Name - Gaurang A Raorane Div - D15A Roll no - 49 Batch - C

**Experiment - 9**

**Aim :-**  Case study on Power BI and Apache Spark.

**Theory:-**

**CASE STUDY ON POWER BI:**

## **Integrating OEE indicators with Power BI:**

Our client, a global company and a leader in the manufacturing of spreads and plant-based foods needed our help with their OEE analysis. Antdata’s task was to implement a comprehensive Business Intelligence tool that would allow them to calculate and analyze the OEE (Overall Equipment Effectiveness) indicator, which is a percentage snapshot of machinery usage efficiency. The challenge was (as it is in many cases) that the data came from 15 factories around the world whose reporting systems were inconsistent.

Our interactive report, thanks to the integration of different systems in Power Query, showed standardized indicators, which had a direct impact on the quality of the data and the repeatability of analysis. The implementation of the Power BI solution provided a clear, up-to-date, and easily accessible view of the effectiveness of individual lines in the factories, including downtime. This helped to focus on what needed to be improved to deliver value to the business. As a result, their performance got better from the bottom up – both globally and from a single production line perspective.

But it doesn’t stop there. Thanks to the use of Power BI reporting it is possible to monitor the trend of availability, productivity, and quality over time. This allows for faster and more accurate identification of areas and processes with high potential for improvement to plan optimization activities in the places which require it most.

Looking at the standardized OEE indicators, it is possible to perform comparative analysis and detect differences in efficiency between individual production lines or entire factories, as well as to detect anomalies that can determine the speed and quality of production.

## **Power BI implementation in a store chain**

Our client, a company owning a store chain and franchise stores, needed a comprehensive Business Intelligence tool to control and analyze sales performance.

The automated reporting system implemented by us on the Power BI platform for the sales department allowed for a quick analysis of sales by receipts. The receipts were supplemented from the SQL database. The data coming from over 70 stores from across Poland were integrated into the database. The resulting interactive sales and margin report showed data from both perspectives – company stores and franchise stores. The data in the report was scheduled to refresh automatically 8 times a day so that the report always showed the most up-to-date results.

The ability to view the sales results of the individual stores on an annual, quarterly, or monthly basis, allowed for more efficient management of the entire sales chain. By utilizing a collection of DAX functions it was possible to create a number of advanced KPIs: budget and margin execution, statistics for different types of sales documents, or the year-to-year comparisons. In addition, tracking trends over time and the ability to compare sales between the individual stores, categories, or products further expanded the field of business analysis.

The report not only showed sales data but also what is strictly connected to sales – marketing data. Measuring the effectiveness of promotional and marketing campaigns facilitates the planning of future actions increasing the company’s profits. Comparison of sales before, during, and after the promotion period allows for an easy evaluation of the effectiveness of marketing campaigns for each product separately.

The whole report provides a wide perspective on sales, which allows for accurate predictions and estimates of future sales. It is possible to predict the average percentage of budget execution and the expected sales value at the end of the current year. This allows taking appropriate actions which can improve sales in the nearest future.

This is a multi-level report which means that thanks to row-level security this one report can be used across all organizational levels. It is available on the Power BI service and can be viewed on all devices – computer, phone, or tablet. The report shows business results both from the company and a single store perspective. All of that hugely simplifies decision-making for the local, regional and central management.

The developed Business Intelligence system is under continuous development and subsequent stages of the project are being systematically implemented. Currently, we are working on further modules of the report within the area of inventory management, orders, and billings.

**CASE STUDY ON APACHE SPARK:**

**Optimizing Data Processing with Apache Spark**

**1. Background:**

* Company Abaca group is a large e-commerce platform handling massive volumes of data related to customer transactions, product catalog, user interactions, and more.
* The existing data processing pipeline faces challenges in scalability, speed, and efficiency due to the increasing volume of data and complex analytical requirements.
* To address these challenges, Company Abaca group decides to explore Apache Spark as a solution for optimizing data processing and analytics.

**2. Objectives:**

* Improve the scalability and performance of data processing tasks, including batch processing and real-time streaming analytics.
* Enable complex data transformations, aggregations, and machine learning algorithms on large datasets.
* Reduce processing time and latency for generating actionable insights and recommendations.
* Enhance the overall data pipeline architecture for better reliability, fault tolerance, and ease of maintenance.

**3. Challenges:**

* Large-scale data processing: Handling terabytes or petabytes of data efficiently and in a timely manner.
* Real-time analytics: Processing streaming data with low latency to enable real-time insights and decision-making.
* Complex transformations: Performing complex data transformations, joins, aggregations, and computations on distributed datasets.
* **Resource optimization:** Efficiently utilizing cluster resources (CPU, memory, storage) to minimize processing time and cost.
* **Scalability and reliability:** Ensuring scalability, fault tolerance, and reliability in the data processing pipeline.

**4. Solution Approach:**

* Apache Spark Adoption: Company Abaca group decides to adopt Apache Spark due to its in-memory processing capabilities, distributed computing model, and support for batch processing and streaming analytics.
* Cluster Setup: Set up a cluster environment with Spark standalone or on a cloud-based platform (e.g., AWS EMR, Azure Databricks) for distributed data processing.
* Data Ingestion: Ingest data from various sources such as databases, data lakes, streaming platforms (Kafka, Kinesis), and external APIs into Spark DataFrames or RDDs.
* Data Processing: Utilize Spark's rich API and libraries (Spark SQL, DataFrame API, MLlib, GraphX) to perform data transformations, aggregations, machine learning tasks, and graph processing.
* Streaming Analytics: Implement real-time streaming analytics using Spark Streaming or Structured Streaming for continuous processing of incoming data streams.
* Optimization Techniques: Apply optimization techniques such as partitioning, caching, and tuning cluster configurations (executor memory, cores, shuffle partitions) to improve performance and resource utilization.
* Monitoring and Management: Monitor job execution, resource usage, and performance metrics using Spark monitoring tools (Spark UI, metrics logging) for performance optimization and troubleshooting.
* Integration with Ecosystem: Integrate Spark with other components of the data ecosystem such as Hadoop, Hive, HBase, Kafka, and external data sources for seamless data integration and analytics workflows

**5. Outcomes:**

* Scalability: Apache Spark enables Company Abaca group to scale data processing tasks horizontally across a cluster of nodes, handling large volumes of data with ease.
* Performance: Spark's in-memory processing and optimization techniques result in significant improvements in processing speed and latency, leading to faster analytics and insights generation.
* Real-time Analytics: Spark Streaming and Structured Streaming empower real-time analytics, enabling Company Abaca group to react swiftly to changing data patterns and events.
* Complex Analytics: Spark's rich API and libraries facilitate complex data transformations, machine learning algorithms, and graph processing, unlocking deeper insights and advanced analytics capabilities.
* Resource Optimization: Through fine-tuning and optimization of Spark jobs, Company Abaca group achieves better resource utilization, cost efficiency, and improved cluster performance.
* Reliability and Fault Tolerance: Spark's fault-tolerant architecture ensures reliable data processing and job recovery in case of failures, enhancing the overall reliability of the data pipeline.

**6. Future Enhancements:**

* Explore advanced analytics and machine learning use cases with Spark MLlib and GraphX for predictive analytics, recommendation systems, and graph algorithms.
* Enhance data governance, security, and compliance measures within the Spark ecosystem.
* Evaluate integration with cloud-native services and serverless architectures for further scalability, flexibility, and cost optimization.
* Continuous optimization and fine-tuning of Spark jobs, cluster configurations, and data processing workflows for ongoing performance improvements.

**CONCLUSION:**

Through this experiment we have successfully studied the case study on PowerBI and Apache Shark.