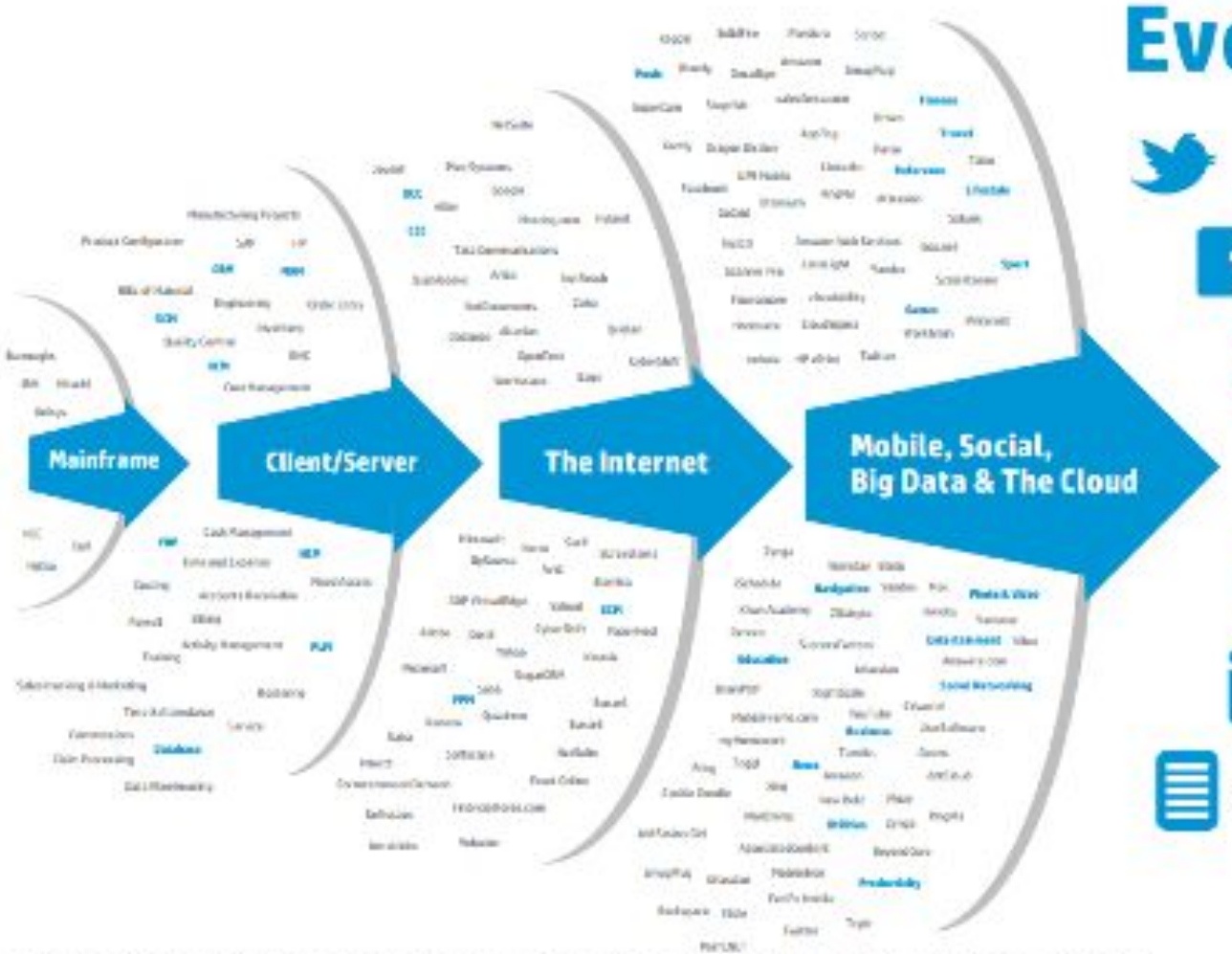
Inteligencia Artificial Avanzada

Big Data

Docente: Junior Fabian Arteaga
(capacitacion@dataminingperu.com)


WHAT WE KNOW...
THE REST...
BIG DATA

2013



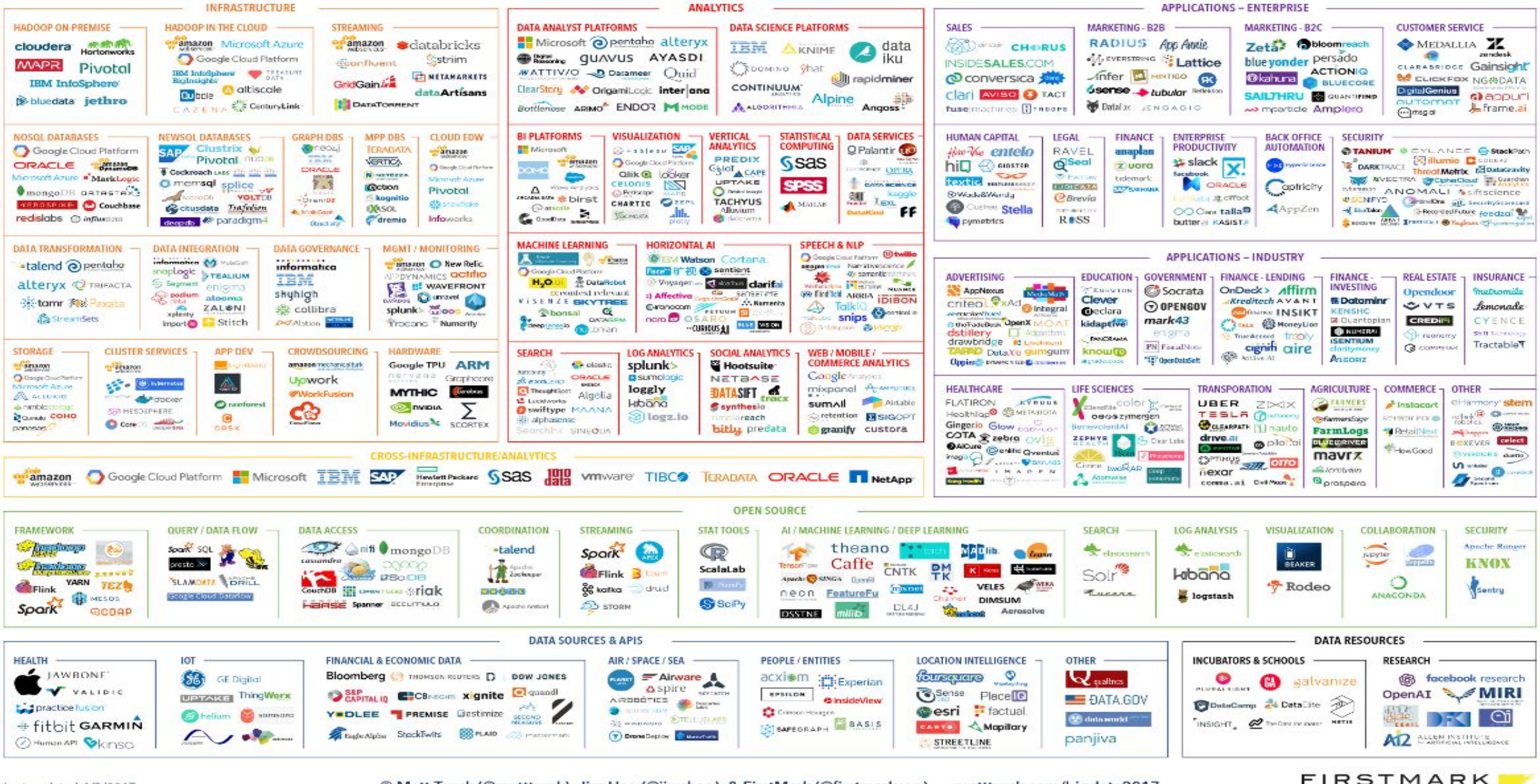
Every 60 seconds

- **98,000+** tweets
- **695,000** status updates
- **11million** instant messages
- **698,445** Google searches
- **168 million+** emails sent
- **1,820TB** of data created
- **217** new mobile web users

→ Facebook:

- (2011) 650 000 posts
- (2016) 3 millones posts
- 510 000 comentarios
- 136 000 fotos subidas
- 293 000 estados
- 4 millones de likes

BIG DATA LANDSCAPE 2017



Last updated 4/5/2017

mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Big Data

"Big Data" hace referencia al conjunto de información que es demasiado compleja como para ser procesada mediante TI tradicionales de manera aceptable

Min Chen, Shiwen Mao, and Yunhao Liu. Big Data: A Survey. Mobile Networks and Applications,19(2):171–209, 2014

Big Data

Actualmente, se generan muchos datos y las empresas exigen saber no sólo lo que sucede en la actualidad, sino también lo que va a pasar en un futuro, requiriendo niveles de servicio mucho más exigentes que hace unos años. Es precisamente en esta coyuntura donde se encuentra Big Data hoy día.

Big Data

Para obtener una definición verdaderamente acertada de lo que significa Big Data debemos abrir la mente y ***romper con el estereotipo de que en "Big" esta la clave***, ya que las exigencias actuales no siempre están basadas en el volumen, sino que éste es sólo una parte del rompecabezas, siendo muy diversos los parámetros que se tienen en cuenta en cada ocasión.

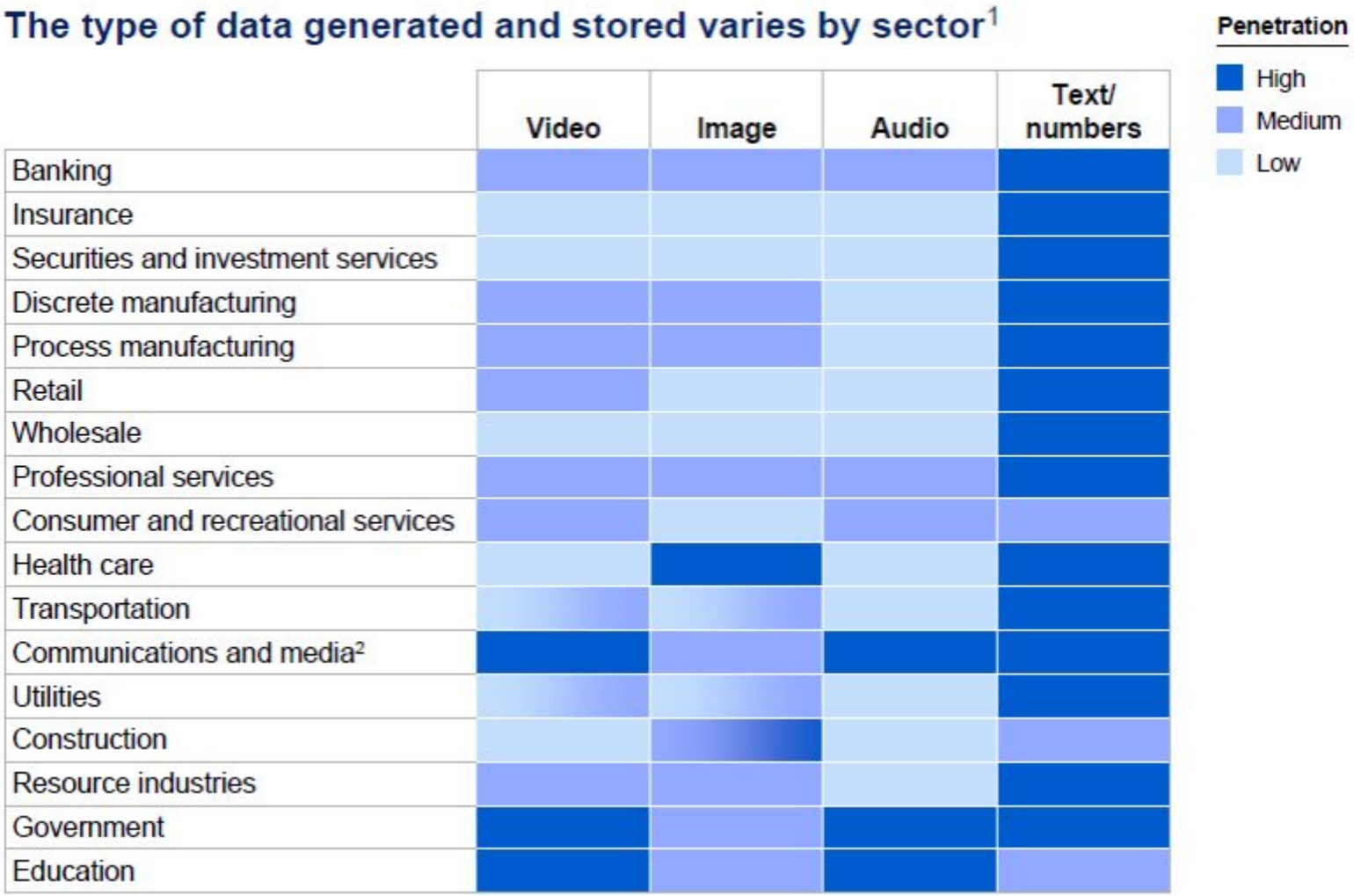
Importancia del Big Data

Impacto tanto en la industria, como en el negocio e incluso en nuestra sociedad y además ofrece una ventaja competitiva considerable.

En efecto, es precisamente en ese tipo de datos donde **las empresas han detectado que se encierra mayor valor**. Hoy en día, para muchas empresas puede llegar a ser más importante detectar al cliente que más influye al resto de posibles compradores, que al que mayor volumen de compra realiza.

En la actualidad, l**a cantidad de datos que se generan es abismal y de una casuística extremadamente compleja para su análisis**. Las empresas cada vez exigen que el análisis sea lo más cercano posible al tiempo real. Y en Big Data está la clave, al traducirse el mismo en las variables de velocidad, variedad y volumen que requiere el mercado actualmente.

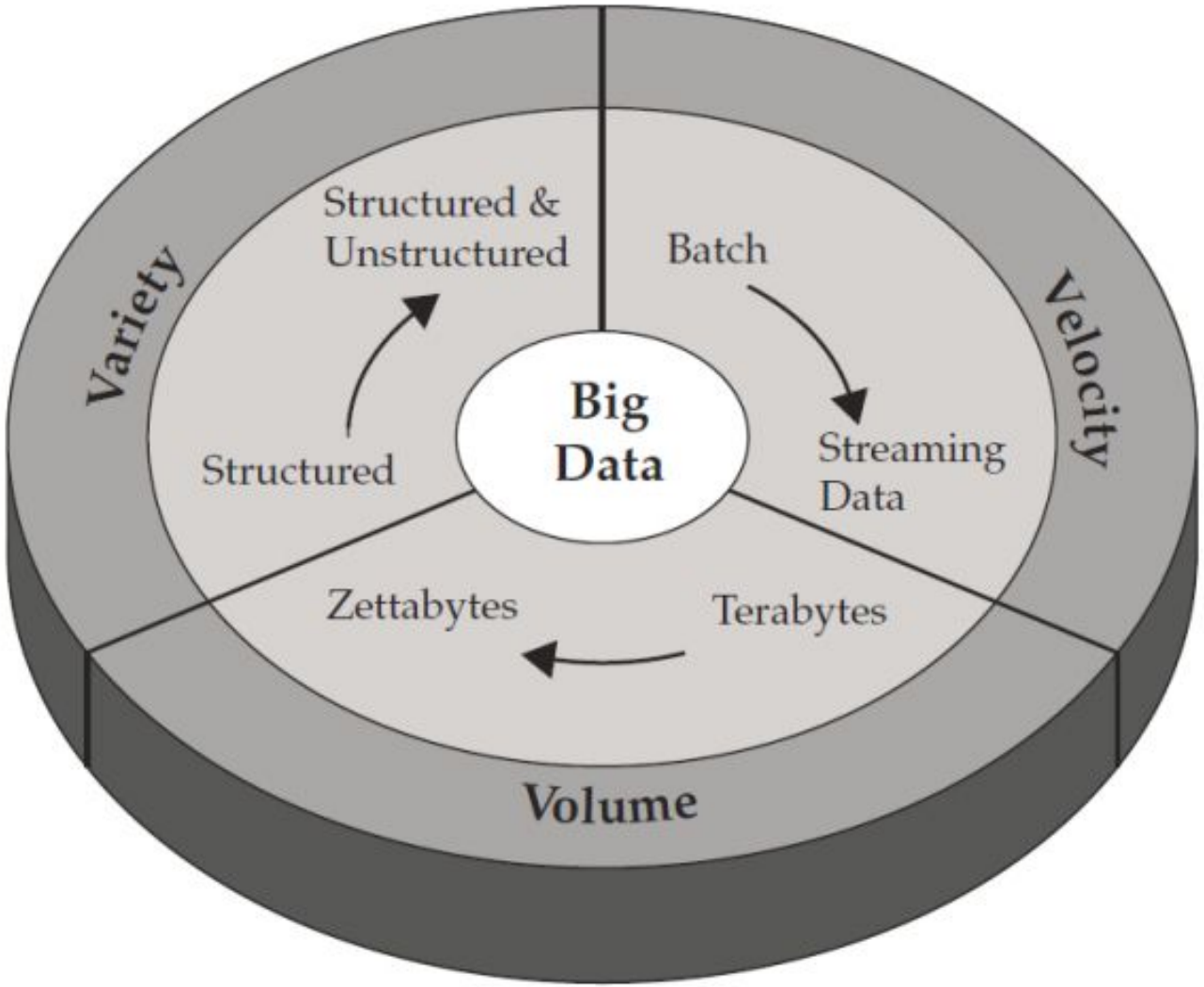
Tipos de datos disponibles en Big Data

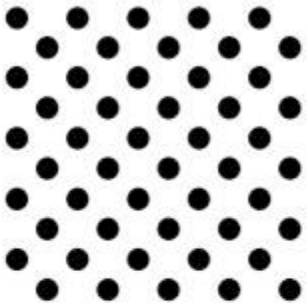
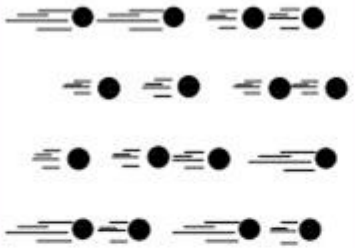
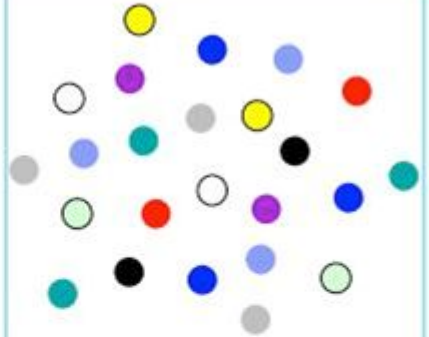
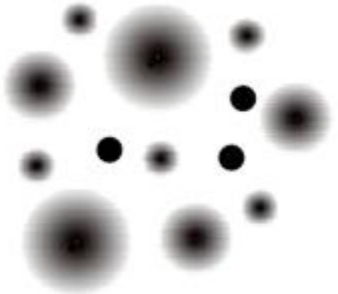


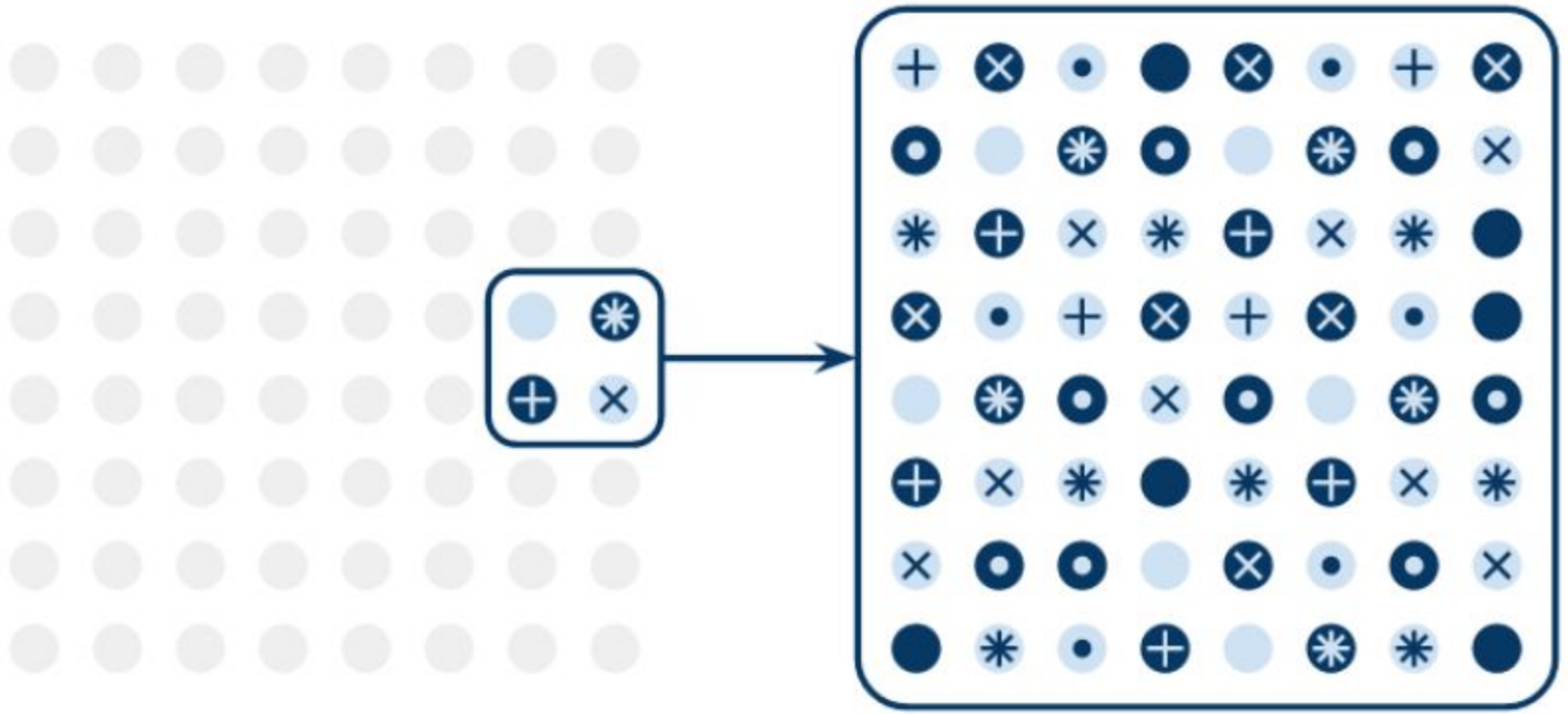
	Video	Image	Audio	Text/ numbers
Banking	Medium	Medium	Medium	High
Insurance	Low	Low	Low	High
Securities and investment services	Low	Low	Low	High
Discrete manufacturing	Medium	Medium	Low	High
Process manufacturing	Medium	Medium	Low	High
Retail	Medium	Low	Low	High
Wholesale	Low	Low	Low	High
Professional services	Medium	Medium	Medium	High
Consumer and recreational services	Medium	Low	Medium	Medium
Health care	Low	High	Low	High
Transportation	Low	Low	Low	High
Communications and media[2]	High	Medium	High	High
Utilities	Low	Low	Low	High
Construction	Low	Medium	Low	Medium
Resource industries	Medium	Medium	Low	High
Government	High	Medium	High	High
Education	High	Medium	High	Medium

Dimensiones y paradigmas...

Volume
velocity
variety

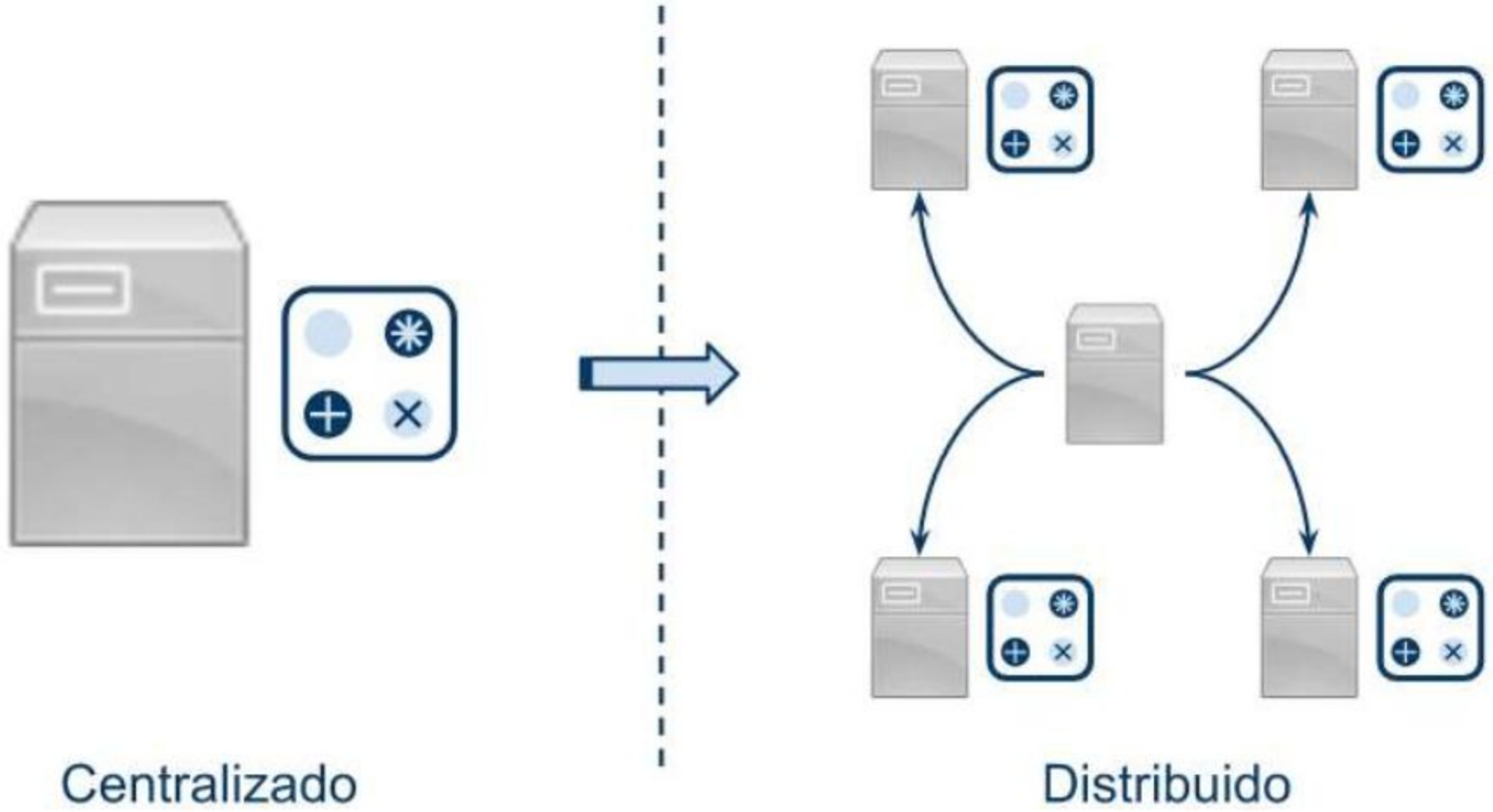

Variety
Structured &
Unstructured
Batch
Velocity
Big
Data
Structured
Streaming
Data
Zettabytes
Terabytes
Volume

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

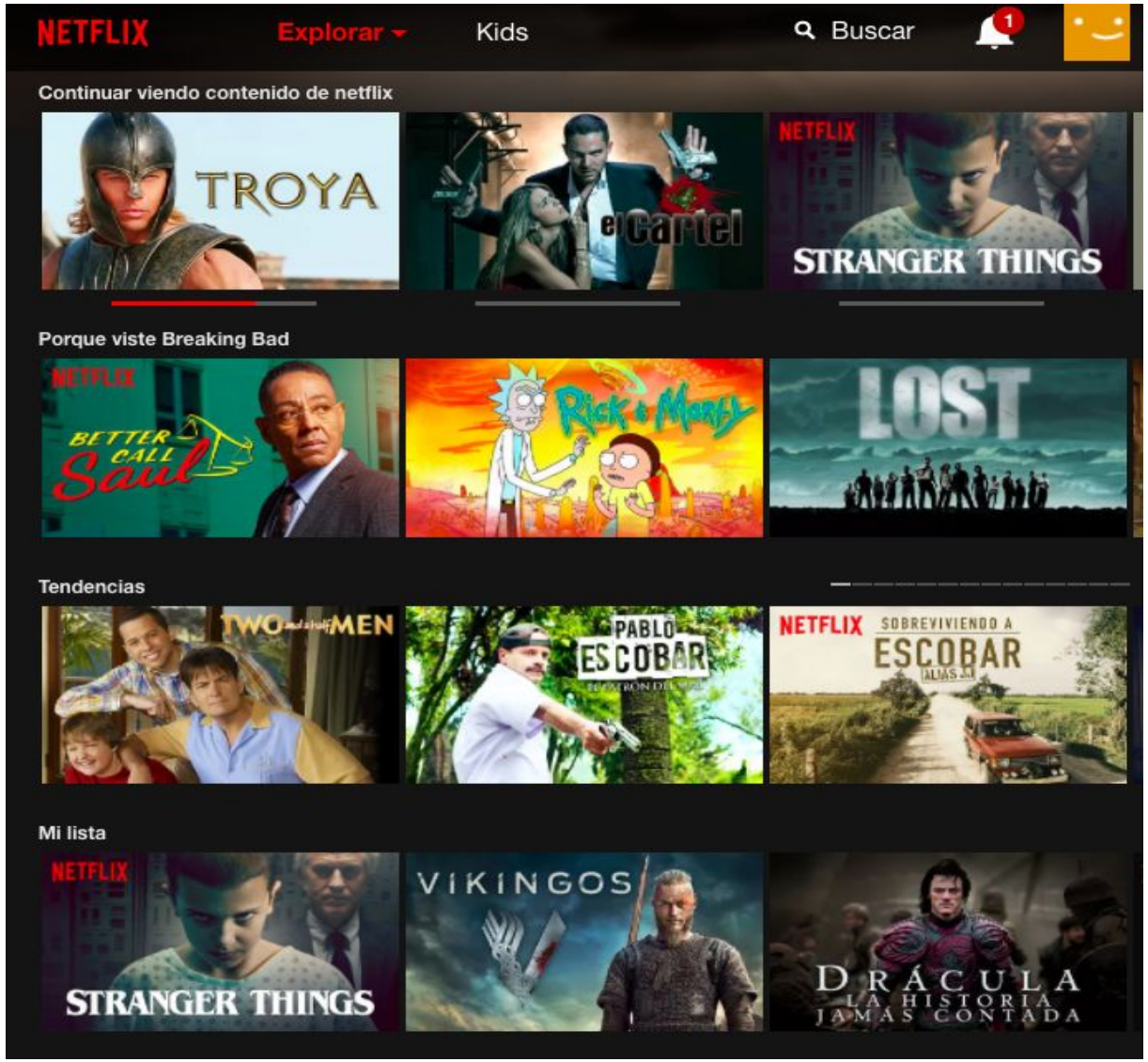


Análisis de un subconjunto de datos

Análisis de todos los datos


Centralizado
Distribuido

Aplicaciones



amazon.com

Recommended for You

Amazon.com has new recommendations for you based on items you purchased or told us you own.

Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop

Google Apps Administrator Guide: A Private-Label Web Workspace

Googlepedia: The Ultimate Google Resource (3rd Edition)

- Domain
- Sub-domain
- Page URL
- URL sub-directories

- Page Meta Tags
- Page Title
- Page Content
- Named Entities

- Has Query
- Referrer Query

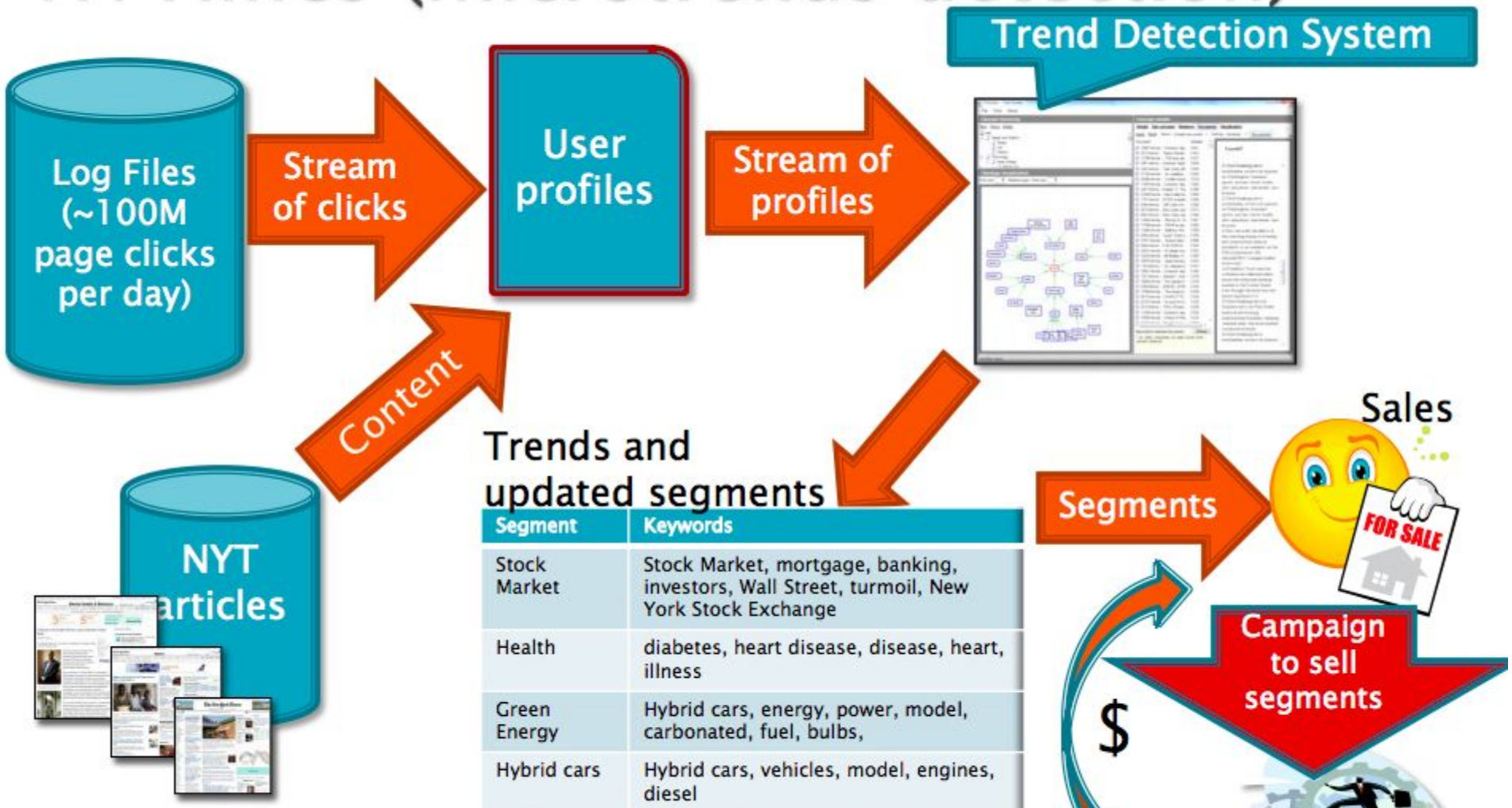
- Referring Domain
- Referring URL
- Outgoing URL

- GeoIP Country
- GeoIP State
- GeoIP City

- Absolute Date
- Day of the Week
- Day period
- Hour of the day
- User Agent

- Zip Code
- State
- Income
- Age
- Gender
- Country
- Job Title
- Job Industry

NYTimes (microtrends detection)

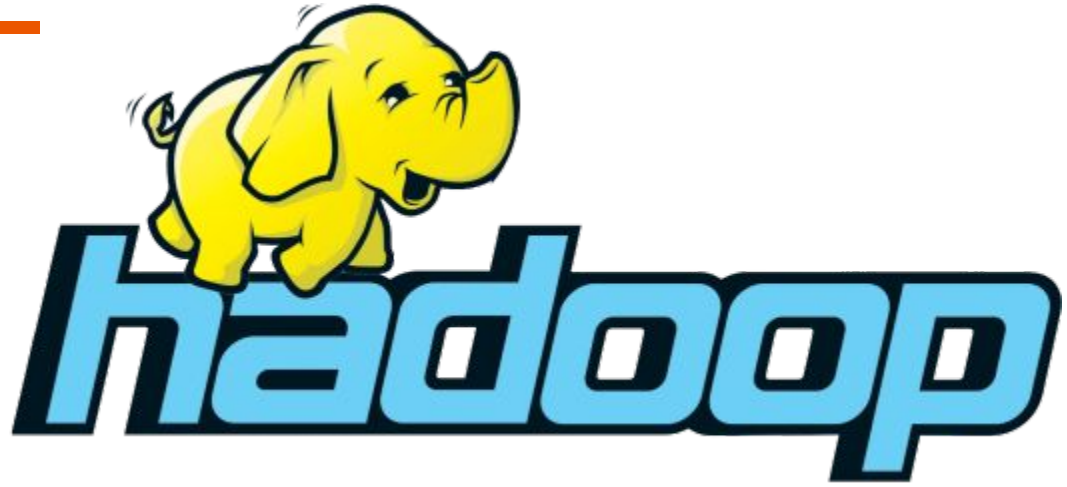


Retail

Retail

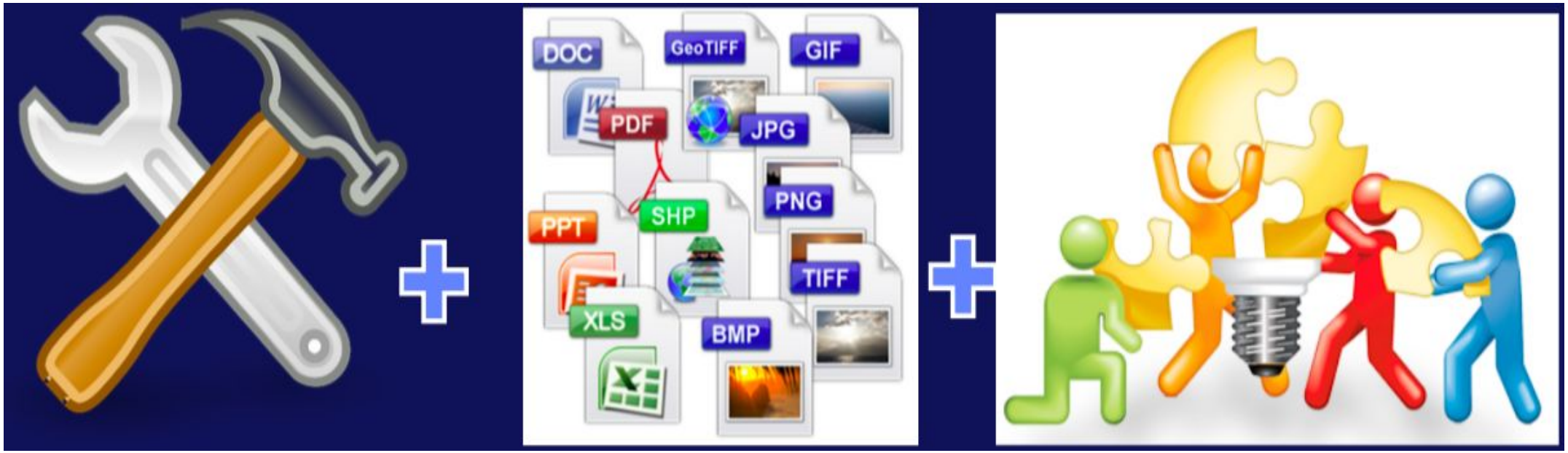

Capturing Store Insights for Timely Engagement
Gesture Recognition
Wi-Fi
Chokepoint / Precision Location
Digital Signage
Integrated Video Surveillance
Associate Tablet
Internet of Things
Customer Mobile
Continuous Channel Customer Interaction
Wisdom from the Cloud
Fog
Secure Big Data Infrastructure, Analysis, & Action
Big Data Repository
Fast Data Processing
Action Enablement

HERRAMIENTAS



$$V = \frac{\Delta x}{\Delta t}$$

INTEGRACIÓN

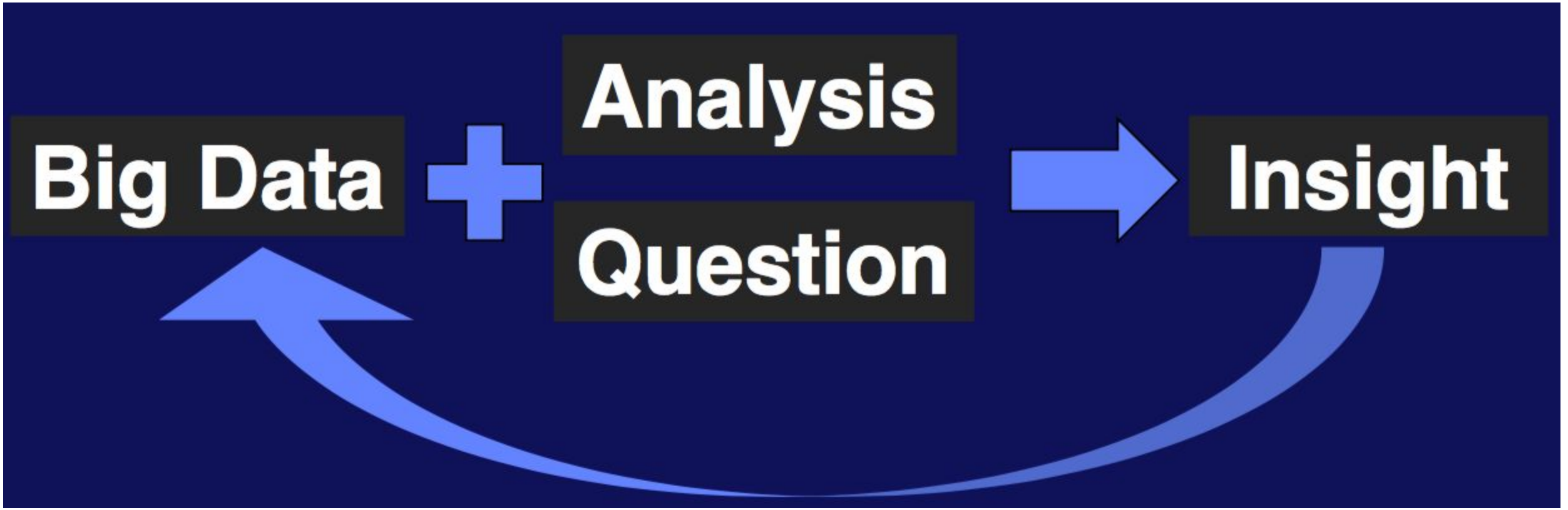


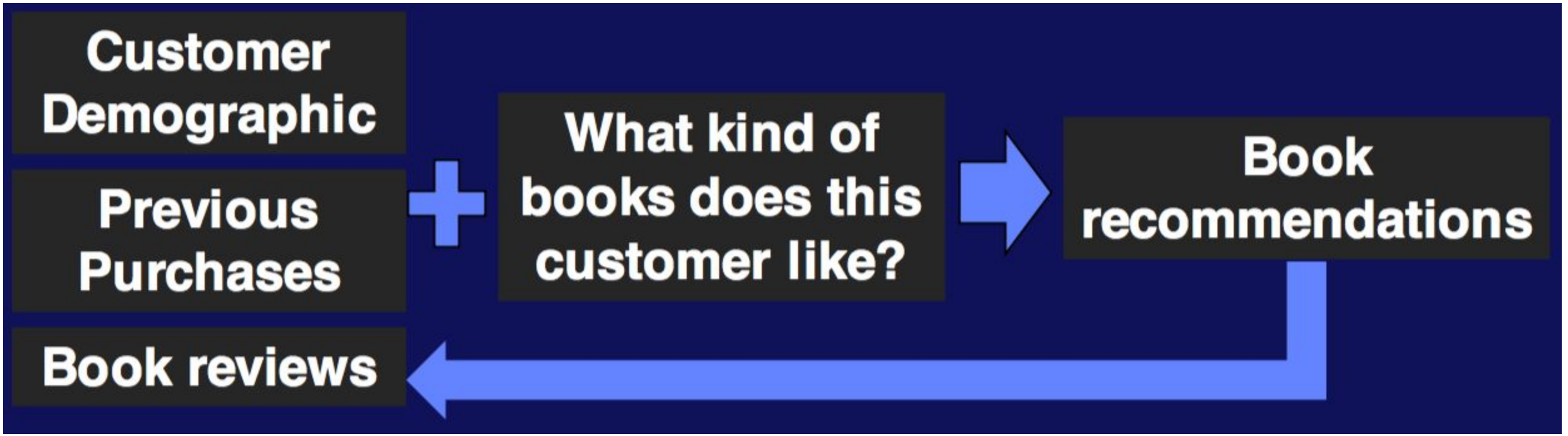
Herramientas + Data + Data Scientist

VALOR

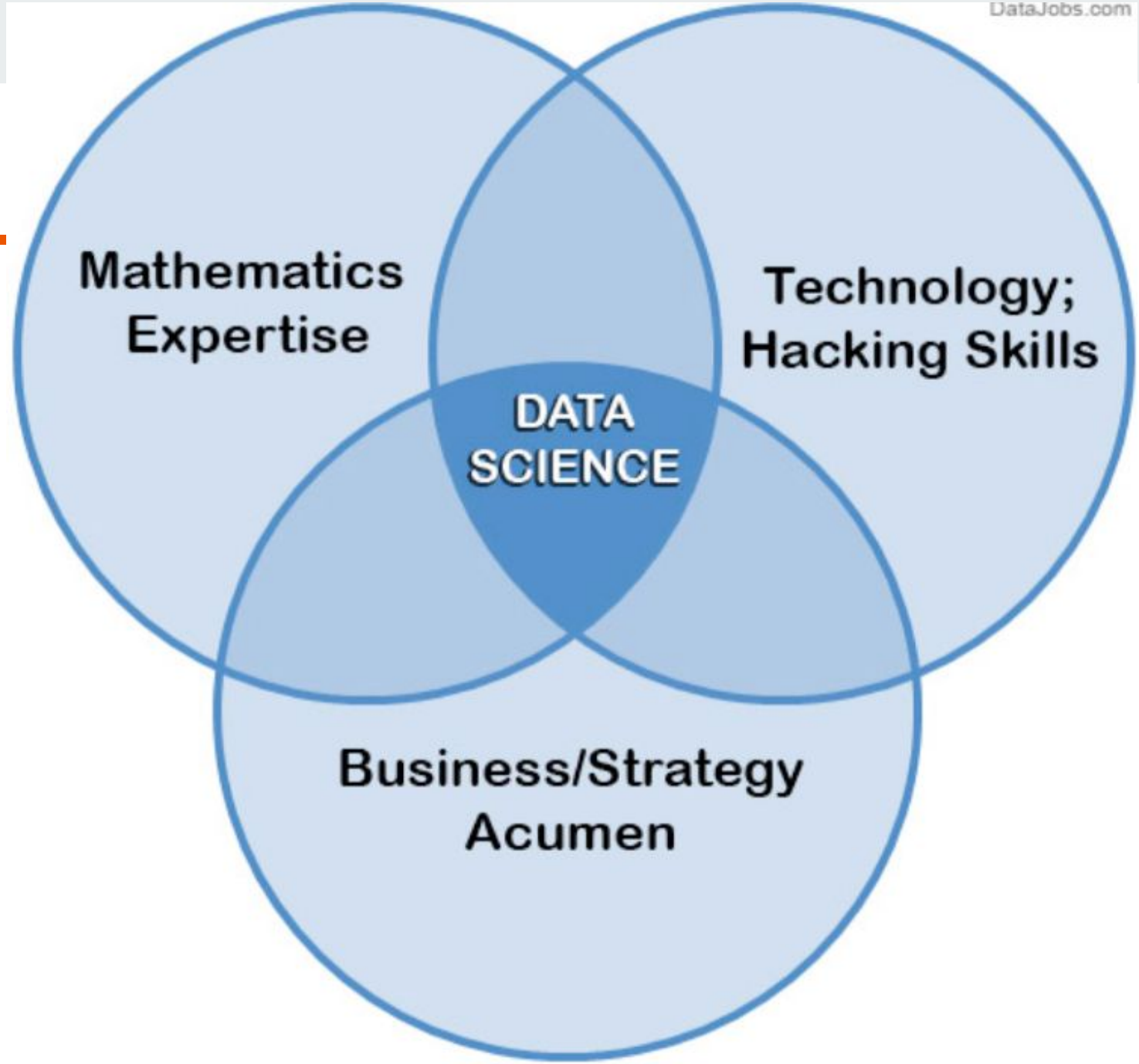
DATA SCIENCE

Data Science: Base del Big Data


Big Data
Analysis
Question
Insight


Customer Demographic
Previous Purchases
Book reviews
What kind of books does this customer like?
Book recommendations

amazon


DataJobs.com
Mathematics Expertise
Technology; Hacking Skills
DATA SCIENCE
Business/Strategy Acumen

Data Science

Data Engineering

Scientific Method

Math

Statistics

Advanced Computing

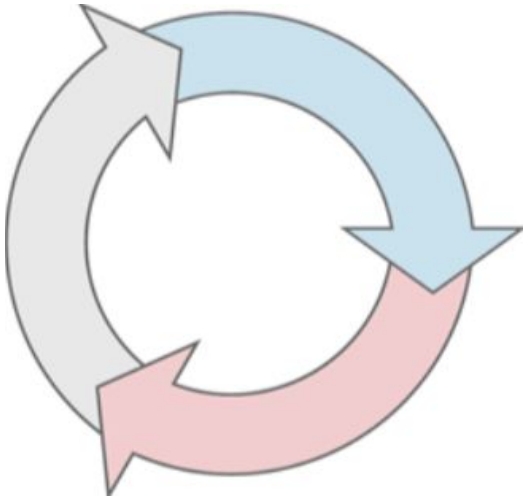
Visualization

Hacker Mindset

Domain Expertise

Etapas del proceso en Data Science

ADQUISICIÓN → PREPARACIÓN → ANÁLISIS → REPORTE → ACT



Proceso Iterativo

ADQUISICIÓN

- Identificar de los data sets
- Recuperar Datos
- Consultar Datos

ADQUISICIÓN | PREPARACIÓN | ANÁLISIS | REPORTE | ACT

Dónde están los Datos?

Bases de Datos Tradicionales

Archivos de Texto

JavaScript

python™

Ruby

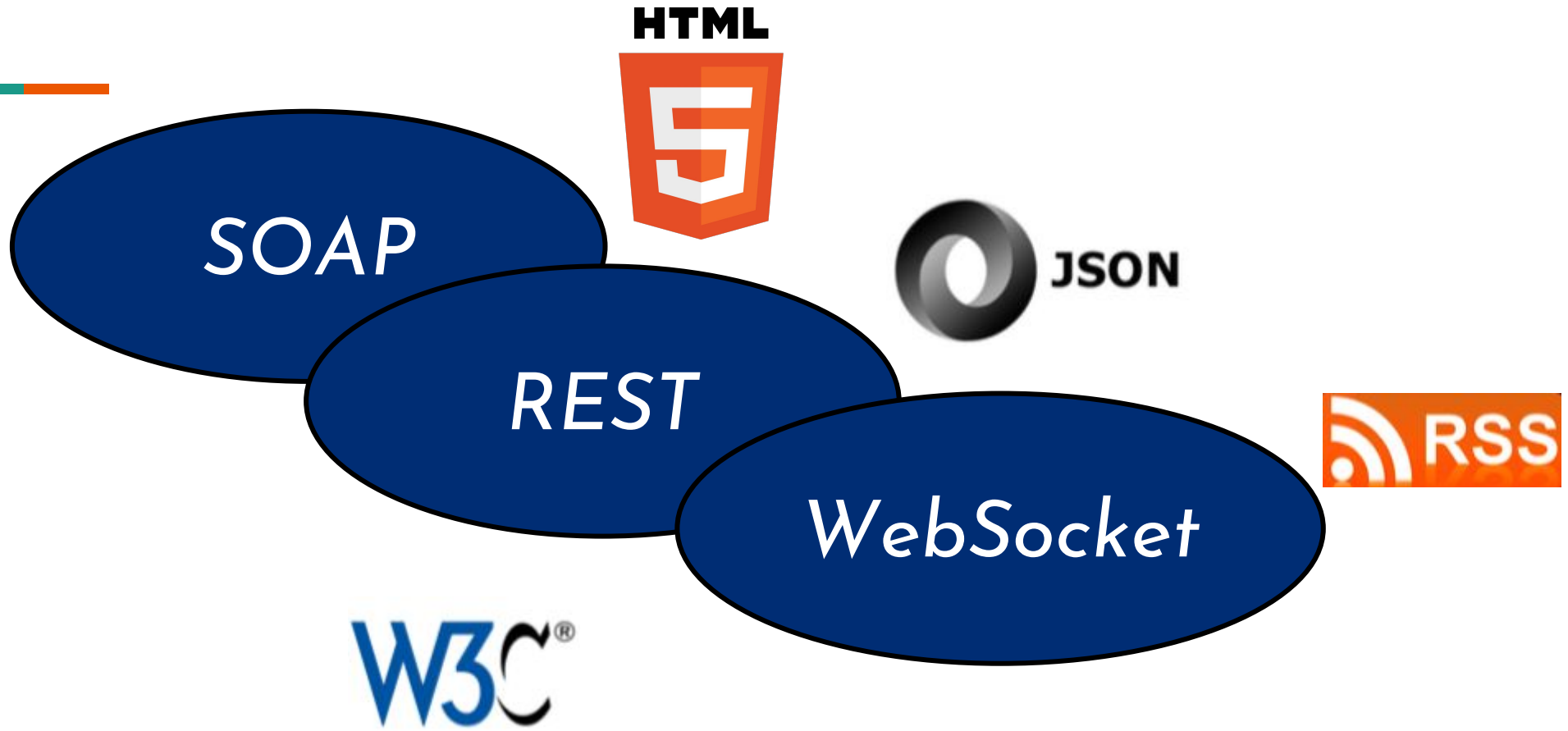
R

php

Perl

Scripting Languages

Datos Remotos



Web Services

Bases de Datos NoSQL

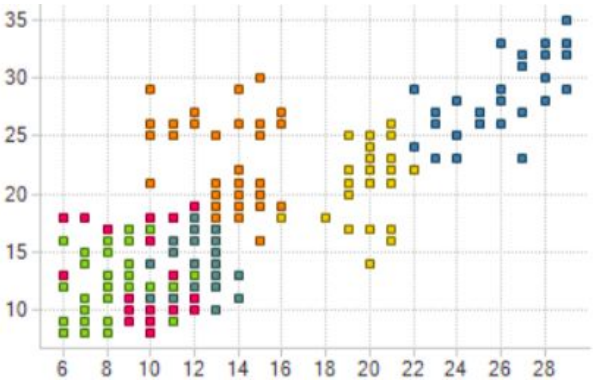
API

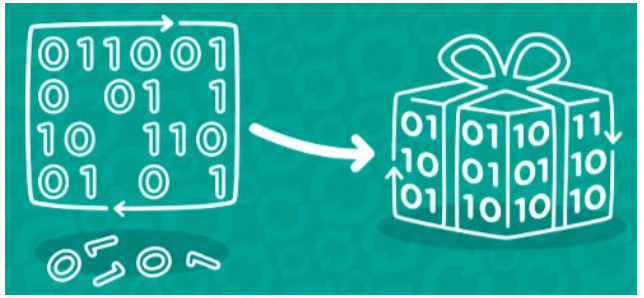
Web Services

Importante ETAPA!

PREPARACIÓN

- Exploración
  - Entender la Naturaleza de los datos
  - Análisis preliminar
- Pre-procesamiento
  - Limpieza -> Integración ->Empaquetado

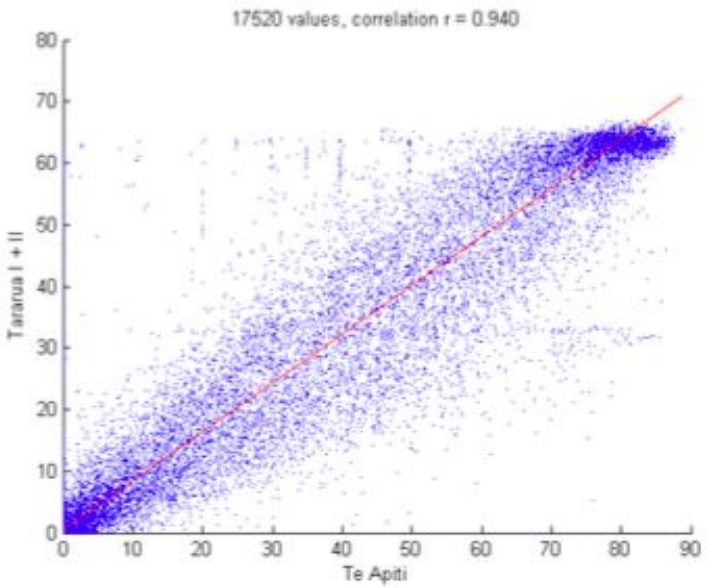


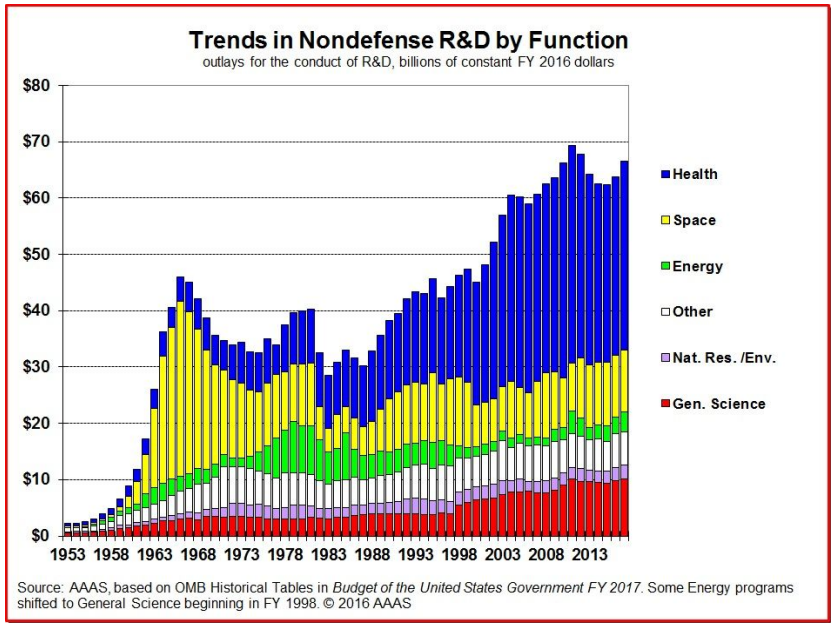


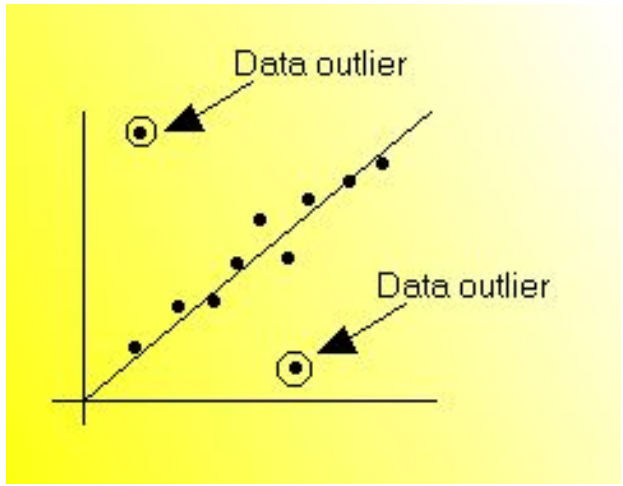
ADQUISICIÓN | PREPARACIÓN | ANÁLISIS | REPORTE | ACT

EXPLORACIÓN

META: Entender los Datos





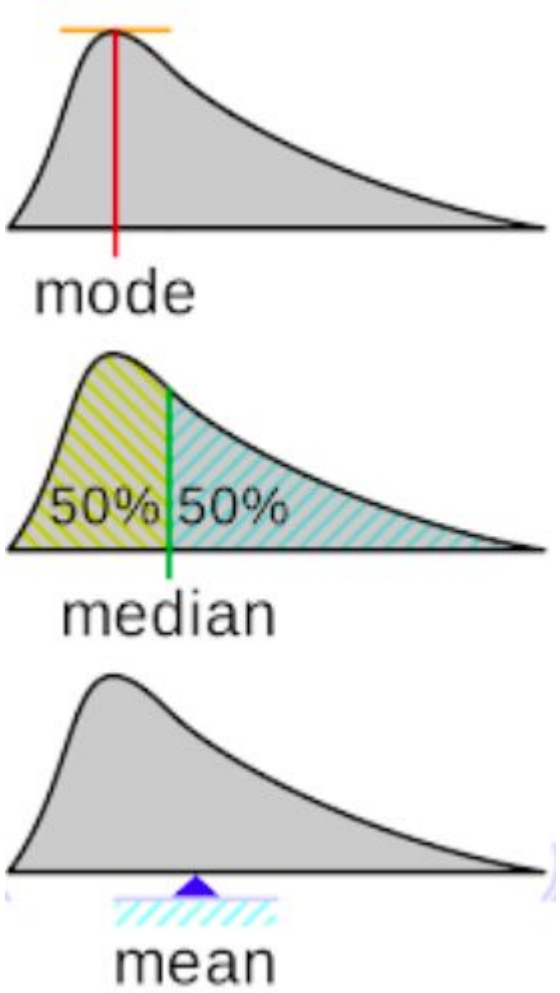


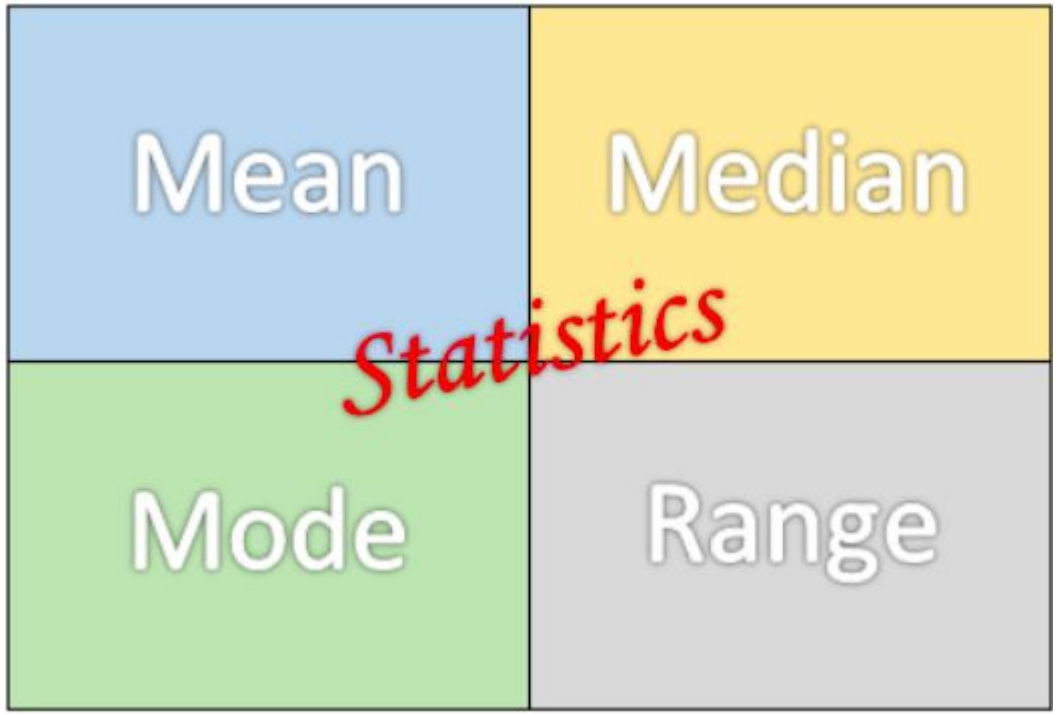
Correlations

Trends

Outliers

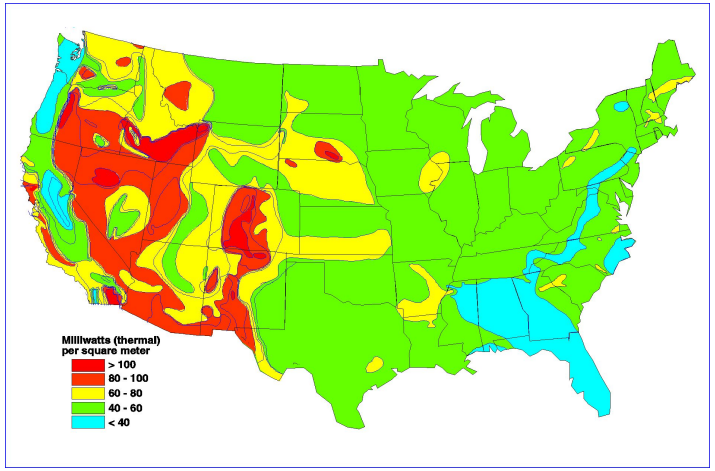
META: Describir los datos

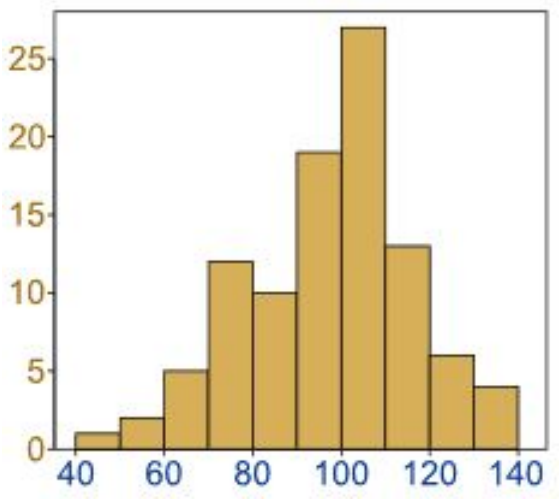




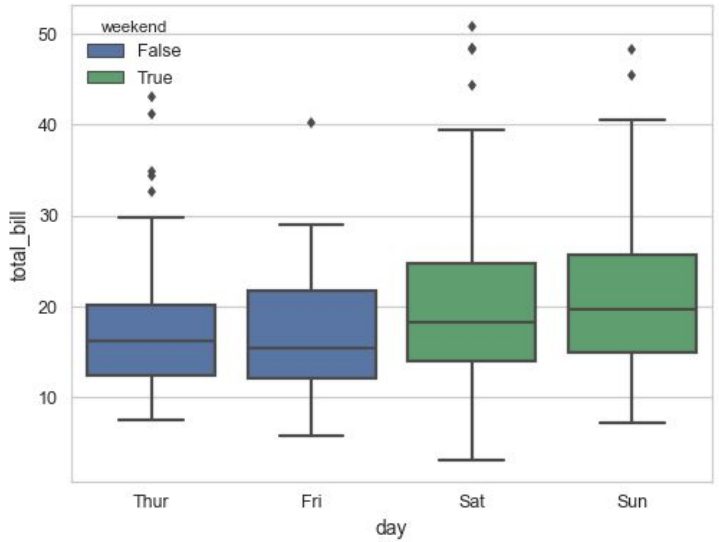
META: Visualizar los Datos

Heat maps



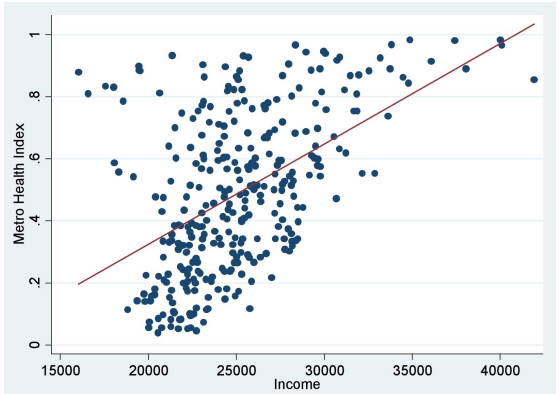


Histograms

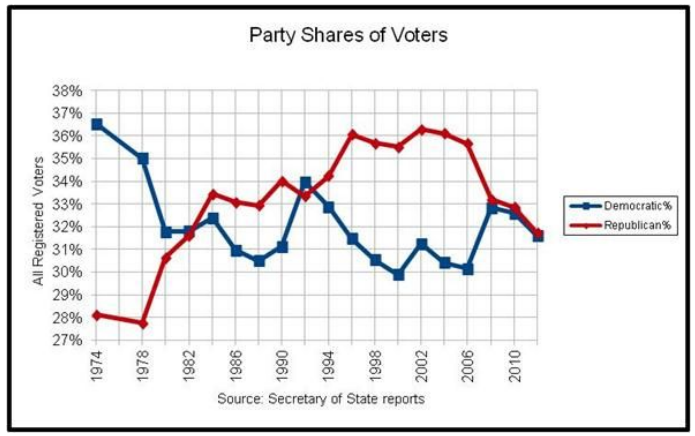


Box plots

Scatter Plots



Line graphs

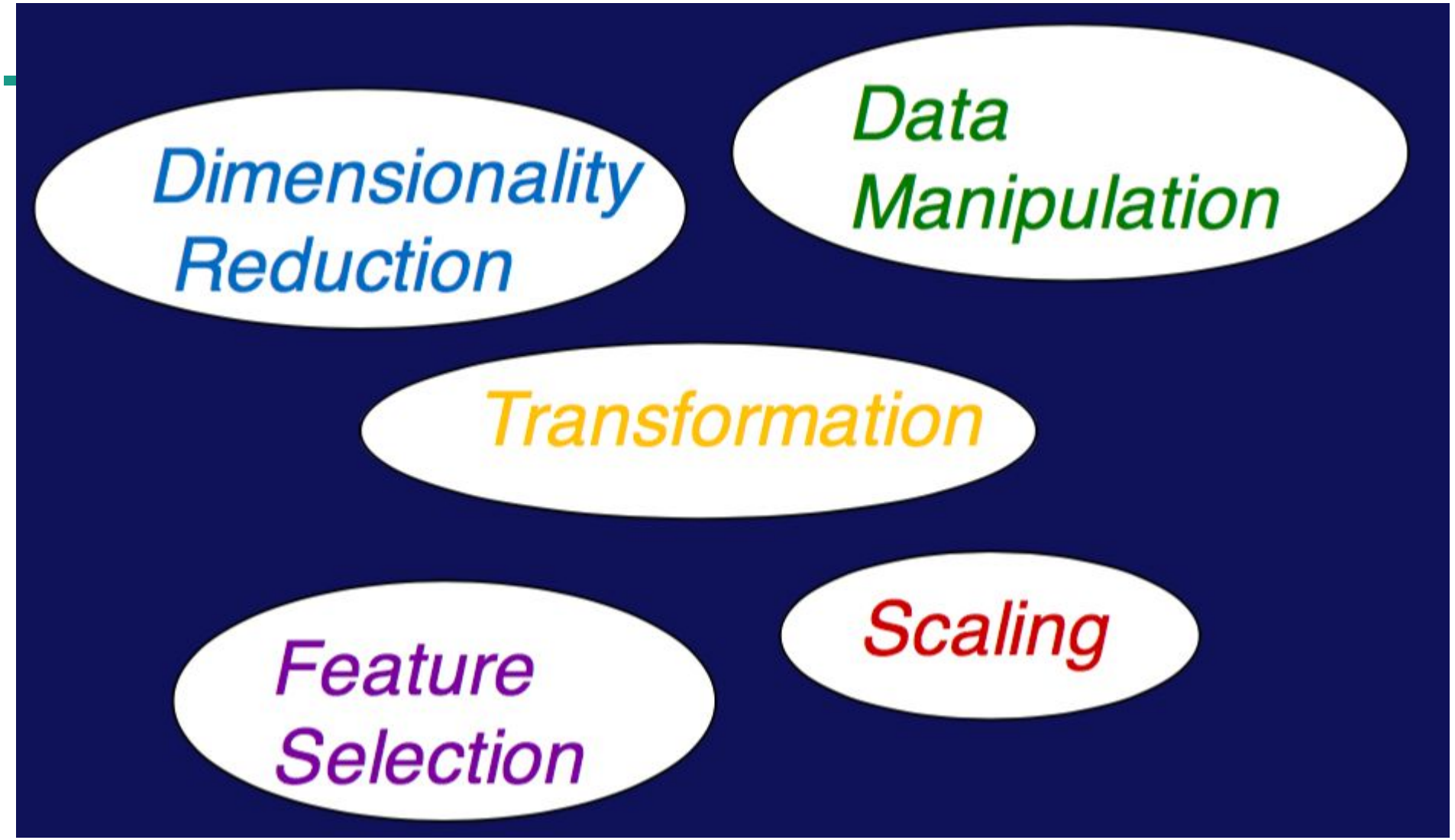


PRE-PROCESAMIENTO

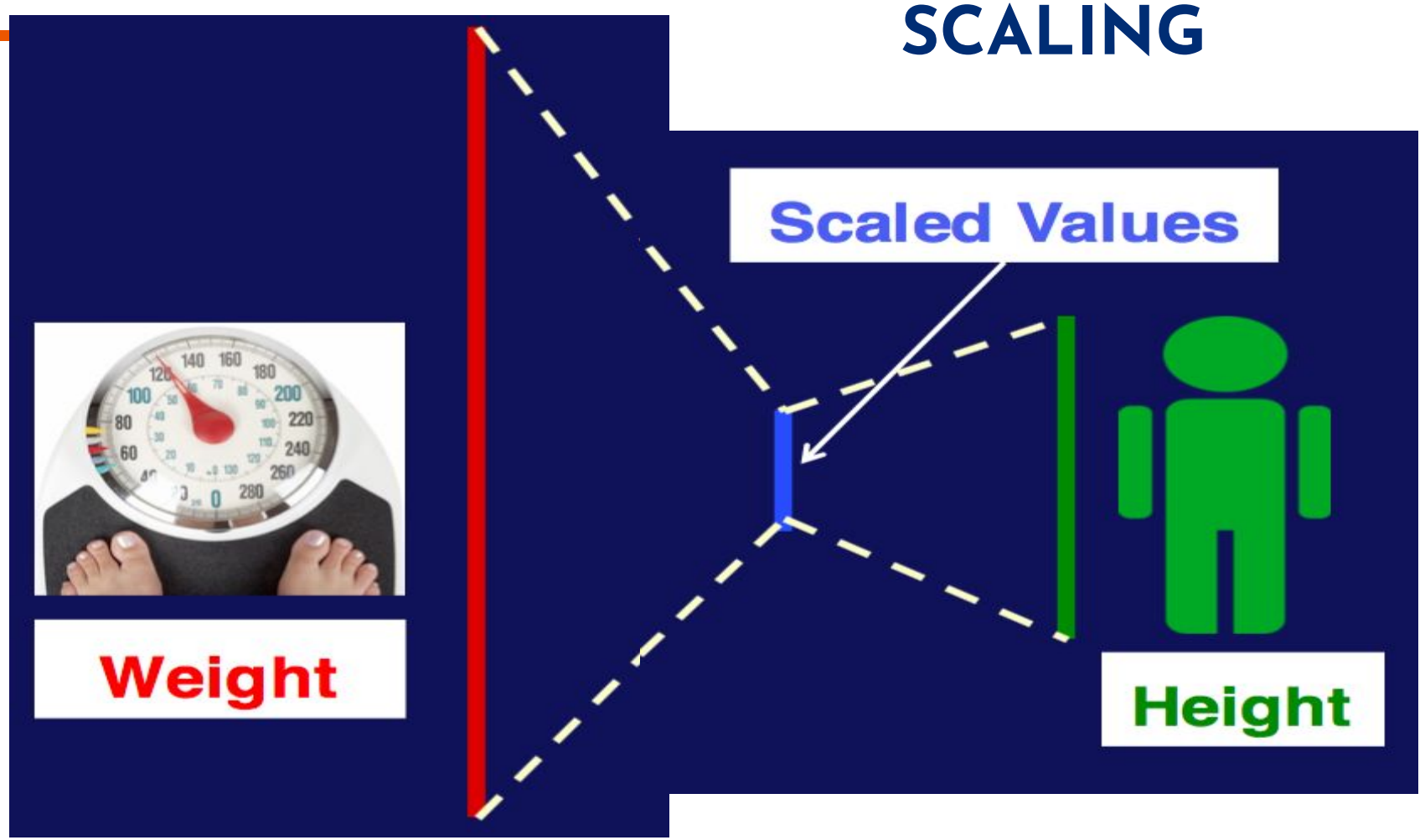
En el mundo real los datos están “Sucios”

Problemas en la calidad de los datos:

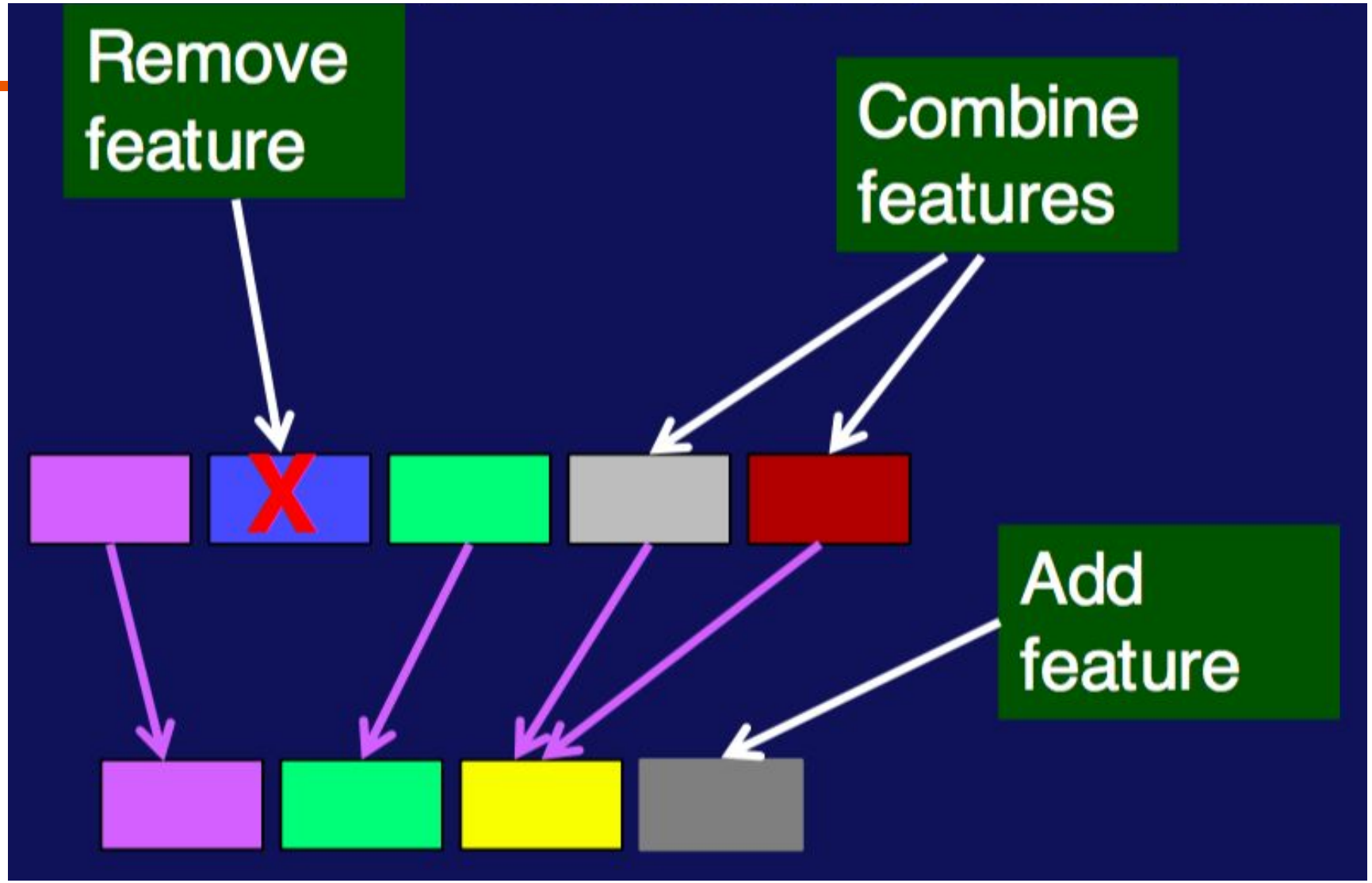
- Registros duplicados
- Valores Missing
- Data Inválida
- Outliers


Dimensionality Reduction
Data Manipulation
Transformation
Feature Selection
Scaling

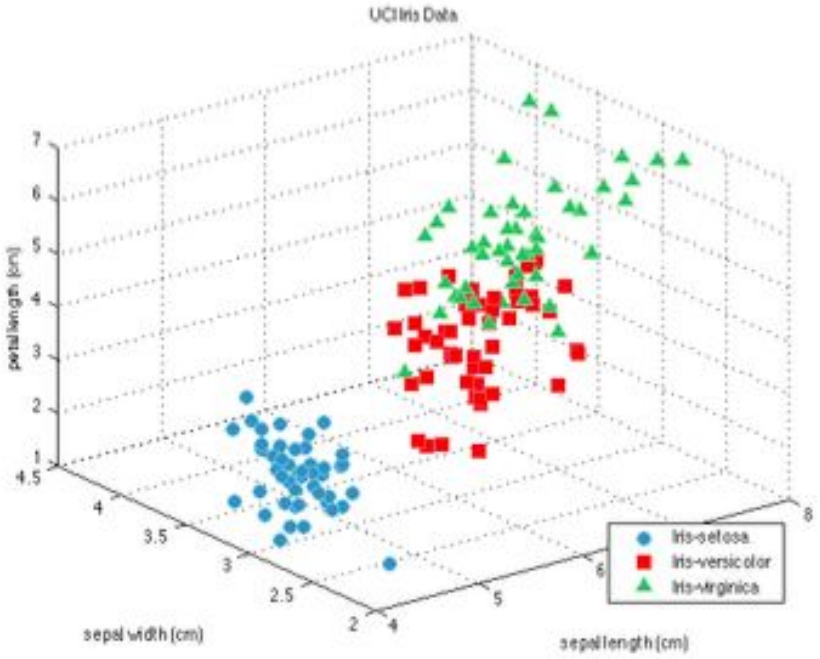
SCALING

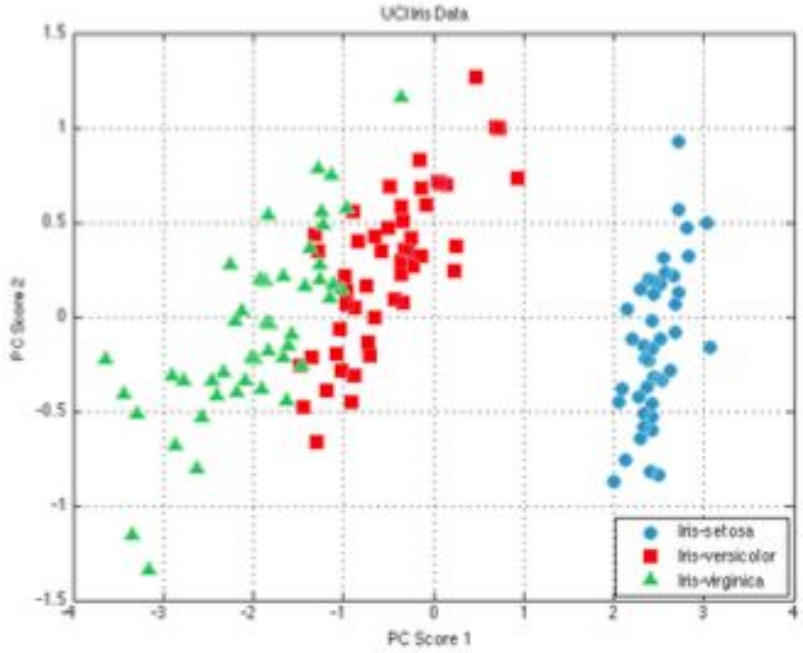


FEATURE SELECTION



DIMENSIONALITY REDUCTION

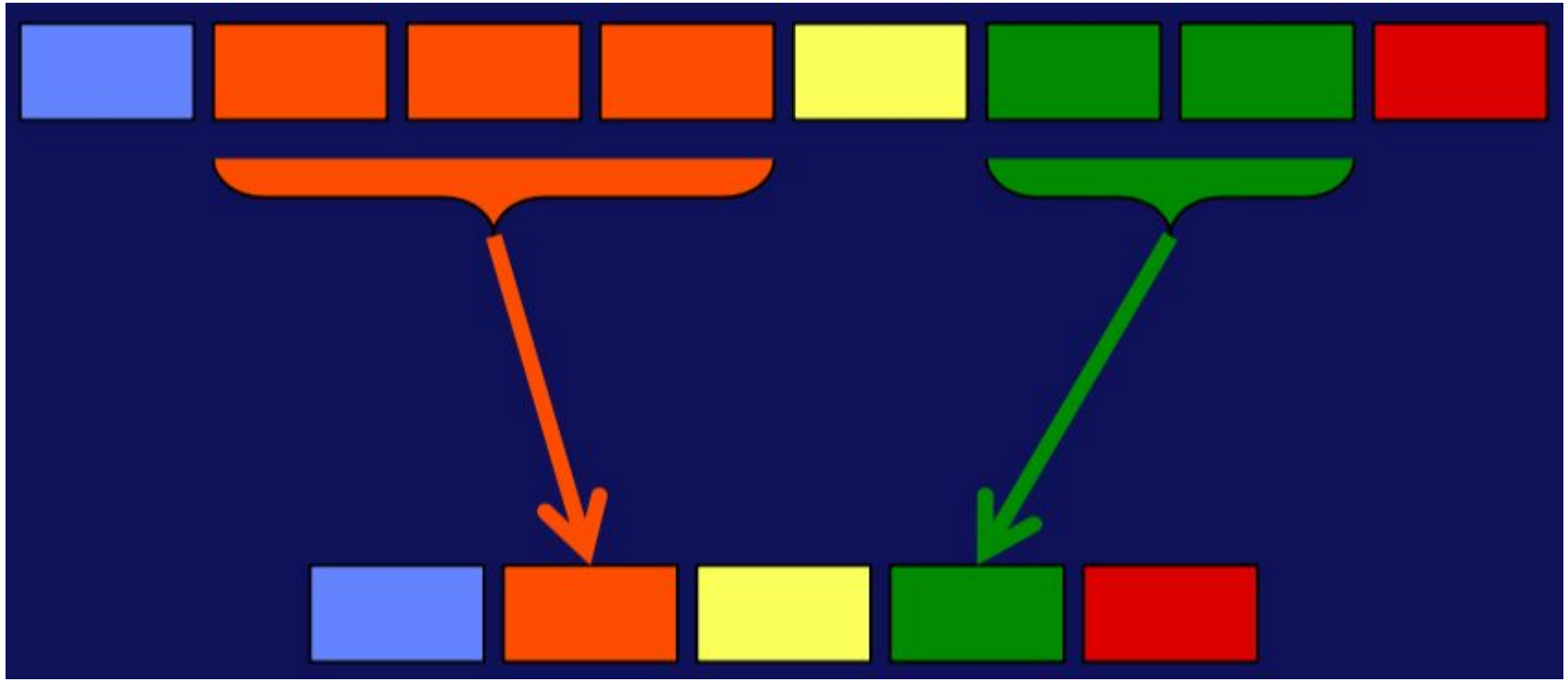




3D

2D

DATA MANIPULATION

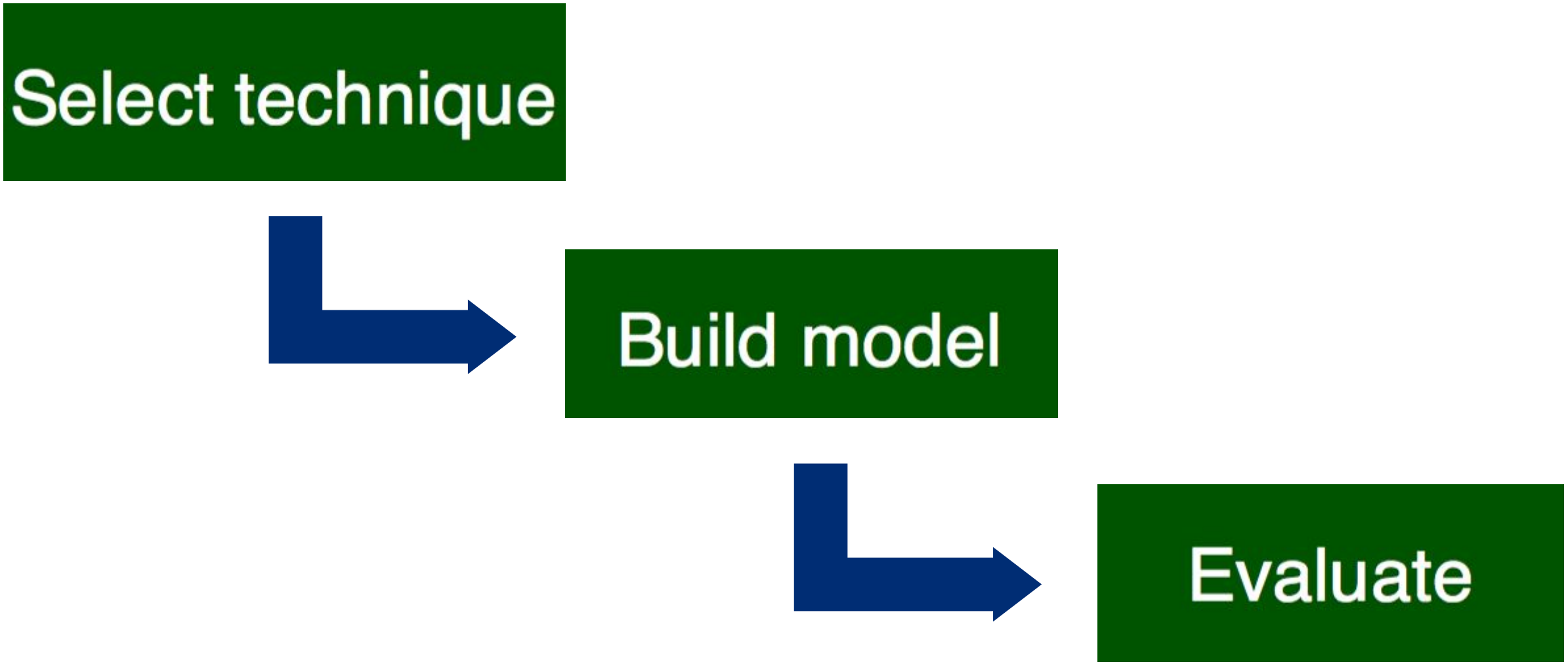


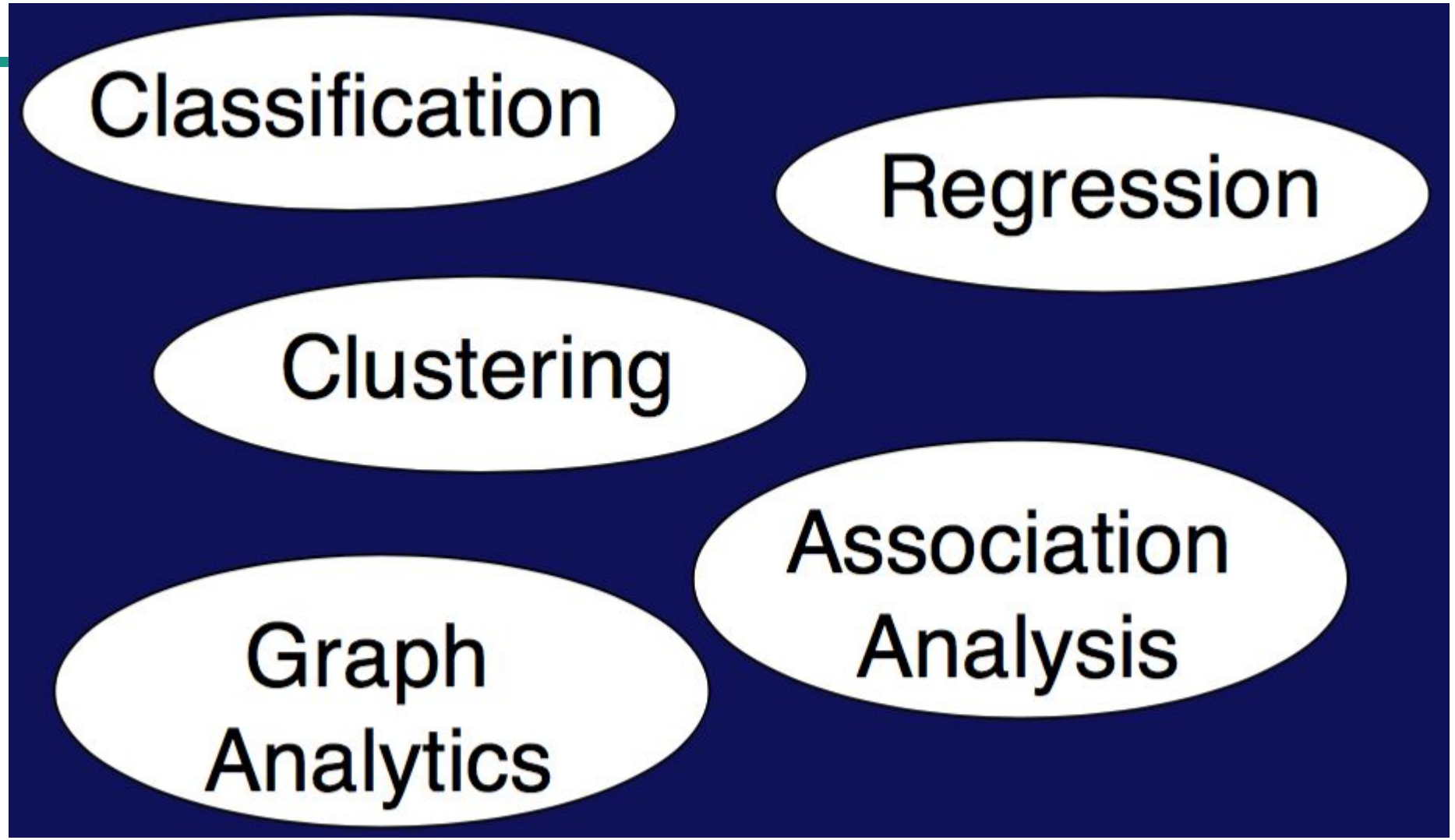
ANÁLISIS

Selección de técnicas de Análisis

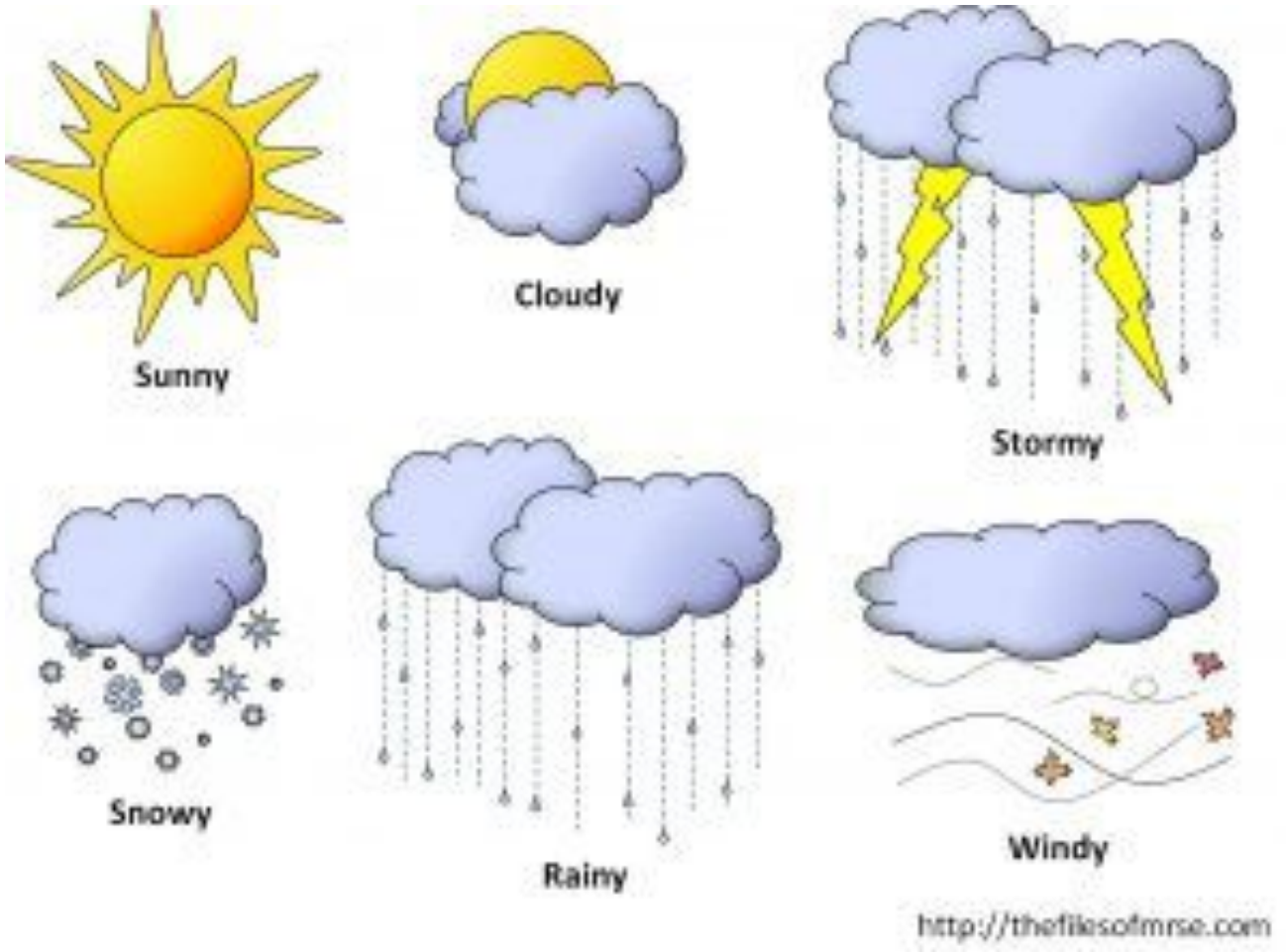
Construcción de modelos

ADQUISICIÓN | PREPARACIÓN | ANÁLISIS | REPORTE | ACT


Select technique
Build model
Evaluate


Classification
Regression
Clustering
Association
Analysis
Graph
Analytics

CLASSIFICATION



Predecir CATEGORÍAS

REGRESSION

Predecir VALORES NUMÉRICOS

CLUSTERING

Organizar items similares en grupos

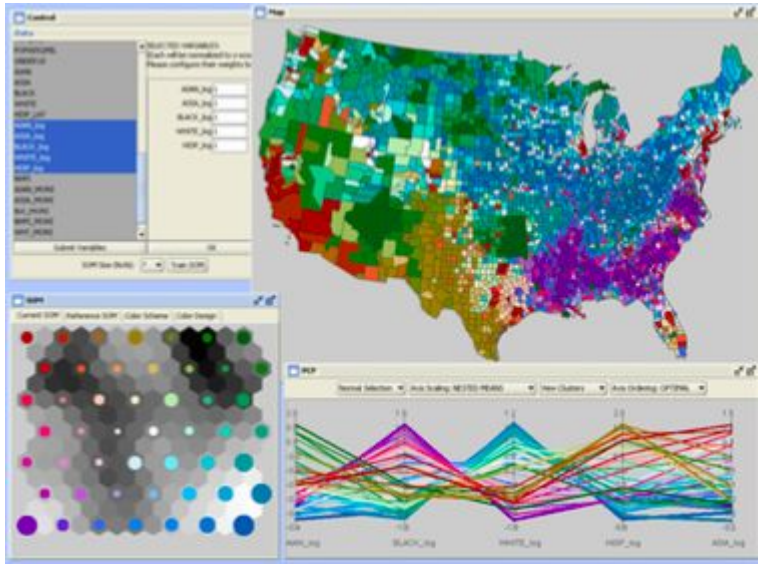
ASSOCIATION ANALYSIS

Descubrir reglas para capturar asociaciones entre variables

GRAPH ANALYTICS

Emplear las estructuras de GRAFOS para encontrar conexiones entre las entidades

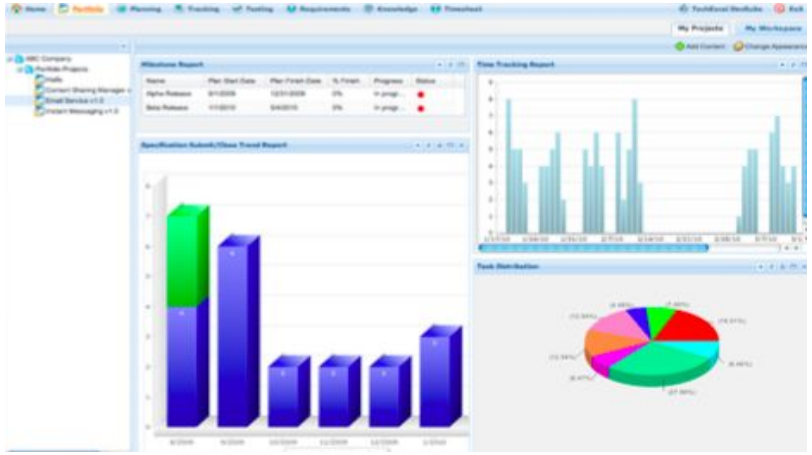
REPORTE



ADQUISICIÓN → PREPARACIÓN → ANÁLISIS → REPORTE → ACT

Presentar

con



usando

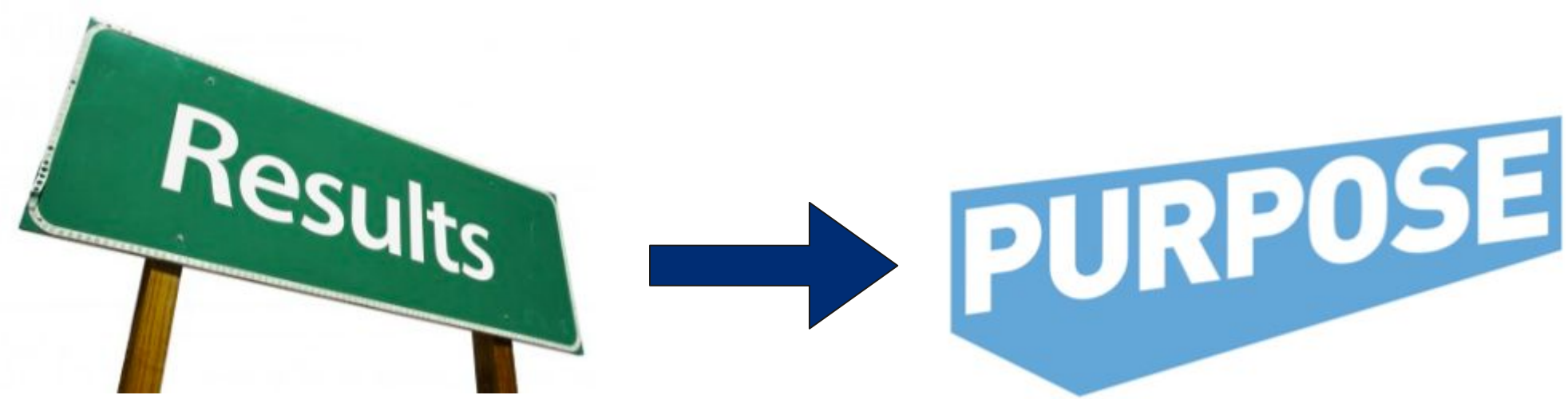
python™

R

tableau

Google Developers CHARTS

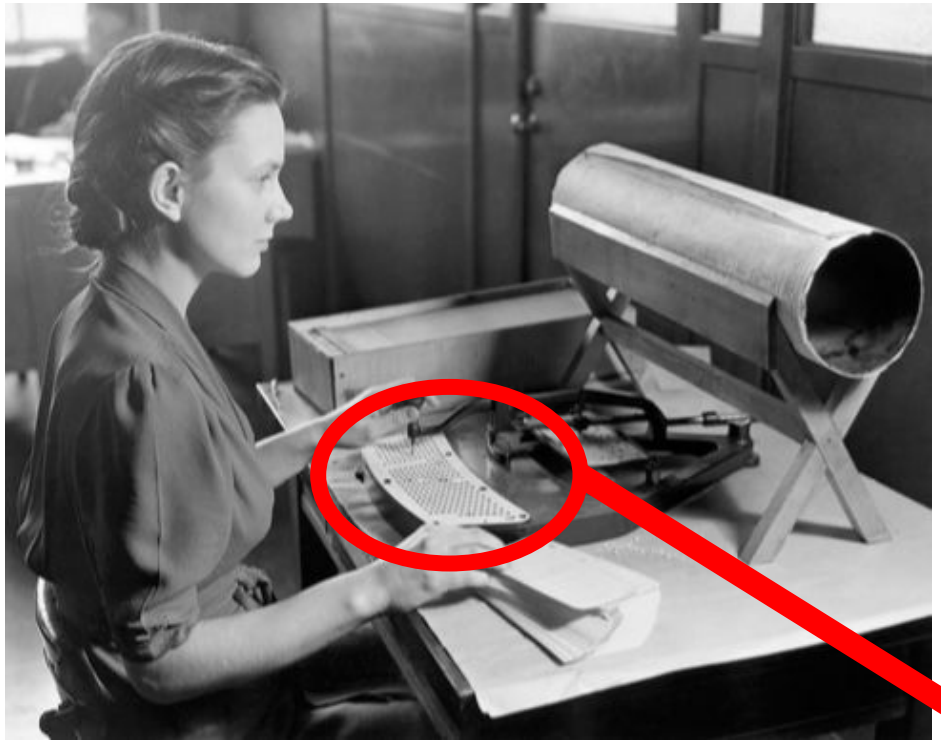
ACT

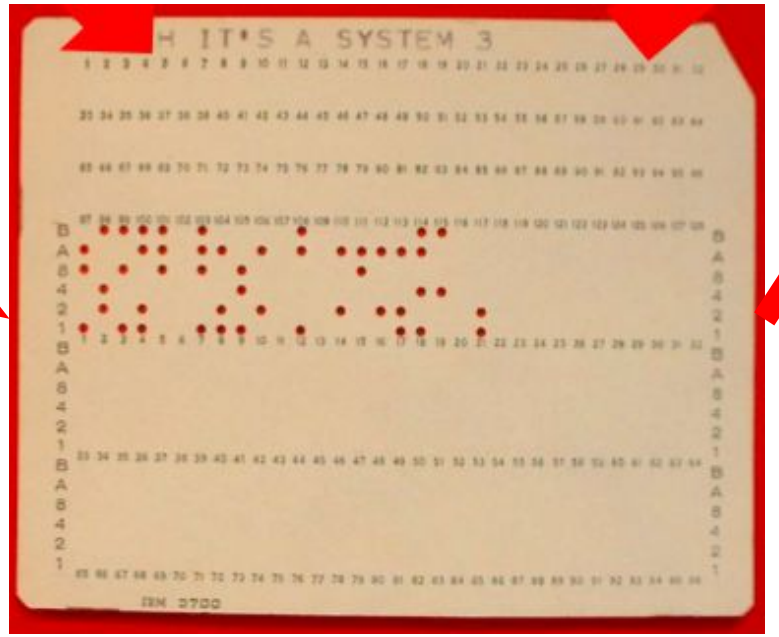


ADQUISICIÓN → PREPARACIÓN → ANÁLISIS → REPORTE → ACT

HDFS -> Hadoop -> Map Reduce

SISTEMA DE FICHEROS (FS)



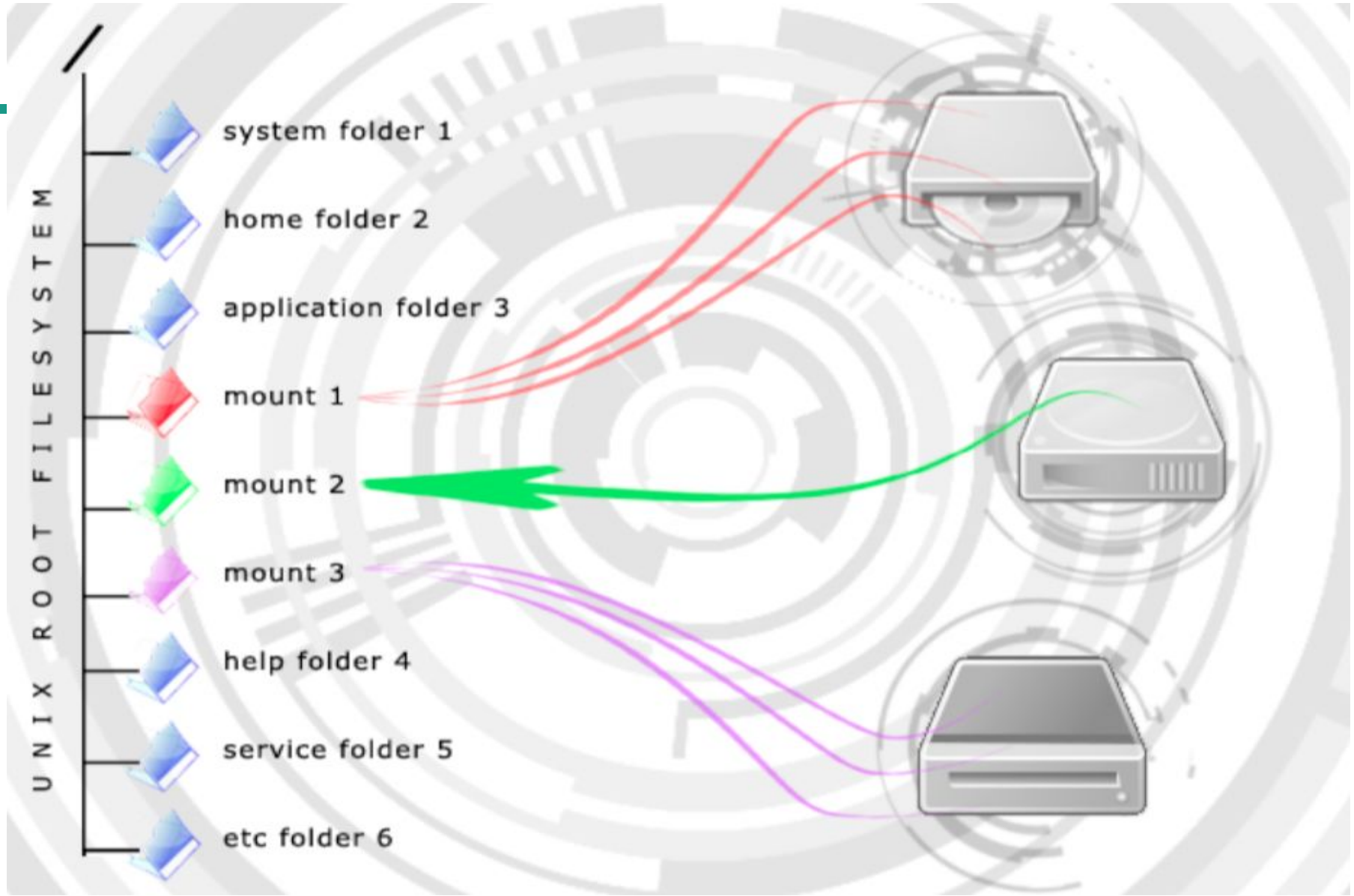

H IT'S A SYSTEM 3
IBM 3700


SIMCP
PROJF
SLOW VER
GDRAW // PROJF DEMO
// XEQ GDRAW
PROJF
FAST VER

SISTEMA DE FICHEROS (FS)

- Almacenamiento de Información a “Largo Plazo”
- Almacenar Grandes Cantidades de Información
- Permitir acceso de múltiples procesos

ARCHIVO


/
UNIX ROOT FILESYSTEM
system folder 1
home folder 2
application folder 3
mount 1
mount 2
mount 3
help folder 4
service folder 5
etc folder 6


64GB
256GB
1TB
5TB

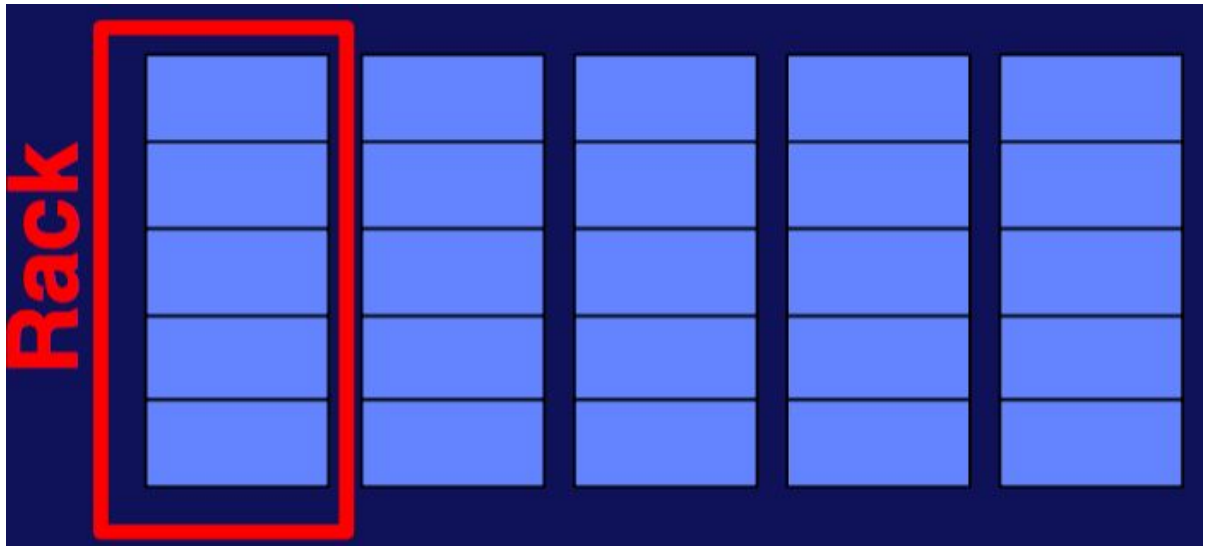
Y si necesitamos almacenar más información?

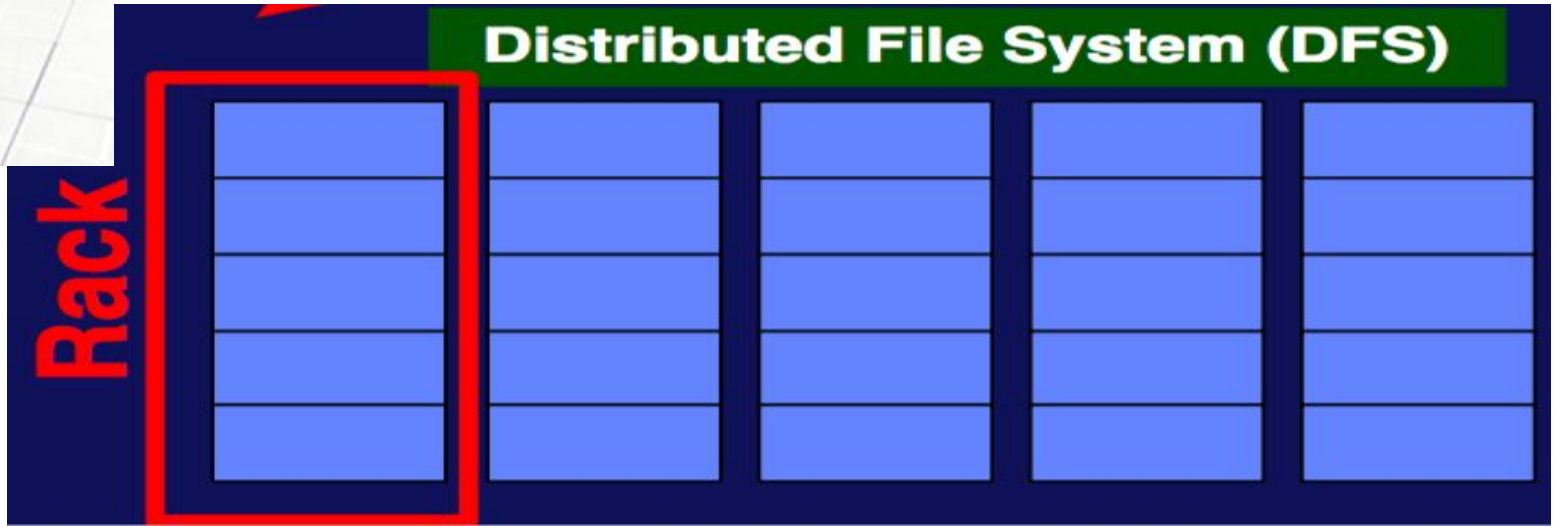
- Comprar discos de mayor capacidad?
- Copiar los datos a un dispositivo externo?

Trabajo

Personal

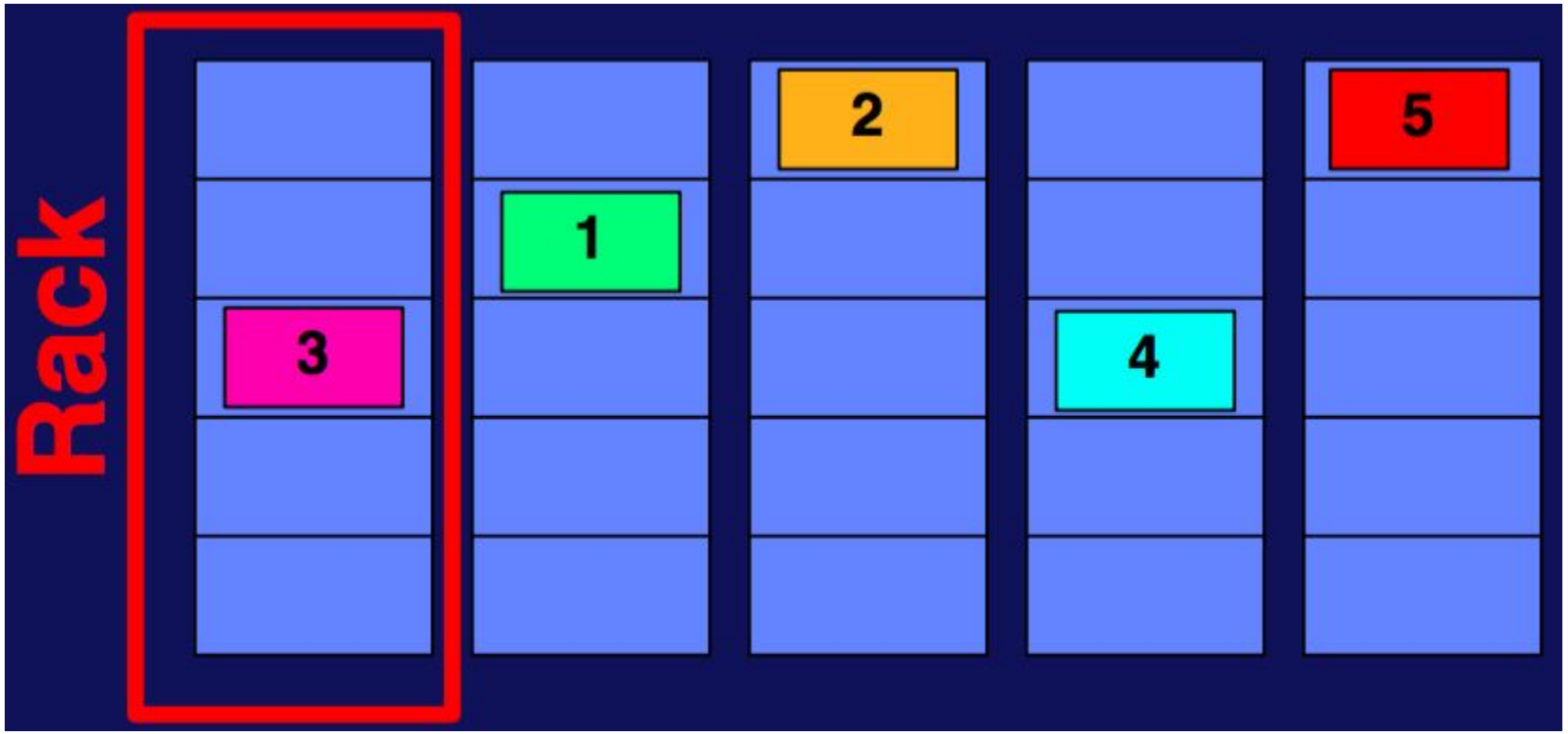
SISTEMA DE FICHEROS DISTRIBUIDOS (DFS)





SISTEMA DE FICHEROS DISTRIBUIDOS (DFS)

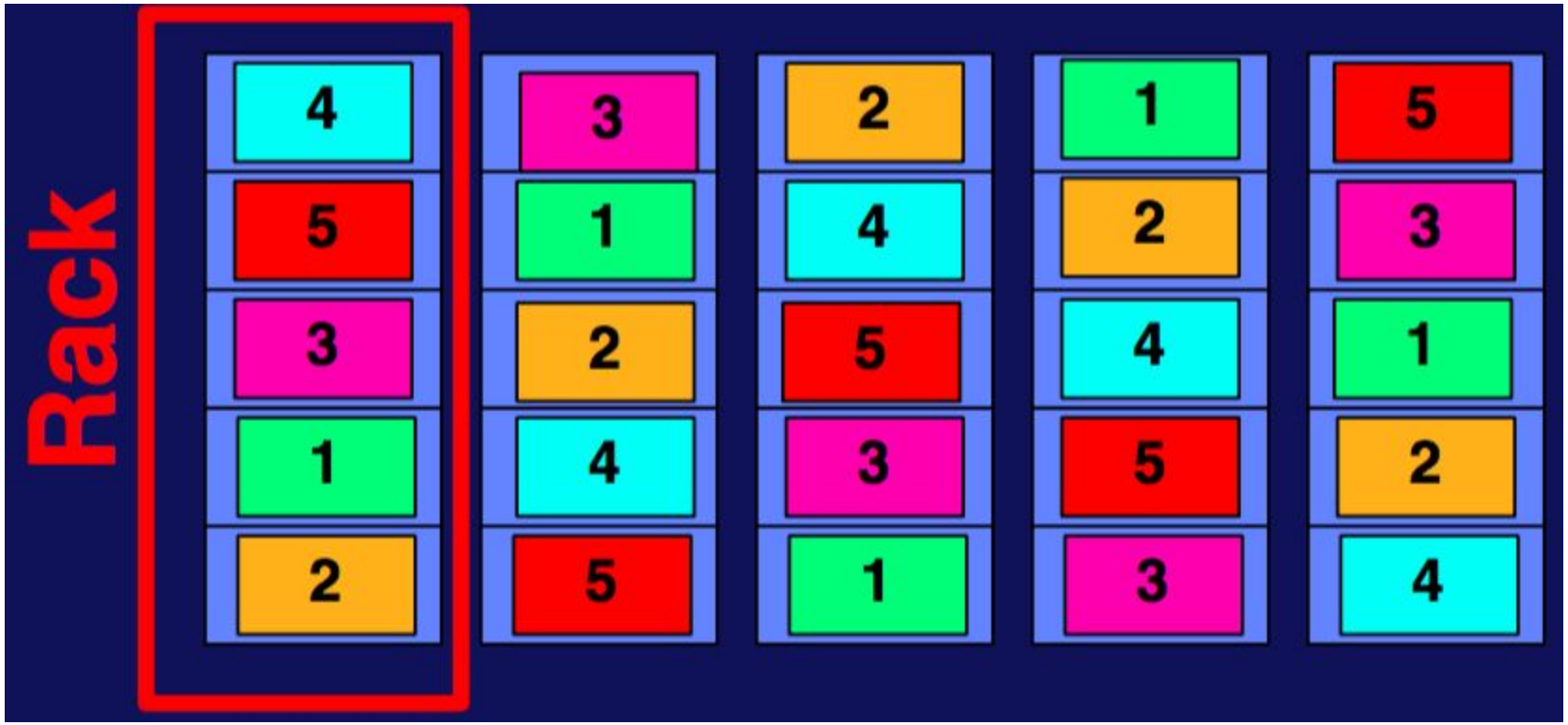
DATOS



Data Partition

SISTEMA DE FICHEROS DISTRIBUIDOS (DFS)

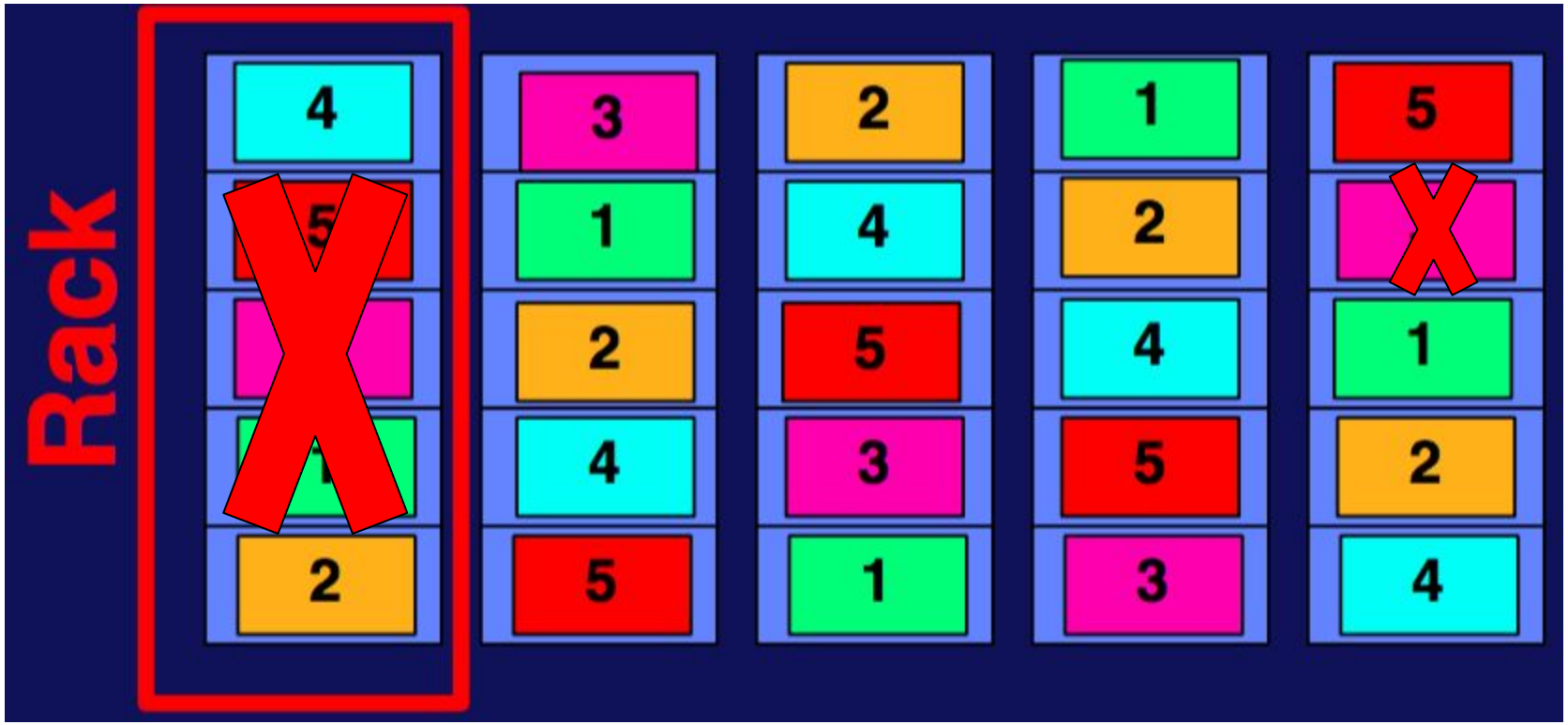
DATOS



Data Replication

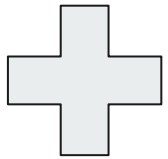
SISTEMA DE FICHEROS DISTRIBUIDOS (DFS)

DATOS

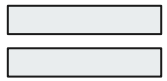


Alta Concurrencia vs. Baja Consistencia

Data Replication



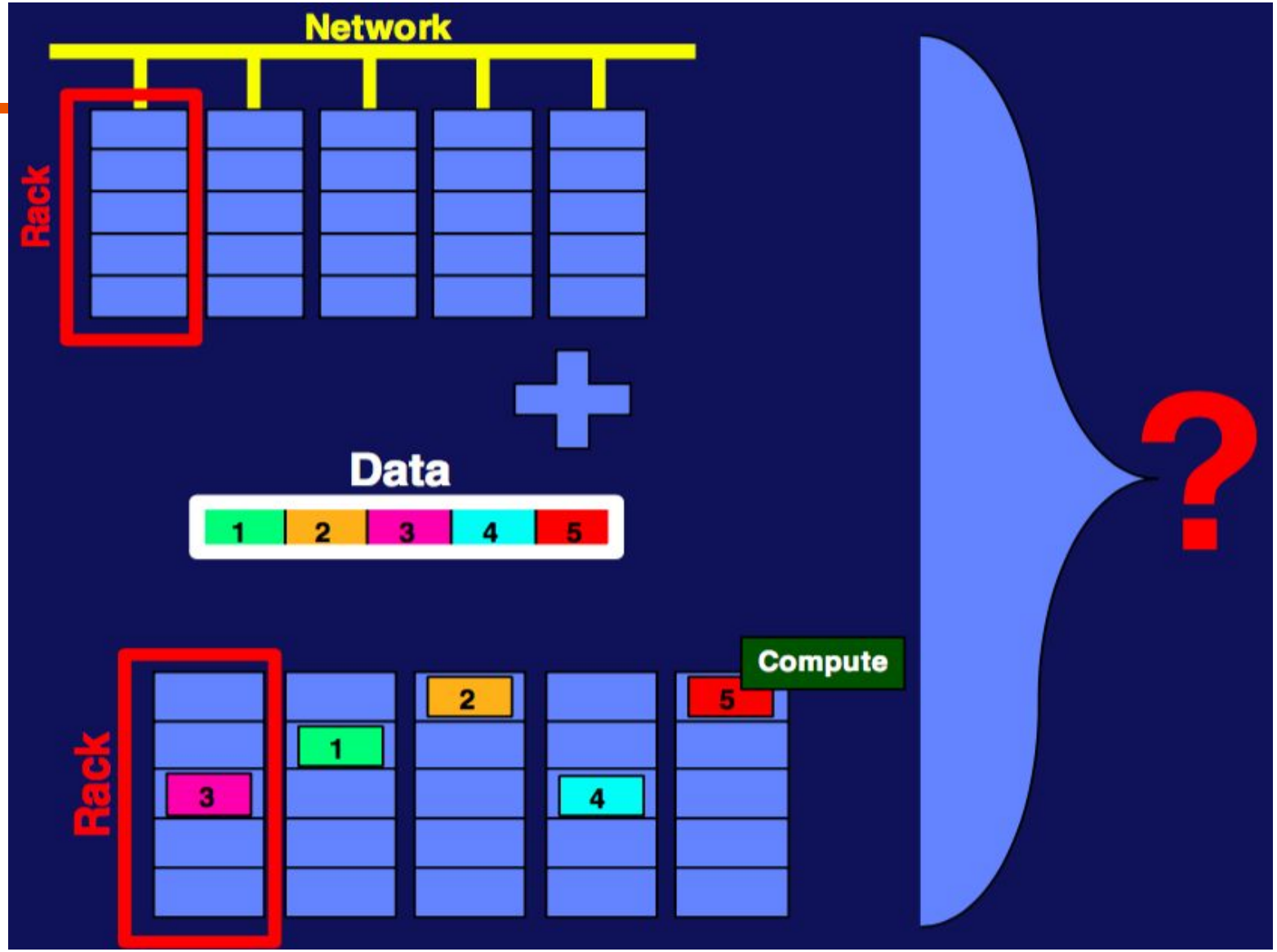
Data Partition



Escalabilidad Tolerancia a Fallas Alta Concurrencia

Paralelismo

Modelos de Programación para Big Data


Network
Rack
Data
1
2
3
4
5
Rack
3
1
2
4
5
Compute
?

Requerimientos para Modelos de Programación en Big Data

1. Soportar operaciones Big Data:
    a. Split Volumes of data
    b. Acceso rápido a los datos
    c. Distribución de carga

Requerimientos para Modelos de Programación en Big Data

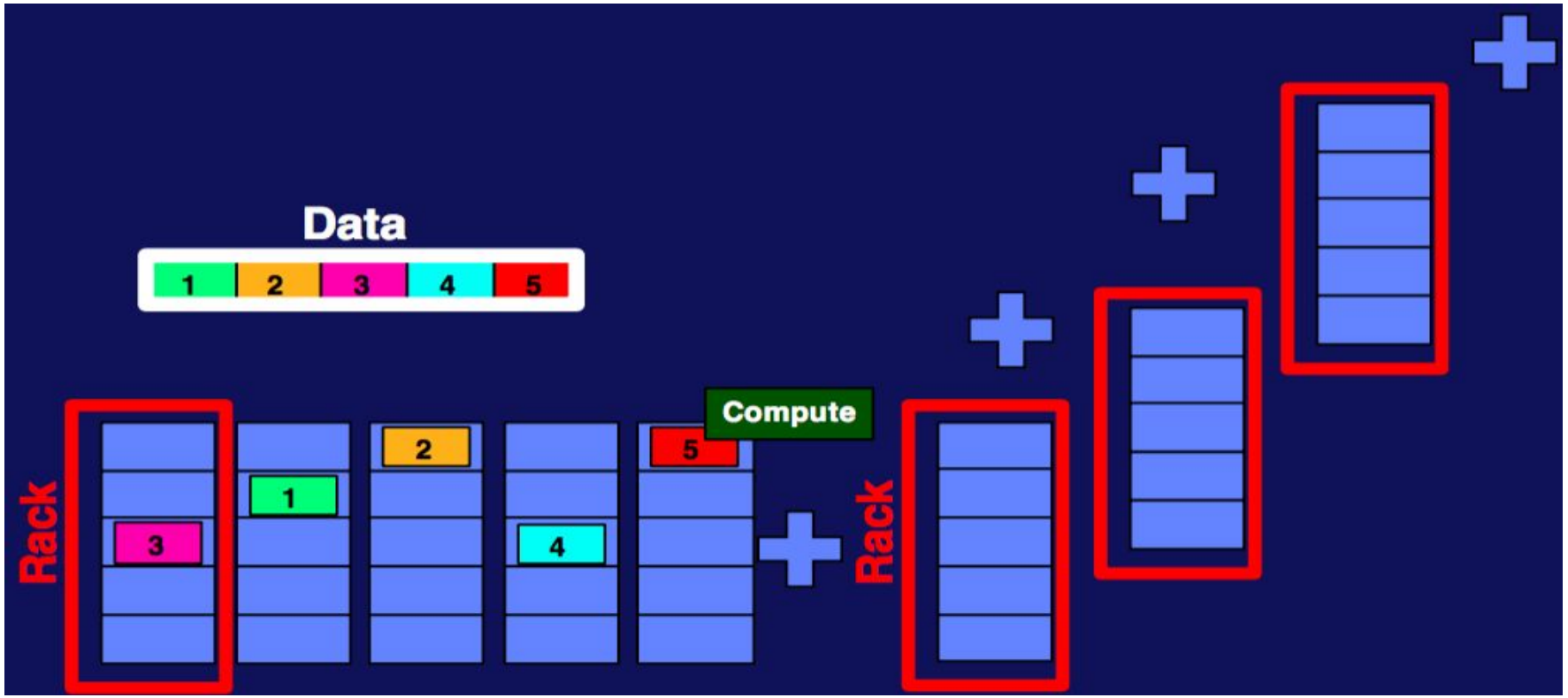
2. Manejo de la Tolerancia a fallas:

   a. Replicar particiones de los datos

   b. Recuperación de Archivos cuando sea “necesario”

Requerimientos para Modelos de Programación en Big Data

3. Capacidad de adicionar más Racks:



Requerimientos para Modelos de Programación en Big Data

4. Optimizado para tipos de datos específicos

Document

Table

Key-value

Graph

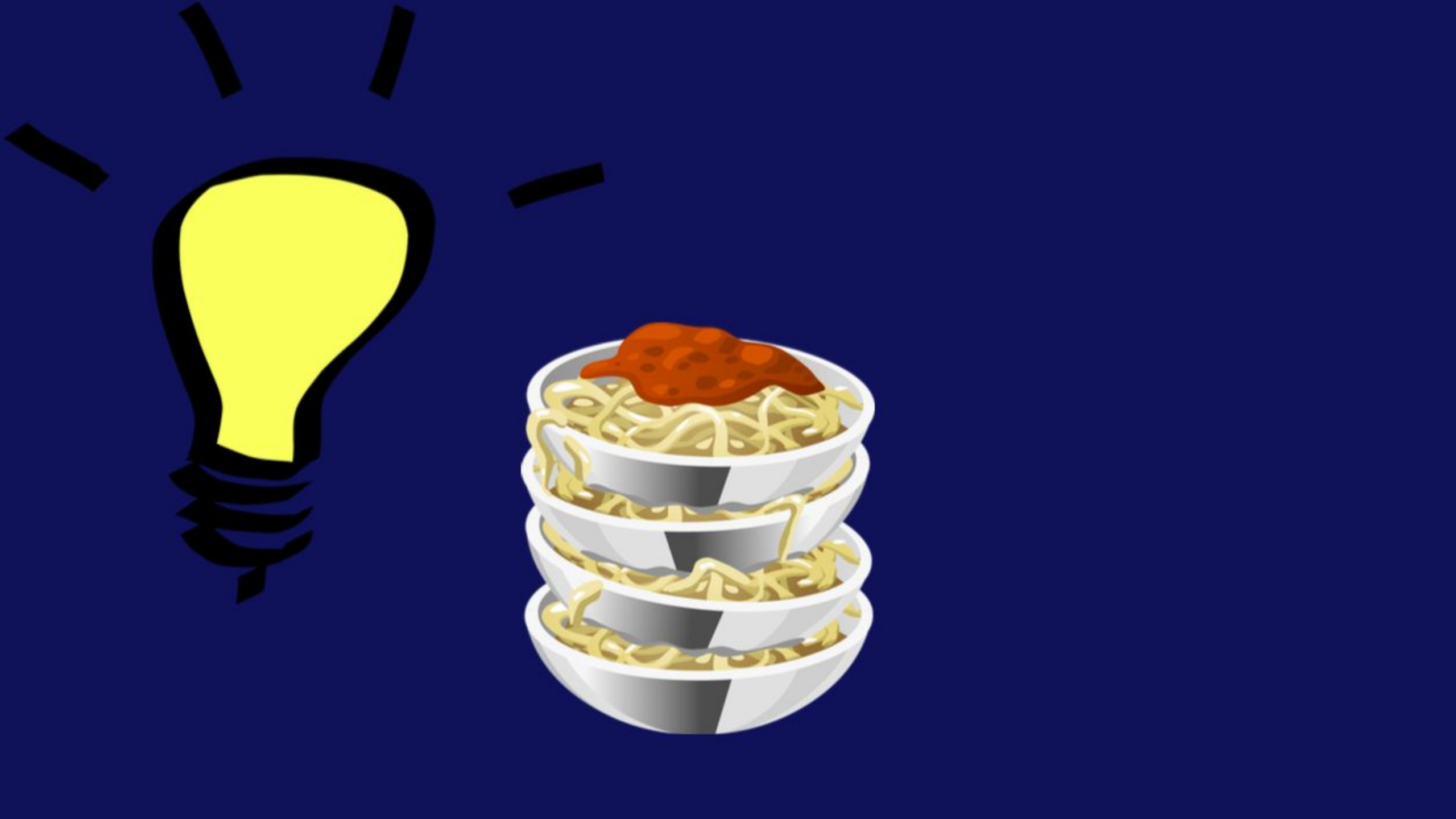
Multimedia

Stream

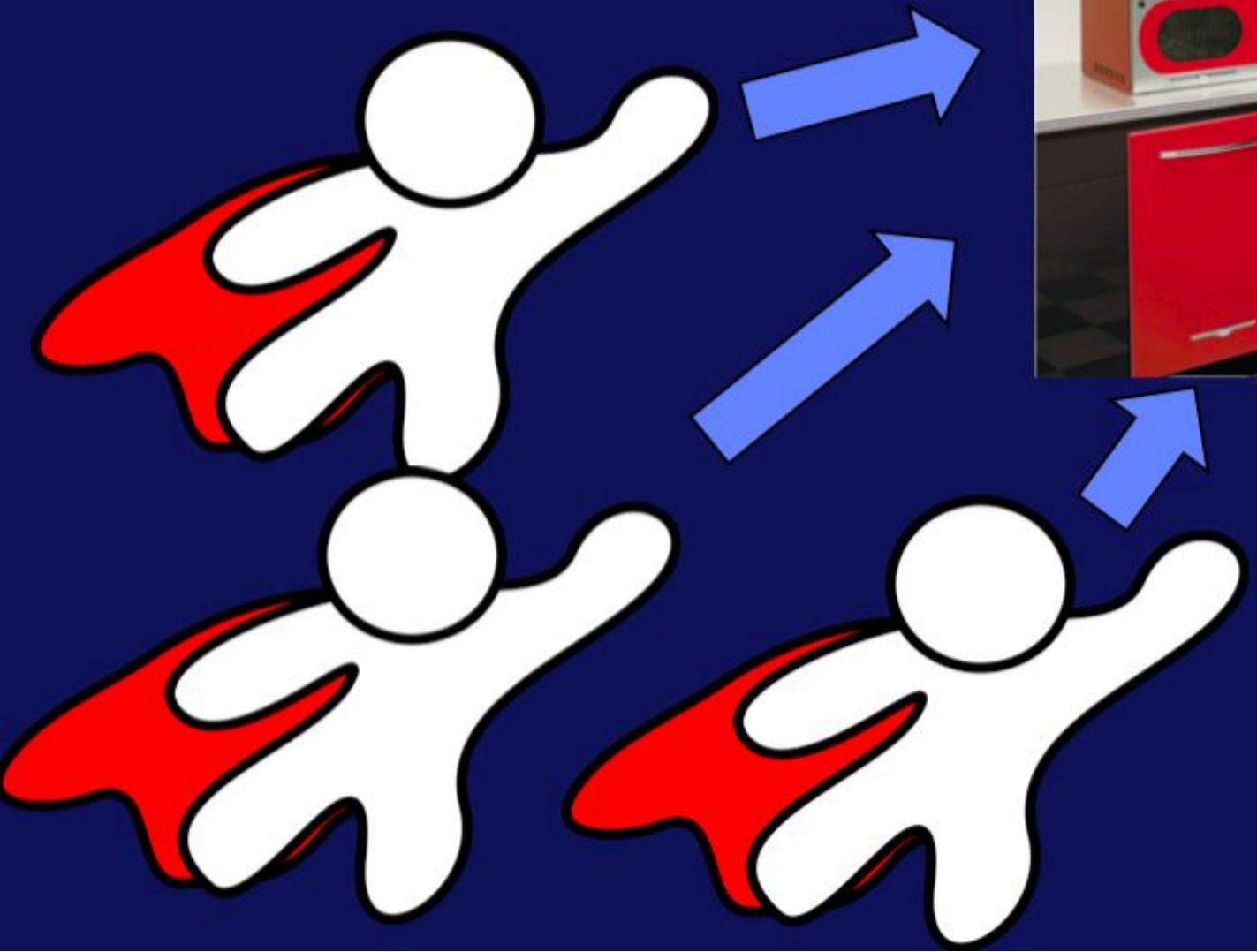
MODELANDO EL PARADIGMA

CASO: PREPARACIÓN DE SALSA DE TOMATE


REUNIÓN EN MEDIA HORA

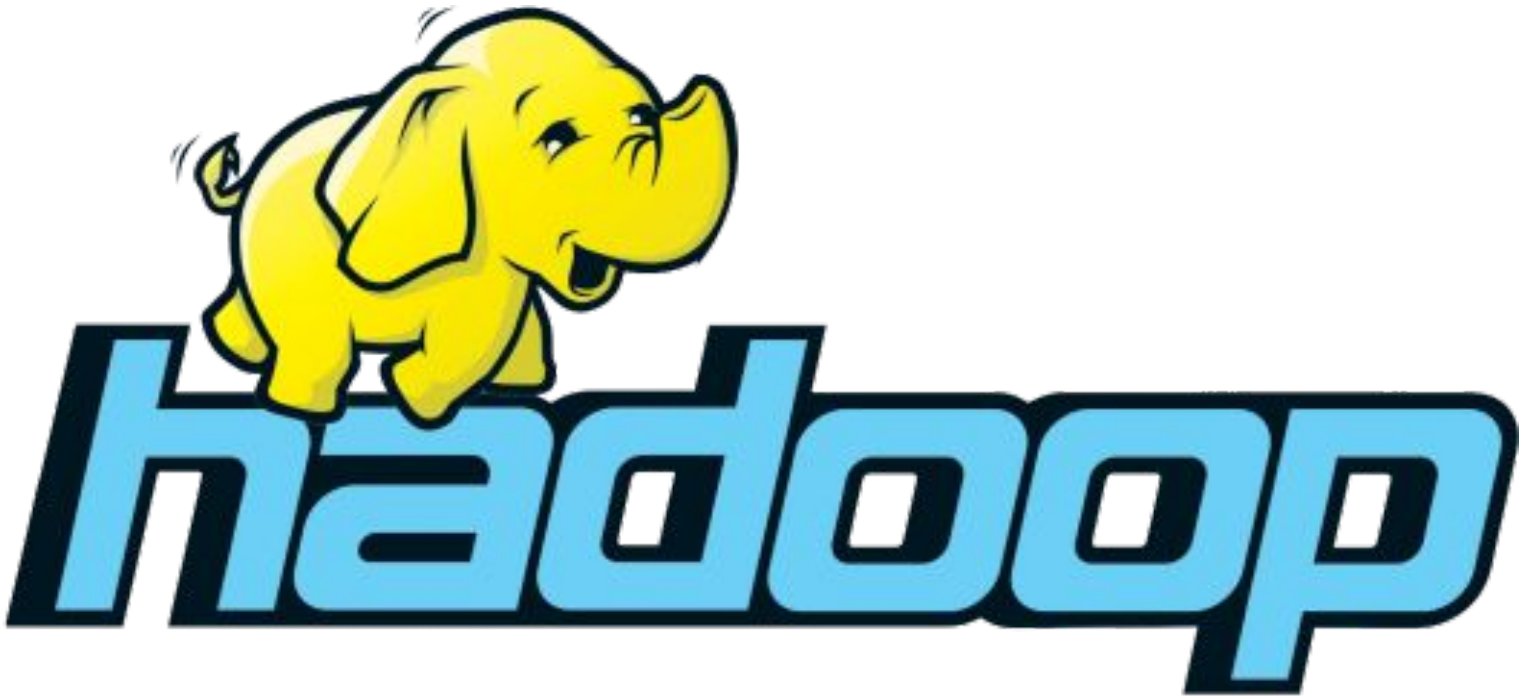


AYUDANTES



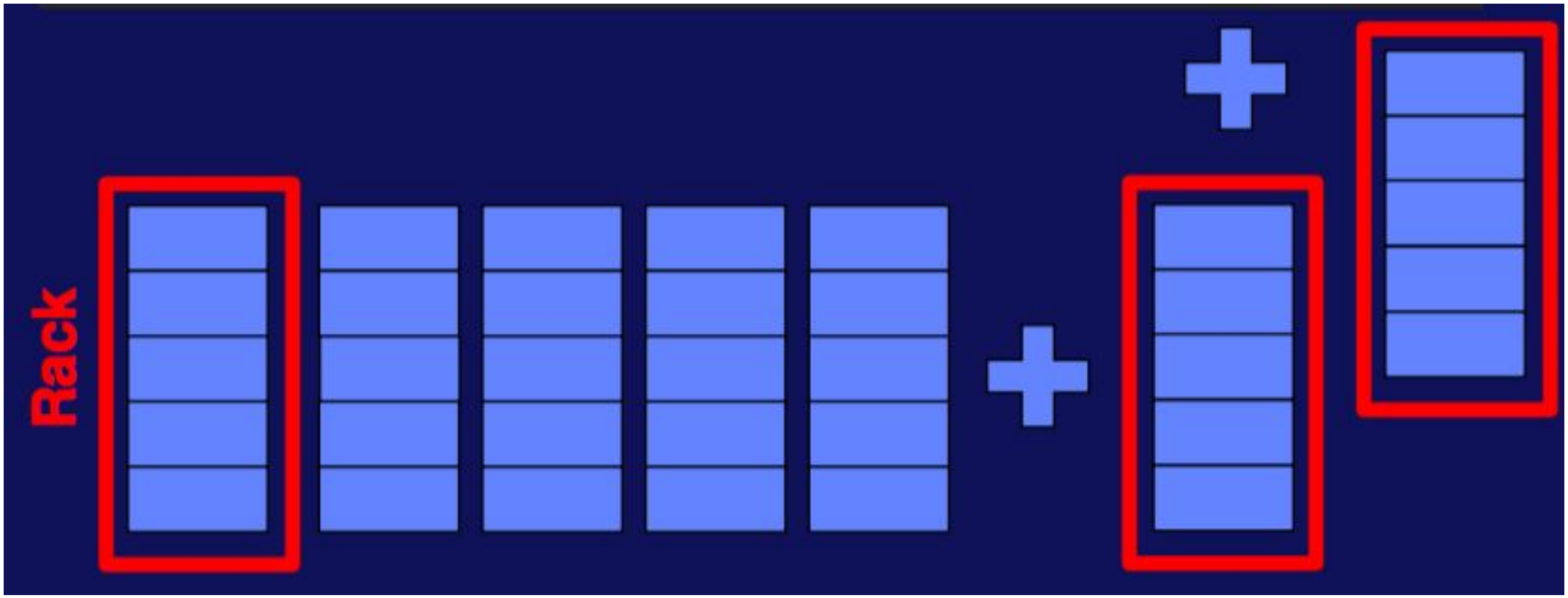
MAP REDUCE

- ➔ Modelo de Programación para Big Data
- ➔ Existen muchas implementaciones
- ➔ Soporta Grandes Volúmenes de Datos
- ➔ Proporciona Tolerancia a Fallas
- ➔ Habilita Escalamiento


hadoop

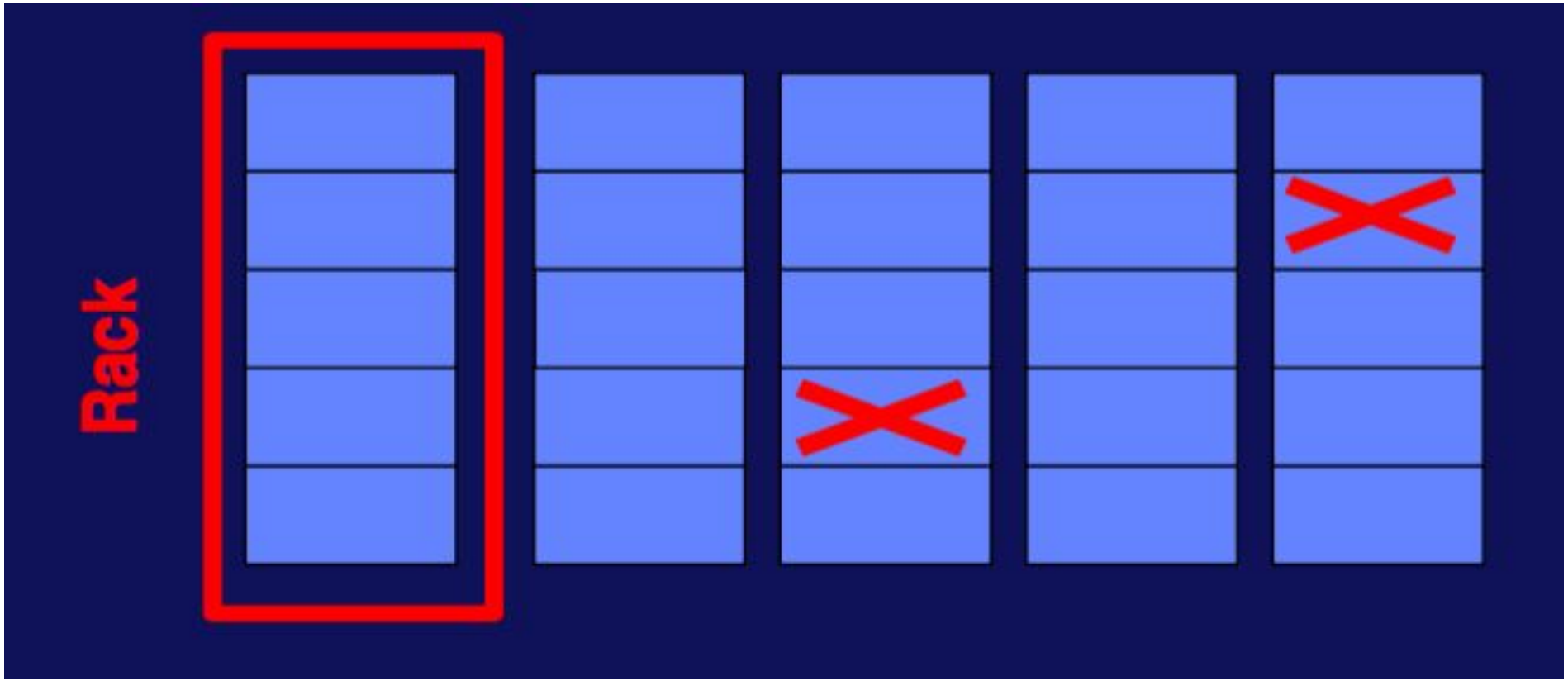
OBJETIVOS

1. Escalabilidad



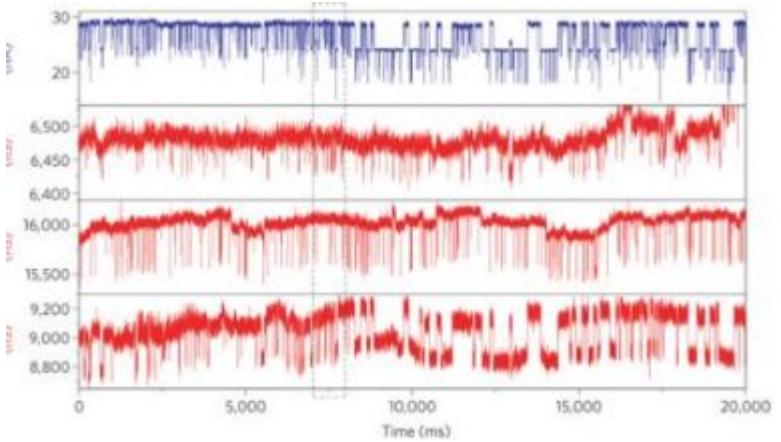
OBJETIVOS

2. Tolerancia a Fallas



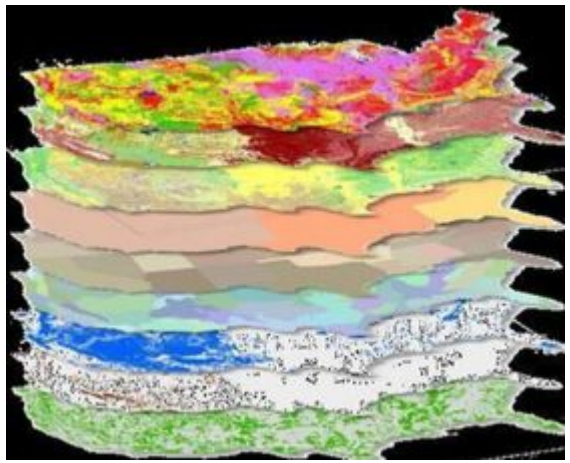
OBJETIVOS

3. Optimizado para distintos Tipos de Datos



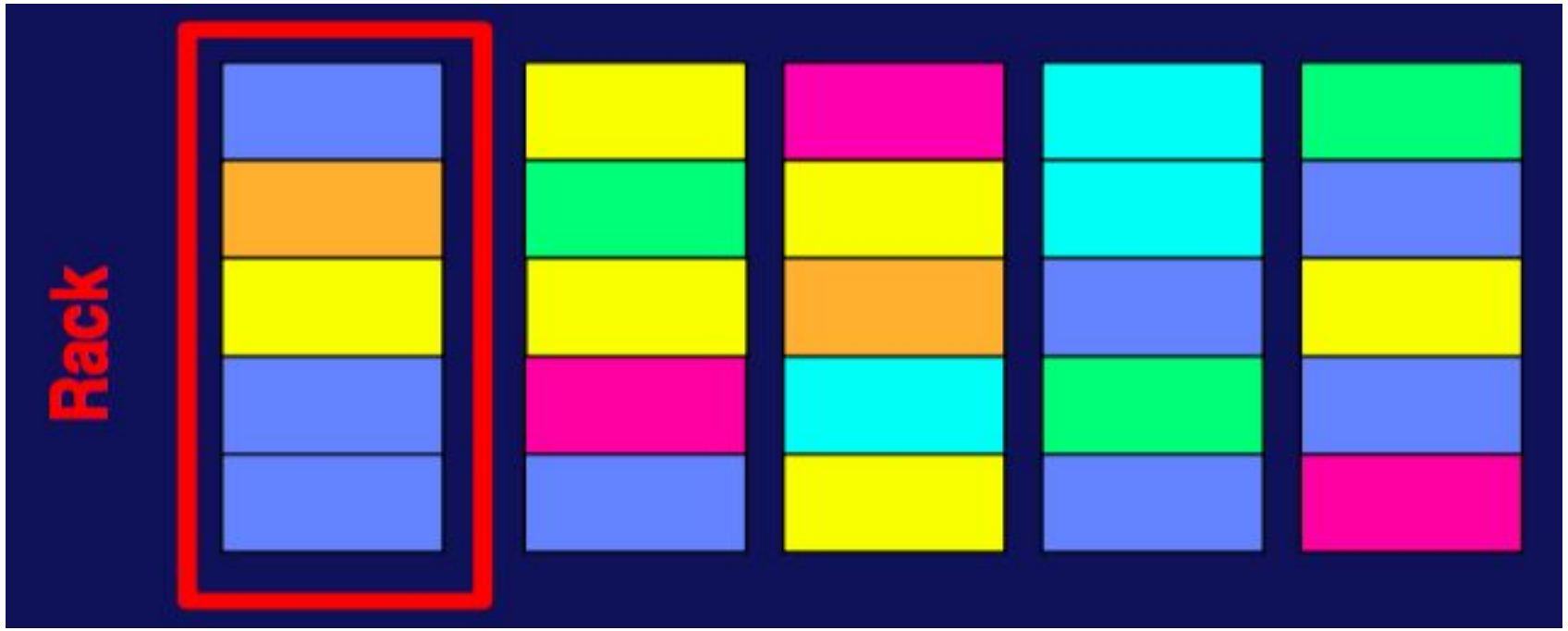
Cars marketplace

vendor	Model	Price	Mileage	VIN Code
Chevrolet	Corvette	17226	25965.0	ILLAKAWAZDZ
Chevrolet	Corvette	34229	46429.0	RCPNSRYGXO
Chevrolet	Corvette	27982	50209.0	NWLGCEVEHG
Chevrolet	Corvette	51825	72998.0	NGVZSCIZGSA
Chevrolet	Corvette	52845	34364.0	PSDRUYYOUG
Chevrolet	Malibu	37874	37273.0	VLFPQPWNEFC
Chevrolet	Malibu	15600	71441.0	EXLJGDWOZSA
Chevrolet	Malibu	52447	46700.0	NLMGJZAKBPC
Chevrolet	Malibu	27129	36254.0	OIPFUEI&LEHSX
Chevrolet	Malibu	28846	77162.0	WRCOOFREZLI
Chevrolet	Malibu	46165	60590.0	HUFTTHQHSFJI
Chevrolet	Malibu	18263	37790.0	3 M9NAFSHVT



OBJETIVOS

4. Facilita un Entorno Compartido



OBJETIVOS

5. Provee Valor a la Empresa

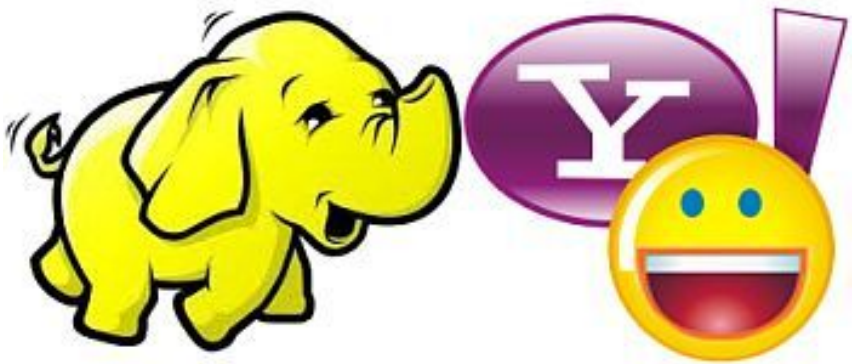
- Comunidad de desarrolladores (http://hadoop.apache.org/)

- Variedad de aplicaciones

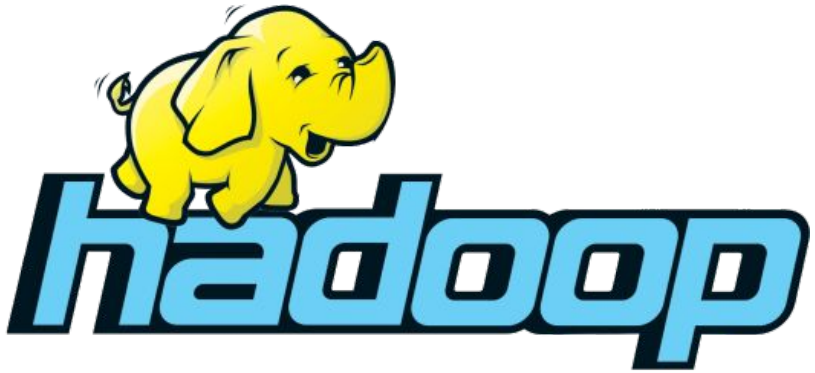
2004

The Google MapReduce

2005



2005+



Hoy en día existen más de 100 proyectos open-source sobre Hadoop

2005+

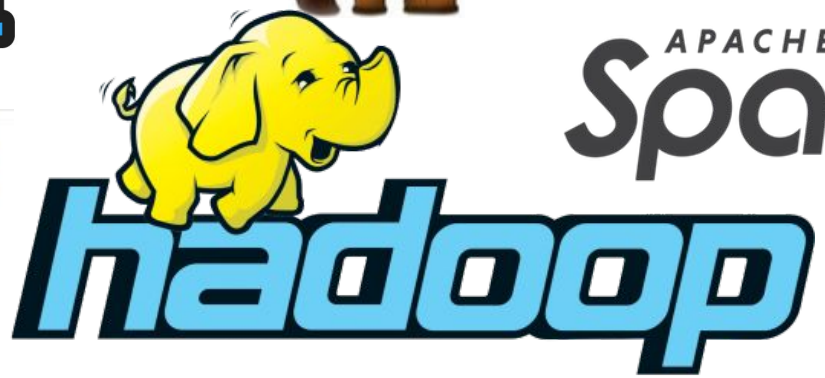
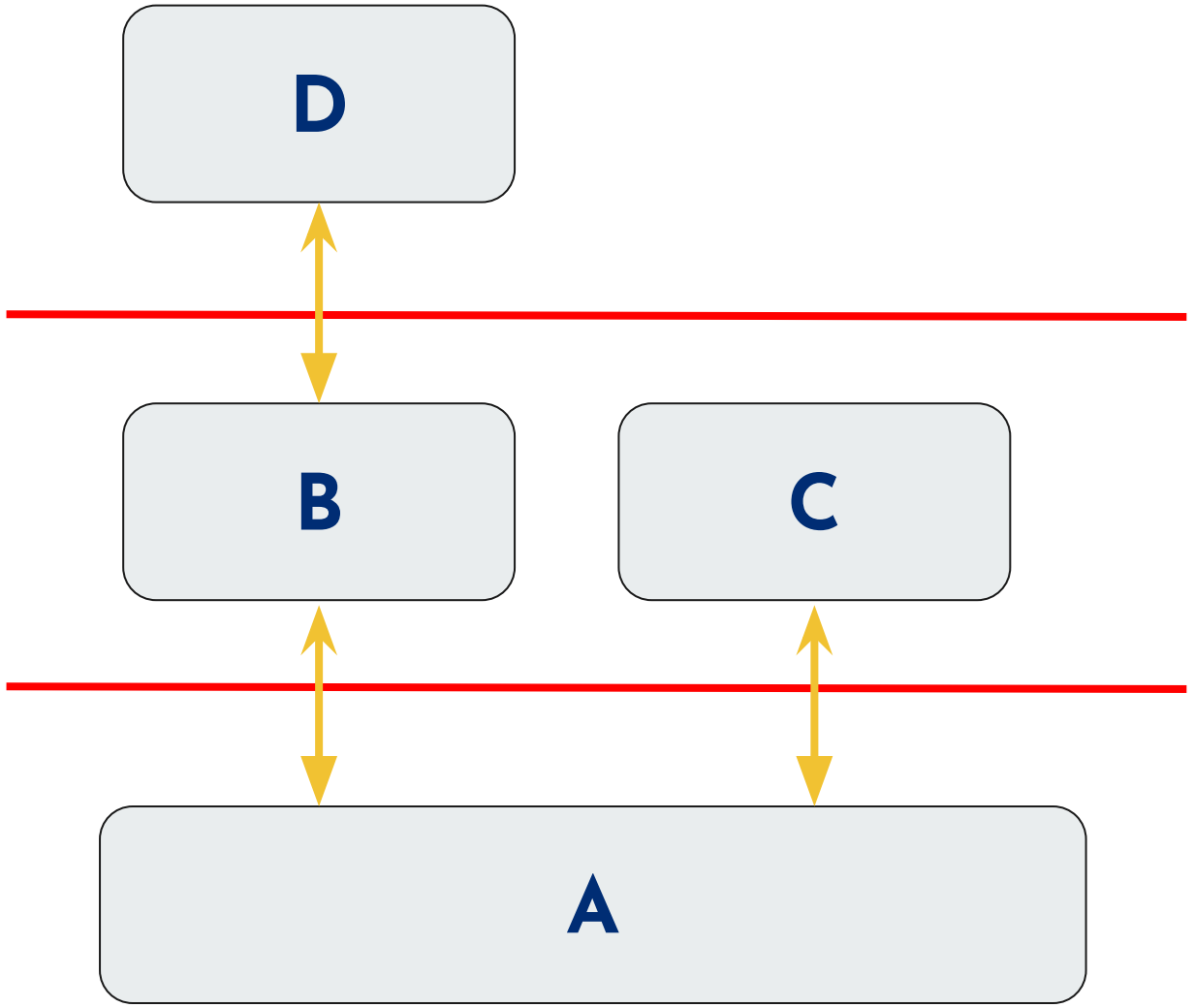
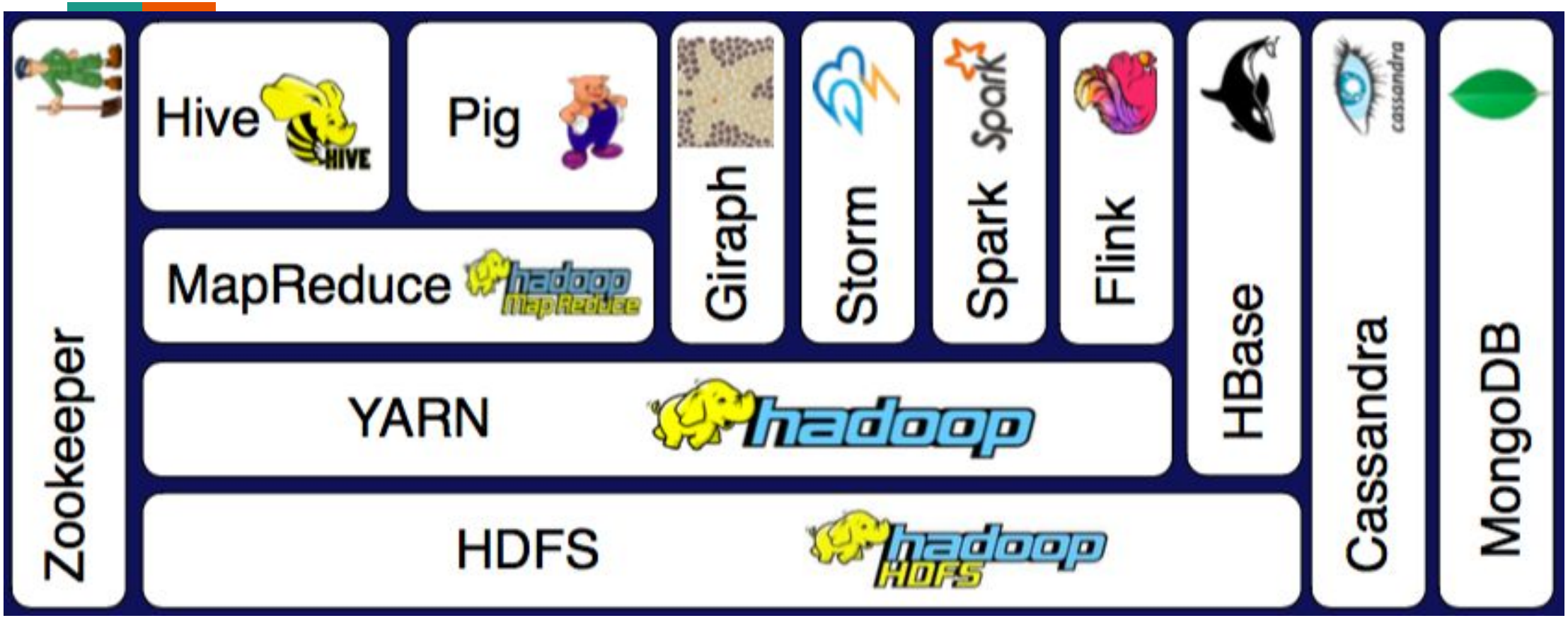
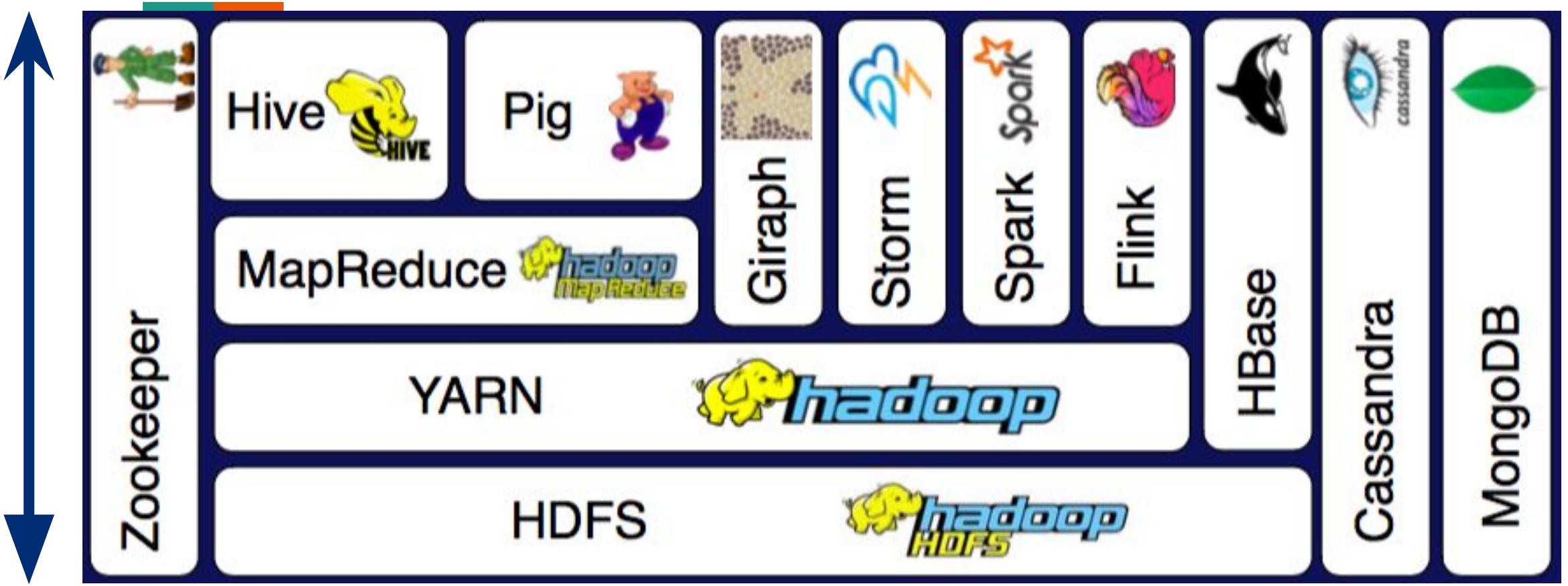


Diagrama de Capas




Zookeeper
Hive
HIVE
Pig
MapReduce
hadoop
MapReduce
YARN
hadoop
HDFS
hadoop
HDFS
Giraph
Storm
Spark
Spark
Flink
HBase
Cassandra
cassandra
MongoDB

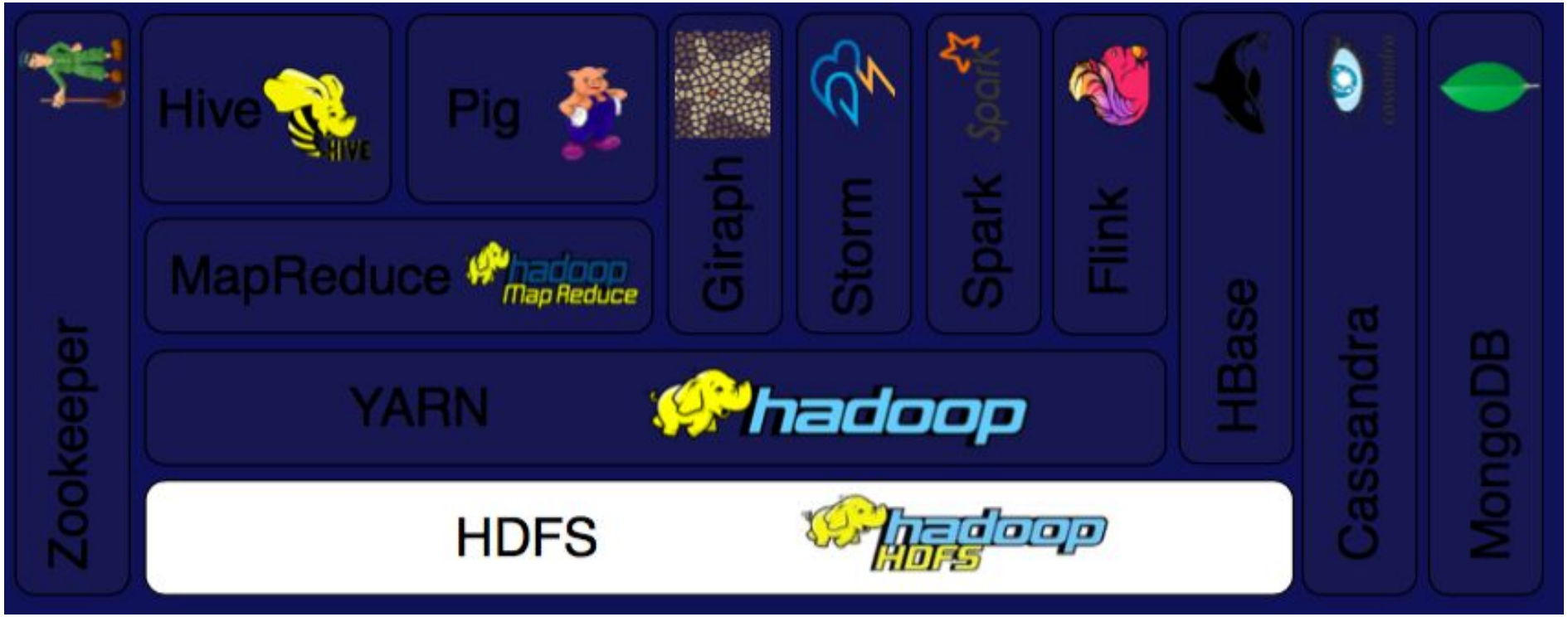
High Level: Interactivity



Low Level: Almacenamiento

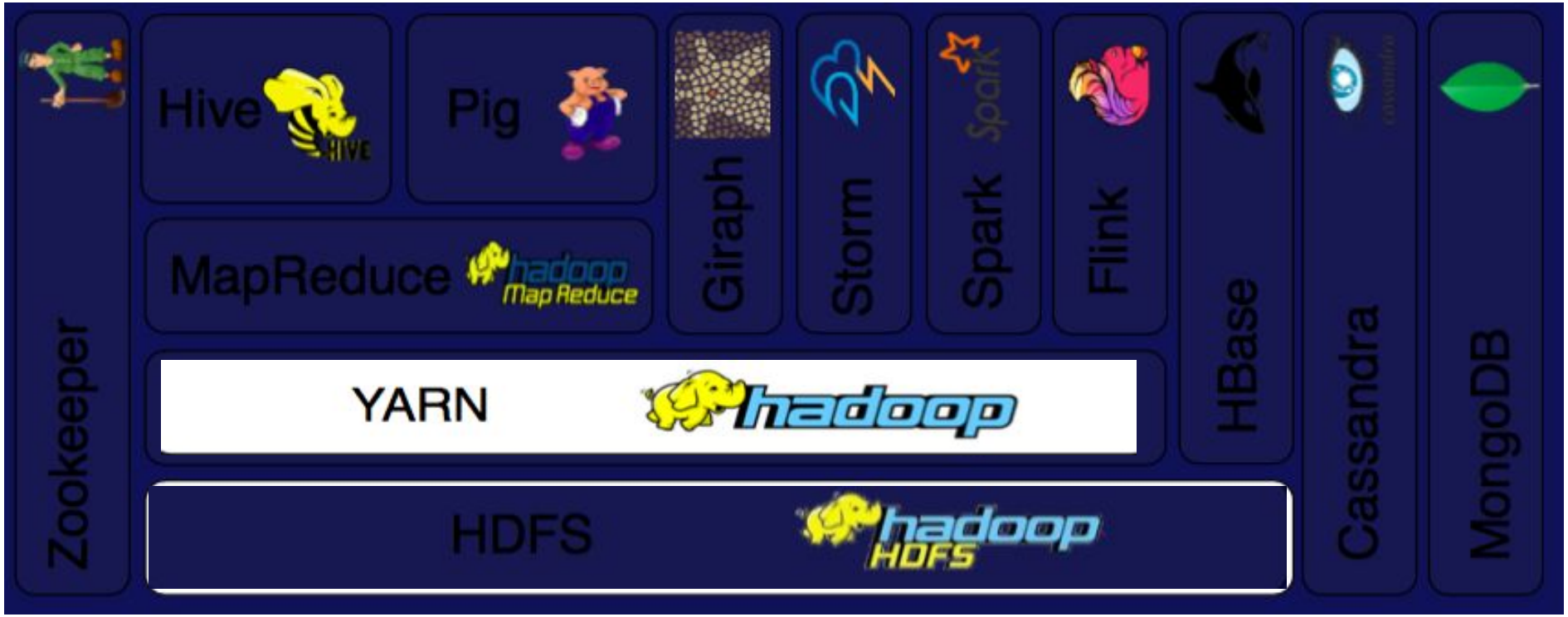
Almacenamiento Escalable

Tolerancia a Fallas



Administrador de los recursos

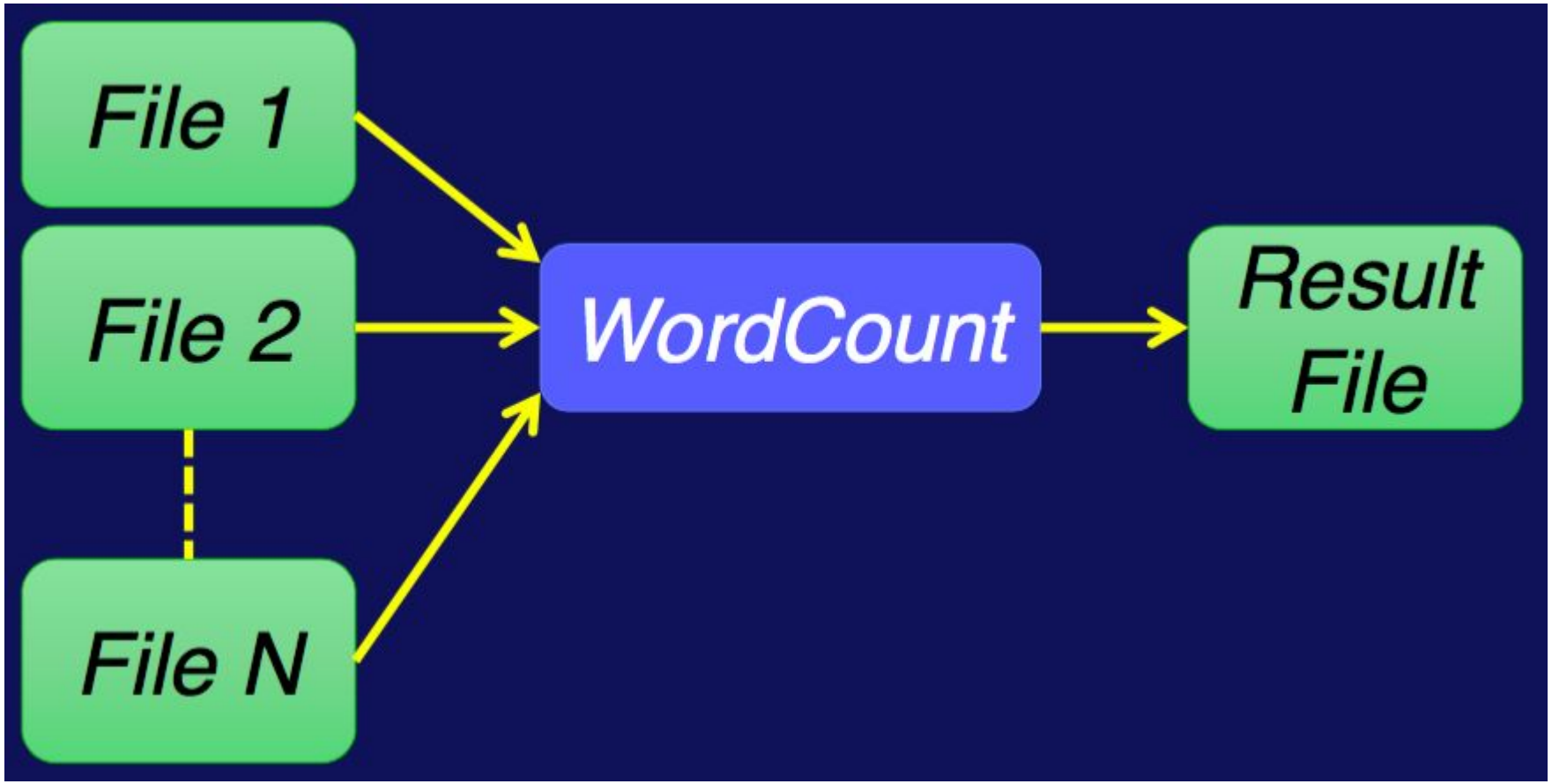
YARN programa jobs en +40 000 servidores en Yahoo



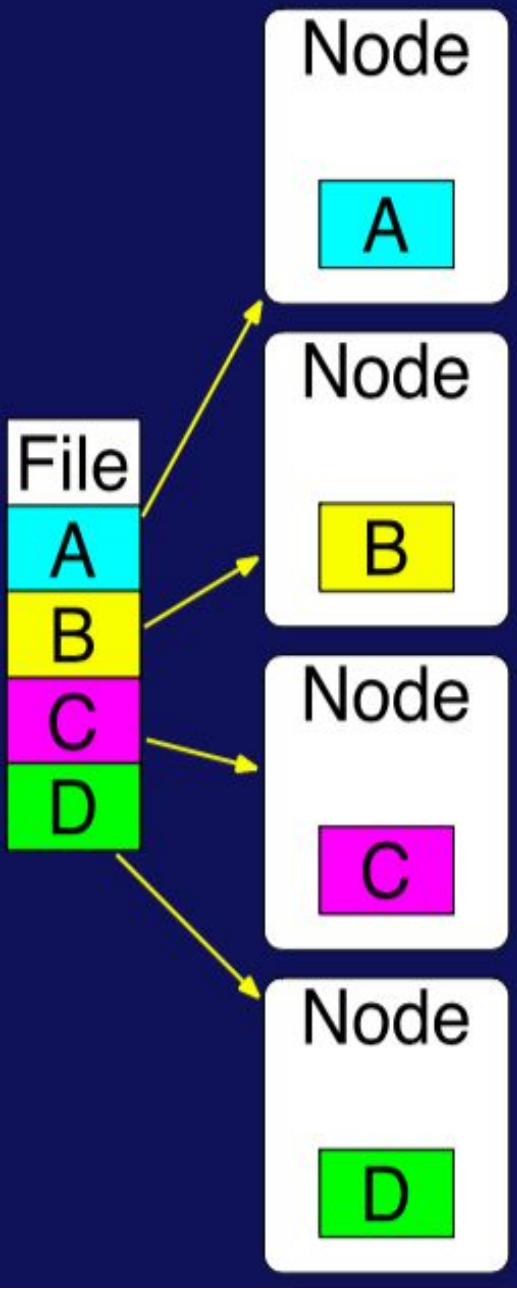
Modelo de programación simplificado

Map -> apply()
Reduce -> summarize()

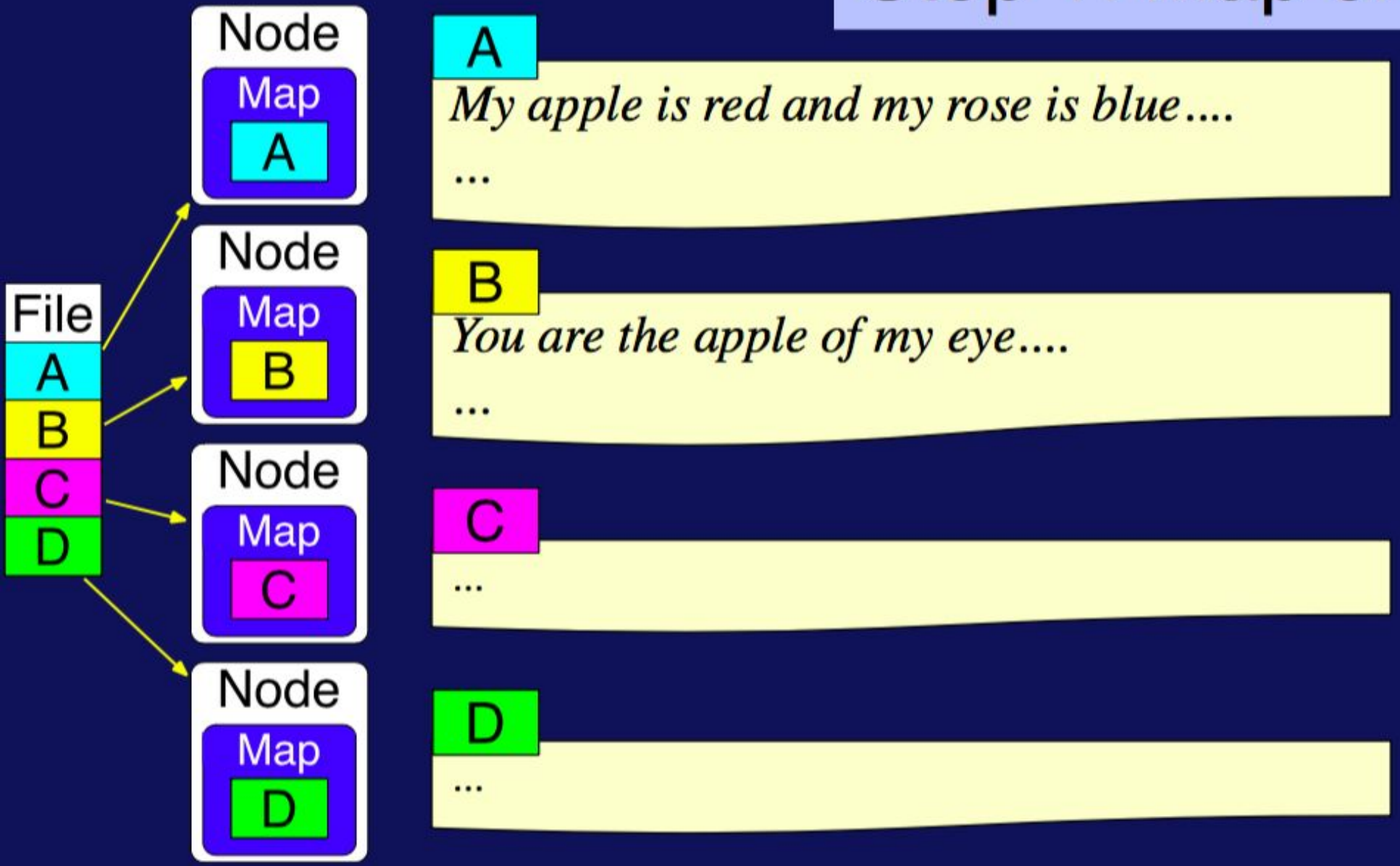
Google lo utiliza para indexar websites

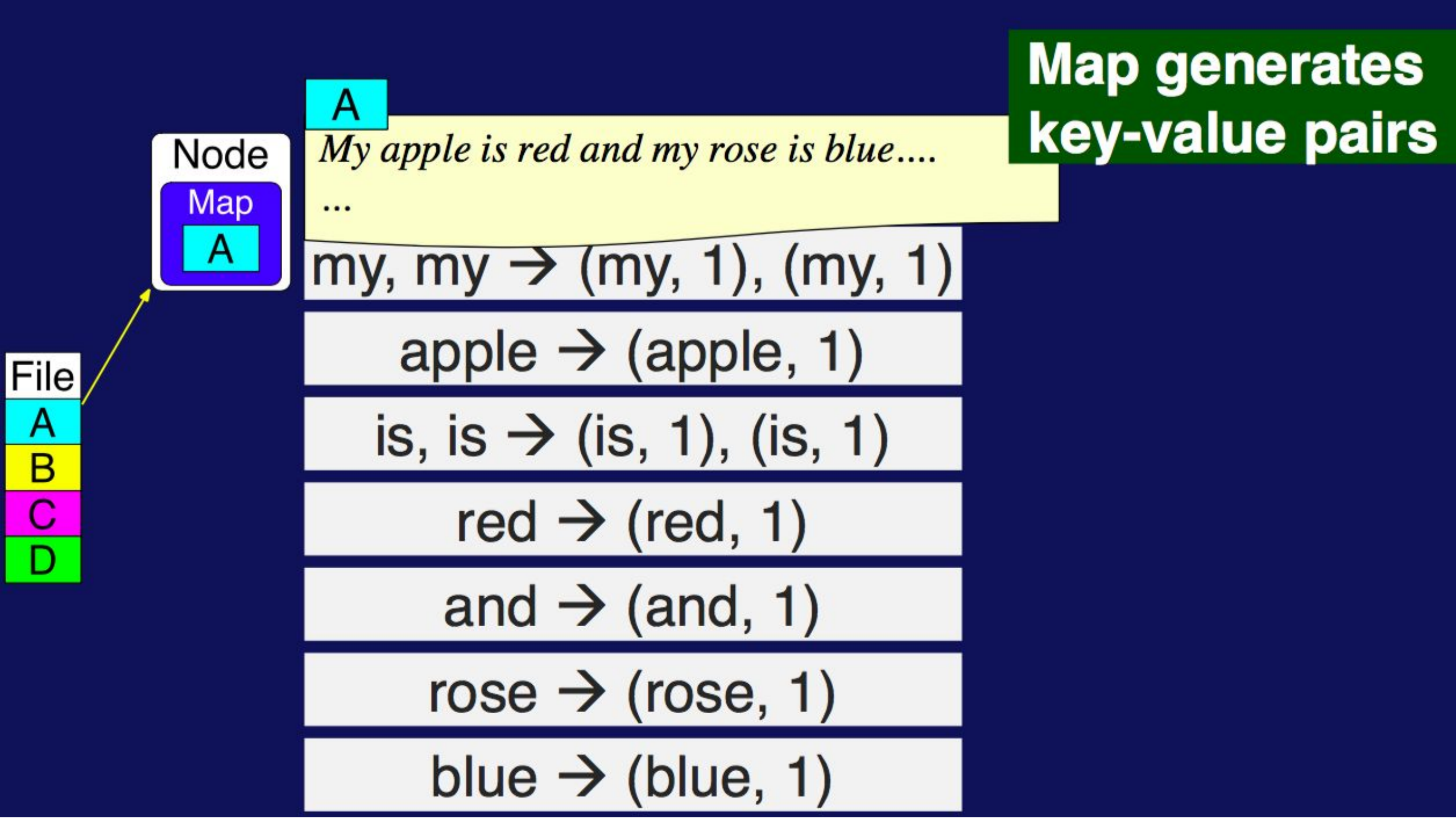

File 1
File 2
File N
WordCount
Result File

Step 0: File is stored in HDFS

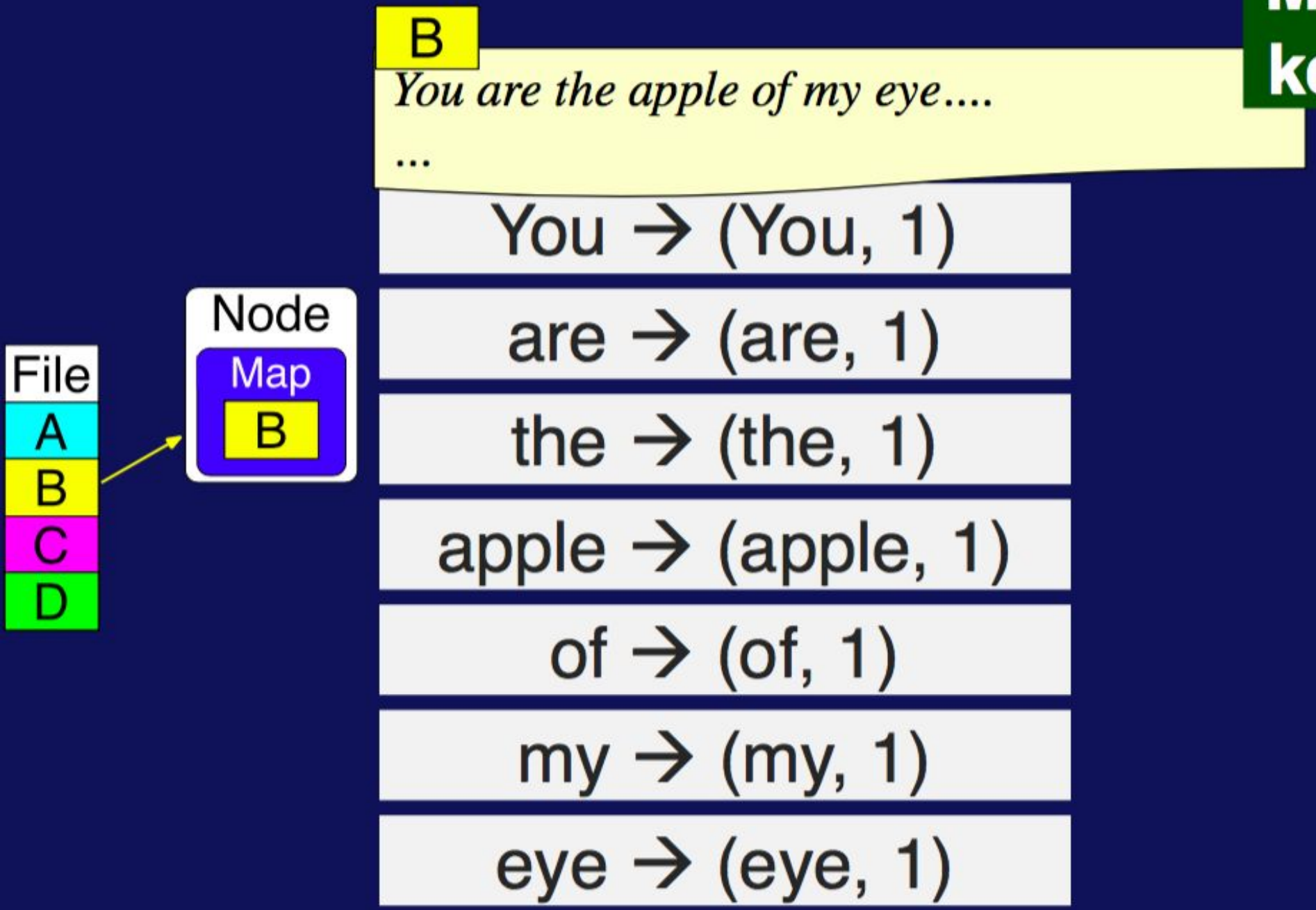


Step 1: Map on each node



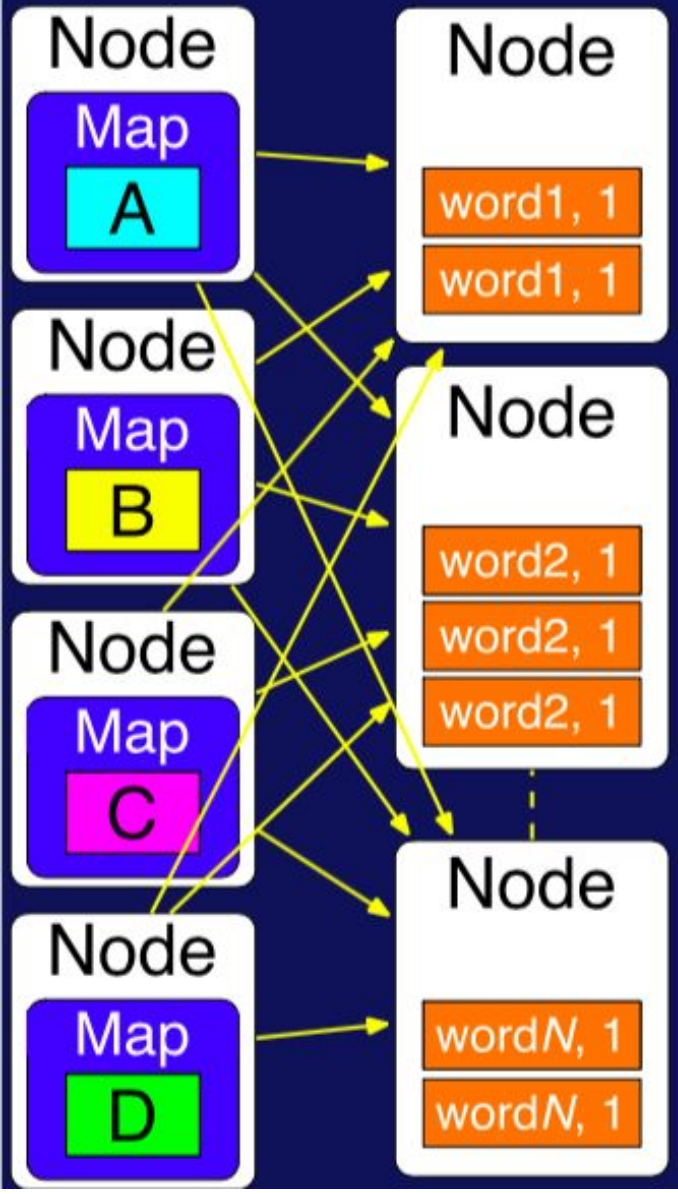

Map generates key-value pairs
Node
Map
A
File
A
B
C
D
A
My apple is red and my rose is blue....
...
my, my → (my, 1), (my, 1)
apple → (apple, 1)
is, is → (is, 1), (is, 1)
red → (red, 1)
and → (and, 1)
rose → (rose, 1)
blue → (blue, 1)

Map generates key-value pairs



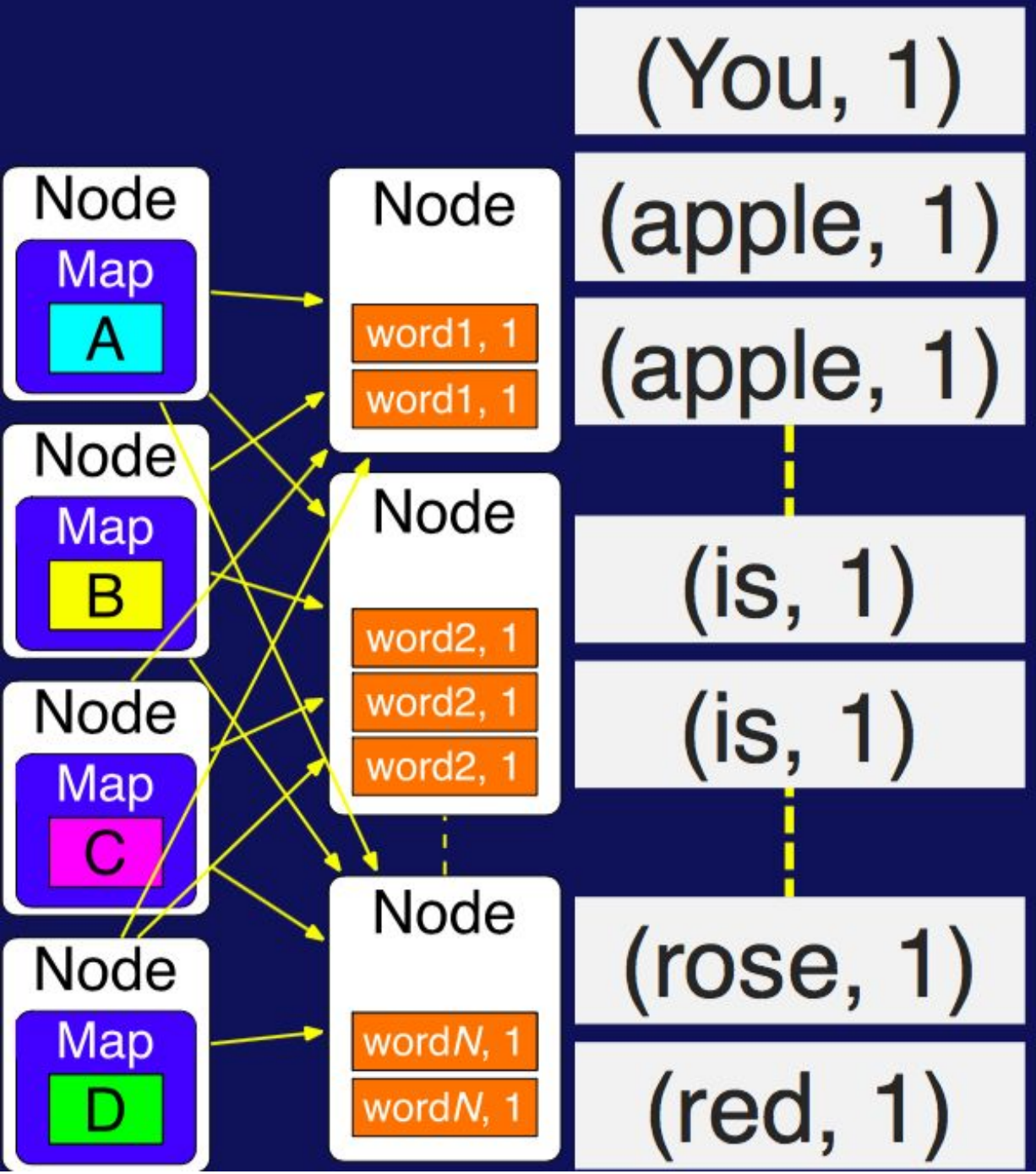
Step 2: Sort and Shuffle

Pairs with same key moved to same node



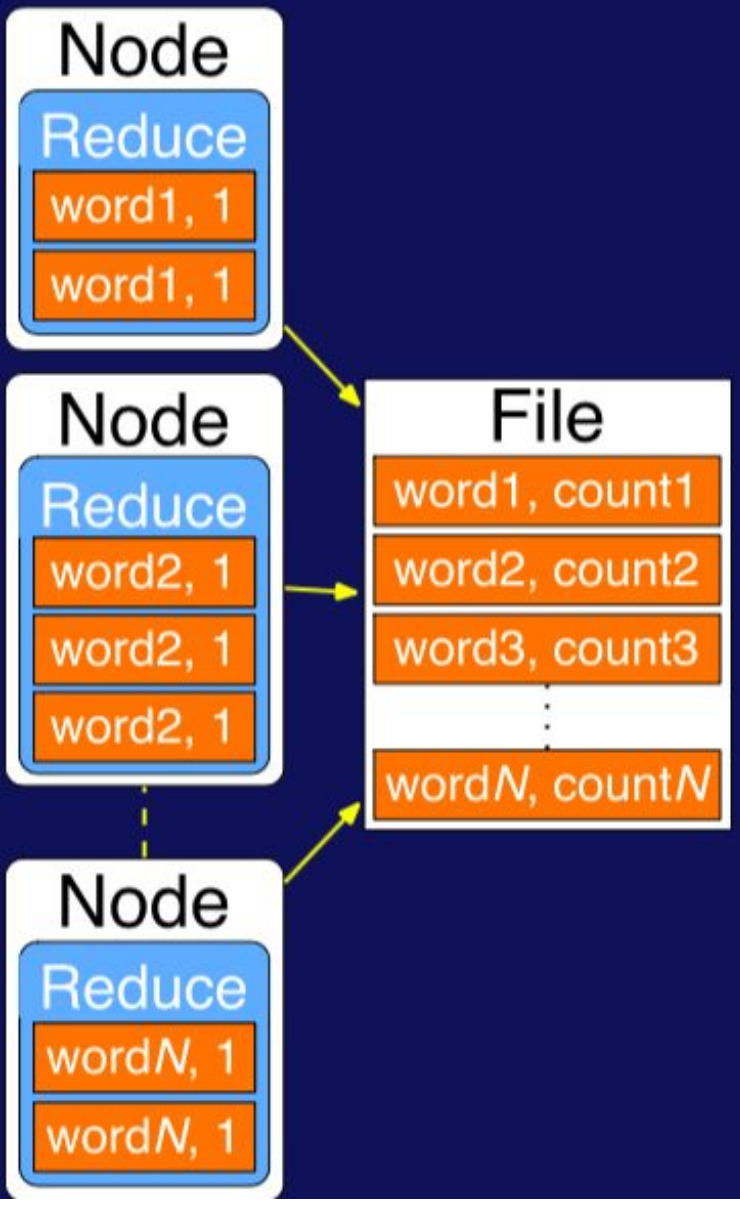
Step 2: Sort and Shuffle

Pairs with same key moved to same node



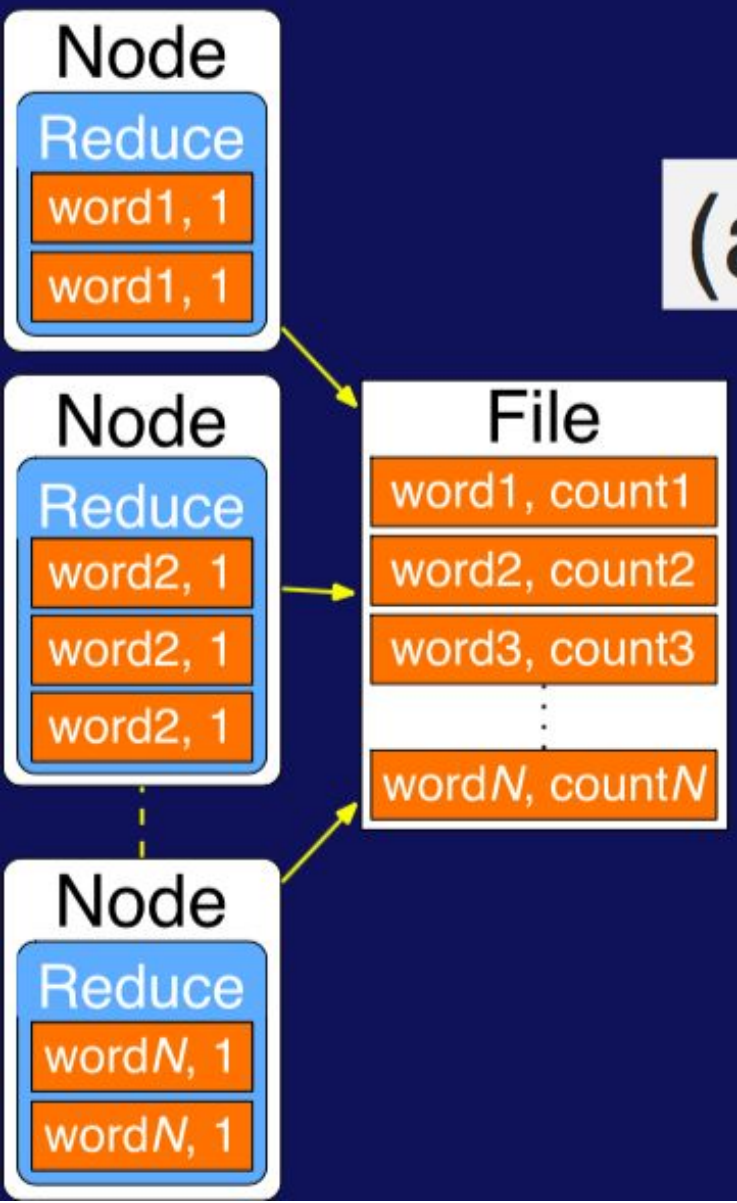
Step 3: Reduce

Add values for same keys



Step 3: Reduce

Add values for same keys



(You, 1) → (You, 1)

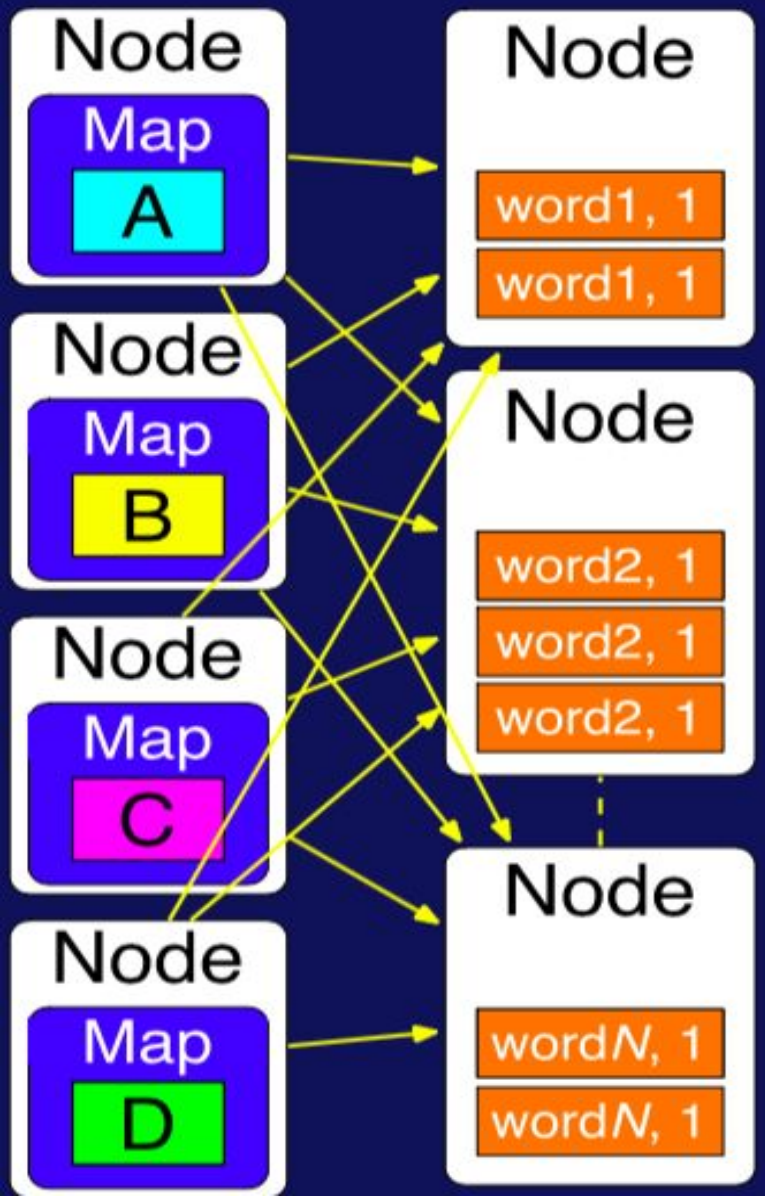
(apple, 1), (apple, 1) → (apple, 2)

(my, 1), (my, 1), (my, 1) → (my, 3)

(red, 1) → (red, 1)

(rose, 1) → (rose, 1)

Sort and Shuffle



(You, http://you1.fake)

(apple, http://apple1.fake)

(apple, http://apple2.fake)

(is, http://apple2.fake)

(is, http://apple2.fake)

(rose, http://apple2.fake)

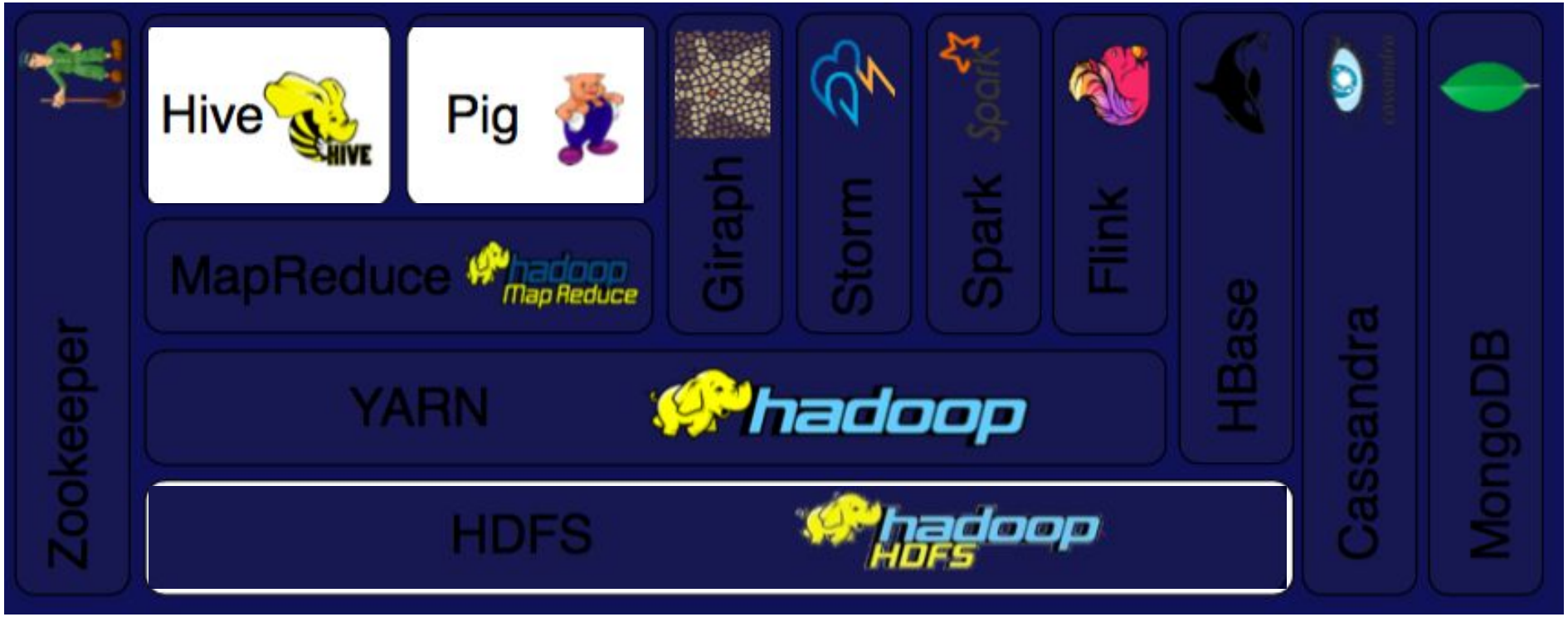
(red, http://apple2.fake)

(apple -> http://apple1.fake,
http://apple2.fake)

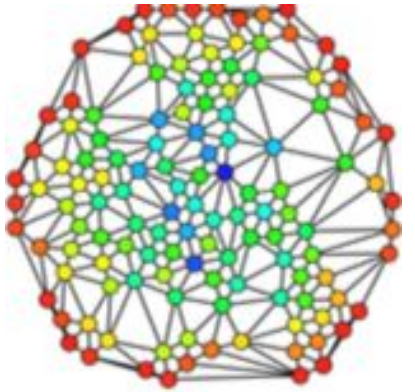
Modelo de programación high-level

Pig = dataflow scripting
Hive = SQL-like queries

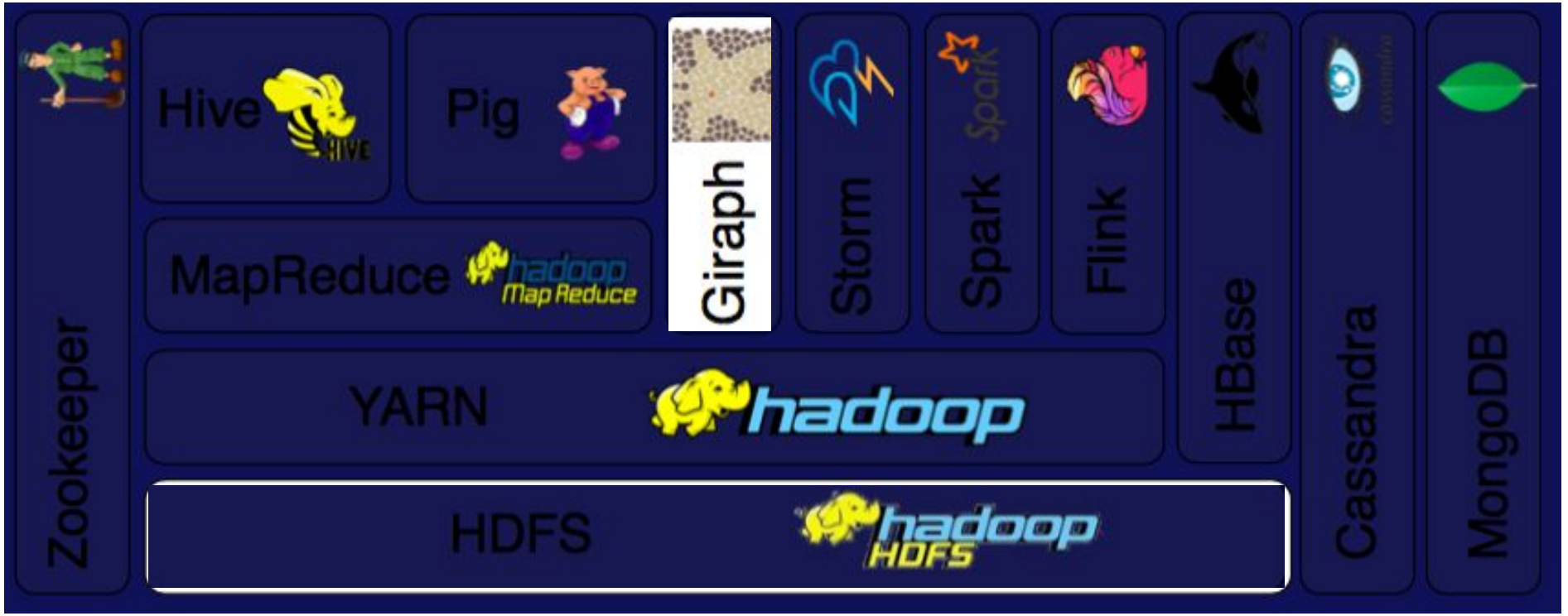
Pig (Yahoo)
Hive (Facebook)



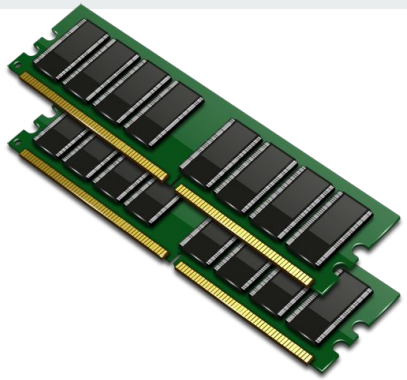
Modelos especializados para procesamiento de grafos



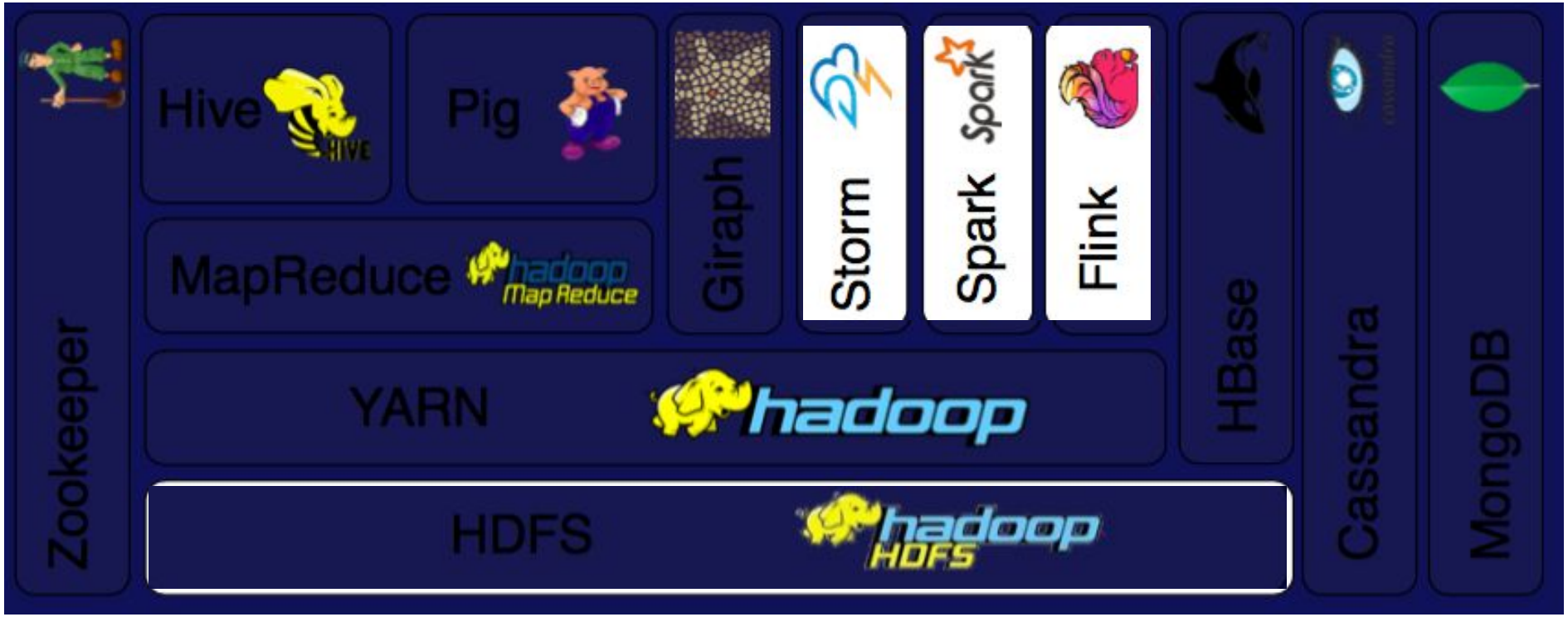
Facebook analiza su grafo usando Giraph



Procesamiento en tiempo real (in-memory)



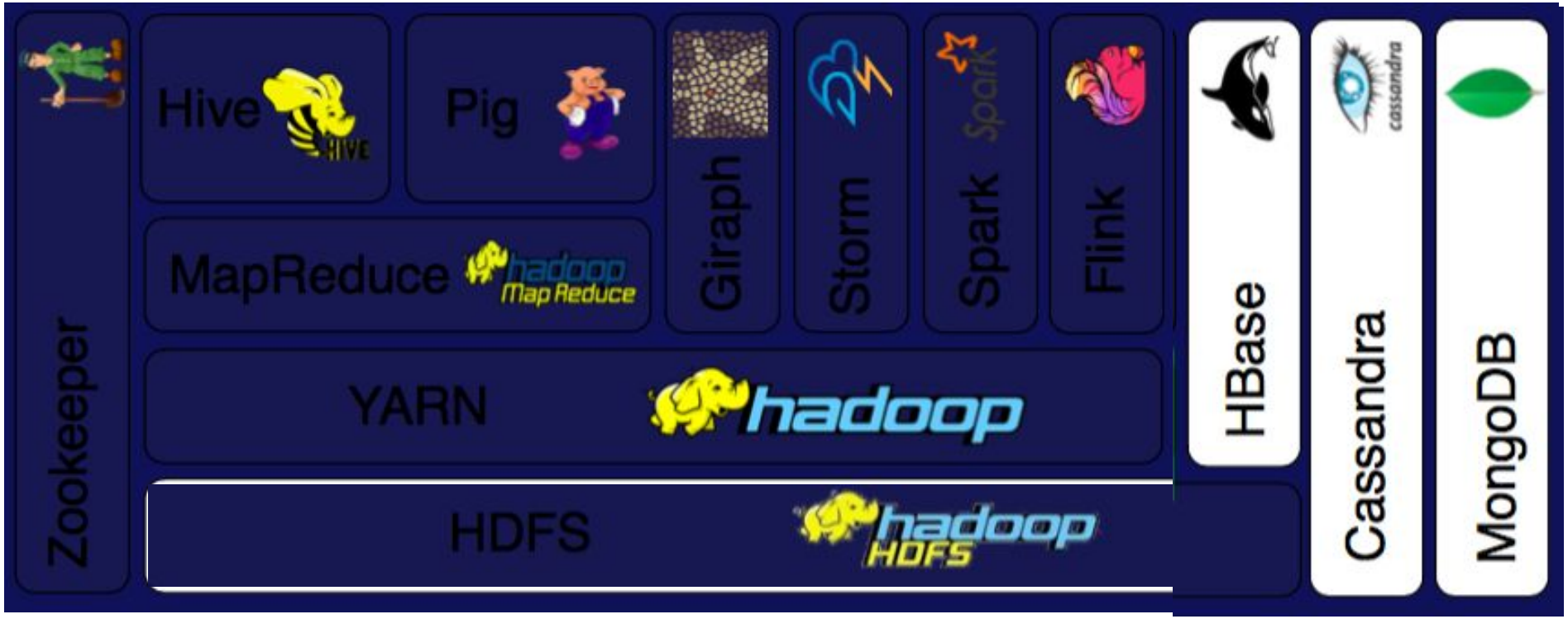
* 100x más rápido en algunas tareas



NoSQL

* Key-Values
* Sparse-tables

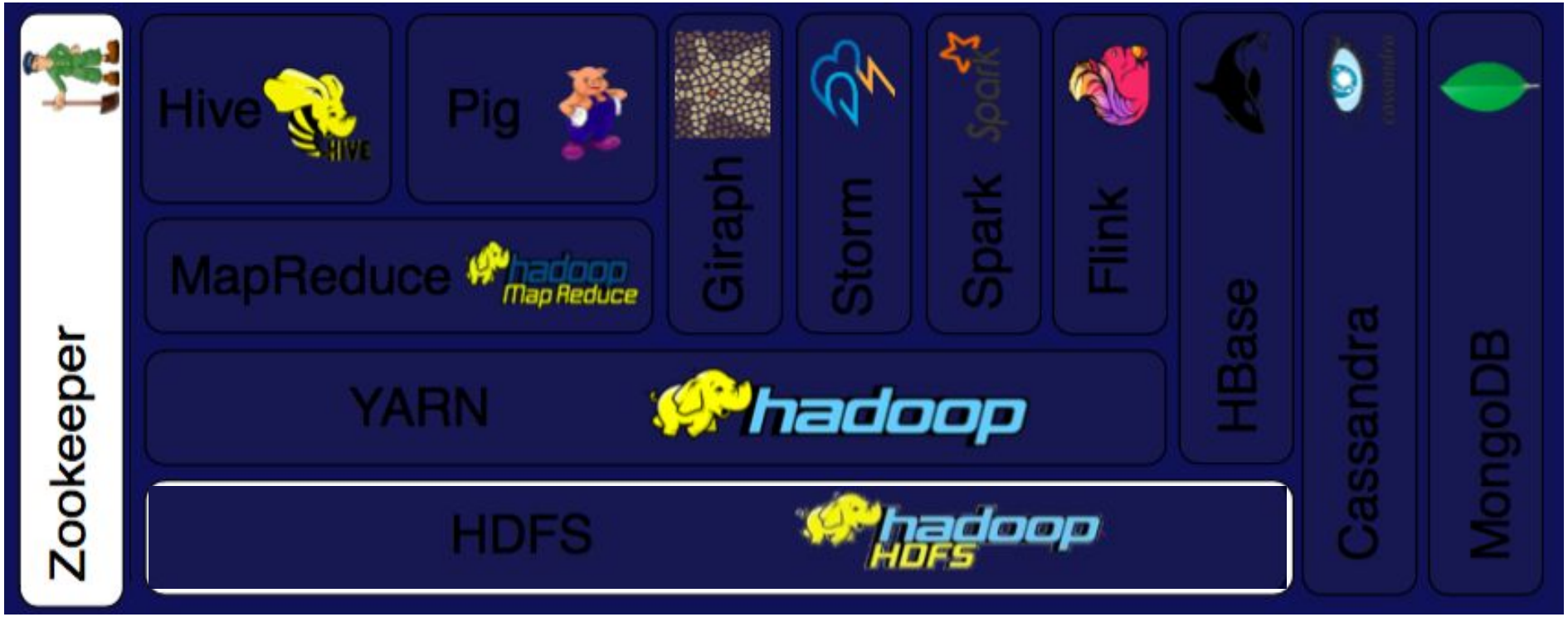
HBase (Facebook Messenger)



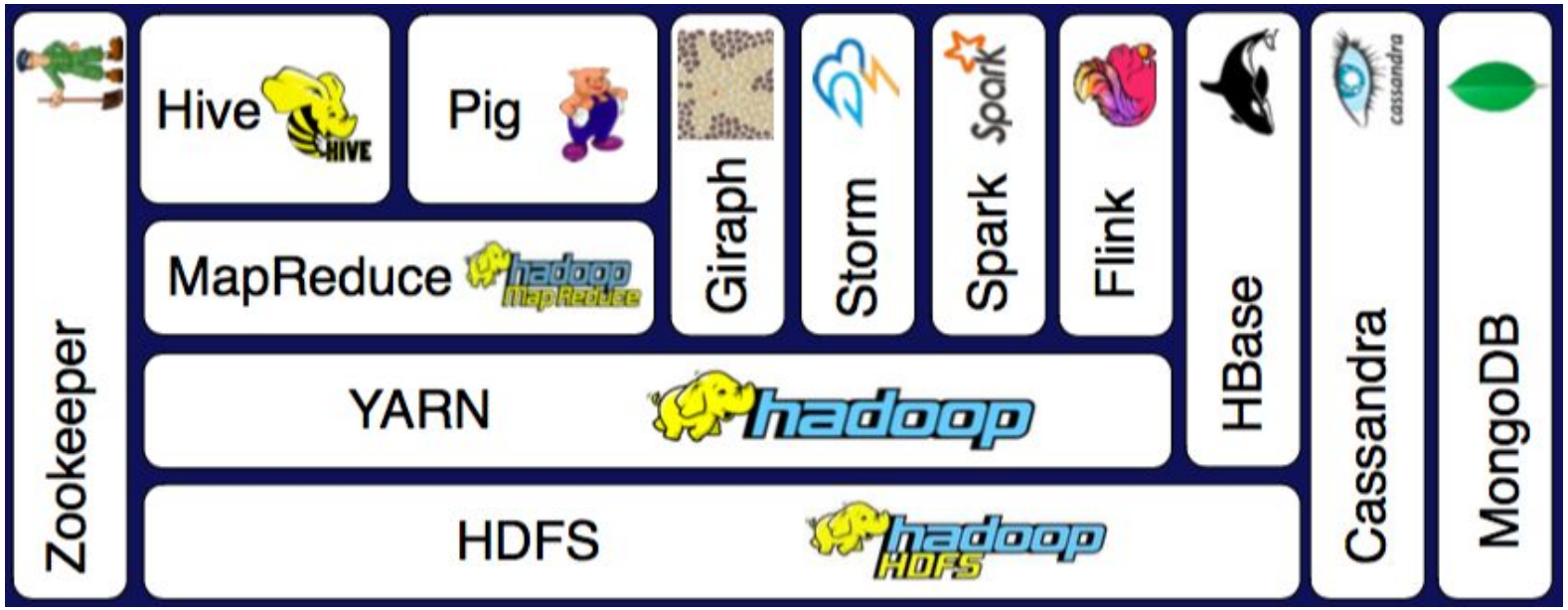
Zookeeper (Administración)

* Sincronización
* Configuración
* Alta disponibilidad

(Facebook)



- Open-source
- Enorme comunidad para soporte
- Independientes para descarga



Dos puntos importantes para reconsiderar utilizar Hadoop:

- Base de datos pequeñas
- Nivel de Paralelismo