|  |  |
| --- | --- |
| Name: Surname, First Name MI. | Lab No. 1 |
| Git Repo/ Colab Link: | Date: |

**Objective**

Clearly state the objective of the lab. What is the purpose of this activity? What are you trying to achieve or learn?

Example:

To understand and implement data preprocessing techniques using Apache Spark, including filtering, aggregation, and handling missing data.

1. Be able to create and understand master-worker works.
2. To implement different kinds of partitions such as hash partitioning,

**Introduction**

Provide background information relevant to the lab. Briefly explain the concepts, tools, or technologies involved. Include any theoretical concepts or algorithms being used.

Example:

This lab explores Apache Spark, a distributed computing framework used for big data processing. The focus is on data transformation techniques and their role in preparing datasets for analysis.

**Methodology**

Describe the steps you followed in the lab. Include algorithms, data flow, or specific Spark transformations/actions if applicable. Use numbered or bullet points for clarity.

Example:

1. Load the dataset into a Spark DataFrame.
2. Perform exploratory data analysis (EDA) to identify missing values and outliers.
3. Apply filtering to remove invalid records.
4. Aggregate data based on specific attributes.

**Results and Analysis**

Present the outcomes of your experiment. Use tables, graphs, or screenshots of results as needed. Provide a detailed analysis of the results, highlighting key findings.

**Challenges and Solutions**

Document any issues encountered during the lab and how you resolved them.

Example:

Challenge: Memory error while loading a large dataset.

Solution: Used Spark’s partitioning feature to load data in chunks.

**Conclusion**

Summarize the key takeaways from the lab. Did you meet the objectives? What did you learn about the tools, techniques, or concepts?

Example:

The lab demonstrated the importance of preprocessing in big data workflows. Using Apache Spark, we efficiently handled large datasets and improved data quality.