Big Data, Machine Learning & Business Intelligence

Por: Carlos Carreño

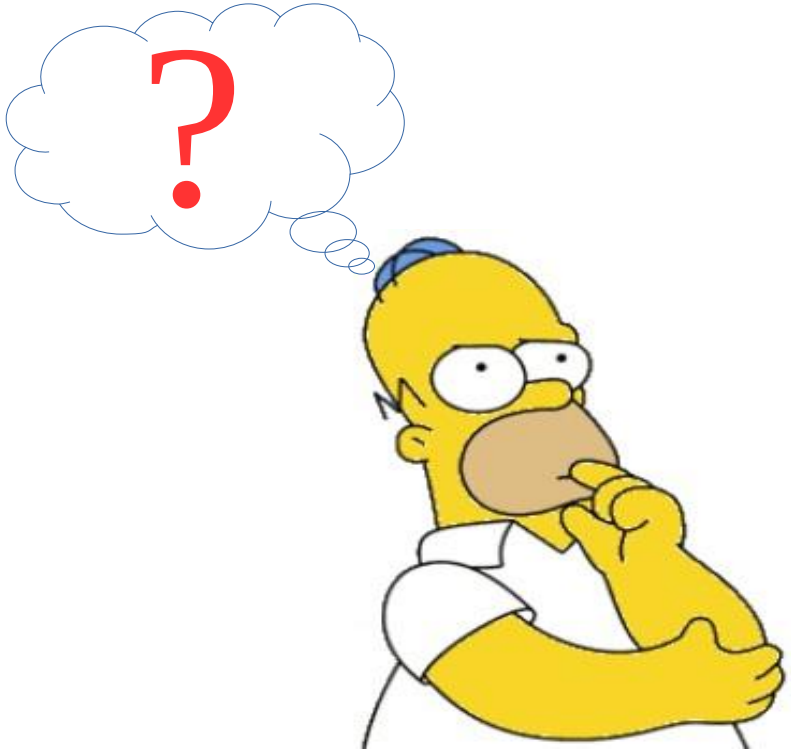
ccarreno@cienciadedatos.es

Unidad 2 Big Data

- Big Data
  - Componentes de ecosistema
- Lenguaje Python y R
- Business Analytics
  - Descriptivo
  - Predictivo
  - Prescriptivo

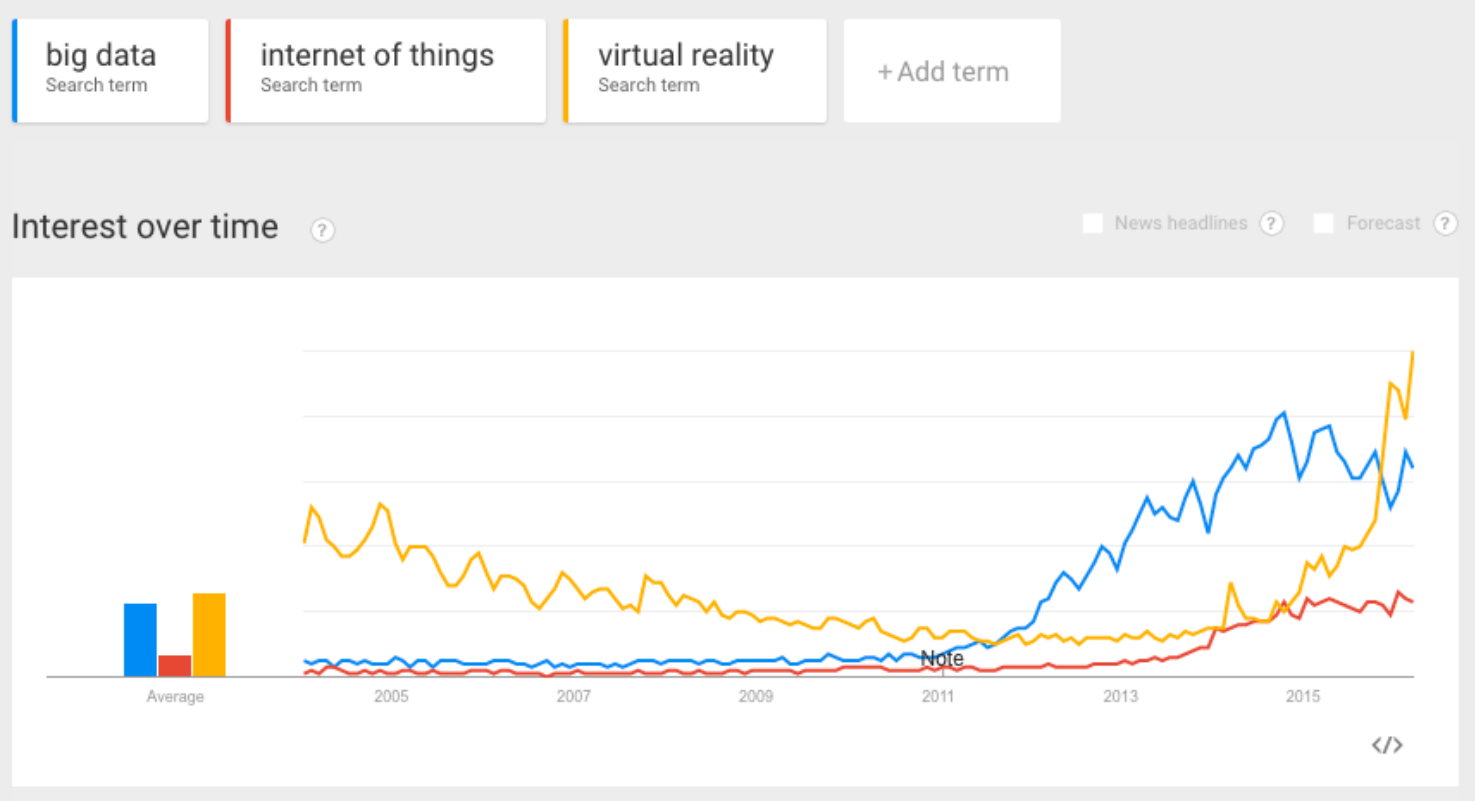
Big Data

- Que es Big Data?

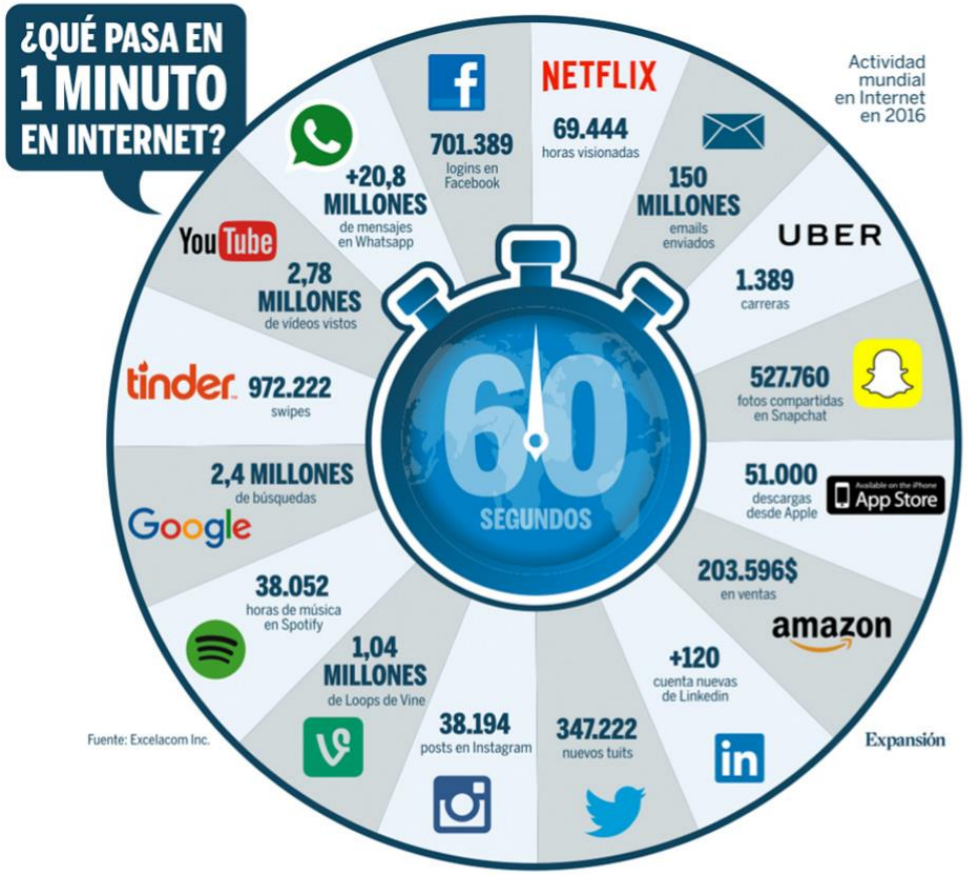


Big Data:

- Desde cuando hablamos de Big Data?



Big Data: 1 Minuto en Datos



Big Data: Cuanto es demasiados Datos

- 1 Gigabyte = 10^9 = 1,000,000,000
- 1 Terabyte = 10^{12} = 1,000,000,000,000
- 1 Peta byte = 10^{15} = 1,000,000,000,000,000
- 1 Exabyte = 10^{18} = 1,000,000,000,000,000,000
- …
- 1 Quintillón
- 10^{30} = 1,000,000,000,000,000,000,000,000,000,000

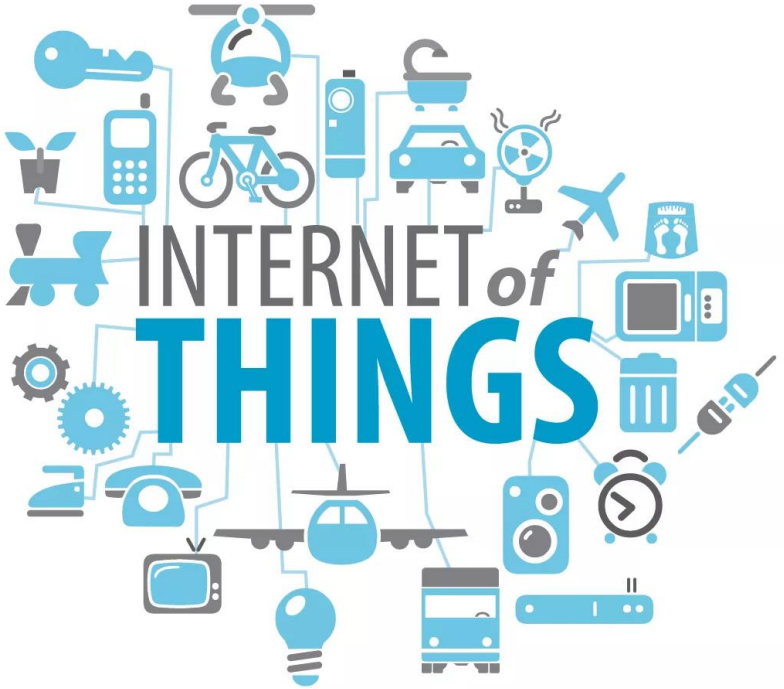
Big Data: Seguimos creciendo!!

- Al 2016 la población creció a **7,400** millones de personas.
- Se prevee:
  - **18.9 billones** de dispositivos.
  - Que el tráfico global de datos móviles alcance **10.8 Exabytes mensuales**

Fuente: ONU
http://exitosanoticias.pe/onu-poblacion-mundial-llego-a-7400-millones/

Big Data: No solo humanos

- No solo los humanos producen datos



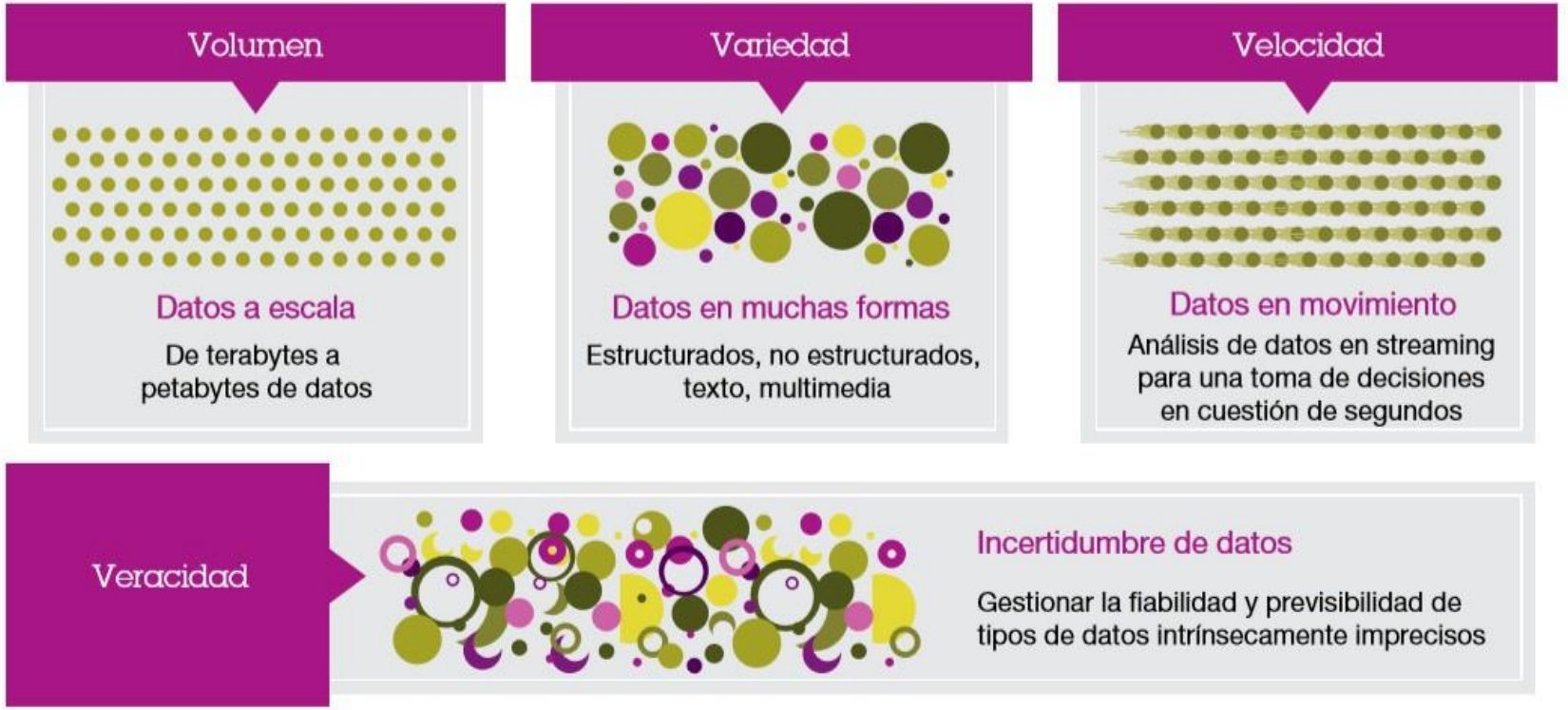
Big Data: Aterrizando el Concepto

- **Big Data** se refiere al procesamiento de **volúmenes** de datos tan grandes que no se pueden realizar con tecnologías tradicionales a una **velocidad** adecuada y a los procedimientos para encontrar **patrones repetitivos** en estos datos.

Referencia: https://es.wikipedia.org/wiki/Big_data

Big Data: Las 3 V

Dimensiones de big data



Big Data: Componentes del Ecosistema - Hadoop

- **Apache™ Hadoop®** es un proyecto de **software libre** que permite el procesamiento distribuido de **grandes volúmenes de datos** en **clusters de servidores básicos**.
- **Hadoop** está diseñado para extender un **sistema de archivos** de servidor único **a miles de máquinas** y a **petabytes** de datos con un muy alto grado de tolerancia a las fallas.

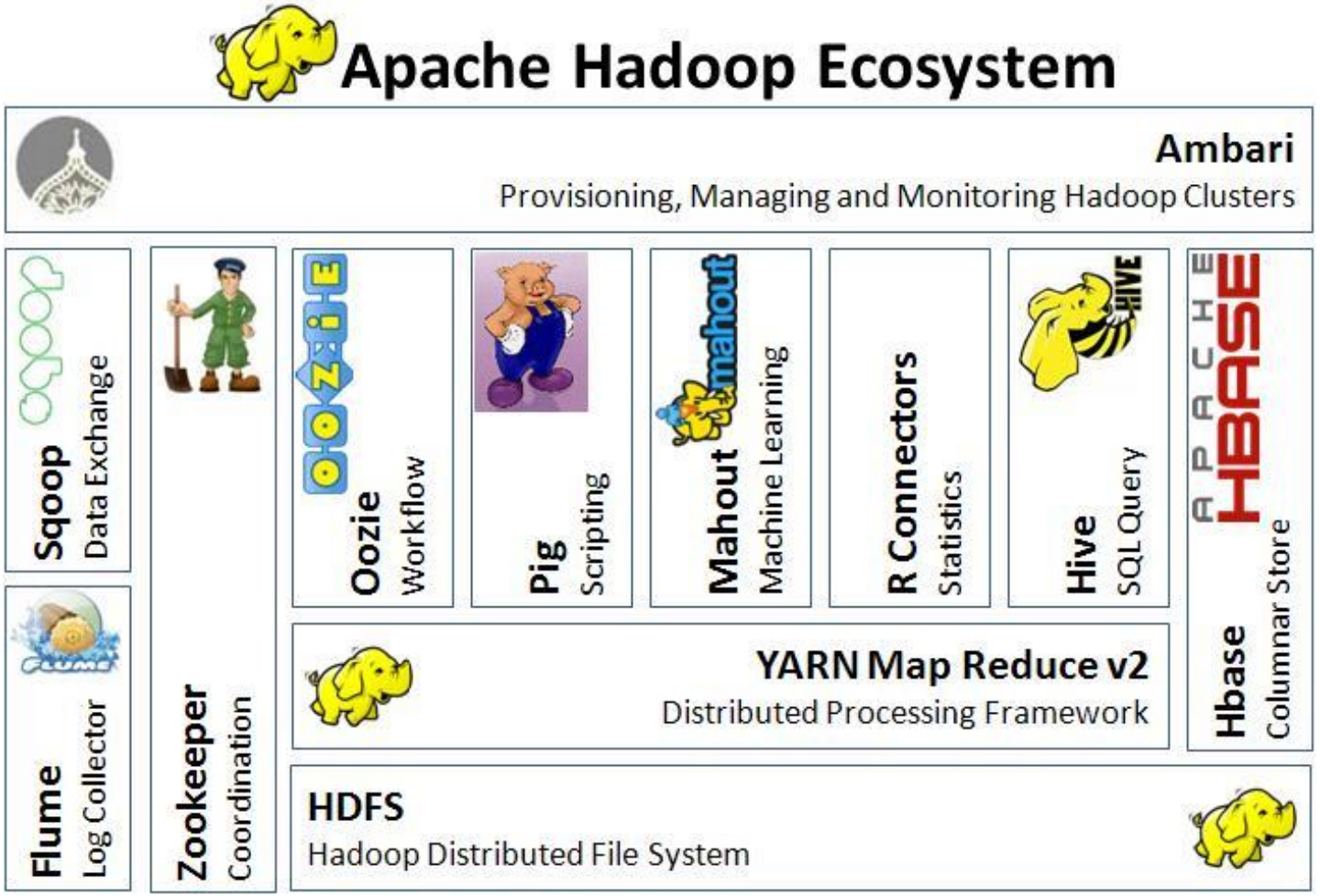
Big Data: Hadoop hace posible el big data

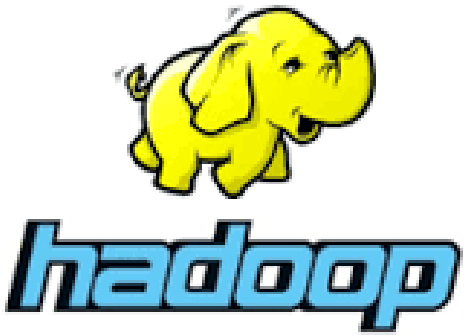
- **Redimensionable**, pueden agregarse tantos nuevos nodos como sea necesario.
- **Rentable**, Hadoop hace posible la computación paralela con servidores básicos.
- **Flexible**, Hadoop funciona sin esquema y puede absorber cualquier tipo de datos.
- **Tolerante a fallas**, si se pierde un nodo, el sistema redirige el trabajo a otra localización de los datos y continúa procesando sin perder el ritmo.

Big Data: Plataformas

- Hadoop ofrece una base para la creación de plataformas o ecosistemas comerciales para el análisis de Big Data.
- Detrás del uso de una plataforma comercial de Big Data esta el propósito de facilitar su adopción, esto es **"Hadoop como servicio"**.

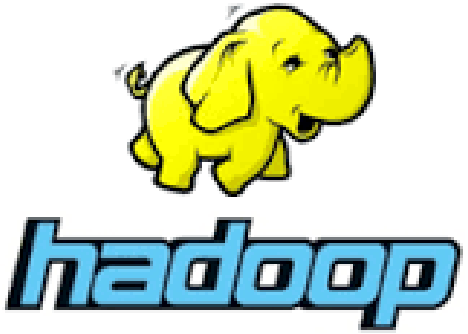
Big Data: Hadoop Tooling



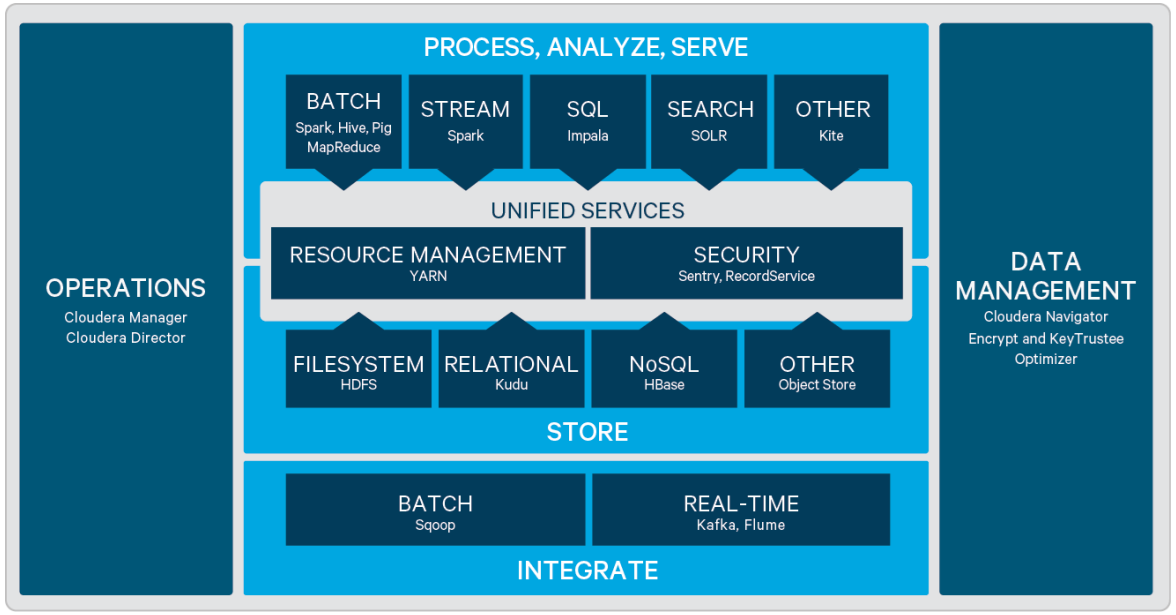


Big Data: Plataformas basadas en Hadoop

- Cloudera
- Amazon Web Services
- Hortonworks
- MapR
- IBM
- Microsoft HDInsight
- Intel Distribution for Apache Hadoop
- Datastax Enterprise Analytics
- Teradata Enterprise Access for Hadoop
- Pivotal HD

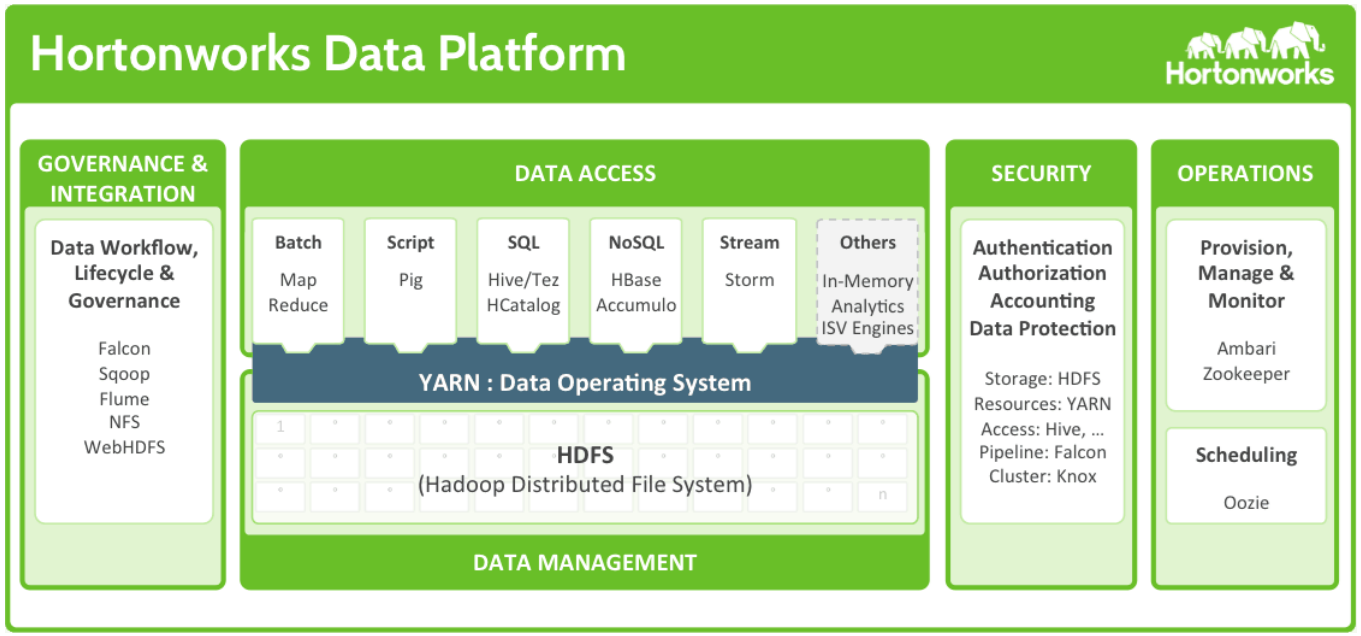


Big Data: Plataforma Cloudera



- **Cloudera una de las primeras ofertas comerciales de Hadoop** una de
- las mas populares.
- **Cloudera** aporta **Impala**, que ofrece en tiempo real el procesamiento masivo
- paralelo de Big Data a Hadoop.

Big Data: Plataforma Hortonworks



.Hortonworks es una de las pocas plataformas 100% de tecnología Hadoop de código abierto sin ninguna modificación propietaria. También fueron los primeros en integrar el soporte para Apache HCatalog, que crea "metadatos", datos dentro de los datos, simplificando el proceso de compartir sus datos a través de otras capas de servicio como Apache Hive o Pig.

Big Data y Modelos de Arquitectura: BD No SQL

- Para dar soporte al Bigdata surgen tres modelos de arquitectura con sus propias tecnologías

Leguaje R

- **Pros**
  - Excelente gama de paquetes de código abierto y de alta calidad, específicos de dominio.
  - La instalación básica viene con funciones y métodos estadísticos integrales muy completos.
  - R también maneja el álgebra de matriz particularmente bien.
  - La visualización de datos es una fortaleza clave con el uso de bibliotecas como ggplot2.
- **Contras**
  - No es un lenguaje rápido.
  - Es fantástico para fines estadísticos y científicos de datos. Pero menos para la programación de propósito general.
  - Tiene algunas características inusuales que pueden atrapar a los programadores con experiencia en otros idiomas. Por ejemplo: indexación desde 1, utilizando, estructuras de datos no convencionales. operadores de asignación múltiple.
- **Veredicto**: "brillante en para lo que está diseñado“.
  - R es un lenguaje poderoso que sobresale en una gran variedad de aplicaciones estadísticas y de visualización de datos, y ser de código abierto permite una comunidad muy activa de contribuyentes.

Python

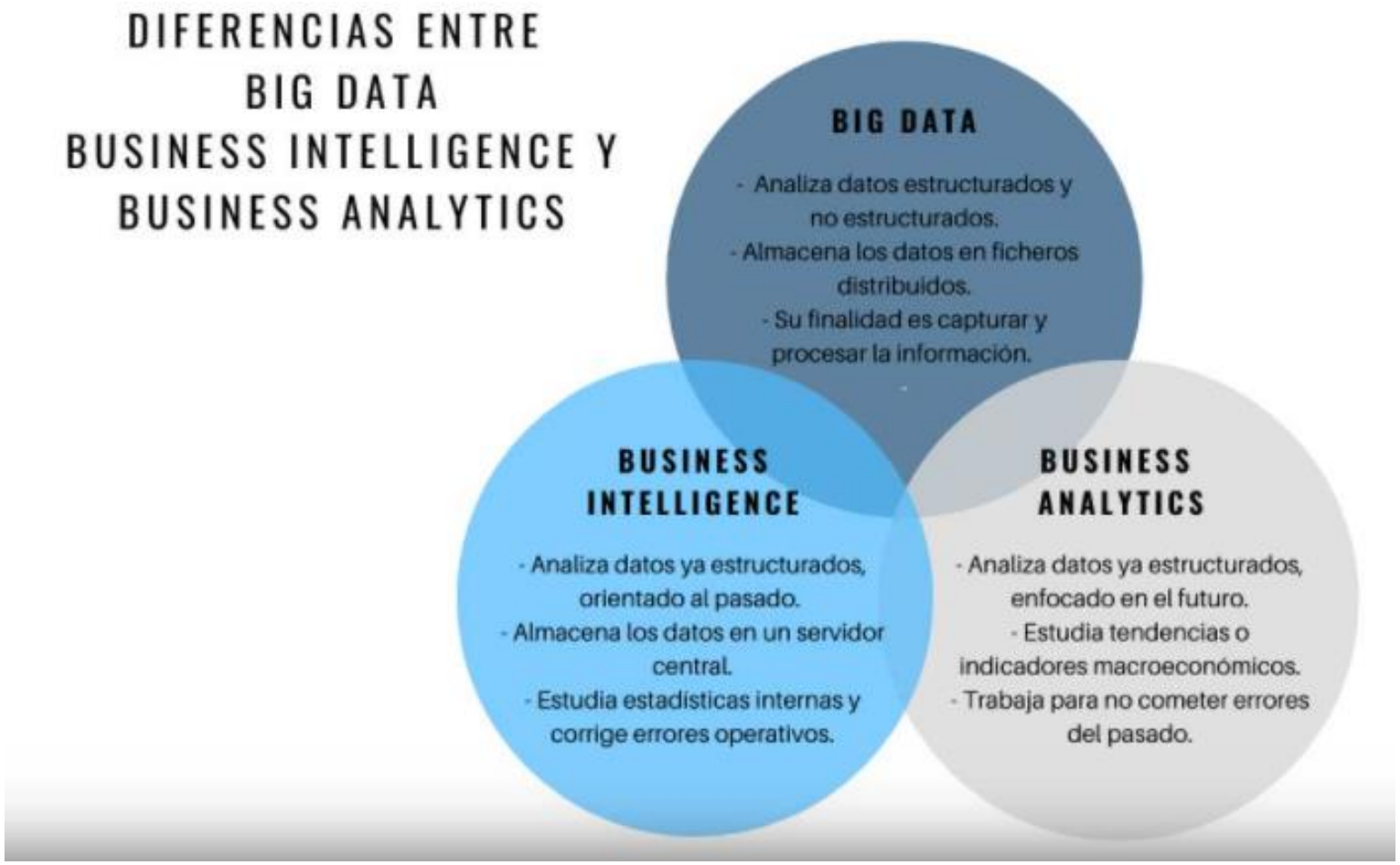
- **Pros**
  - Es un lenguaje de programación de propósito general.
  - Python tiene una curva de aprendizaje muy rápida.
  - Posee buenos paquetes para el manejo de datos y ML.
- **Contras**
  - Seguridad de tipos: Python es un lenguaje de tipo dinámico.
  - Para fines específicos de análisis estadístico y de datos, la amplia gama de paquetes de R le da una ligera ventaja sobre Python. Para los lenguajes de propósito general, hay alternativas más rápidas y seguras a Python.
- **Veredicto** - "excelente todo terreno"
  - Python es una muy buena opción de lenguaje para la ciencia de datos, y no solo en el nivel de entrada. Gran parte del proceso de ciencia de datos gira en torno al proceso ETL (extracción-transformación-carga). Esto hace que la generalidad de Python sea ideal. Las bibliotecas como TensorFlow de Google o Keras hacen de Python un lenguaje muy útil para el aprendizaje automático.

Business Analytics

Que es la Analítica?

- Es el descubrimiento e interpretación de los patrones del comportamiento de los datos empresariales por medio de las herramientas:
  - Estadística
  - Matemática
  - Informática
  - Ciencia de los datos

Business Analytics y Big Data



Inteligencia Empresarial

Sistemas de Información de la Analítica

Campos de Aplicación

Modelos de la Analítica

- Analítica Descriptiva
- Analítica de Prescriptivo
- Analítica Predictivo

Business Analytics: Descriptivo

- Analítica Descriptiva : Estudian eventos pasados.
- donde?
- Cuando?

Business Analytics: Predictivo

- Analítica Predictiva : Estudian eventos probables en el futuro.
- Que pasaría si?

Business Analytics: Prescriptivo

- Analítica prescriptiva o diagnostico: Estudian eventos pasados.
- Porque sucedió?
- Cuales fueron las condiciones?

Video

https://www.youtube.com/watch?v=-ZuqWq25YFg&feature=youtu.be

Laboratorio

- Lab 3 Programación en Lenguaje R
- Lab 4 Programación con Python