|  |  |
| --- | --- |
| **AIDS** | **Lab: BDH Lab** |

|  |  |
| --- | --- |
| **Name of Students: Himanshu Dhomane** | **Semester/ Section: 7/A** |
| **Roll No: 43** | **Enroll No: 21070297** |

# Practical No. 8

|  |
| --- |
| **Aim:** Pig Operations: Load &amp; Store Data, Aggregation Operations,Filtering Data and Joining Datasets. |
| **Introduction:**  **Introduction to Apache Pig:**  Apache Pig is a high-level platform for processing large datasets in Hadoop. It provides a scripting language, Pig Latin, which simplifies data manipulation and analysis. Pig abstracts the complexity of writing MapReduce programs by allowing users to perform data transformations and analysis using simple and expressive commands.  **Core Pig Operations:**   1. **Loading and Storing Data**   Loading and storing data in Pig allows users to import data from external sources (e.g., HDFS, local file systems) and export the processed data after applying transformations.   * + **LOAD**: The LOAD statement reads data from the file system into Pig for processing. Data can be loaded from various formats like plain text, CSV, JSON, etc.     - **Syntax**:   pig  Copy code  relation = LOAD 'file\_path' USING PigStorage('delimiter') AS (field1:datatype, field2:datatype, ...);   * + **STORE**: The STORE statement writes the transformed data back to a storage system (e.g., HDFS or the local file system). It supports multiple output formats such as plain text or specific delimited formats.     - **Syntax**:   pig  Copy code  STORE relation INTO 'output\_path' USING PigStorage('delimiter'); |

## Aggregation Operations

Aggregation operations in Pig allow users to perform data summarization and analytics over grouped data. Common aggregation functions include SUM, COUNT, MAX, MIN, and AVG.

* + **GROUP**: The GROUP operation is used to group records based on a specific field. Grouping is necessary for performing aggregations.

## Syntax:

pig

Copy code

grouped\_relation = GROUP relation BY field; Other Operations included are Count , SUM.

## Filtering Data

Filtering operations in Pig allow users to selectively remove or retain records based on specific conditions.

* + **FILTER**: The FILTER operation is used to select rows from a dataset that meet a particular condition.

## Syntax:

pig

Copy code

filtered\_relation = FILTER relation BY condition;

## Joining Datasets

Joining operations in Pig enable users to combine data from multiple datasets based on common fields. This is useful for combining related information stored across different sources.

* + **JOIN**: The JOIN operation is used to combine two or more datasets based on a common field, such as a key.

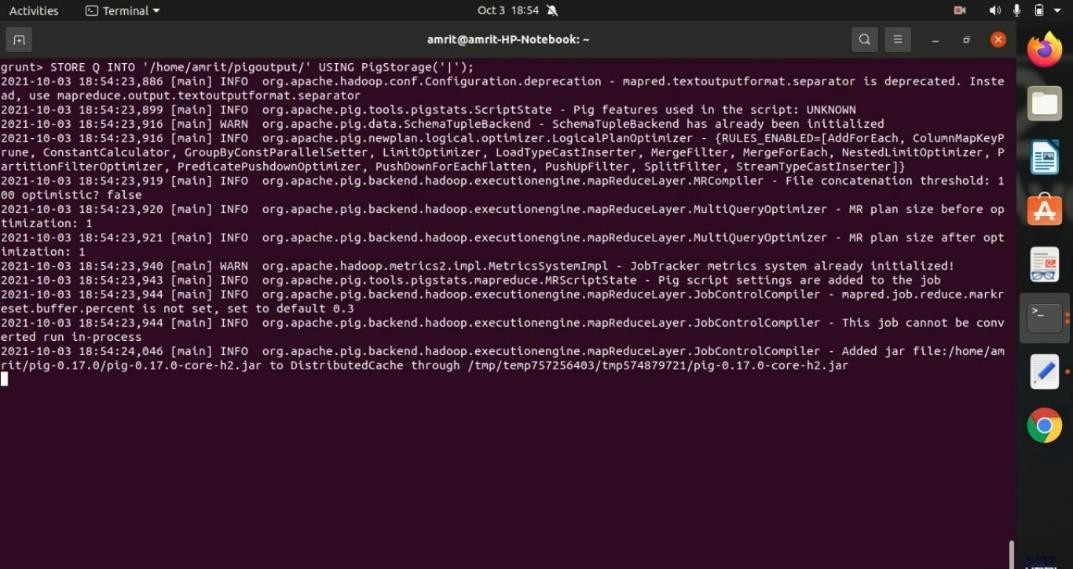
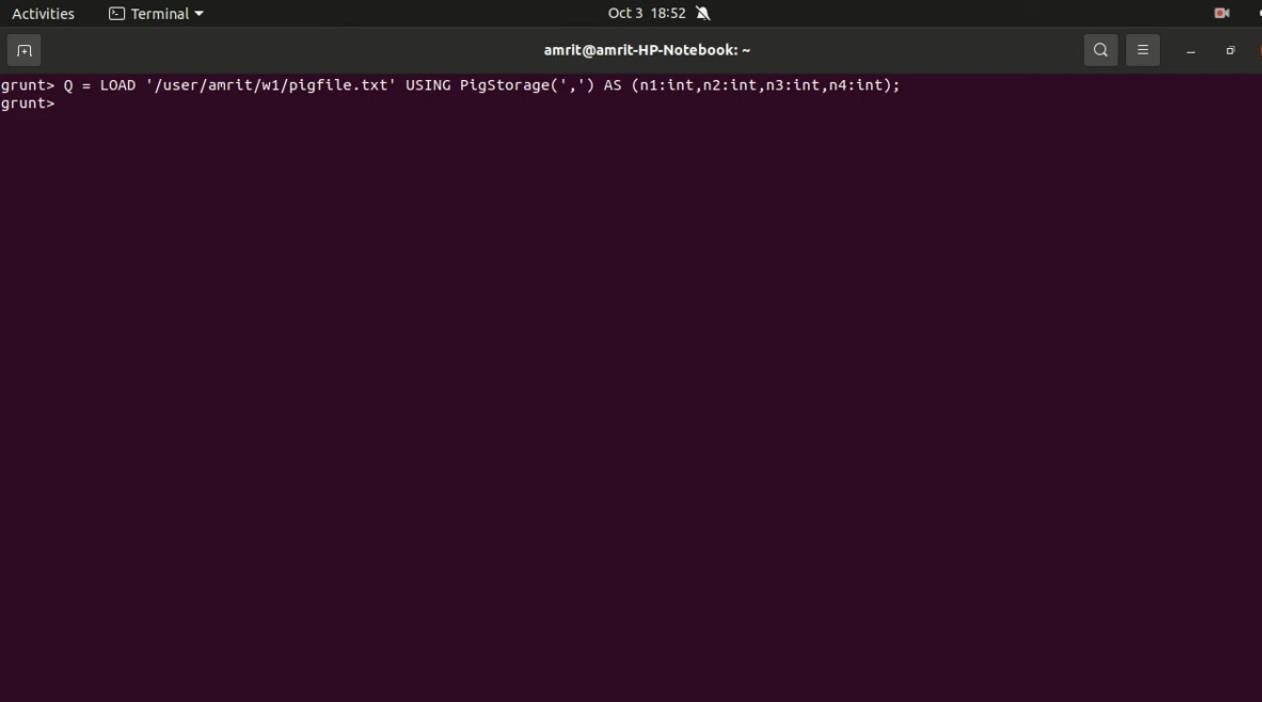
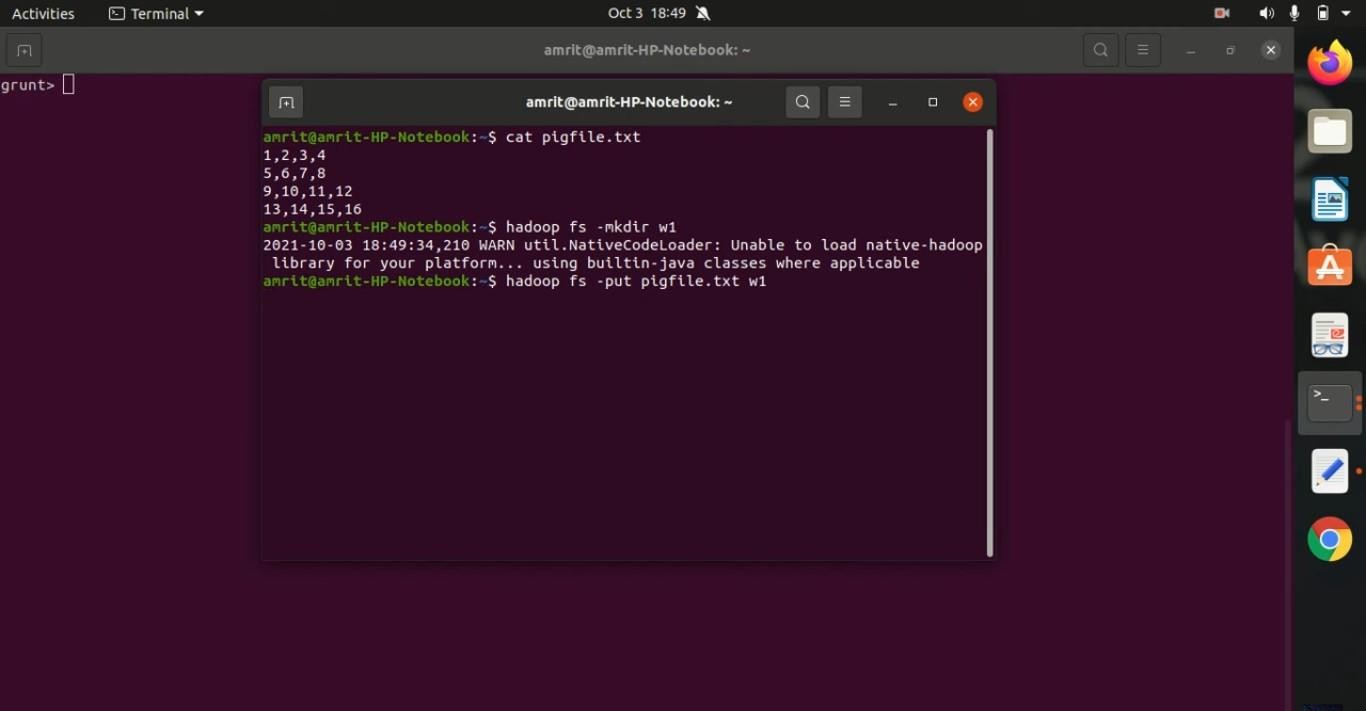
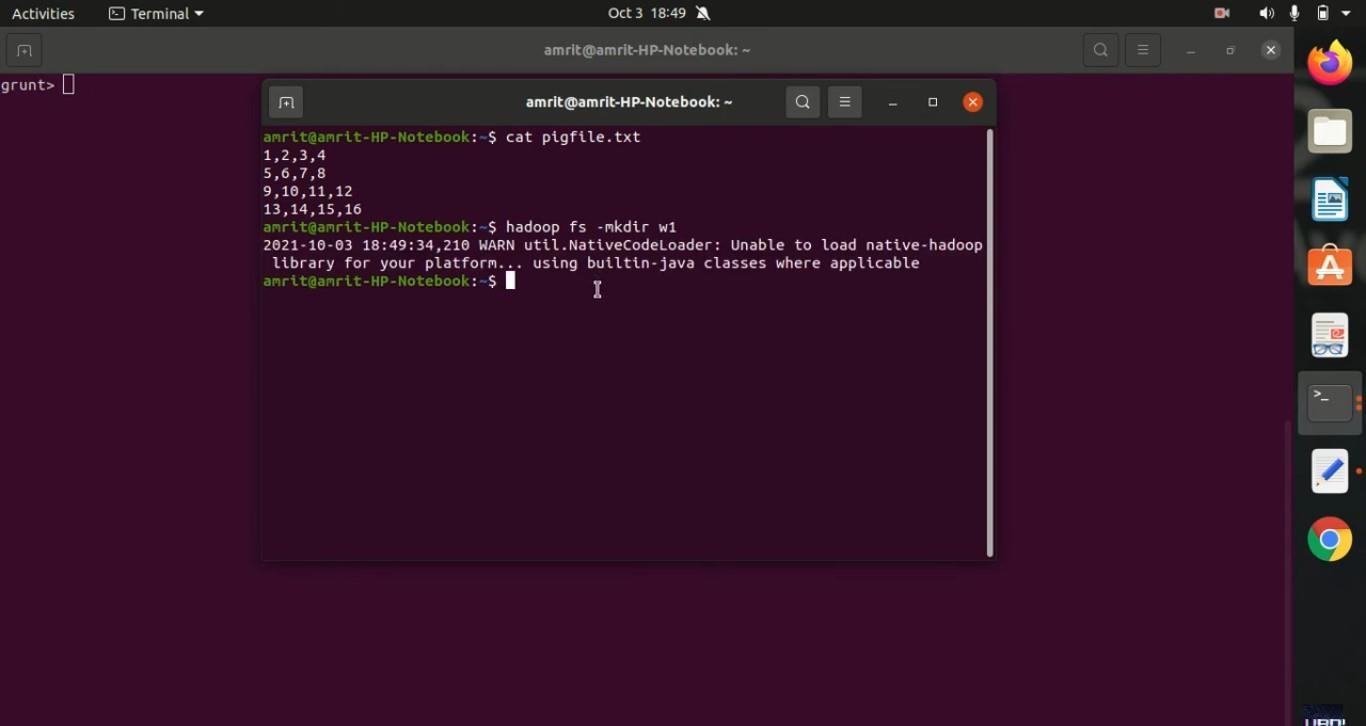
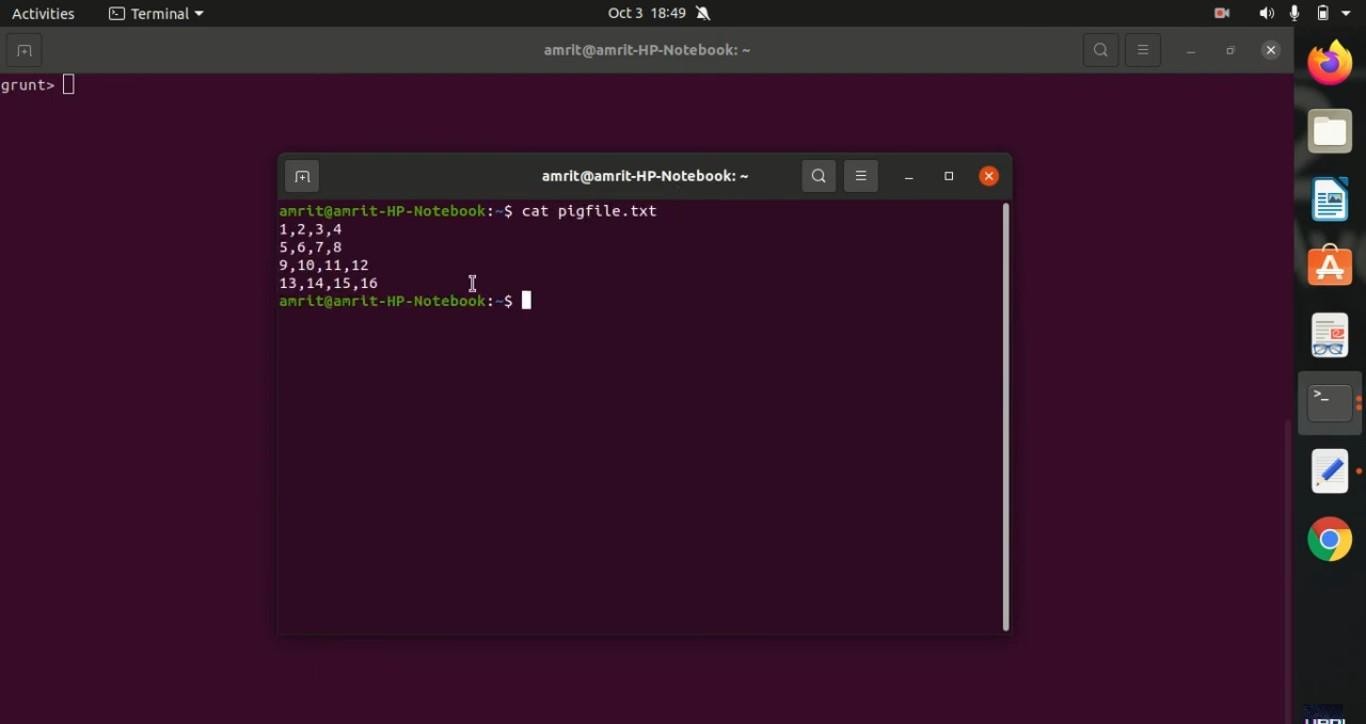
## Syntax:

pig

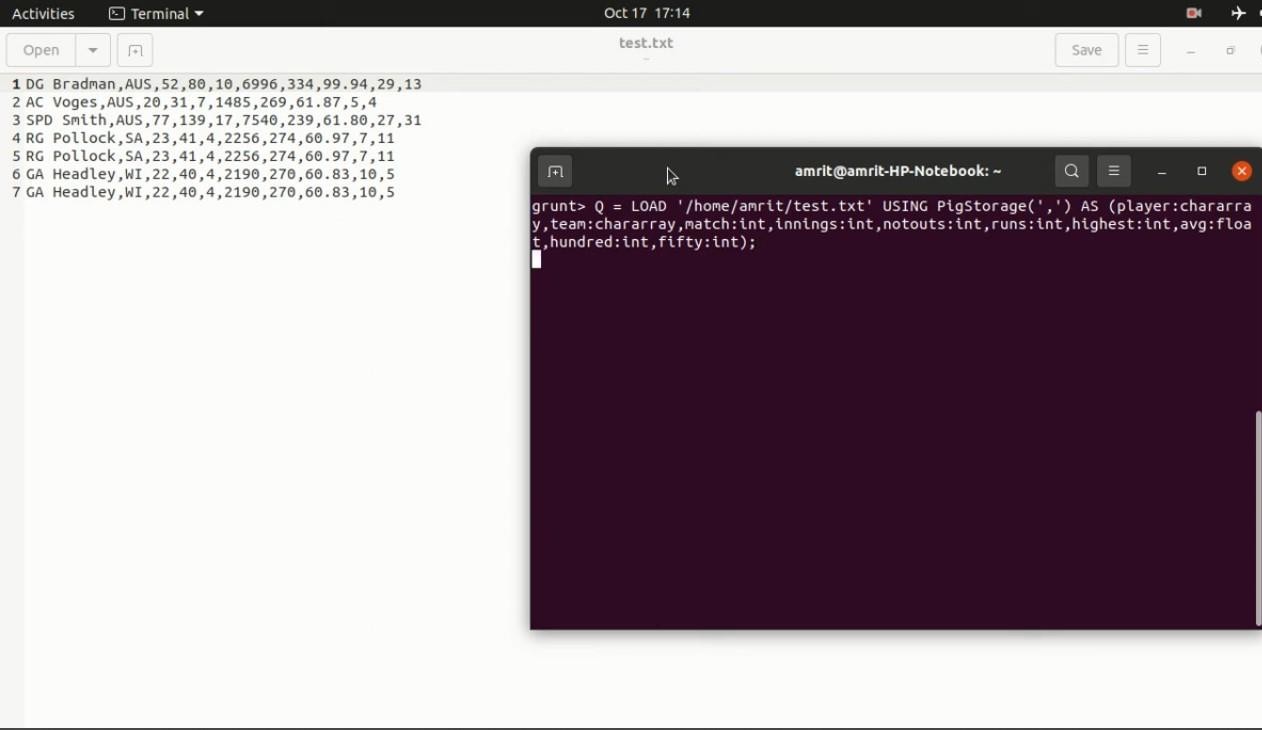
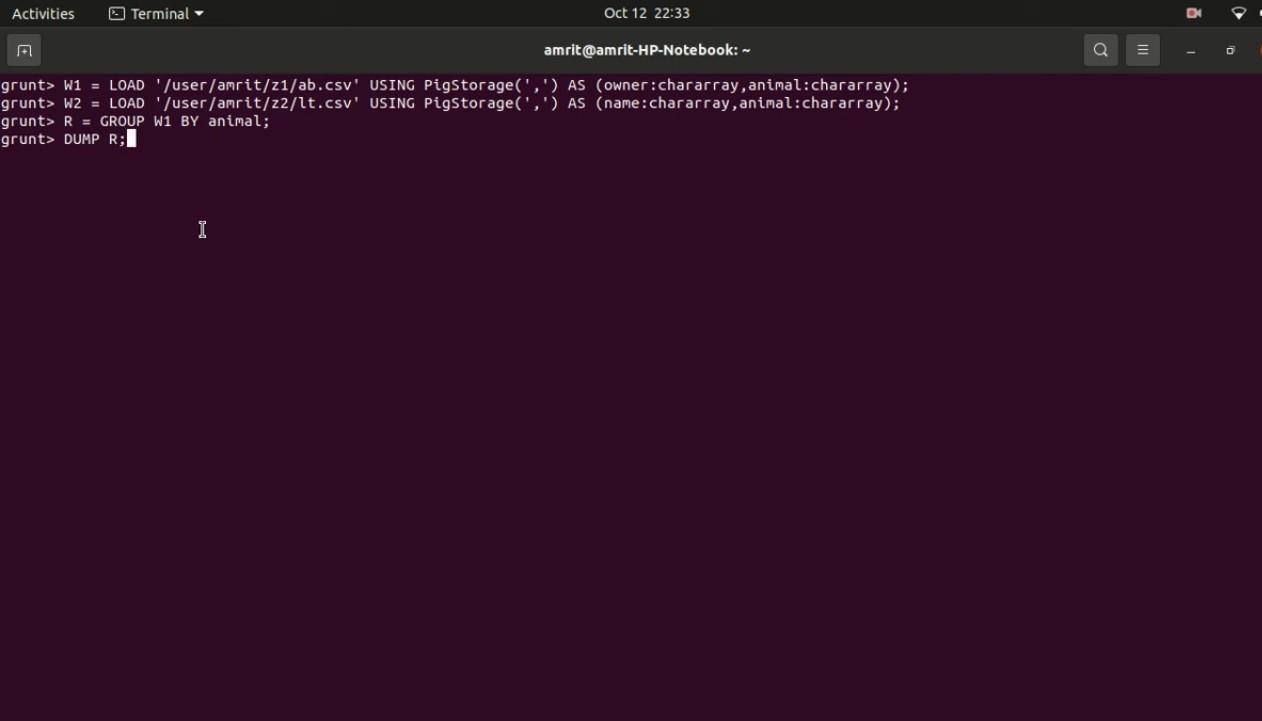
Copy code

joined\_relation = JOIN relation1 BY key1, relation2 BY key2;

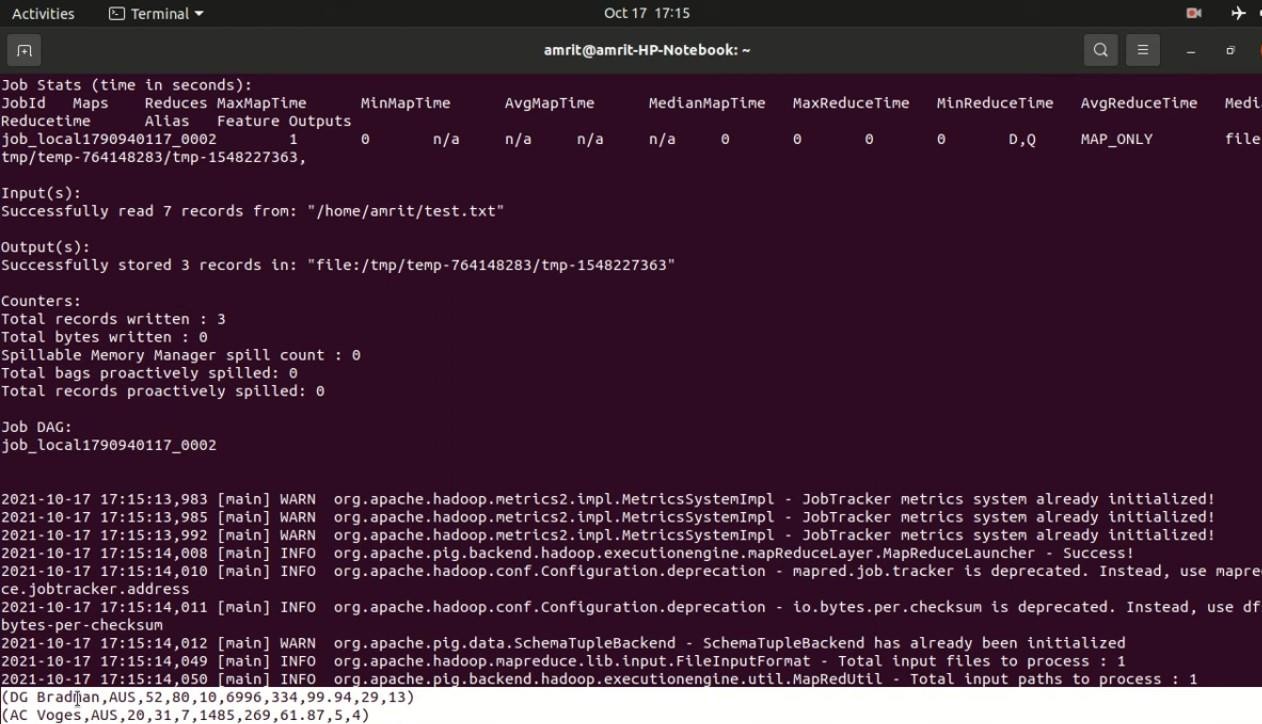
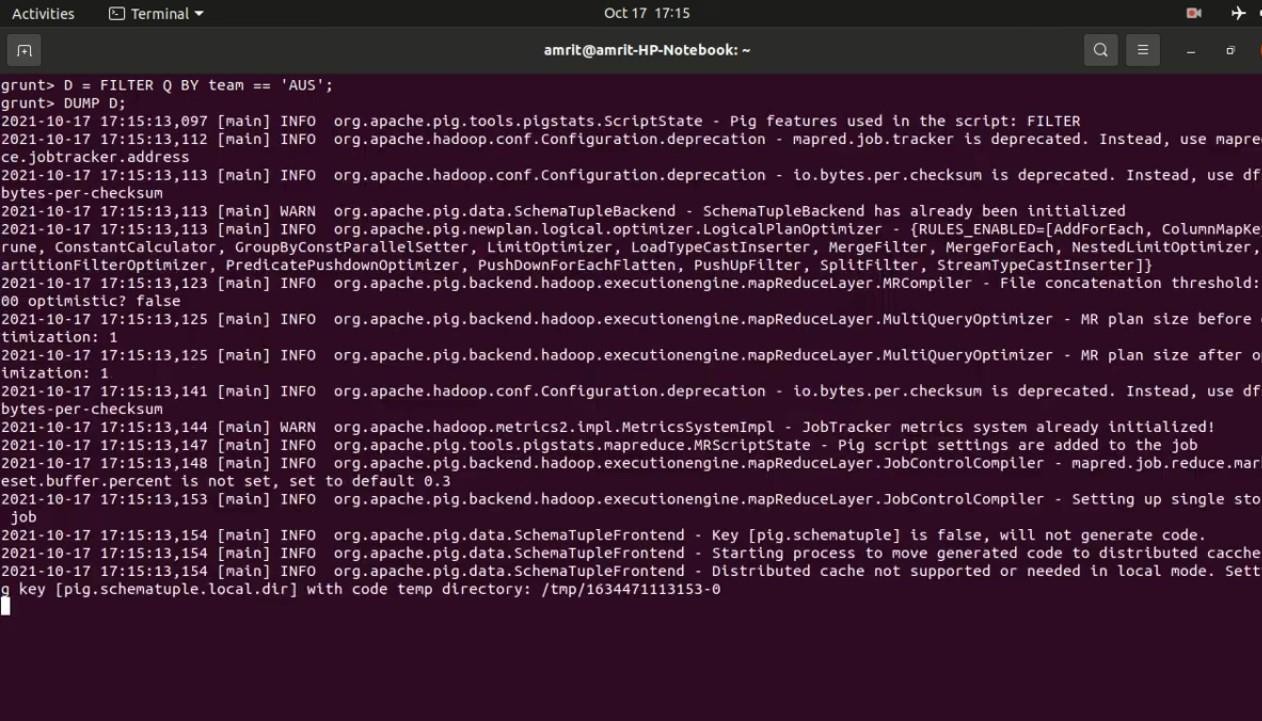
# Output Screenshots:



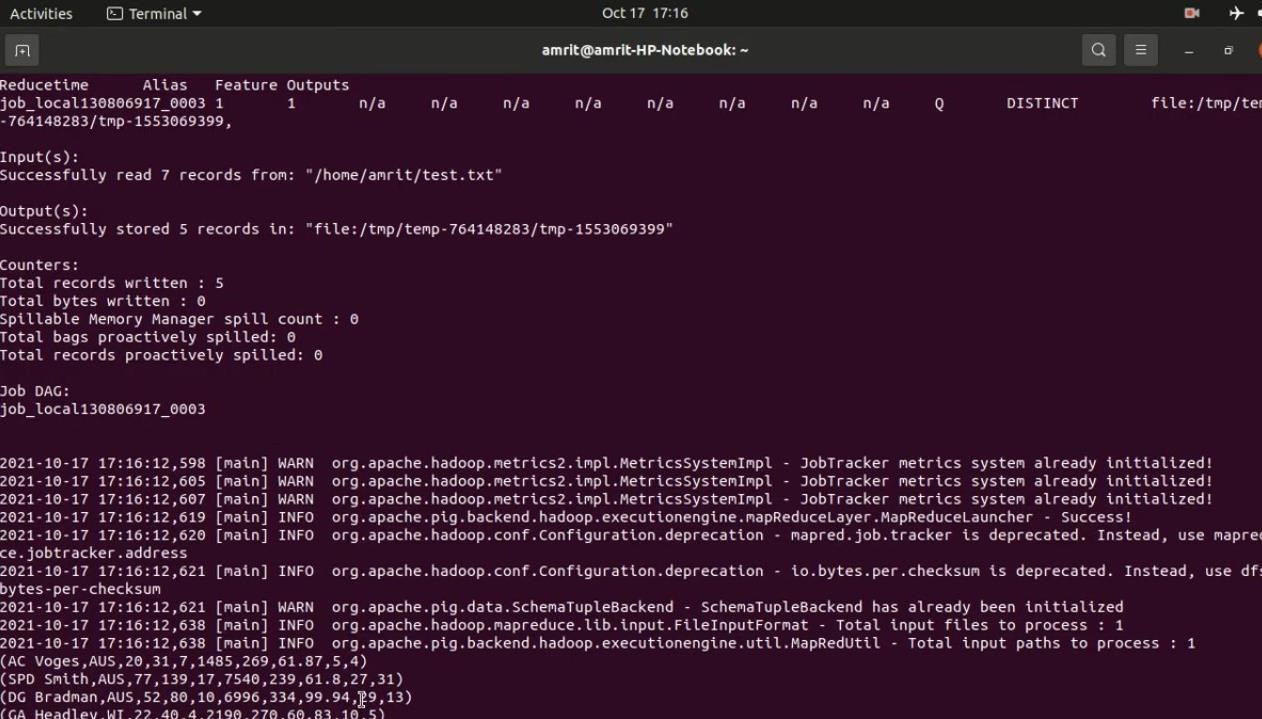
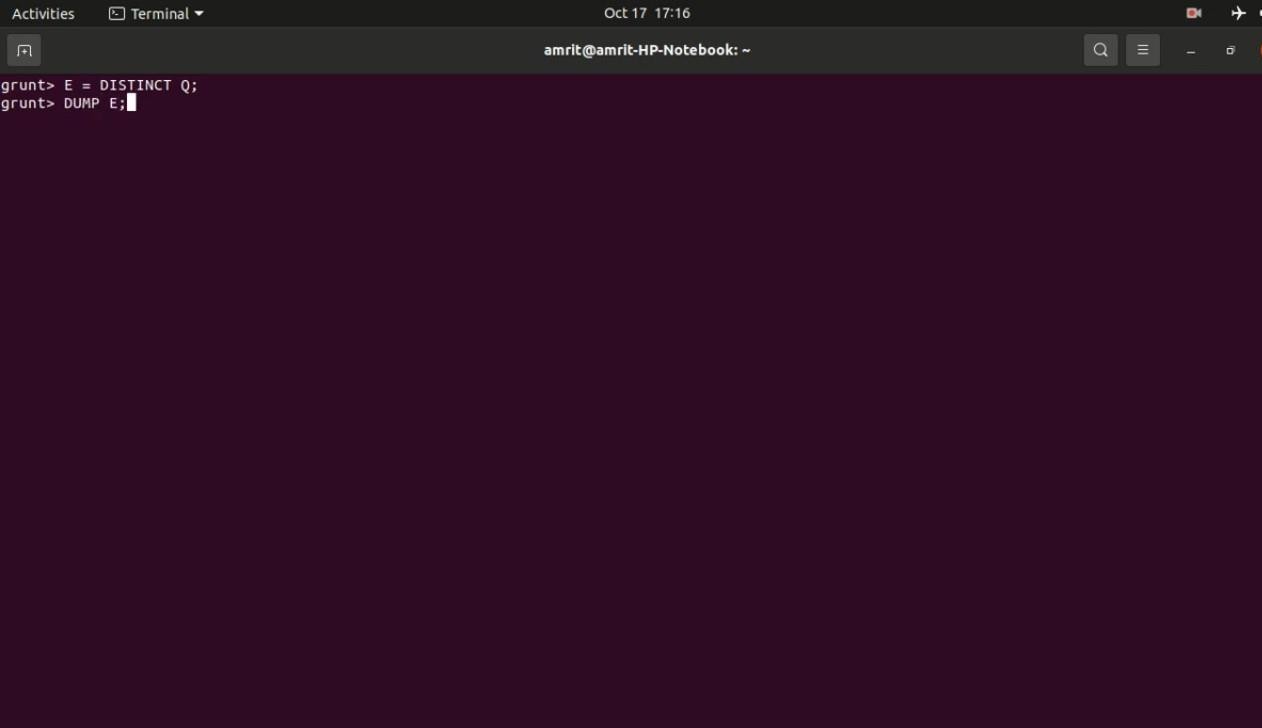
1. LOAD:
2. STORE:
3. GROUP:



1. FILTER:



1. DISTINCT:



**Conclusion:**

By using Apache Pig's intuitive commands like LOAD, STORE, GROUP, FILTER, and JOIN, data processing becomes more accessible and efficient, especially for large-scale datasets. These operations allow users to perform ETL tasks, aggregation, filtering, and joining datasets without needing to write complex MapReduce code, making it an essential tool in the Hadoop ecosystem.