**Restaurant Sales Analysis Using PySpark**

**🔹 1. What is PySpark?**

**PySpark** is the Python API for **Apache Spark**, an open-source distributed computing system used for big data processing. It allows you to write Spark applications using Python and can handle **large-scale datasets** efficiently through parallel processing.

PySpark is a powerful alternative to Pandas, especially when dealing with large datasets that don’t fit in memory.

**🔹 2. How is Data Analysis Done in PySpark?**

Data analysis in PySpark follows these major steps:

* **Reading data** using .read() method with a defined schema
* **Transforming data** using functions like withColumn(), groupBy(), agg()
* **Joining datasets** with .join()
* **Aggregating** using sum, count, countDistinct, etc.
* **Sorting** and **filtering** using .orderBy() and .filter()
* **Displaying results** using display() in Databricks

In this analysis, two datasets were used:

* sales.csv – Customer order records
* menu.csv – Product details with pricing

**🔹 3. Concepts and Operations Used in This Analysis**

**Data Ingestion**

* Used StructType, StructField to define schemas
* Loaded CSVs into DataFrames using spark.read.format("csv")

**Data Transformation**

* Added new time-based columns: order\_month, order\_year, order\_quarter using withColumn() and date functions (month, year, quarter)

**Data Joins**

* Merged sales\_df and menu\_df on product\_id to combine order info with prices

**Aggregations and Grouping**

* groupBy() to group data by customer, product, date, location
* Aggregations:
  + sum('price') – for total sales
  + count() – to count orders
  + countDistinct() – to find unique customer visits

**Filtering**

* Used .filter() to isolate only restaurant orders (source\_order == 'Restaurant')

**Sorting & Limiting**

* .orderBy() for sorting
* .limit(3) for top-3 products

**🔹 4. KPIs Computed in the Analysis**

1. Total amount spent by each customer
2. Total amount spent by each food category
3. Total amount of sales in each month
4. Yearly sales
5. Quarterly sales
6. Total number of orders by each category
7. Top 3 ordered items
8. Frequency of customer visits in Restaurants
9. Total sales by each country
10. Total sales by order source

**🔹 5. What is Databricks and Why Did You Use It?**

**Databricks** is a cloud-based data platform built on Apache Spark. The **Databricks Community Edition** offers a free platform to write, run, and share big data analytics using notebooks.

You used Databricks for:

* 🚀 Easy Spark cluster setup
* 📝 Notebook-based code execution
* 📊 Built-in support for display() and data visualizations
* ☁️ Cloud storage and UI for managing CSVs

**🔹 6. Real-World Applications / Use Cases of PySpark**

* **Retail & Restaurants**: Customer segmentation, sales forecasting, order trend analysis
* **Finance**: Fraud detection, credit risk modeling
* **Healthcare**: Patient record processing, treatment prediction
* **E-commerce**: Product recommendation systems, inventory management
* **IoT & Logs**: Processing large-scale sensor and log data in real time

**🔹 7. Bonus: Why PySpark over Pandas?**

| **Feature** | **Pandas** | **PySpark** |
| --- | --- | --- |
| Data Volume | Up to GBs | TBs and beyond |
| Processing Type | Single machine | Distributed |
| Speed | Slower on large data | Fast due to parallelism |
| Cloud Compatibility | Limited | Built for the cloud |

**Summary :**

This project not only showcases your technical skills but also highlights your ability to leverage big data tools for real-world business analysis.