NAME: SHEJAL TIWARI DIV/ROLL NO: D15C/57

Experiment No: 9

Aim: To perform Exploratory data analysis using Apache Spark And Pandas.

Theory -

1. What is Apache Spark and How It Works?

Apache Spark is an open-source data processing engine designed for big data workloads. It supports fast and general-purpose cluster computing.
It uses in-memory computation, meaning data is processed in RAM instead of slow disk drives, which greatly speeds up performance.
Spark supports multiple languages like Python (PySpark), Scala, Java, and R, making it flexible for developers and data scientists.
It follows a DAG (Directed Acyclic Graph) execution model, where transformations on data are organized as a sequence of stages.
Spark can run on various cluster managers like YARN, Mesos, or standalone, and supports tools like Hadoop HDFS and Apache Hive.

2. How is Data Exploration Done in Apache Spark? Explain Steps.

Load the data: Use Spark's read API to load datasets from sources like CSV, JSON, or databases into DataFrames.
Inspect schema: Use .printSchema() and .dtypes to understand the structure and data types of each column.
View sample data: Use .show() or .head() to preview the top rows and check for obvious issues or patterns.
Summary statistics: Use .describe() and .summary() to get basic statistics like mean, count, min, and max for numeric columns.
Check for nulls and duplicates: Use .filter(), .dropna(), or .dropDuplicates() to handle missing values or redundant records.

Conclusion:

Apache Spark is a robust, open-source distributed computing system designed for fast processing of large-scale data. Its in-memory computation model, support for multiple languages, and compatibility with various data sources make it an efficient choice for big data analytics. Data exploration in Apache Spark is performed through structured steps such as loading the dataset, inspecting schema, analyzing statistics, and handling missing or duplicate values. These steps help in gaining initial insights and preparing the data for further processing. Overall, Apache Spark simplifies complex data operations and is a valuable tool in the big data ecosystem.