**Employee Review of Comment Analysis**

DESCRIPTION

**Objective**: To use Hive features for data analysis and sharing the actionable insights into the HR team for taking corrective actions.

**Problem Statement:**The HR team is surfing social media to gather current and ex-employee feedback or sentiments.  
This information gathered will be used to derive actionable insights and take corrective actions to improve the employer-employee relationship. The data is web-scraped from Glassdoor and contains detailed reviews of 67K employees from Google, Amazon, Facebook, Apple, Microsoft, and Netflix.

**Domain:** Human Resource

**Analysis to be done:** Exploratory analysis, to determine features and relationships impacting employee satisfaction and derive actionable insights by learning from the historical data

**Content:** This data set contains employee\_review\_data.csv separated into the following categories:

* Index: Index number.
* Company: Name of the company being reviewed.
* Location: This dataset is global, as it includes the country&#39;s name in parenthesis [for example,

"Toronto, ON(Canada)"]. However, if the location is the USA then it will only include the city and state [i.e. "Los Angeles, CA"].

* Date Posted: Date posted will be in the following format MM DD, YYYY.
* Job-Title: This string will also include whether the reviewer is a "Current" or "Former" Employee at the time of review. Both are fixed-length strings (“Current Employee “and “Former Employee “) followed by the role of reviewer.
* Summary: Short summary of employee review.
* Pros: Pros
* Cons: Cons
* Overall Rating: 1-5
* Work/Life Balance Rating: 1-5
* Culture and Values Rating: 1-5
* Career Opportunities Rating: 1-5
* Comp and Benefits Rating: 1-5
* Senior Management Rating: 1-5
* Helpful Review Count: A count of how many people found the review to be helpful/

**Steps to perform:**

* Create a hive table partitioned by country and bucketed by year and also load the review.csv file.

**Note**: Ensure that the right hive environment variable is set for bucket insert.

* Impute the missing value (none) for all rating columns with a numerical value between 0 and 5.

**Note**: For imputation, calculate the median for each of the 5 rating fields and create a new table.

1. Work-balance stars
2. Culture values stars
3. Career opportunities-stars
4. Comp-benefit-stars
5. Senior-management-stars

* Write the final relation schema to review.csv file in your HDFS home directory.
* Using the over-all rating fields display trend:

1. Globally by company

* Identify trends at 25%, 50%, 75%

      2. Globally by company per year

* Identify trends at 25%, 50%, 75%

      3. By company by country (Identify trends for each company by country

* Identify trends at 25%, 50%, 75%

* Display the impact of employee status on rating a company using the overall-ratings field by the company by year.
* Display the impact of job role on rating a company using the overall-ratings field by the company by year.
* Display the relationship between the overall rating score vs. the rest of the rating field scores by company. Also, document your findings.
* Overall-ratings

Versus

* Work-balance stars
* Culture values stars
* Career opportunities-stars
* Comp-benefit-stars
* Senior-management-stars

* Document your findings for the following:

a) Which corporation is worth working for  
b) Classification of satisfied or unsatisfied employees