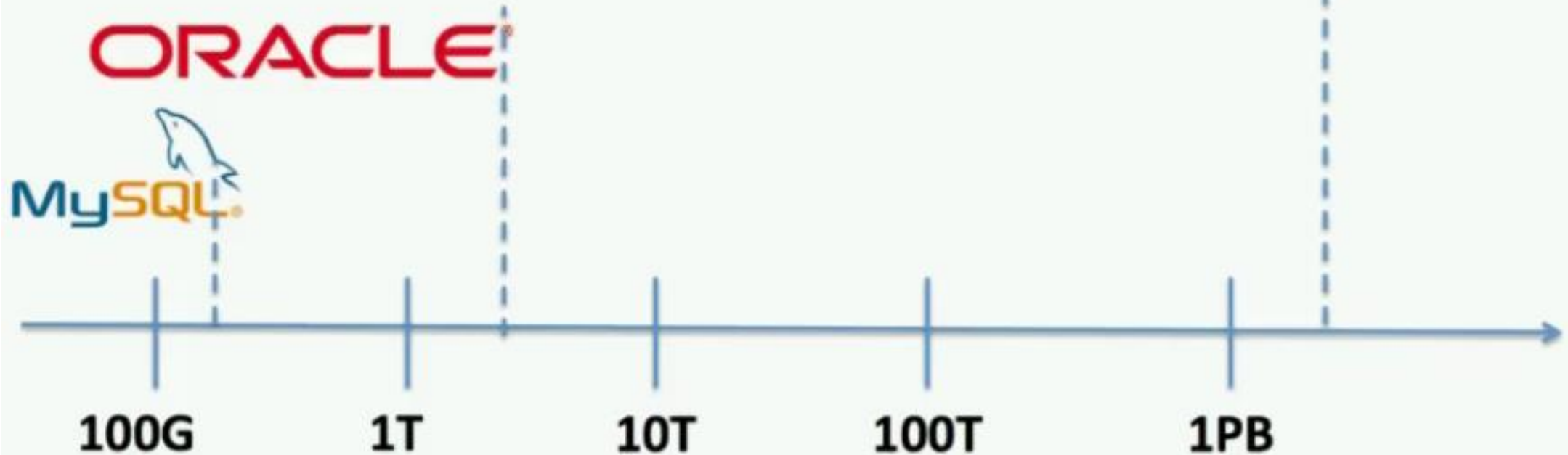


mahout



Hablaremos en especial de una de las tecnologías de Big Data mas populares, pero no la única, como se vio anteriormente.





Creado por el desarrollador Doug Cutting en 2006.

## 7. Plataforma de Código Abierto



- ▶ Apache Hadoop es un framework desarrollado en Java y de licencia libre y código abierto que permite el desarrollo de aplicaciones distribuidas con grandes cantidades de datos.

2006 Doug Cutting & Cafarella crea Hadoop para Nutch de Yahoo!

2007 Hadoop se mueve al proyecto de Apache.

2008 Hadoop gana la marca Terasort

<https://www.forbes.com/sites/oracle/2018/12/03/hadoop-pioneer-says-developers-should-build-open-source-into-their-career-plans/#296fb78379bc>





## 7. Plataforma de Código Abierto

► Inspirada en:

Google Map-Reduce

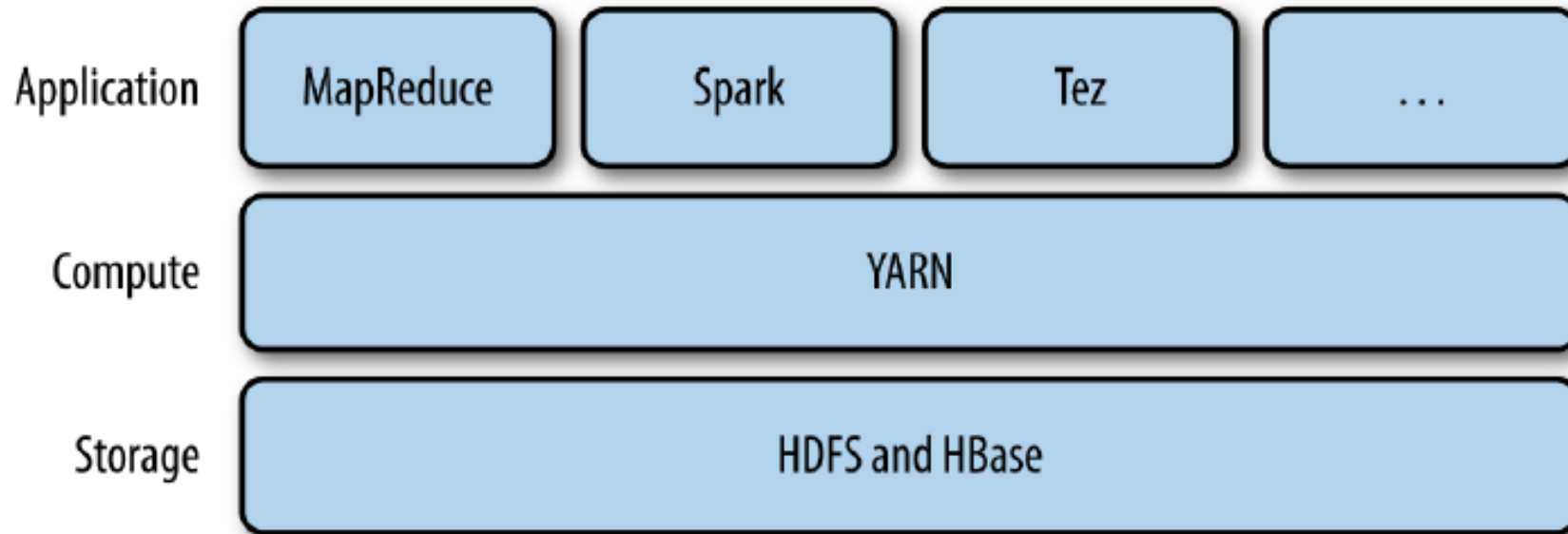
GFS(Google File System)



<https://httpd.apache.org/>

### Principales características

- ❑ HDFS(Hadoop Distributed File System)
- ❑ Hadoop MapReduce



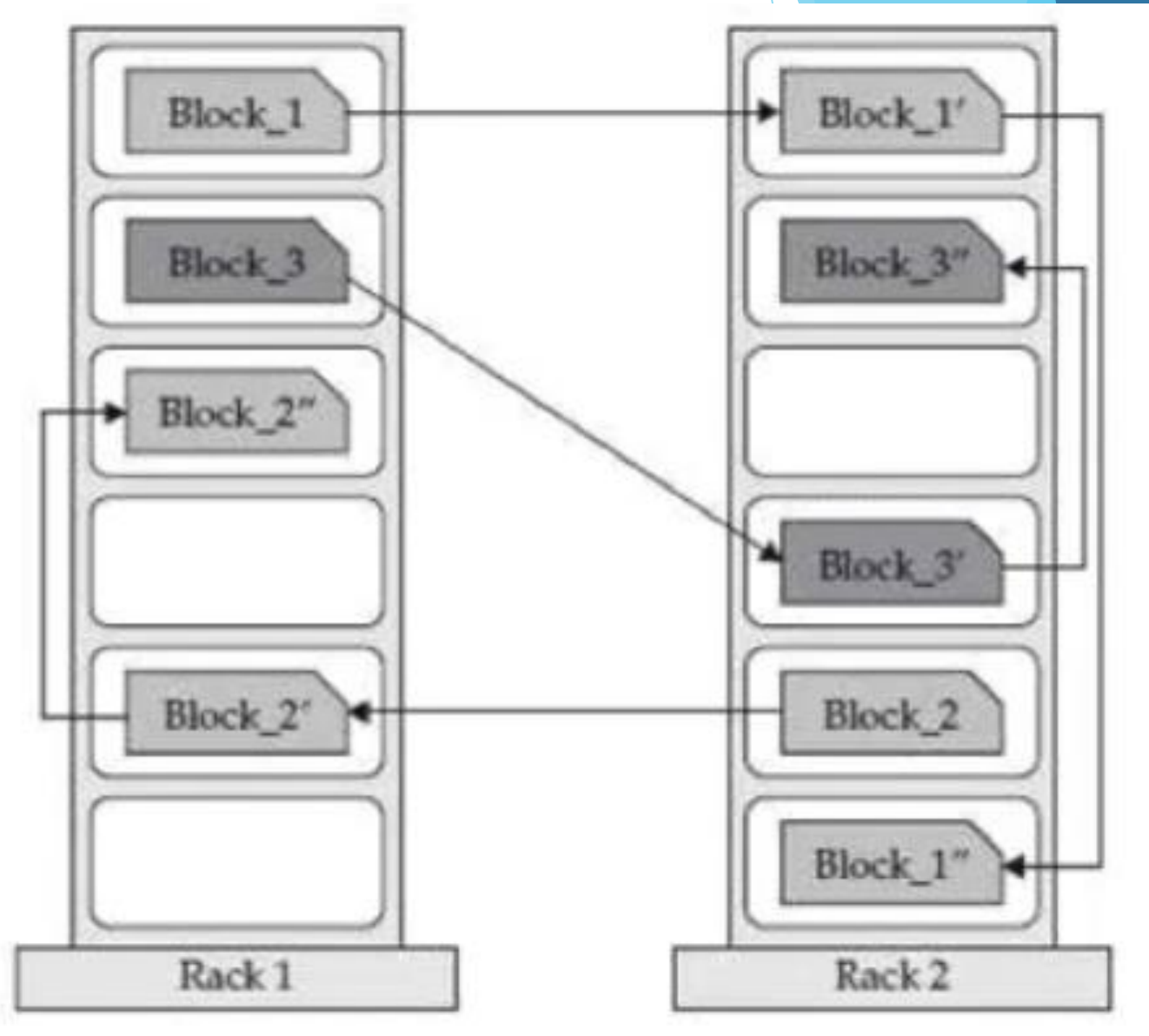


# Hadoop Distributed File System (HDFS)



## Almacenamiento

- ▶ En la figura se ejemplifica como los bloques de datos son escritos hacia HDFS. Observe que cada bloque es almacenado tres veces y al menos un bloque se almacena en un rack diferente para lograr redundancia.





# Componentes de HDFS



- ▶ **NameNode:** es el maestro (master) un sistema HDFS. Mantiene los directorios, archivos y manejo de bloques(blocks) que están presentes sobre los DataNodes.
- ▶ **DataNode:** son nodos esclavos(slaves) que son movidos por cada máquina y provee el almacenamiento real. Son responsables de atender las solicitudes de datos de lectura y escritura para los clientes.
- ▶ **Secondary NameNode:** este es responsable de realizar puntos de control periódicos. Por lo tanto, si el NameNode falla en cualquier momento, puede reemplazarse con una imagen instantánea almacenada por los puntos de control Secondary NameNode.

# Hadoop MapReduce

## Procesamiento

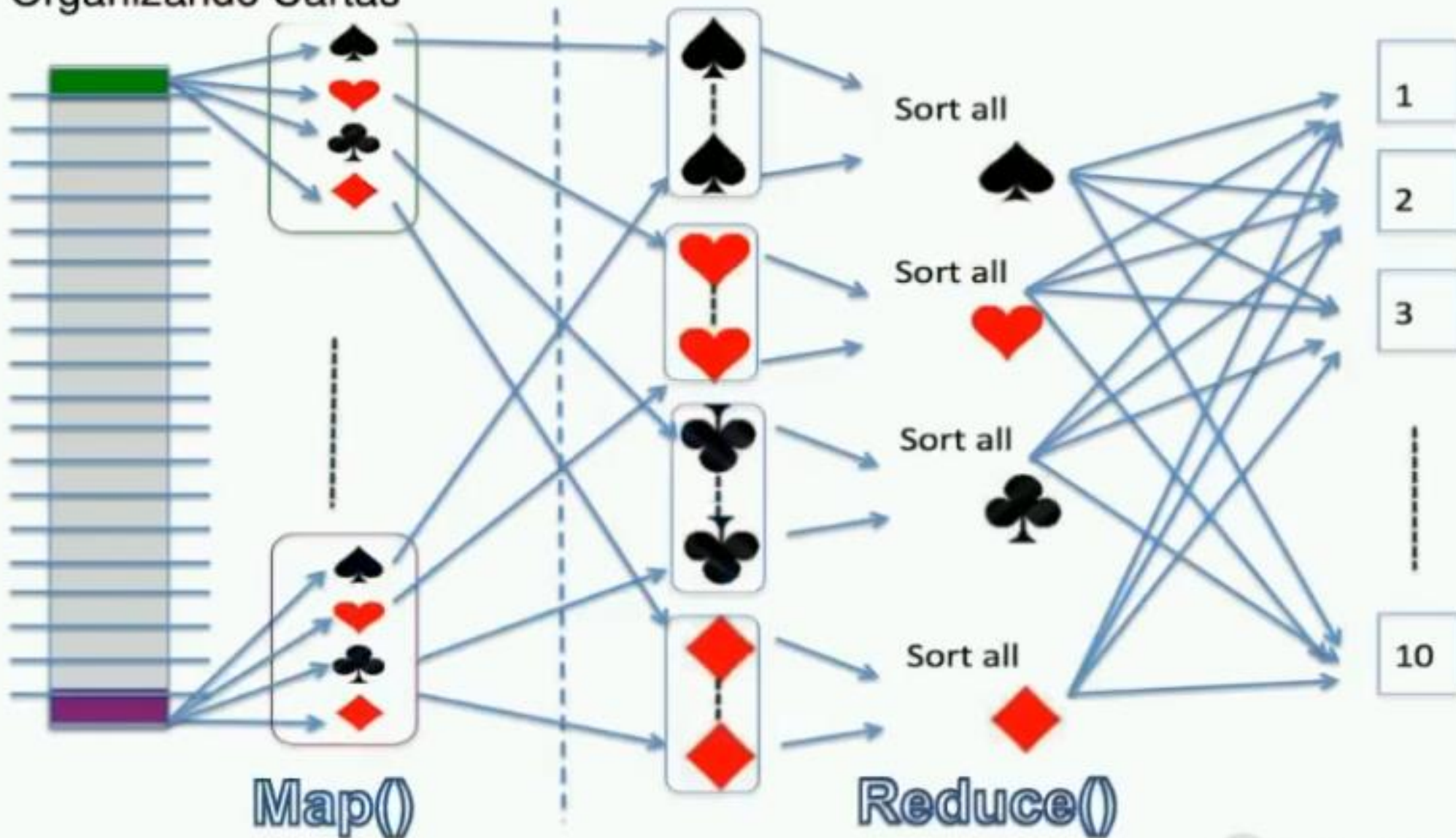
- ▶ **MapReduce** es el núcleo de Hadoop. El término MapReduce en realidad se refiere a dos procesos separados que Hadoop ejecuta.
- ▶ El primer proceso map, el cual toma un conjunto de datos(Data Block) y los convierte en otro conjunto, donde los elementos individuales son separados en tuplas (pares de llaves/valor).
- ▶ El proceso reduce obtiene la salida del map como datos de entrada y combina las tuplas en un conjunto más pequeño de las mismas.
- ▶ Una frase intermedia es la denominada **Shuffle(bajar o barajar)** la cual obtiene las tuplas del proceso map y determina que nodo procesará estos datos dirigiendo la salida a una tarea reduce en específico.

Example 1:

# MapReduce



Organizando Cartas



## Apache Spark

[spark.apache.org](http://spark.apache.org)



<b>Tipo de programa</b>	framework machine learning Framework computación en la nube software libre
<b>Desarrollador</b>	Apache Software Foundation AMPLab
<b>Lanzamiento</b>	30 de mayo de 2014
<b>Género</b>	Data analytics, machine learning algorithms
<b>Programado en</b>	Scala, Java, Python, R
<b>Sistema operativo</b>	Microsoft Windows, macOS, Linux
<b>Plataforma</b>	Java
<b>Licencia</b>	Apache License 2.0
<b>Estado actual</b>	Activo
<b>Idiomas</b>	inglés
<b>En español</b>	No

[\[editar datos en Wikidata\]](#)

# ¿Que más hay?



CLUDERA

Why Cloudera

Products

Solutions

Services & Support

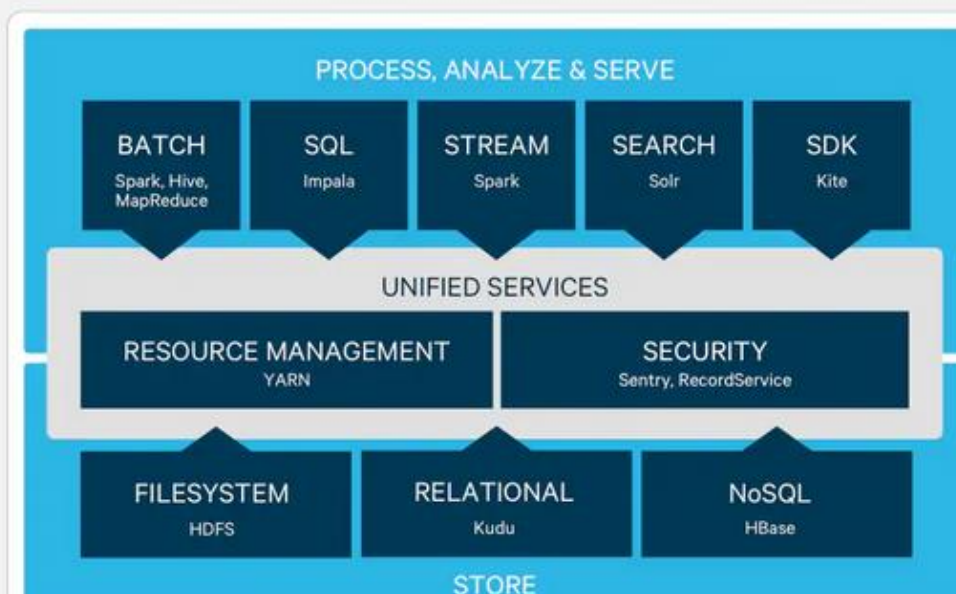


# Apache Hadoop Ecosystem

Hadoop is an ecosystem of open source components that fundamentally changes the way enterprises store, process, and analyze data. Unlike traditional systems, Hadoop enables multiple types of analytic workloads to run on the same data, at the same time, at massive scale on industry-standard hardware. CDH, Cloudera's open source platform, is the most popular distribution of Hadoop and related projects in the world (with support available via a Cloudera Enterprise subscription).

[Try now >](#)

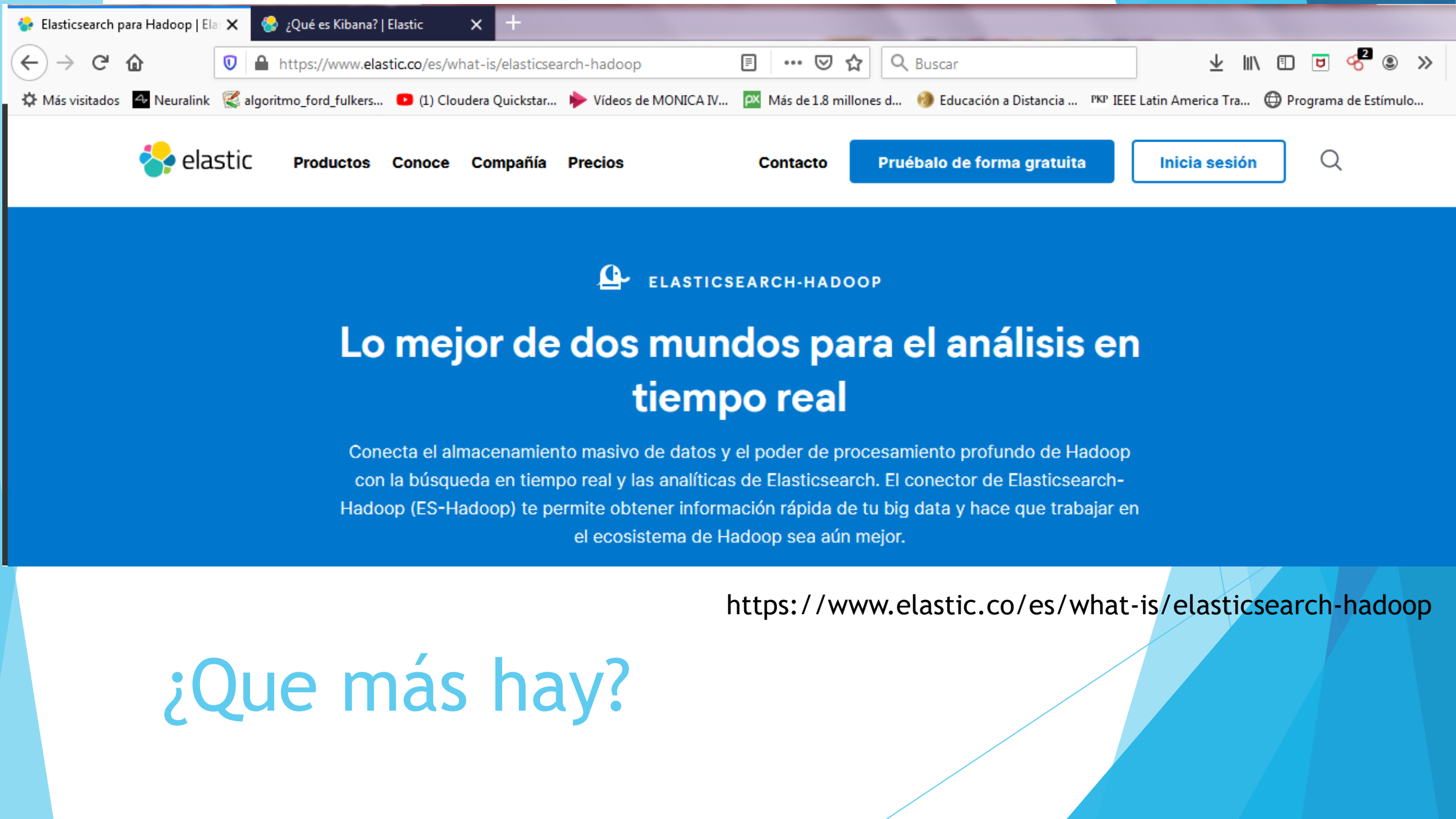
[Hadoop ecosystem in the Engineering Blog >](#)



<https://en.wikipedia.org/wiki/Cloudera>

<https://www.cloudera.com/products/open-source/apache-hadoop.html>

<https://www.cloudera.com/>



## ELASTICSEARCH-HADOOP

# Lo mejor de dos mundos para el análisis en tiempo real

Conecta el almacenamiento masivo de datos y el poder de procesamiento profundo de Hadoop con la búsqueda en tiempo real y las analíticas de Elasticsearch. El conector de Elasticsearch-Hadoop (ES-Hadoop) te permite obtener información rápida de tu big data y hace que trabajar en el ecosistema de Hadoop sea aún mejor.

<https://www.elastic.co/es/what-is/elasticsearch-hadoop>

¿Que más hay?



Archivo

Editar

Ver

Historial

Marcadores

Herramientas

Ayuda

Create your account — Elastic X ¿Qué es Kibana? | Elastic X +

←

→

↺

🏠

🔒

https://cloud.elastic.co/registration?elektra=es-elastic-stack-page

...

🛡️

☆

🔍

Buscar

⬇️

📄

📅

🔗


👤

2

⌵

☰

Más visitados Neuralink algoritmo\_ford\_fulkers... (1) Cloudera Quickstar... Videos de MONICA IV... Más de 1.8 millones d... Educación a Distancia ... PKP IEEE Latin America Tra... Programa de Estímulo... >>





Elastic Cloud

Create your account

Already have an account? [Log in](#)

Email

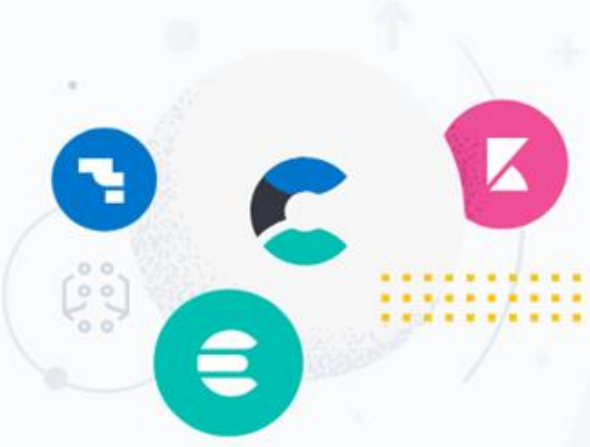
Password 



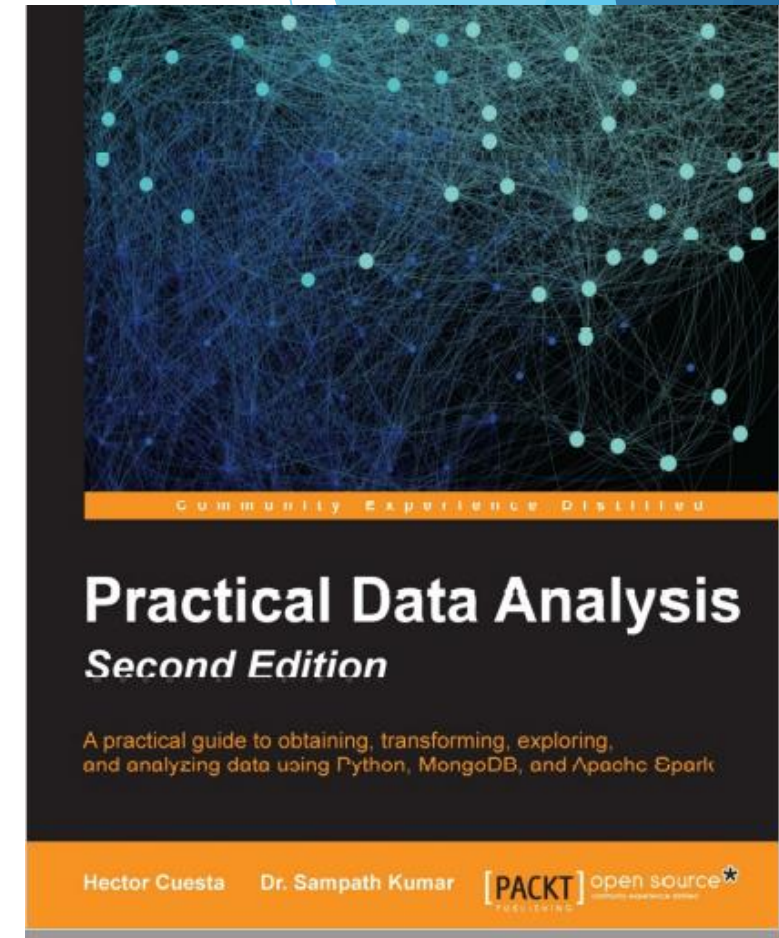
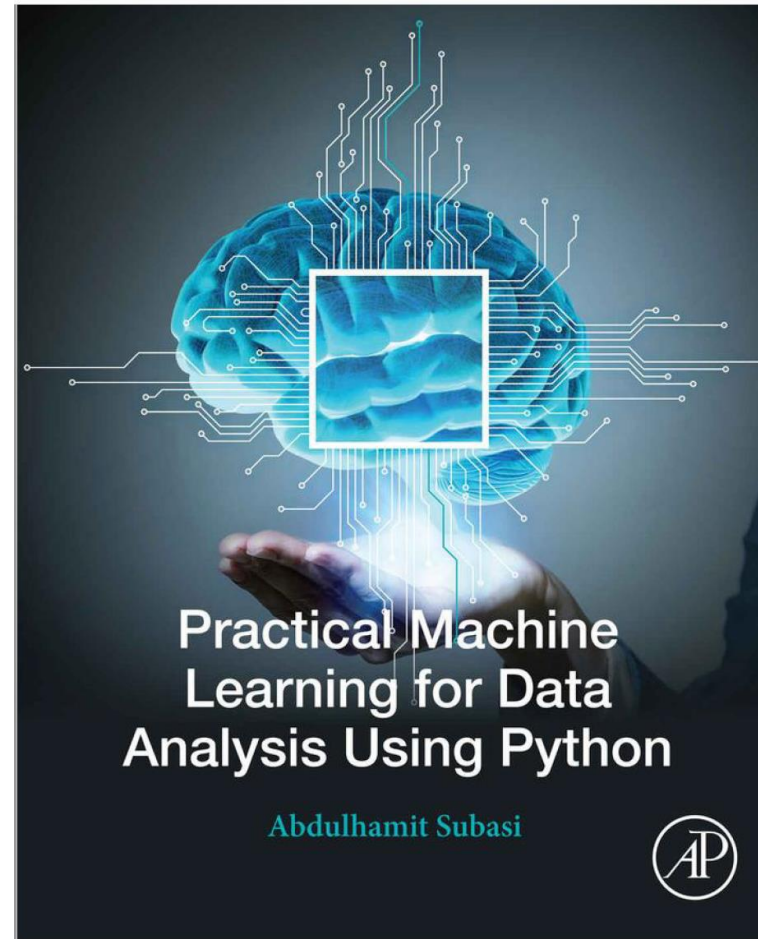
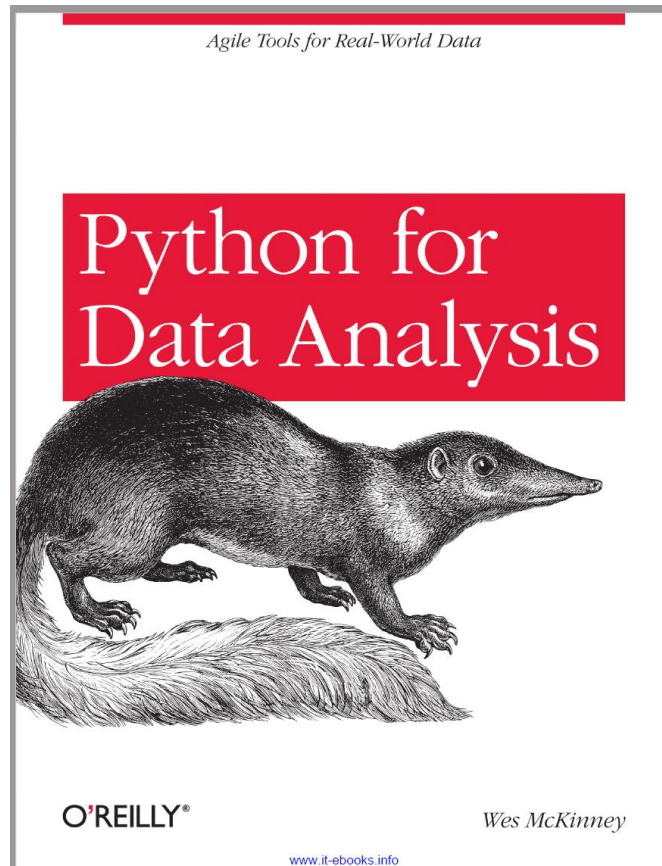
Password strength meter

Create account

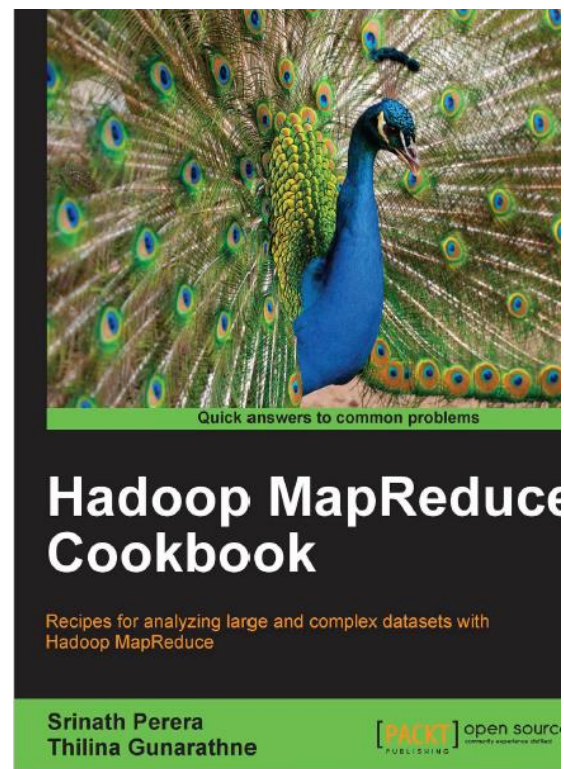
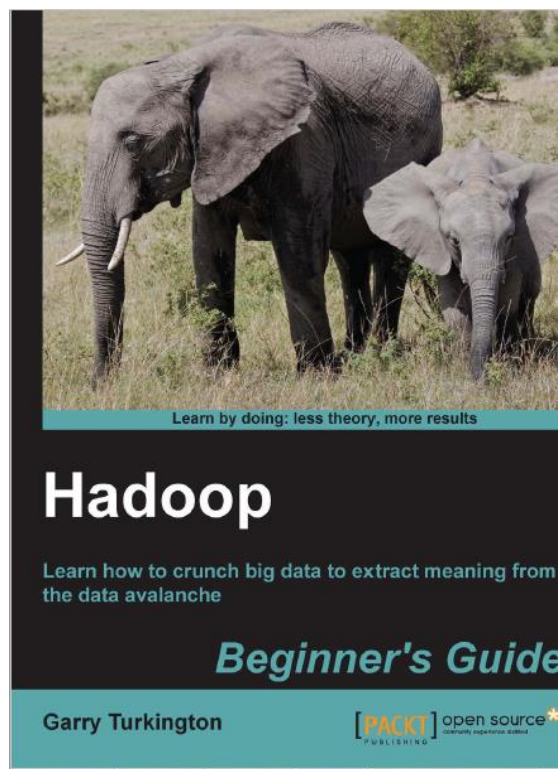
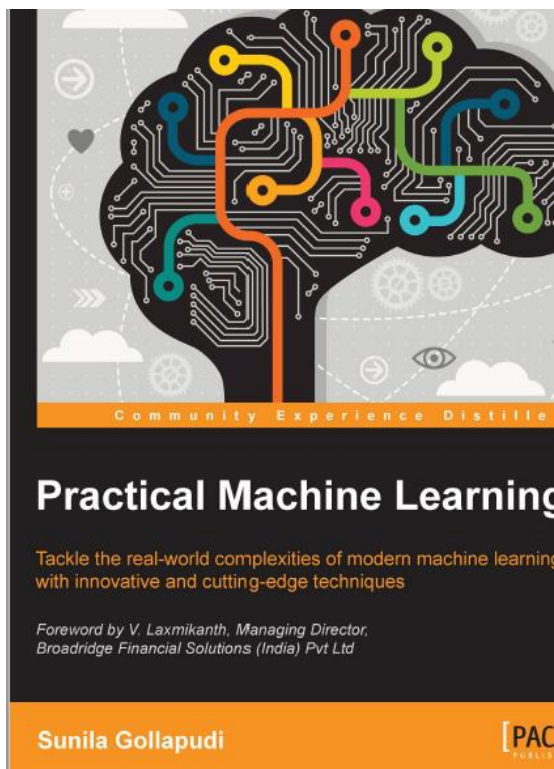
Or



# Algunas referencias

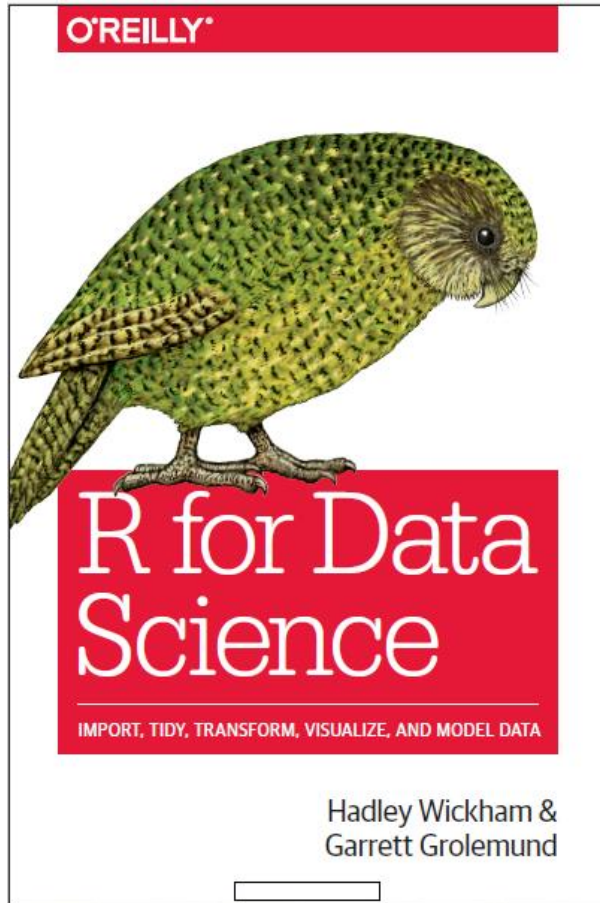






# Algunas referencias

# Algunas referencias





# Algunas referencias

