Name:Kaushik Kotian Roll No.:30 Div:D15B Batch:B

Experiment No : 9

Aim: Case study on Power BI and Apache Spark :

Theory:

What is PowerBI ?

Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. Your data might be an Excel spreadsheet, or a collection of cloud-based and on-premises hybrid data warehouses. Power BI lets you easily connect to your data sources, visualize and discover what's important, and share that with anyone or everyone you want.

Power BI consists of several elements that all work together, starting with these three basics:

- A Windows desktop application called Power BI Desktop.
- An online software as a service (SaaS) service called the Power BI service. ● Power BI Mobile apps for Windows, iOS, and Android devices.

Advantages of Power BI:

1. **User-Friendly Interface:** Power BI offers an intuitive interface that allows users to create visually appealing reports and dashboards without extensive technical expertise.
2. **Data Connectivity:** It provides seamless connectivity to a wide range of data sources, enabling users to import data from multiple sources and combine them for analysis.
3. **Interactive Visualizations:** Power BI offers a variety of interactive visualization options such as charts, graphs, maps, and slicers, allowing users to explore data dynamically and gain deeper insights.
4. **Real-Time Analytics:** Users can perform real-time analysis of their data and set up automatic data refresh schedules to ensure that reports are always up-to-date.
5. **Collaboration and Sharing**: Power BI enables users to publish reports to the Power BI Service, share them with colleagues or clients, and collaborate on dashboards in real-time.
6. **Security and Governance:** It offers robust security features, including role-based access control, row-level security, and data encryption, to ensure that sensitive information is protected.

Working of PowerBI:

1. **Data Acquisition:** Power BI connects seamlessly to diverse data sources like databases, files, and cloud services, allowing users to import data effortlessly.

2. **Data Preparation:** Within Power BI Desktop, users clean, transform, and model the data to ensure accuracy and relevance for analysis.
3. **Report Creation:** Utilizing intuitive drag-and-drop functionality, users craft visually engaging reports comprising various charts, graphs, and tables to represent insights effectively.
4. **Data Analysis:** Power BI empowers users to delve into data intricacies, create calculated fields, and apply filters dynamically, enabling comprehensive analysis and discovery of trends.
5. **Dashboard Creation:** Users amalgamate key visualizations into interactive dashboards, providing a consolidated view of critical metrics for holistic monitoring and decision-making.
6. **Publishing and Sharing:** Completed reports and dashboards are published to the Power BI service, accessible to designated users for viewing and collaboration.
7. **Consumption and Collaboration:** Stakeholders access published content via web browsers or mobile apps, fostering collaboration and enabling data-driven decisions across the organization.

Disadvantages of Power BI:

1. **Data Size Limitations:** Power BI has limitations on the size of datasets that can be imported and processed, particularly in the free or lower-tier versions. Large datasets may require additional processing or filtering before being loaded into Power BI.
2. **Complexity for Advanced Analytics:** While Power BI offers robust capabilities for visual analytics and reporting, it may lack some advanced analytics features compared to dedicated statistical or data science tools. Performing complex analyses or building advanced machine learning models directly within Power BI may require additional integration with other tools or platforms.
3. **Dependency on Internet Connectivity:** Power BI's cloud-based nature means that users heavily rely on internet connectivity for accessing and refreshing reports and dashboards. Offline access or limited connectivity scenarios may pose challenges for users who need to access and analyze data in remote or disconnected environments.

Real-Life Applications of Power BI:

Supply Chain Analytics at Amazon:

Amazon, being one of the world's largest e-commerce and logistics companies, relies heavily on efficient supply chain management to ensure timely delivery of millions of products to customers worldwide. Power BI plays a crucial role in analyzing and optimizing various aspects of Amazon's supply chain operations:

1. **Inventory Management**: Amazon uses Power BI to track inventory levels in real-time across its vast network of fulfillment centers, warehouses, and distribution hubs.

Visualizations and reports generated by Power BI provide insights into inventory turnover rates, stock levels, and demand forecasting, enabling Amazon to optimize inventory replenishment strategies and reduce stockouts.

2. **Logistics Optimization:** Power BI helps Amazon analyze transportation routes, shipment volumes, and delivery times to optimize its logistics operations. By visualizing data on shipping costs, carrier performance, and delivery routes, Amazon can identify inefficiencies, minimize shipping delays, and reduce transportation costs.
3. **Demand Forecasting:** Power BI enables Amazon to forecast demand for products based on historical sales data, market trends, and seasonal fluctuations. Accurate demand forecasting allows Amazon to anticipate customer demand, optimize inventory allocation, and ensure that popular items are readily available for purchase.
4. **Product Lifecycle Management:** Amazon uses Power BI to track the lifecycle of products from procurement to disposal, analyzing sales performance, customer feedback, and product returns. Insights derived from Power BI help Amazon make data-driven decisions regarding product pricing, promotion, and discontinuation.
5. **Warehouse Optimization:** Power BI helps Amazon optimize the layout and operations of its fulfillment centers and warehouses by analyzing data on order processing times, storage capacities, and workforce productivity. By identifying bottlenecks and inefficiencies, Amazon can streamline warehouse operations and improve order fulfillment speed.

Apache Spark:

Apache Spark is an open-source distributed computing system that provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. It is designed for fast and efficient processing of large-scale data sets across clusters of computers. Spark offers a wide range of libraries and tools for various tasks such as data processing, machine learning, streaming, and graph processing.

Working of Apache Spark:

1. **Resilient Distributed Dataset (RDD):** The fundamental data structure in Spark is RDD, which represents an immutable collection of objects that can be processed in parallel across a cluster. RDDs are created from external data sources or by transforming existing RDDs through operations like map, filter, and reduce.
2. **Distributed Computing:** Spark distributes data across nodes in a cluster and processes it in parallel. It achieves fault tolerance through lineage tracking, where each RDD records the transformations used to create it, allowing lost data to be recomputed from the original data source.

3. **Lazy Evaluation:** Spark uses lazy evaluation, meaning that transformations on RDDs are not executed immediately. Instead, they are optimized and compiled into an execution plan, which is then executed when an action is triggered.
4. **Directed Acyclic Graph (DAG) Execution**: Spark constructs a DAG of transformations and actions based on the user's code. This DAG represents the logical flow of data processing and is optimized for efficient execution across the cluster.
5. **In-Memory Processing:** Spark leverages in-memory computing to cache intermediate results in memory, reducing the need for disk I/O and improving performance. This is particularly advantageous for iterative algorithms and interactive data analysis.

Advantages of Apache Spark:

1. **Speed:** Spark is known for its high performance and speed, thanks to in-memory processing and efficient distributed computing across clusters.
2. **Ease of Use:** Spark provides a simple and expressive API in various programming languages like Scala, Java, Python, and R, making it accessible to a wide range of developers.
3. **Versatility:** Spark offers a rich set of libraries for diverse tasks such as batch processing, stream processing, machine learning, graph processing, and SQL queries, allowing users to build complex data pipelines within a single framework.
4. **Scalability:** Spark scales horizontally to handle large-scale data processing tasks across clusters of thousands of nodes, making it suitable for big data applications.
5. **Fault Tolerance:** Spark provides built-in fault tolerance through RDD lineage tracking, enabling resilient data processing even in the presence of node failures.

Disadvantages of Apache Spark:

1. **Steep Learning Curve:** Apache Spark has a relatively steep learning curve, especially for users who are new to distributed computing and parallel processing concepts. Understanding Spark's programming model, APIs, and cluster management can require significant time and effort.
2. **Resource Intensive:** Spark clusters require substantial computational and memory resources, particularly for large-scale data processing tasks. Deploying and managing Spark clusters may incur higher infrastructure costs compared to traditional data processing frameworks.
3. **Complexity of Cluster Management:** Setting up and managing Spark clusters can be complex, especially in large-scale production environments. Administrators need to configure and optimize cluster settings, monitor resource usage, and troubleshoot performance issues to ensure efficient and reliable operation.
4. **Data Shuffling Overhead:** Spark's distributed nature involves shuffling data across cluster nodes during various stages of data processing, such as data partitioning, shuffling, and aggregation. This data shuffling overhead can introduce latency and network bottlenecks, especially for operations involving large datasets or complex

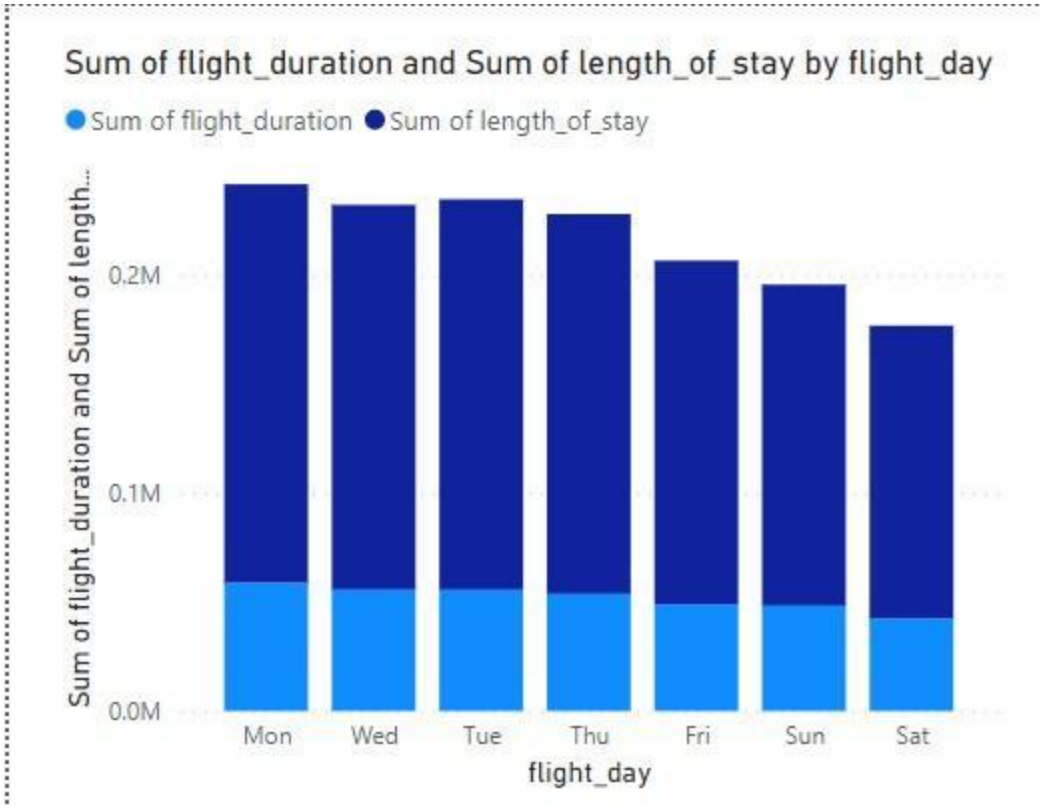
transformations. Efficient data partitioning and cluster tuning are essential to mitigate these performance issues.

Apache Spark at Netflix:

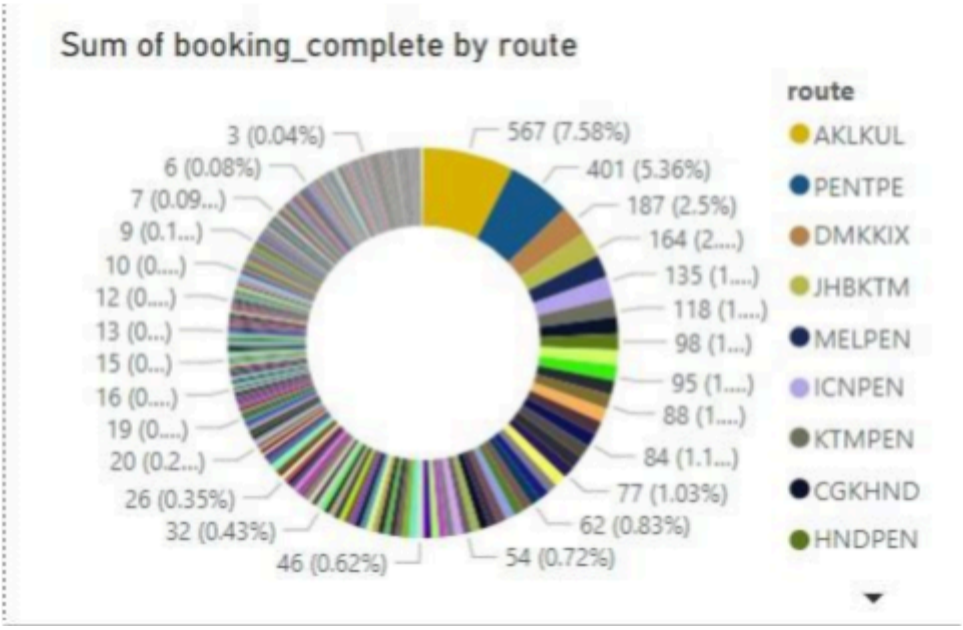
Netflix, a leading streaming service provider, relies on Apache Spark for various data processing and analytics tasks to enhance user experience and drive content recommendations. Here's how Netflix applies Apache Spark:

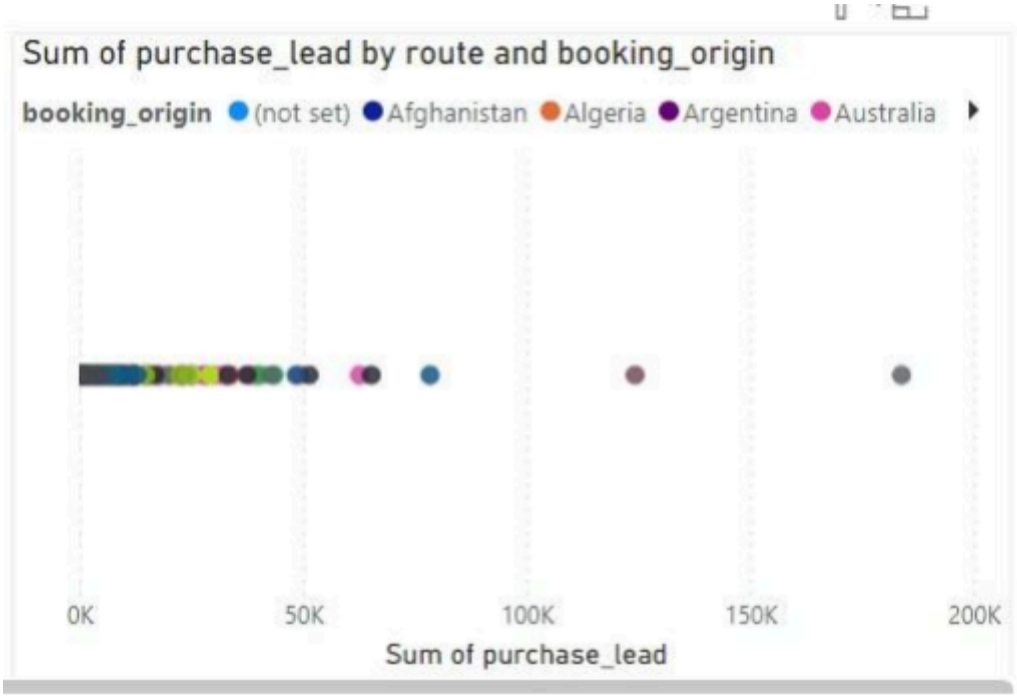
1. **Content Recommendation Engine:** Apache Spark enables Netflix to analyze vast amounts of user interaction data, including viewing history, ratings, search queries, and browsing behavior. By processing this data in real-time using Spark's streaming capabilities, Netflix can generate personalized recommendations for each user based on their preferences and viewing habits. Spark's machine learning libraries allow Netflix to build and deploy sophisticated recommendation models that continuously learn and adapt to user feedback.
2. **Content Optimization:** Netflix utilizes Spark to analyze content performance metrics such as viewer engagement, retention rates, and audience demographics. By correlating this data with user preferences and viewing patterns, Netflix can identify successful content attributes and optimize its content catalog accordingly. Spark's distributed computing capabilities enable Netflix to process and analyze large volumes of data quickly and efficiently, facilitating data-driven decision-making for content acquisition and production.
3. **Operational Analytics:** Netflix leverages Spark for operational analytics to monitor and optimize its streaming infrastructure, including server performance, network latency, and system reliability. Spark enables Netflix to aggregate and analyze log data from thousands of servers and devices in real-time, providing insights into system health, resource utilization, and user experience. This allows Netflix to identify and address performance bottlenecks, improve scalability, and ensure high availability of its streaming services.
4. **A/B Testing and Experimentation:** Netflix conducts A/B testing and experimentation using Spark to evaluate new features, algorithms, and user interface changes. Spark enables Netflix to design and execute large-scale experiments, analyze experimental results, and measure the impact on user engagement and retention. This iterative approach to product development allows Netflix to continuously improve its service and deliver a more personalized and satisfying user experience.

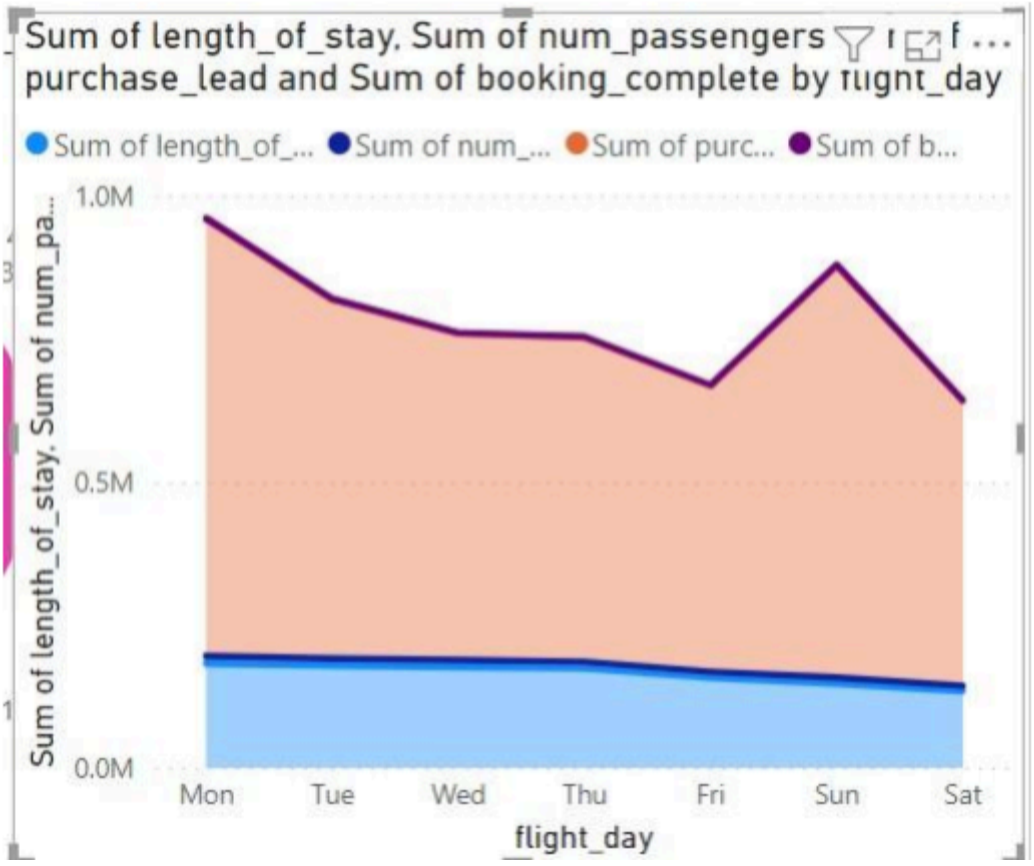
Conclusion: I have successfully Studied about POWER BI & Apache Spark along with its working, advantages and real-life applications of both the tools.


Sum of flight_duration and Sum of length_of_stay by flight_day
Sum of flight_duration
Sum of length_of_stay
Sum of flight_duration and Sum of length...
0.2M
0.1M
0.0M
Mon
Wed
Tue
Thu
Fri
Sun
Sat
flight_day


Sum of flight_duration by booking_origin
booking_origin
Australia
Malaysia
South Korea
Japan
China
Indonesia
Thailand
Taiwan
New Zealand
India
Singapore
United States
142.42K (39.14%)
49.39K (13.57%)
31.52K (8.66%)
28.63K (7.87%)
20.17K (5.5...)
16.11K (4.4...)
14... (...)
12.... (3....)
9.08K (2.5%)
2K (0.55%)


Sum of booking_complete by route
route
AKLKUL
PENTPE
DMKKIX
JHBKTM
MELPEN
ICNPEN
KTMPEN
CGKHND
HNDPEN
567 (7.58%)
401 (5.36%)
187 (2.5%)
164 (2....)
135 (1....)
118 (1....)
98 (1...)
95 (1....)
88 (1....)
84 (1.1...)
77 (1.03%)
62 (0.83%)
54 (0.72%)
46 (0.62%)
32 (0.43%)
26 (0.35%)
20 (0.2...)
19 (0....)
16 (0....)
15 (0...)
13 (0...)
12 (0....)
10 (0....)
9 (0.1...)
7 (0.09...)
6 (0.08%)
3 (0.04%)


Sum of purchase_lead by route and booking_origin
booking_origin (not set) Afghanistan Algeria Argentina Australia
0K
50K
100K
150K
200K
Sum of purchase_lead


Sum of length_of_stay, Sum of num_passengers
purchase_lead and Sum of booking_complete by flight_day
Sum of length_of_... Sum of num_... Sum of purc... Sum of b...
1.0M
0.5M
0.0M
Sum of length_of_stay, Sum of num_pa...
Mon
Tue
Wed
Thu
Fri
Sun
Sat
flight_day