

Experiment 8

employee_analysis.pig file

-- Load the employee data

```
employees = LOAD 'employee.csv' USING PigStorage(',')
  AS (emp_id:int, name:chararray, dept_id:int, salary:int);
```

-- Display original data

```
DUMP employees;
```

-- 1. FILTERING: Filter employees with salary greater than 55000

```
high_salary_employees = FILTER employees BY salary > 55000;
DUMP high_salary_employees;
```

-- 2. PROJECTION: Select only name and salary columns

```
employee_salaries = FOREACH employees GENERATE name, salary;
DUMP employee_salaries;
```

-- 3. SORTING: Sort employees by salary in descending order

```
sorted_employees = ORDER employees BY salary DESC;
DUMP sorted_employees;
```

-- 4. GROUPING: Group employees by department and calculate statistics

```
dept_groups = GROUP employees BY dept_id;
dept_stats = FOREACH dept_groups GENERATE
  group AS dept_id,
  COUNT(employees) AS employee_count,
  MIN(employees.salary) AS min_salary,
  MAX(employees.salary) AS max_salary,
  AVG(employees.salary) AS avg_salary;
DUMP dept_stats;
```

-- 5. FILTERING + PROJECTION: Employees in department 1 with specific fields

```
dept1_employees = FILTER employees BY dept_id == 1;
dept1_details = FOREACH dept1_employees GENERATE name, salary;
DUMP dept1_details;
```

-- 6. SORTING within GROUP: Get highest paid employee in each department

-- First group by department

```
dept_employee_groups = GROUP employees BY dept_id;
```

-- Then for each department, order employees by salary and take the first one

```
dept_top_earners = FOREACH dept_employee_groups {
  sorted = ORDER employees BY salary DESC;
  top_earner = LIMIT sorted 1;
  GENERATE group AS dept_id, FLATTEN(top_earner);
```

```

}
DUMP dept_top_earners;

-- 7. PROJECTION with calculations: Add bonus calculation
employees_with_bonus = FOREACH employees GENERATE
    emp_id,
    name,
    dept_id,
    salary,
    (salary * 0.10) AS bonus,
    (salary + (salary * 0.10)) AS total_compensation;
DUMP employees_with_bonus;

-- 8. FILTERING with multiple conditions: Specific salary range and department
mid_range_employees = FILTER employees BY
    salary >= 52000 AND salary <= 60000 AND dept_id != 3;
DUMP mid_range_employees;

-- 9. GROUPING with FILTER: Department statistics only for departments with more than 1 employee
dept_groups_filtered = GROUP employees BY dept_id;
large_depts = FILTER dept_groups_filtered BY COUNT(employees) > 1;
large_dept_stats = FOREACH large_depts GENERATE
    group AS dept_id,
    COUNT(employees) AS employee_count,
    AVG(employees.salary) AS avg_salary;
DUMP large_dept_stats;

-- 10. SORTING by multiple fields: Sort by department then by salary
sorted_by_dept_salary = ORDER employees BY dept_id ASC, salary DESC;
DUMP sorted_by_dept_salary;

-- Store the results
STORE sorted_employees INTO 'sorted_employees';
STORE dept_stats INTO 'department_statistics';
STORE employees_with_bonus INTO 'employees_with_bonus';

employees.csv
101,John Doe,1,50000
102,Jane Smith,2,60000
103,Mike Johnson,1,55000
104,Sarah Brown,3,65000
105,David Lee,2,52000

```

Command to run the program :

```
pig -x local employee_analysis.pig
```