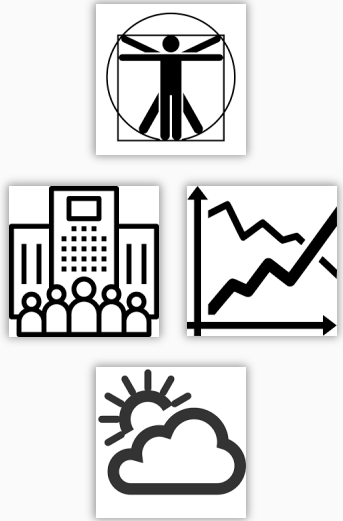
Big Data 101
Overview

Hoy

~26 B Dispositivos
~47 Zb de Datos
16% Estructurados



Insights

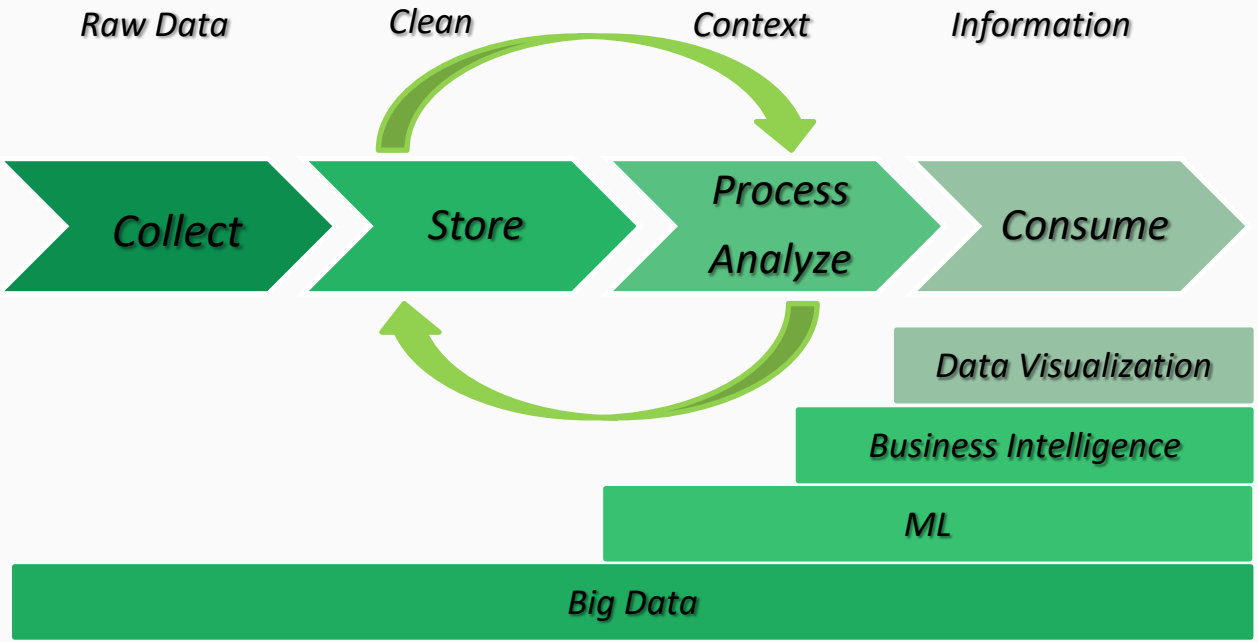
- *Mejorar Calidad de Vida*
- *Diagnosticar & Detectar*
- *Optimizar Procesos*
- *Nuevos Productos & Servicios*
- *Mejorar el Mercado Objetivo*
- *Innovar a Bajo Costo*

Monetizar Datos
(Obtener Valor)

Data Value Path

Raw Data -> *Clean* + *Context* = *Information* + *Knowledge* = *Value*

Data Value Path



+ Knowledge = Value

Challenges

MB<GB<TB
TB<PB<EB
ZB<YB

- Store
- Cost
- Filter
- Process

VOLUME

Estructurados
Semi Estructurados
No Estructurados

- Store (RF)
- Time/Cost
- Filter (Vol)
- Integrate

VARIETY

Batch – Periodic
Near Real Time(min)
Streaming (ms)

- Store
- Process
- Time to market

VELOCITY

2001 – Doug Laney (Gartner)

Challenge

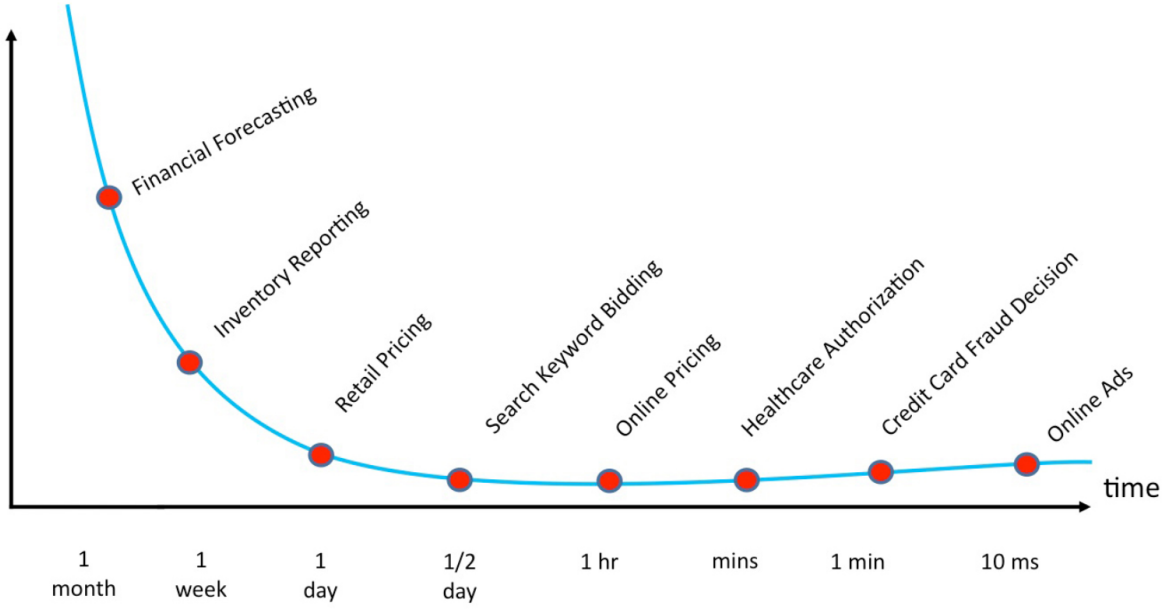


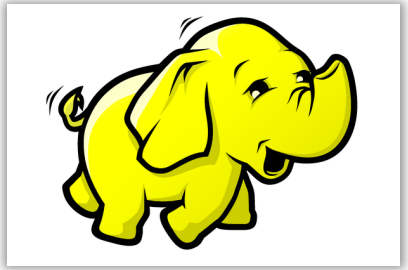
Figure 3. Graphic showing average time-to-decision across various industries. Image courtesy of Evangelos Simoudis.

Definiendo Big Data

Data Sets tan grandes y complejos que se necesitan nuevas formas de ser procesados para obtener valor.

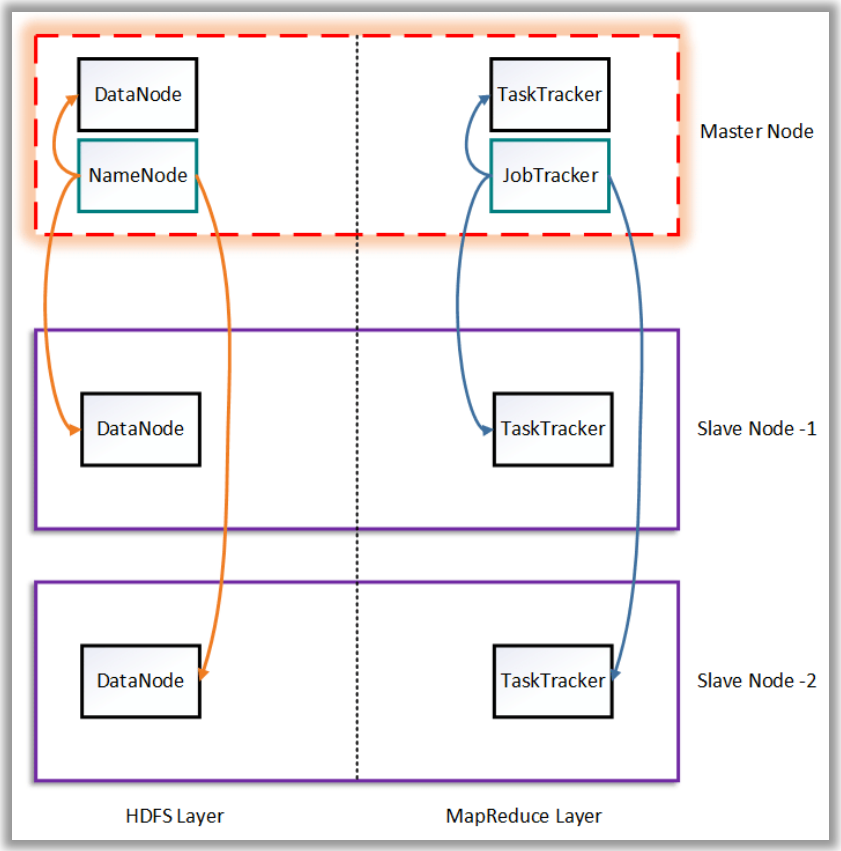
Hadoop

Doug Cutting – Mike Cafarella
2003 – Google File System
2006 - Hadoop

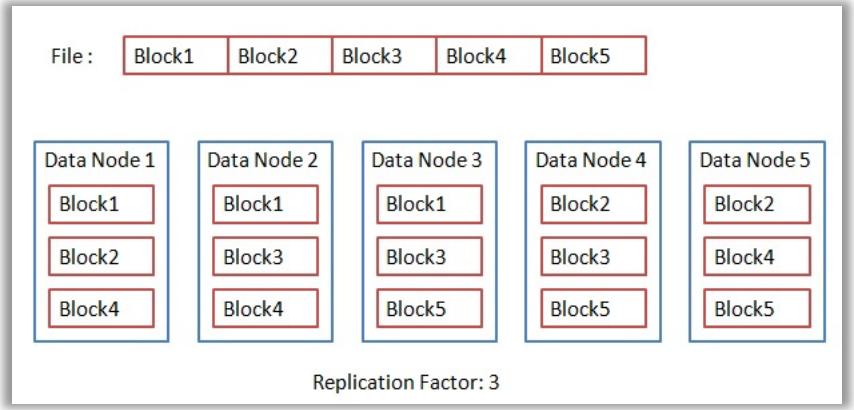


- *Framework que permite el amacenamiento y procesamiento distribuido de grandes datasets.*
- *Construido en Java*
- *Facilidad de escalar vertical y horizontalmente. 1 a 1000>*
- *Utiliza commodity hardware (low cost).*
- *Basado en la premisa todo el hardware puede fallar, alta disponibilidad a nivel de capa de aplicaciones.*
- *Arquitectura Shared Nothing / Master-Slaves*
- *Datos Redundantes.*
- *Puede almacenar todo tipo de datos y tamaño. (Bloques, compresión).*
- *Write Once, Read Many*
- *Schema on Read*

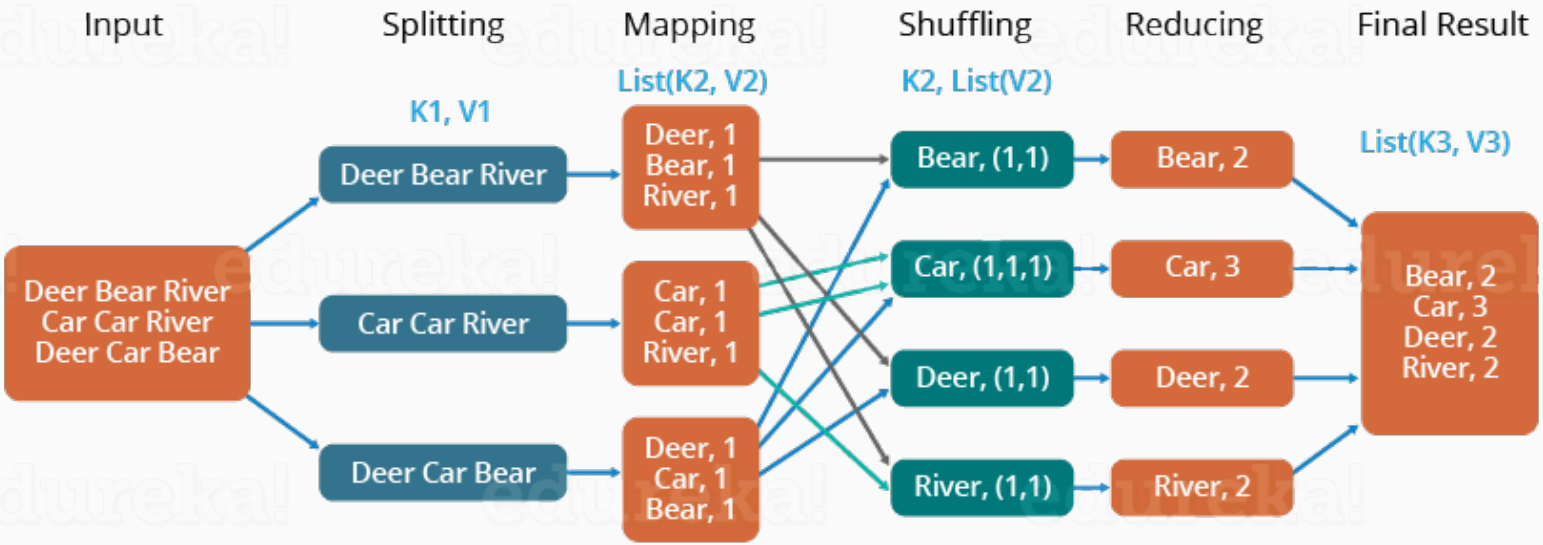
HADOOP – Arch & Layers



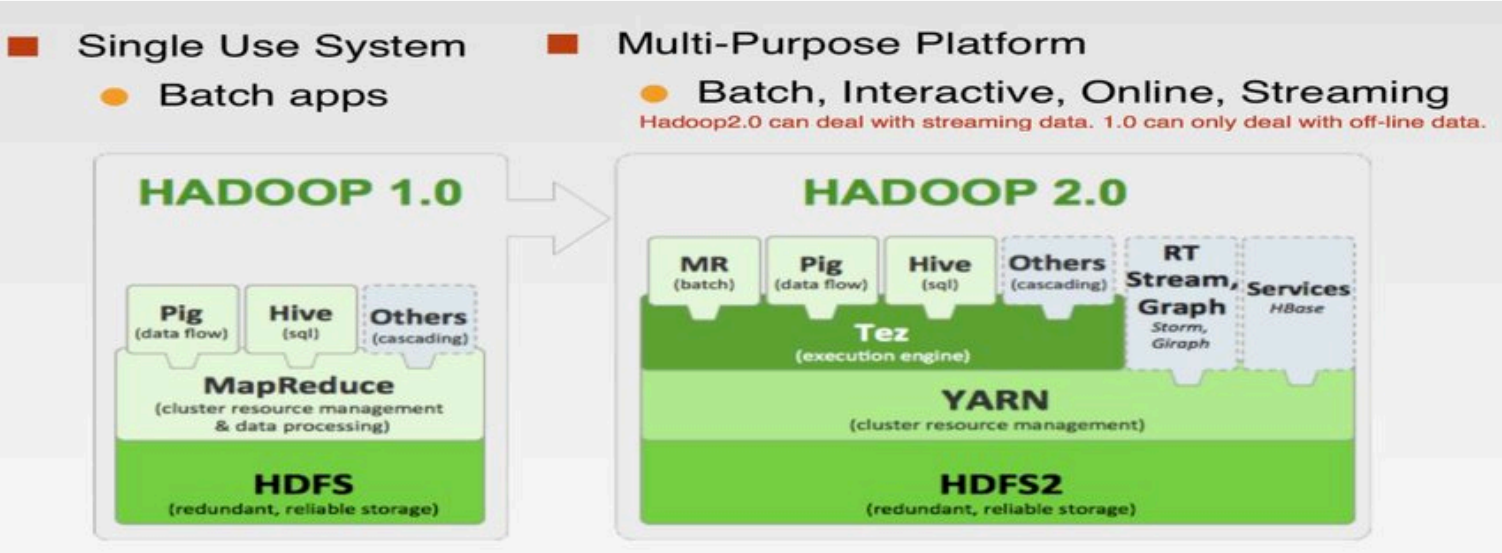
Hadoop – Replication



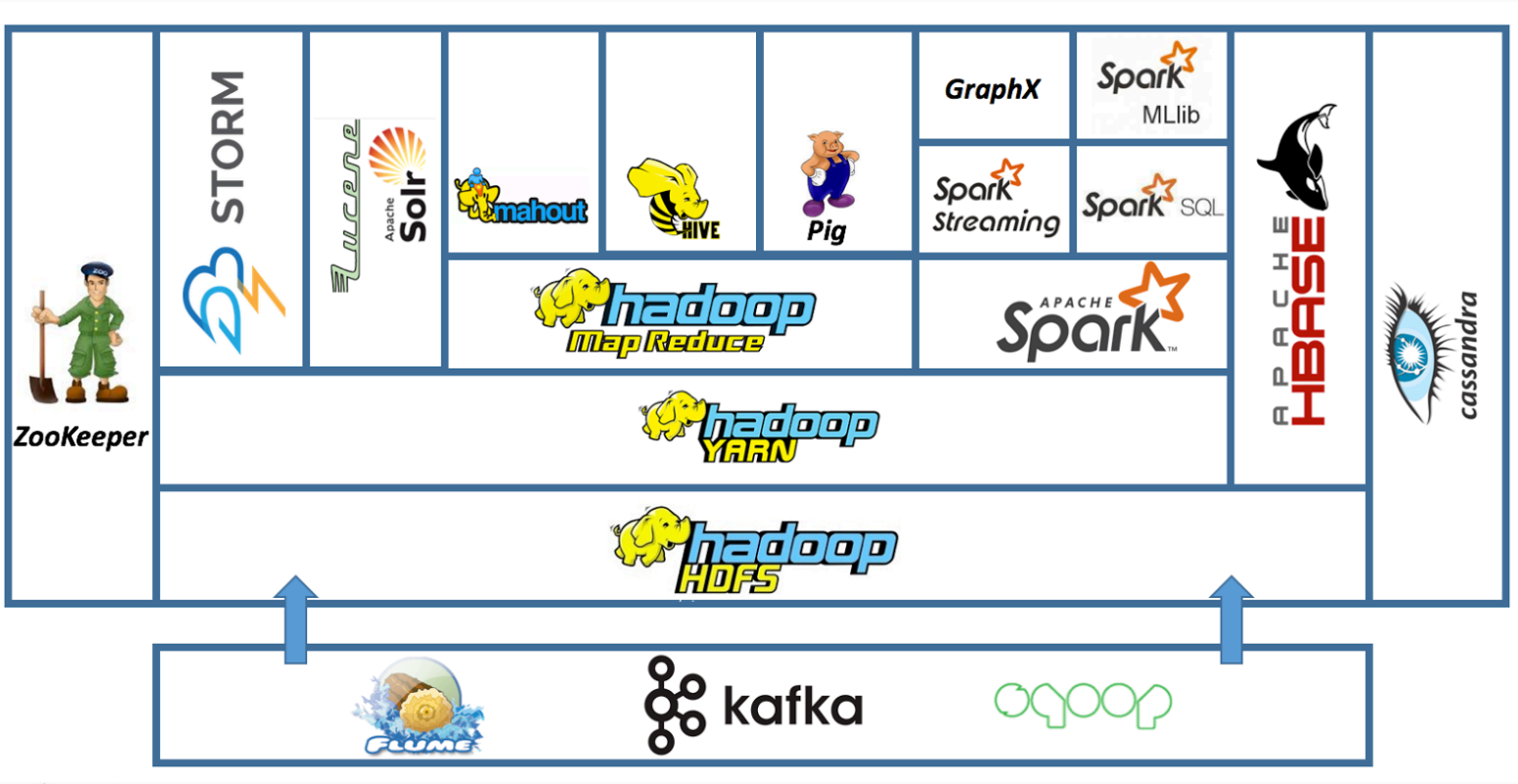
Hadoop – Map Reduce



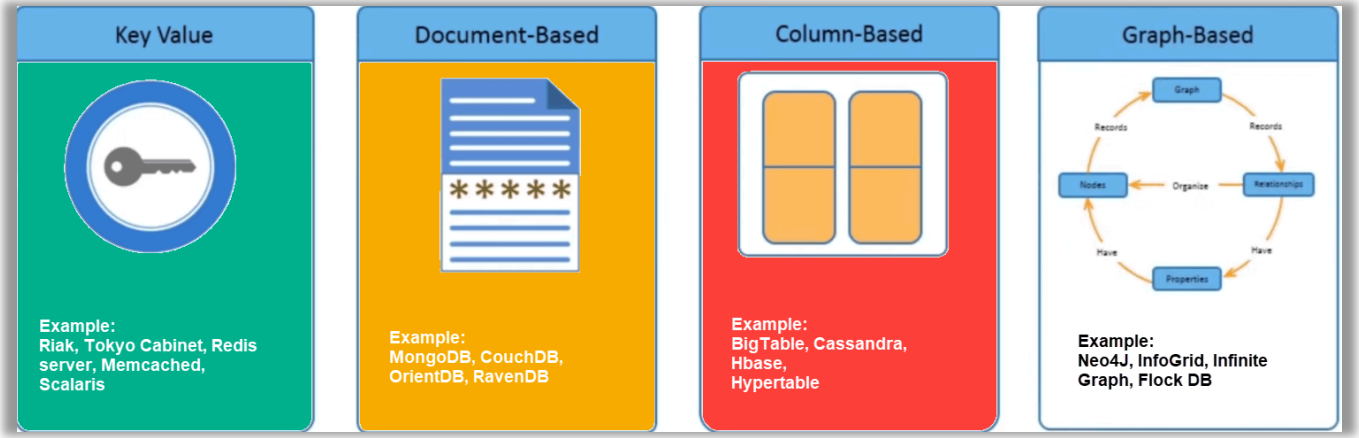
Hadoop - Evolution



Hadoop Framework



No SQL


Key Value
Example:
Riak, Tokyo Cabinet, Redis server, Memcached, Scalaris
Document-Based
*****
Example:
MongoDB, CouchDB, OrientDB, RavenDB
Column-Based
Example:
BigTable, Cassandra, Hbase, Hypertable
Graph-Based
Graph
Records
Records
Nodes
Organize
Relationships
Have
Have
Properties
Example:
Neo4J, InfoGrid, Infinite Graph, Flock DB

Data Lake

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

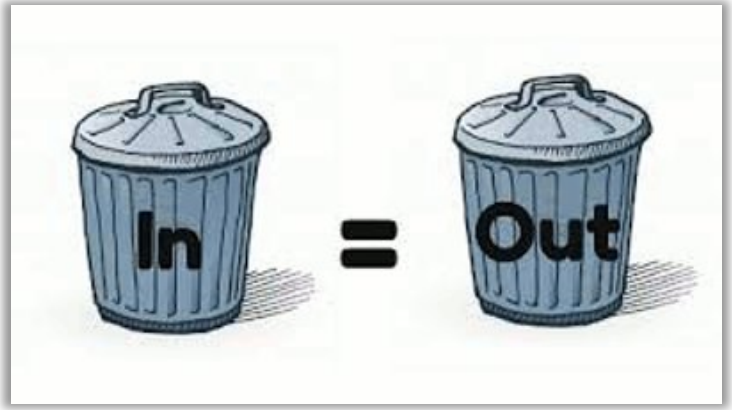
Hints & Tips

1º Why?

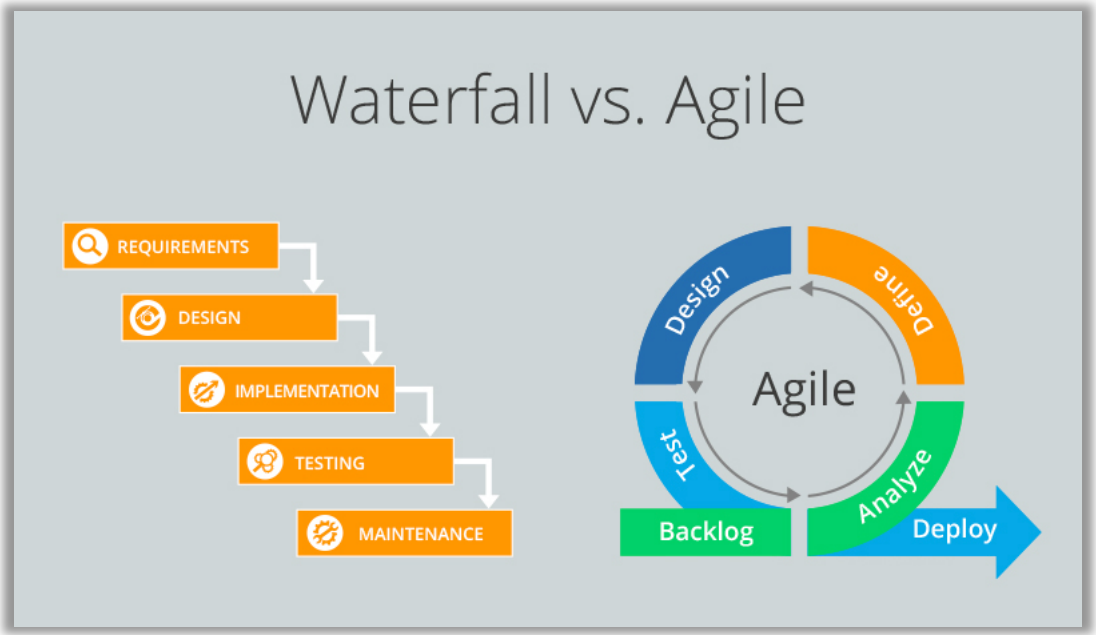
2ºWhat?

Hints & Tips

Hints & Tips



Hints & Tips


Waterfall vs. Agile
REQUIREMENTS
DESIGN
IMPLEMENTATION
TESTING
MAINTENANCE
Agile
Design
Define
Analyze
Deploy
Test
Backlog

Next Steps

idea from Victor Silva

Cursos Introductorios a HADOOP y Big Data on AWS (Free)

Intro to Hadoop and MapReduce by Udacity

Getting Started with Hadoop by Cloudera

Big Data Fundamentals by AWS

Data Analytics Fundamentals by AWS

Iniciar con Big Data en AWS by AWS

Capa gratuitua de AWS

Maximiliano Kretowicz

Linkedin