SAMPLE HIVE PROJECT

# TABLE OF CONTENTS

PROJECT DESCRIPTION 3

[INPUT FILES 3](#_TOC_250012)

[PROBLEM STATEMENT 3](#_TOC_250011)

[ENVIRONMENT SETUP 3](#_TOC_250010)

[PROJECT MODULES 4](#_TOC_250009)

[Placing the given dataset in HDFS 5](#_TOC_250008)

Create directory in HDFS 5

Placing the input files in the HDFS directory 5

[Implementation in HIVE 5](#_TOC_250007)

Creating HIVE DB 5

Creating & Loading the HIVE tables with the given datasets 5

[Problem Scenario 1 - What is the total number of consumers for Branch1? 8](#_TOC_250006)

Type 1: Creating single physical table with sub queries 8

Type 2: Creating multiple physical tables 9

[Solution 10](#_TOC_250005)

[Problem Scenario 2 – What is the number of consumers for the Branch2? 10](#_TOC_250004)

Type 1: Sub queries selection without table creation 10

Type 2: Creating multiple physical tables 11

[Solution 12](#_TOC_250003)

[Problem Scenario 3 - What is the most consumed beverage on Branch1? 12](#_TOC_250002)

Explanation 12

[Solution 12](#_TOC_250001)

[Problem Scenario 4 - What are the beverages available on Branch10, Branch8, and Branch1? 12](#_TOC_250000)

Explanation 12

**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

## 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 – centos-6.3-i386-server
  + Hadoop version – Hadoop 2.0.0 –cdh4.7.1
  + Hive version – Hive version 0.8.0
  + WinSCP – version 5.9.4

## PROJECT MODULES

1. Placing the given dataset in HDFS
   1. Create directory in HDFS
   2. Placing the input files in the HDFS directory
2. Implementation in HIVE
   1. Creating HIVE DB
   2. Creating & Loading the HIVE tables with the given datasets
3. Problem Scenario 1 - What is the total number of consumers for Branch1?
   1. Type 1 - Creating single physical table with sub queries
   2. Type 2 - Creating multiple physical tables
   3. Solution
4. Problem Scenario 2 – What is the number of consumers for the Branch2?
   1. Type 1 - Sub queries selection without table creation
   2. Type 2 - Creating multiple physical tables
   3. Solution
5. Problem Scenario 3 - What is the most consumed beverage on Branch1?
   1. Explanation
   2. Solution
6. Problem Scenario 4 - What are the beverages available on Branch10, Branch8, and Branch1?
   1. Explanation
   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 “projinput” as follows,

$ hadoop fs -mkdir /user/hive/projinput

*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 Bev\_BranchA.txt /user/hive/projinput/Bev\_BranchA.txt

$ hadoop fs -copyFromLocal Bev\_BranchB.txt /user/hive/projinput/Bev\_BranchB.txt

$ hadoop fs -copyFromLocal Bev\_BranchC.txt /user/hive/projinput/Bev\_BranchC.txt

$ hadoop fs -copyFromLocal Bev\_ConscountA.txt /user/hive/projinput/Bev\_ConscountA.txt

$ hadoop fs -copyFromLocal Bev\_ConscountB.txt /user/hive/projinput/Bev\_ConscountB.txt

$ hadoop fs -copyFromLocal Bev\_ConscountC.txt /user/hive/projinput/Bev\_ConscountC.txt

**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 “hadoophiveproj” in HIVE. The code as follows,

Hive> create database if not exits hadoophiveproj with dbproperties(‘creator’=’Haripriya’,’date’=’15/10/2017’);

*Creating & Loading the HIVE tables with the given datasets*

**Step 1:** Create separate raw tables for the Beverages-Counsumercount different datasets each in “hadoophiveproj” 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 hadoophiveproj;

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/projinput/Bev\_ConscountA.txt’ into table BevcountA; Hive> load data inpath ‘/user/hive/projinput/Bev\_ConscountB.txt’ into table BevcountB; Hive> load data inpath ‘/user/hive/projinput/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/projinput/Bev\_BranchA.txt’ into table BevbranchA; hive> load data inpath ‘/user/hive/projinput/Bev\_BranchB.txt’ into table BevbranchB; hive> load data inpath ‘/user/hive/projinput/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 |

## Problem Scenario 1 - What is the total number of consumers for Branch1?

*Type 1: Creating single physical table with sub queries*

### Explanation

Step 1: Creating single table “Branch1Constotcount” with sub queries in which the Branch1 consumers alone selected and counted. The code as follows,

Hive> create table if not exists Branch1Constotcount as select beverage,totalcount from (select BevcountA.beverage,sum(BevcountA.count) totalcount from (select beverage,branch from ( select \* from BevbranchA where branch='BRANCH1' union all select

* from BevbranchB where branch='BRANCH1' union all select \* from BevbranchC where branch='BRANCH1')unionResult )a join BevcountA on(a.beverage=BevcountA.beverage) group by BevcountA.beverage

union all

select BevcountB.beverage,sum(BevcountB.count) totalcount from ( select beverage,branch from ( select \* from BevbranchA where branch='BRANCH1' union all select

* from BevbranchB where branch='BRANCH1' union all select \* from BevbranchC where branch='BRANCH1')unionResult )b join BevcountB on(b.beverage=BevcountB.beverage) group by BevcountB.beverage

union all

select BevcountC.beverage,sum(BevcountC.count) totalcount from ( select beverage,branch from ( select \* from BevbranchA where branch='BRANCH1' union all select

* from BevbranchB where branch='BRANCH1' union all select \* from BevbranchC where branch='BRANCH1')unionResult )c join BevcountC on(c.beverage=BevcountC.beverage) group by BevcountC.beverage

)unionResult ;

**Step 2:** The created table have the following table structure.

|  |  |  |
| --- | --- | --- |
| Table | Fields | Input Type |
| Branch1Constotcount | beverage | string |
| totalcount | int |

**Step 3:** Finally the summation of the table gives the end result. Code as follows, Hive> select sum(totalcount) from Branch1Constotcount;

*Type 2: Creating multiple physical tables*

Explanation

**Step 1:** Creating physical table “BRANCH1Branch” from the previous tables where the branch name is equal to BRANCH1. The code as follows,

Hive > create table if not exists BRANCH1Branch as select \* from BevbranchA where branch

= 'BRANCH1';

Hive > insert into table BRANCH1Branch select \* from BevbranchB where branch = 'BRANCH1';

Hive > insert into table BRANCH1Branch select \* from BevbranchC where branch = 'BRANCH1';

**Step 2:** Creating another table “BRANCH1branchcount” which merge tables which has the number of consumers with Beverages. The code as follows,

Hive> create table if not exists BRANCH1branchcount(beverage string, count int);

Hive> insert into table BRANCH1branchcount select BevcountA.beverage,sum(BevcountA.count) from BRANCH1Branch join BevcountA on(BRANCH1Branch.beverage = BevcountA.beverage ) group by BevcountA.beverage ;

Hive> insert into table BRANCH1branchcount select BevcountB.beverage,sum(BevcountB.count) from BRANCH1Branch join BevcountB on(BRANCH1Branch.beverage = BevcountB.beverage ) group by BevcountB.beverage ;

Hive> insert into table BRANCH1branchcount select BevcountC.beverage,sum(BevcountC.count) from BRANCH1Branch join BevcountC on(BRANCH1Branch.beverage = BevcountC.beverage ) group by BevcountC.beverage ;

**Step 3:** Finally summation of the fields from total number of consumers in the newly created table “BRANCH1branchcount”. The code as follows,

Hive> select sum(count) from BRANCH1branchcount;

**Step 4:** Created tables schema definition is as follows,

|  |  |  |
| --- | --- | --- |
| Table | Fields | Input Type |
| BRANCH1Branch | beverage | string |
| branch | string |

|  |  |  |
| --- | --- | --- |
| Table | Fields | Input Type |
| BRANCH1branchcount | beverage | string |
| Count | int |

#### Solution

1115974

## Problem Scenario 2 – What is the number of consumers for the Branch2?

*Type 1: Sub queries selection without table creation*

### Explanation

**Step 1:** Selecting data from previous tables with sub queries in which the branch BRANCH2 consumers alone selected and counted. The code as follows,

Hive> select sum(totalcount) from (select BevcountA.beverage,sum(BevcountA.count) totalcount from (select beverage,branch from ( select \* from BevbranchA where branch='BRANCH2' union all select \* from BevbranchB where branch='BRANCH2' union all select \* from BevbranchC where branch='BRANCH2')unionResult )a join BevcountA on(a.beverage=BevcountA.beverage) group by BevcountA.beverage

union all

select BevcountB.beverage,sum(BevcountB.count) totalcount from ( select beverage,branch from ( select \* from BevbranchA where branch='BRANCH2' union all select

* from BevbranchB where branch='BRANCH2' union all select \* from BevbranchC where branch='BRANCH2')unionResult )b join BevcountB on(b.beverage=BevcountB.beverage) group by BevcountB.beverage

union all

select BevcountC.beverage,sum(BevcountC.count) totalcount from ( select beverage,branch from ( select \* from BevbranchA where branch='BRANCH2' union all select

* from BevbranchB where branch='BRANCH2' union all select \* from BevbranchC where branch='BRANCH2')unionResult )c join BevcountC on(c.beverage=BevcountC.beverage) group by BevcountC.beverage

)unionResult ;

*Type 2: Creating multiple physical tables*

Explanation

**Step 1:** Creating physical table “BRANCH2Branch” from the previous tables where the branch name is equal to BRANCH1. The code as follows,

Hive > create table if not exists BRANCH2Branch as select \* from BevbranchA where branch

= 'BRANCH2';

Hive > insert into table BRANCH2Branch select \* from BevbranchB where branch = 'BRANCH2';

Hive > insert into table BRANCH2Branch select \* from BevbranchC where branch = 'BRANCH2';

**Step 2:** Creating another table “BRANCH2branchcount” which merge tables which has the number of consumers with Beverages. The code as follows,

Hive> create table if not exists BRANCH2branchcount(beverage string, count int);

Hive> insert into table BRANCH2branchcount select BevcountA.beverage,sum(BevcountA.count) from BRANCH2Branch join BevcountA on(BRANCH2Branch.beverage = BevcountA.beverage ) group by BevcountA.beverage ;

Hive> insert into table BRANCH2branchcount select BevcountB.beverage,sum(BevcountB.count) from BRANCH2Branch join BevcountB on(BRANCH2Branch.beverage = BevcountB.beverage ) group by BevcountB.beverage ;

Hive> insert into table BRANCH2branchcount select BevcountC.beverage,sum(BevcountC.count) from BRANCH2Branch join BevcountC on(BRANCH2Branch.beverage = BevcountC.beverage ) group by BevcountC.beverage ;

**Step 3:** Finally summation of the fields from total number of consumers in the newly created table “BRANCH2branchcount”. The code as follows,

Hive> select sum(count) from BRANCH2branchcount;

**Step 4:** Created tables schema definition is as follows,

|  |  |  |
| --- | --- | --- |
| Table | Fields | Input Type |
| BRANCH2Branch | Beverage | String |
| Branch | String |

|  |  |  |
| --- | --- | --- |
| Table | Fields | Input Type |
| BRANCH2branchcount | Beverage | string |
| Count | Int |

#### Solution

5099141

## Problem Scenario 3 - What is the most consumed beverage on Branch1?

*Explanation*

**Step 1:** Selecting the aggregate count from the previously created table Branch1Constotcount and ordering the data in descending. The code as follows,

Hive> select beverage,sum(totalcount) totcount from Branch1Constotcount group by beverage order by totcount desc limit 1;

#### Solution

Special\_cappuccino 108163

## Problem Scenario 4 - What are the beverages available on Branch10, Branch8, and Branch1?

*Explanation*

**Step 1:** Creating a table “BRANCH10BRANCH8BRANCH1beverage” from the previous tables which has the records from the BRANCH10, BRANCH8, BRANCH1 branches. The code as follows,

Hive> Create table if not exists BRANCH10BRANCH8BRANCH1beverage as select beverage,branch from (

select \* from BevbranchA where branch = 'BRANCH10' or branch = 'BRANCH8' or branch = 'BRANCH1' union all select \* from BevbranchB where branch = 'BRANCH10' or branch = 'BRANCH8' or branch = 'BRANCH1' union all select \* from BevbranchC where branch = 'BRANCH10' or branch = 'BRANCH8' or branch = 'BRANCH1' )unionResult ;

**Step 2:** “BRANCH10BRANCH8BRANCH1beverage” table schema definition:

|  |  |  |
| --- | --- | --- |
| Table | Fields | Input Type |
| BRANCH10BRANCH8BRANCH1beverage | beverage | string |
| Branch | string |

**Step 3:** Storing the output data table in a local comma separated file in hadoop. The code as follows,

[training@hadoop -]$ hive –e ‘select distinct(beverage),branch from BRANCH10BRANCH8BRANCH1beverage order by beverage’ | sed ‘s/[\t]/./g’ >

/home/training/workspace/output.txt

### The output is stored in the local file is attached here.