BIG DATA ANALYSIS

BY ANKITA PATIL

OBJECTIVES -

- Analyze a large dataset using pyspark
- Demonstrate scalable big data processing
- Extact meaningful business insisghts
- Improve decision – making using analytics

DATASET OVERVIEW -

- Fields includes -
- Order ID
- Product category
- Region
- Sales Amount
- Quantity
- Order Date

TOOLS & TECHNOLOGIES -

- PySpark
- Python
- Apache Spark
- Jupyter Notebook

DATA CLEANING -

- Removed missing values
- Removed duplicate records
- Standardized data
- Verified consistency

ANALYSIS PERFORMED

- Total sales by region
- Average sales by category
- Order volume analysis
- Category performance comparison

BUSINESS IMPACT -

- Helps improve sales strategy
- Supports inventory planning
- Identifies growth opportunity
- Enhances customer targeting
- Enables data driven decisions

CONCLUSION -

- PySpark is powerful for big data analysis
- Distributed computing increases efficiency
- Insights support smarter business planning
- Scalable analytics is essential in modern industries