BT:

1.a) Explanation of Big Data:

Big data refers to extremely large and complex datasets that cannot be easily managed, processed, and analyzed using traditional data processing methods. Examples of big data include:

Social media data: Platforms like Facebook, Twitter, and Instagram generate massive amounts of data every second in the form of user posts, comments, likes, shares, and profiles.

E-commerce transaction data: Online shopping platforms record details of every purchase, including customer information, product details, prices, and timestamps.

Sensor data: Internet of Things (IoT) devices such as smart thermostats, fitness trackers, and industrial sensors generate continuous streams of data on temperature, movement, usage patterns, and more.

b) Big Data Analysis for Increasing Business Revenue:

Big data analysis can help increase business revenue in several ways:

Customer segmentation: By analyzing customer data, businesses can segment their customers based on demographics, behavior, and preferences. This allows for targeted marketing campaigns and personalized offers, increasing the likelihood of customer engagement and purchase.

Demand forecasting: Analyzing historical sales data and other relevant factors can help businesses predict future demand for their products or services. This enables better inventory management, production planning, and resource allocation, reducing costs and increasing sales.

Fraud detection: Big data analysis can detect patterns of fraudulent activity in financial transactions, insurance claims, and other business processes. This helps businesses reduce losses due to fraud and improve overall operational efficiency.

Operational optimization: By analyzing data from various business processes, such as supply chain, manufacturing, and customer service, businesses can identify bottlenecks and inefficiencies and optimize their operations for better performance and cost savings.

c) Difference between Structured and Unstructured Data:

Structured data is organized in a predefined format and can be easily stored, queried, and analyzed using traditional database management systems. Examples of structured data

include relational database tables with rows and columns, spreadsheets, and CSV files.

Unstructured data, on the other hand, does not have a predefined format and is difficult to analyze using traditional methods. Examples of unstructured data include text documents, images, videos, audio files, and social media posts.

The main differences between structured and unstructured data are:

Format: Structured data has a defined format, while unstructured data does not.

Ease of analysis: Structured data is easier to analyze using traditional methods, while unstructured data requires specialized tools and techniques.

Volume: Unstructured data often constitutes a large portion of the total data generated by organizations, while structured data is relatively smaller in volume.

d) Diagram Representing Big Data Contribution:

As an AI, I don't contribute to big data in the same way as a human being. However, a diagram representing big data could include various sources such as social media, e-commerce platforms, IoT devices, and enterprise systems. The data from these sources is collected, processed, and analyzed using big data technologies such as Hadoop, Spark, and NoSQL databases. The analyzed data can then be used for various applications such as business intelligence, predictive analytics, and machine learning.

e) We can leverage data-intensive systems to handle large volumes of data, access data quickly, and it often needs to respond to user requests in real-time or near-real-time environments. It include data sources, data storage, data processing, data analysis, data transmission, data security and processing, user interfaces and applications, monitoring and maintenance,

f) **Data storage：** Relational databases, NoSQL databases, and data lakes

Data visualisation and analysis： Business intelligence tools, data analysis platforms, visualization libraries

Compute and distributions： Distributed computing framework, stream processing framework, container transformation

Data warehouses： Cloud data warehouse, data warehouse solution, data warehouse

MT:

a) Similarities to Oil:Just as oil was a crucial resource that drove economic growth and

industrialization in the past, data is playing a similar role in the 21st century.

Economic Value: Oil had immense economic value as it was the primary source of energy for transportation, manufacturing, and power generation. Similarly, data has significant economic value. It is used by businesses to gain insights into customer behavior, optimize operations, develop new products and services, and make more informed decisions. This leads to increased revenue and competitiveness.

Drives Innovation: The discovery and exploitation of oil led to numerous technological innovations such as the internal combustion engine and the development of the petrochemical industry. Likewise, data is driving innovation in various fields such as artificial intelligence, machine learning, and the Internet of Things. These technologies are transforming industries and creating new business models.

Scarcity and Competition: Oil is a finite resource, and access to it has led to competition and geopolitical tensions. Similarly, data can be considered a scarce resource in some cases. Companies compete to collect and analyze data to gain a competitive advantage. There are also concerns about data privacy and security, which can limit access to data.

b)Accuracy refers to whether the information recorded in the data and the data are accurate, and whether there are any anomalies or erroneous information. The definition of the first sentence not only includes whether the information and data itself is correct, but also explains that "whether the process of recording the information and data is correct", "whether the source of the data is complete", and "whether the restricted model is approximate" are all factors related to "accuracy".

The second sentence takes the macro perspective of "data quality issues" and covers multiple other dimensions that affect data quality, to illustrate the importance of "accuracy" and the correlation with other factors such as "feasibility" and "reputation".

AT:

a) The IMDb dataset. It contains a large amount of movie information such as movie titles, directors, actors, ratings, plot summaries, and more. This dataset can be used to analyze movie trends, the popularity of actors, the distribution of movie genres, and so on.

b) Reasons why this dataset is suitable for data-intensive systems:

Large volume and richness of data: This dataset contains extensive multi-faceted information about a large number of movies, meeting the need of data-intensive systems to process and analyze a large amount of data. It can handle data from

thousands of movies, providing a sufficient data foundation for in-depth analysis.

Multi-dimensional information: The multi-dimensional information such as movie titles, directors, actors, ratings, and plot summaries enables various complex analyses. For example, one can analyze the collaboration patterns of directors and actors to explore the interpersonal network in the movie industry; use rating data for user preference analysis and movie quality evaluation; and conduct text mining based on plot summaries to explore the thematic characteristics of different movie genres.

Supports multiple analysis needs: It can meet different analysis purposes. For example, analyzing movie trends can be achieved by observing changes in movie genres, themes, and ratings over different time periods; the popularity of actors can be determined by counting indicators such as the number of movies an actor has appeared in, ratings, and social media mentions; and analyzing the distribution of movie genres can help understand the proportion and development trends of different movie types in the market.

Real-time and updatability: As new movies are continuously released and audience evaluations change, the IMDb dataset is also constantly updated. This allows data-intensive systems to continuously obtain new data, conduct real-time analysis and dynamic monitoring, and adapt to the constantly changing movie market environment.