Department Name – Computer Science & Engineering

Roll no – 20BCE057

Name – Devasy

Subject Name and Code – 2CS702 Big Data Analytics

Practical No – 2

AIM: Identify the data sources for big data. Find the technological limitations of conventional data analysis algorithms to perform analytics on big data. Justify your answer with any one of the applications.

1) Data sources for Big Data

Big Data refers to the accumulation of data in large pools and quantities. It is a collection of organised, semi-structured, and unstructured data gathered by businesses with the objective of improving their services. It provides valuable information that drives the innovation and decision-making in any company in any sector. Few sources of Big Data are as follows:

- **Social Media Platforms**: Billions of people engage with social media daily, sharing their thoughts, experiences, and preferences. It enables the companies to better understand their consumers and accordingly, personalize the plans offered by creating a user profile.

- **Internet of Things (IoT)**: IoT devices, such as smart sensors, wearables, and connected appliances, gather real-time data. From health and fitness data collected by fitness trackers to environmental data like Air Quality Index captured by smart cities' sensors, varied information is generated by IoT devices This data can be analysed to optimize processes, improve efficiency, and enhance overall experiences for individuals and communities.

- **E-Commerce Platforms:** E-Commerce platforms like Amazon and Flipkart have made shopping much easier, everything is available at the click of a thumb. The consumer also demands everything ready and fast. Thus, it is of prime importance to study the user’s past preferences for suitable personalized recommendations. Thus, they analyse the clicks, searches, likes, comments and past purchases plus their trackers also scan other websites.

- **Government Agencies:** Government agencies and public organizations play a crucial role in generating Big Data. Census data, public health records, and administrative data make up a major source of information that helps policymakers in decision-making and resource allocation. Analysing this data assists in identifying societal trends, addressing public issues, and efficient planning for the future.

- **Banks and Financial Institutions**: Banks, credit card companies, and financial service providers collect extensive data on transactions, customer behavior, and economic trends. Analyzing this data helps in detecting fraudulent activities, assessing credit risks, and offering tailored financial solutions to customers.

Dataset Chosen: Market Correlation of Data
Market Segmentation Data involves tracking and recording customer purchase patterns, including the frequency of purchases and which products are often bought together. This information is used to create user profiles and generate relevant product recommendations.

Analyzing Big Data

1) The Clustering algorithm is a widely used technique for association rule mining, but it has certain limitations when applied to Big Data analytics:
2) Scalability: The Clustering algorithm becomes computationally expensive when dealing with large datasets, such as gigabytes (GB) and terabytes (TB) of data. Its performance degrades significantly as the dataset size increases.
3) Diverse Datasets: In the retail domain, shopping malls offer a diverse range of products, with new items continually introduced and old ones phased out. This diversity increases the number of candidate itemsets, making it challenging to generate meaningful rules.
4) Rapid Data Generation: Big Data is generated at an astonishing rate, and the Apriori algorithm may struggle to keep up with the pace of data generation. Real-time processing of such data can be a significant challenge.
5) Rule Quality: The Apriori algorithm generates strong rules based on support and confidence metrics. However, not all strong rules are necessarily interesting. Some rules may appear misleading when

For instance, a high-confidence rule may suggest that customers who purchase bananas are likely to buy apples. Still, in reality, a significant portion of all customers may buy apples, making the rule less valuable.

In summary, while the Apriori algorithm is a valuable tool for association rule mining, it may face scalability and interpretability challenges when dealing with Big Data in dynamic and diverse domains like retail.
Custom implementations and optimizations
are often required to address these limitations
effectively.