

Deep Diving in the World of Data Warehousing



BluePi Consulting [Follow](#)

Aug 23, 2017

“Information is the oil of the 21st century, and analytics is the combustion engine”.—Peter Sondergaard, Senior Vice President, Gartner.

Number of companies are venturing into Big Data Analytics today and industries are investing money, effort and looking to utilize the technical advancements for the business growth.

Till date RDBMS, excels have been favored by industries to fulfill their reporting, analytics needs, but as data is growing at rapid pace, there are roadblocks posed by these traditional measures like scalability, performance etc.

In this blog, we will try to deep dive into Analytics space and understand various terms prevalent with help of a real-world scenario.

Architecture Vision

The concept of data warehousing has evolved out of the need of improving the ongoing decision making process. Over a period, organizations collect vast amounts of data but find it increasingly difficult to access and make sensible use of it. This is mainly because data is collected at different points, is in different formats, stored on many platforms. Once initial findings are made, digging deeper into data is costly, inefficient, and very time consuming.

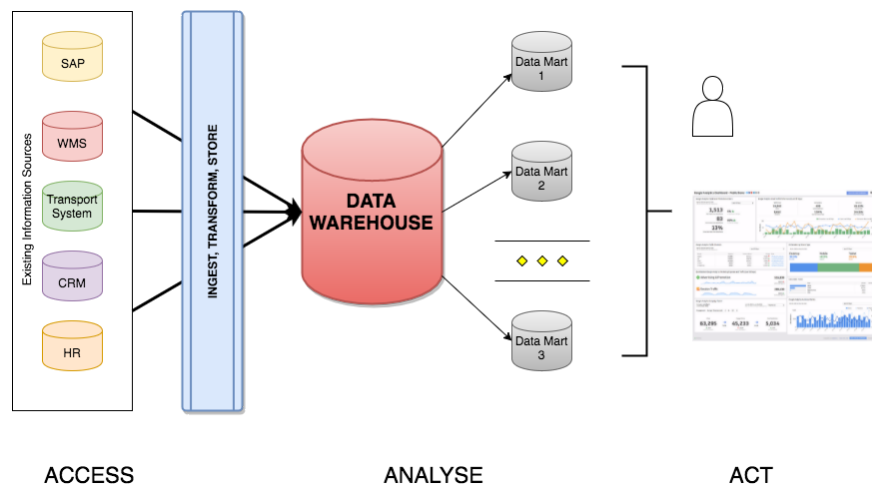
Data warehousing helps in implementing the process of:

- Fetching data from heterogeneous data sources.

- Clean, filter, and transform the data for insights.

- Store the data in a form and manner that is easy to access, understand, and use.

The processed data is then used for query, reporting, and data analysis. As such, the access, use, technology, and performance requirements are completely different from those in a transaction-oriented operational environment.



Above diagram shows the key components of the proposed solution:

Data Warehouse Database

Extract, Transform, Load (ETL)

Meta Data

Reporting and Dashboard Tools

Data Marts

Data Warehouse Database

A data warehouse is, by definition, a subject-oriented, integrated, time-variant collection of data.

There are number of solutions present in market like Teradata's EDW Platform, Oracle Exadata Machine, or our in-house favourite and most competent of them all Amazon Redshift.

One of the most basic concepts of data warehousing is to clean, filter, transform, summarize, and aggregate the data, and then put it in a structure for easy access and analysis by those users. But, that structure must first be defined and that is the task of the data warehouse model.

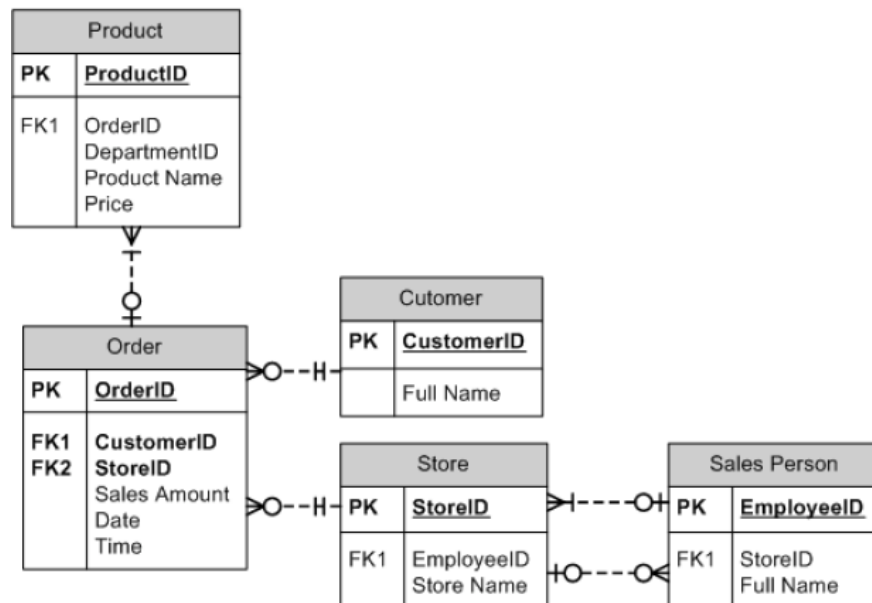
There are two basic data modeling techniques

ER Modeling

Dimensional Modeling

ER Modelling

ER Modelling consists of three basic concepts i.e. entities, attributes and relationships.

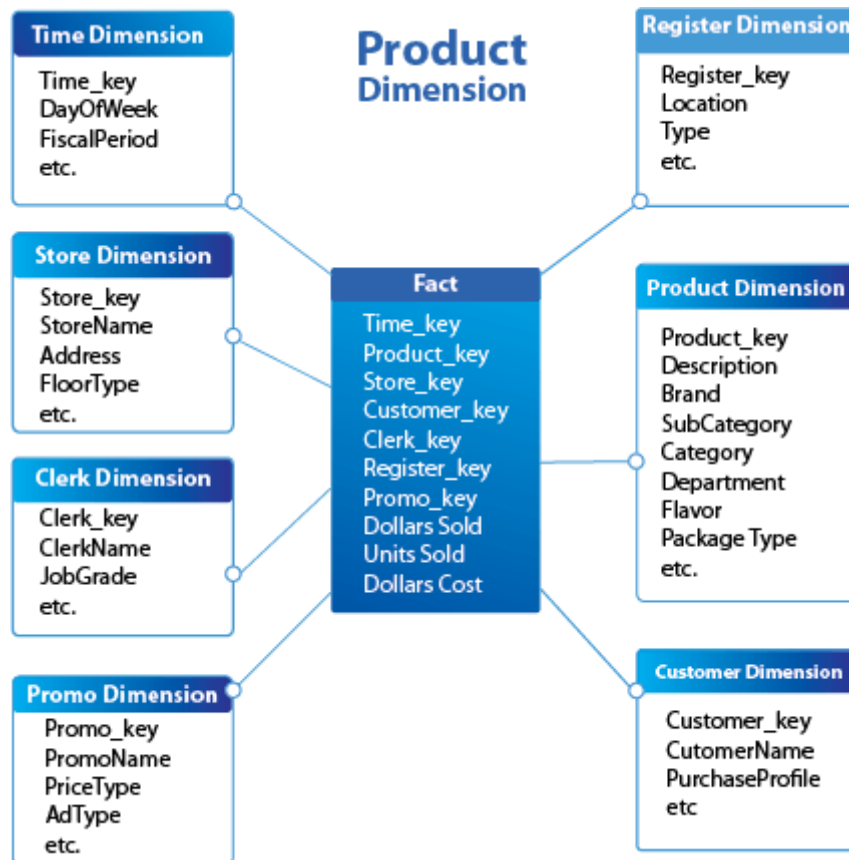


It's better when we want to keep data in normalized form and since entities are the core concept the data load and query revolves around the logical, real world objects like Product, Customer etc.

This data modelling is followed by the RDBMS data stores and are very efficient for the transaction based systems but fail to scale when it comes to analytical systems.

Dimensional Modelling

Dimensional modeling is simpler, more expressive, and easier to understand. There are 3 basic concepts in dimensional modeling i.e. **facts**, **dimensions** and **measures**.



A fact is a collection of related data items, which consist of measures and context data. Each fact typically represents a business item, a business transaction, or an event that can be used in analyzing the business or business processes.

A dimension is a collection of members or units of the same type of views. Dimensions are the parameters over which we want to perform Online Analytical Processing (OLAP). For example, in a database for analyzing all sales of products, common dimensions could be:

Time

Location/region

Customers

Salesperson

A measure is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions. For example, measures are the sales in money, the sales volume, the quantity supplied, the supply cost, the transaction amount etc.

Dimensional modeling is primarily used to support OLAP and decision making while ER modeling is best fit for OLTP where results consist of detailed information of entities rather an aggregated view.

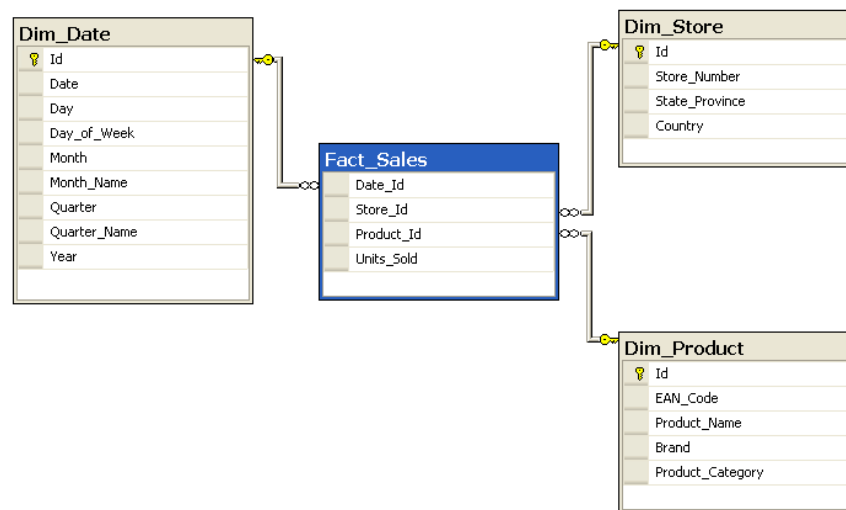
It provides four types of operations: **Drill down**, **Roll up**, **Slice** and **Dice**.

Drill down and roll up are the operations for moving the view down and up along the dimensional hierarchy levels to get more refined or bird-eye views. With drill-down capability, users can navigate to higher levels of detail. With roll-up capability, users can zoom out to see a summarized level of data.

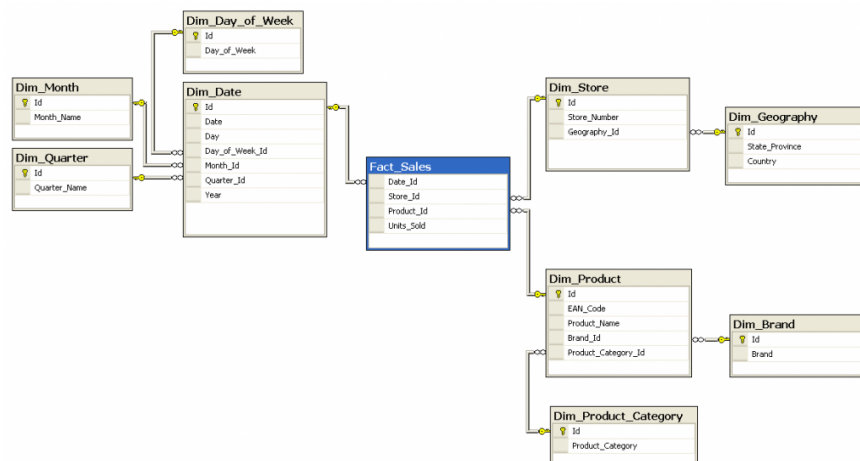
Slice and dice are the operations for browsing the data through the visualized cube. Slicing cuts through the cube so that users can focus on some specific perspectives. Dicing rotates the cube to another perspective so that users can be more specific with the data analysis.

Star and Snowflake Schema

The star model is the basic and most widely used structure for a dimensional model. It typically has one large central table (called the fact table) and a set of smaller tables (called the dimension tables) arranged around the fact table, as shown below.



The snowflake model is like star schema but results from decomposing one or more of the dimensions, which sometimes have hierarchies themselves. This leads to normalization of data but may increase query cost. Example is shown below:



Hope that with our detailed analysis and description, you have gathered some interesting and knowledgeable insights, and now these various terminologies used in a typical IT setup, imparts you more clarity and correct usage.

In our next blog of this series, we will see various other components used to build an end-to-end analytics solution.

Till then, keep innovating!

If you have any feedback for us or have some query, feel free to drop us an email at info@bluepi.in or visit our site at www.bluepiit.com

This blog has been written by **Mrs. Aashu Aggarwal**, one of our core Leadership Team's Member.