**Big Data Analytics**

**BAD601**

**Experiment 7:**

Develop Pig Latin scripts to sort, group, join, project, and filter the data.

**Start Hadoop Services**

start-dfs.sh

start-yarn.sh

**Start Pig in Local or MapReduce Mode**

* **Local Mode (for testing):**

pig -x local

* **MapReduce Mode:**

pig

**Assume we have a CSV file: students.csv**

101,John,CS,80

102,Alice,EC,90

103,Bob,CS,70

104,David,ME,85

**Suppose we have another file: dept\_info.csv**

CS,Computer Science

EC,Electronics

ME,Mechanical

**Open Notepad.**

**Save the following script as** student\_analysis.pig

**-- Load student data**

students = LOAD 'students.csv' USING PigStorage(',')

AS (id:int, name:chararray, dept:chararray, marks:int);

**-- Load department data**

departments = LOAD 'departments.csv' USING PigStorage(',')

AS (code:chararray, dept\_name:chararray);

**-- 1. Projection: Select name and marks**

projected = FOREACH students GENERATE name, marks;

**-- 2. Filtering: Students with marks > 80**

filtered = FILTER students BY marks > 80;

**-- 3. Sorting: Students sorted by marks descending**

sorted = ORDER students BY marks DESC;

**-- 4. Grouping: Group by department**

grouped = GROUP students BY dept;

**-- Calculate average marks per department**

avg\_marks = FOREACH grouped GENERATE group AS dept, AVG(students.marks) AS avg\_marks;

**-- 5. Joining: Join student with department info**

joined = JOIN students BY dept, departments BY code;

**-- Show outputs**

DUMP projected;

DUMP filtered;

DUMP sorted;

DUMP avg\_marks;

DUMP joined;

**How to Execute the Script**

From the terminal, navigate to the directory where the script is saved and run:

**In local mode (no need for Hadoop):**

pig -x local student\_analysis.pig

**In MapReduce mode (uses Hadoop cluster):**

pig student\_analysis.pig