

Experiment 8

employee_analysis.pig file

-- Load the employee data

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

-- Display original data

```
DUMP employees;
```

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

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

-- 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

```
deptgroups = GROUP employees BY deptid;  
deptstats = FOREACH deptgroups 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

```
dept1employees = FILTER employees BY deptid == 1;  
dept1details = FOREACH dept1employees GENERATE name, salary;  
DUMP dept1_details;
```

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

-- First group by department

```
deptemployeegroups = GROUP employees BY dept_id;
```

```
-- Then for each department, order employees by salary and take the first one
depttopearners = FOREACH deptemployeegroups {
  sorted = ORDER employees BY salary DESC;
  top_earner = LIMIT sorted 1;
  GENERATE group AS deptid, FLATTEN(top_earner);
}
DUMP depttopearners;
```

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

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

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

```
-- 10. SORTING by multiple fields: Sort by department then by salary
sortedbydeptsalary = ORDER employees BY deptid ASC, salary DESC;
DUMP sortedbydept_salary;
```

```
-- Store the results
STORE sortedemployees INTO 'sortedemployees';
STORE deptstats INTO 'departmentstatistics';
STORE employeeswithbonus INTO 'employeeswithbonus';
```

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
```