MACHINE LEARNING & BIG DATA

Conceptos básicos

José Nelson Zepeda

San Salvador, octubre 2018

Fundamentos Machine Learning

Conceptos Básicos

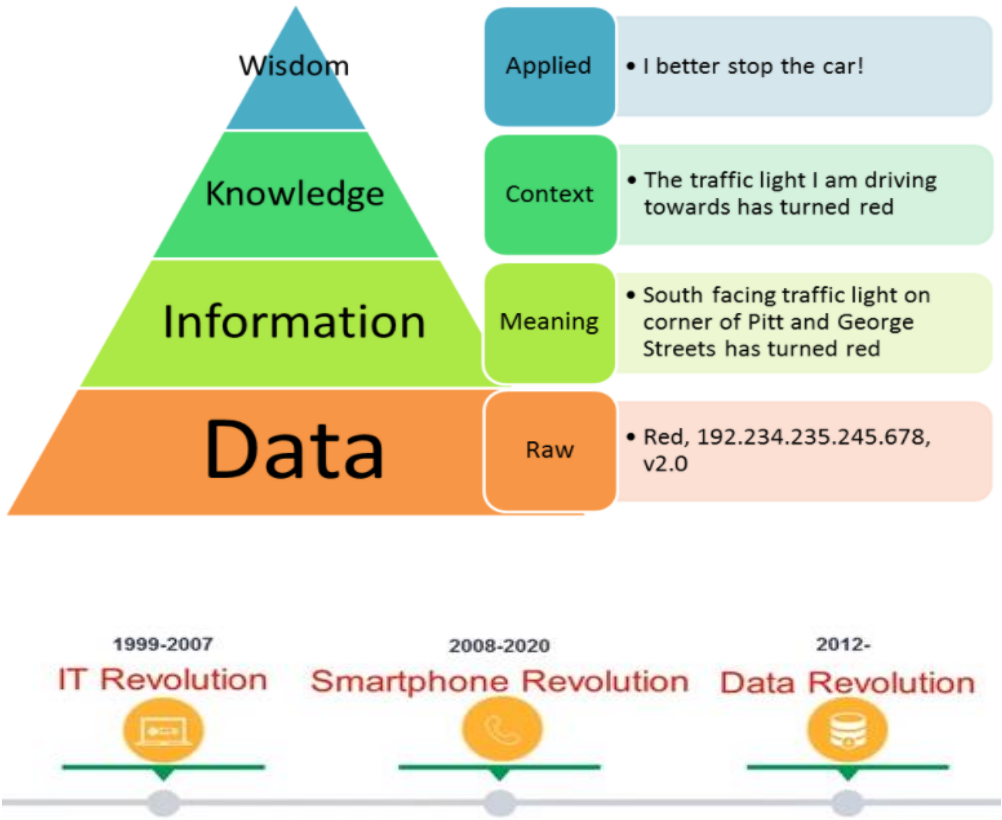
Estadística Básica

Análisis Exploratorio

Data Quality

Conceptos Básicos

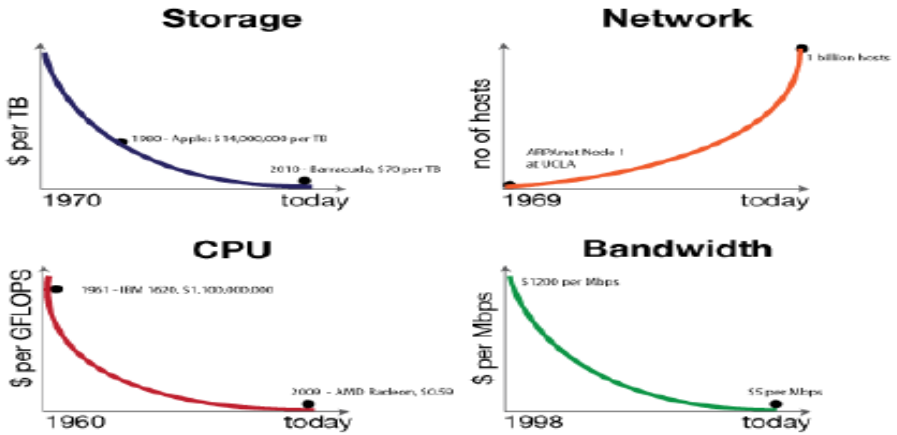
¿Qué es Data?



Data is the **seed** from which information, knowledge and wisdom sprouts and blossoms.

Data is the **key** to answer the right question

Data is a **set of values** of qualitative or quantitative variables.

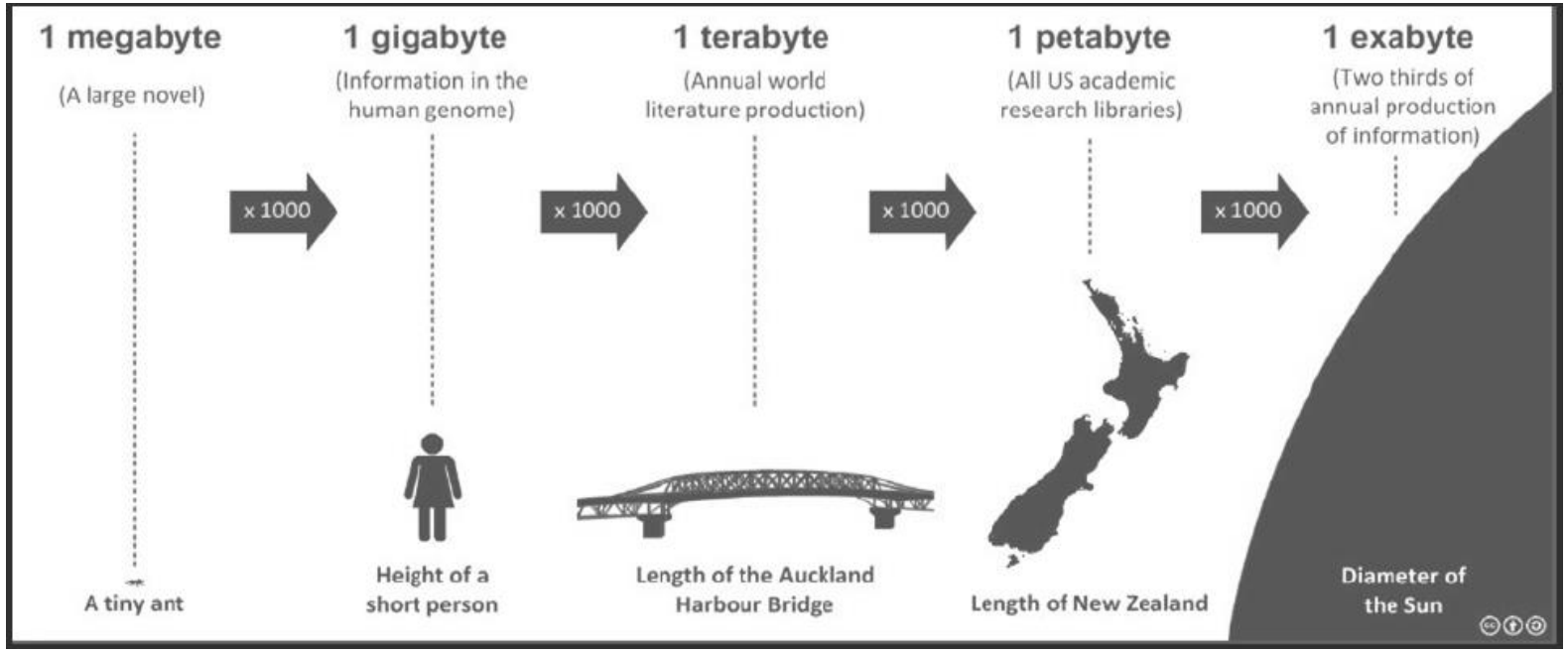


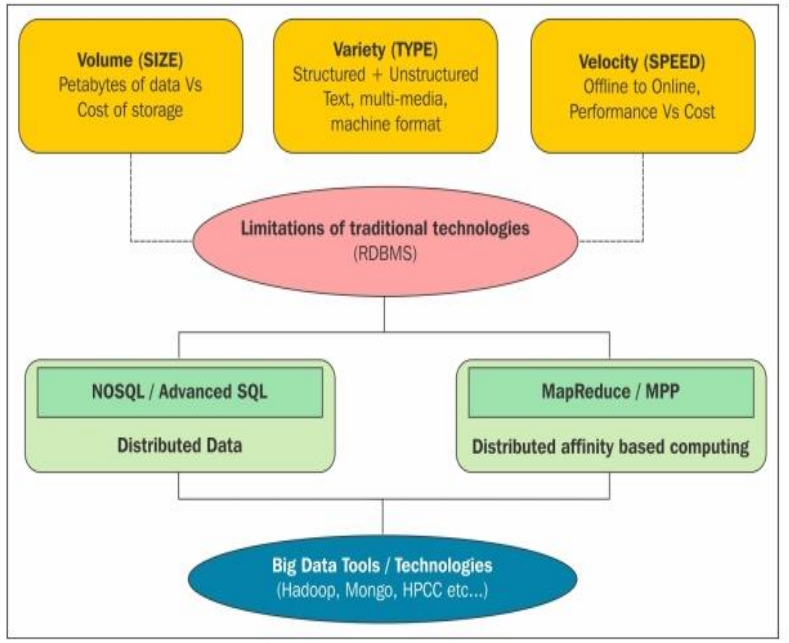
https://www.youtube.com/watch?v=jbkSRLYSojo&t=2s

¿Qué es Big Data?

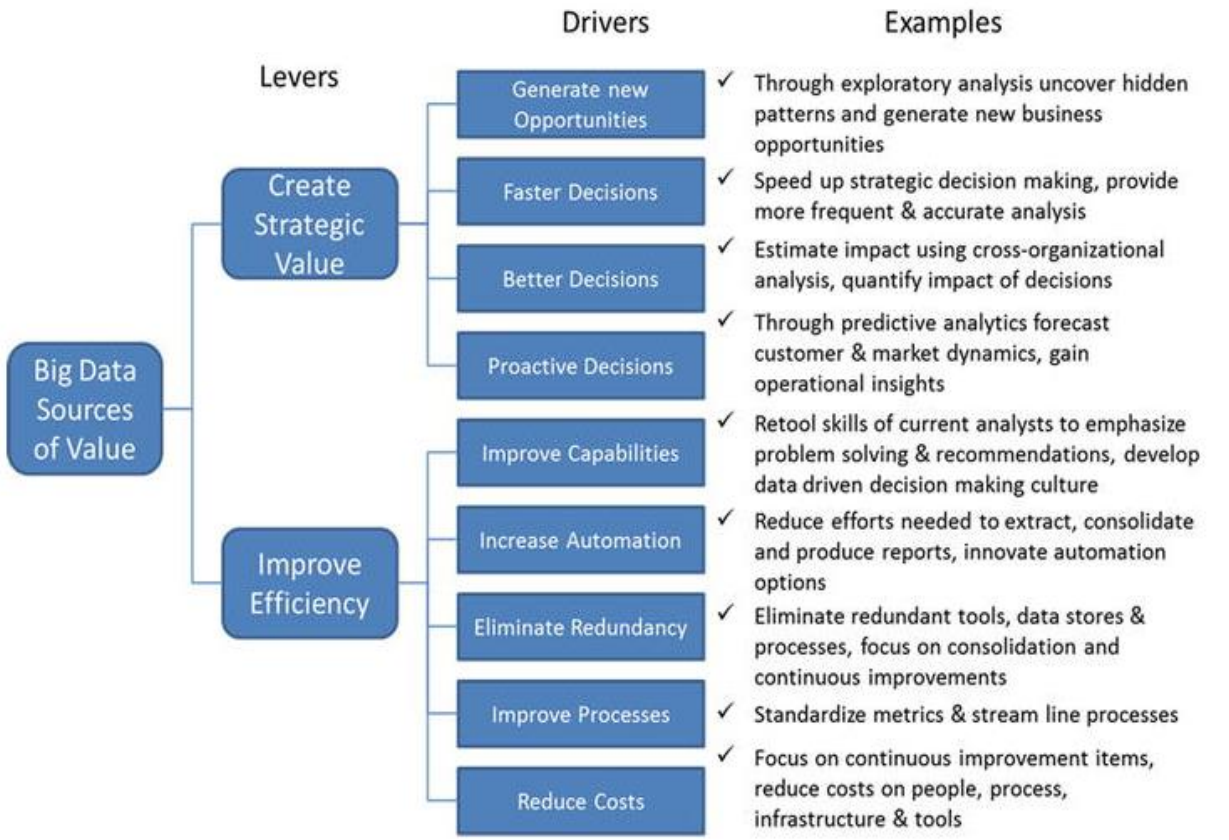
"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it..."

Dan Ariely, Duke University





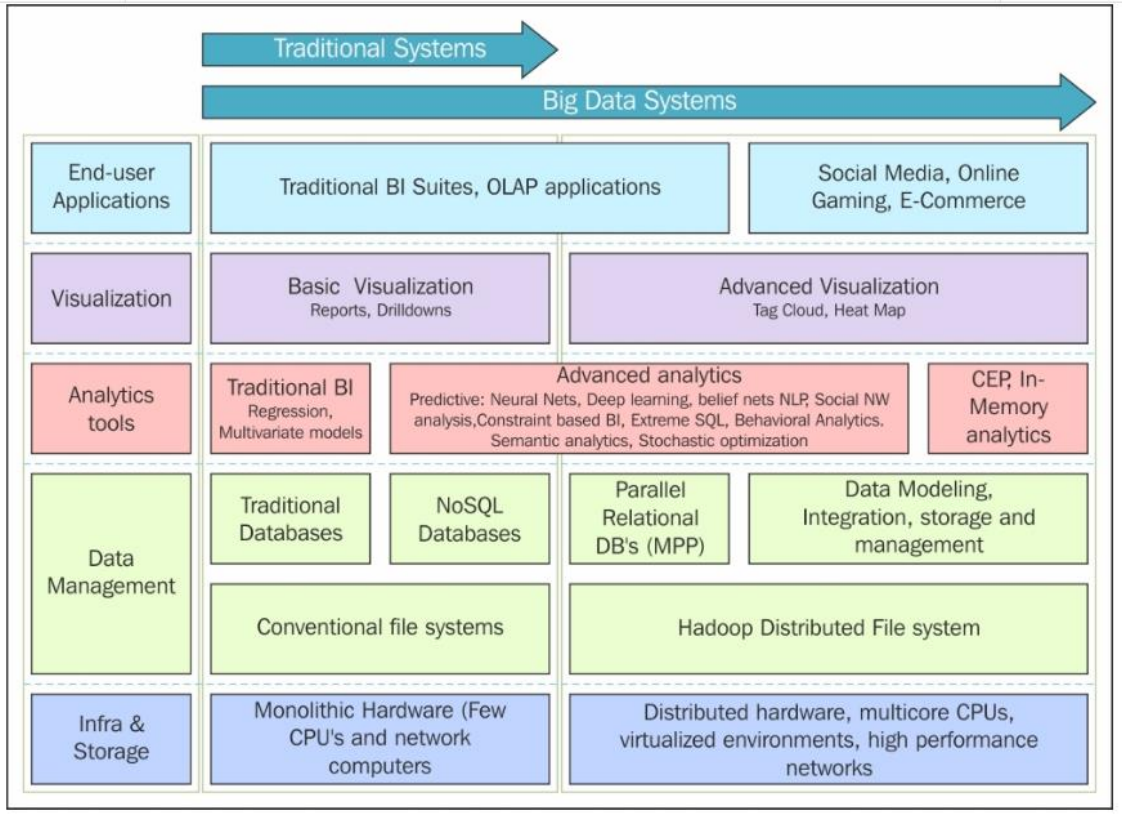
Casos de Uso de Big Data



Retail		Manufacturing	
✓ Customer Relationship Management ✓ Store Location & Layout	✓ Fraud Detection & Prevention ✓ Supply-Chain optimization ✓ Dynamic Pricing	✓ Product Research ✓ Engineering Analysis ✓ Predictive Maintenance	✓ Process & Quality Metrics ✓ Distribution Optimization
Financial Services		**Media & Telecommunications**	
✓ Algorithmic Trading ✓ Risk Analysis	✓ Fraud Detection ✓ Portfolio Analysis	✓ Network Optimization ✓ Customer Scoring	✓ Churn Prevention ✓ Fraud Prevention
Advertising & Public Relations		**Energy**	
✓ Demand Signaling ✓ Targeted Advertising	✓ Sentiment Analysis ✓ Customer Acquisition	✓ Smart Grid ✓ Exploration	✓ Operational Modeling ✓ Power-Line Sensors
Government		**Healthcare & Life Sciences**	
✓ Market Governance ✓ Weapon Systems & Counter Terrorism	✓ Econometrics ✓ Health Informatics	✓ Pharmacogenomics ✓ Bioinformatics	✓ Pharmaceutical Research ✓ Clinical Outcomes Research

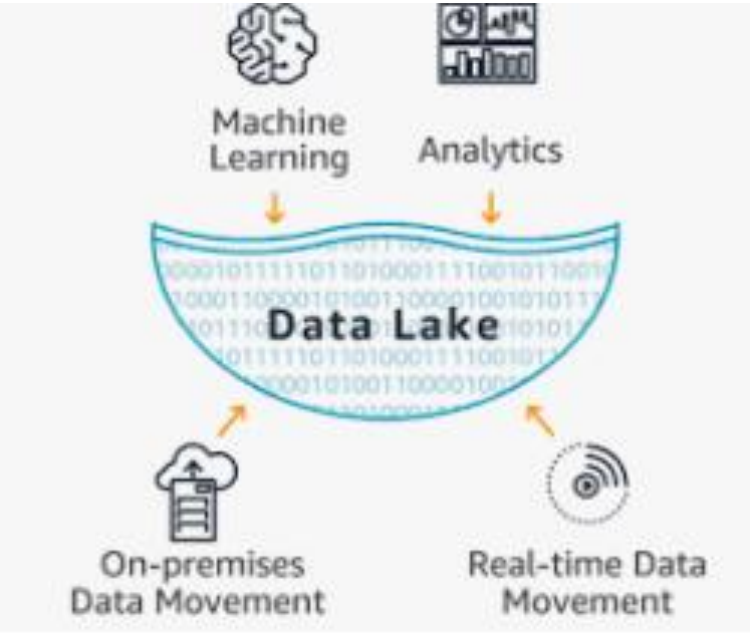
Diferencias entre DWH/BI y Big Data

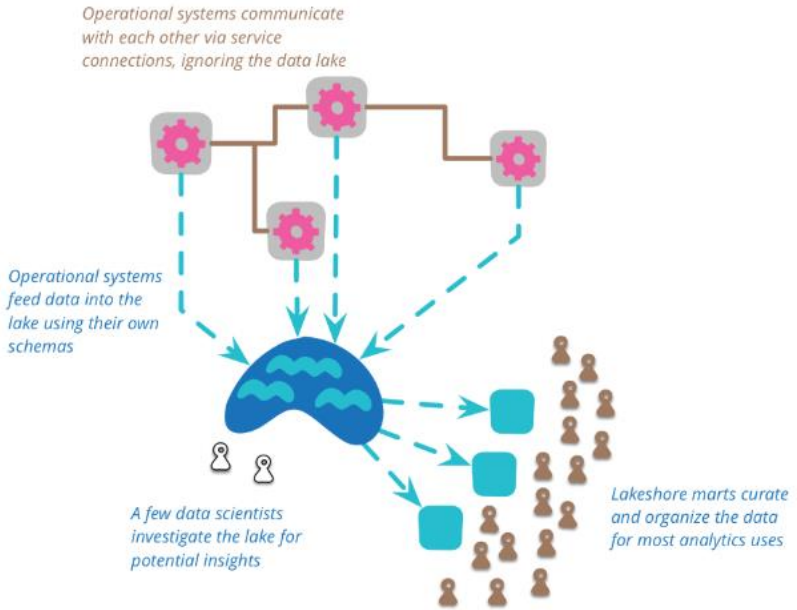
- La escalabilidad del almacenamiento y el poder de procesamiento son diferentes.
- En el enfoque tradicional, la data proviene de sistemas relacionales y estructurados, en la nueva era del Big Data la data puede provenir de todo tipo de fuentes incluyendo las no estructuradas.
- La velocidad de procesamiento de los sistemas tradicionales es menor.
- La complejidad de los algoritmos que se pueden aplicar sobre la data.
- El enfoque tradicional ofrece reporteria y cubos con drill-downs, el nuevo enfoque es mucho más visual incluyendo mapas de calor, graficas de N dimensiones, etc. El Story teller es una realidad y una necesidad.



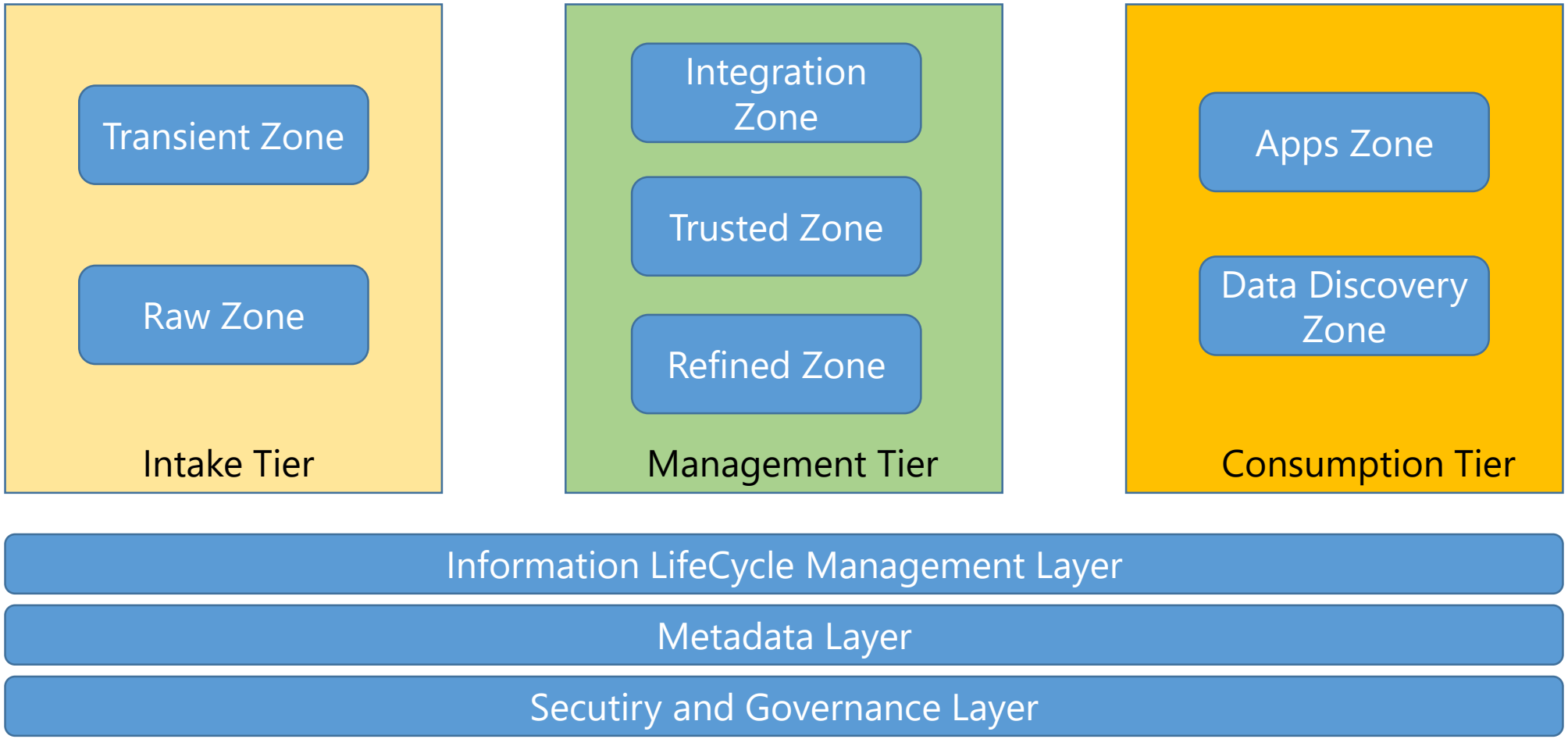
Que es un Datalake

Un Data Lake, es un repositorio que almacena una gran cantidad de datos estructurados, semi-estructurados y no estructurados en su formato natural, es decir todo está almacenado de forma plana y los datos se van procesando/preparando según sea necesario. Debe ser reconocido como un punto de integración de la data para propósitos de análisis, no como un puente o colaboración entre los sistemas operacionales



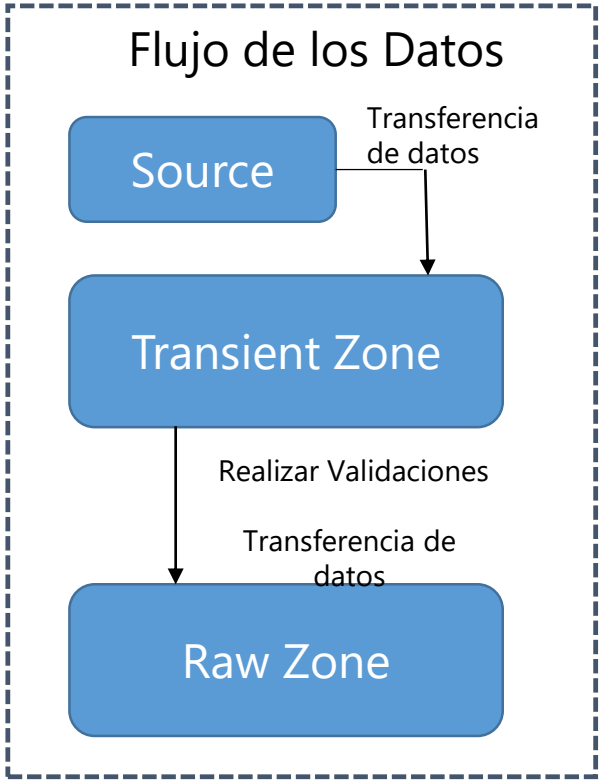


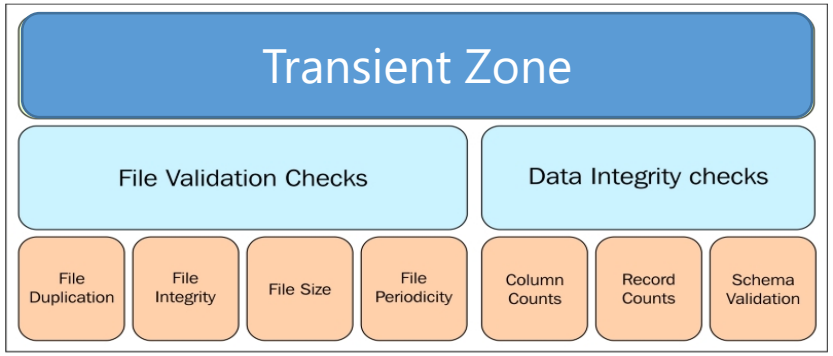
Distribución Lógica del Data Lake



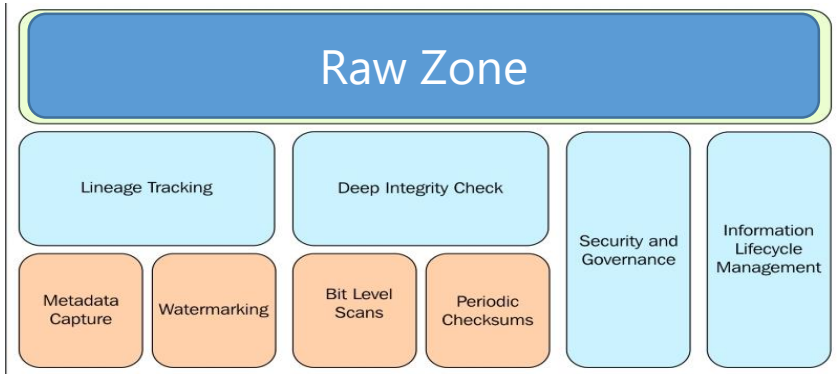
Intake Tier

- La capa de ingestion de datos se utiliza para proveer todas las interfaces necesarias para establecer las conexiones a las diferentes Fuentes (**Pull-based**) y guardar la data en su **formato y estado original**





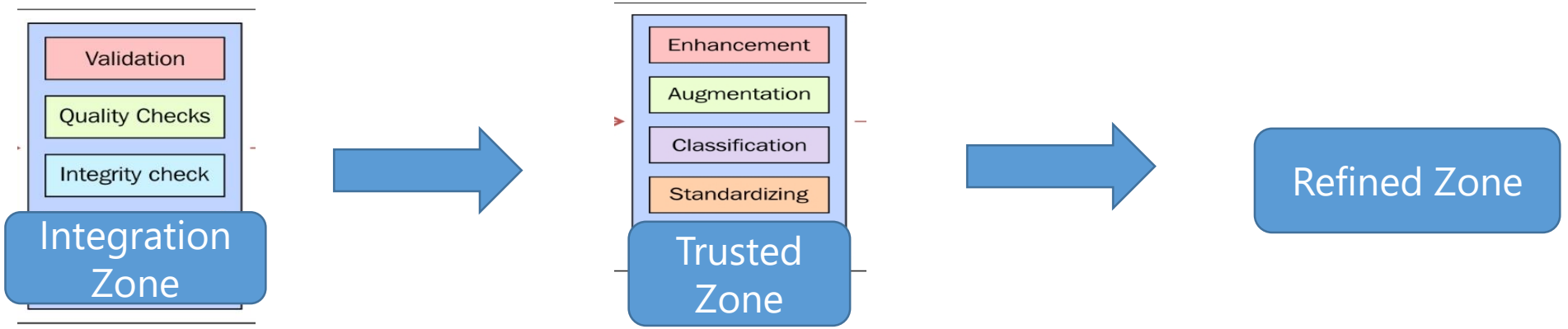
- Tambien conocida como Landing Zone es una zona en la cual sera almacenada toda la data proveniente de las diferentes Fuentes. Generalmente esta zona esta organizada en función de las Fuentes de datos.
- En esta zona se llevan a cabo las validaciones mas basicas como conteo de registros, tamaño de los archivos, etc.



- Luego de hacer las validaciones necesarias la data debe ingestarse en la Raw Zone, la cual contendra la data en su forma original y con la fidelidad original segun la fuente de datos.
- La data alojada en esta zona generalmente corresponde a datos usados muy activamente + datos historicos (persistencia)

Management Tier

- Su proposito es adquirir la data desde la Intake Tier para procesarla de modo que quede lista para su consumo, es sumamente importante cuando los modelos analiticos y analisis exploratorios no pueden consumir la data cruda.



- La zona de integración sirve ademas de integrar las diversas Fuentes para aplicar las **transformaciones comunes.** La data queda estandarizada, limpia y estructurada.
- En esta zona se llevan a cabo tareas de data **quality checks, integrity checks**

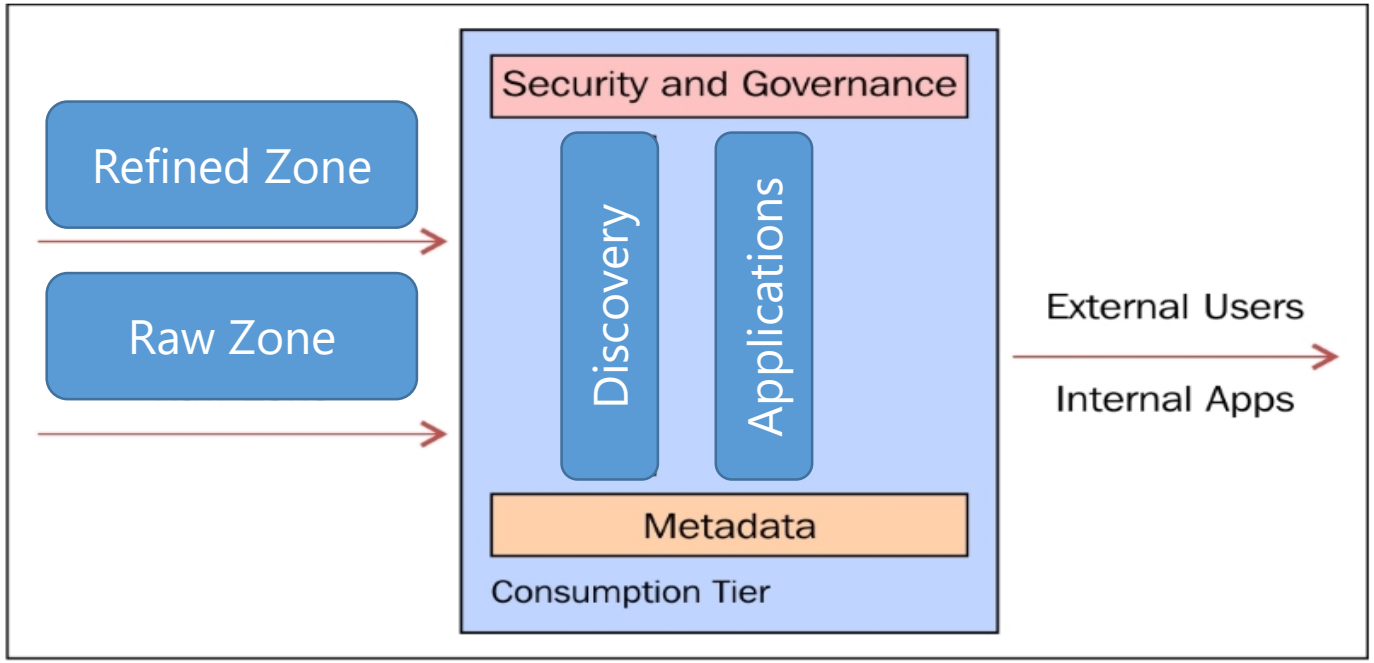
- La data en esta zona ya ha pasado por los procesos que **aplican reglas de negocio** para derivar o agregar nuevos atributos

- Esta es la zona final para la data limpia, procesada y que cumple con todas las reglas de negocio y si hay necesidad de tener data agregada esta es la zona indicada

Consumption Tier

- La capa de consumo es la puerta de acceso hacia la data ya sea en la raw zone o bien a las zonas con formatos estructurados, desde aca se controlan accesos y otros aspectos de seguridad.

- La Data Discovery Zone es la "**Sandbox**" para los analistas y cientificos de datos
- La Apps Zone es utilizada generalmente por diferentes aplicaciones o bien herramientas de Inteligencia de Negocio



Bases NoSql

- Hadoop/Hbase
- Cassandra
- Hypertable
- Accumulo
- Amazon SimpleDB
- Flink
- MongoDB
- Amazon DynamoDB
- Redis
- Oracle NoSQL Database
- Neo4J

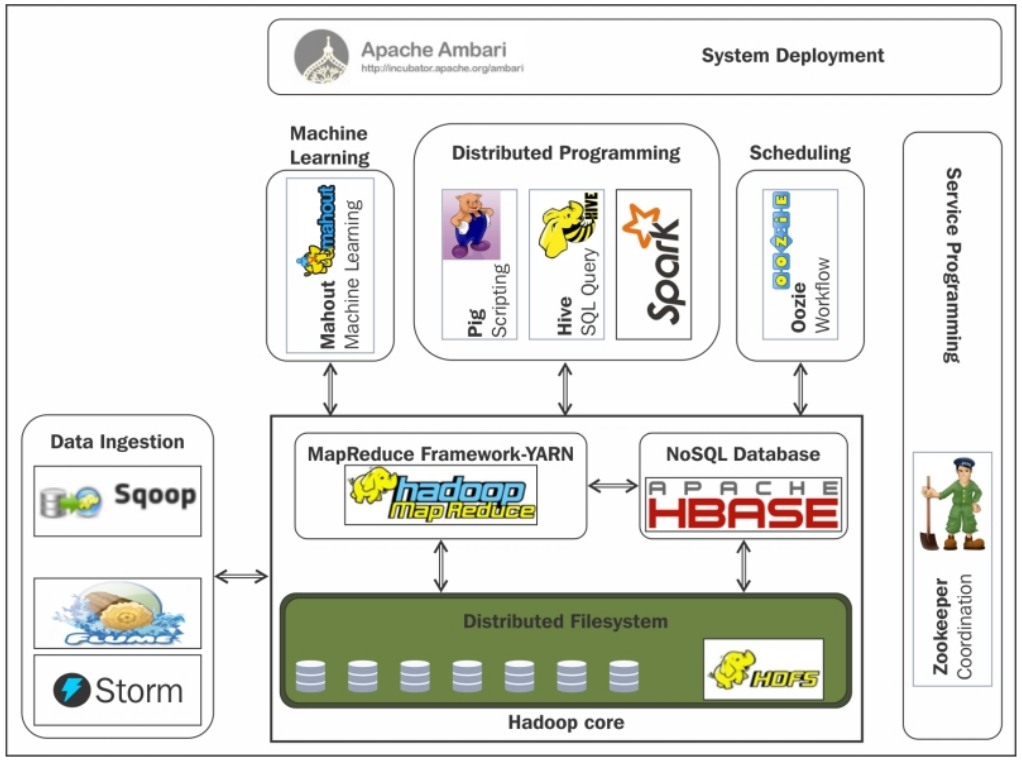
Distribuciones de Hadoop

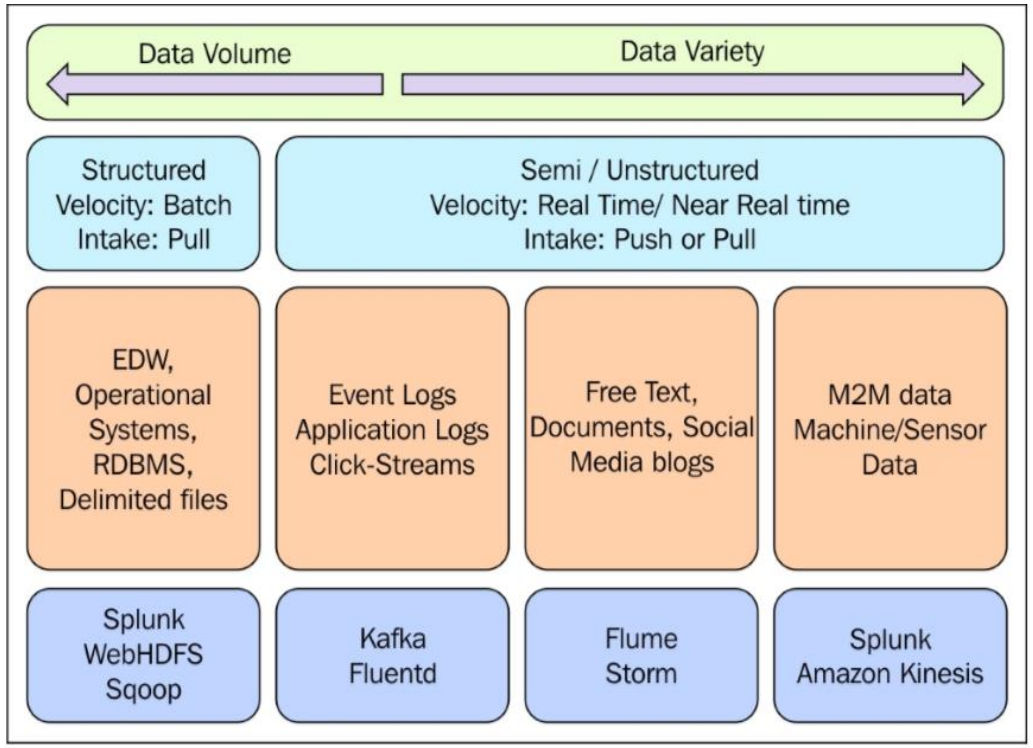
cloudera

MAPR

Hortonworks

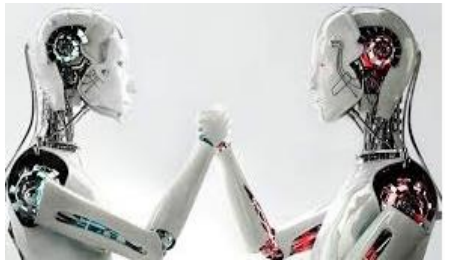
Ecosistema en el Datalake (Hadoop)





Que no es Machine Learning

Supongamos que tienes un problema de Machine Learning que debes resolver, sin embargo, no conoces que es Machine Learning. Empezaremos por decirte lo que no es:



No es una investigación sobre las capacidades de un algoritmo.

No es el desarrollo de un algoritmo o de alguna teoria.

No es una investigación esoterica de algun tipo de aprendizaje.

No es la construcción de un agente de inteligencia artificial

No es la construcción de un circuito que emita señales

https://machinelearningmastery.com/think-machine-learning/
http://blogs.teradata.com/data-points/todays-machine-learning-ai-not/

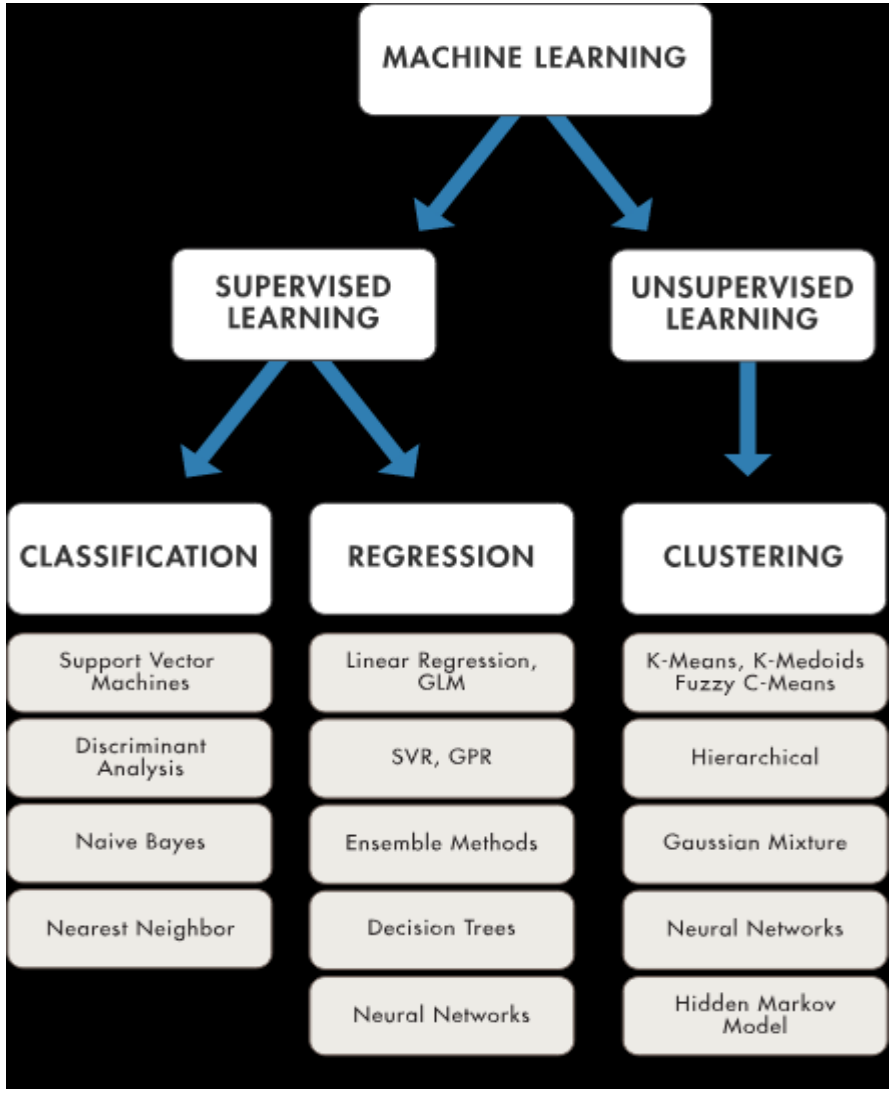
Machine Learning

Machine Learning son un conjunto de métodos/algoritmos diseñados para encontrar patrones y tendencias en los datos. Se encuentra en la intersección entre las matemáticas y estadística con la ingeniería de software y ciencias de la computación.

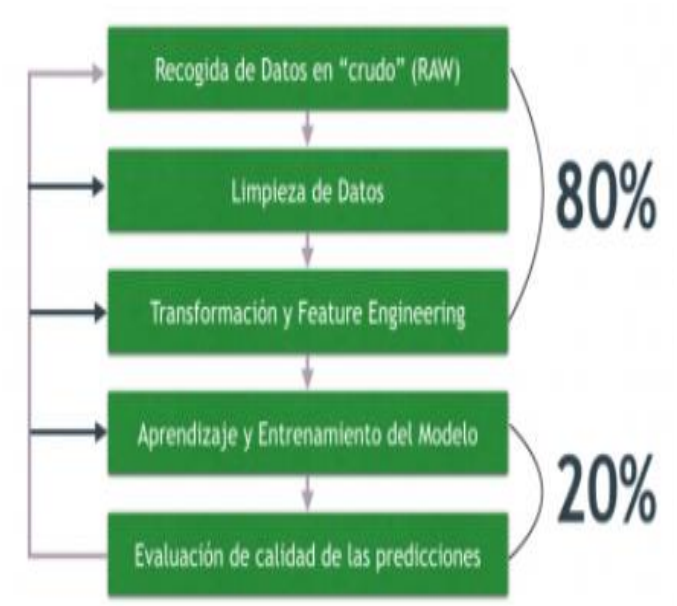
Familias de técnicas de ML

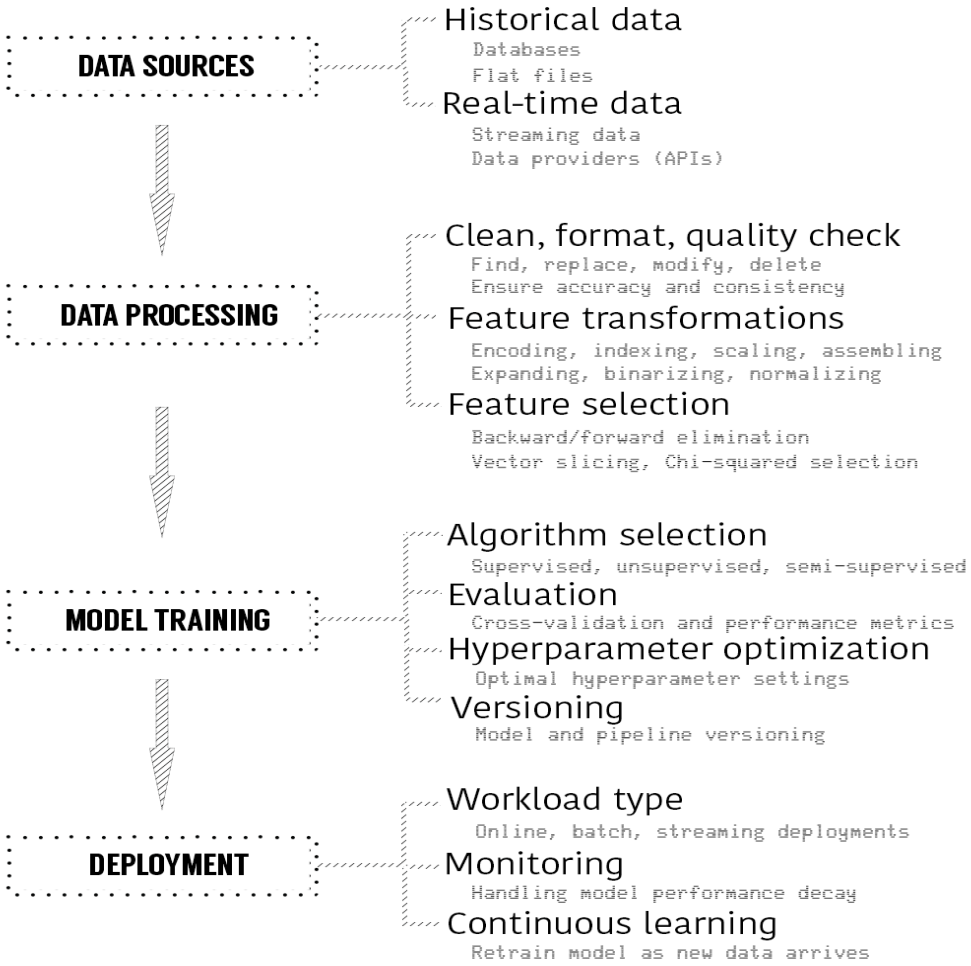
1. Aprendizaje Supervisado: En este proceso de aprendizaje la variable de salida está bien definida (variable objetivo), es decir estas técnicas nos son útiles cuando nos interesa hacer predicciones sobre una variable objetivo.

2. Aprendizaje No Supervisado: Este proceso de aprendizaje no implica tener una variable objetivo bien identificada, su objetivo no es hacer predicciones.



Ciclo Vida Machine Learning

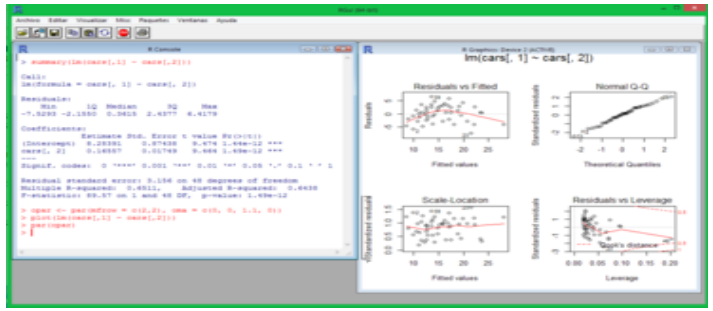




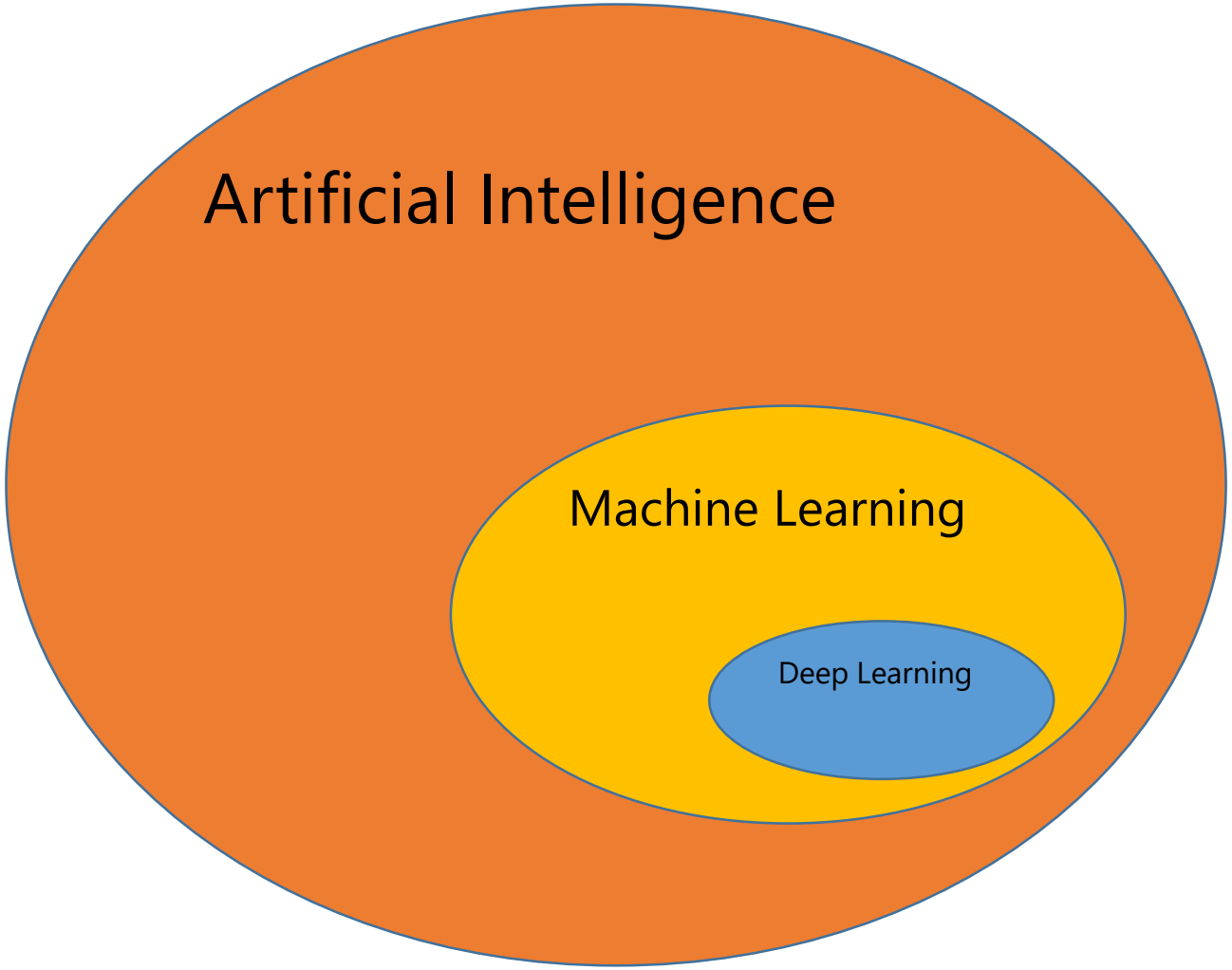
Herramientas para Machine Learning

Las herramientas para soportar las actividades de ML son una gran cantidad, entre las más populares destacan:

- Lenguaje R
- Python
- Weka
- Knime
- RapidMiner
- Azure ML Studio
- TensorFlow
- BigML
- SkyTree
- IBM Watson
- MLIB Spark
- Julia
- Jupyter



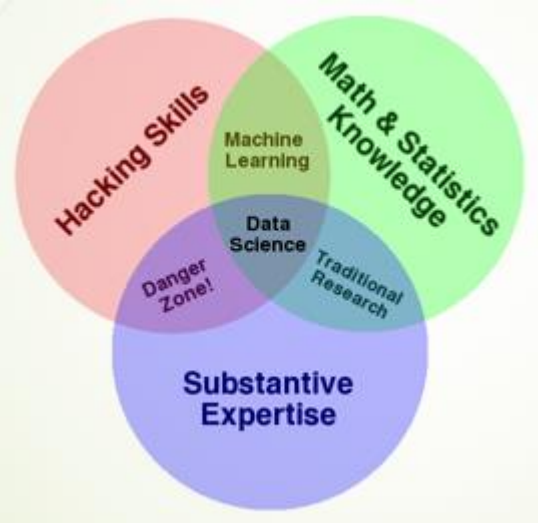
AI vs ML vs DL

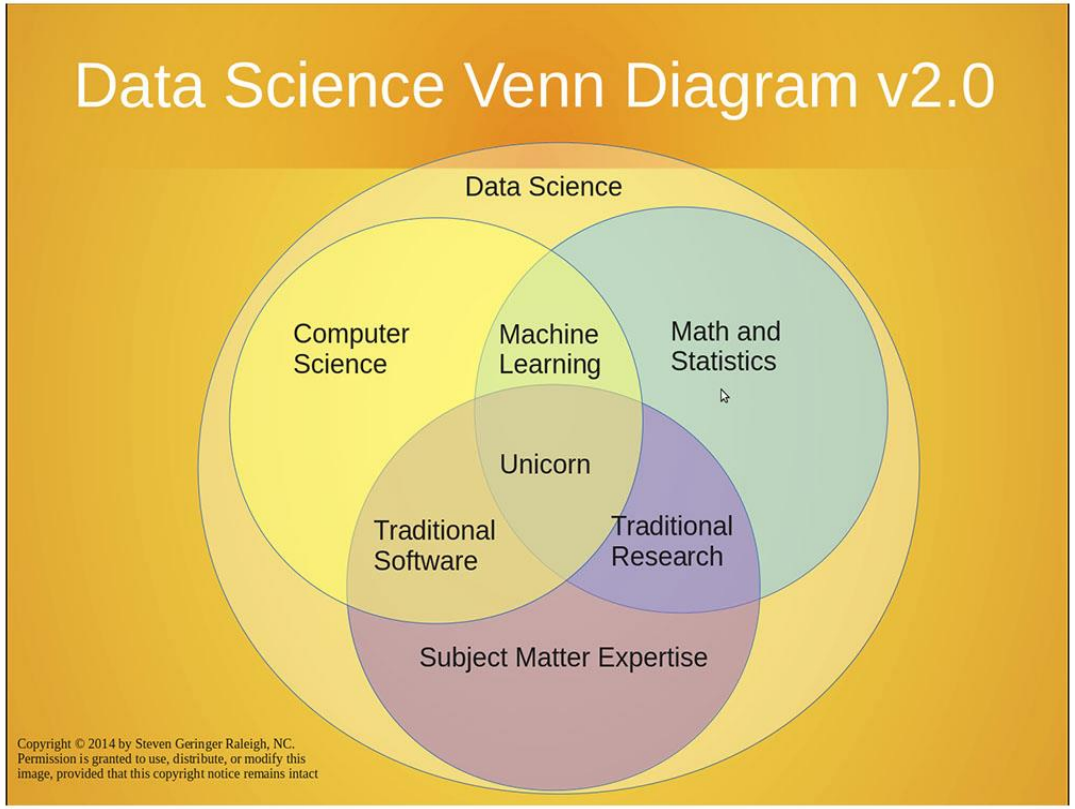


*Machine Learning is a **current application of AI** based around the idea that we should really just be able to give machines access to data and let them learn for themselves*

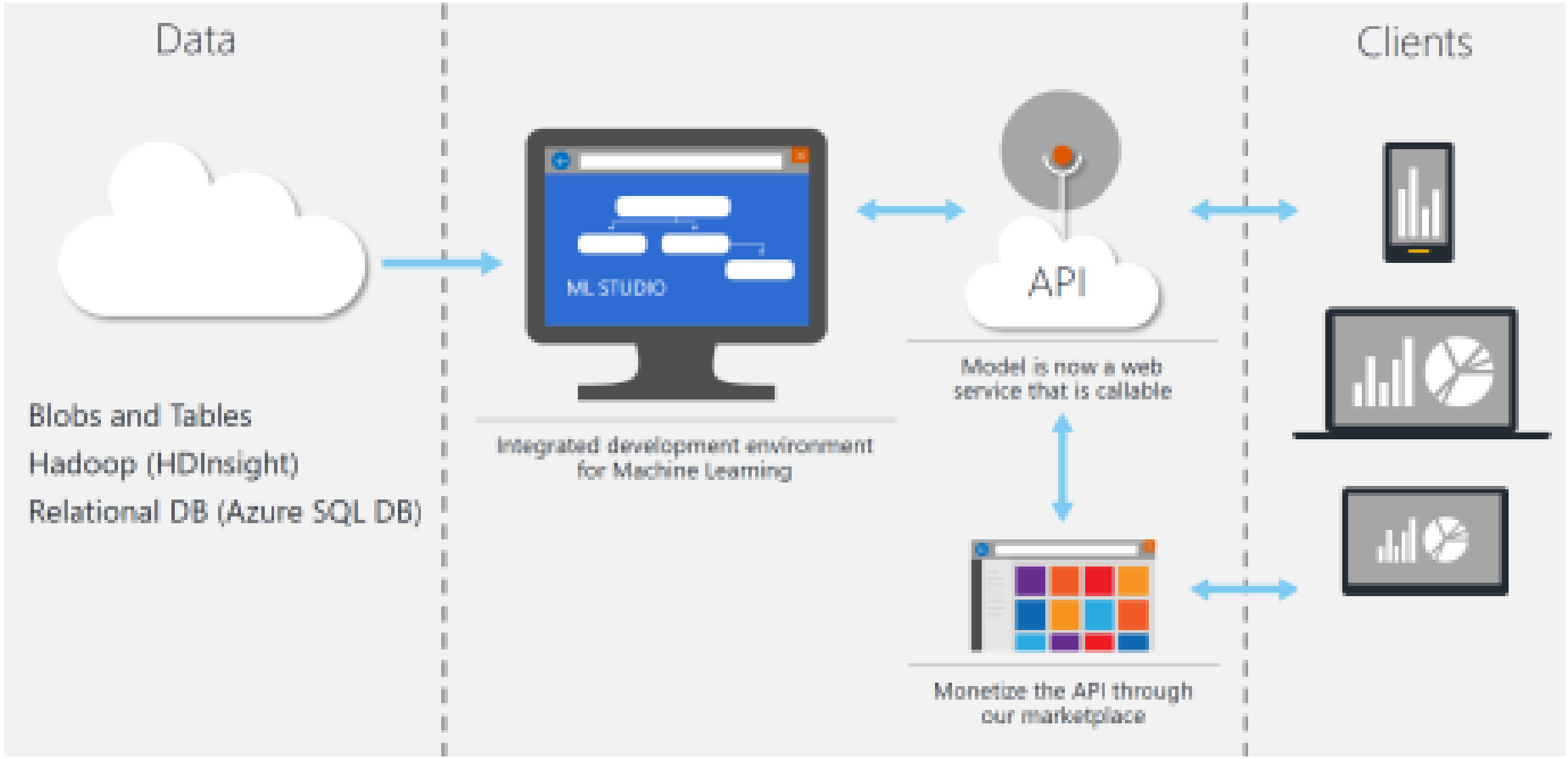
Deep Learning — A Technique for Implementing Machine Learning

El Científico de Datos





ML + Big Data



https://biz-excellence.com/technologies/azure-machine-learning/

Q & A

Bibliografía

Big Data Analytics: Turning Big Data into Big Money

by Frank J. Ohlhorst, November 2012

Hadoop Essentials

by Swizec Teller, April 2015

Scalable Big Data Architecture: A Practitioner's Guide to Choosing Relevant Big Data Architecture

by Bahaaldine Azarmi,2016

A Framework for Analysis of Data Quality Research

by Richard Y. Wang,1995