SAMPLE HIVE PROJECT

Omar M.Yehia

19/10/2022

SAMPLE HIVE PROJECT By Omar M.yehia

TABLE OF CONTENTS

PROJECT DESCRIPTION

The project aims to display Beverages-to-Multiple Branches of one Coffee Shop (many-to-many relation) using Hive. In other words, each Beverage might be available on many Branches, and each Branch of the Coffee shop might distribute many Beverages.

Assuming each branch send their sales report as a csv file. The project aims to stage them to HDFS and further analysis to be performed using Hive for the given problem statement below.

The description of the data is as below

- Beverages Name does not have spaces.
- Coffee shop Branches have mentioned as Branch1, Branch2 and etc.
- Beverages can be ordered many times, with different counts
- A Beverage and Branch combination might appear multiple times
- Beverages could be available on multiple Branches
- The output should have no commas or punctuation, only 1 space between the Beverages and Count of Consumed people.

INPUT FILES

Dataset for the project.zip

PROBLEM STATEMENT

- What is the total number of consumers for Branch1?
- What is the number of consumers for the Branch2?
- What is the most consumed beverage on Branch1?
- What are the beverages available on Branch10, Branch8, and Branch1?

ENVIRONMENT SETUP

- Software Specification
  - VM Used – cloudera
  - Hadoop version
  - Hive version
  - WinSCP

SAMPLE HIVE PROJECT By Omar M.yehia

PROJECT MODULES

1. Placing the given dataset in HDFS
   1.1. Create directory in HDFS
   1.2. Placing the input files in the HDFS directory
2. Implementation in HIVE
   2.1. Creating HIVE DB
   2.2. Creating & Loading the HIVE tables with the given datasets
3. Problem Scenario 1 - What is the total number of consumers for Branch1?
   3.1. Type 1 - Creating single physical table with sub queries
   3.2. Type 2 - Creating multiple physical tables
   3.3. Solution
4. Problem Scenario 2 – What is the number of consumers for the Branch2?
   4.1. Type 1 - Sub queries selection without table creation
   4.2. Type 2 - Creating multiple physical tables
   4.3. Solution
5. Problem Scenario 3 - What is the most consumed beverage on Branch1?
   5.1. Explanation
   5.2. Solution
6. Problem Scenario 4 - What are the beverages available on Branch10, Branch8, and Branch1?
   6.1. Explanation
   6.2. Solution

Placing the given dataset in HDFS

Create directory in HDFS

Step 1: Before creating directories in HDFS, ensure all the daemons in hadoop are started. The below code is for creating directory called "hiveproject" as follows,

```
$ hadoop fs -mkdir /user/hive/ hiveproject
```

Placing the input files in the HDFS directory

Step 1: Copying all the given dataset files from local to HDFS directory in a separate directory. The code as follows,

```
$ hadoop fs -copyFromLocal  /home/cloudera/hive/Bev_BranchA.txt  /user/hive/hiveproject/
$ hadoop fs -copyFromLocal  /home/cloudera/hive/Bev_BranchB.txt  /user/hive/hiveproject/
$ hadoop fs -copyFromLocal  /home/cloudera/hive/Bev_BranchC.txt  /user/hive/hiveproject/
$ hadoop fs -copyFromLocal  /home/cloudera/hive/Bev_ConscountA.txt  /user/hive/hiveproject/
$ hadoop fs -copyFromLocal  /home/cloudera/hive/Bev_ConscountB.txt  /user/hive/hiveproject/
$ hadoop fs -copyFromLocal  /home/cloudera/hive/Bev_ConscountC.txt /user/hive/hiveproject/
```

Step 2: After the Step 1, check whether the files got placed in the HDFS in browser.

Implementation in HIVE

Creating HIVE DB

Step 1: Create a database in the name "hadoophiveproject" in HIVE. The code as follows,

```
Hive> create database  hadoophiveproject ;
```

Creating & Loading the HIVE tables with the given datasets

Step 1: Create separate raw tables for the Beverages-Counsumercount different datasets each in "hadoophiveproject" database. The given file (Bev_Conscount *.txt) consist of <Beverages, Consumercount> (A Beverage and the number of consumers).

Example Bev_Conscount***.txt:**

```
Special_Lite, 21
Triple_Espresso, 38
Mild_LATTE, 73
LARGE_Coffee, 144
Cold_cappuccino, 287
SMALL_cappuccino, 574
 ...
```

The codes for creating tables are as follows,

```
Hive> use hadoophiveproject;

Hive> create table if not exists BevcountA (beverage string,count int) row format delimited fields terminated by ",";

Hive> create table if not exists BevcountB(beverage string,count int) row format delimited fields terminated by ",";

Hive> create table if not exists BevcountC (beverage string,count int) row format delimited fields terminated by ",";
```

Step 2: Loading the Beverage -Number of consumers' raw tables from the given text files individually. The code as follows,

```
Hive> load data inpath "/user/hive/hiveproject/Bev_ConscountA.txt" into table BevcountA;
Hive> load data inpath "/user/hive/hiveproject/Bev_ConscountB.txt" into table BevcountB;
Hive> load data inpath "/user/hive/hiveproject/Bev_ConscountC.txt" into table BevcountC;
```

Schema definition for the created tables as follows,

Beverage-Number of consumers Relationship

Tables	Fields	Input Type
BevcountA	beverage	string
	count	int
BevcountB	beverage	string
	count	int
BevcountC	Beverage	string
	count	int

Step 3: Create separate raw tables for the Beverages-Branches different datasets each in “hadoophiveproj” database. The given file (Bev_Branch*.txt) consist of <Beverages, Branches> (A Beverages and the Branches it was on).

Example Bev_Branch***.txt:**

```
Special_Lite, Branch6
MED_LATTE, Branch2
Triple_cappuccino, Branch9
ICY_LATTE, Branch5
SMALL_Espresso, Branch1
Double_cappuccino, Branch6
LARGE_Espresso, Branch2
Mild_Espresso, Branch9
…
```

The codes for creating tables are as follows,

```
Hive> create table if not exists BevbranchA(beverage string,branch string) row format delimited fields terminated by ",";

Hive> create table if not exists BevbranchB(beverage string, branch string) row format delimited fields terminated by ",";

Hive> create table if not exists BevbranchC(beverage string, branch string) row format delimited fields terminated by ",";
```

Step 4: Loading the Beverage type-Branch raw tables from the given text files individually. The code as follows,

```
hive> load data inpath "/user/hive/hiveproject/Bev_BranchA.txt" into table BevbranchA
hive> load data inpath "/user/hive/ hiveproject /Bev_BranchB.txt" into table BevbranchB
hive> load data inpath "/user/hive/ hiveproject /Bev_BranchC.txt" into table BevbranchC
```

Schema definition for the created tables as follows,

Beverage Type-Branch Relationship

Tables	Fields	Input Type
BevbranchA	Beverage	String
	Branch	String
BevbranchB	Beverage	String
	Branch	string
BevbranchC	Beverage	string

	Branch	string