Vinaya Rajaram Nayak

U74501808

Hive / Impala Assignment

1. Clone (or pull) your data into the Cloudera VM

git clone <https://github.com/nayakvinaya/HDFSHive>

1. In the VM, use the terminal to create a HDFS folder called data16 located within the /user/cloudera folder.
2. Copy the data from the local filesystem into your HDFS /user/cloudera/data16 folder

hdfs dfs -mkdir /user/cloudera/data16

hdfs dfs -put /home/cloudera/HDFSHive/layoffs\_data.csv /user/cloudera/data16

hdfs dfs -chmod 777 /user/cloudera/data16/layoffs\_data.csv

1. Come up with a question that you want to be answered using the data - Create an SQL query to answer this question (do more than select \* -- demonstrate filtering and aggregation (such as count or average)
2. **What is the total laid off count by industry?**

SELECT industry,SUM(laid\_off\_count) AS total\_count

FROM data16.layoffs\_data

GROUP BY industry

ORDER BY total\_count DESC;

Graphical user interface, text, application

Description automatically generated

1. **What are the top 5 companies in each industry with highest layoffs?**

SELECT

Industry,

Company,

Laid\_Off\_Count,

Company\_Rank

FROM

(

SELECT

Industry,

Company,

Laid\_Off\_Count,

RANK() OVER (PARTITION BY Industry ORDER BY Laid\_Off\_Count DESC) AS Company\_Rank

FROM

data16.layoffs\_data

WHERE

Laid\_Off\_Count > 0

) ranked\_data

WHERE

Company\_Rank <= 5

ORDER BY

Industry ASC,

Company\_Rank ASC;



1. **What is the average layoffs by industry?**

SELECT

Industry,

AVG(Laid\_Off\_Count) AS Average\_Layoffs

FROM data16.layoffs\_data

WHERE

industry is not null

GROUP BY

Industry

ORDER BY

Average\_Layoffs desc ;

Text

Description automatically generated

1. **What are the top 10 countries with highest layoffs?**

SELECT

Country,

SUM(Laid\_Off\_Count) AS Total\_Layoffs

FROM

data16.layoffs\_data

GROUP BY

Country

ORDER BY

Total\_Layoffs DESC

LIMIT 10;

Text

Description automatically generated with low confidence