Análisis de Grandes Volúmenes de Datos

POSGRADOS

Presentación del Curso

Agenda

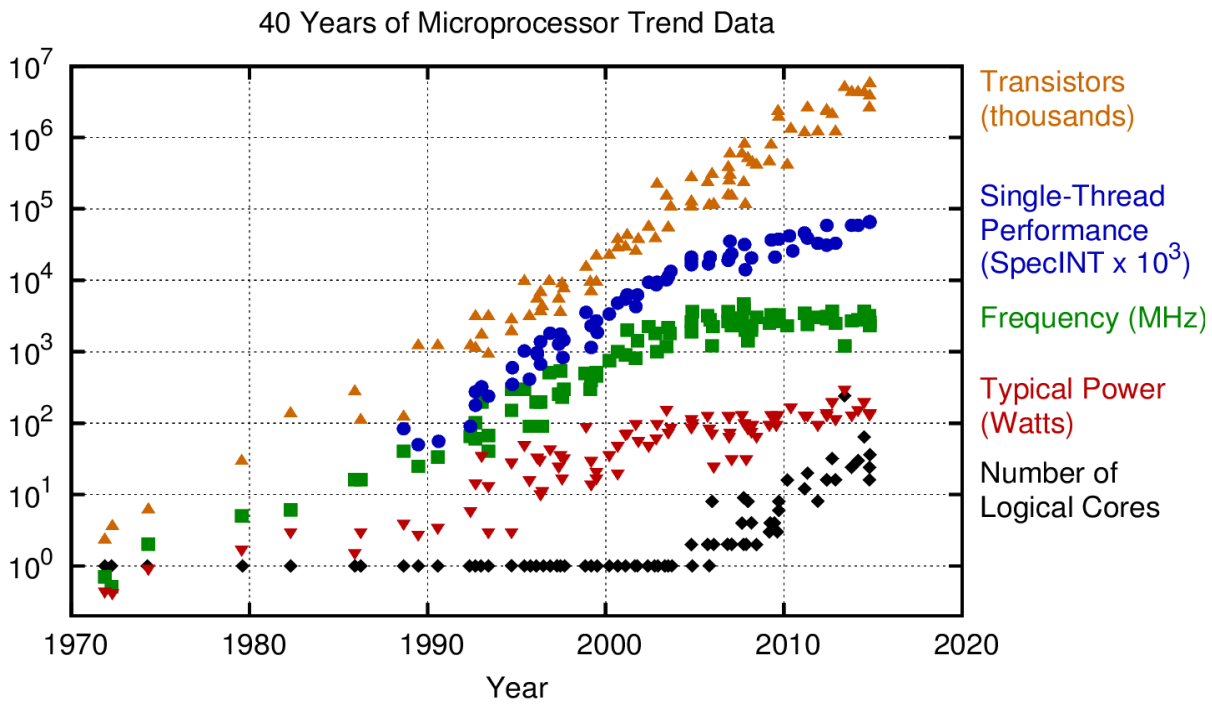
1. Antecedentes
2. Grandes volúmenes de datos: Big Data
3. Plataformas
4. PySpark
5. Conclusiones

Antecedentes

Evolución de las Computadoras

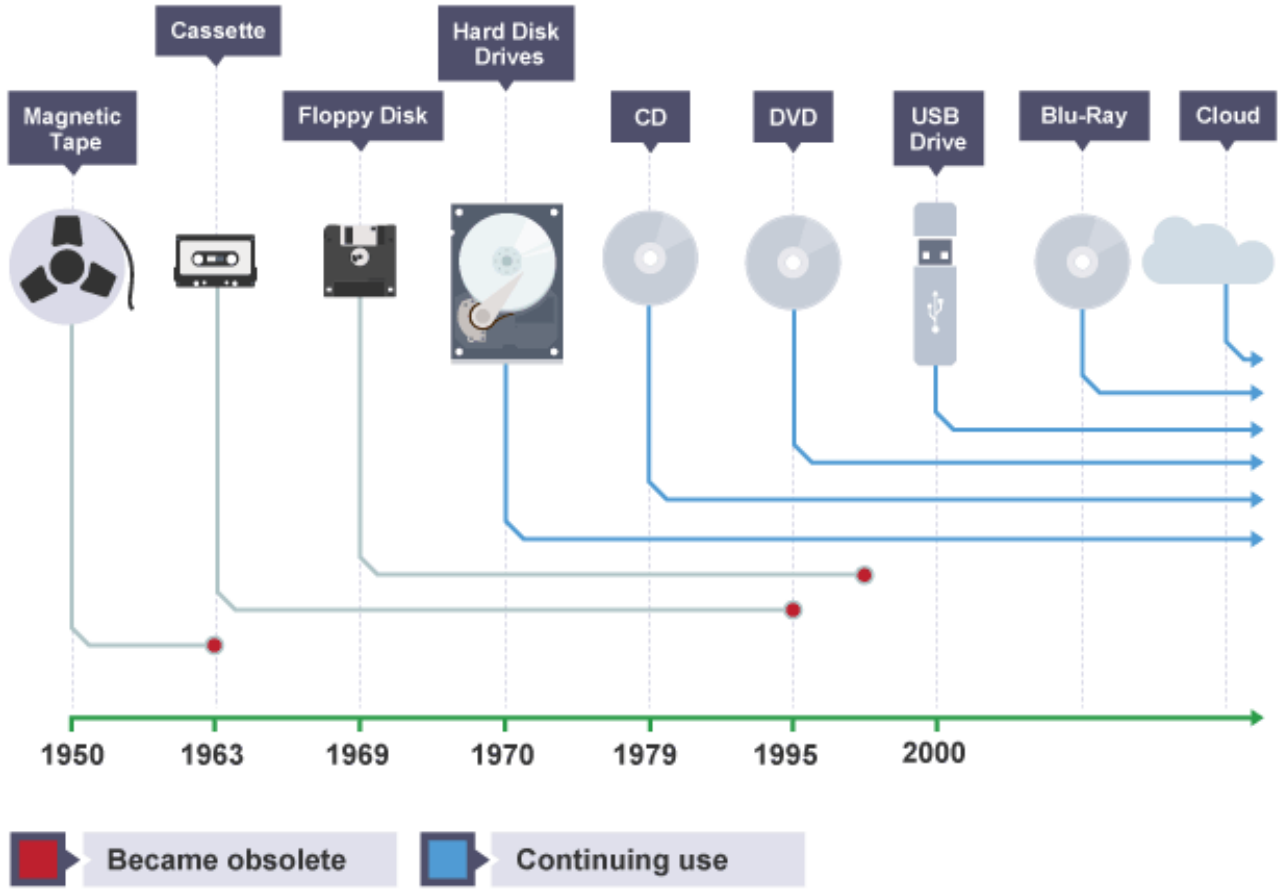
- Cuando se habla de la capacidad de las computadoras, existen típicamente dos dimensiones a considerar
  - Capacidad de procesamiento: velocidad con la cual son procesados datos en un procesador. Se usan diferentes métricas como velocidad de reloj (Hz), operaciones de punto flotante (Flops), operaciones en memoria por segundo (Mops)
  - Capacidad de almacenamiento: capacidad para almacenar (en memoria volátil o no volátil) datos. La unidad de medida son los bytes (Kbytes, Mbytes, Gbytes, Tbytes, …)

Evolución en el poder de procesamiento



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

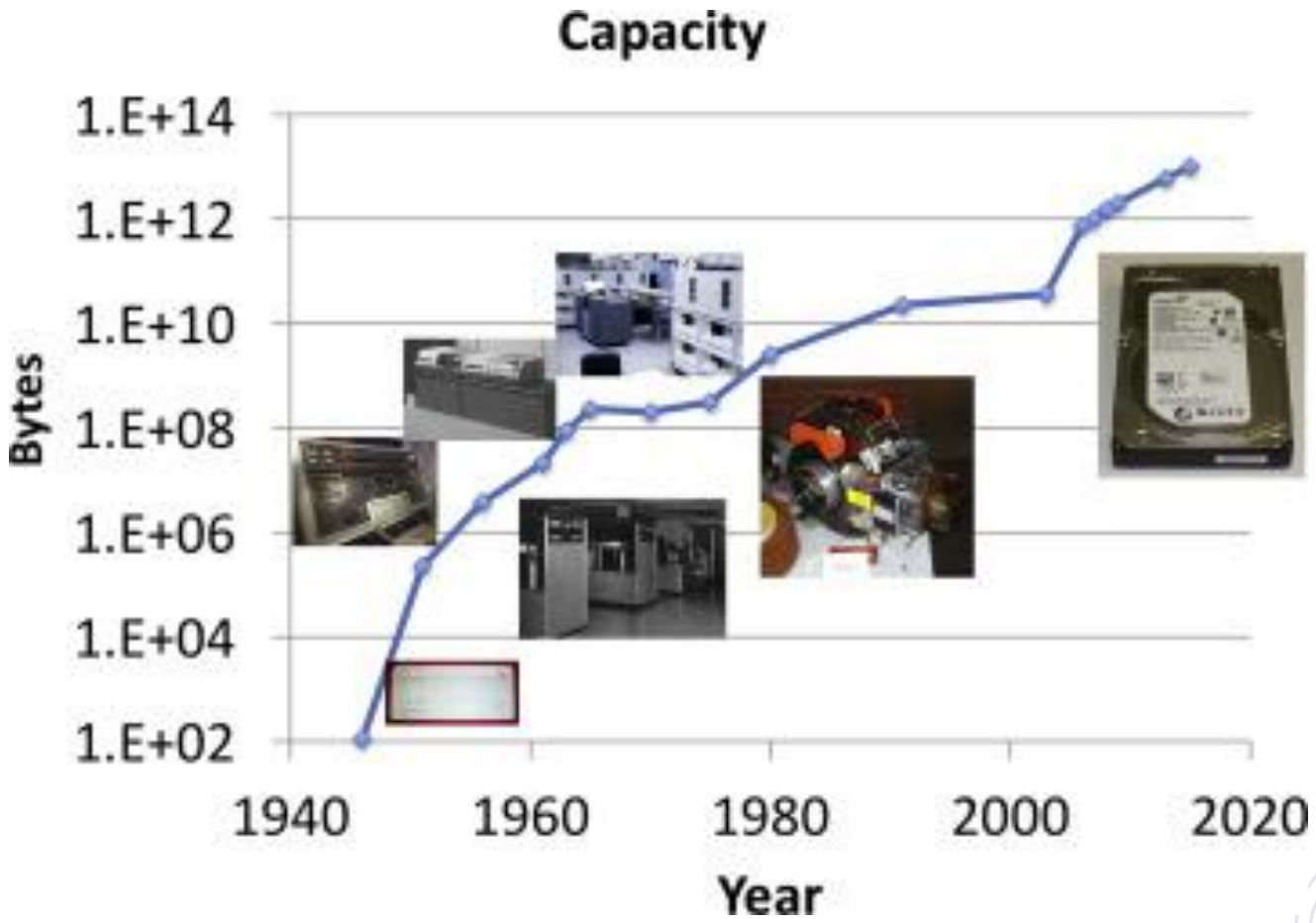
Evolución en la capacidad de almacenamiento (tecnología)



Evolución en la capacidad de almacenamiento (volumen)

Data quantity	In bytes
Kilobyte	1,024
Megabyte	1,048,576
Gigabyte	1,073,741,824
Terabyte	1,099,511,627,776
Petabyte	1,125,899,906,842,624
Exabyte	1,152,921,504,606,846,976
Zettabyte	1,180,591,620,717,411,303,424
Yottabyte	1,208,925,819,614,629,174,706,176

Evolución en la capacidad de almacenamiento (volumen)



Grandes Volúmenes de Datos

Grandes Volúmenes de Datos

Históricamente, ha existido el reto de procesar grandes volúmenes de datos a partir de los sistemas de cómputo existentes, pero ¿que significa un gran volumen de datos?

Dataset type	Fits in RAM?	Fits on local disk?
Small dataset	Yes	Yes
Medium dataset	No	Yes
Big dataset	No	No

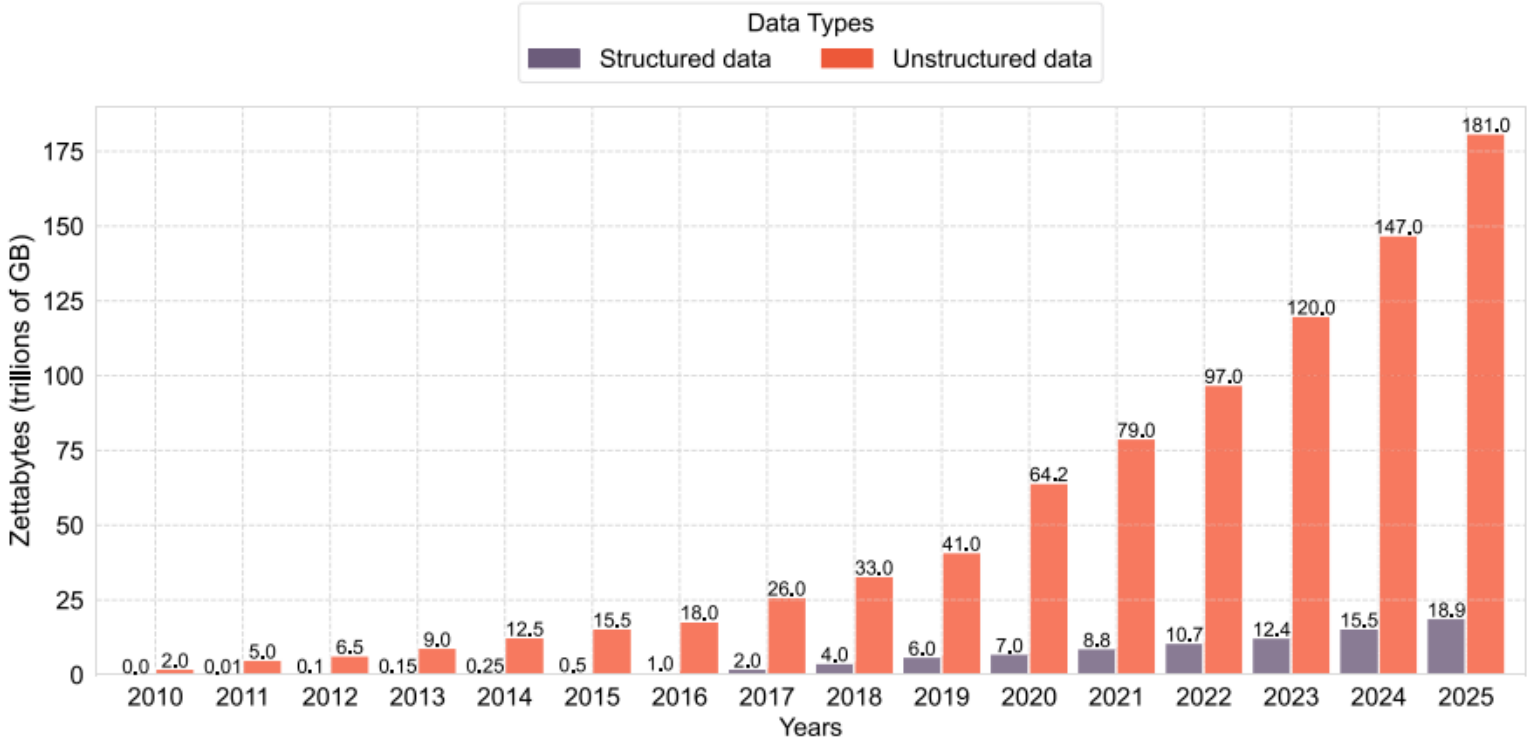
Big Data

- El término Big Data surge en el contexto de la astronomía y genética: cuando la cantidad de datos no puede ser almacenada y/o procesada en un sistema de cómputo
- Existen diferentes autores que en diferentes contextos han hablado del término “big data”, sin embargo en 2001, Doug Laney describió a este concepto a partir de tres características esenciales : volumen, velocidad y variedad (las 3 Vs del Big Data)
  - Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

Las tres V’s

1. **Volumen**: Se refiere a la cantidad masiva de datos que se generan constantemente. Por ejemplo, datos provenientes de redes sociales, sensores, transacciones, etc.
2. **Variedad**: Los datos pueden ser estructurados (como bases de datos), semiestructurados (como XML o JSON) o no estructurados (como imágenes o videos). La diversidad de formatos y fuentes es un desafío para el procesamiento.
3. **Velocidad**: Los datos se generan y deben procesarse rápidamente para tomar decisiones en tiempo real, como en el análisis de tendencias o la detección de fraudes.

Crecimiento de datos estructurados / no estructurados



Retos en Big Data

- Gestión de los datos: procedimientos efectivos para el almacenamiento, procesamiento y recuperación
- Diversidad de los datos provoca retos en su integración, control de calidad en los datos. Además, lo anterior genera la necesidad de contar con especialistas en técnicas para los diferentes formatos de datos (estructurados, no estructurados), investigación para el uso de algoritmos optimizados para la limpieza, selección de datos, el procesamiento para identificar patrones de interés
- Velocidad con la cual se generan los datos
- Infraestructura: empleo de tecnologías avanzadas como el cómputo distribuido para lograr escalabilidad
- Seguridad en la administración de los datos, acceso a los datos

Plataformas para Big Data

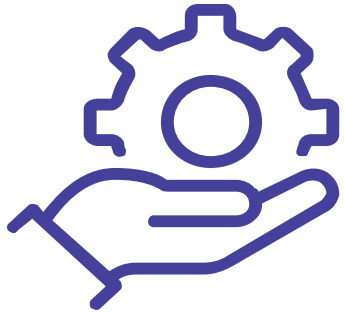
Plataformas para el Big Data

- Las tecnologías para la administración del Big Data se soportan en diferentes tecnologías, las cuales se implementan en infraestructura propia o a partir de servicios de empresas del sector tecnológico:
  - Amazon EMR (antes Elastic MapReduce)
  - Cloudera Data Platform
  - Google Cloud Dataproc
  - HPE Ezmeral Data Fabric (formerly MapR Data Platform)
  - Microsoft Azure HDInsight
  - IBM Cloud

Ventajas y desventajas

- Ventajas
  - Infraestructura probada
  - Soporte técnico
  - Escalabilidad *
  - Herramientas diversas

- Desventajas
  - Costo económico



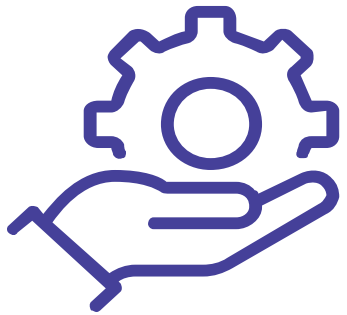
Infraestructura propietaria

- Para las entidades que desean administrar sus propios sistemas, se puede optar por la instalación de herramientas tecnológicas a partir de la implementación de clusters de computadoras (procesamiento distribuido / paralelo)
  - Apache Hadoop: Es un framework de código abierto que permite el procesamiento distribuido de grandes volúmenes de datos. Es ideal para análisis en lote y es ampliamente utilizado por empresas como Facebook
  - Apache Spark: Una herramienta poderosa para el procesamiento en tiempo real y el análisis de datos. Es conocida por su velocidad y capacidad de manejar datos en memoria

Ventajas y desventajas

- Ventajas
  - Control sobre la infraestructura física
  - Diseño “ad hoc”
  - Sin dependencia de entidades externas

- Desventajas
  - Costo en inversión de infraestructura
  - Costo en personal especializado
  - Tiempo de implementación
  - Escalabilidad *



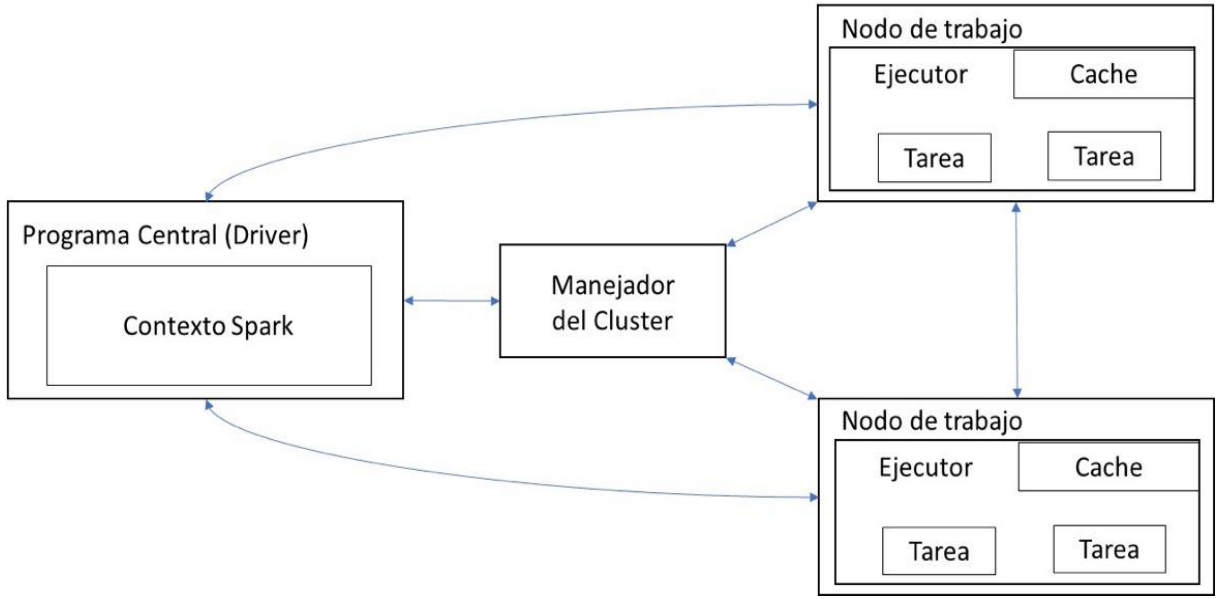
PySpark

Spark

- Biblioteca de código abierto diseñada para el procesamiento de grandes volúmenes de datos, con una alta optimización para ser eficiente y veloz
  - Soporte de múltiples lenguajes de programación (Python, Java, Scala, R)
  - Implementa el soporte de estructura de datos distribuidas (dataframes Spark)
  - Soporte de operadores propios o a través de lenguaje de consulta SQL
  - Bibliotecas de aprendizaje máquina para el cómputo distribuido

Arquitectura Spark

- Implementa una arquitectura jerárquica “maestro – esclavo”



Elementos clave de Spark

- Spark driver: proceso dedicado que corre en el nodo principal (driver machine), que convierte el código del usuario en diversas tareas que son distribuidas entre los workers
- Executor: proceso lanzado por una aplicación Spark en un nodo Worker
- Worker Node: encargado de la ejecución de programas. Múltiples executors pueden correr en un Worker Node
- Cluster Manager: encargado de organizar el sistema distribuido, Asignadndo Executors a Worker Node, recursos, comunicación, etc.

Spark Core

- Es la base de ejecución para todos los procesamientos de datos, ofreciendo cómputo eficiente en memoria, permite referenciar las bases de datos, entre otras tareas:
  - Administración de memoria y recuperación de fallas
  - Planificación, distribución y monitoreo de tareas en el cluster
  - Interacción con sistemas de almacenamiento
  - Abstracción de datos
  - Soporta módulos orientados al aprendizaje máquina en Big Data (ML, MLlib)

Estructuras de Datos en Spark

- RDD
  - Los RDD (Resilient Distributed Datasets) son conjuntos de datos resilientes distribuidos, almacenados en memoria con tolerancia a fallos, que pueden ser distribuidos entre múltiples nodos, que se pueden procesar en paralelo
  - Abstracción más básica de Spark, introducida en las primeras versiones
  - No contiene información estructural sobre los datos (Spark no conoce los nombres de las columnas ni los tipos de datos)
  - Colecciones de datos inmutables

Estructuras de Datos en Spark

- Dataframe
  - Introducida a partir de la versión 1.3 (2015)
  - Son colecciones de datos organizados por columnas, ofreciendo operadores como filtrado, agrupación, entre otros
  - Más flexible y eficiente que los RDDs, debido a su capacidad para aprovechar optimizaciones internas
  - Es el estándar que hoy se usa para el uso de modelos de aprendizaje máquina

Spark SQL

- Módulo de Apache Spark diseñado para el trabajo con datos estructurados, permitiendo el soporte de consultas SQL directamente sobre datos distribuidos

  - Compatibilidad con SQL estándar
  - Integración con Dataframes
  - Optimización de consultas
  - Compatibilidad con múltiples fuentes de datos

Aprendizaje Máquina en Spark

- ML/MLIB son las bibliotecas dónde se implementan algoritmos de aprendizaje supervisado / no supervisado con soporte para cómputo distribuido en Spark
  - MLIB opera sobre RDDs
  - ML opera sobre Spark Dataframes
- Contiene algoritmos de:
  - Clasificación
  - Regresión
  - Clustering

Configuración

- Para trabajar con Spark, se puede optar por una:
  - Instalación local (por ejemplo, corriendo sobre Anaconda)
  - Google Colab

Instalación local

- Se deben de instalar una serie de bibliotecas
  - Anaconda (creando un entorno de trabajo)
  - OpenJDK
  - PySpark
  - FindSpark
  - Ipykernel

- Nota: el proceso de instalación se encuentra detallado en “Manual_de_instalacion_pyspark.pdf” (disponible en la plataforma)

Ejecución sobre Google Colab

```
[1] !apt-get install openjdk-8-jdk-headless -qq > /dev/null
    !wget -q http://archive.apache.org/dist/spark/spark-3.1.1/spark-3.1.1-bin-hadoop3.2.tgz
    !tar xf spark-3.1.1-bin-hadoop3.2.tgz
    !pip install -q findspark
```

```
[2] import os
    os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
    os.environ["SPARK_HOME"] = "/content/spark-3.1.1-bin-hadoop3.2"
```

```
[3] !ls
```

```
sample_data  spark-3.1.1-bin-hadoop3.2  spark-3.1.1-bin-hadoop3.2.tgz
```

```
import findspark
findspark.init()
from pyspark import SparkContext, SparkConf, SQLContext
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[*]").getOrCreate()
spark.conf.set("spark.sql.repl.eagerEval.enabled", True) # Property used to format output tables better
spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
v3.1.1
Master
local[*]
AppName
pyspark-shell

Conclusiones

Conclusiones Generales

- El Big Data es un área de investigación / desarrollo orientado al manejo y procesamiento de grandes volúmenes de datos
- Existen diferentes retos que se tiene que afrontar al trabajar en Big Data
- Existen plataformas propietarias / libres, que soportan el Big Data, cada una con ventajas y desventajas
- Spark es un framework gratuito que permite el trabajo en Big Data, con una alta optimización para el manejo de operaciones en cluster de computadoras

Tecnológico de Monterrey