System Requirements Specification Index

For

Pyspark Usecase

Employee Salary and Tenure AnalysisVersion

1.0

IIHT Pvt. Ltd.
fullstack@iiht.com

Problem Statement : Employee Salary and Tenure Analysis
Description : Use relevant methods operations toperform specified activities which are given in the instructions.

XYZ Corporation is conducting an internal analysis to understand the distribution of salaries across departments, identify the highest and lowest earners, and gain insights into the tenure and demographics of its employees. The company wants to make data-driven decisions regarding salary adjustments, promotions, and department-level resource allocation.
They have collected data on employees' names, dates of birth, dates of joining, salaries, and departments. The company's data science team has been tasked with analyzing this data using PySpark to efficiently handle and process large datasets.

Objective:
The goal is to answer several key business questions based on the provided employee data:

1. **Identifying the Employee(s) with the Maximum Salary:**
   XYZ Corp wants to recognize high-performing employees, particularly those who are the top earners in the company. The management needs to know who earns the highest salary to consider them for leadership roles or other recognitions.
2. **Identifying the Employee(s) with the Minimum Salary:**
   The company is concerned about pay equity and wants to identify those who are at the lower end of the salary spectrum. Understanding who earns the least can help HR review compensation structures and ensure fair pay across roles.
3. **Identifying the Youngest Employee:**
   The management is keen on understanding the demographics of its workforce. Identifying the youngest employee helps in creating mentoring programs where experienced employees can guide the younger ones.
4. **Identifying the Senior-most Employee:**
   Seniority often correlates with experience and loyalty to the company. Identifying the longest-serving employee is crucial for recognizing their contributions and possibly involving them in key decisions or as part of the company's legacy initiatives.
5. **Identifying the Department with the Highest Average Salary:**
   To assess budget allocation, the management wants to identify which department has the highest average salary. This information could be used for future budgeting, talent acquisition, or re-evaluating departmental pay scales.

Questions Based on the Code:

1. **Who are the employees with the maximum salary, and how might their roles impact their earnings?**
2. **Who are the employees with the minimum salary, and are their roles potentially undervalued in the company?**
3. **Who is the youngest employee in the company, and what department do they belong to?**
4. **Who is the senior-most employee, and how has their experience possibly contributed to their department or the company as a whole?**
5. **Which department has the highest average salary, and what factors could contribute to this?**

Use the same dataset

(John Doe', '1990-05-15', '2010-01-01', 50000, 'Engineering'),

('Jane Smith', '1985-07-20', '2008-03-15', 75000, 'HR'),

('Mike Johnson', '1970-10-10', '1995-10-20', 95000, 'Finance'),

('Emily Davis', '1995-12-25', '2020-06-10', 60000, 'Engineering'),

('Robert Brown', '1980-11-05', '2005-07-25', 85000, 'Finance')

Execution Steps to Follow:

1. All actions like build, compile, running application,running test cases will be through Command Terminal.
2. To open the command terminal the test takers, need to go to Application menu (Three horizontal lines at left top) -> Terminal -> New Terminal
3. This editor Auto Saves the code
4. If you want to exit(logout) and continue the coding later anytime (using Save & Exit
option on Assessment Landing Page) then you need to use CTRL+Shift+B-command compulsorily on code IDE. This will push or save the updated contents in the internal git/repository. Else the code will not be available in the next login.
5. These are time bound assessments the timer would stop if you logout and while logging in back using the same credentials the timer would resume from the same time it was stopped from the previous logout.
6. To setup environment:
You can run the application without importing any packages
7. To launch application:
Pip install pyspark
python3 empspark.py
8. To run Test cases:
python3 -m unittest
9. Before Final Submission also, you need to use CTRL+Shift+B-command compulsorily
on code IDE. This will push or save the updated contents in the internal git/repository for code quality analysis graph.

-----x-----