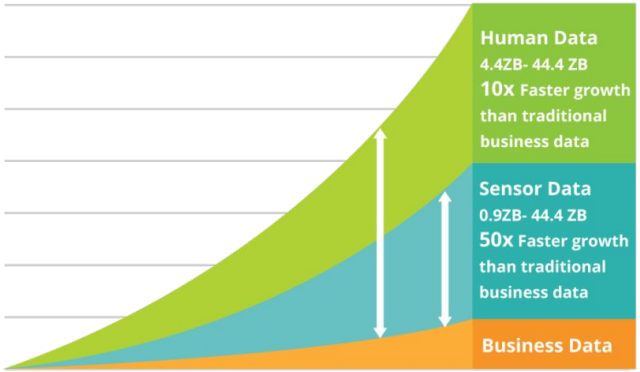
Doing Data Analytics I

24 Dec 2022

Introduction to Big Data

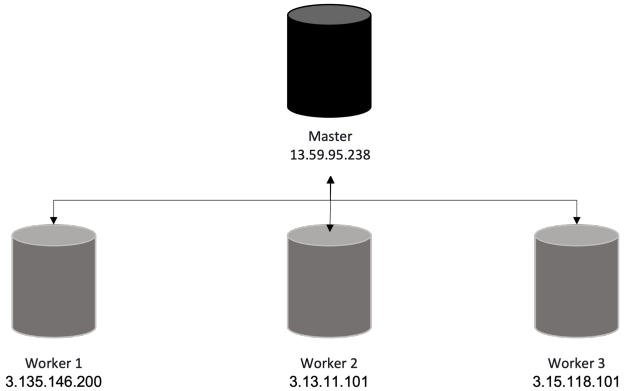
- With the development of e-commerce, businesses could capture not only the sales transactions but also the customer behaviour data.
- Customer behaviour data includes products viewed and added to the cart with product attributes such as brand, price and category etc.
- Since this capture data points at multiple touch points, the volume of data collected can be significantly larger than the sales transactions.
- For example, a multi category e-commerce website could generate sales transactions data of GB level but customer behaviour data of TB or even PB level.
- We'll be using customer behaviour data of an e-commerce website in the month of October in 2019.
- The size of this dataset is ~5GB which can't be handled using our R Studio or Co-Lab IDEs.



Source: Inside big data

Introduction to Distributed Computing

- When you have a data set which can not be processed or analysed using a **single machine**, you can consider it as **big data**.
- At this point, you need big data tools and technologies to process and analyse data.
- Today the main technology behind the big data tools is **Distributed Computing**.
- Distributed computing is the method of making **multiple computers** work together to solve a common problem.
- It makes a computer network appear as a powerful single computer that provides large-scale resources to deal with complex challenges.



Distributed Big Data Technologies

Data companies such as Google, Yahoo, Facebook and Amazon adopted distributed computing technology to deal with big data.

Currently, there are several distributed big data tools used for big data use cases.

This session discusses three distributed big data platforms developed by Apache (www.apache.org)

1. **Spark**
2. **Hive**
3. **Hadoop MapReduce**

Apache Spark

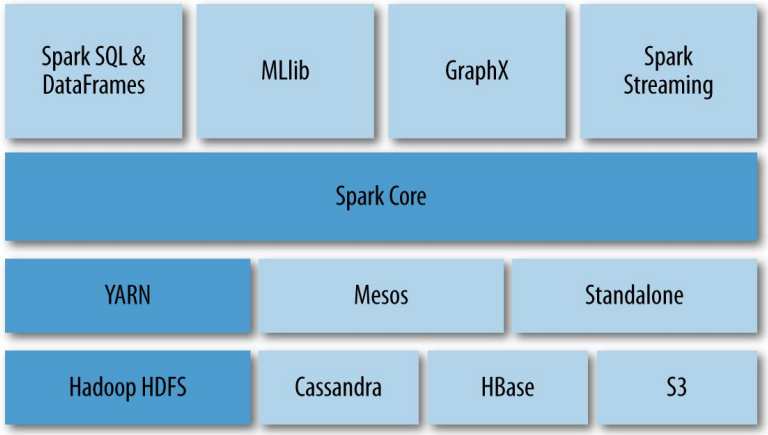
Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

Spark is a fast in-memory data processing engine.

It has batch processing, streaming and SQL support to handle big data

Spark requires a resource manager such as YARN, Mesos or Kubernetes

Spark integrates with multiple distributed file systems such as HDFS, S3, Cassandra, Mongodb and HBase etc.



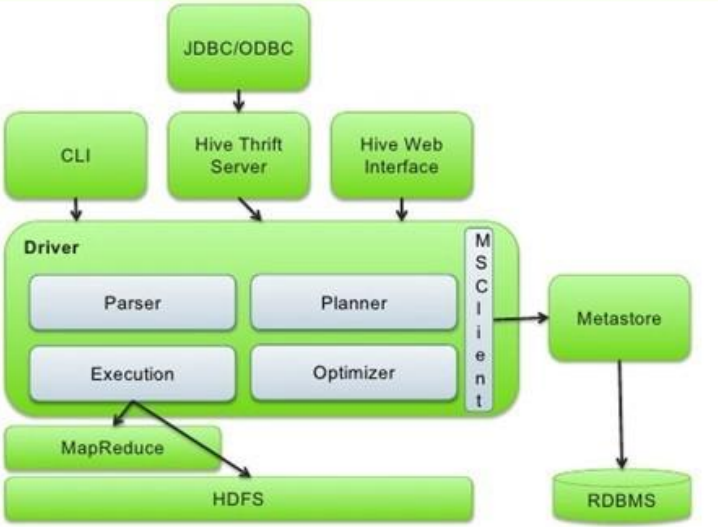
Apache Hive

The Apache Hive™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.

Structure can be projected onto data already in storage.

A command line tool and JDBC driver are provided to connect users to Hive.

A user can connect to a Hive server using JDBC driver similar to connecting to any other RDBMS such as MySQL, SQL Server or PostgreSQL.



Questions & Answers

Use your Google Classroom stream to post any questions or start discussions.
https://classroom.google.com/c/NDk3MzI3NzgxNDM5?cjc=ltzuypk