**44517 Exam 1: Study Guide**

1. Given a scenario, be able to write the examples of 5Vs. Ensure your answer to be specific to the scenario.

A: Volume – Huge amount of data.

Velocity – High speed of accumulation of data.

Variety – Different formats of data from various sources.  
 Value – Extract useful data. How the organization derives value from that data should be unique to them.

Veracity – Inconsistency and uncertainty of data. Quality and accuracy of data, overall it refers to the level of trust there is in the collected data.

1. Be able to the know the terms of the memory in bytes (Ex: Is Yotta bytes greater than peta bytes? - false )

Asc order (MEGA, GIGA, TERA, PETA, EXA, ZETTA, YOTTA)

1. Define various terms related to Big Data
   * IoT : The collective network of collective devices and the technology that facilitates communication between devices and the cloud as well as between the devices themselves.
   * CPS(Cyber Physical systems) integrate sensing, computation, control and networking into physical objects and infrastructures connecting them to internet and to each other.
   * Data Streaming
   * Other terms
2. What are the various applications of Big Data in different industries?

Finance, Health Care, Agriculture, Media & entertainment.

Impacts : Advance healthcare, Presidential Campaigns, Sports prediction, Easier commutes, Smartphones, Personalized advertising.

1. What are the disadvantages of relational databases?
2. Maintenance Problem · 2 – Cost · 3 – Physical Storage · 4 – Lack of Scalability · 5 – Complexity in Structure · 6 – Decrease in performance over time. 7. Isolated Information.
3. What are the differences between a centralized database and a distributed database?

|  |  |
| --- | --- |
| **Centralized Database** | **Distributed Database** |
| It is a database that is stored, located as well as maintained at a single location only. | It is a database that consists of multiple databases which are connected with each other and are spread across different physical locations. |
| It has more data consistency. | It may have some data replications thus data consistency is less. |
| A centralized database is less costly. | This database is very expensive. |
| The management, modification, and backup of this database are easier as the entire data is present at the same location. | The management, modification, and backup of this database are very difficult as it is spread across different physical locations. |
| The data access time in the case of multiple users is more. | The data access time in the case of multiple users is less. |

1. What are distributed databases? Give some examples.

A distributed database is a place where data is stored across different networked computers. These databases will operate on two or more interconnected servers on a computer network. The database is stored across multiple locations but appears to users and apps as a single database. An example can be the Walmart store, the data is stored in various branches in multiple locations, and they are operated on different servers on a computer network.

1. Be able to know the difference between homogenous and heterogenous distributed databases.

* In a Homogeneous distributed database, all the sites use similar database management systems and the transaction process is quick. In a Heterogeneous distributed database, different sites have different database management systems and transaction processing is very complex.
* Similar type of software was used by all sites and each site is aware of all other sites in homogeneous whereas in heterogeneous different sites will use different software and each site is not aware of all other sites.

1. What is the difference between software engineering and data engineering?
   * Software engineering is the process of designing, creating, testing, and maintaining software. It involves the application of engineering principles and practices to software development. Software engineers focus on the development of software systems, including software architecture, design, coding, testing, deployment, and maintenance.
   * Data engineering, on the other hand, is the process of designing, building, testing, and maintaining the systems and infrastructure required to manage and analyze large volumes of data. Data engineers focus on developing the tools and systems necessary to store, process, and analyze large amounts of data efficiently and effectively. This involves working with databases, data warehouses, data lakes, big data platforms, and other data-related technologies.
   * While both software engineering and data engineering involve designing, building, and maintaining complex systems, they differ in terms of their focus and goals. Software engineering is primarily concerned with building software systems that meet the needs of users, while data engineering is primarily concerned with building systems that can store, process, and analyze large amounts of data efficiently and effectively.
2. What are the primary job responsibilities of data engineer?
   * Designing, building, and maintaining data pipelines and data infrastructure: Data engineers are responsible for designing and building data pipelines that extract data from various sources, transform it into a usable format, and load it into a target system. They also maintain data infrastructure, including databases, data warehouses, and data lakes.
   * Data modeling: Data engineers create data models that enable data to be stored, processed, and accessed in an efficient manner. This involves designing the schema of the database or data warehouse, selecting appropriate data types, and defining relationships between tables.
   * Performance tuning: Data engineers optimize data infrastructure to ensure that data can be accessed and analyzed in a timely and efficient manner. This may involve tuning database queries, optimizing ETL processes, or implementing caching mechanisms.
   * Data quality management: Data engineers ensure the accuracy, completeness, and consistency of data by implementing data quality checks and data validation rules.
   * Collaboration with data scientists and analysts: Data engineers work closely with data scientists and analysts to understand their data requirements and to provide them with access to the data they need. They may also be involved in designing and building data visualizations and dashboards.
   * Security and compliance: Data engineers are responsible for ensuring that data is secure and compliant with regulations such as GDPR and HIPAA. They implement security measures such as encryption and access controls, and they ensure that data is stored and processed in a way that complies with relevant regulations.
3. What is your favorite job title in data science related field? Why?
   * **Hadoop Developer**
   * Proficient in Big data application development toolset to design, develop, test, deploy, maintain, and improve software, extract and isolate data clusters. Strong understanding of techniques such as Continuous Integration, Continuous Delivery, Test Driven Development, and Cloud Development.
4. What are the advantages and disadvantages of Hadoop?
   * **Cost:** Hadoop is open-source and uses cost-effective commodity hardware which provides a cost-efficient model.

• **Scalability:** Hadoop is a highly scalable model. A large amount of data is divided into multiple inexpensive machines in a cluster which is processed parallelly.

• **Flexibility:** Hadoop is designed in such a way that it can deal with any kind of dataset like structured(MySql Data), Semi-Structured(XML, JSON), Un-structured (Images and Videos) very efficiently.

* **Speed:** Hadoop uses a distributed file system to manage its storage i.e. HDFS (Hadoop Distributed File System).
* **Fault Tolerance**: Hadoop uses commodity hardware(inexpensive systems) which can be crashed at any moment. data is replicated on various data nodes in a Hadoop cluster which ensures the availability of data if any of your systems get crashed.
* **Minimum Network Traffic:** each task is divided into various small sub-task which is further assigned to each data node available in the Hadoop cluster.
  + Disadvantages of Hadoop:
  + **Problem with Small files:** Hadoop can efficiently perform over a small number of files of large size. Hadoop stores the file in the form of file blocks which are from 128MB in size (by default) to 256MB. Hadoop fails when it needs to access a small size file in a large amount. These small files in large number surcharge the name node and make it difficult to work.
  + **Vulnerability:** Hadoop is a framework that is written in java, and java is one of the most used programming languages which makes it more insecure as it can be easily exploited by any cyber-criminal.
  + **Lack of Security:** Data is everything for an organization, by default the security feature in Hadoop is made un-available. Hadoop uses Kerberos for security purpose which is not easy to manage.
  + **High Up Processing:** Read/Write operation in Hadoop is immoderate since we are dealing with large size data which is in TB or PB.
  + **Supports Only Batch Processing:** The batch process is nothing but the processes that run in the background and does not have any kind of interaction with the user. The engines used for these processes inside the Hadoop core is not much efficient.

1. What are the two questions you need answer for a Map-Reduce problem?
   * Map: In the map job, we split the input dataset into chunks

• Map task processes these chunks in parallel.

* + On the map we use outputs as inputs for the reduce tasks.

• Reducer process the intermediate data from the maps into smaller tuples, that reduces the tasks, leading to the final output of the framework

* + What is the transformation that needs to be applied to the input data in order to obtain a set of intermediate key-value pairs?
  + Reduce question: What is the transformation that needs to be applied to the intermediate key-value pairs in order to obtain the final output?

1. Given a scenario, explain the different stages to solve Map-Reduce problem.
2. Be able to know how to write source code for a Map-Reduce problem.
3. Given a source code snippet, be able know what the functionality of the code. (Similar to our code comprehension activiti es.)