Q1.

"Datasets are generally quite large, taxing capacities of main memory, local disk, and even remote disk. We call this problem as Big data," was valid when the term "big data" first came into use, and it still is today, but with some considerations and context.

Volume: As innovation advances, an ever-increasing number of information is being produced across different spaces, from online entertainment cooperations and IoT gadgets to logical examination and deals. This rising information volume remains a main trait of enormous information.

Variety: Enormous information isn't just about size yet in addition about the range of information types. It incorporates organized conventional data sets, semi-organized information, and unstructured information. Overseeing and removing esteem from this different scope of information sources is a critical test.

Velocity: The speed at which information is produced and should be handled is one more part of large information. Ongoing and close constant information handling has become critical in different applications, like monetary exchange, misrepresentation discovery, and observing virtual entertainment patterns.

Veracity: Veracity alludes to the quality and dependability of information. With the wealth of information from different sources, guaranteeing information exactness and dependability stays a test. Information might contain blunders, irregularities, or predispositions that should be tended to.

Value: Removing significant bits of knowledge and worth from large information is a definitive objective. This requires progressed examination, AI, and information mining methods. In order to stay ahead of the competition and make well-informed decisions, businesses must master the complexities of big data.

The idea of enormous information has developed past these conventional "3Vs" (Volume, Assortment, and Speed) to incorporate extra aspects like:

Variability: Organizations must adapt to deal with the variability of data, which can be inconsistent or unpredictable in its flow.

Complexity: Enormous information frequently includes complex connections, and the information might be appropriated across numerous areas or stages.

Security and privacy: Data security and privacy concerns have grown to be major concerns as a result of the large amount of data being collected. Guidelines like GDPR and CCPA mirror the significance of safeguarding people's information.

Ethics: The moral utilization of information has acquired huge consideration. Associations need to consider the moral ramifications of their information rehearses, particularly while managing delicate or individual data.

Cost: Putting away and handling huge volumes of information can be expensive, so cost administration is an element to consider.

Q2.

Value Creation: Big data offers insightful and useful information on consumer preferences, behaviours, and market trends. Businesses can use this information to improve decision-making, create cutting-edge goods and services, and streamline operations.

Competitive Advantage: Big data can be a useful resource that is difficult for rivals to duplicate or replace. Businesses with a competitive advantage in the market use big data effectively.

Innovation: Big data helps businesses find new business prospects and fosters innovation. It can be used to enhance product design, streamline operations, and spot new trends that might inspire the creation of fresh goods and services.

Customer Engagement: Big data enables organizations to more fully comprehend their clients, personalize services, and enhance client relationships. This results in more customers.

Efficiency in Operations: Big data may be utilized to save costs and increase efficiency in Operations. It can, for instance, aid in the optimization of inventory control, resource allocation, and supply chain management.

Q3.

The article emphasizes that data storage can be expensive for enterprises because of a number of variables with regard to cost:

Data Volume: As a result of continuously rising data production and collection, enterprises now require more storage. Large data sets can be expensive to store, especially if they are unstructured.

Infrastructure costs: Keeping servers, storage devices, and data centers, all essential components of the infrastructure required for data storage up and running can be expensive.

Data Security: Investing in data security techniques and technologies is necessary to protect data from loss, damage, or theft, which raises the overall cost.

Costs associated with compliance: Many firms are subject to data protection and privacy requirements, and maintaining compliance with these regulations frequently entails extra expenses for managing and storing data.

Reputational Costs: Data breaches or poor data handling can result in legal penalties and reputational harm, raising the cost of data management even higher.

Q4.

The article highlights a number of detrimental outcomes that big data can have on businesses, such as:

Costs: Businesses may not always see a return on investment when collecting, storing, and analyzing big data.

Privacy Issues: Using big data creates privacy issues, particularly when it entails gathering and analyzing personal data. Data handling errors might result in privacy violations and legal problems.

Reputation Risk: As evidenced by situations like the Facebook-Cambridge Analytica incident discussed in the article, improper data practices can harm a company's reputation.

Data quality: It might be difficult to guarantee the dependability, correctness, and quality of huge data. Decisions and insights may be wrong as a result of poor data quality.

Volume: The sheer amount of big data can be overwhelming, making it challenging for enterprises to get value and insightful information from it.

Lack of Expertise: Big data analysis calls for specialized knowledge and abilities, and many firms may not have them in order to fully utilize their data assets.

Conclusion: Big data presents huge opportunities for businesses, but it also carries dangers and problems that must be carefully managed if its full potential is to be realized.

Q5.

Companies struggle to manage their tremendous volume of information for various reasons, as indicated by the article "The hard bits of insight about the information supportability in the midst of the enormous information frenzy," including:

Data Overload: How much information is being created and assembled is growing dramatically. It very well may be hard for some organizations to deal with and examine such enormous measures of information enough.

Data Types: There are various kinds of information, for example, organized, unstructured, and semi-organized information. It tends to be testing and asset concentrated to oversee and get bits of knowledge from different information types.

<u>Costs</u>: Overseeing and putting away critical measures of information can be expensive. Organizations might spend a great deal on information capacity, especially when a sizable part of the information is excess or pointless.

<u>Data quality</u>: Accurate and high-quality data are essential for making wise decisions. Inaccuracies and inefficiency in data processing can result from poor data quality.

<u>Privacy and Compliance</u>: Data privacy laws, like the GDPR, oblige businesses to manage data sensibly and uphold the rights of individuals to privacy. Data management becomes more challenging when compliance is achieved.

Q6.

In regards to the acronym "ROT," it stands for "redundant, obsolete, and trivial data." Companies frequently have a lot of ROT data because:

Data redundancy occurs when there are several copies or locations of the same piece of information within an organization.

Data can become out-of-date or irrelevant over time, yet organizations may still keep it.

Triviality: Some data may be stored despite having little to no value or significance.

For a variety of reasons, including a lack of data management standards, ambiguity around data ownership, and a historical unwillingness to delete data out of concern for losing potentially useful information, companies amass ROT data. Therefore, ROT data uses up storage space and adds to the problems with data management and sustainability.

Q7.

I believe Apache HBase is an essential tool for organizations. Here are a few reasons why Apache HBase is crucial:

<u>Scalability</u>: Apache HBase is highly scalable and built to handle enormous volumes of data. Petabytes of data may be accommodated, and by adding additional servers to the cluster, it is simple to extend horizontally. This scalability is essential for businesses with constantly increasing data volumes.

Distributed Architecture: HBase was developed in a distributed setting and is based on the Hadoop Distributed File System (HDFS). Because of its high availability and fault tolerance, this architecture is appropriate for mission-critical applications.

Low-latency: Low-latency data retrieval and real-time data processing are two areas where HBase shines. This is vital for applications like fraud detection, recommendation engines, and monitoring tools where quick access to data is required.

HBase uses a columnar storage style, which is effective for retrieving specific data. It enables organizations to swiftly get particular columns or attributes from big datasets. This function is useful for reporting and analytics.

Strong Consistency: HBase has strong consistency guarantees that make sure all of the cluster's data is accurate and up to date. For applications like financial services and healthcare where data integrity is crucial, this is essential.

Schema Flexibility: HBase provides some degree of schema flexibility while being recognized for its columnar storage. Data models are adaptable for changing business needs since organizations can do so without significant interruptions.

Integration with the Hadoop Ecosystem: HBase combines easily with other parts of the Hadoop ecosystem, including Apache Spark and Hadoop MapReduce. Organizations are now able to analyze and analyze complicated data on data stored in HBase thanks to this connectivity.

HBase is open source and cost-effective because it is free to use, making it a good choice for businesses. It eliminates the requirement for pricey proprietary database solutions, making it an appealing option for organizations on a budget.

Numerous Use Cases: Apache HBase is adaptable and may be used in a wide range of settings and use cases. Applications like log storage, time-series data analysis, recommendation systems, and fraud detection are frequently utilized in sectors including finance, e-commerce, social media, and more.