Big Data

What is Big Data ?

- Big Data is defined as data that is huge in size. Bigdata is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.

- The term **Big Data** refers to a dataset which is too **large** or too complex for ordinary computing devices to process. As such, it is relative to the available computing power on the market.

Big Data Classification

1. Volume
2. Velocity
3. Variety
4. Veracity

Volume

- Volume refers to the **amount of data (Size of the data).**
- Today data size has increased to size of terabytes in the form of records or transactions.

Velocity

- The Velocity is the **speed at which the data is created, stored, analyzed and visualized.**
- In the past, when batch processing was common practice, it was normal to receive an update from the database every night or even every week. Computers and servers required substantial time to process the data and update the databases.
- In the big data era, data is created in real-time or near real-time.

Variety

- Variety refers to the many **sources and types** of data. In the past, all data that was created was structured data, it neatly fitted in columns and rows but those days are over.
- Nowadays, 90% of the data that is generated by organizations unstructured data. Data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data.

Veracity

- Big data veracity refers to the biases, noise and abnormalities, ambiguities, latency in data.
- Keep your data clean and processes to keep 'dirty data' from accumulating in your systems.
- Having a lot of data in different volumes coming in at high speed is worthless if that data is incorrect. Incorrect data can cause a lot of problems for organizations as well as for consumers. Therefore, organizations need to ensure that the data is correct as well as the analyses performed on the data are correct. Especially in automated decision-making, where no human is involved anymore, you need to be sure that both the data and the analyses are correct.

Data Types

Structured Data

Any data that can be stored, accessed and processed in the form of fixed format is termed as a structured data. Structured data refers to kinds of data with a high level of organization, such as information in a relational database.

E.g. Relational Data

Semi-structured Data

Semi-structured data is information that doesn’t reside in a relational database but that does have some organizational properties that make it easier to analyze. With some process you can store them in relation database. Examples of semi-structured: XML and JSON documents are semi structured documents, NoSQL databases are considered as semi structured.

Data Types

Unstructured Data

Any data with unknown form or unknown structure is classified as unstructured data. It often includes text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents.

Important 2V's of Big Data

1. **Validity**

The correct data and accurate are intended to use for taking decisions.

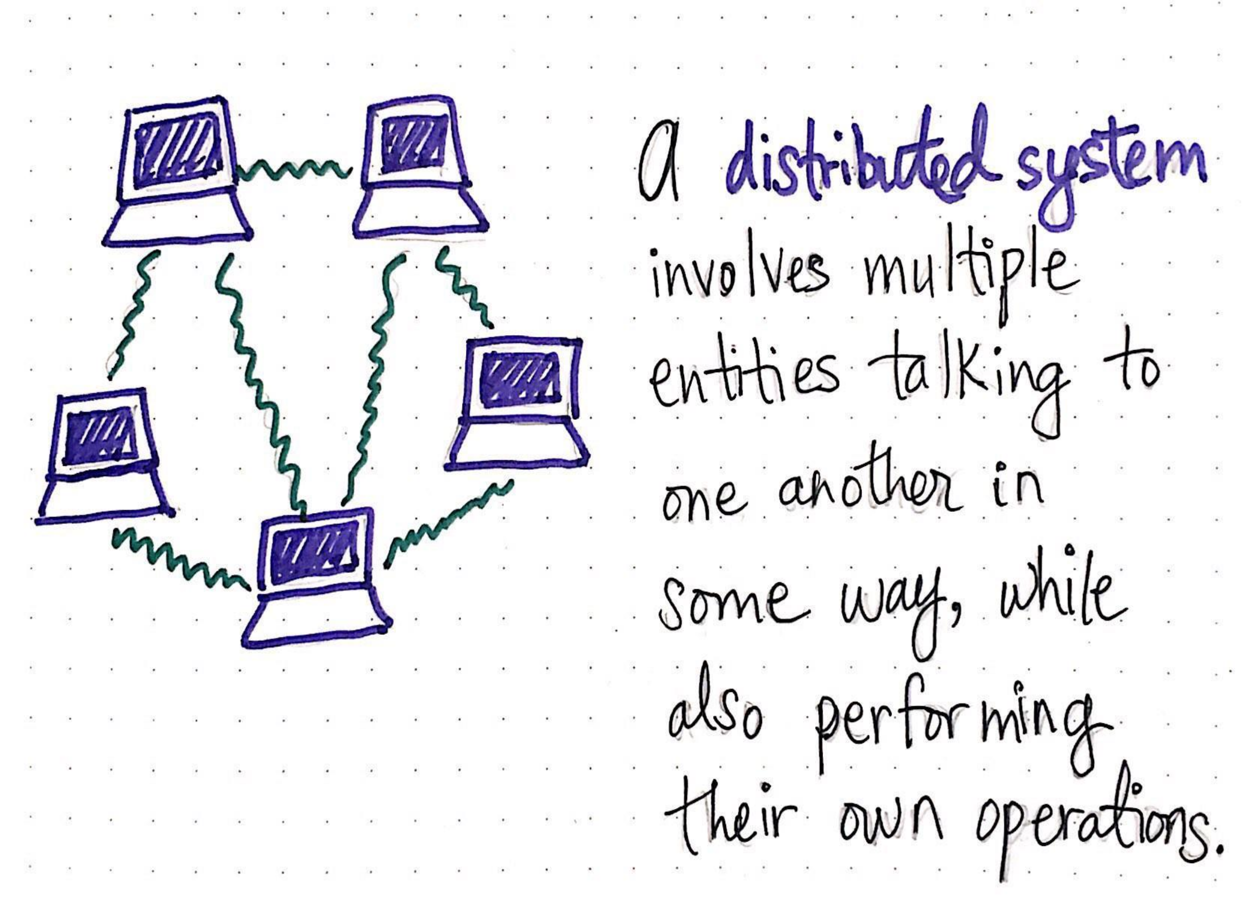
2. **Volatility**

Big data volatility refers to how long is data valid and how long should it be stored. In this world of real-time data, you need to determine at what point the data is no longer relevant to the current analysis.

Distributed System

- A distributed system contains multiple nodes that are physically separate but linked together using the **network**.

- All the nodes in this system communicate with each other and handle processes in tandem. Each of these nodes contains a small part of the distributed operating system software.



a distributed system involves multiple entities talking to one another in some way, while also performing their own operations.

Drawbacks of Single System

- High Chance of System Failure
- Limited Bandwidth
- High Complexity
- High Dependency on Single System
- Not Scalable

Solution for Big Data

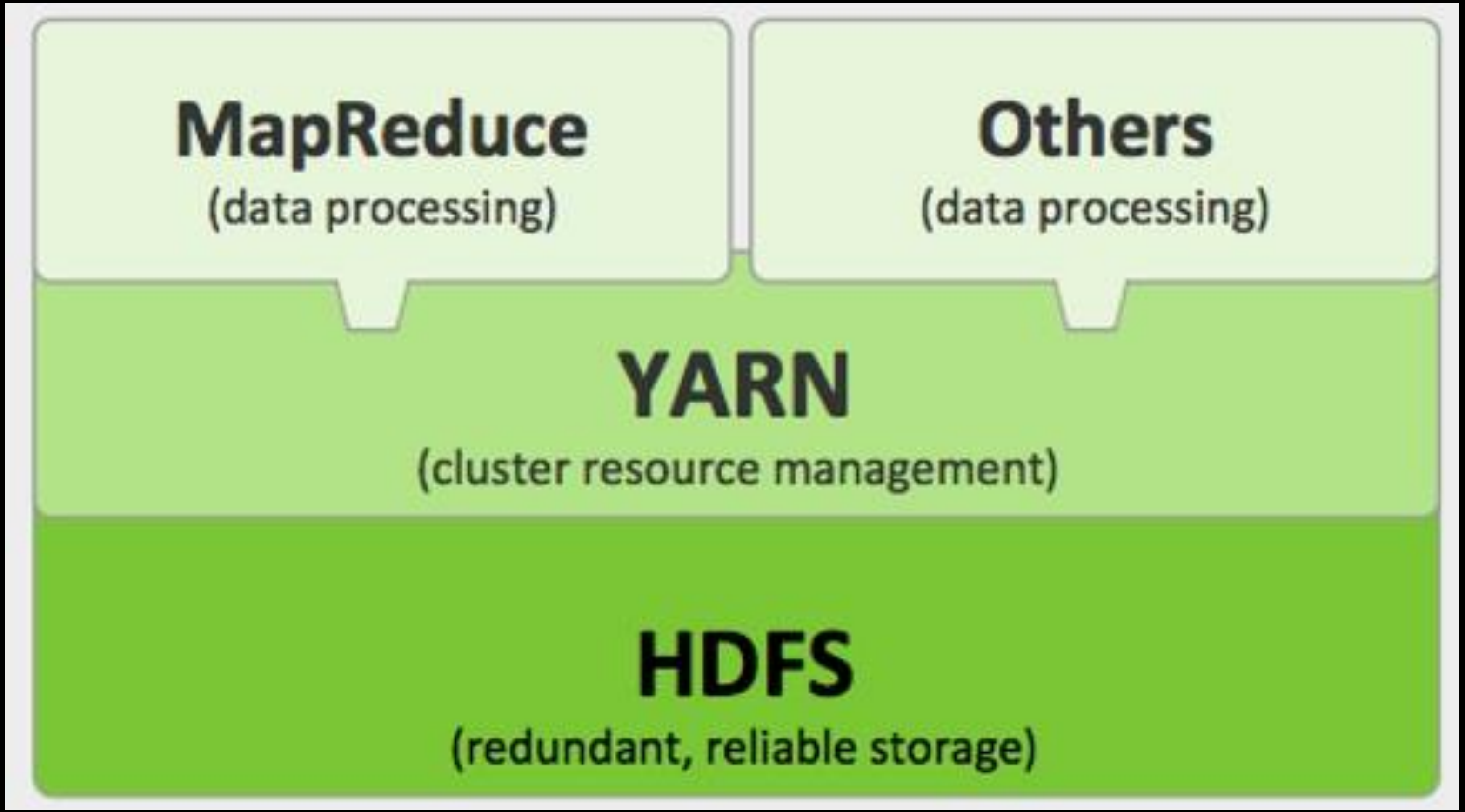
Hadoop (High Availability Distributed Object Oriented Platform)

- Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

Key Characteristics of Hadoop

- System Failure
- Programming Complexity
- Bandwidth
- Scalable
- Stores Multiple Copies
- Reliable

Hadoop Core Components



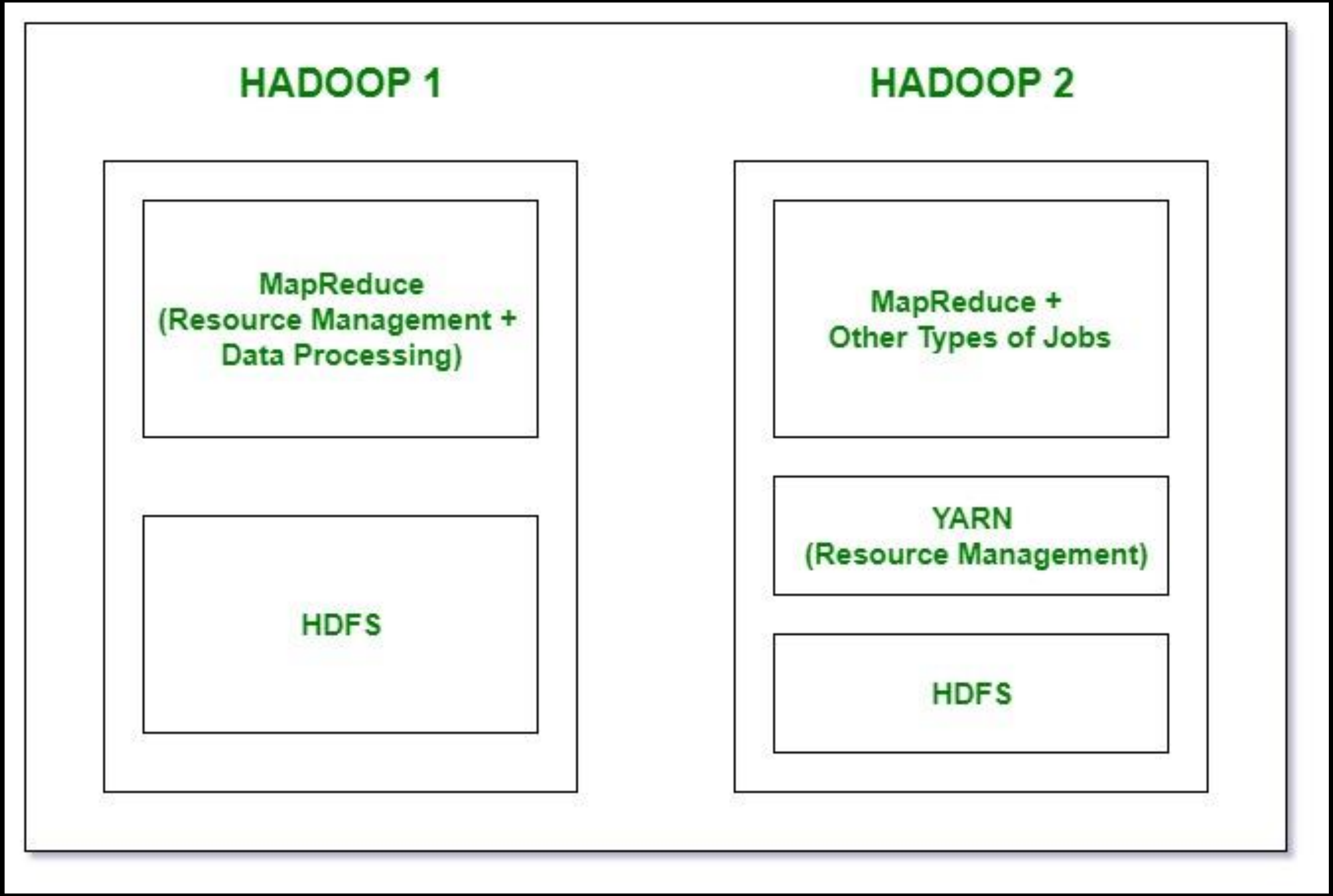
MapReduce
(data processing)

Others
(data processing)

YARN
(cluster resource management)

HDFS
(redundant, reliable storage)

Hadoop 1 vs Hadoop 2



HADOOP 1
MapReduce
(Resource Management +
Data Processing)
HDFS
HADOOP 2
MapReduce +
Other Types of Jobs
YARN
(Resource Management)
HDFS