Analysis of clinical trial data

# Part 1 – Setup required for analysis

To analyze the clinical trial dataset, I have used Databricks to work with Pyspark and HQL using the Python & SQL notebooks to execute the code in a web based interactive interface. Different python libraries are used to extract the conclusions. I have also attached my code for each answer from Q1 to Q6. Each code segment is written with comments to explain the answers. Clinical trial data is analyzed to get the desired results from different problem statements.

# Part 2 – Data Cleaning and Preparation

Dataset for clinical trial was provided in compressed csv file. To use the dataset in databricks environment I have uploaded the dataset in databricks FileStore>tables directory.

![Graphical user interface, application, email

Description automatically generated]()

After data uploading, I have uncompressed the dataset to csv format for performing the actions. To uncompressing the data I have used the unix command in python notebook and first I moved the compressed file to databricks FileStore/tmp directory and after uncompressing I again moved the file to its original directory dbfs/tables directory. For this whole process I have I have used a fileroot variable which performed the taking file path, moving and uncompressing actions.

![Graphical user interface, text, application, email

Description automatically generated]()

![Graphical user interface, text, application, Teams

Description automatically generated]()

These steps are done in PySpark both RDD & Dataframe. To abstract the data, data cleaning of the csv file was compulsory. So, I have I went through the dataset and removed/skipped the header and split the dataset using delimiters pipe.

For running the dataset throughout the file I have declared two variables one (fileroot) for file name and other (Year) for year as dataset year for relevant and accurate results.

# Part 3 – Problem Answers

1. Problem Statement I :

The number of studies in the dataset. You must ensure that you explicitly check distinct studies.

**Answer**:

(dataset used to produce answer: clinicaltrial\_2021.csv)

Count: 387261 #Total number of distinct studies

**Code Segments Used in RDD Implementation:**

According to problem statement dataset has studies which involves the clinical trial analysis. To extract the distinct studies from the dataset I have assigned the dataset to the RDD2021 and for accurate results I have skipped the first row as it’s the header by filtering the RDD2021. Then I have taken the distinct count of the dataset studies.

RDD2021= sc.textFile('FileStore/tables/'+fileroot)

RDD2021\_New = RDD2021

RDD2021\_Header = RDD2021\_New.first()

RDD2021\_Count = RDD2021.filter(lambda Header: Header!=RDD2021\_Header)

RDD2021\_Count.distinct().count()

**Screenshot:**

![Graphical user interface, text, application, email

Description automatically generated]()

**Code Segments Used in DF Implementation:**

In dataframe implementation I have load the file in DF2021 and made further steps for header, schema and delimiter pipe. Then took the count for studies and results are accurate as per RDD.

ClinicalTrial2021= ('dbfs:/FileStore/tables/'+fileroot)

DF2021 = spark.read.format('csv')\

.options(header='true', inferSchema='true')\

.options(delimiter='|')\

.load(ClinicalTrial2021)

DF2021.count()

**Screenshot:**

![Graphical user interface, text, application, email

Description automatically generated]()

**Code Segments Used in HQL Implementation:**

In Hive SQL implementation I have made an external table clinicaltrial\_2021 by specifying the schema and inserted the data from the dataset. And choose the studies count from the table.

CREATE EXTERNAL TABLE IF NOT EXISTS clinicaltrial\_2021

( Id STRING ,

Sponsor STRING ,

Status STRING ,

Start STRING ,

Completion STRING ,

Type STRING ,

Submission STRING ,

Conditions STRING ,

Interventions STRING )

USING CSV

OPTIONS (path 'dbfs:/FileStore/tables/$fileroot',

delimiter "|",

header "true");

select count(\*) from clinicaltrial\_2021;

**Screenshot:**

![Graphical user interface, text, application, email

Description automatically generated]()

1. Problem Statement II :

You should list all the types (as contained in the Type column) of studies in the dataset along with the frequencies of each type. These should be ordered from most frequent to least frequent.

**Answer**:

(dataset used to produce answer: clinicaltrial\_2021.csv)

[('Interventional', **301472**),

('Observational', **77540**),

('Observational [Patient Registry]', **8180**),

('Expanded Access', **69**)]

**Code Segments Used in RDD Implementation:**

For problem statement implementation in RDD I have splitted the RDD2021\_Split by pipe ‘|’. Then I have taken the 6th index column name ‘type’ to check the types with frequencies. Then I used reduceByKey function for grouping the same keys and adding their values. As per requirement I sorted the final results in descending order for most frequent types at top.

RDD2021\_Split=RDD2021.map(lambda Word:Word.split('|'))

RDD2021\_Type=RDD2021\_Split.map(lambda index: index[5])\

.map(lambda x:(x,1)).reduceByKey(lambda v1,v2:v1+v2).sortBy(lambda v:v[1],ascending =False)

RDD2021\_Type.take(4)

**Screenshot:**

**![Graphical user interface, text, application, email

Description automatically generated]()**

**Code Segments Used in DF Implementation:**

In dataframe implementation I grouped the Type column by taking their count and then ordered it in descending.

DF2021\_Count1=DF2021.groupby(['Type']).count()

DF2021\_Count=DF2021\_Count1.orderBy(['count'], ascending=False)

DF2021\_Count.display()

**Screenshot:**

**![Graphical user interface, table

Description automatically generated]()**

**Code Segments Used in HQL Implementation:**

In HQL implementation I made a query of clinicaltrial\_2021 table abd grouped all the types from type column by taking the count and then ordered it as per desired output.

select Type, count(Type) as count from clinicaltrial\_2021

group by Type

order by count

desc limit 4

**Screenshot:**

![Graphical user interface, table

Description automatically generated]()

1. Problem Statement III :

The top 5 conditions (from Conditions) with their frequencies.

**Answer**:

(dataset used to produce answer: clinicaltrial\_2021.csv)

[('Carcinoma', 13389),

('Diabetes Mellitus', 11080),

('Neoplasms', 9371),

('Breast Neoplasms', 8640),

('Syndrome', 8032)]

**Code Segments Used in RDD Implementation:**

For problem statement 3 I have mapped the each value from conditions column and splitted it against comma delimiter and skipped the nulled values by filtering the RDD. The I have mapped the RDD and assigned 1 to each key. reduceByKey function is used for grouping the same keys and adding their values and sorted the RDD on values in descending order.

RDD2021\_Disease1=RDD2021\_Split.flatMap(lambda index: index[7].split(','))\

.filter(lambda c: c != '')

RDD2021\_Disease=RDD2021\_Disease1.map(lambda x:(x,1)).reduceByKey(lambda v1,v2:v1+v2).sortBy(lambda v:v[1],ascending =False)

RDD2021\_Disease.take(5)

**Screenshot:**

Graphical user interface, text, application, email

Description automatically generated

**Code Segments Used in DF Implementation:**

In dataframe implementation I split the column condition on comma and used the explode function to list all the items on row from the column to new column named conditions. Then I grouped the conditions and make the order descending of the conditions count.

from pyspark.sql.functions import col, split, explode

DF2021\_Split\_Conditions= DF2021.withColumn('conditions', explode(split(DF2021['conditions'],',')))

DF2021\_Conditions\_Count1=DF2021\_Split\_Conditions.groupby(['conditions']).count()

DF2021\_Conditions\_Count=DF2021\_Conditions\_Count1.orderBy(['count'], ascending=False)

DF2021\_Conditions\_Count.display()

**Screenshot:**

![Table

Description automatically generated]()

**Code Segments Used in HQL Implementation:**

For HQL implementation I created a view in which I splitted the column and used explode function for converting the list into row in a new column named explode\_condition. I skipped the nulled values by checking the length of each row and make a count by grouping the conditions and sorted it as per desired results.

CREATE OR REPLACE VIEW ConditionSplit\_view

AS

SELECT id, explode\_codition

FROM clinicaltrial\_2021

lateral VIEW explode(split(Conditions,",")) condition\_view AS explode\_codition

where length(explode\_codition) > 0 ;

select explode\_codition,count(explode\_codition) as count

from ConditionSplit\_view

group by explode\_codition

order by count desc limit 6

**Screenshot:**

![Graphical user interface, table

Description automatically generated]()

1. Problem Statement IV :

Each condition can be mapped to one or more hierarchy codes. The client wishes to know the 5 most frequent roots (i.e. the sequence of letters and numbers before the first full stop) after this is done.

**Answer**:

(dataset used to produce answer: clinicaltrial\_2021.csv and Mesh.csv)

[('C04', 143994),

('C23', 136079),

('C01', 106674),

('C14', 94523),

('C10', 92310),

('C06', 85646),

('C08', 70720),

('C13', 42599),

('C18', 41276),

('C12', 40161)]

**Code Segments Used in RDD Implementation:**

Each condition has its own hierarchy code. To get access with it I have used mesh file and cleaned it by skipping the header and splitting it on comma and full stop to get the desired part of the code related to condition. Then I joined the clinicaltrial\_20211’s conditions with cleaned mesh file to join conditions and codes. Then I used reducebykey to group the similar keys and added their values. And sorted it in descending order to show the most frequent at top.

RDDmesh\_Header = RDDmesh.first()

RDDmesh\_Header\_Remove = RDDmesh.filter(lambda Header: Header!=RDDmesh\_Header)

RDDmesh\_Split = RDDmesh\_Header\_Remove.map(lambda x:(x.split(",")[0], x.split(",")[1].split(".")[0]))

RDDmesh\_Split.take(2)

RDD2021\_Recount= RDD2021\_Disease1.map( lambda x:[x,1] )

RDD2021\_DiseaseCode = RDD2021\_Recount.join(RDDmesh\_Split).map(lambda y: (y[1][1], 1))\

.reduceByKey(lambda v1, v2: v1+ v2).sortBy( lambda x: -x[1] )

RDD2021\_DiseaseCode.take(10)

**Screenshot:**

**![Graphical user interface, text, application, email

Description automatically generated]()**

**Code Segments Used in DF Implementation:**

In dataframe implementation I have loaded the mesh file into DFmesh by defining the schema and delimiter. Then to take code part from mesh file I have used substr function by taking first three letters from the code with a new column name NewCode. Then I compared the condition file and mesh file on the basis of conditions and joined both dataframes. Now each condition has its own code and then I have group them and made the count and sorted it in descending order.

mesh=('dbfs:/FileStore/tables/mesh.csv')

DFmesh = spark.read.format('csv')\

.options(header='true', inferSchema='true')\

.options(delimiter=',')\

.load(mesh)

display(DFmesh)

DFmesh\_Code= DFmesh.withColumn('NewCode', DFmesh.tree.substr(1,3))

DFmesh\_Join1= DF2021\_Split\_Conditions.conditions == DFmesh\_Code.term

DFmesh\_Join= DF2021\_Split\_Conditions.join(DFmesh\_Code, on = DFmesh\_Join1, how='inner')

DFmesh\_Disease\_Code=DFmesh\_Join.select('NewCode').groupby(['NewCode']).count().orderBy(['count'],

ascending= False)

DFmesh\_Disease\_Code.display()

**Screenshot:**

**![Graphical user interface, application

Description automatically generated]()**

**Code Segments Used in HQL Implementation:**

For HQL implementation I have created table using the mesh file. Then I created a view from code by using substring function . then I joined the condition column and disease code where both conditions are same. Then I made a count, grouped by same keys and sorted it according to requirement.

CREATE External TABLE IF NOT EXISTS mesh (

term STRING ,

tree STRING )

USING CSV

OPTIONS ( path "dbfs:/FileStore/tables/mesh.csv",

delimiter ",",

header "true" );

CREATE OR REPLACE VIEW disease\_code As

select term,tree,substr( tree,1,3 ) as disease\_code from mesh

CREATE OR REPLACE VIEW joinDiseasecodeView As

SELECT c.id,c.explode\_codition,d.disease\_code

FROM ConditionSplit\_view as c

JOIN disease\_code as d ON c.explode\_codition=d.term

select disease\_code, count(disease\_code) as count

from joinDiseasecodeView

group by disease\_code

order by count desc limit 10

**Screenshot:**

**![Table

Description automatically generated]()**

1. Problem Statement V

Find the 10 most common sponsors that are not pharmaceutical companies, along with the number of clinical trials they have sponsored. Hint: For a basic implementation, you can assume that the Parent Company column contains all possible pharmaceutical companies.

**Answer**:

(dataset used to produce answer: clinicaltrial\_2021.csv and Pharma.csv)

[('National Cancer Institute (NCI)', 3218),

('M.D. Anderson Cancer Center', 2414),

('Assistance Publique - Hôpitaux de Paris', 2369),

('Mayo Clinic', 2300),

('Merck Sharp & Dohme Corp.', 2243),

('Assiut University', 2154),

('Novartis Pharmaceuticals', 2088),

('Massachusetts General Hospital', 1971),

('Cairo University', 1928),

('Hoffmann-La Roche', 1828)]

**Code Segments Used in RDD Implementation:**

In this implementation Pharma file has all the pharmaceutical companies list but in clinical trial file there are some sponsors which were not pharmaceutical company. So I used the sponsor column from clinicaltrial file and skipped the nulled fields by filtering the RDD. I used leftouterjoin function to pick the unmatched sponsor companies from clinicaltrial file which were not in Pharma file. Then I again filtered the resulted RDD to skipped the null fields. Reducebykey function and count function used to find the frequency of each sponsor for trial study.

RDD2021\_Sponsor = RDD2021\_Split.map(lambda s: s[1]) \

.filter(lambda c: c != '') \

.map(lambda x: (x, 1))

RDDpharma\_Split=RDDpharma.map(lambda x: x.replace('"','').split(',')[1]) \

.map(lambda v: (v,1))

RDD2021\_Joined = RDD2021\_Sponsor.leftOuterJoin(RDDpharma\_Split).filter(lambda x : x[1][1] == None) \

.map(lambda y: (y[0], 1)) \

.reduceByKey(lambda v1, v2: v1+ v2).sortBy(lambda x: -x[1])

RDD2021\_Joined.take(10)

**Screenshot:**

**![Graphical user interface, text, application, email

Description automatically generated]()**

**Code Segments Used in DF Implementation:**

In dataframe I loaded the pharma file and read it as csv file by defining the schema and delimiter.Then compared and joined the clinicaltrial’s sponsor column and pharma file’s parent company column. By using the where statement I checked the values where sponsor is not matched with parents company column. Then I grouped by sponsor and order it to extract the desired results.

pharma=('dbfs:/FileStore/tables/pharma.csv')

DFpharma=spark.read.format('csv') \

.options(header='true', inferSchema='true') \

.options(delimiter=',') \

.load(pharma)

DFpharma.display()

DFpharma\_Join1=DF2021.Sponsor == DFpharma.Parent\_Company

DFpharma\_Join= DF2021.join(DFpharma, on = DFpharma\_Join1, how='left')

DFpharma\_Join.display()

DFpharma\_Sponsor1= DFpharma\_Join.select('Sponsor', 'Parent\_Company').where('Parent\_Company is null')

DFpharma\_Sponsor=DFpharma\_Sponsor1.groupby(['Sponsor']).count().orderBy(['count'], ascending=False)

DFpharma\_Sponsor.display()

**Screenshot:**

**![Graphical user interface, text, application, email

Description automatically generated]()**

**Code Segments Used in HQL Implementation:**

In HQl implementation I have created table using the pharma file. Then I created to view to check the different values by using the leftouterjoin function which were not in pharma file but in clinicaltrial file’s sponsor column. By grouping the sponsor and making the count I got the desired output as per client’s requirement.

CREATE External 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 )

USING CSV

OPTIONS (path "dbfs:/FileStore/tables/pharma.csv",

delimiter ",",

header "true" ) ;

CREATE OR REPLACE VIEW joinPharmaCompany

AS

select c.\*,p.Parent\_Company

from clinicaltrial\_2021 as c LEFT OUTER JOIN pharma p

ON (c.Sponsor=p.Parent\_Company)

select

Sponsor,count(Sponsor)as count

from joinPharmaCompany

where

Parent\_Company is null group by sponsor order by count(Sponsor) desc Limit 10

**Screenshot:**

![Table

Description automatically generated]()

1. Problem Statement VI :

Plot number of completed studies each month in a given year – for the submission dataset, the year is 2021. You need to include your visualization as well as a table of all the values you have plotted for each month.

**Answer**:

(dataset used to produce answer: clinicaltrial\_2021.csv, Mesh.csv and Pharma.csv)

[('Jan', 1131),

('Feb', 934),

('Mar', 1227),

('Apr', 967),

('May', 984),

('Jun', 1094),

('Jul', 819),

('Aug', 700),

('Sep', 528),

('Oct', 187)]

**Code Segments Used in RDD Implementation:**

In this implementation Pharma file has all the pharmaceutical companies list but in clinical trial file there are some sponsors which are not pharmaceutical company. So I used the sponsor column from clinicaltrial file and skipped the nulled fields by filtering the RDD. I used leftouterjoin function to pick the unmatched sponsor companies from clinicaltrial file which were not in Pharma file. Then I again filtered the resulted RDD to skipped the null fields. Reducebykey function and count function used to find the frequency of each sponsor for trial study.

RDD1 = RDD2021\_Split.first()

RDD2021\_Count = RDD2021\_Split.filter(lambda Header: Header!=RDD1)

RDD2021\_Study\_Completion=RDD2021\_Count.map(lambda x: (x[4].split(),x[2]))\

.filter(lambda x: len(x[0])!=0)\

.filter(lambda y:(y[1]=='Completed'))\

.filter(lambda z: (z[0][1]==Year))\

.map(lambda a: (a[0][0],1))\

.reduceByKey(lambda v1, v2: v1+ v2)

RDD2021\_Study\_Completion.take(12)

import calendar

Sorting = {a:z for z,a in enumerate(calendar.month\_abbr[1:],1)}

RDD2021\_Study\_Completion\_Sorting=RDD2021\_Study\_Completion.sortBy(lambda g: Sorting.get(g[0]))

RDD2021\_Study\_Completion\_Sorting.collect()

**Screenshot:**

![Graphical user interface, text, application, email

Description automatically generated]()

**Code Segments Used in DF Implementation:**

DF2021\_Study1=DF2021.withColumn('Month', split(DF2021['Completion'], ' ').getItem(0))\

.withColumn('year', split(DF2021['Completion'], ' ').getItem(1))

DF2021\_Study1.show()

DF2021\_Study\_Filter=DF2021\_Study1.filter((DF2021\_Study1.Status=='Completed') &

(DF2021\_Study1.year==Year)).groupby('Month').count()

DF2021\_Study\_Filter.show()

from pyspark.sql.functions import col,when

DF2021\_Study\_Month\_Arrangment1= DF2021\_Study\_Filter.orderBy(when(col('Month') == 'Jan',1).when(col('Month') == 'Feb',2).when(col('Month') == 'Mar',3).when(col('Month') == 'Apr',4).when(col('Month') == 'May',5).when(col('Month') == 'Jun',6).when(col('Month') == 'Jul',7).when(col('Month') == 'Aug',8).when(col('Month') == 'Sep',9).when(col('Month') == 'Oct',10).when(col('Month') == 'Nov',11).when(col('Month') == 'Dec',12))

DF2021\_Study\_Month\_Arrangment=DF2021\_Study\_Month\_Arrangment1.withColumnRenamed('count', 'Year '+Year)

DF2021\_Study\_Month\_Arrangment.show()

**Screenshot:**

Graphical user interface, text, application, Word

Description automatically generated

**Code Segments Used in HQL Implementation:**

CREATE OR REPLACE VIEW month\_year

AS

select \*, split(completion,' ')[0] as Month, split(completion,' ')[1] as Year from clinicaltrial\_2021 where completion is not null

CREATE OR REPLACE VIEW counting\_view

AS

select month,count(month)as count from month\_year where Status=='Completed' and Year=='$Year' group by month

CREATE OR REPLACE VIEW final\_month\_view AS

SELECT month, count,

CASE

WHEN month ='Jan' THEN 1 WHEN month ='Feb' THEN 2 WHEN month ='Mar' THEN 3 WHEN month ='Apr' THEN 4 WHEN month ='May' THEN 5 WHEN month ='Jun' THEN 6

WHEN month ='Jul' THEN 7 WHEN month ='Aug' THEN 8 WHEN month ='Sep' THEN 9 WHEN month ='Oct' THEN 10 WHEN month ='Nov' THEN 11 WHEN month ='Dec' THEN 12

END as month\_sorting

FROM counting\_view ORDER by month\_sorting asc;

select month,count from final\_month\_view

**Screenshot:**

![Table

Description automatically generated]()

# Instructions to Run:

Use web based cloud platform Databricks for running the codes. This has the python and sql notebook interface. Upload/import the .ipynb file into databricks workspace and execute it.

Remember that the file location of dataset must be uploaded in FileStore/tables. If dataset is uploaded in any other location then location must be changed inside the code pertaining the dataset location. This is same for implementation in RDD and Dataframe. But for .sql file the dataset must be uploaded unzipped containing file name clinicaltrial\_2021.csv .