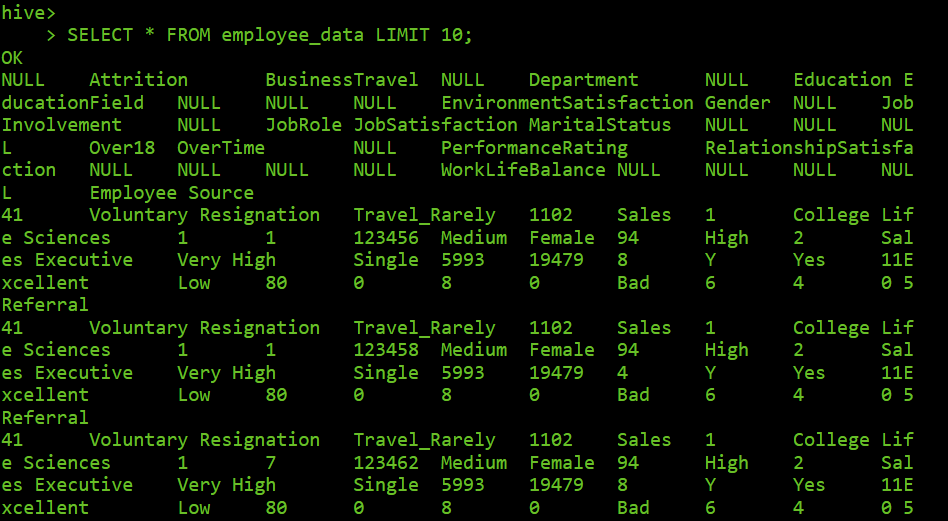
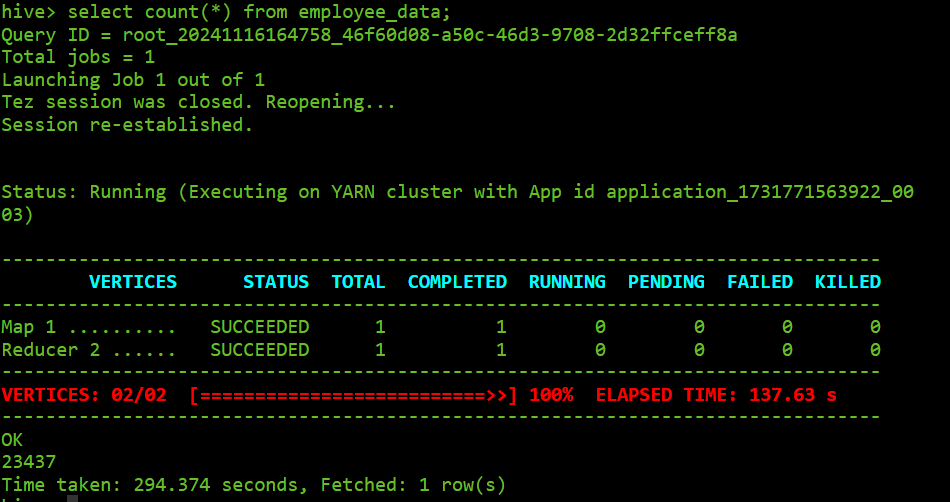
Hive commands and their outputs:

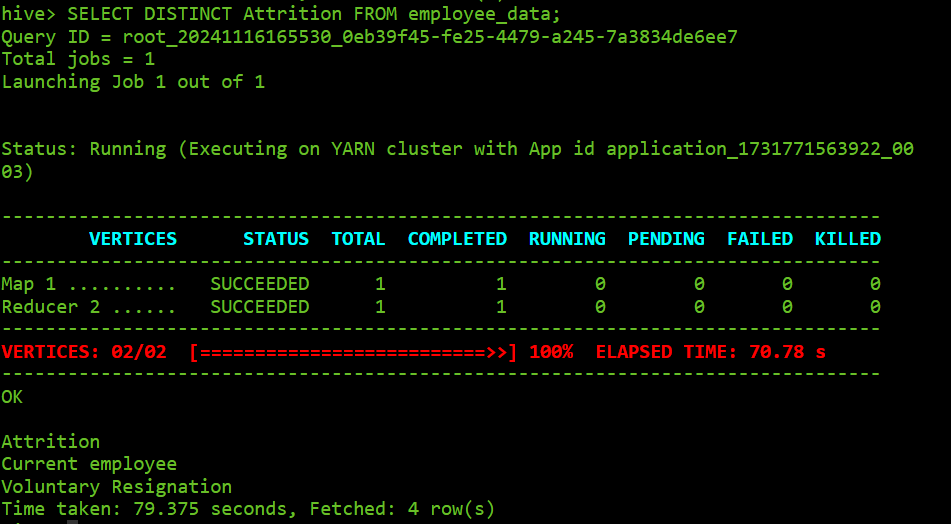
1. Check the first few rows of the dataset:



1. Total number of records



1. Distinct values in the 'Attrition' column



4)null values checking in all coulmns:

SELECT

COUNT(CASE WHEN Age IS NULL THEN 1 END) AS Age\_NULLs,

COUNT(CASE WHEN Attrition IS NULL THEN 1 END) AS Attrition\_NULLs,

COUNT(CASE WHEN BusinessTravel IS NULL THEN 1 END) AS BusinessTravel\_NULLs,

COUNT(CASE WHEN DailyRate IS NULL THEN 1 END) AS DailyRate\_NULLs,

COUNT(CASE WHEN Department IS NULL THEN 1 END) AS Department\_NULLs,

COUNT(CASE WHEN DistanceFromHome IS NULL THEN 1 END) AS DistanceFromHome\_NULLs,

COUNT(CASE WHEN Education IS NULL THEN 1 END) AS Education\_NULLs,

COUNT(CASE WHEN EducationField IS NULL THEN 1 END) AS EducationField\_NULLs,

COUNT(CASE WHEN EmployeeCount IS NULL THEN 1 END) AS EmployeeCount\_NULLs,

COUNT(CASE WHEN EmployeeNumber IS NULL THEN 1 END) AS EmployeeNumber\_NULLs,

COUNT(CASE WHEN ApplicationID IS NULL THEN 1 END) AS ApplicationID\_NULLs,

COUNT(CASE WHEN EnvironmentSatisfaction IS NULL THEN 1 END) AS EnvironmentSatisfaction\_NULLs,

COUNT(CASE WHEN Gender IS NULL THEN 1 END) AS Gender\_NULLs,

COUNT(CASE WHEN HourlyRate IS NULL THEN 1 END) AS HourlyRate\_NULLs,

COUNT(CASE WHEN JobInvolvement IS NULL THEN 1 END) AS JobInvolvement\_NULLs,

COUNT(CASE WHEN JobLevel IS NULL THEN 1 END) AS JobLevel\_NULLs,

COUNT(CASE WHEN JobRole IS NULL THEN 1 END) AS JobRole\_NULLs,

COUNT(CASE WHEN JobSatisfaction IS NULL THEN 1 END) AS JobSatisfaction\_NULLs,

COUNT(CASE WHEN MaritalStatus IS NULL THEN 1 END) AS MaritalStatus\_NULLs,

COUNT(CASE WHEN MonthlyIncome IS NULL THEN 1 END) AS MonthlyIncome\_NULLs,

COUNT(CASE WHEN MonthlyRate IS NULL THEN 1 END) AS MonthlyRate\_NULLs,

COUNT(CASE WHEN NumCompaniesWorked IS NULL THEN 1 END) AS NumCompaniesWorked\_NULLs,

COUNT(CASE WHEN Over18 IS NULL THEN 1 END) AS Over18\_NULLs,

COUNT(CASE WHEN OverTime IS NULL THEN 1 END) AS OverTime\_NULLs,

COUNT(CASE WHEN PercentSalaryHike IS NULL THEN 1 END) AS PercentSalaryHike\_NULLs,

COUNT(CASE WHEN PerformanceRating IS NULL THEN 1 END) AS PerformanceRating\_NULLs,

COUNT(CASE WHEN RelationshipSatisfaction IS NULL THEN 1 END) AS RelationshipSatisfaction\_NULLs,

COUNT(CASE WHEN StandardHours IS NULL THEN 1 END) AS StandardHours\_NULLs,

COUNT(CASE WHEN StockOptionLevel IS NULL THEN 1 END) AS StockOptionLevel\_NULLs,

COUNT(CASE WHEN TotalWorkingYears IS NULL THEN 1 END) AS TotalWorkingYears\_NULLs,

COUNT(CASE WHEN TrainingTimesLastYear IS NULL THEN 1 END) AS TrainingTimesLastYear\_NULLs,

COUNT(CASE WHEN WorkLifeBalance IS NULL THEN 1 END) AS WorkLifeBalance\_NULLs,

COUNT(CASE WHEN YearsAtCompany IS NULL THEN 1 END) AS YearsAtCompany\_NULLs,

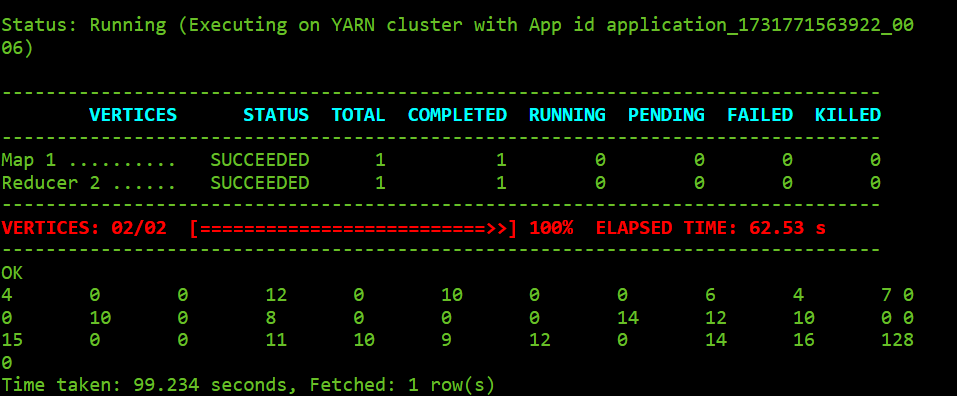
COUNT(CASE WHEN YearsInCurrentRole IS NULL THEN 1 END) AS YearsInCurrentRole\_NULLs,

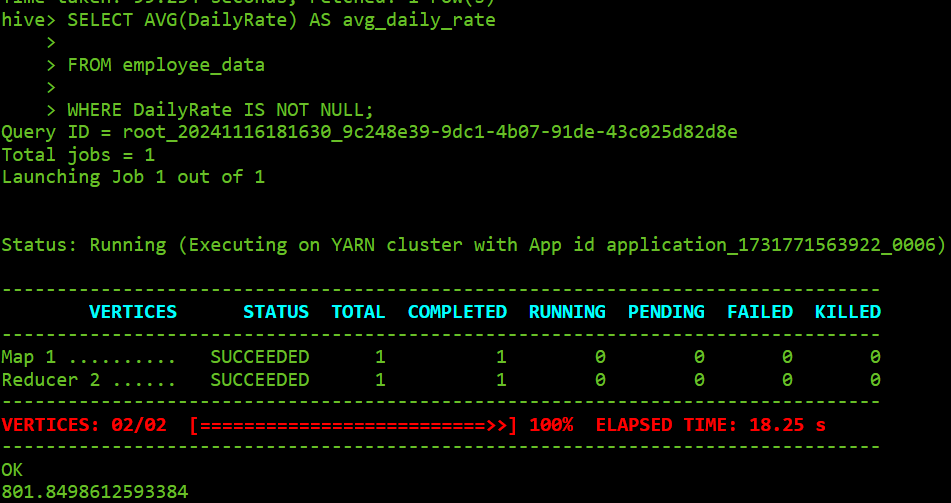
COUNT(CASE WHEN YearsSinceLastPromotion IS NULL THEN 1 END) AS YearsSinceLastPromotion\_NULLs,

COUNT(CASE WHEN YearsWithCurrManager IS NULL THEN 1 END) AS YearsWithCurrManager\_NULLs,

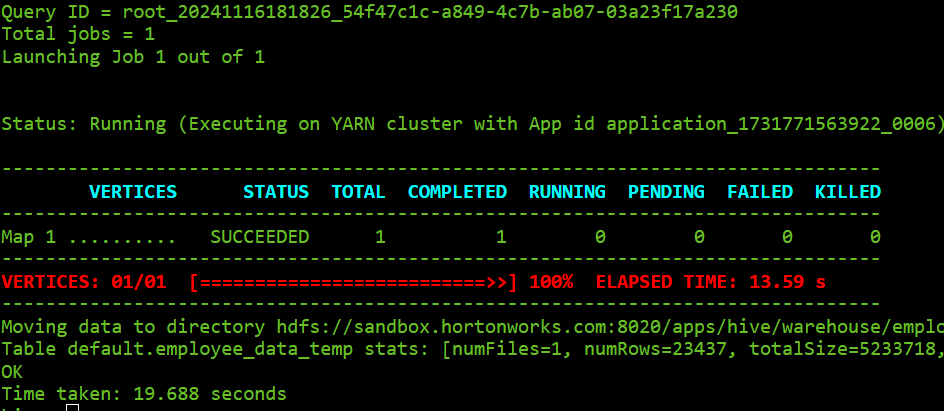
COUNT(CASE WHEN EmployeeSource IS NULL THEN 1 END) AS EmployeeSource\_NULLs

FROM employee\_data;

  
  
and then filling their columns with their mean ,mode and some with unknown:

->mean:  
getting mean:  


Filling with this mean:  
(here we create temp table with mean values filled ,which is used furthur)



This new table is renamed as employee table with old table deleted,  
  
so similar process is repeated for other columns,filing with mean,mode and unknown