International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

Design of Data Warehouse

Author: Vipul W. Dhakate Mentor: Mrs.Sweta Nigam

MCA 6th Semester

Tilak Maharashtra Vidyapeeth Pune

ABSTRACT :

Data warehouse (DW) is pivotal and central to BI applications in that it integrates several diverse data sources, mainly structured transactional databases. However, current researches in the area of BI suggest that, data is no longer always presented in only to structured databases or format, but they also can be pulled from unstructured sources to make more power the managers' analysis. Consequently, the ability to manage this existing information is critical for the success of the decision making process.
The operational data needs of an organization are addressed by the online transaction processing (OLTP) systems which is important to the day-to-day running of its business. Nevertheless, they are not perfectly suitable for sustaining decision-support queries or business questions that managers normally needs to address. Such questions involve analytics including aggregation, drilldown, and slicing/dicing of data, which are best supported by online analytical processing (OLAP) systems. Data warehouses support OLAP applications by storing and maintaining data in multidimensional format. Data in an OLAP warehouse is extracted and loaded from multiple OLTP data sources (including DB2, Oracle, SQL Server and flat files) using Extract, Transfer, and Load (ETL) tools.

Introduction :

The retail sector was one of the first sectors to make significant investments in collecting and integrating customer data in data warehouse. The project is required in order to assist the management of the retail company in making better decision using the historical data available within the organization. The business users (decision makers) lack the ability to access data easily when needed. In an attempt to address this shortcoming, several departments within the retail company find their own resources, use different data available and hired consultants to solve their individual short-term data needs. In many cases, the same data was extracted from the same source systems to be accessed by separate departments without any strategic overall information-delivery strategy. The management realized the negative effect the different sources of the data has on the reports presented by the managers as the lack of integration. Given the importance of the information for the retail company, the management was motivated to deal with the problem of data inconsistence by introducing a central data warehouse and to ensure that data is available to all users irrespective of their department. The data harmonization and the need for consistent and quality report gave birth to the project of data warehouse in the company. Business need and business strategy is the need for the build of the data warehouse and business intelligence. A data warehouse is the foundation for powerful data analysis, it supports business decision by encouraging manager and other users of the company to examine data and carry out analysis in a better way. Because the data has been gathered in a repository, it can facilitate measurement of the effect of various combinations of factors like supply chain, customer's preference, demography, geographic and can assist the analyst in working out the customer retention process and trend. The levers that a retailer can use to optimize performance include: price, promotion, markdown, assortment, space, allocation and replenishment. Data driven decision making is key to successful decisions regarding all of these levers. Competition in the retail sector is becoming increasingly fierce as the complexities of global expansion, rapid product cycles, currency fluctuation and changing customer preferences continue to transform many segments. Hence, Crystal entertainment must be able to make strategic decision that would influence the position of the organization in the entertainment industry and stay on top of the game.

Aim and Objectives :

This project has only one aim and that it can be achieved through the actualization of the objectives as stated below;

- To examine the importance of Data warehouse and business intelligence system in an entertainment industry.
- To design and develop a data warehouse and business intelligence system in an entertainment industry.
- To evaluate how the decision tools would assist the decision maker in taking better decision about the company.
- To validate the design of data warehouse and business intelligence using the case study.

Data Warehouse Concepts :

Data warehousing is the process of collecting data to be stored in a managed database in which the data are subject-oriented and integrated, time variant, and nonvolatile for the support of decision-making. Data from the different operations of a corporation are reconciled and stored in a central repository (a data warehouse) from where analysts extract information that enables better decision making.
Data can then be aggregated or parsed, and sliced and diced as needed in order to provide information. The approach by Inmon is top down design while that of Kimball is bottom up design. Most of the practitioners of Data warehouse subscribe to either of the two approaches.
According to Inmon, a Data Warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data used in support of decision making processes. “Subject Oriented” means that a data warehouse focuses on the high-level entities of the business and the data are organized according to subject “Integrated” means that the data are stored in consistent formats, naming conventions, in measurement of variables, encoding structures, physical attributes of data.For example, whereas an organization may have four or five unique coding schemes for ethnicity, in a data warehouse there is only one coding scheme.Data warehouse is the conglomerate of all data marts within the enterprise. Information is always stored in the dimensional model.Data marts are focused on delivering business objectives for departments in the organization. And the data warehouse is a conformed dimension of the data marts. The data warehouse is the sum of all the data marts, each representing a business process in organization by a means of a star schema, or a family of star schemas of different granularity. There is no right or wrong between these two ideas, as they represent different data warehousing philosophies. This is because most data warehouses started out as a departmental effort, and hence they originated as a data mart. Only when more data marts are built later do they evolve into a data warehouse.
The DW characteristics to include;

- It is subject-oriented.
- It is non-volatile.
- It allows for integration of various application systems.
- It supports information processing by consolidating historical data.
- Data is stored in a format that is structured for querying and analysis.
- Data is summarized. DWs usually do not keep as much detail as transaction-oriented systems.

The Data Warehouse Data Model:

There are three levels in data modeling process: High-level modeling (called the ERD, entity relationship level) which features entities, attributes and relationships, Mid-level modeling (called the data item set) which is data set by department, and Low-level modeling (called the physical model) optimize for performance.
After the high-level data model is created, the next level is established—the midlevel model. For each major subject area, or entity, identified in the high level data model, a midlevel model is created. Each area is subsequently developed into its own midlevel model.
The physical data model is created from the midlevel data model just by extending the midlevel data model to include keys and physical characteristics of the model. At this point, the physical data model looks like a series of tables, sometimes called relational tables.
Data is conformed (Data elements are conformed so that the definitions of "customer" or "revenue" mean the same thing no matter where the originated), Data is historical (view of the business at a particular point in time), Data is shared (Can be queried or otherwise accessed has little value), Data is comprehensive (Can be captured and consolidated from multiple systems).

DW Database Design Modeling:

There are three levels of data modeling. They are conceptual, logical, and physical. For the purpose of this thesis, we would discuss only the first two. Conceptual design manages concepts that are close to the way users perceive data; logical design deals with concepts related to a certain kind of DBMS; physical design depends on the specific DBMS and describes how data is actually stored.
The main goal of conceptual design modeling is developing a formal, complete, abstract design based on the user requirements.
DW logical design involves the definition of structures that enable an efficient access to information. The designer builds multidimensional structures considering the conceptual schema representing the information requirements, the source databases, and non functional (mainly performance) requirements. This phase also includes specifications for data extraction tools, data loading processes, and warehouse access methods. At the end of logical design phase, a working prototype should be created for the end-user.

Developing Data Warehouse:

Demarest was explicit when it say that planning the developing and deployment of a standard data warehouse should be taken as an IT project, hence what made IT project fail applies also applies when developing data warehouse; thus the need for Project Planning and following the system development life cycle. There is the need for careful planning, requirements specification, design, prototyping and implementation. The cyclical model entails five stages which are described below;

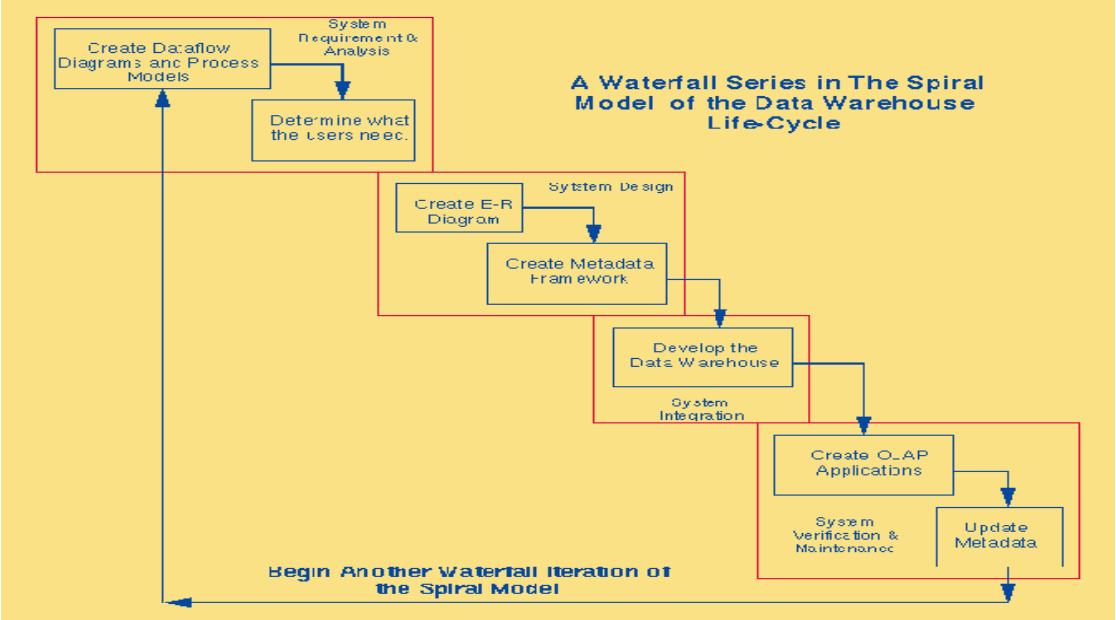

Figure 1: DW Development Lifecycle (DWLC) Model

Where the Design stage takes information from both available data inventories and analyst requirements and analytical needs, of robust data models and turns it into data marts and intelligent information. The Prototype deployment stage, where group of opinion-makers and certain end-user clientele, are brought in contact with a working model of the data warehouse or data mart design, suitable for actual use. The purpose of prototyping shifts, as the design team moves back and forth between design and prototype. Deploy stage is the stage of formalization of user-approved prototype for actual production use. The Operation is the day-to-day maintenance of the data warehouse or mart, the data delivery services and client tools that provide analysts with their access to warehouse and the management of ongoing extraction, transformation and loading processes that keep the warehouse current with respect to the authoritative transactional source systems. Enhancement stage is where external business conditions change discontinuously, or organizations themselves undergo discontinuous changes enhancement moves seamlessly back into fundamental design, if the initial design and implementation didn't meet requirements.

Top-Down Model:

These was Introduced by Bill Inmon, The process begins with an Extraction, Transformation, and Loading (ETL) process working from legacy and/or external data sources. Extraction transformation, process data from these sources and output it to a centralized Data Staging Area. Following this, data and metadata are loaded into the Enterprise Data Warehouse and the centralized metadata repository. Once these are constituted, Data Marts are created from summarized data warehouse data and metadata. In the top-down model, integration between the data warehouse and the data marts is automatic as long as the discipline of constituting data marts as subsets of the data warehouse is maintained.

Bottom-Up Model:

The central idea in Bottom-up model is to construct the data warehouse incrementally over time from independently developed data marts. The process begins with ETL for one or more data marts. No common data staging area is required. There is generally a separate area for each data mart. There may not even be standardization on the ETL tool. The Model was introduced by Ralph Kimball.
For the purpose of this project, Bottom-up model approach would be adopted, which is the Kimball's development lifecycle, this states with one data mart (e.g. Sales) later on further data mart are added e.g. Marketing and Collection. Data flows from sources into data marts, then into the data warehouse. It is also implemented in stages (faster) Due to the time constraint and project limitation, it is easier to complete a process for a subset of a company based on the data mart and link it up as the business grows. The stages proposed for the process include Investigation, Analysis of the current environment, identify requirements, and identify architecture, data warehouse design, implementation and ongoing data administration. Kimball et al (1998).

Methodology:
The following methods were used;

- Secondary Research: Due to the time constraint, it allows us to move close to the aim by examining the existing data collated by the company.

- Field based Research: To better understand the nucleus of the project we did little of field research in the form of question which is anonymous.

- Case study – to examine the objective of the research project in order to formulate the strategy. We look at the Crystal music industry as the case study.

System Analysis and Research Methods:
In this stage, we expect to analyze data which has been compiled by Crystal Entertainment over the years. This phase also involves outlining the functions that the DW will achieve; and an ideal working environment in which the data warehouse will be delivered. Whatever the business requirements are, the overall goal is to get a perception of the core utilization of the initial data and to identify other stakeholders who may need access to the data. Also at this stage, the business analysis/user requirements are to understand the workings of users in relations to the business and how they want to use the solution, what data they currently make use of, and what they would like to do with such data. This data can then be used in different manner to decompose this information into Business entities and their attributes, and manage relationship between the entities, and hierarchies. The requirements can be gathered through a chain of interviews with the different stakeholders. Answers from these users will generate the requirements needed for further development of the data warehouse.

Secondary Research:
In this research, the company has data collected through the OLTP (Online Transaction Process) and would be use in the thesis. Data warehouse and business intelligence is all about using the existing data to enable the users, managers and decision makers in the organization to make insightful decision about the business. We have chosen to employ the secondary analysis because with secondary analysis, there is more time for data analysis as data collection could be very time consuming. Some of the limitation we come across includes lack of familiarity with the data, complexity of the data, no control over the quality of data and absence of key variables.

Field Research:
In order for us to take an informed decision during the design stage of the project, we find it very important to ask the business users (used interchangeably in this project to mean Decision Makers/Managers) what their expectations are regarding the design of the data warehouse for the company. We will be using an online questionnaire which is strictly anonymous to get more information about the data usage and what report is important to the business user.Questionnaire is a research instrument consisting of a series of questions and other prompts for gathering information from respondents (*Wikipedia Extract*). With a self-completion questionnaire, respondents' answers question by completing the questionnaire themselves. With the self-completion questionnaire, there is no interviewer to ask question; instead, respondents much read each question themselves and answer the questions themselves.

Case Study:
Case study is the method of choice when the phenomenon understudy is not readily distinguishable from its context. Such a phenomenon may be project or program in an evaluation study. Sometimes, the definition of this project or program may be problematic, as in determining when the activity started or ended. The inclusion of the context as a major part of a study, however, creates distinctive technical challenges. First, the richness of the context means the ensuing study will likely have more variable than data points. Second, the richness means that the study cannot rely on a single data collection method but will likely need to use multiple source of evidence. Third, even if all the relevant variables are quantitative, distinctive strategies will be needed for research design and for analysis. We have decided to do a case study as it gives an in depth study of a particular situation rather than a sweeping statistics survey. It is a method used to narrow down a very broad field of research topic. Case study provides more realistic responses than a purely statistics survey. They are more detail than the statistical method. Case study in this project will give us the opportunity to study the aim of the project using some past project in the same industry and compare it with aim of this project.

Implementation:

- **System Analysis:**

Analysis involved a detailed study of the current system, leading to specifications of a new system. Analysis is a detailed study of various operations performed by a system and their relationships within and outside the system. During analysis, we studied the activities of the company and we choose 3 departments to design the data mart for and data were collected on the available files, decision points and transactions handled by the present system. Interviews, on-site observation and questionnaire are the tools used for system analysis. Using the following steps it becomes easy to draw the exact boundary of the new system under consideration:

• Keeping in view the problems and new requirements

• Workout the pros and cons including new areas of the system

- **Retail Data:**

In the future, firms will need to continue to be cost effective but increasingly will need to focus on using data to drive revenue by better understanding their customer's needs. This understanding will come from supplementing internal collected data with the vast quantities of external data generated or made accessible by internet. Organisation with latest BI technology tools to integrate his cross enterprise, inter enterprise and external data in order to achieve insight and transparency, across all channels. Any company that can effectively harness the vast quantities of information that the IT systems generate- both within the corporation and outside its walls are poised to gain competitive advantage. The simple definition of a transaction can reveal significant discrepancies across department and users. By the time a particular transaction is completed, so many deductions, rebates, discount and other trade spending had occurred that it is almost impossible to specifically identify profit centre at a granular level (i.e by customer, by product, by channel). And without this level of detail, planning for profitable volume is no more than guess; the challenge lies in the insight, not in the availability of raw data.

References:

- Alan, B, Emma B, (2007), 2nd Edition Business research methods, Oxford, Oxford Press.
- Balaji, P. & Alexander T. (2003). On the Use of Optimization for Data Mining: Theoretical Interactions and ECRM Opportunities. Management Sciences, New York.
- Ballard Chuck, et al (1998*)* Data Modeling Techniques for Data Warehousing IBM International Technical Support Organization.
- Başaran, Beril P (2005)**,** a Comparison of Data Warehouse Design Models, the Graduate School of Natural and Applied Sciences, Atilim University.
- Başaran, Beril P (2007), Developing a Data Warehouse for a University Decision Support System, The Graduate School of Natural and Applied Sciences, Atilim University.
- Burton, P et al (2010), Meta Data: The Key to Data Warehouse Design (A Systems Engineering Approach) ENSE623 Project Institute of System Research.