

Slide 1: Title

■ SUPERSTORE SALES ANALYSIS USING PYSPARK

Prepared by:

Anwar Basha Shaik

Malla Reddy University

A data analytics project on sales performance using big data tools.

Slide 2: Introduction

Superstore Sales Analysis is a PySpark-based project that processes and visualizes large-scale retail data. It identifies sales patterns, profit trends, and product performance across various regions and categories.

Slide 3: Objectives

- Analyze sales and profit across categories and regions.
- Identify top-performing products.
- Understand business performance using PySpark.
- Create visual reports for insights.

Slide 4: Importance of Analysis

- Enables data-driven decision-making.
- Identifies profitable products and regions.
- Helps in optimizing pricing and inventory.
- Supports accurate sales forecasting.

Slide 5: Tools & Technologies Used

- Apache PySpark – for big data processing.
- Pandas – for data manipulation.
- Matplotlib – for visualization.
- Python – core programming language.
- Jupyter / VS Code – for execution.

Slide 6: Dataset Details

- File: Superstore_utf8.csv
- Rows: ~10,000 records
- Columns: Category, Product Name, Sales, Profit, Region, etc.
- Encoding: Converted from cp1252 to UTF-8 for Spark compatibility.

Slide 7: Data Cleaning

- Removed null and duplicate records.
- Trimmed whitespaces in text columns.
- Converted Sales & Profit to DoubleType.
- Verified schema using Spark printSchema().

Slide 8: Analysis Performed

- Total Sales by Category – `groupBy + sum`.
- Top 10 Products by Sales – `orderBy(desc())`.
- Average Profit by Region – `avg()`.
- Visualized all metrics using Matplotlib charts.

Slide 9: Challenges Faced

- File encoding issues during data import.
- Handling numeric type conversions.
- Processing large CSV files efficiently.
- Memory and performance tuning in PySpark.

Slide 10: Conclusion

- PySpark processed and analyzed the dataset efficiently.
- Gained insights into sales, profit, and trends.
- Exported cleaned data for BI tools.
- Output File: Processed_Superstore.csv

Big Data Analytics helps businesses make smarter decisions!