

Sales Data Analysis using Apache Spark on Azure

Project Overview

This project demonstrates large-scale sales data analysis using Apache Spark on Azure Databricks.

Sales data stored in Azure Blob Storage is processed to generate business insights.

Technologies Used

- Azure Databricks
- Apache Spark (PySpark)
- Azure Blob Storage
- Python

Project Workflow

1. Created Azure Blob Storage and uploaded sales data
2. Connected Azure Blob Storage to Azure Databricks
3. Loaded CSV data using Spark DataFrame API
4. Cleaned and transformed sales data
5. Generated insights:
 - Total sales
 - Region-wise sales
 - Top-selling products
 - Monthly sales trends
6. Visualized results using Databricks built-in tools

Sample Analysis Code

```
```python
df.groupBy("region").sum("price").show()
```