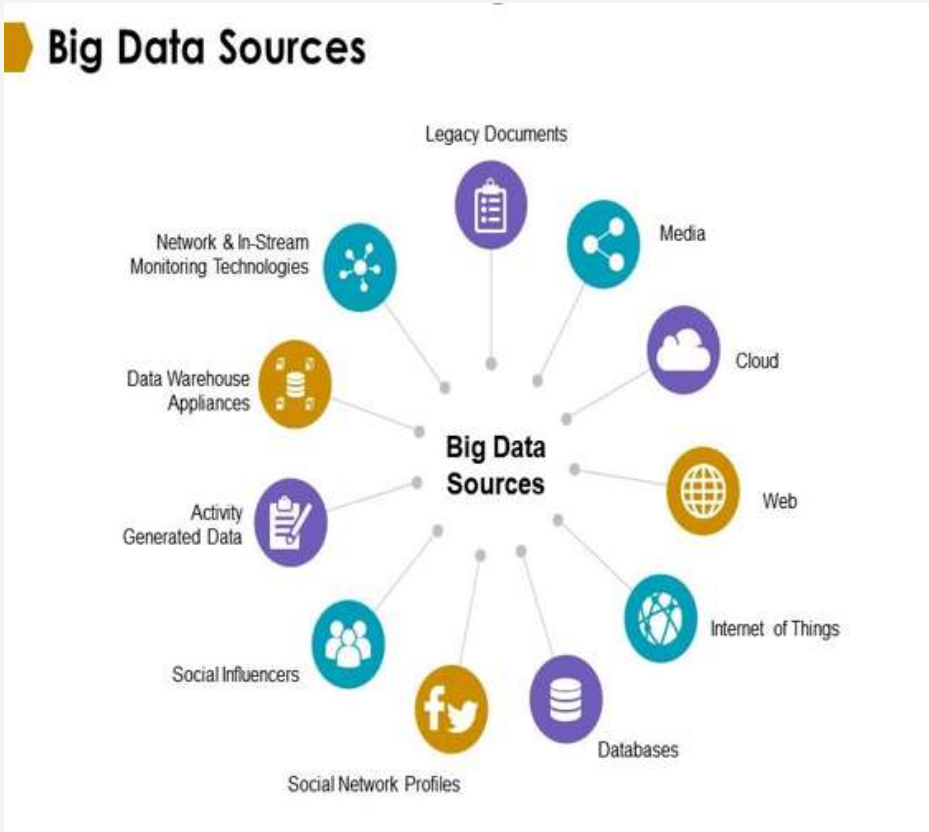
What is Big Data?

- **Big data** is a collection of massive and complex data sets.
- It is about data volume and large data set's measured in terms of terabytes or petabytes.
- There exist large amounts of heterogeneous digital data. Has data management capabilities.
- **Big data analytics** is the process of examining large amounts of data.
- Data volume that include the huge quantities of data ex. Social media analytics and real-time data.

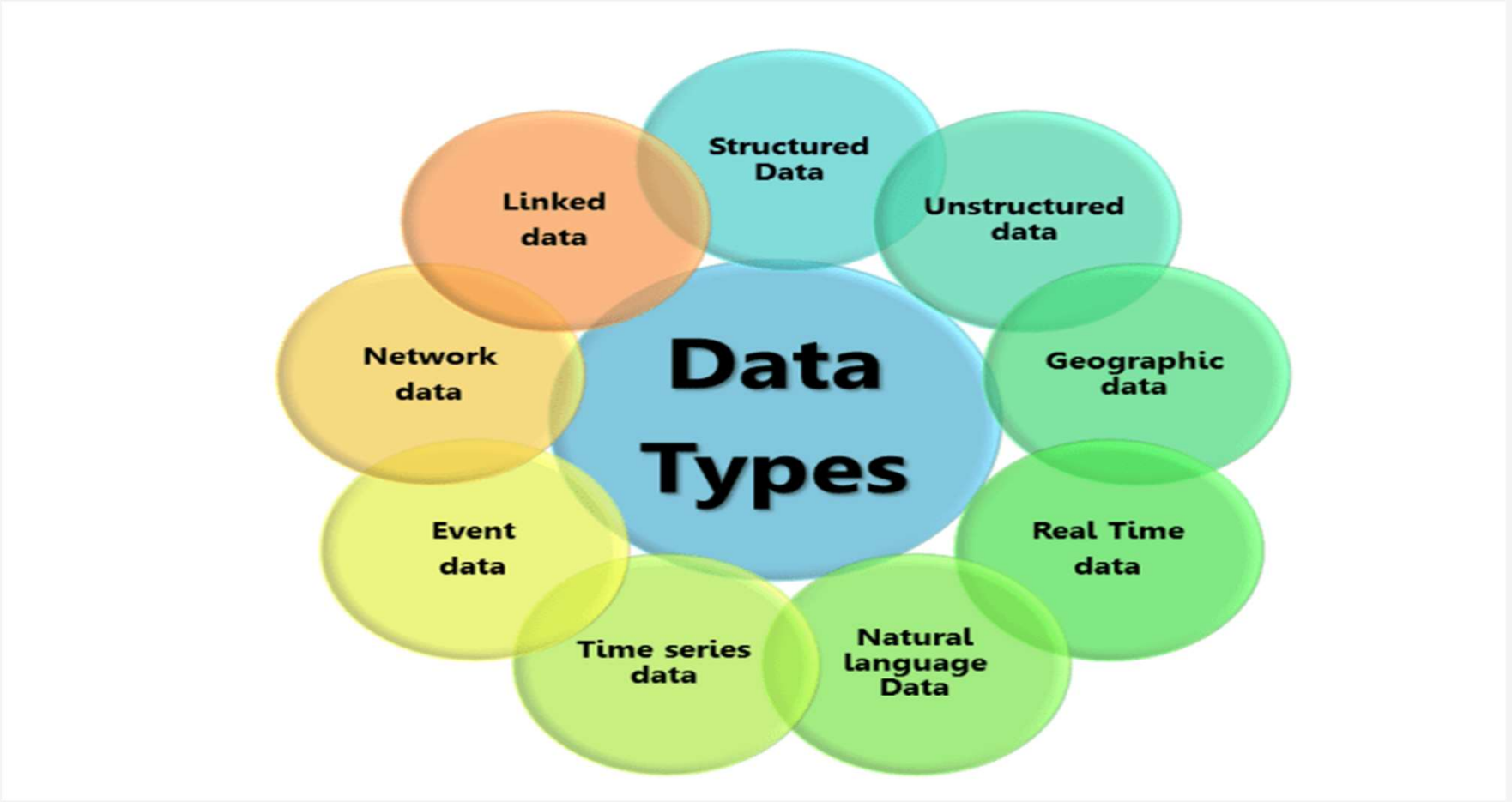


What are the Challenges of Big Data ?

- Capturing data and Storing the huge amount of data.
- Second problem is storing heterogeneous data.
- Third problem is accessing and processing speed.
  - ✓ Curation
  - ✓ Searching
  - ✓ Sharing
  - ✓ Transfer
  - ✓ Analysis
  - ✓ Presentation

Five Vs of Big Data

Data Type


Data Types
Structured Data
Unstructured data
Linked data
Network data
Geographic data
Event data
Real Time data
Time series data
Natural language Data

Use Cases

- **Black Box Data** − Component of helicopter, airplanes, and jets, etc. captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data** − Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data** − The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Power Grid Data** − The power grid data holds information consumed by a particular node with respect to a base station.
- **Transport Data** − Transport data includes model, capacity, distance and availability of a vehicle.
- **Search Engine Data** − Search engines retrieve lots of data from different databases.

And So on…

Traditional Approach

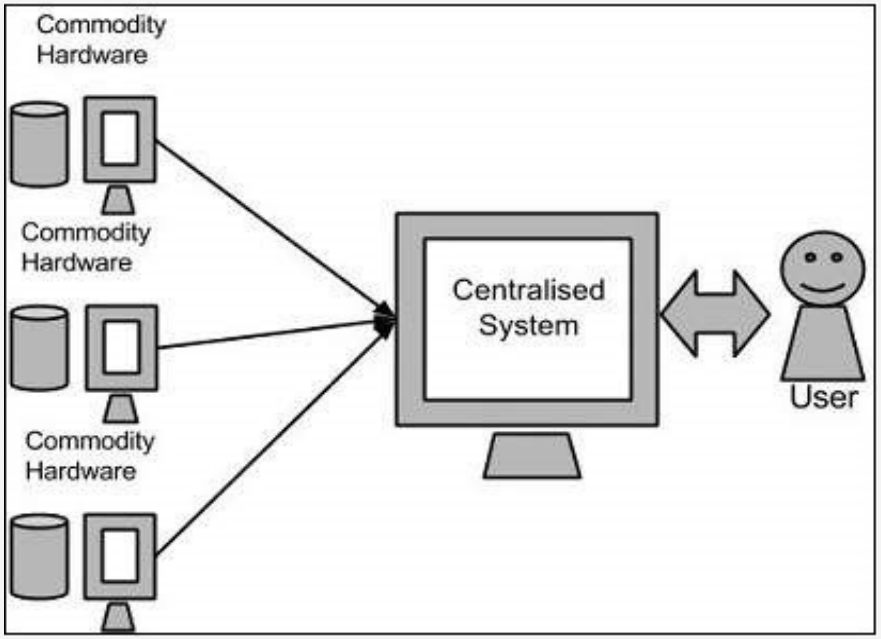
- An enterprise will have a computer server to store and process big data.
- For storage purpose, the programmers will take the help of their choice of database vendors such as Oracle, IBM, etc.
- The user interacts with the application, which in turn handles the part of data storage and analysis.
- Limitation:

  Approach works fine applications processes less voluminous data can be accommodated by standard database servers, or up to the limit of the processor that is processing the data.

  When it comes to dealing with huge amounts of scalable data, it is a hectic task to process such data through a single database bottleneck.

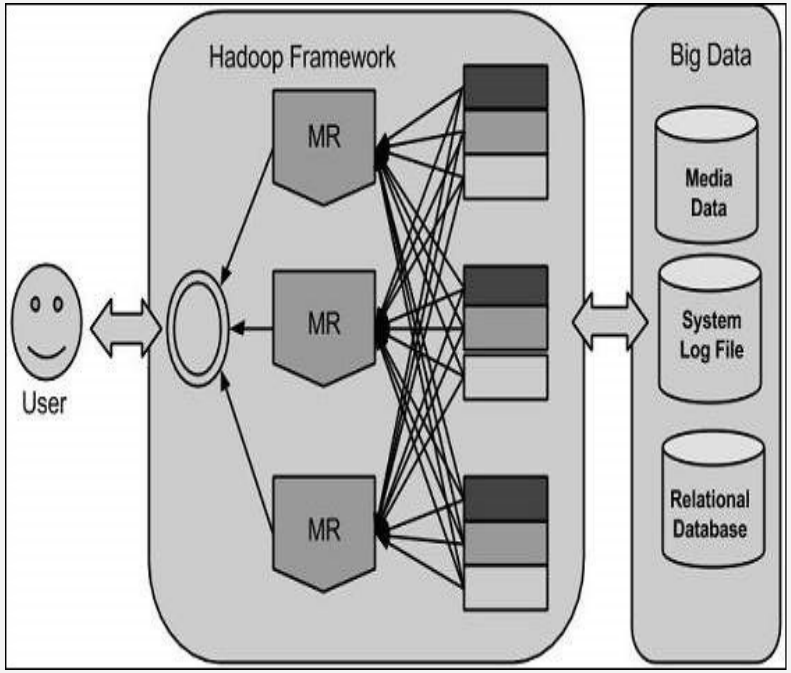
Google's Solution

- Google solved this problem using an algorithm called MapReduce.
- This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset.



What is Hadoop?

- Using the solution provided by Google, Doug Cutting and his team developed an Open Source Project called HADOOP.
- Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others. In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

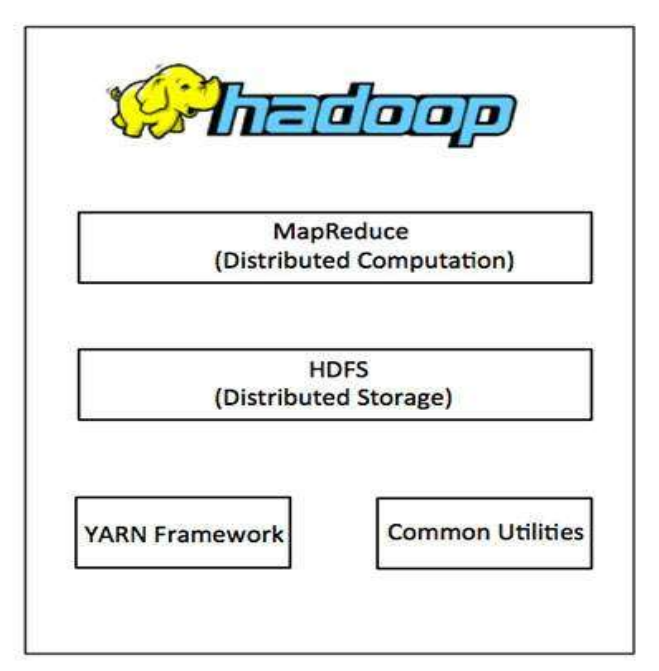


What is Hadoop?

- Hadoop is an **open-source framework** that allows to **store and process big data** in a **distributed environment across clusters of computers** using **simple programming models**.
- It is designed to **scale from single servers** to **thousands of machines**, each offering local computation and storage.

Hadoop Architecture

- At its core, Hadoop has two major layers namely
  - Processing/Computation layer (MapReduce)
  - Storage layer (Hadoop Distributed File System).

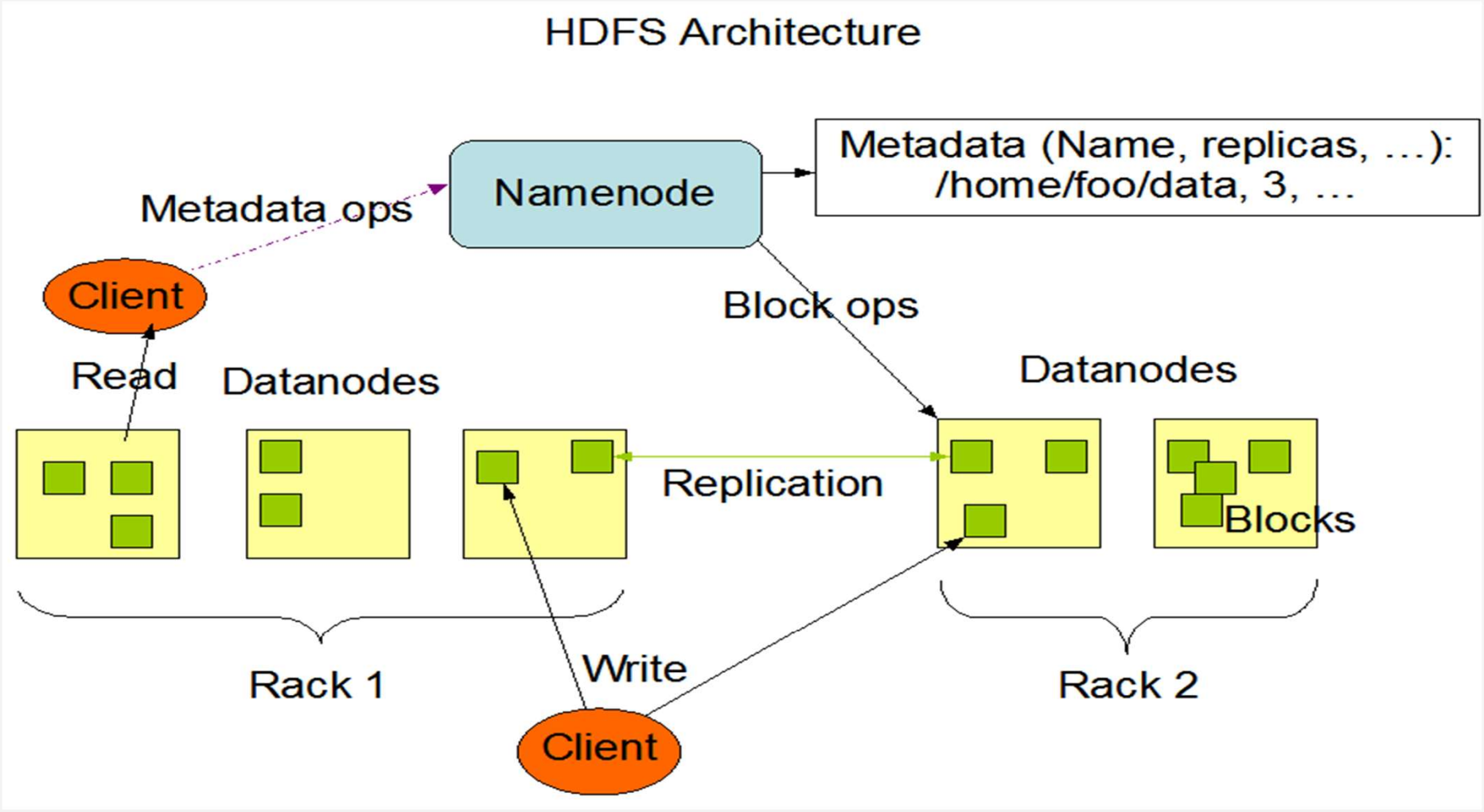


MapReduce – The Data Processing

- **MapReduce** is a **parallel programming model** for writing **distributed applications.**
- **Devised at Google** for efficient processing of large amounts of data (multi-terabyte data-sets), on **large clusters (thousands of nodes)** of **commodity hardware** in a **reliable**, **fault-tolerant manner**.
- The MapReduce program runs on Hadoop which is an Apache open-source framework.

HDFS - Data Storage and Ingest

- The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications.
- HDFS employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters.
- Hadoop itself is an open-source distributed processing framework that manages data processing and storage for big data applications.
- HDFS is a key part of the many Hadoop ecosystem technologies. It provides a reliable means for managing pools of big data and supporting related big data analytics applications.
- **Two core components**, Hadoop framework also includes the following two modules
  - **Hadoop Common** – These are Java libraries and utilities required by other Hadoop modules.
  - **Hadoop YARN** – This is a framework for job scheduling and cluster resource management.


HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

HDFS Features

- **Data replication**.
  - Ensures that the data is always available and prevents data loss.
- **Fault tolerance and reliability**.
  - HDFS' ability to replicate file blocks and store them across nodes in a large cluster ensures fault tolerance and reliability.
- **High availability**.
  - As mentioned earlier, because of replication across notes, data is available even if the NameNode or a DataNode fails.
- **Scalability**.
  - HDFS stores data on various nodes in the cluster, as requirements increase, a cluster can scale to hundreds of nodes.
- **High throughput**. HDFS stores data in a distributed manner, the data can be processed in parallel on a cluster of nodes. It increases high throughput.
- **Data locality**. With HDFS, computation happens on the DataNodes where the data resides. this approach decreases network congestion and boosts a system's overall throughput.

Components of Hadoop Ecosystems

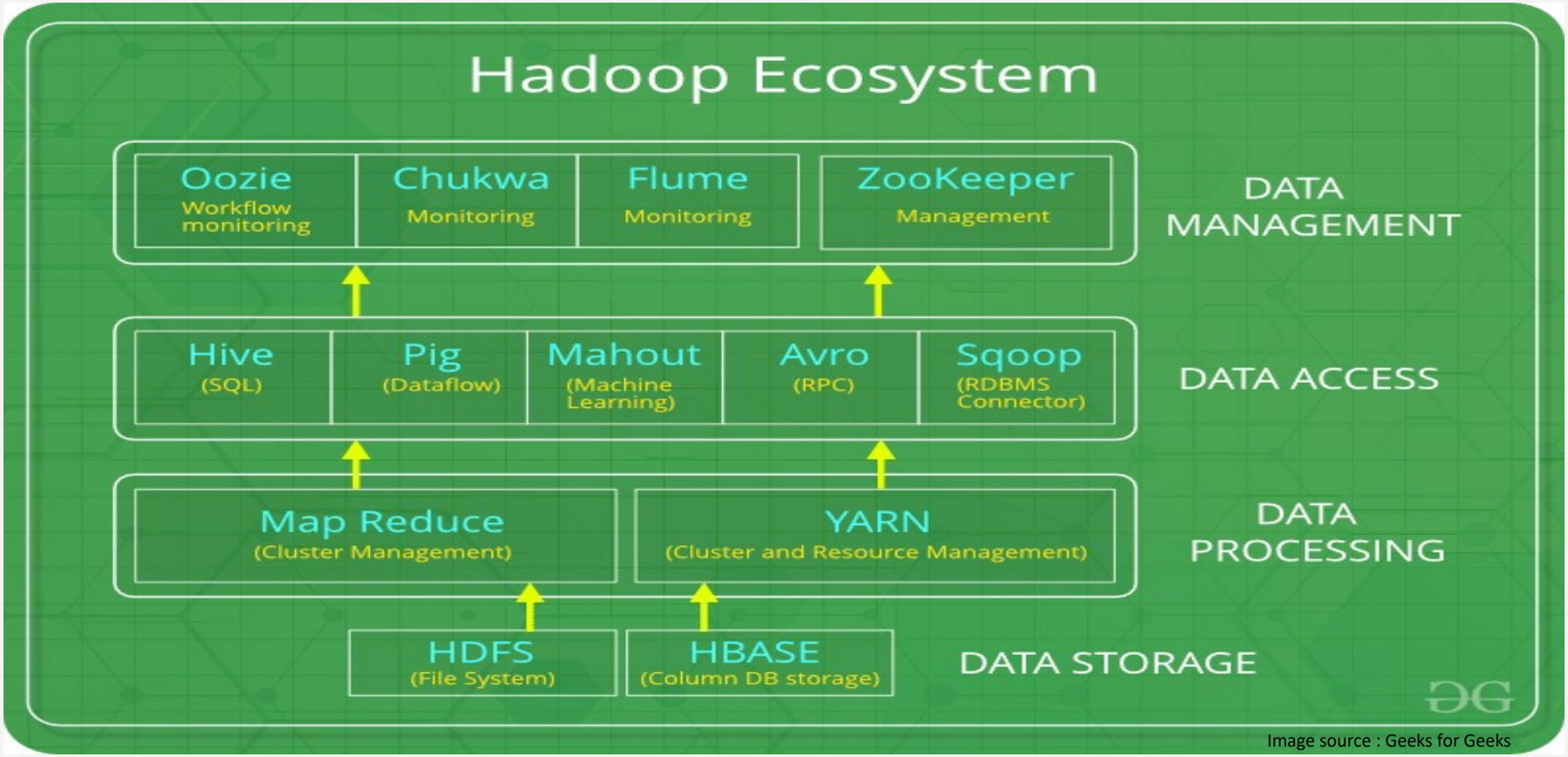


Image source : Geeks for Geeks

Case Study

- Stock Market
  - ✓ The New York Stock Exchange is an example of Big Data that generates about one terabyte of new trade data per day.

- Social Media
  - ✓ The statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, every day.
  - ✓ This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

facebook

- Airlines
  - ✓ A single Jet engine can generate 10+terabytes of data in 30 minutes of flight time.
  - ✓ With many thousand flights per day, generation of data reaches up to many Petabytes.

References

- Hadoop – The Definitive Guide By Tom White.
- Hadoop In Action. By Chuck Lam

Thank You…