**Big Data**

Información que requiere el uso de nuevas técnicas de arquitecturas y analítica para las nuevas fuentes del valor del negocio.

* Volumen
* Variedad
* Velocidad
* Veracidad
* Valor

**Análisis**

* Descriptivo, ¿Que paso?
* Diagnóstico, ¿Qué fue lo que paso?
* Predictivo, ¿Qué puede pasar?
* Prescriptivo, ¿Que debería de hacer?

**Metodologías**

* CRISP-DM

**Hadoop**

* Es el ecosistema de los datos distribuidos.
* HDFS,Hadoop,Hbase,Hive,Pig

**Resume**

1. Construye un clúster de máquinas.
2. Distribución de los datos HDFS.
3. Procesamiento de los datos en paralelo.

**HDFS – La forma de cómo se almacenan los datos**

Sistema de archivos capas de almacenar grandes bloques de archivos.

**NameNode**: Registra la info de donde este cada bloque y poder describirlos.

**DataNode**: Puede guardar el dato.

- Cada bloque son de 128 megabytes.

- Cada bloque es independiente.

- Este puede ser replicado en 3 nodos diferentes y se puede mover en diferentes racks.

- Es mas barato mover la computación que mover los datos entre computadores.

**HDFS Client**

* Este interactúa con NameNode y DataNode

El cliente o aplicación se comunica con el NameNode.

El NameNode dirige el paquete y sabe donde esta el paquete en el DataNode

**MAP-REDUCE – Como se procesan los datos**

* Esta es una de las técnicas.
* Es la indexación de los datos de manera rápida.
* Es un paradigma de programación.
* Utiliza un enfoque de divide y vencerás.

**Map** toma una llave de entrado y genera una lista.

**Reduce** toma una lista generada y devuelve una lista de llaves y valores.

* **Se clasifican y se agrupan.**

Cloudera es una compañía que ya brinda el soporte.