Assignment –Data Warehouse

Startup

If you have SQL Server 2012 installed on your personal computer, you can find the Adventure WorksDW2012 data warehouse database files on the class Blackboard site. They are located in the *Course Content/Software* folder.

If you want to check your SQL query answers by computing the checksum (a digital signature of sorts) for your query results you can download my *Generate Checksum* program from the *Course Content/Software* folder. **Be sure to read the instructions** for using the program paying special attention to the information pertaining to result set column headers. All checksums below are generated from result set data that does not include column headers.

Problems 1 through 4 all require the use of the *AdventureWorksDW2012* data warehouse database. Diagrams of the tables in the data warehouse can be found in the *Database Diagrams* folder inside of the database itself.

You may complete this exercise in teams of at most 2 students. Assignment documents must be submitted by a single member of your team. Be sure both team members' names appear in the write-up document.

Problem 1 (15 pts)

The VP of marketing is thinking of purchasing some print ads and would like to better understand the demographics of our customers.

Question 1: Who buys more (by count) jerseys from the Internet, men or women? Paste your SELECT syntax for your solution and the 1st 10 rows (if there are more than 10 result set rows) from your result set into your submission Word document. You should not upload any .SQL syntax files.

SELECT CLAUSE COLUMN ORDER:
`gender, number of purchases`

RESULT SET ORDER:
`order by column 1 ascending`

CHECKSUM: **158388583**

Question 2: What is the count of the number of jerseys purchased over the Internet by gender and year? Use the order date to determine sales year. Paste your SELECT syntax for your solution and the 1st 10 rows (if there are more than 10 result set rows) from your result set into your submission Word document. You should not upload any .SQL syntax files.

SELECT CLAUSE COLUMN ORDER:
`year, gender, number of purchases`

```
RESULT SET ORDER:
order by column 1 ascending then column 2 descending
```

CHECKSUM: **1358337909**

Problem 2 (10 pts)

The VP of Sales is thinking of over-hauling the recent sales promotions used to promote AWC products to resellers. First she needs some data. To answer this question you are to considering reseller sales facts only.

Question 1: For sales where the promotion category is 'Reseller' and excluding, 'No Discount' and 'Volume Discount' promotion types, which sales promotions were most successful in driving dollar sales between 2005 and 2008? Note, this question requires the use of a different fact table than the previous question. In this fact table, the *SalesAmount* attribute contains the total sales for each fact. Since we have no metadata, you will need to poke around in some of the tables to find where attributes such as promotion type, promotion category, etc. reside. Paste your SELECT syntax for your solution and the 1st 10 rows (if there are more than 10 result set rows) from your result set into your submission Word document. You should not upload any .SQL syntax files.

SELECT CLAUSE COLUMN ORDER:

```
English promotion name, total sales amount for the promotion
```

RESULT SET ORDER:

```
order by column 2 descending
```

CHECKSUM: **17146125**

Problem 3 (10 pts)

Examine the schema and data for the Adventure Works Cycles 2012 (AWC) data warehouse. Formulate a realistic business question that can be answered given the data that is available. <u>You must incorporate the date dimension into your question</u>. Since we have no metadata for this data warehouse, you should explore the tables comprising the data warehouse to become familiar with the nature of the data so that you can create a creative (and useful) business question. Now, answer your own question. Paste the SQL required for your solution and the <u>first 10 rows only</u> of your result set into your write-up document.

Problem 4 (40 pts)

The *AdventureWorksDW2012* employs a snowflake design. Here's the question, **what does this approach buy you in terms of storage requirements?** We are going answer this question for a single dimension – *DimCustomer*. The customer dimension is snowflaked into 3 related tables – *DimCustomer*, *DimGeography*, and *DimSalesTerritory*.

Question 1: What are the storage data space requirements (i.e., the size of the data in the table in megabytes) for these 3 tables as currently constituted? Simply add the 3 data space requirements numbers.

Question 2: What would be the storage data space requirements for a de-normalized version (1NF) of the customer dimension? Your 1NF customer dimension must contain all attributes from the 3 tables above with the following exceptions. That is, the following attributes are to be excluded from you new table.

DimCustomer.GeographyKey
DimGeography.GeographyKey
DimGeography.SalesTerritoryKey
DimSalesTerritory.SalesTerritoryKey
DimSalesTerritory.SalesTerritoryAlternateKey
DimSalesTerritory.SalesTerritoryImage

Clearly, answering this question will require that you actually build such a dimension, populate the dimension with data, and determine the storage data space requirements. For this question, save the SQL syntax file that creates and populates the 1NF table using the name ***Problem3.SQL***. Upload the ***Problem3.SQL*** along with your submission's Word document

Note: I am aware that you may not know exactly how to perform some of the required steps. Actually, part of the reason for asking this question is to require you to search, scrounge, and be otherwise resourceful in order to figure out how to perform the required steps. Solutions to this problem will range from extremely simple to extremely convoluted and complex. That depends on just how resourceful you are.

Hint: your 1NF customer dimension should have 18,484 rows.

Question 3: What is the % change in data space requirements in going from the snowflaked approach to DimCustomer and the 1NF approach? Does the added design and query complexity of the snowflake design seem worth while?

Problem 5 (25 pts)

Over the past 5 years, data warehouse size has skyrocketed; new physical data management paradigms and products have evolved to keep pace. Many, if not most, of these projects are so-called open source projects. Do some web research into the following 5 products **Hadoop**, **Hive**, **Pig**, **Piglatin**, and **Zookeeper**. For each:

- Describe the purpose of the product. I.e., what is it designed to do or what deficiency was it designed to overcome. If possible, include in your discussion the origins of the product. I.e., why was it originally invented and where or by whom?

- List at least two companies that have embraced one or more of these technologies from a user perspective. That means I'm not talking about how Microsoft and IBM re-bundle and sell these products, I want you to find users of these products. Users of the re-bundled versions of the products are fine. For example, Microsoft's HDInsight product offers 100% compatibility with Hadoop. Any Microsoft customer using this technology would be fair game. For each company, briefly describe how/where they use the product in there organization and any benefits that they claim to receive from its use.