
Big Data

Chapter 1: Introduction to Big Data

What is Data?

The quantities, characters, or symbols on which operations are performed by a computer,
which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

What is Big Data?

Big Data is also data but with a huge size. Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

"Extremely large data sets that may be analyzed computationally to reveal patterns , trends and association, especially relating to human behavior and interaction are known as Big Data."

Examples Of Big Data

Following are some the examples of Big Data-
The New York Stock Exchange 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.

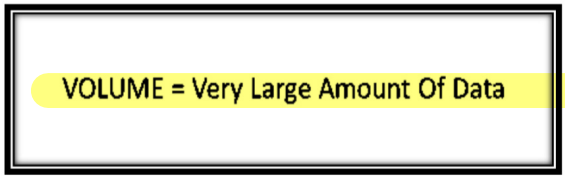
Name	Equal To	Size(In Bytes)
Bit	1 bit	1/8
Nibble	4 bits	1/2 (rare)
Byte	8 bits	1
Kilobyte	1024 bytes	1024
Megabyte	1, 024kilobytes	1, 048, 576
Gigabyte	1, 024 megabytes	1, 073, 741, 824
Terrabyte	1, 024 gigabytes	1, 099, 511, 627, 776
Petabyte	1, 024 terrabytes	1, 125, 899, 906, 842, 624
Exabyte	1, 024 petabytes	1, 152, 921, 504, 606, 846, 976
Zettabyte	1, 024 exabytes	1, 180, 591, 620, 717, 411, 303, 424
Yottabyte	1, 024 zettabytes	1, 208, 925, 819, 614, 629, 174, 706, 176

❑ Characteristics Of Big Data

- The following are known as "Big Data Characteristics".
  1. Volume
  2. Velocity
  3. Variety
  4. Veracity

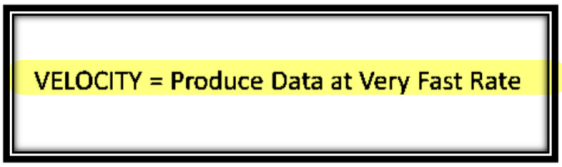
1. Volume:

Volume means "How much Data is generated". Now-a-days, Organizations or Human Beings or Systems are generating or getting very vast amount of Data say TB(Tera Bytes) to PB(Peta Bytes) to Exa Byte(EB) and more.



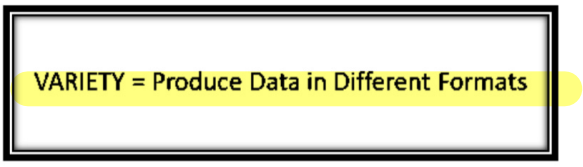
2. Velocity:

Velocity means "How fast produce Data". Now-a-days, Organizations or Human Beings or Systems are generating huge amounts of Data at very fast rate.



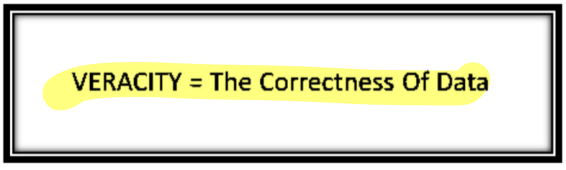
3. Variety:

Variety means "Different forms of Data". Now-a-days, Organizations or Human Beings or Systems are generating very huge amount of data at very fast rate in different formats. We will discuss in details about different formats of Data soon.



4. Veracity

Veracity means "The Quality or Correctness or Accuracy of Captured Data". Out of 4Vs, it is most important V for any Big Data Solutions. Because without Correct Information or Data, there is no use of storing large amount of data at fast rate and different formats. That data should give correct business value.



❖ Types of Digital Data

1. Structured
2. Unstructured
3. Semi-structured

❑ Structured

- Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.
- Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it.
- However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the range of multiple zettabytes.

> ➢ ***Do you know? 10^{21} bytes*** equal to ***1 zettabyte*** or ***one billion terabytes*** forms ***a zettabyte***. Looking at these figures one can easily understand why the name Big Data is given and imagine the challenges involved in its storage and processing.

> ➢ ***Do you know?*** Data stored in a relational database management system is one example of a **'structured'** data.

- **Examples Of Structured Data**

An 'Employee' table in a database is an example of Structured Data

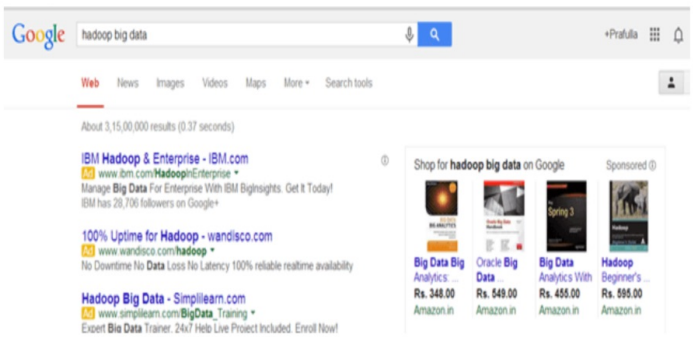
Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

❑ Unstructured

- Any data with unknown form or the structure is classified as unstructured data.
- In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.
- A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.
- Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

- **Examples Of Un-structured Data**

The output returned by 'Google Search'



❑ Semi-structured

- Semi-structured data can contain both the forms of data.
- We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.
- Example of semi-structured data is a data represented in an XML file.

➢ Examples Of Semi-structured Data

Personal data stored in an XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

❖ Big Data Analytics

➢ Big Data Analytics:

- Big Data analytics is the process of collecting, organizing and analyzing large sets of data (*called* Big Data) to discover patterns and other useful information.
- Big Data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with Big Data typically want the *knowledge* that comes from analyzing the data.

➢ High-Performance Analytics Required:

- To analyze such a large volume of data, Big Data analytics is typically performed using specialized software tools and applications for predictive analytics, data mining, text mining, forecasting and data optimization.
- Collectively these processes are separate but highly integrated functions of high-performance analytics.
- Using Big Data tools and software enables an organization to process extremely large volumes of data that a business has collected to determine which data is relevant and can be analyzed to drive better business decisions in the future.

➢ The Challenges:

- For most organizations, Big Data analysis is a challenge. Consider the sheer volume of data and the different formats of the data(both structured and unstructured data) that is collected across the entire organization and the many different ways different types of data can be combined, contrasted and analyzed to find patterns and other useful business information.
- The first challenge is in breaking down data silos to access all data an organization stores in different places and often in different systems.
- A second challenge is in creating platforms that can pull in unstructured data as easily as structured data.
- This massive volume of data is typically so large that it's difficult to process using traditional database and software methods.

➢ How Big Data Analytics is Used Today:

- As the technology that helps an organization to break down data silos and analyze data improves, business can be transformed in all sorts of ways.
- Today's advances in analyzing big data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook.
- Another example comes from one of the biggest mobile carriers in the world.
- France's Orange launched its Data for Development project by releasing subscriber data for customers in the Ivory Coast.
- The 2.5 billion records, which were made anonymous, included details on calls and text messages exchanged between 5 million users.
- Researchers accessed the data and sent Orange proposals for how the data could serve as the foundation for development projects to improve public health and safety.
- Proposed projects included one that showed how to improve public safety by tracking cell phone data to map where people went after emergencies; another showed how to use cellular data for disease containment. (source)

➢ The Challenges:

- For most organizations, Big Data analysis is a challenge. Consider the sheer volume of data and the different formats of the data(both structured and unstructured data) that is collected across the entire organization and the many different ways different types of data can be combined, contrasted and analyzed to find patterns and other useful business information.
- The first challenge is in breaking down data silos to access all data an organization stores in different places and often in different systems.
- A second challenge is in creating platforms that can pull in unstructured data as easily as structured data.
- This massive volume of data is typically so large that it's difficult to process using traditional database and software methods.

➢ How Big Data Analytics is Used Today:

- As the technology that helps an organization to break down data silos and analyze data improves, business can be transformed in all sorts of ways.
- Today's advances in analyzing big data allow researchers to decode human DNA in minutes, predict where terrorists plan to attack, determine which gene is mostly likely to be responsible for certain diseases and, of course, which ads you are most likely to respond to on Facebook.
- Another example comes from one of the biggest mobile carriers in the world.
- France's Orange launched its Data for Development project by releasing subscriber data for customers in the Ivory Coast.
- The 2.5 billion records, which were made anonymous, included details on calls and text messages exchanged between 5 million users.
- Researchers accessed the data and sent Orange proposals for how the data could serve as the foundation for development projects to improve public health and safety.
- Proposed projects included one that showed how to improve public safety by tracking cell phone data to map where people went after emergencies; another showed how to use cellular data for disease containment. (source)

➢ The Benefits of Big Data Analytics:

- Enterprises are increasingly looking to find actionable insights into their data. Many big data projects originate from the need to answer specific business questions. With the right big data analytics platforms in place, an enterprise can boost sales, increase efficiency, and improve operations, customer service and risk management.
- Webopedia parent company, QuinStreet, surveyed 540 enterprise decision-makers involved in big data purchases to learn which business areas companies plan to use Big Data analytics to improve operations. About half of all respondents said they were applying big data analytics to improve customer retention, help with product development and gain a competitive advantage.
- Notably, the business area getting the most attention relates to increasing efficiency and optimizing operations. Specifically, 62 percent of respondents said that they use big data analytics to improve speed and reduce complexity.

❖ Application of Big Data

✓ Here is the list of top Big Data applications in today's world:

- Big Data in Healthcare
- Big Data in Education
- Big Data in E-commerce
- Big Data in Media and Entertainment
- Big Data in Finance
- Big Data in Travel Industry
- Big Data in Telecom
- Big Data in Automobile

1. Big Data in Retail

- The retail industry is the one that faces the most fierce competition of all. Retailers constantly hunt for ways that will give them a competitive edge over others. Customers are the real king sounds legit for the retail industry in particular.
- For retailers to thrive in this competitive world, they need to understand their customers in a better way. If they are aware of their customers' needs and how to fulfill those needs in the best possible way, then they know everything.

- ***Check how Big Data act as a weapon for retailers to connect with their customers – Big Data in Retail.***
  - Through advanced analysis of their customer's data, retailers are now able to understand them from every angle possible. They gather this data from various sources such as social media, loyalty programs, etc.

2. Big Data in Healthcare

- Big Data and healthcare are an ideal match. It complements the healthcare industry better than anything ever will. The amount of data the healthcare industry has to deal with is unimaginable.
- Gone are the days when healthcare practitioners were incapable of harnessing this data. From finding a cure to cancer to detecting Ebola and much more, Big Data has got it all under its belt and researchers have seen some life-saving outcomes through it.
- Big Data and analytics have given them the license to build more personalized medications. Data analysts are harnessing this data to develop more and more effective treatments. Identifying unusual patterns of certain medicines to discover ways for developing more economical solutions is a common practice these days.

3. Big Data in Education

- When you ask people about the use of the data that an educational institute gathers, the majority of the people will have the same answer that the institute or the student might need it for future references.
- Even you had the same perception about this data, didn't you? But the fact is, this data holds enormous importance. Big Data is the key to shaping the future of the people and has the power to transform the education system for better.
- Some of the top universities are using Big Data as a tool to renovate their academic curriculum. Additionally, universities can even track the dropout rates of the students and are taking the required measures to reduce this rate as much as possible.

4. Big Data in E-commerce

- One of the greatest revolutions this generation has seen is that of E-commerce. It is now part and parcel of our routine life. Whenever we need to buy something, the first thought that provokes our mind is E-commerce. And not your surprise, Big Data has been the face of it.
- Some of the biggest E-commerce companies of the world like Amazon, Flipkart, Alibaba, and many more are now bound to Big Data and analytics is itself an evidence of the level of popularity Big Data has gained in recent times.
- Big Data is now as important as anyone else in these organizations. *Amazon, the biggest E-commerce firm in the world and one of the pioneers of Big Data and analytics, has Big Data as the backbone of its system. Flipkart, the biggest E-commerce firm in India, has one of the most robust data platforms in the country.*
- ***See how Flipkart used Big Data to have one of the most robust data platforms.***
- Big Data's recommendation engine is one of the most amazing applications the Big Data world has ever witnessed. It furnishes the companies with a **360-degree** view of its customers.
- Companies then suggest customers accordingly. Customers now experience more personalized services than they have ever had. Big Data has completely redefined people's online shopping experiences.

5. Big Data in Media and Entertainment

- Media and Entertainment industry is all about art and employing Big Data in it is a sheer piece of art. Art and science are often considered to be the two completely contrasting domains but when employed together, they do make a deadly duo and Big Data's endeavors in the media industry are a perfect example of it.



- Viewers these days need content according to their choices only. Content that is relatively new to what they saw the previous time. Earlier the companies broadcasted the Ads randomly without any kind of analysis.
- But after the advent of Big Data analytics in the industry, companies now are aware of the kind of Ads that attracts a customer and the most appropriate time to broadcast it for seeking maximum attention.
- Customers are now the real heroes of the Media and entertainment industry - courtesy to Big Data and Analytics.

6. Big Data in Finance

- The functioning of any financial organization depends heavily on its data and to safeguard that data is one of the toughest challenges any financial firm faces. Data has been the second most important commodity for them after money.
- Even before Big Data gained popularity, the finance industry was already conquering the technical field. In addition to it, financial firms were among the earliest adopters of Big Data and Analytics.
- **Digital banking** and **payments** are two of the most trending buzzwords around and Big data has been at the heart of it. Big Data is bossing the key areas of financial firms such as **fraud detection**, **risk analysis**, **algorithmic trading**, and **customer contentment**.
- This has brought much-needed fluency in their systems. They are now empowered to focus more on providing better services to their customers rather than focussing on security issues. Big Data has now enhanced the financial system with answers to its hardest of the challenges.

7. Big Data in Travel Industry

- While Big Data is spreading like wildfire and various industries have been cooking its food with it, the travel industry was a bit late to realize its worth. Better late than never though. Having a stress-free traveling experience is still like a daydream for many.
- And now Big Data's arrival is like a ray of hope, that will mark the departure of all the hindrances in our smooth traveling experience.



- ***See how Big Data is revolutionizing the travel & tourism sector.***
- Through Big Data and analytics, travel companies are now able to offer more customized traveling experience. They are now able to understand their customer's requirements in a much-enhanced way.
- From providing them with the best offers to be able to make suggestions in real-time, Big Data is certainly a perfect guide for any traveler. Big Data is gradually taking the window seat in the travel industry.

8. Big Data in Telecom

- The telecom industry is the soul of every digital revolution that takes place around the world. With the ever-increasing popularity of smartphones, it has flooded the telecom industry with massive amounts of data.
- And this data is like a goldmine, telecom companies just need to know how to dig it properly. Through Big Data and analytics, companies are able to provide the customers with smooth connectivity, thus eradicating all the network barriers that the customers have to deal with.
- Companies now with the help of Big Data and analytics can track the areas with the lowest as well as the highest network traffics and thus doing the needful to ensure hassle-free network connectivity.
- Big Data alike other industries have helped the telecom industry to understand its customers pretty well.
- Telecom industries now provide customers with offers as customized as possible.
- Big Data has been behind the data revolution we are currently experiencing.

9. Big Data in Automobile

- *"A business like an automobile, has to be driven, in order to get results." B.C. Forbes*
- And Big Data has now taken complete control of the automobile industry and is driving it smoothly. Big Data is driving the automobile industry towards some unbelievable and never before results.
- The automobile industry is on a roll and Big Data is its wheels or I must say Big Data has given wings to it. Big Data has helped the automobile industry achieve things that were beyond our imaginations



- From analyzing the trends to understanding the supply chain management, from taking care of its customers to turning our wildest dream of connected cars a reality, Big Data is well and truly driving the automobile industry crazy.

Questions:

a- Can you describe the limitations of traditional (local) file system?

- Scalability: limited storage capacity
- Data access: cannot efficiently concurrent access by multiple user
- Fault tolerance: vulnerable to data loss if local machine fail
- Performance: performce degrade with large volume of data
- Data sharing: difficult to share & collaborate on data access location

b- Can you explain what object stroge is?

is data storage architecture for Rondling large amount of structured data, it treats each piece of data as an object with unique ID (identified with identifier and use large amount of cheap machine rather than a supa pc.

c_ Can you explain what are the benefits of object storage?

no-hierarchy _ make metadata flexible _ use a very simple model _ use very large number of cheap machine instead of a super pc.

d_ Can you contrast or relate object storage with block storage, a file system and the Key-value model?

- Object storage includes meta data in better for unstructured data while block storage is faster & better for data needing high performance
- object storage is more scalable and suitable for large scale data while while local file systems are easier for everyday file management
- both can handle unstructured data, but object storage is more suited for large file with extensive meta data while Key-value stores excel in Fast, simple data retrieval (Key value / Obj. stor.)

e- Can you explain the 3 ≠ ways, on the physical level to deal with more data (scale up, scale out, write better code)?

scale up: buy bigger machine, more storage, more cores and faster

scale out: buy more similar machines and share the work access them

improve your code: improve efficiency of code to handle data more effectively (optimizing algo, reduce complexity)

f- Can you explain why scaling out is less expensive than scaling up?

because it involves adding more typically lower cost, server to handle increased work loads, rather than upgrading a single server with more expensive components

g- Can you explain what aspects of the design of object storage enable scaling out and why?

non hierarchical flexible meta data aspects enable scaling out by allowing data to be distibuted accross multiple nodes without complex hierarchy structure because each object is stored with multiple unique identifier, meta data associated with each object help manage & locate data efficiently

h- Do you Know what data center is made of?

- Racks: physical forms hold multiple servers, storage, devices, & networking components
- Servers: node that provide power for processing task
- storage: manage & store data
- networking equipments: switches, router, cables that enable communicat: between data centers and external sources and connects storage & server together
- Power supply system: infrastructure to provide consistent and reliable power (UPS - -)
- cooling sys: equipment to regulate temperature and ensure optimal operating conditions.
- Security sys: physical & digital security mesures to prokat data

i- Do you Know the difference between storage, memory CPU and network and how the three are paramount in a cluster ?

storage : provide but a storage

RAM : random access memory offer fast access to actively processed data

CPU : brain of pc, execute tasks

networks : facilitate communication between nodes.

these components work together to enable distributed computing processing performance accross cluster to make each one efficient operation

j- Do you Know rough, typical numbers (per-node storage capacity, memory, nb of cores...) ?

CPU : 1 - 64 cores

memory size : 16GB → 6TB

storage : 1 - 20TB

Bandwidth : 1Gb - 100 MB/s

K. Do you know how storage and memory technologies (HDD - SSD - RAM) compare in terms of capacity and throughput?

HDD provide highest capacity but slower in throughput
SSD provide balance between capacity & throughput
RAM: Fastest throughput with lowest capacity

l. sketch (Key-value) data model behind object storage? can you spot instances of the Key-value model in several other technologies seen in the future?

object storage uses Key-value data model to retrieve data. in this model each object is signed a unique key and the value associated with that key is the data itself (it allows efficient retrive for the data without needing to know its physical location example HDF, AS3, ---)

m. Difference betwen data & meta data:

metadata is the data about the data
data represents actual content or info stored, metadata provide a descriptive inf. about the data, helping organization, management, and understanding.

n- Do you know the order of magnitude that can be achieved in terms of number of objects, and object size?

in terms of nb of objects: Terrabytes
while object size typically ranges up to few hundred of KB.

o- Can you name a few big players for cloud based object storage?

Vendors: Amazon S3, Azure, IBM client, google
Consumers: Netflix, spotify, NASA

p- Can you describe the feature of S3 and Azure Blob Storage on a high level? Do you know what a bucket and object are? What block blob storage, append blob storage and page blob storage

P- Can you describe the feature of S3 and Azure Blob Storage on a high level? Do you know what a bucket and object are? What block blob storage, append blob storage and page blob storage

ex Amazon S3:

called simple storage service where objects are stored into buckets, buckets identified with bucketID, and each object is identified with objID

==> object: objectID + bucketID

we say object rather than file

- object store has no hierarchies
- object can be at most 5TB (upload in a single chunk if less than 5TB otherwise in several blocks)
- by default users have 100 buckets.

ex: Azure BlobB storage: fully managed cloud storage

differ from S3 in:

- its architecture is publicly documented → easier
- objects has 3IDs : account, container, BlobB.
- objects are divided in several blocks more prominently than S3.
- 3 between block blobs and Append block
- max size 195 GB for Append and 190.7 TB for Block

9- Can you describe what the most important SLA (Service level Agreement) parameters mean (e.g. latency, availability, durability) as well as their typical range?

- Latency (time it takes for request to be processed) 99.9% of requests will be processed in the 300ms
- Availability (% of time that service is operational
- durability :

10- Can you describe a typical use case for object storage?

streaming platform : Netflix for example

- CAP Theorem:
  - C : Consistency
  - A : Availability
  - P : Partition tolerance

11- Can you explain why, for large amounts of data CAP becomes relevant over ACID?

ACID requires consistency & isolation
CAP: is more suited for distributed system with large amount of data

Questions chap 3:

a- Can you explain the difference between block storage and object storage ?

- **Block storage:**
  - split data into blocks fixed size and store them as separated peices.
  - For structured data
  - More expensive, Fast, less scalable
  - File sys access to data
- **object storage:**
  - store data as whole object with customizable meta data
  - For large unstructured data
  - cheap, slower, scalable
  - accessed through API, URL

b- Can you explain the difference beteen logical Key value model and a file system ?

logical Key values store data using Key value pair, where each key is unique and used to retrieve its associated value.

File system organize data hierarchically using directories and files with a path structure to locate and manage files (Structur data / Traditional sys)

c- Can you contrast block storage and object storage in terms of maximum number of object/ file?

Block storage sys might impose more limits on the number of block due to its structured nature & the limitation of capacity
Where object storage is designed for massive scalability and can support as extremely large nb of object more spaces and metadata capabilities allow for efficient management of vast amount of data

d_ Can you contrast block storage and object storage in terms of maximum size of object/ file?

Block storage allows very large volumes often measured in Petabyte but max size is generally constrainted by file sys used.
Object storage support very large individual object size with many sys allowing object up to several Terabytes (AS3 object 5TB)

c- Do you know the order of magnitude of a block size for a local filesystem & for a distributed file system? Can you explain the rationale behind them with respect to latency & throughput?

Local filesystem: 4Kb to 64Kb.
Latency: smaller blocksizes reduces the latency of accessing file
DFS: 64, 128MB
Latency: DS are optimized throughput rather than latency
Larger block help minimize overhead associated with managing large nb of small blocks across network reducing nb of block transaction
Larger throughput maximize blocks by allowing more data to be transfered

f- Do you Know where HDFS shines and why ?

HDFS shine in scenarios that require handling large scale data with throughput and reliability its design principles scalability fault tolerance and cost effectiveness make it powerful in solution

g- Do you Know that HDFS files are updated by appending atomically & why ?

HDFS design choice to support atomic append

h. Do you Know that HDFS performs in terms of throughput and latency ?

HDFS performs exceptionally well in scenarios requiring high throughput and can handle large data structures but its less optimal for app that requires low- latency data access.

i. Do you know the main benefit of HDFS (commodity hardware, ---?)

Fault tolerance, monitoring itself, automatically recovering, failure detection and scalability

j. Can you contrast centralized architectures to decentralized (peer-to-peer) architectures?

In decentrilized network each machine to any machine

In contrast in centralized architecture, where one node is special interchangable and connected to it.

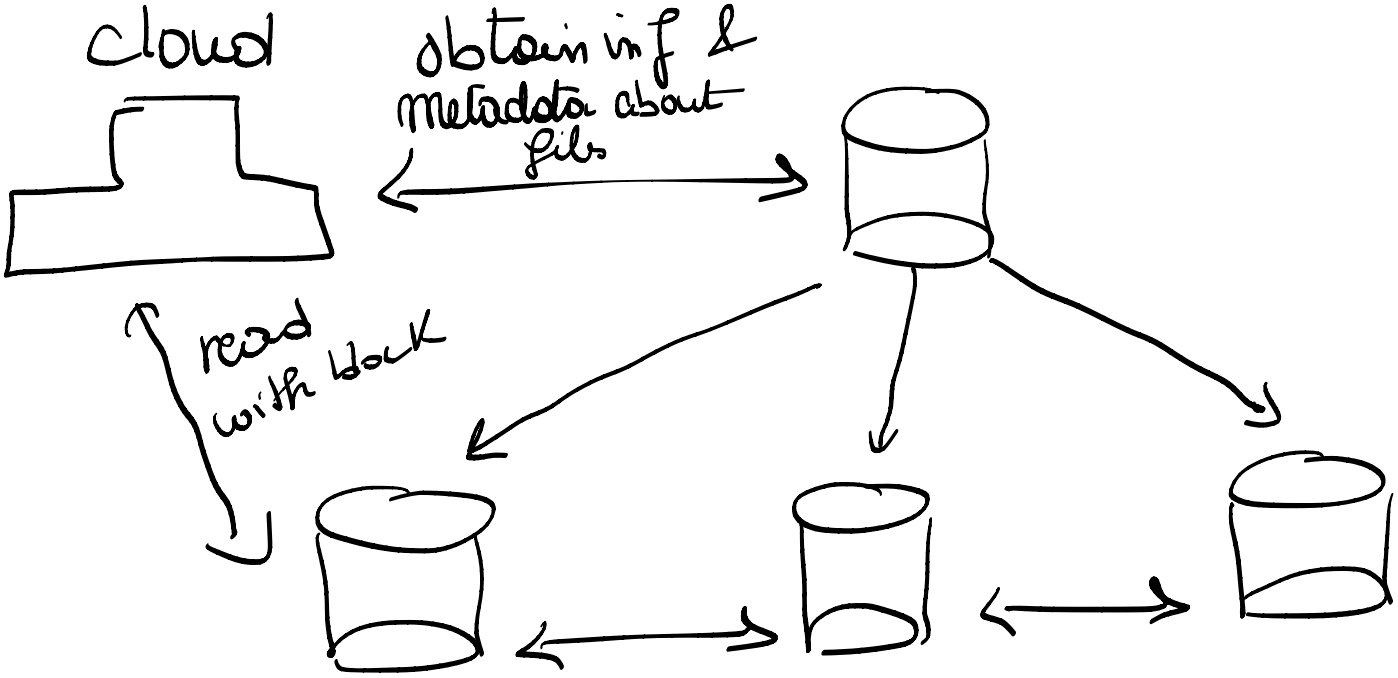
k. Can you explain the HDFS architecture, what a NameNode is and what a DataNode is, how blocks are replicated?

HDFS is centralized where NameNode is masternode responsible for managing File system, namespace and meta data material

Data nodes are slave stores and

Data replication ensure fault tolerance and availability by replicating 3 times as default 1 in server working with and 2 in different rack one and node.

l. Can you sketch how various components communicate with each other? (Client, NameNode, DataNode)?



m. Can you point the single point of failure of HDFS and explain how they can be addressed?

NameNode by backup and recovery to address this metadata file to block mapping and file namespace is backed up periodically to snapshot. Additionally update File sys after the snapshot are stored
If namenode crashed it can be restarted. edit log can be used to restore file sys state.

p. Are you able to use the HDFS shell (creating directories, reading, uploading and downloading files, etc)?

Create dir
mkdir
list reading
Copy from local
Copy to local