Roll No: 20BCE204
Course Name and Course 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. Big Data:

The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs.

- Volume
- Velocity
- Variety

Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems, you wouldn't have been able to tackle before.

2. Big Data in Transportation:

Big Data has revolutionized the travel industry by leveraging the vast amount of data generated from travelers, booking platforms, hotels, airlines, and other related sources. Its applications have significantly transformed how travel companies operate, enhance customer experiences, and optimize their services.

1. Personalized Travel Experience: Big Data enables travel companies to analyse vast amounts of customer data, including preferences, past behaviours, and feedback, to create personalized travel experiences. This personalization ranges from tailored travel recommendations and customized itineraries to targeted marketing campaigns.
2. Dynamic Pricing: The travel industry extensively uses dynamic pricing models, where the cost of flights, hotel rooms, and other travel services changes based on demand and availability. Big Data analytics plays a crucial role in analysing real-time data and market trends to set optimal pricing strategies.
3. Fraud Detection and Security: With the rise of online booking platforms, fraud and security concerns have also increased. Big Data analytics can identify suspicious transactions, patterns, and anomalies, helping travel companies mitigate fraud risks and ensure secure transactions.
4. Demand Prediction and Capacity Planning: Big Data allows travel companies to predict demand patterns, identify peak travel periods, and optimize capacity planning for airlines, hotels, and other travel services. This ensures better resource allocation and maximizes revenue opportunities.
5. Sentiment Analysis and Customer Feedback: Social media and online review platforms generate massive amounts of customer feedback. Big Data tools can analyse this unstructured data, providing valuable insights into customer sentiments and helping companies improve their services and offerings.

3. Data Sources:

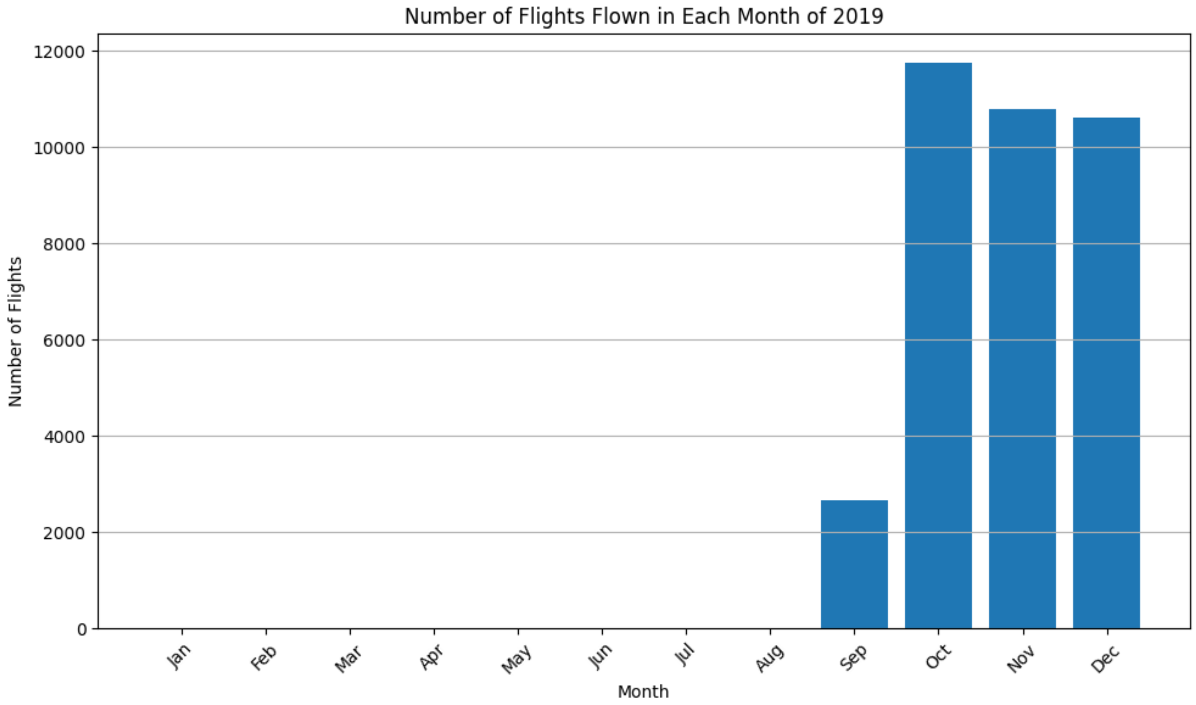
Booking Platforms	Online travel agencies (OTAs) gather massive amounts of data on hotel bookings, flight reservations, vacation rentals, and more.
Customer Reviews and social media	Platforms like TripAdvisor, Yelp, and social media networks are valuable sources of customer reviews, ratings, and feedback on hotels, restaurants, attractions, and destinations.
Airline and Airport Data	Airlines collect vast amounts of data on flight operations, passenger bookings, loyalty programs, and frequent flyer data. Airports generate data on passenger traffic, flight schedules, baggage handling, security wait times, and retail sales.
Location-Based Services	Mobile apps and platforms that provide location-based services offer data on travelers' movements, preferences, and interactions with local businesses.
Government and Tourism Boards	Governments and tourism boards collect data on visitor arrivals, hotel occupancy rates, tourist spending, and tourism-related policies
Weather and Climate Data	Weather data is crucial for travel planning, especially for outdoor activities and popular tourist destinations
Traffic and Transportation Data	Data on traffic conditions, public transportation schedules, and transportation infrastructure are important for travelers navigating within cities or between destinations.
Geospatial Data	Geographic information systems (GIS) like FlightRadar, FlightAware, etc provide spatial data about landmarks, points of interest, and geographical features, enriching travel applications with maps and navigation services.

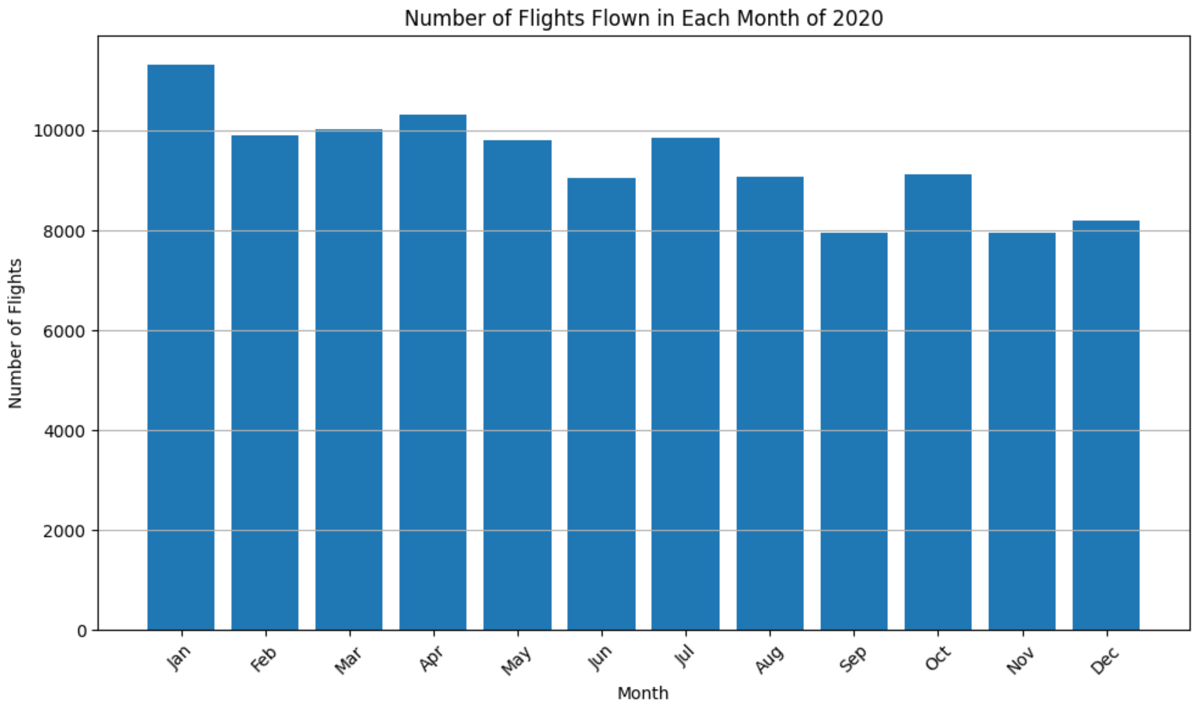
4. Technological limitations of conventional data analysis algorithms:

- **Memory Constraints**: Conventional algorithms assume that data can fit entirely in memory. With big data, datasets often exceed available memory, leading to performance degradation or crashes as systems struggle to process and analyse data that cannot be loaded entirely into RAM.
- **Processing Speed**: Traditional algorithms may not be optimized for parallel processing or distributed computing, which are essential for handling the immense volume and complexity of big data. As a result, processing times can become impractical and hinder real-time or timely insights.
- **Scalability**: Traditional algorithms may lack the scalability required to handle data that grows exponentially. Adding more data may lead to diminishing returns or complete breakdowns in analysis, rendering the algorithms inadequate for the demands of big data.

- **Storage Efficiency**: Storing and managing large datasets is a challenge for conventional algorithms, especially if they rely on certain data structures or storage formats that are inefficient for big data storage and retrieval.
- **Data Variety**: Big data often comes in various formats, including structured, semi-structured, and unstructured data. Conventional algorithms might struggle to handle diverse data types effectively, limiting their applicability in extracting meaningful insights from such data.
- **Dimensionality and Feature Space**: Big data can have high dimensionality, where the number of features or attributes is vast. Traditional algorithms might face the "curse of dimensionality," leading to increased computational complexity and decreased accuracy.
- **Noise and Quality**: With large and diverse data sources, big data often contains noisy, incomplete, or low-quality information. Conventional algorithms may struggle to handle such noise and might produce unreliable results.
- **Real-Time Analytics**: Many conventional algorithms are designed for batch processing and may not be suitable for real-time or streaming data analytics, limiting their ability to provide immediate insights.
- **Algorithmic Efficiency**: Some conventional algorithms might have suboptimal time or space complexity for big data analytics, making them inefficient and resource-intensive.
- **Interactivity and Visualization**: Big data analytics often require interactive exploration and visualization. Traditional algorithms might not provide the responsiveness and interactivity required for effective exploration of large datasets.

5. Problems with the big data:





This graph contains data about number of flights flown in each month of particular year.

The problems are :

Scalability

It is a crucial requirement in the field of big data analytics, as it addresses the challenges posed by the ever-increasing volume and speed of data generation. As data accumulates rapidly, it becomes essential for systems to adapt and expand their resources dynamically. By ensuring that computing infrastructure can scale up to handle growing data sizes, scalability prevents performance bottlenecks and ensures timely processing and analysis. This is particularly vital in industries like finance and e-commerce, where real-time insights are critical for staying competitive.

Furthermore, scalability optimizes resource utilization by allowing workloads to be distributed across multiple machines. This harnesses the power of parallel processing, reducing processing times and expediting data analysis. As a result, decision-making processes are enhanced, and organizations can extract valuable insights more efficiently. Scalability also enables seamless integration of new data sources and types without requiring a complete overhaul of existing infrastructure. This agility allows organizations to adapt to evolving data requirements without disruption.

In summary, scalability empowers big data analytics to operate seamlessly with massive datasets, driving real-time insights, efficient resource utilization, and strategic agility across diverse industries and sectors.

High Availability

High availability (HA) is a critical necessity in big data analytics, ensuring continuous and reliable access to data and services. With vast volumes of data being processed, any interruption or downtime can lead to significant losses and missed opportunities. HA achieves this by providing redundant and resilient components, safeguarding against failures and disruptions. It enables seamless failover to backup systems, minimizing downtime and ensuring uninterrupted access to real-time insights, supporting timely decision-making.

In the dynamic realm of big data, a robust HA strategy is essential to accommodate changing demands. Organizations increasingly rely on real-time analytics and AI-driven insights, making HA crucial for business operations and customer experiences. By ensuring data availability during hardware failures or maintenance, HA reinforces data integrity, compliance, and customer trust.

In essence, High Availability is a linchpin in big data analytics, safeguarding against interruptions and consistently harnessing the potential of data-driven decision-making across industries and sectors.

Fault Tolerance

It is a critical component in big data systems, protecting against unexpected failures that can disrupt data processing and analytics. In the complex world of big data, where failures are inevitable, fault tolerance ensures that such incidents do not lead to catastrophic consequences. By using redundancy and backup mechanisms, fault-tolerant systems can seamlessly switch to alternative resources when failures occur, preserving data integrity and availability for critical decision-making processes.

The significance of fault tolerance is magnified in large-scale and distributed big data environments, where the likelihood of failures increases. Without fault tolerance, the entire analytics pipeline could be affected, impeding operations and delaying insights. It extends to data storage, where redundant systems safeguard against data loss or corruption. As organizations rely more on big data for decision-making and innovation, integrating fault tolerance strategies becomes essential. By preparing for failure and designing systems to handle disruptions gracefully, fault tolerance ensures data continuity and reinforces the resilience of big data infrastructures.

Overall, fault tolerance is a crucial safeguard in big data systems, providing resilience and adaptability in an uncertain technological landscape.