Student name: CHRISTOS VLASSIS, ID: f2822204

Task 1

Among Hive, Impala, and Drill, which is the one the implements more precisely the concept of data virtualization? Elaborate.

As data virtualization we mean a different way to data management that makes ETLs procedures obsolete. With ETL and ELT we have to move the data from one or many data source to another but with data virtualization we can retrieve and manipulate the data more directly without actually moving them. We can can Imagive data virtualization such as taking a View of a table in SQL. We don't really move the data we just take a picture.

If we compare Hive, Impall and Drill for the data virtualization concept, Drill is better suited for this task. Apache Drill can interact with many different data sources such as HDFS, MongoDB, Amazon, csv and many more. So, the schema of the database models that it can query has a great variety. Also, there is no need for data transformations since it does not require this pre defined schema.

On the other hand we have Impala that is a Massive Parallel Processing SQL query engine for huge datasets. It an process HDFS, Hbas, kudu, Amazon data but it does not have the same schema freedom as drill. So, it can not handle JSON or XML for example.

Finally we have Hive that is similar to Impala. It can query only inside the Hadoop ecosystem such as the HDFS and Hbase.

Overall, Apache Drill is better suited for a data virtualization engine since it can connect with many types of Storage-engines.

Task 2

You started working for a large bookstore company. Your client has a large data center containing data in various formats. More specifically, all client data (e.g., personal information, orders) are stored in a Mongo DB database, e-books are stored on HDFS, and social media metadata (likes, ratings, reviews) are stored in a Hive database. They would like to simplify the queries used by various User Interface elements. What would you suggest for their case? Elaborate.

In this case we can see that the user uses two main systems for storing data. The first is the MongDB database for orders and personal informations and he also uses the Hadoop ecosystem for likes, ratings and reviews of the products. I would not suggest a ETL since the data are many and will have a big cost impact in the entreprise. I would suggest them to use Apache Drill to be able to query both Mongo DB and HDFS since it will come with less cost. After the Apache Drill is installed and the queries runned I would also use Tableau or some reporting software for the business user to have a great picture of all of their information.

Task 3

3a) Create the Impala database & the required tables.

```
CREATE TABLE Student (
sid INT,
name VARCHAR (50) ) ;

CREATE TABLE Attended (
sid INT,
cid INT,
ac_year CHAR (9),
grade NUMERIC (1),
FOREIGN KEY (sid) REFERENCES Student(sid)
FOREIGN KEY (cid) REFERENCES Course(cid) ) ;

CREATE TABLE Course (
cid INT,
title VARCHAR (50),
description TEXT ) ;
```

3b) Give an example command that inserts an entry to the Student table (use your own details for that entry).

```
INSERT INTO Student (sid, name)
VALUES (f2822204, Christos Vlassis) ;
```

3c) Write a statement that retrieves all the names of the students that have attended the course having title "Artificial Intelligence" during the academic year "2021-2022".

```
SELECT
        st.name
FROM Student st
JOIN Attended at ON at.sid = st.sid
JOIN Course co ON co.cid = at.cid
WHERE co.title = 'Artificial Intelligence'
AND at.ac_year = '2021-2022' ;
```

3d) Write a statement that retrieves the titles and the average grades of all the courses for which the average grade of the students that attended them is lower than 6.

```
SELECT
        co.titles,
        AVG (at.grade)
FROM Attended at
JOIN Course co ON co.cid = at.cid
GROUP BY co.title
HAVING AVG (at.grade) < 6 ;
```

Task 4

A particular query in the previous Impala database is too slow. Describe what you are going to do to investigate what is going wrong and what can be done to improve efficiency. Provide any commands that you are going to run.

I would run the following command to take all the relative statistics about the query

PROFILE <query_id>;

Here I will expect to see big difference between the start and end time the query.

Also, I would check the resource usage like the CPU and memory, because, maybe, the system is running out of memory. Also, the data skewness maybe creates some problems. Meaning, that some nodes are doing a lot more work than others and this makes the queries to run slow.

To provide efficiency we could update the statistics of the table. If the problem is from the lack of resources we would have to use more resources to impala. Finally we could optimize the schema, meaning partition the tables based on the columns that are more frequently used in the queries. But, before trying to make the queries more efficient we have to understand the problem because some problems need different solutions.