Imp Questions of Big Data

Qu-1: Discuss RDBMS, No SQL and difference.
Qu-2: What are Big Data Processing tools?
Qu-3: Discuss the Applications and characteristics of Big Data.
Qu-4: Explain Data Marts, Data lakes, Data pipelines etc types of Data.
Qu-5: Explain 6V's of Big Data.
Qu-6: What are different file formats of Big Data?
Qu-7: How will you improve Quality and storage of data.
Qu-8: What is the process of security of BD
Qu-9: What is usability in B.D. → Applications
Qu-10: Explain ETL in detail.

Q1) RDBMS and NoSQL.

→ RDBMS.
- Relational Database Management Systems.
- Most Popular Database.
- data is stored in form of rows & (tuples).
- there are huge numbers of tables in it.
- Data can be easily accessed because it is stored in the tables.
- This Model was proposed by E.F. Codd.

→ NO SQL
- non-SQL Database.
- doesn't use tables to store the data
- used for storing and fetching the data in database.
- generally used to store large amount of data.

GOOD WRITE

- it supports query language and provides better performance.

Diff (Expensive, uses big servers) (easy, uses cheap servers)

RDBMS	No SQL
(1) used to handle data coming in low velocity.	(1) high velocity.
(2) gives only read scalability.	(2) gives both read and write scalability.
(3) manages structured data.	(3) manages all type of data.
(4) Data arrives from one ~~and~~ or few locations.	(4) many locations.
(5) supports Complex transactions.	(5) simple transactions.
(6) has single point of failure.	(6) No single point of failure.
(7) handles data in less volume.	(7) high volume.
(8) Transactions written in one location.	(8) many locations.
(9) Support ACID properties.	(9) doesn't support.
(10) difficult to make changes in Database once it is defined.	(10) Enables easy and frequent changes to database.

(11) Schema is mandatory to store the data.	(11) Schema design is not required.
(12) Deployed in Vertical fashion.	(12) Deployed in Horizontal fashion.

(2) Big Data Processing Tools :-

→ Big Data Tool is a software that extracts info from various complex data types. and sets and then process these to provide meaningful ~~insights~~ insights. Traditional databases cannot proces huge data hence best big data tools that manage big data easily are used by businesses.

→ Tools are :-

(1) Apache Hadoop.

- open source software.
- process data set of big data with the help of MapReduce programming model.
- written in Java.
- provides cross-platform support.
- Most popular big data Tool.
- used by IBM, Amazon Web Services, Microsoft and Facebook etc.

Pros :-

- offers flexibility and faster data processing.
- Highly scalable, provides fast access to data.
- HDFS (Hadoop Distributed File System), which is its core strength, has ability to hold all data types like videos, images, XML, JSON and ~~text~~ plain text over the same file system.

Cons:
- Gives disk space problems sometimes.

Pricing: Free to use under Apache License.

(2) Apache Cassandra.

- open source (free of cost).
- it employees CQL (Cassandra Structure language) to interact with database.
- used by fortune 500 companies including facebook, Honeywell, Yahoo, Accentabe and American Express etc.

Pros:
- Handles Huge data volume very fast without any single point of failure.
- comes with simple ring Architecture and dog-structured storage.
- offers automated replication.

Cons:
- clustering needs improvement.
- Troubleshooting and maintenance are not easy.

Pricing: Free of Cost.

(3) Mongo DB.

- Free open-source.
- supports multiple operating systems, including windows Vista(), OS X, Linux, Solaris and Free BSD.
- it is a Document oriented database.
- is written in C, C++ and JavaScript.
- used by brands like 'Facebook, ebay, Google.

Pros:
- Reliable, low-cost, easy to learn.
- Smooth installation and maintenance.

Cons:
- Has limited analytics and is slow for certain use cases.

Pricing: Pricing for enterprise and SMB versions is available on request.

4. Tableau:
- Famous software with three diff. options:
  1. Tableau Desktop (for analyst).
  2. Tableau Server (for enterprise use).
  3. Tableau online (cloud-based)

  Tableau Reader and Tableau Public are recent additions.
- used mainly for data visualization, exploration and understanding.
- Can handle all sizes of data and quite easy to handle.

Pros:
- offers huge flexibility to create various types of visualizations as desired.
- provides superb data blending.
- Mobile friendly.

Cons:
- Scope of improvement area include formatting controls and built-in tools for deployment and migration b/w other tableau environment servers.

Pricing: - is not free and pricing starts from $35/mon

5. Datawrapper:
- open source.

GOOD WRITE

- helps in Data visualization and prepares precise and simple charts fast.
- used by many big brands like Twitter, Bloomberg, The times and Fortune to name a few.

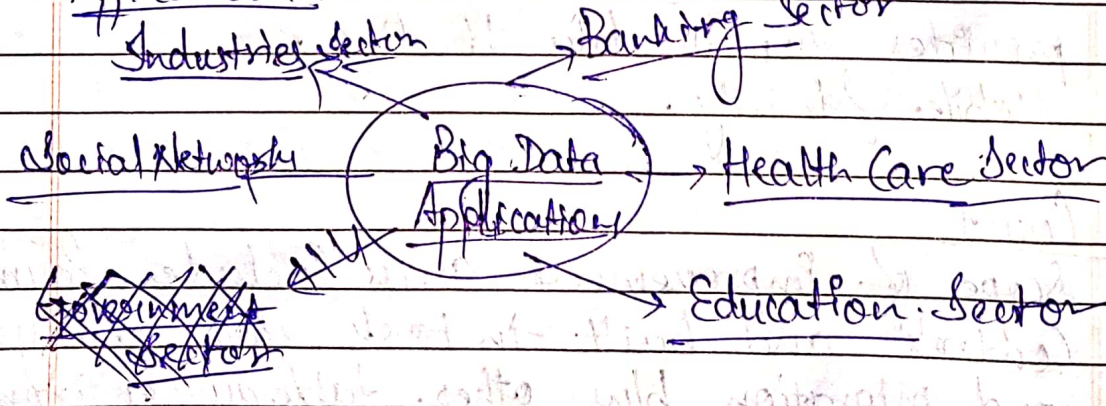
Pros:
- Device Independent and works well on all types of devices - mobile, tablet or desktop.
- Very fast, fully responsive, interactive
- No Coding required.

Cons:
- Nothing serious. Just offers a limited color palette.

Pricing: Offers both free and pricing models.

3- Application and characteristics of Big Data

Applications:



1- Banking Sector:
- Banks are creating a Digital Database in which all data related to deposits or withdrawals, transactions and various Customer Service records is stored.

By analyzing the whole data they can enhance cybersecurity and can provide innovative and personal offers to each individual.

(2) Big retails use Big data to Track Customer spending Habit, and Shopping Behavior to provide recommendations.

(3) Smart traffic system : Data about condition of traffic of dif. road, collected through Camera kept beside road, GPS device placed in vehicle. (ola, uber, Cab etc). All these data are analysed and jam-free or less jam way, less time taking ways are recommended.

(4) Virtual Personal Assistant Tool : Big Data helps virtual personal Assistant Tool (Siri, Google Assistant etc) to provide answer of various questions asked by user.

(5) Education Sector : Online Educational Course Conducting organization utilize big data to search Candidate, Interested in that Course.

(6) Media and Entertainment Sector : Netflix, Spotify, Amazon prime do analysis on data collected from their users to set the next business strategy.

(7) HealthCare Sector : Big data is used to improve services and better understand the overall health situation to solve it quickly.

(8) Industries : Big Data helps Companies enhance the quality and efficiency of their various products

helps companies make more flexible business decisions and quickly resolve problems.

(9.) Social Network →

- By this data companies can predict the customers demand before they request it.

(10.) ~~Government Sector~~ →

⇒ Characteristics of Big Data: and 6 Vs of Big Data (Que-5)

- Volume, Velocity, Variety, Veracity, Value, Variability.

(1.) Volume →

- Big Data = enormous data.
- Volume is a huge amount of data.
- To determine volume of data, size of data plays a very crucial role.
  → If volume of data is very large, then it is actually considered as a 'Big Data'.
- Hence, while dealing with 'Big Data' it is necessary to consider a characteristic "Volume".

(2.) Velocity →

- refers to high speed of accumulation of data.
- In Big Data, Data flows in from sources like machines, networks, social media, mobile phones etc.
- There is massive and continuous flow of data. This determines the potential of data that how fast the data is

generated and processed to meet the demands.
- Sampling data can help in dealing with the pace like 'Velocity'.

(3) Variety :-
- refers to nature of data that is structured, semi-structured and unstructured data.
- also refers to heterogeneous sources.
- it is basically the arrival of data from new sources that are both inside and outside of an enterprise.

→ Structured - organized data.
→ Semi-structured - semi-organized data.
Ex- log files.
→ Unstructured - unorganized data.
Ex:- Text, Images, videos etc.

(4) Veracity :-
- refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- Ex:- Data in bulk can create confusion whereas less amount of data could convey half or incomplete info.

(5) Value :-
- The bulk of Data having no value is of no ~~use~~ good to company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract information.
- Hence it is the most important V of all 6 V's.

(6) Variability : (tendency to change or shift).
~~• How fast or available data~~

• How fast you data is changing or shifting.
• How often the meaning or shape of your data changes.
• Ex: If you are eating same ice-cream daily and the taste just keep changing.

Qu-4. Types of Data :

Structured Data	Unstructured data	Semi-structured data
(1) well-organized data.	(1) Not organized at all.	(1) Partially organized.
(2) less flexible and difficult to scale. it is schema dependent.	(2) is flexible and scalable. it is schema independent.	(2) is more flexible and simpler to scale than structured data. but denser than unstructured data.
(3) based on RDBMS.	(3) based on character and binary data.	(3) based on XML / RDF.
(4) Easy analysis.	(4) Difficult analysis.	(4) Difficult analysis compared to structured data but easier when compared to unstructured data.
(5) Ex: financial data, bar codes.	(5) Ex: Media logs, videos, audios, images.	(5) Ex: Tweets organized by hashtags, folders organized by topics.

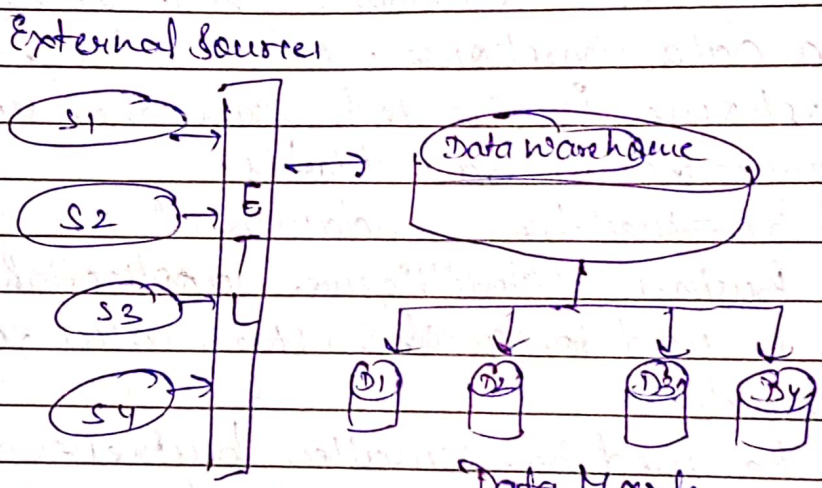
→ Reason for Creating Data Mart :

- Easy to Create.
- Creates Collective data by group of users.
- Easy access to frequently needed data.
- Improves end-user response time.
- Lower Cost than data warehouse.
- Contains only essential business data.

→ Types of Data Marts :

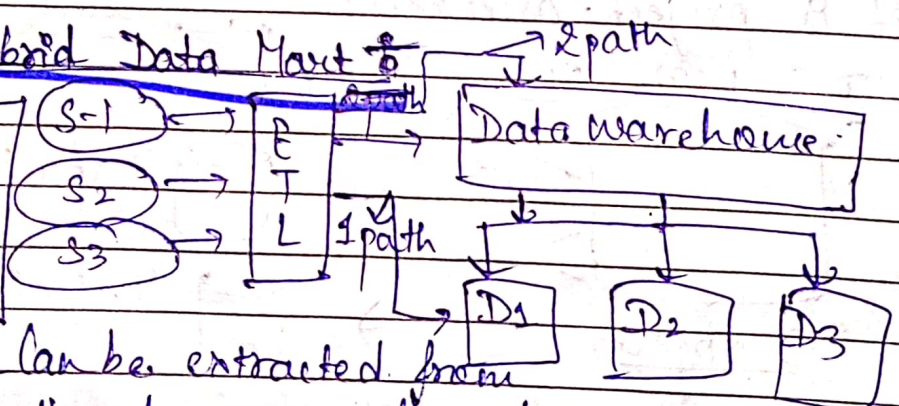
- Dependent Data Marts.
- Independent " "

(1) Dependent Data Marts.



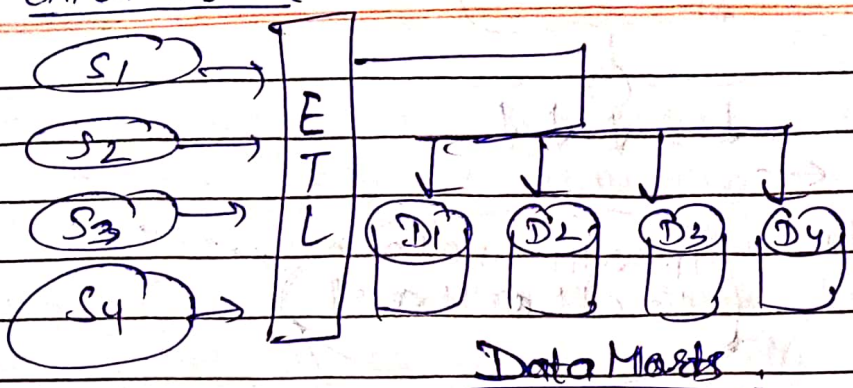
- This model is used by big organizations.
- it is created in Top-down approach.

(3) Hybrid Data Mart :



- 2 path reflects dependent data model.
- data can be extracted from operational source or from data warehouse.
- 1 path reflects accessing data directly from external source.

(2) Independent data Mart :



Data Marts :

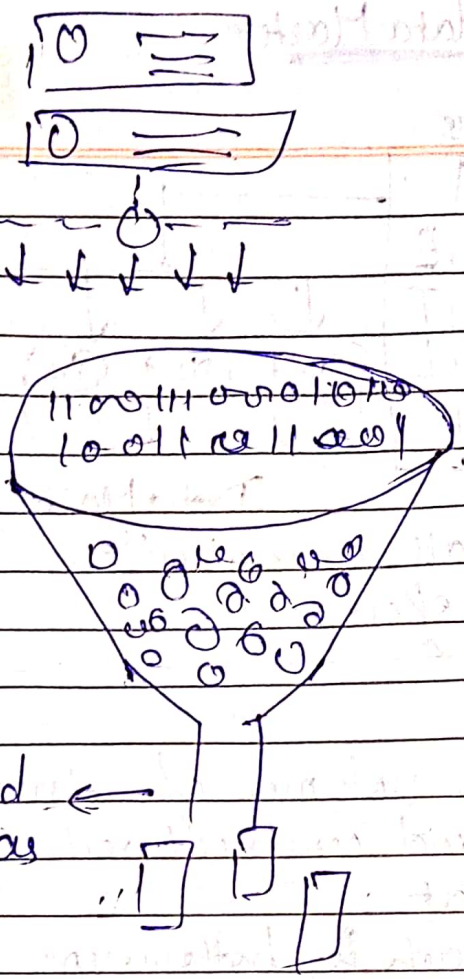
- used by small organizations.
- it is Cost effective.

→ Data Lakes :

- stores large volumes of structured, semi-structured and unstructured data. in its native format.
- Design approach is bottom-up.
- data lakes Captures anything the organization deems valuable for future use.
- data Can be images, videos, PDFs anything.
- Can be used for data analytics and report Creation.
- technology used in Data lake is much more Complex than in a data warehouse.
- Because of level of Complexity and skill required to leverage, a data lake requires users who are experienced in programming languages and data science techniques.

→ characteristics of Data Lake :

- Collects all data from many disparate data sources over an extended period.
- Meets the needs of various users in organization
- it is uploaded without an established methodology.
- Process and cleans data and stores it in the data lake.



→ Data pipelines :

- Can work with any type of data.
- It is a means of moving data from one place (the source) to a destination (such as data warehouse).
  Along the way data is transformed and optimized, arriving in a state that can be analyzed and used to develop business insights.
- It is a step involved in aggregating, organizing and moving data.

→ Elements :

Data pipelines consist of three essential elements

(1) Sources : Sources are where data comes from.

- Common sources include RDBMS systems like MySQL, CRMs such as Salesforce and HubSpot.

(2) processing steps : transformation, augmentation, filtering, grouping and aggregation.

(3) Destination : where data ~~arrives~~ arrives at the end of its processing, typically a data lake or data warehouse for analysis.

→ characteristics : (to look when considering a data pipeline include)

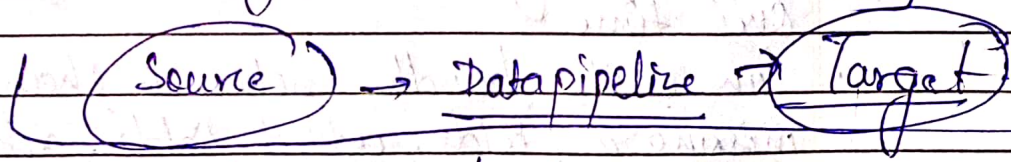
- Continuous and extensible data processing.
- The elasticity and agility of cloud.
- Isolated and independent resources for data processing.
- High availability and disaster recovery.

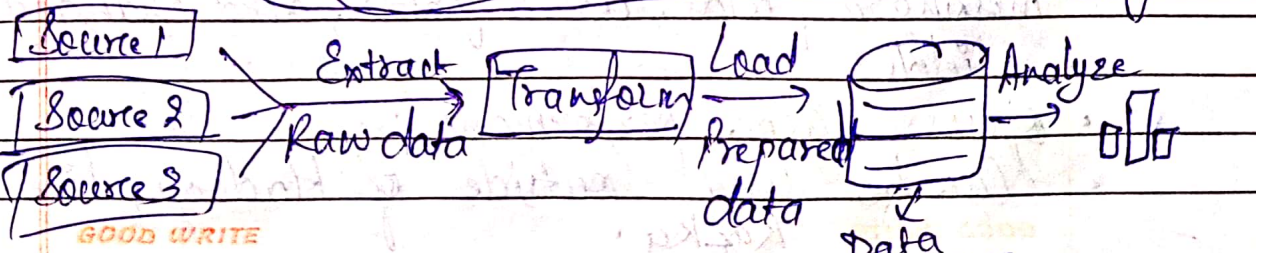
→ challenges to building Data Pipelines :

- Netflix has built its own Data pipeline. However, Building your own pipeline is very difficult and time consuming.

challenges are :

- Connection.
- Flexibility.
- Centralization → (Ex: McDonald's (exact no. of pickles is put on each burger. no matter where you are in the world).)
- Latency (delay)





GOOD WRITE

Ques-6. Dif. file formats of Big Data?

(1) CSV →
- Comma-separated values.
- here data is stored in row-based file format
- mostly used for exchanging tabular data.
- Easier to use
- human-readable.
- widely used format for tabular data representation, but lacks many other capabilities which other formats provides.
- Very slow.

(2) PARQUET →
- open-source file format for Hadoop.
- helps to achieve efficient storage and performance.
- is Column-oriented format.
- it is default format for ~~spark~~ Spark.
- widely used in Hadoop ecosystem to query data.

(3) AVRO →
- row-based storage format.
- widely used for serialization.
- supports dynamic data schemas that change over time
- Can easily handle schema changes like missing fields, added fields, and changed fields.
- offers good performance.
- Can be used outside of Hadoop, like in Kafka.

(4) ORC :-

- optimized Row Columnar.
- provides highly efficient way to store data.
- was designed to overcome the limitations of other file formats.
- improves overall performance when Hive (SQL kind of Interface, built on hadoop) reads, writes, and process the data.
- stores Collection of Rows.
- Mostly used in Hive.

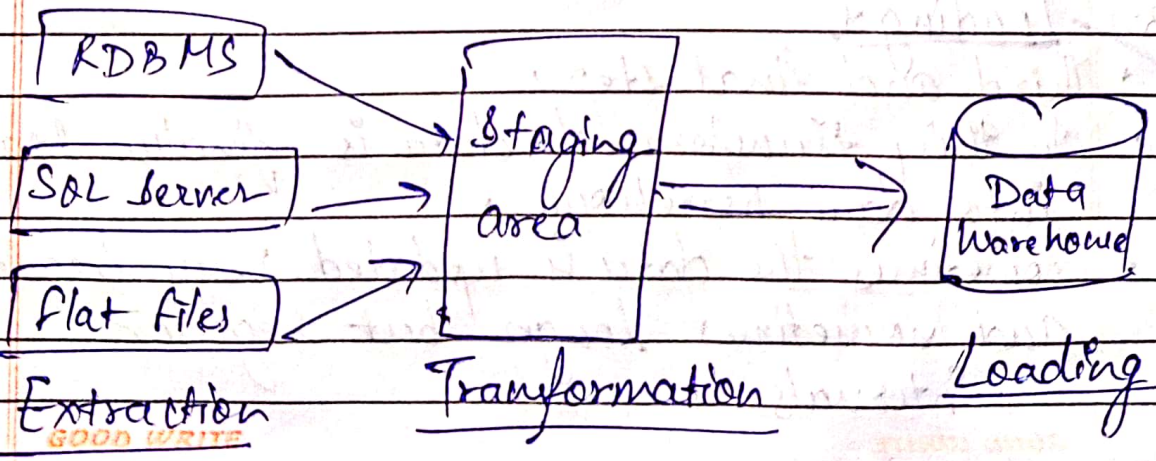
(5) Feather :-

- is a portable file format.
- used for storing Arrow tables or data frames
- ~~Feat~~ is fast
- Lightweight and easy to use binary file format for ~~easy to use~~ storing data frames.
- files are generally faster in read and write performance when used with solid state drives, due to its simpler compression scheme.

Ques. Explain ETL in detail :-

→ ETL stands for Extract, Transform, and Load.

- It used to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse.



(1) Extraction :-

- First step of ETL process.
- ~~Data is~~ In this data is extracted from various sources systems which can be in various formats like Relational databases, NoSQL, XML and flat files in staging Area.
- It is important to extract data from various systems in staging area because data can be corrupted.

So, directly loading it in data warehouse may damage it.

Therefore it is most important step of ETL.

(2) Transformation :-

- second step of ETL.
- In this, a set of rules or functions are applied on the extracted ddta to convert it into a single standard format.
- may involve following process/tasks :-
  - → filtering
  - → cleaning
  - → Joining (joining multiple attributes).
  - → Splitting (splitting a single attribute into multiple attributes).
  - → Sorting (sorting tuples on the basis of some attribute (generally key-attribute)).

(3) Loading :-

- Third and final step.
- In this, transformed data is finally loaded into data warehouse.
- Sometimes the data is updated very frequently and sometimes longer but regular intervals.

GOOD WRITE

- The rate and period of loading solely depends on the requirements and varies from system to system.

→ ETL process can also use pipeline concept ↑

Extract → Transform → Load

→ ETL Tools:

- Hevo, Sybase, Oracle Warehouse builder, Clover ETL, and MarkLogic.

→ Advantages ↑

- Improved data Quality.
- Better data Integration.
- Increased data security.
- Improved scalability.
- Increased Automation.

} Geeks.

→ Disadvantages ↑

- High Cost.
- Complexity.
- Limited flexibility.
- Limited scalability.
- Data privacy Concerns.

} Geeks.