Garanti BBVA Teknoloji

Introduction to Big Data

Altemur Celikayar – Mert Bilgin

Kasım 2023

Garanti BBVA Teknoloji

Content

Garanti BBVA Teknoloji

01

What is Big Data?

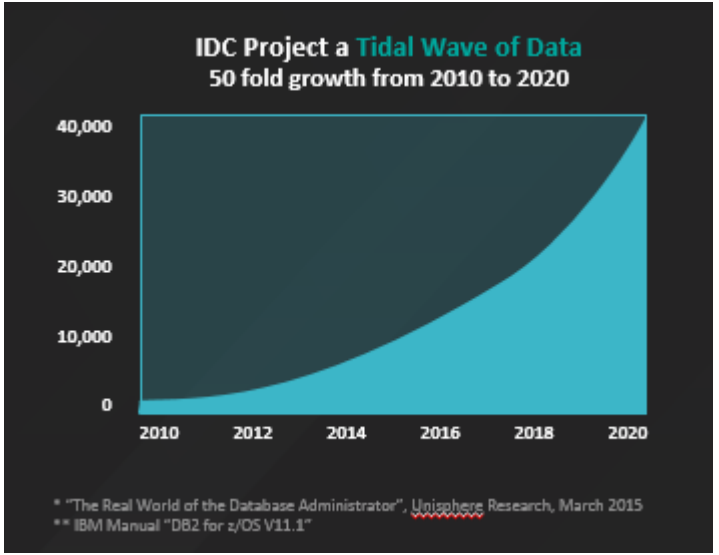
Ekim 2022

What is Big Data?

History of Big Data

"The World is one big data problem. " Andrew McAfee

- The term ‘Big Data’ has been in use since the early 1990s
- 4.4 zettabytes data in 2013 in the world.
- 44 zettabytes by 2020 (44 trillion gigabytes)



1 zettabytes = 1.000.000.000.000.000.000.000 bytes

What is Big Data?

History of Big Data

In 2010 Eric Schmidt at the Techonomy conference in Lake Tahoe in California and he states that "there were 5 exabytes of information created by the entire world between the dawn of civilization and 2003. Now that same amount is created every two days."

The ancient Egyptians around 300 BC already tried to capture all existing ‘data’ in the library of Alexandria.

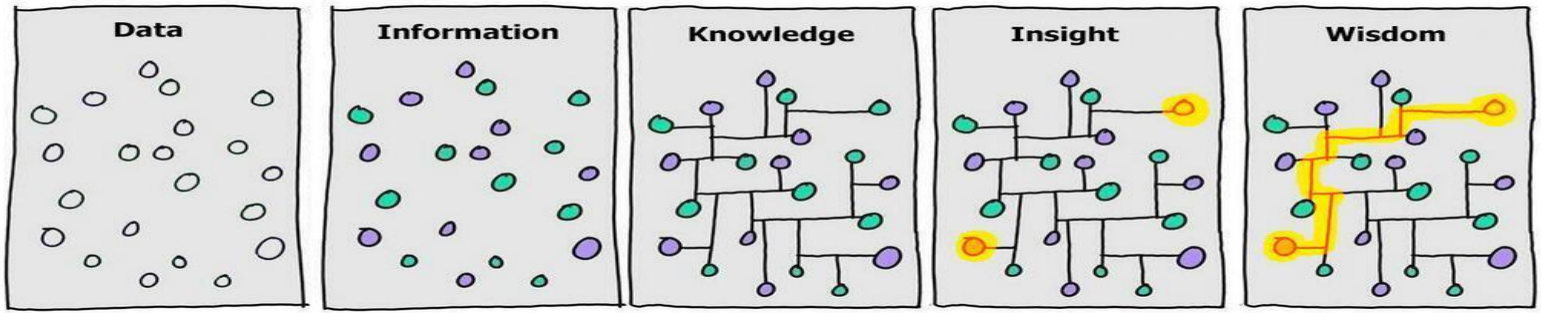
The Roman Empire used to carefully analyze statistics of their military to determine the optimal distribution for their armies.

What is Big Data?

Why Big Data?

More efficient. Higher profits. Happier customer!

- Impossible to manage and process using traditional business intelligence tools
- Also not fast enough!
- Gaining insights



What is Big Data?

Comparision

	TRADITIONAL DATA	BIG DATA
01.	Traditional data is generated in enterprise level.	Big data is generated in outside and enterprise level.
02.	Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
03.	Traditional database system deals with structured data.	Big data system deals with structured, semi structured and unstructured data.
04.	Traditional data is generated per hour or per day or more.	But big data is generated more frequently mainly per seconds.
05.	Traditional data source is centralized and it is managed in centralized form.	Big data source is distributed and it is managed in distributed form.

What is Big Data?

Industries & Benefits

Education

- Career prediction
- Reframing course material
- Reducing dropout rates

Insurance

- Gaining customer insight
- Fraud detection

Healthcare

- Early detection of diseases
- Pandemic prediction

Banking

- The misuse of credit cards
- Money laundering
- Risk Mitigation

Media

- Predictive data analysis
- Ads more clear

Retail

- Future purchase prediction
- Demand forecasting

What is Big Data?

In conclusion

Two main problems;

- Storing the data
- Processing the data

What is Big Data?

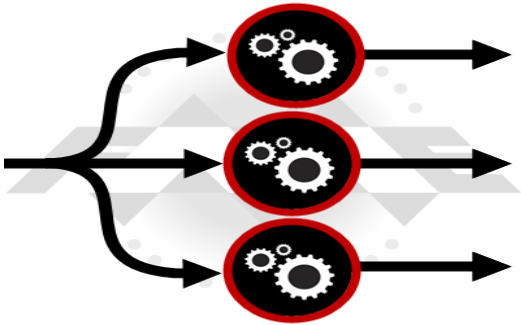
4Vs

Those needs need to deal with a concept. Doug Laney articulated definition of big data as the four V's

- Volume: Size of the data
- Veracity: Quality of the data
- Variety: From different resources
- Velocity: Time is of the essence

What is Big Data?

Technical Approach



- 1 strong worker or 100 weaker worker.
- 1 huge computer or many small computers?
- 1 Node = 100 Computers
- Reliability = Distribute the data into several pieces and dublicate them
- Parallel processing

Garanti BBVA Teknoloji

02

What is Hadoop?

What is Hadoop?

Hadoop History

- Hadoop was created by Yahoo! in 2005
- Named the project after his son's toy elephant.
- Has many modules

What is Hadoop?

Hadoop

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

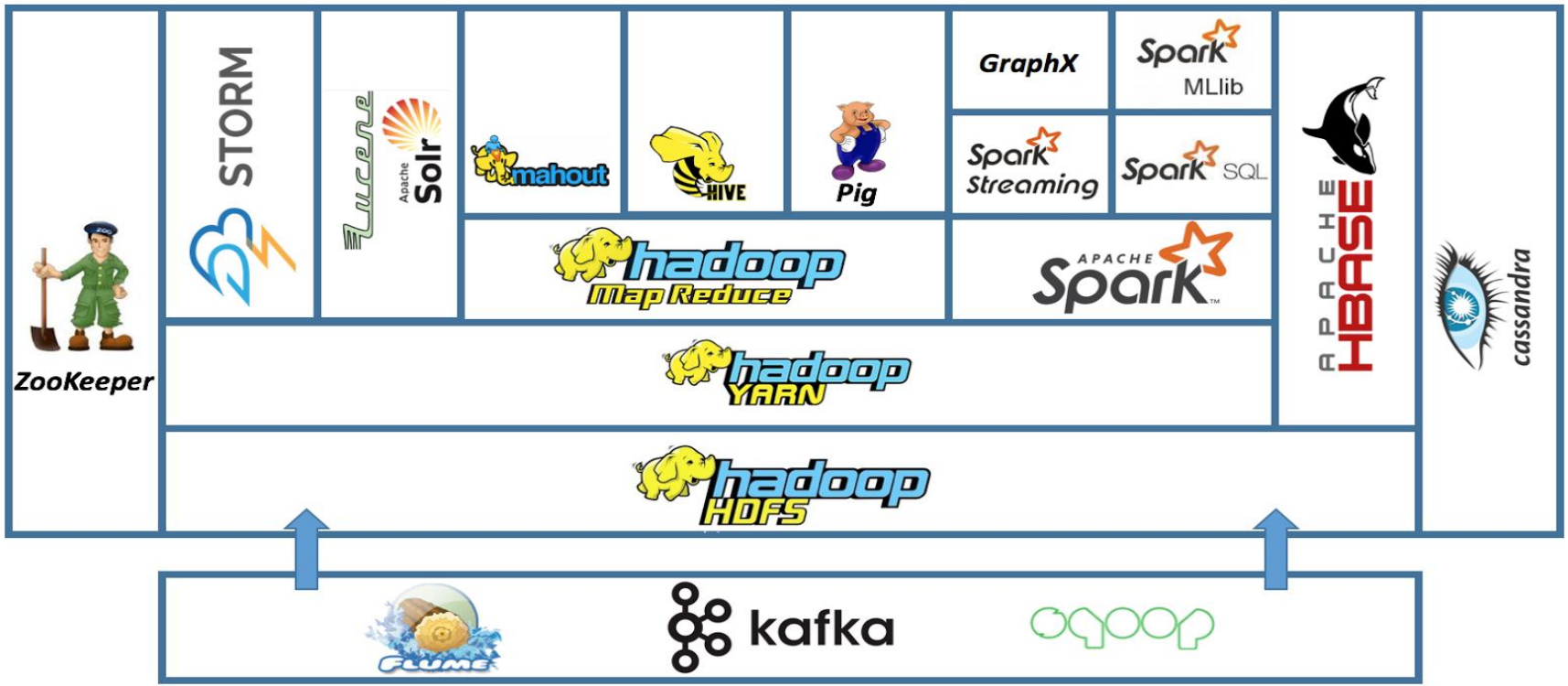
Garanti BBVA Teknoloji

03

Hadoop Ecosystem

Ekim 2022

Hadoop Ecosystem



Garanti BBVA Teknoloji

04

Hadoop's Core

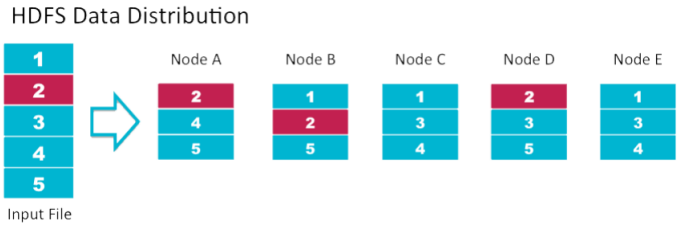
Ekim 2022

Hadoop's Core

HDFS

HDFS is a distributed file system that handles large data sets running on commodity hardware.

- Hadoop Distributed File System
- Hardware Failure
- Huge Datasets
- Data Replication

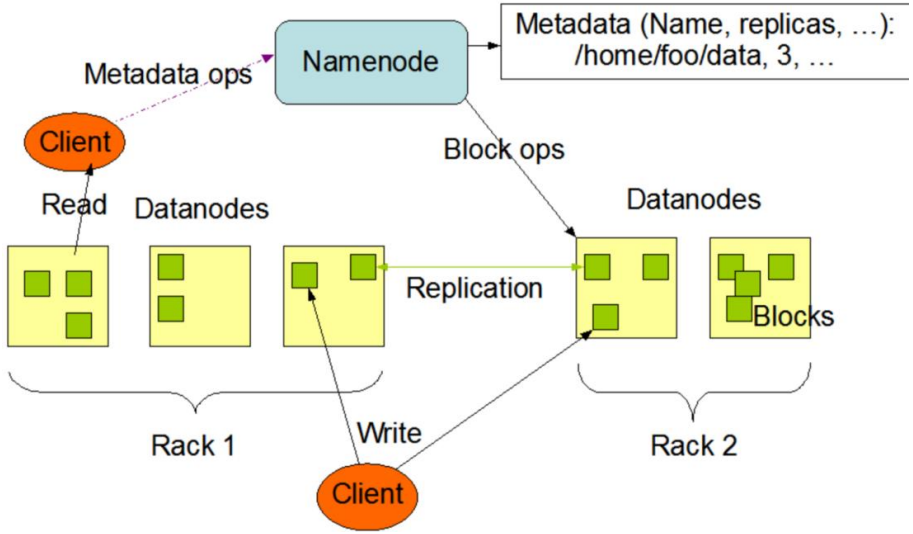


Hadoop's Core

How does HDFS work?

- Data is splitted into seperate blocks.
- Those blocks are distributed among the nodes.
- All blocks are replicated
- HDFS has chunks (default 128 MB)
- Nodes have a Master (NameNode) and Slaves (DataNodes)

HDFS Architecture



Hadoop's Core

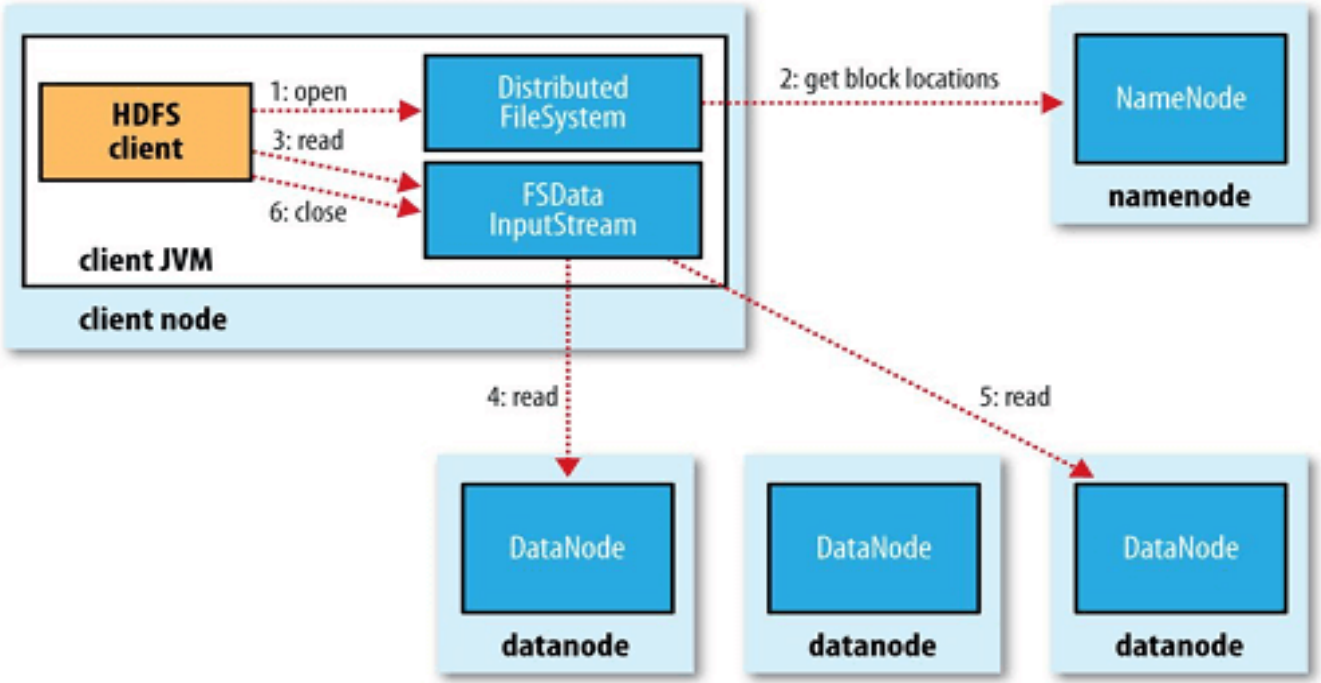
Namenode

- Also know as Master
- Stores Metadata of Actual Data
- Assign work to Namenode

Datanode

- Also know as Slave
- Stores Actual Data
- Performs Actions

Hadoop's Core



Hadoop's Core

MapReduce

Parallel execution programming paradigm that enables massive scalability across hundreds or thousands of servers.

- Processing Layer of Hadoop
- Input -> Output
- Divides into tasks

Hadoop's Core

Map

- A function defined by user
- Takes Key/Value Pair as Input

  Key = Reference

  Value = Data

Reduce

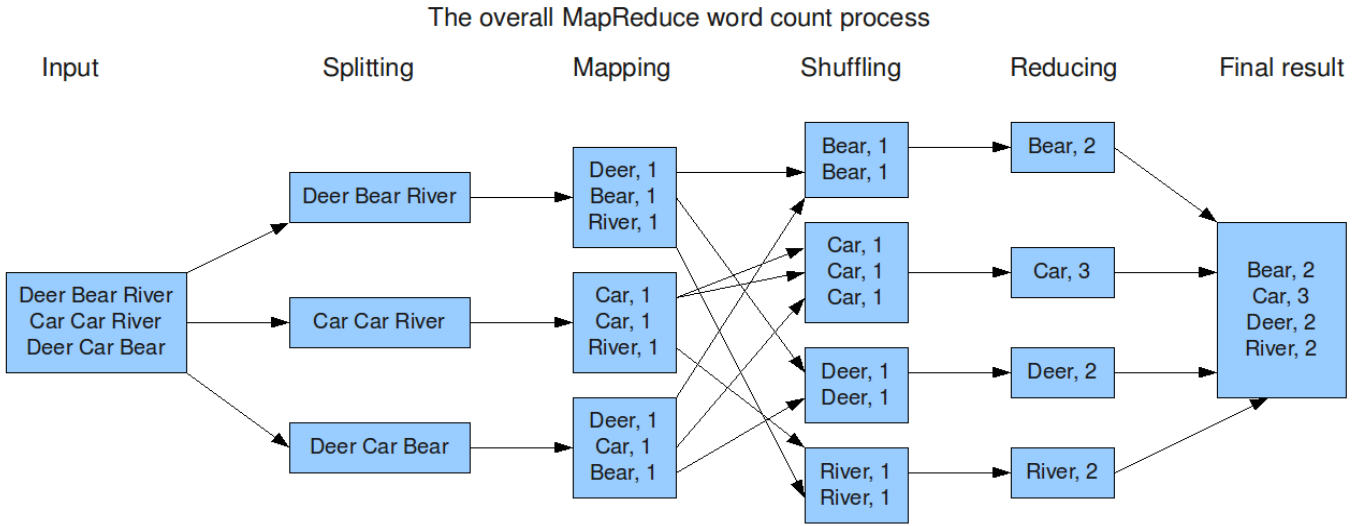
- A function defined by user
- Input <- Map
- Reduce by Key

Shuffle

- Map -> Shuffle -> Reduce
- Prepares data for Reducer

Hadoop's Core

MapReduce



Garanti BBVA Teknoloji

05

Managing the Cluster

Ekim 2022

Managing the Cluster

What we need?

- How can I coordinate the cluster?
- How do I estimate resource management?
- We have bunch of applications run in a sequence. How can I queue them?

Managing the Cluster

Zookeeper

ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

- Heart of Distributed System
- Coordination

Managing the Cluster

YARN

Apache Hadoop YARN is the resource management and job scheduling technology

- Yet Another Resource Negotiator
- Brain of Distributed System
- Resource Management

Managing the Cluster

Oozie

Oozie is a workflow scheduler system to manage Apache Hadoop jobs

- Scheduler
- Workflow

Garanti BBVA Teknoloji

06

Storage Level

Ekim 2022

Storage Level

Relational Database

A relational database is a type of database that stores and provides access to data points that are related to one another

- Access Data in Relation
- Table
- Row & Column

name	age	country
Natalia	11	Iceland
Ned	6	New York
Zenas	14	Ireland
Laura	8	Kenya

Storage Level

Hive

- ETL and Data Warehousing Tool
- On top of HDFS
- Dealing with Structured Data
- SQL-inspired language

Storage Level

Non-relational Databases

- Non-tabular Form
- Ideal for Frequently Changes
- Columnar

Key	Document
1001	{ "CustomerID": 99, "OrderItems": [{ "ProductID": 2010, "Quantity": 2, "Cost": 520 }, { "ProductID": 4365, "Quantity": 1, "Cost": 18 }], "OrderDate": "04/01/2017" }
1002	{ "CustomerID": 220, "OrderItems": [{ "ProductID": 1285, "Quantity": 1, "Cost": 120 }], "OrderDate": "05/08/2017" }

Storage Level

HBase

A column-oriented database management system that runs on top of HDFS.

- Not support SQL
- Real-time Read/Write

Garanti BBVA Teknoloji

07

Processing Tools

Ekim 2022

Processing Tools

Brief Instruction

Data is stored in a cluster but needs to be processed.

Batch processing

Data is collected over time

Once data is collected, it's sent for processing

Batch processing is lengthy and is meant for large quantities of information that aren't time-sensitive.

Stream processing

Data streams continuously.

Data is processed piece-by-piece.

Stream processing is fast and is meant for information that's needed immediately.

Batch Processing

Spark

An open-source, distributed processing system used for big data workloads

- Engine for large-scale data process
- Runs on RAM
- Java, Scala, Python

Batch Processing

Sqoop

Apache Sqoop is a command-line interface application used for transferring data between relational databases and Hadoop.

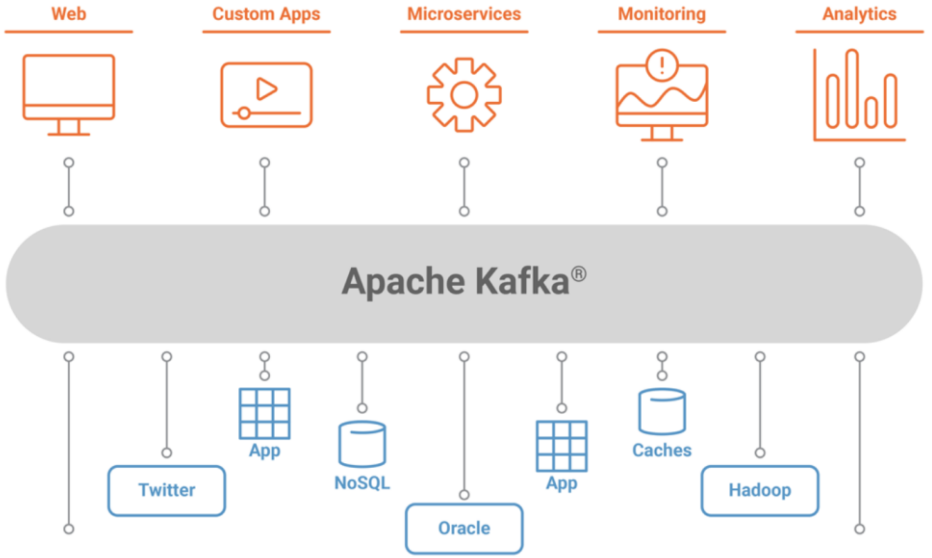
- RDBMS -> Hadoop

Stream Processing

Kafka

Apache Kafka® is an event streaming platform.

- Messaging System
- Distributed Streaming Platform
- Publish-Subscribe



Stream Processing

Spark Streaming

Spark Streaming is an extension of the core Spark API that allows data engineers and data scientists to process real-time data

- Big Data Framework
- RAM