**Analysis of clinical trial data**

-- Hive Queries for Assignment

-- Create Table clinicaltrial\_2021

CREATE TABLE IF NOT EXISTS

clinicaltrial\_2021(Id STRING, Sponsor STRING,

Status STRING, Start STRING,

Completion STRING,

Type String, Submission STRING,

Conditions STRING, Interventions STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'

LINES TERMINATED BY '\n'

TBLPROPERTIES("skip.header.line.count"="1");

-- Loading data in clinicaltrial\_2021 table

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/clinicaltrial\_2021.csv'

INTO TABLE clinicaltrial\_2021;

SELECT \* FROM clinicaltrial\_2021 LIMIT 5;

Text

Description automatically generated

-- Create pharama Table

CREATE TABLE IF NOT EXISTS pharma(

Company STRING,

Parent\_Company STRING,

Penalty\_Amount STRING,

Subtraction\_From\_Penalty STRING,

Penalty\_Amount\_Adjusted\_For\_Eliminating\_Multiple\_Counting STRING,

Penalty\_Year STRING,

Penalty\_Date STRING,

Offense\_Group STRING,

Primary\_Offense STRING,

Secondary\_Offense STRING,

Description STRING,

Level\_of\_Government STRING,

Action\_Type STRING,

Agency STRING,

Civil\_Criminal STRING,

Prosecution\_Agreement STRING,

Court STRING,

Case\_ID STRING,

Private\_Litigation\_Case\_Title STRING,

Lawsuit\_Resolution STRING,

Facility\_State STRING,

City STRING,

Address STRING,

Zip STRING,

NAICS\_Code STRING,

NAICS\_Translation STRING,

HQ\_Country\_of\_Parent STRING,

HQ\_State\_of\_Parent STRING,

Ownership\_Structure STRING,

Parent\_Company\_Stock\_Ticker STRING,

Major\_Industry\_of\_Parent STRING,

Specific\_Industry\_of\_Parent STRING,

Info\_Source STRING,

Notes STRING)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES ( "separatorChar" = ",", "quoteChar" = "\"", "escapeChar" = "\\" )

TBLPROPERTIES("skip.header.line.count"="1");

-- Load Data into pharama Table

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/pharma.csv'

INTO TABLE pharma;

SELECT \* FROM pharma LIMIT 5;

Text, letter

Description automatically generated

-- Create mesh table

CREATE TABLE IF NOT EXISTS mesh(

term STRING, tree STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

TBLPROPERTIES("skip.header.line.count"="1");

-- Loading Data into mesh Table

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/mesh.csv'

INTO TABLE mesh;

SELECT \* FROM mesh LIMIT 5;

Text

Description automatically generated

-- Answer 1

-- 1. The number of studies in the dataset

SELECT COUNT(DISTINCT Id) FROM clinicaltrial\_2021;

Text, letter

Description automatically generated

-- Answer 2

-- 2. Type of Studies

SELECT Type, COUNT(Type) as Count FROM clinicaltrial\_2021

GROUP BY Type ORDER BY Count DESC;

Text, letter

Description automatically generated

Chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

-- Answer 3

-- 3. Top 5 Conditions

SELECT Condition, COUNT(Condition) as Count FROM (SELECT EXPLODE(SPLIT(Conditions, ","))

as Condition FROM clinicaltrial\_2021) as Condition

GROUP BY Condition HAVING Condition != "" ORDER BY Count DESC LIMIT 5;

Text, letter

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, funnel chart

Description automatically generated

-- Answer 4

-- 4. Most Frequent roots

SELECT tree.root, COUNT(tree.root) as Count FROM

(SELECT EXPLODE(SPLIT(Conditions, ",")) as Condition FROM clinicaltrial\_2021) as Condition

LEFT JOIN (SELECT term, SPLIT(tree, "\\.")[0] as root FROM mesh) as tree ON

(Condition.Condition = tree.term)

GROUP BY tree.root ORDER BY Count DESC LIMIT 10;

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Chart, timeline, bar chart

Description automatically generated

-- Answer 5

-- 5. Most Common sponsors that are not pharmaceutical companies,

-- with clinical trials they have sponsored

SELECT trial.Sponsor, COUNT(trial.Sponsor) as Count FROM clinicaltrial\_2021 AS trial

LEFT OUTER JOIN pharma ON trial.Sponsor = pharma.Parent\_Company

GROUP BY trial.Sponsor, pharma.Parent\_Company HAVING pharma.Parent\_Company IS NULL

ORDER BY Count DESC LIMIT 10;

Text, letter

Description automatically generated

Text

Description automatically generated

Chart, bar chart

Description automatically generated

-- Answer 6

-- 6. Completed Case Studies in 2021

SELECT month, COUNT(month) as count FROM (SELECT SPLIT(Completion, " ")[0] as month,

SPLIT(Completion, " ")[1] as year, Status FROM clinicaltrial\_2021) as temp WHERE

temp.year = "2021" AND Status = "Completed" GROUP BY month ORDER BY count DESC;

Text, letter

Description automatically generated

Text

Description automatically generated

Chart, bar chart

Description automatically generated

Timeline

Description automatically generated with medium confidence