Session 8

Assignment 1

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 8 – Assignment 1 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 8 – Assignment 1** |
|  |  |
| **Start Date:** | 23/09/2017 |
|  |  |
|  |  |

# 

# Contents

[Contents 2](#_Toc493918936)

[Change History 3](#_Toc493918937)

[1. Problem Statement 4](#_Toc493918938)

[2. Solutions 5](#_Toc493918939)

[2.1. Get a list of employees who receive a salary less than 100, compared to their immediate employee with higher salary in the same unit 5](#_Toc493918940)

[2.2. List of all employees who draw higher salary than the average salary of that department 6](#_Toc493918941)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 23/09/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

* Get a list of employees who receive a salary less than 100, compared to their immediate employee with higher salary in the same unit
* List of all employees who draw higher salary than the average salary of that department

# Solutions

## Get a list of employees who receive a salary less than 100, compared to their immediate employee with higher salary in the same unit

**Data set**

**Users.txt**

1 Amit 600 DNA

2 Sumit 200 DNA

3 Yadav 229 DNA

4 Sunil 500 FCS

5 Kranti 100 FCS

6 Bob 201 FCS

7 Duncan 150 DNA

8 Sue 220 DNA

9 Billy 340 DNA

10 Norbert 700 FCS

11 Stuart 84 FCS

12 Mahoor 190 FCS

**Using database custom create a table users**

*hive (custom)> CREATE TABLE users*

*> (*

*> id INT,*

*> name STRING,*

*> salary INT,*

*> unit STRING*

*> )*

*> ROW FORMAT DELIMITED*

*> FIELDS TERMINATED BY '\t';*

*OK*

*Time taken: 0.045 seconds*

**Load data**

*hive (custom)> LOAD DATA LOCAL INPATH '/home/acadgild/Datasets/users.txt'*

*> INTO TABLE users;*

*Loading data to table custom.users*

*Table custom.users stats: [numFiles=1, totalSize=194]*

*OK*

*Time taken: 0.103 seconds*

**Check data has been imported**

*hive (custom)> select \* from users;*

*OK*

*1 Amit 600 DNA*

*2 Sumit 200 DNA*

*3 Yadav 229 DNA*

*4 Sunil 500 FCS*

*5 Kranti 100 FCS*

*6 Bob 201 FCS*

*7 Duncan 150 DNA*

*8 Sue 220 DNA*

*9 Billy 340 DNA*

*10 Norbert 700 FCS*

*11 Stuart 84 FCS*

*12 Mahoor 190 FCS*

*Time taken: 0.028 seconds, Fetched: 12 row(s)*

**Determine what results should be returned**

*Hive (custom)> SELECT rank() OVER (PARTITION BY unit ORDER BY salary DESC) AS rank, id, name, salary, unit*

*FROM users;*

*Total MapReduce CPU Time Spent: 3 seconds 540 msec*

*OK*

*1 1 Amit 600 DNA*

*2 9 Billy 340 DNA*

*3 3 Yadav 229 DNA*

*4 8 Sue 220 DNA*

*5 2 Sumit 200 DNA*

*6 7 Duncan 150 DNA*

*1 10 Norbert 700 FCS*

*2 4 Sunil 500 FCS*

*3 6 Bob 201 FCS*

*4 12 Mahoor 190 FCS*

*5 5 Kranti 100 FCS*

*6 11 Stuart 84 FCS*

**The highlighted records in yellow should be returned**

**Query to produce results**

*SELECT name, salary, lead from*

*(select lead(salary) OVER (PARTITION BY unit ORDER BY salary) as lead,id,name,salary,unit*

*FROM users)*

*temp*

*where (lead-salary) <100;*

**Results**

*OK*

*Duncan 150 200*

*Sumit 200 220*

*Sue 220 229*

*Stuart 84 100*

*Kranti 100 190*

*Mahoor 190 201*

*Time taken: 20.191 seconds, Fetched: 6 row(s)*

*hive (custom)>*

**\*\*returned as expected**

## List of all employees who draw higher salary than the average salary of that department

**Using the same database and table as in solution 1.1**

Query used

*select users.\**

*from (select users.\*, avg(salary) over (partition by unit) as avgsalary*

*from users*

*) users*

*where users.salary > users.avgsalary;*

**Determine what results should be returned**

*Hive (custom)> SELECT rank() OVER (PARTITION BY unit ORDER BY salary DESC) AS rank, id, name, salary, unit*

*FROM users;*

*Total MapReduce CPU Time Spent: 3 seconds 540 msec*

*OK*

*1 1 Amit 600 DNA*

*2 9 Billy 340 DNA*

*3 3 Yadav 229 DNA*

*4 8 Sue 220 DNA*

*5 2 Sumit 200 DNA*

*6 7 Duncan 150 DNA*

*1 10 Norbert 700 FCS*

*2 4 Sunil 500 FCS*

*3 6 Bob 201 FCS*

*4 12 Mahoor 190 FCS*

*5 5 Kranti 100 FCS*

*6 11 Stuart 84 FCS*

The highlighted records in green should be returned as they have salaries greater than the average for each unit.

**The averages**

***For DNA = 289.83***

***For FCS = 295***

**The highlighted records should be returned**

**Results**

OK

9 Billy 340 DNA 289.8333333333333

1 Amit 600 DNA 289.8333333333333

10 Norbert 700 FCS 295.8333333333333

4 Sunil 500 FCS 295.8333333333333

Time taken: 20.291 seconds, Fetched: 4 row(s)

hive (custom)>

**\*\*returned as expected**