## Aggregating and Analyzing Your Data

- What it takes to build a BI/OLAP solution

*By Dr. Chris Zhao*

This brief document is written for those who would like to aggregate and analyze his/her data in support of his/her business evaluation, judgment, and business decision making. In Mathematics, data is abstracted into numbers and character strings and therefore does not carry any business meaning. However, in the business world within a business context, data always carries business meaning.

For example, the number 98 could mean 98 customers, the year 1998, or sales price of $98, or growth rate of 98%, or a comparison of 98 times. Therefore, in order for me to aggregate and analyze your data, it is very important to know:

1. The description of your data: i.e. the business meaning of each table and data columns involved in the aggregation and analysis.
2. The mapping, relationship, or hierarchies between your tables and columns;
3. Your requirements: what you want out of your data and how you want your data to be aggregated and analyzed
4. The supply of your data

The more I know your data and the more I understand your requirements, the more meaningful and fruitful my analytical effort would be. Once I have intimate knowledge of your data, it will also help me to be creative when analyzing your data, for example, providing some ideas and insights, as well as additional measures and dimensions that you haven’t thought about. Without intimate knowledge of your data, it is impossible to be creative. Therefore, ***knowing data is the key to analyzing data***.

When describing your data, the best way to follow is a **data dictionary**. For each table, you should provide a paragraph describing what each record in the table represent. For example, a record in Sales table might represent a sales transaction together with all the attributes associated with a sales transaction such as sales date, sales price, customer ID, … etc. Then, you need to identify either a column or a combination of columns in the table that would or should uniquely identify each record in the table. This is so-called **unique identifier**. This unique identifier would help to determine the **granularity of your data** in the table. Generally speaking, such a column or a combination of columns represents the characteristic columns of each record, and forms the so-called “Primary Key” of the table. In some cases, you may have an identity column or a database-generated unique identifier column in the table that serves as the Primary Key. This type of column, in general, does not carry any business meaning, they should not be used as unique identifier for the purpose of determining the granularity of your data records.

Next, you need to describe each column in all your tables, i.e. the business meaning of each attribute represented by the column.Finally, you need to describe any existing relationship between any two tables by showing:

* What column(s) in one table relate to what column(s) in another table, and
* What is the relationship? Is it one-to-one, one-to-many, many-to-one, or many-to-many?

Now let us talk about item 3 - your requirements. It would be helpful if you know some basic OLAP concepts – especially the dimension and measure concepts when describing your requirements. Let us use Walmart business as an illustration. If you ask: “What is my total revenue, profit, and expenses by Store City, Year, and Product Category”, then you are requesting an analysis based upon 3 measures and 3 dimensions. The 3 measures are:

* [Total Revenue],
* [Total Profit],
* [Total Expense],

And the 3 dimensions are

* [Store] – where store city is an attribute,
* [Time] where Year is an attribute,
* [Product] where Product Category is an attribute.

Basically, **measures** describe “What you want?”, and **dimensions** describe “How you want, from what perspectives?”

You can imagine Dimension as an entity or object, and the attributes of the dimension are simply properties associated with the entity or object. For example, employee can be a dimension which would contain First Name, Last Name, Gender, Age, DOB, Hire Date, Salary, Education, … as attributes.

For the easiest and simplest approach, you can start gather your requirements by **making a list of questions that you and your users would be interested to ask**, and these questions should be addressed or answered by an OLAP solution or a data mining solution. To give you some idea about what questions can be addressed or answered by OLAP cubes, here are some illustrative examples based on a retail business:

* What are my total sales revenue by product category, year, and region?
* What is my sales trend – annual growth percentage by country and year and quarter?
* Is my customer base expanding or shrinking?
* What are the geographical or industry distribution of my customers?
* What characteristics do my customers share for certain products?
* What types (characteristics) of potential customers should I target to sell my products?
* …… etc.

You can make a long list of these questions. This kind of questions can be helpful in drawing a big picture of your requirements. Another approach would be making a list of measures and dimensions (and their attributes) based on your business, if you have some basic understanding on measures and dimensions.

Using Walmart business, here I make a partial list of possible measures and dimensions (and their attributes):

**Measures:**

Revenue

Profit

Expense

Employee Salary

Rent

Tax

Average Profit Per Store

Average Salary

# of employees

# of customers

Quantities of Product Sold

**Dimensions and their Attributes**

Store

StoreID

Name

Location

City

State

Openning\_Date

Square Feet

Product

ProductID

Name

Unit Price

Weight

Size

Product Category

Barcode

Date/Time

Year

Quarter

Month

Date

Employee

EmployeeID

SupervisorEmployeeID

First Name

Last Name

Hire Date

Annual Salary

Department

StoreID

Gender

Another concept that you may be interested to know is called **key performance indicators**, or simply called **KPI**. KPIs are customizable business metrics or measures that represent an organization’s status, progress, and trends toward achieving its predefined goals in an easily understandable format. After a business defines its mission or objectives, KPIs can bedefined to measure its progress toward those objectives.

In general, each KPI has a target value and an actual value. The target value represents aquantitative goal considered important to the success of a business or organization. The goalsof organizations can vary widely for different types of businesses, because their aims areoften dissimilar. For example, a business might have KPIs concerning sales, net profit, anddebt ratio, whereas a school might define a KPI relating to graduation rate.To determine the health of the business, the actual values are compared to the targetvalues. KPIs provide a clear description of organizationalgoals, and distill large amounts of data to a single value that can be used by senior managementto monitor business performance and progress toward organizational benchmarks.KPIs can be rolled into a scorecard - a group of KPIs - that shows the overall health of thebusiness, where each KPI is a metric or measure of one aspect of business growth.

In conclusion, here are the important initial considerations in planning and designing a data warehouse and BI solution:

1. **Know the data and the business knowledge associated with the data.**

* Description of each table - what each record represents? At what granular level?
* Unique Identifier of each table – what columns combined would be able to uniquely identify each record in the table?
* Description of each table column that will be used – the business meaning
* Any existing hierarchies or mapping between columns in any table
* Table Relationship - The relationship between every two tables that are related: what column(s) in what table is linked to what column(s) in what table? What type of relationship is this? Is it one-to-many, one-to-one, or many-to-one?

1. **Understand the business needs or requirements.**

One way of gathering the requirements is to collect a list of sample questions that business managers / analysts would be interested to get answers.