================

\*\* Question 1 \*\*

================

# INSTRUCTIONS:

Use the 'hr\_records' data to generate a report of all employees that

joined the company in 2015.

# DATA DESCRIPTION:

Employee records are in a Parquet file format, compressed using gzip, and

stored in the HDFS directory:

/verulam\_blue/tests/data/parquet/hr\_records/

# OUTPUT REQUIREMENTS:

• Place the result data in the hdfs directory:

/verulam\_blue/test1/problem1/solution/

• Output should have the following 4 columns:

'first\_name', 'last\_name', the month of joining 'month\_of\_joining'

and the year of joining 'year\_of\_joining' .

• Output should be stored in a single file using a JSON file format

• Compress the output file using gzip compression.

• Order results by month of joining, from January through to December.

================

\*\* Question 2 \*\*

================

# INSTRUCTIONS:

Find GP practices that have less than 300 patients.

# DATA DESCRIPTION:

GP practice details are stored as tab-delimited text files,

compressed using gzip, and stored in the HDFS directory:

/verulam\_blue/tests/data/tab\_text/gp\_address

The second column is 'practice\_code'

The third column is 'surgery\_name'

GP practice demographics records are stored as tab-delimited text files,

compressed using gzip, and stored in the HDFS directory:

/verulam\_blue/tests/data/tab\_text/gp\_demographics

The fist column is 'practice\_code'

The third column is 'nbr\_of\_patients'

Consider the 'practice\_code' field a foreign key in 'gp\_demographics' data

which references the primary key 'practice\_code' from the 'gp\_address' data.

# OUTPUT REQUIREMENTS:

• Place the result data in the hdfs directory:

/verulam\_blue/test1/problem2/solution/

• Output should have the following columns:

'practice\_code', 'surgery\_name' and 'nbr\_of\_patients'.

• Save results as gzip compressed JSON files.

# SAMPLE RESULTS:

+-------------------+-------------------------------------+---------------------+

|practice\_code |surgery\_name |nbr\_of\_patients|

+-------------------+-------------------------------------+---------------------+

..... ..... ......

|A82619 |NELSON STREET SURGERY |149 |

|A83068 |JUPITER HOUSE SURGERY |164 |

|B81671 |128 CHELMSFORD AVENUE |139 |

================

\*\* Question 3 \*\*

================

# INSTRUCTIONS:

Convert the EMR data to a single pipe delimited text file ready for exporting.

# DATA DESCRIPTION:

Emergency department visits records are stored as Parquet files, compressed

using gzip, and stored in the HDFS directory:

/verulam\_blue/tests/data/parquet/emr\_data

# OUTPUT REQUIREMENTS:

• Place the result data in the hdfs directory:

/verulam\_blue/test1/problem3/solution/

• Use a text format with a pipe character “|” as the columnar delimiter.

• Keep all of the original 6 columns.

• Compress the output file using a gzip compression.

# SAMPLE RESULTS:

|2020-07-25|2020-07-14|11420|40|1|1|

|2020-07-25|2020-07-02|10038|27|0|0|

|2020-07-24|2020-05-31|10019|33|0|0|

================

\*\* Question 4 \*\*

================

# INSTRUCTIONS:

The HR department have asked you to generate a report showing employees'

user names used to login to systems.

# DATA DESCRIPTION:

Employee records are stored as tab-delimited text files,

compressed using gzip, and stored in the HDFS directory:

/verulam\_blue/tests/data/tab\_text/hr

There are a total of 18 columns.

The 4th column is the employee's last name.

The 6th column is the employee's date of birth.

The last column is the employee's user name.

# OUTPUT REQUIREMENTS:

• Place the result data in the hdfs directory:

/verulam\_blue/test1/problem4/solution/

• Use a text format with a pipe character “|” as the columnar delimiter.

• The report should contain the employee's last name, date of birth and

their user name.

# SAMPLE RESULTS:

|Suhr |1992-09-04|hjsuhr |

|Joy |1989-12-24|aqjoy |

|Moores|1980-09-23|sjmoores|

================

\*\* Question 5 \*\*

================

# INSTRUCTIONS:

Health administrators have asked for a report showing the top 10

prescribers of the medication 'Diazepam'.

Generate a report showing the GP practices & the number of 'Diazepam'

prescriptions they made.

# DATA DESCRIPTION:

• GP prescribing records are stored in the metastore table 'gp\_rx' in the

database 'gp\_db'.

• The column 'bnf\_name' describes the type of medication prescribed,

how it is to be taken and its dosage.

• The column 'items' represents the total number of items prescribed.

• GP surgery details are stored in the metastore table 'gp\_address' in the

database 'gp\_db'.

• Consider the 'practice\_code' field a foreign key in 'gp\_address' table

which references the primary key 'practice\_code' from the 'gp\_rx' table.

# OUTPUT REQUIREMENTS:

• Place the result data in the hdfs directory:

/verulam\_blue/test1/problem5/solution/

• Output should only consist of the columns 'surgery\_name', the number

of medications prescribed and the rank.

• Sort results by descending order of number of prescriptions made.

• Save results as a tab delimited text file compressed using bzip2

compression.

# SAMPLE RESULTS:

|DRUG AND ALCOHOL ACTION TEAM 991 1 |

|MAPLE ACCESS PARTNERSHIP LLP 865 2 |

|MIDLANDS MEDICAL PARTNERSHIP 742 3 |

================

\*\* Question 6 \*\*

================

# INSTRUCTIONS:

Taxi spotters only ever need to access a few of the fields in the green

taxi data. You have been commissioned to set up smaller dataset to be

converted to a high-performance file format & stored in a metastore table.

# DATA DESCRIPTION:

The green taxi records are stored as gzip compressed JSON files

in the HDFS directory:

/verulam\_blue/tests/data/json/taxi\_data

# OUTPUT REQUIREMENTS:

• Place the result data in the hdfs directory:

/verulam\_blue/test1/problem6/solution/

• The data should be stored as snappy compressed parquet files.

• The results should have three columns only 'vendor\_id',

'pickup\_datetime' and 'dropoff\_datetime'

• Order results in descending order of 'vendor\_id'.

• The schema of the data should be as below:

|-- Column: Type

|-- vendor\_id: string

|-- dropoff\_datetime: date

|-- dropoff\_datetime: date

# SAMPLE RESULTS:

+--------------+------------------------+-----------------------+

|vendor\_id |pickup\_datetime |dropoff\_datetime|

+--------------+------------------------+-----------------------+

|2 |2018-01-01 |2018-01-01 |

|2 |2018-05-01 |2018-05-01 |

|2 |2018-01-01 |2018-01-01 |

================

\*\* Question 7 \*\*

================

# INSTRUCTIONS:

The data security department are looking to create aliases for employees

for login purposes.

They have tasked you with generating a custom alias.

This alias should be made up as follows:

First 3 numbers from employee's SSN ('ssn' column).

Employee's zip code ('zip' column).

First 3 numbers from employee's phone number ('phone\_nbr' column).

All separated using a colon (':').

Sample alias: 275:72453:479

# DATA DESCRIPTION:

Employee records are stored in the metastore table hr\_records

in the database hr\_db.

# OUTPUT REQUIREMENTS:

• Place the result data in the hdfs directory:

/verulam\_blue/test1/problem7/solution/

• The data should be stored in snappy compressed parquet files.

• Output should be the following 3 columns:

employee's first name, last name and alias.

# SAMPLE RESULTS:

+--------------+--------------+-------------------+

|first\_name|last\_name|alias |

+--------------+--------------+-------------------+

|Hermila |Suhr |275:72453:479|

|Antonio |Joy |646:30455:229|

|Sebastian |Moores |499:13608:212|

================

\*\* Question 8 \*\*

================

# INSTRUCTIONS:

Use the EMR data to find the total number of emergency department visits,

for each of the 31 days in May, that were due to influenza-like illness

and/or pneumonia.

# DATA DESCRIPTION:

Emergency department visits records are stored as Parquet files,

compressed using gzip, and stored in the HDFS directory:

/verulam\_blue/tests/data/parquet/emr\_data

• The 'date\_of\_visit' column represents the date of hospital visit.

• The 'column ili\_pne\_visits' represents the count of influenza-like illness

and/or pneumonia visits.

# OUTPUT REQUIREMENTS:

• Place the result data in the hdfs directory:

/verulam\_blue/test1/problem8/solution/

• Use a text format with a pipe character “|” as the columnar delimiter.

• Compress the output file using a gzip compression.

• Order results by the date of hospital visit.

# SAMPLE RESULTS:

|2020-05-01|28716|

|2020-05-02|22765|

|2020-05-03|21788|

================

\*\* Question 9 \*\*

================

# INSTRUCTIONS:

As part of an exercise to check stock levels, health administrators

have asked to see a report showing the number of prescriptions of the

medication 'Diazepam', broken down by how the medication is to be taken

and its dosage.

# DATA DESCRIPTION:

GP prescribing records are stored in the metastore table 'gp\_rx' in

the database 'gp\_db'.

• The column 'bnf\_name' describes the type of medication prescribed,

how it is to be taken and its dosage.

• The column 'items' describes the number of medications prescribed.

# OUTPUT REQUIREMENTS:

• Place the result data in the hdfs directory:

/verulam\_blue/test1/problem9/solution/

• Output should only consist of the columns 'bnf\_name' and

'nbr\_prescribed' (the number of medications prescribed).

• Sort results by the most prescribed.

• Save results as a single gzip compressed Parquet file.

# SAMPLE RESULTS:

+-------------------------------------------------+---------------------+

|bnf\_name |nbr\_prescribed |

+-------------------------------------------------+---------------------+

|Diazepam\_Tab 2mg |237539 |

|Diazepam\_Tab 5mg |185622 |

|Diazepam\_Tab 10mg |29520 |

|Diazepam\_Oral Soln 2mg/5ml S/F |3038 |

================

\*\* Question 10 \*\*

================

# INSTRUCTIONS:

Management are looking to reward long serving employees.

You have been asked to generate a report containing records of employees

with more than 10 years' service working for the company.

• Assume a current date of 1st December 2020 (2020-12-01).

• Assume 365 days in a year.

# DATA DESCRIPTION:

Employee records are stored in the metastore table hr\_records in the database hr\_db.

# OUTPUT REQUIREMENTS:

• Create a database called 'test1 \_solutions' (if one doesn't already exist)

and save results in the metastore table named 'q10\_soln' in the

'test1\_solutions' database.

• Store the table data in the custom table path:

/verulam\_blue/test1/problem10/solution/

• The data should be stored as snappy compressed avro files.