ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Advanced Topics in Data Engineering

Assignment 1

Name : Ilias Dimos

AM : f2822102

Professor : Athanasios Vergoulis

Task 1

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

Answer

Data virtualization is to retrieve and manipulate data without requiring knowledge about the format and the physical location of the data.

The best program that implements the concept of the data virtualization more precisely is the Apache Drill.

Drill does not require schema or type specification for the data to start the query execution process. Also, it is capable to handle Self-describing data formats like Parquet, JSON, AVRO, NoSQL database formats. It provides a flexible hierarchical columnar data model that can represent complex, highly dynamic and evolving data models.

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.

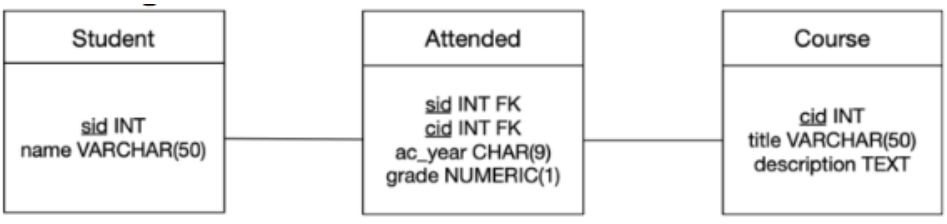
Answer

Because of the various data formats, they have they are forced to run the queries in different programming environments. My first suggestion would be to educate the working staff on the Apache Drill . In this way the company staff will be able to handle every type of data via one program platform making them saving time and simplify their working routing as they will not run different programs to answer business questions.

Also, the ecosystem of the Apache Drill is the ANSI SQL meaning that the users do not need to learn a new programming language but the can use it by programming in a simple SQL.

Task 3

Another client of yours has an Impala database for their needs. They want to have a new database with the following schema:



Please provide detailed answers to the following:

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

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

Answer

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

Database

```
create database if not exists assignment_impala;
```

Tables

```
5
6  CREATE TABLE `Student` (
7  `sid` int NOT NULL,
8  `name` varchar(50) NOT NULL,
9   PRIMARY KEY (`sid`)
10 );
11
12 CREATE TABLE `course` (
13  `cid` int NOT NULL,
14  `title` varchar(50) NOT NULL,
15  `description` TEXT NOT NULL,
16  PRIMARY KEY (`cid`)
17  );
18
19 CREATE TABLE `attended` (
20  `ac_year` char(9) NOT NULL,
21  `grade` NUMERIC(1) NOT NULL,
22  `sid` int NOT NULL,
23  `cid` int NOT NULL,
24  key `sid_index` (`sid`),
25  key `cid_index` (`cid`),
26  constraint `cid` foreign key (`cid`) references `course` (`cid`),
27  constraint `sid` foreign key  (`sid`) references `student` (`sid`)
28  );
29
```

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

```
INSERT INTO student
 VALUES
(102,'Ilias');
```

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  name
from student
join attended
on student.sid = attended.sid
join course
on attended.cid = course.cid
where course.title ='Artificial Intelligence' and attended.ac_year in (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 title, round(avg(grade),1) as average_grade
from course
join attended
on attended.cid = course.cid
group by title
having average_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.

Answer

In order to investigate what is going on with the query we can use the **profile** command which displays information about the last query and find what is going wrong. We can also use the **explain** command that returns the execution plan for a given query.

In order to fix any problems that increase the run time of the query we can do to the following improvements :

1. **Overview table and column statistics** : Gather table and column statistics, using the COMPUTE STATS statement, helps Impala automatically optimize the performance for join queries, without requiring changes to SQL query statements.
2. **Partitioning for Tables** : This technique physically divides the data based on the different values in frequently queried columns, allowing queries to skip reading a large percentage of the data in a table.

Commands

Assuming that the query that is slow is from the table “attended” we will display the commands that will be used to find what is going wrong and solve the problem.

1. ***Profile command***

When we want to investigate the performance of a query, we have to run it first and then we write the command “profile;” so as to calculate the query’s performance.

2. ***Explain command***

We can use the command by using the following syntax :

- explain select * from attended

3. ***Overview table and column statistics***

The Impala COMPUTE STATS statement automatically gathers statistics for all columns, because it reads through the entire table relatively quickly and can efficiently compute the values for all the columns. It can be used by writing the command :

- compute stats attended;

In order to see the stats, we use :

- show table stats attended;

4. ***Partitioning for Tables***

- create table attended (ac_year **CHAR**, grade **NUMERIC ,** sid **INT,**cid **INT**) partitioned by (year **char**);
- insert into attended partition (year =2020) values (2020,9,2822102,3);