School of Computing
Blekinge Institute of Technology

Design of Data Warehouse and Business Intelligence System

A case study of a Retail Industry

Temitope Adeoye Oketunji
Raufu Olalekan Omodara

Thesis submitted for completion of Master of Science (60 credits)
Main field of study: Computer Science
Specialization: Informatics

June 2011

School of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona
Sweden

This thesis is submitted to the School of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science (60 credits) in Computer Science with specialization in Informatics. The thesis is equivalent to 10 weeks of full time studies.

Contact Information:
Author(s):
Temitope Adeoye Oketunji
Address: If applicable
E-mail: toketunji@gmail.com

Contact Information:
Author(s):
Raufu Olalekan Omodara
Address: If applicable
E-mail: rolekan@yahoo.com

External advisor(s): if applicable
Firstname Lastname
Company/Organization name, do not forget AB
Address:
Phone: Only exchange number, international standard, e.g. use +

University advisor(s):
Han Taps PhD
School of Computing, Blekinge Institute of Technology

School of Computing
Blekinge Institute of Technology
SE-371 41 Karlskrona
Sweden

Internet : www.bth.se/com
Phone : +46 455 38 50 00
Fax : + 46 455 38 50 57

ABSTRACT

Business Intelligence (BI) concept has continued to play a vital role in its ability for managers to make quality business decision to resolve the business needs of the organization. BI applications comes handy which allows managers to query, comprehend, and evaluate existing data within their organizations in order to obtain functional knowledge which then assist them in making improved and informed decisions.

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.

This thesis seeks to develop DW and BI system to support the decision makers and business strategist at Crystal Entertainment in making better decision using historical structured or unstructured data.

Keywords: Business Intelligence, Data Warehouse and OLTP, OLAP, ETL, SQL Server

EXECUTIVE SUMMARY

The entertainment industry is not left out of technological advancement which is significantly transforming the way business is being conducted in and around the industries. It has opened up better opportunities for managers to have better grip and insights into their business viability.

Crystal Entertainment[1] faces some challenges which include; which is most profitable brand of their business? How do they reuse existing operational data to support business operations? What are the buying patterns of customers? Which one of the distribution channels is most profitable to the organization? What is the revenue trend of the business processes in the current financial year? And so on. These are some of the unanswered question, and answers to them are required to make better decision about the future of the organization. They get data from various sources and it is left to the collation and analyses of such data to make managers make informed decision.

This is a ground-breaking opportunity for Crystal Entertainment to take advantage and opportunities that the field of Business intelligence presents. Starting from exhaustive market and customer behavior analysis, to summary level sales and profit trending, to real-time operational intelligence, business intelligence offers insight into the opportunities and how to harness them.

We can categorically state that Business Intelligence would ultimately increase revenue for crystal entertainment by improving; the ability to leverage customer data to understand buying patterns, have insight into the success of new products, features, and services, show visibility of the most profitable customers, target promotions and marketing campaigns towards failing products.

[1]Crystal Entertainment is a fictitious name for the purpose of reporting this thesis work in line with the best practices of the entertainment industry.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DB	-	Database
DW	-	Data Warehouse
ER	-	Entity Relationship
ERD	-	Entity Relationship Diagram
ETL	-	Extracting, Transforming, Loading
OLAP	-	Online Analytical Processing
OLTP	-	Online Transaction Processing
RDBMS	-	Relational Database Management System
SSAS	-	SQL Server Analysis Services
SSIS	-	SQL Server Integration Services

CHAPTER 1

1.0 Introduction

Crystal Entertainment is an entertainment retailer specializing in music, films, and audio books. It has eight online stores operating in the United States, Germany, France, the United Kingdom, Spain, Australia, Japan, and India. It has 96 offline stores operating in those countries as well.

Customers can buy individual products such as a song, an audio book, or a film, or they can subscribe to a certain package, which enables them to download a certain quantity of products in a certain period. For example, In the UK, with the Primer package, a customer can download 100 songs, 50 books, and 50 films a month for £150. Customers can also listen or watch a song, a book, or a film once for 1⁄10 the cost of purchasing it. So if a film is £5, to watch it once it's only 50 pence.

Customers use online streaming for this, so a good Internet connection is required. Crystal Entertainment has four main delivery channels: Internet, mobile phone, cable TV, and post. There are several payment models for customer subscriptions, such as annual, in advance, and monthly direct debit. The company purchases the products in bulk, such as any 10,000 songs from a record company, of any title, for a certain cost. For online streaming, the company pays a central provider (XXX Broadcasting PLC.) based on usage (monthly invoice).

The company uses different applications to carry out their business activities. This includes a custom-developed .NET-based system for dynamic web sites, multimedia trading, broadcasting, sales order processing, and subscription management, all running on an different database platform. The back-end enterprise resource planning (ERP)

1.1 Reason for the Project

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.

1.2 Area and Scope of Research

The research is based on an understanding of the outline of both academic and industry theories under consideration and the background on why there is need to carry out the research.

The scope of this project will focus on the design, development of a data warehouse and business intelligence, which will include analysis of data and the presentation of the data in the form of information using analysis and reporting services tools. Without the data warehouse, it is almost impossible to have a business intelligence system. Due to the time constraints and information available, the project would not cover all the areas of the business, the focus will be on two data marts that represent major department within the organization which are for the sales and marketing department. The system will be developed using Microsoft SQL server 2008 R2.

1.3 Aim and Objectives

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

1.3.1 Aims

To develop a Data warehouse and business intelligence system to support the decision makers and business strategist in making better decision using historical data

1.3.2 Project Objectives

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

1.4 Scope

The aim of Crystal Entertainment is to be a significant player in its chosen markets. Expertise, service and execution skills will differentiate the strategic business unit (SBU) from its peers.

The project is required in order to help the decision makers of the company making an effective decision. Effective decisions are choices that move an organization closer to an agreed-on set of goals in a timely manner. (Larson, 2006)

Effective decision is important at all organizational levels. Timely foundation and feedback information is needed as part of that effective decision-making. Therefore, we need to make business intelligence available throughout the organization.

CHAPTER 2

2.0 Critical Literature Review

This section focuses on providing the necessary information about the relevant literature that connects to the research. It will provide theories about Data warehouse and business intelligence that covers the research problem.

2.1 Definition

Business intelligence is the delivery of accurate, useful information to the appropriate decision makers with necessary timeframe to support effective decision-making. Larson (2006)

Data warehouse is a system that retrieves and consolidates data periodically from the source systems into a dimensional or normalized data store. It usually keeps years of history and is queried for business intelligence or other analytical activities. It is typically updated in batches, not every time a transaction happens in the source system Rainardi (2008)].

Data Mart is a subset of data warehouse and is defined as body of historical data in electronic repository that doest not participate in the daily operations of the organization. Instead, this data is used to create business intelligence. The data in the data mart usually applies to a specific area of organization. (Larson, 2006)

Fact Table is the primary table in a dimensional model where the numerical performance measurements of the business are stored. We try to store the measurement data resulting from a business process in a single data mart.

Dimension Table is an integral companion to a fact table. The dimension tables contain the textual descriptors of the business. In a well-designed dimensional model, dimension tables have many columns or attributes. These attributes describe the rows in the dimension table. Dimension tables tend to be relatively shallow in terms of the number of rows (often far fewer than 1 million rows) but are wide with many large columns. Dimension tables are the entry points into the fact table. The dimensions implement the user interface to the data warehouse.

Online analytic processing (OLAP) database is a technology for storing, managing, and querying data specifically designed to support business intelligence uses.

Extract, Transformation, and Load (ETL) system is a set of processes that clean, transform, combine, de-duplicate, archive, conform, and structure data for use in the data warehouse.

2.2 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 (Inmon, 1993). 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 (Cho and Ngai, 2003).

Data can then be aggregated or parsed, and sliced and diced as needed in order to provide information (Fox, 2004). There are two main authors that are known in the world of data warehouse design, their approaches to some area of the data warehousing are different; William Inmon and Ralph Kimball. 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 (1993), 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 (Chan, 1999) and the data are organized according to subject (Zeng et. al., 2003; Ma et. al., 2000)

”Integrated” means that the data are stored in consistent formats, naming conventions, in measurement of variables, encoding structures, physical attributes of data, or domain constraints O’Leary, (1999). 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 (Chan, 1999).

“Time-variant” means warehouses provide access to a greater volume of more detailed information over a longer period (Zeng et. al., 2003) and that the data are associated with a point in time (Chan, 1999; O’Leary, 1999) such as month, quarter, or year. Warehouse data are non-volatile in that data that enter the database are rarely, if ever, changed once they are entered into the warehouse. The data in the warehouse are read-only; updates or refresh of the data occur on a periodic, incremental or full refresh basis (Zeng et. al., 2003) Finally,” nonvolatile” means that the data do not change (Chan, 1999).

According to Kimball (2002), Data warehouse is the conglomerate of all data marts within the enterprise. Information is always stored in the dimensional model. Kimball views data warehousing as a constituency of data marts. 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. Kimball (1996) describes a data mart as a subset of data warehouse. 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.

The main difference between the approach of Kimball et al. (1996) and that of Inmon (1993) is that Kimball’s conformed dimensions are de-normalized, whereas Inmon uses a highly normalized central database model. Inmon's data marts stores a second copy of the data from the centralized data warehouse tables, whereas the dimensions of Kimball used in the data marts, are not copies of the conformed dimensions, but the dimension table themselves. Kimball et al (1996) refers to the set of conformed dimensions as the data warehouse bus.

There is no right or wrong between these two ideas, as they represent different data warehousing philosophies. In reality, the data warehouse in most enterprises is closer to Ralph Kimball's idea. 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.

Basaran (2005) reveals some of 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.

2.2.1 The Data Warehouse Data Model

Inmon (1993) argues that 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. Basaran (2005)

Stuart Mullins (2007) in his blog titled “Data Warehouse Data Model Design” explains what can be used to differentiate the DW from an ordinary archive database which can easily become a dumping ground. 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).

2.2.2 DW Modeling Techniques

Ballard (1998) gave an exploration of the evolution of the concept of data warehousing, as it relates to data modeling for the data warehouse, they defined database warehouse modeling is the process of building a model for the data in order to store in the DW. There are two data

modeling techniques that are relevant in a data warehousing environment are Entity Relationship (ER) modeling and dimensional modeling.

ER modeling produces a data model of the specific area of interest, using two basic concepts: entities and the relationships between those entities. Detailed ER models also contain attributes, which can be properties of either the entities or the relationships.

Dimensional modeling uses three basic concepts: measures, facts, and dimensions. Dimensional modeling is powerful in representing the requirements of the business user in the context of database tables. Measures are numeric values that are can be added and calculated. Ballard (1998)

2.2.3 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. Basaran (2005)

2.3 Developing Data Warehouse

Demarest (2008) 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;

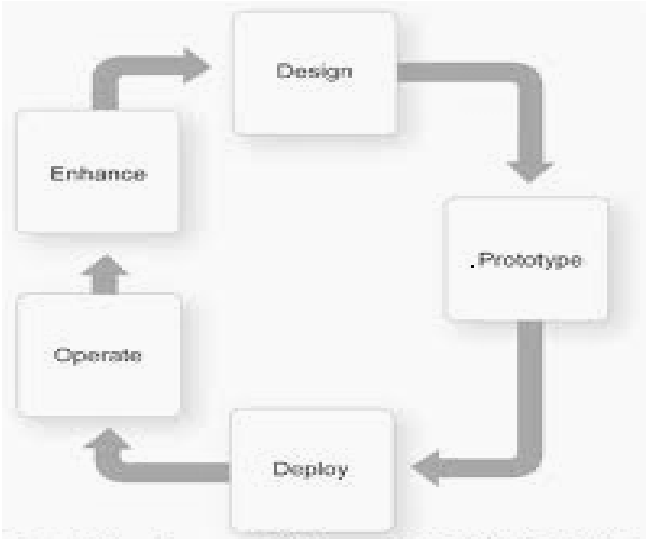


Figure 1: DW Development Lifecycle (DWLC) Model (Demarest, 2008)

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.

2.4 Business Intelligence Concepts

Initially, BI was coined as a collective term for data analysis tools. Meanwhile, the understanding broadened towards BI as an encompassment of all components of an integrated decision support infrastructure. In BI systems, data from OLTP is combined with analytical front ends to "present complex and competitive information to planners and decision makers". A central component of BI systems is the data warehouse (DW), which integrates data OLTP for analytical tasks.

From the managerial approach, BI is seen as a process in which data from within and out the organisation are consolidated and integrated in order to generate information that would facilitate quick and effective decision-making. The role of BI here is to create an informational environment and process by which operational data gathered from transactional systems and external sources can be analyzed and to reveal the "strategic" business dimensions. From this perspective emerge concepts such as "intelligent company": one that uses BI to make faster and smarter decisions than its competitors (Liautaud, 2000). "Intelligence" means reducing a huge volume of data into knowledge through a process of filtering, analyzing and reporting information.

The technological approach presents BI as a set of tools that supports the storage and analysis of information. The focus is not on the process itself, but on the technologies that allow the recording, recovering, manipulation and analysis of information. For instance, Scoggins (1999) classifies data mining (DM) as a BI technique; Hackathorn (1999) includes all resources (DW, DM, hypertext analysis and web information) in the creation of a BI system; and finally, linking BI and the Internet, Giovinazzo (2002) posit the integration of DW and customer relationship management (CRM) applications.

Whether managerial or technological, there is one shared idea among all these studies: (1) the core of BI is information gathering, analysis and use and (2) the goal is to support the decision making process, helping the company's strategy. Taking into account the scarce literature, we looked for other areas that could help us reach a more comprehensive understanding of BI. We find contributions in three distinct topics: information planning, balanced score card and competitive intelligence.

Here are some benefits that business intelligence offers and how they can help the entertainment industry to make and distribute creative substance and stay aloft of the game: [21]

- Product profitability: How much profit does a particular item contribute? How does item's profit break down across business units, media and distribution channels? What are the specific costs and expenses associated with producing the item? What percent of revenue or profit do they represent?

- Customer and market analysis: What are the key demographic characteristics of customers by product? Which other products do they tend to buy? Does the data indicate that an underserved market segment has greater revenue potential?

- Channel analysis: Which channels reach what types of consumers? How profitable is each channel? How will channels be affected by changing technologies, as well as the emergence of new channels?

- Forecasting and planning: What are the market potential of a new product, and how much investment should be made? How will a new release perform and what will its profit contribution be? What level of supply will adequately meet demand?

The result – employees can now access detailed sales data from around the world, which was previously not possible, and they are also able to run sophisticated self-service reports that provide granularity and a near real-time view into sales performance, ultimately helping these users make informed decisions that drive the results of the business. In addition to sales data, Media companies can measure marketing and promotion effectiveness and monitor corporate performance and results.

BI not only converts raw data into intelligent information, but also allows business users to access the right information at the right time and able to transform it into smart decisions. Media companies with its business processes based on such intelligent information can disrupt its competitor's moves, strategize a sustainable competitive edge, tap into new customer bases, retain existing customer bases, increase operational efficiencies and be better prepared for the future. [17]

2.5 Data Warehousing versus Online Transactional Processing (OLTP)

Data warehouse are also known as Online Analytical processing (OLAP) system because they serve managers and knowledge workers in the field of data analysis. Online transaction processing (OLTP) systems or operational systems are those information systems that support the daily processing that an organizational does. OLTP system's main purpose is to capture information about economic activities of an organization. On might argue that the purpose of OLTP system is to get data into computers, whereas the purpose of data warehouse is to get data or information out of computers

Han and Kamber (2001) argue that an OLTP system is customer-oriented as opposed to a data warehouse that is market-oriented. It is a bit difficult to combine data warehousing (OLAP) and OLTP capabilities in one system. The dimensional data design model used in data warehouse is much more effective for querying that the relational model used in OLTP systems. Furthermore, data warehouses may use more than one database as a data source. The dimensional design is not suitable for OLTP systems mainly due to redundancy and loss of referential integrity of the data. Organisations choose to have two separate information systems, one OLTP and One OLAP system

Poe et al (1998) stress the fact that analysis using OLAP systems are primarily done through comparisons, or by analyzing patterns and trends. For example, sales trends are analyzed along with marketing strategies to determine the relative success of specific marketing strategies with regard to sales patterns; such analysis may not be possible with OLTP. Kimball supported same idea but Inmon (1993) was a bit different on the approach to the development of data warehouse system. He argue that although OLTP are developed from requirements as a starting point, data warehousing starts at implementing the data warehouse and ends with a clear understanding of the requirements. The data warehouse development lifecycle is data-driven and OLTP are requirements driven. Kimbal (1996) differ from this approach by following a requirements-driven development lifecycle.

2.6 Data Warehouse and Business Intelligence High Level Architecture

Eckerson (2003) from the Data warehouse institute did study on the success factor in implementing BI, systems in organisations and the role of data warehouse in this process.

Eckerson (2003) views the BI process holistically as a "data refinery" Data from different OLTP systems are integrated, which leads to a new product called information. The data warehouse staging process is responsible for the transformation. Users equipped with program such as specialized reporting tools, OLAP tools and data mining tools transform the information into knowledge. Kimball (1996) includes this as part of the data warehouse. According to Kimball, the aim of the data warehouse is to give end-users (mostly managers) easy access to data in the organization. In order to do this, it is necessary to capture everyday operational data from the operational systems of the organization. Theses are the OLTP system. The data from the source systems go through a process called data staging to the presentation servers (Kimball et al 1996). The data at the staging process involves four processes namely Extract, Transformation, Loading and finally presentation. It is on the presentation stage that the data marts, which represent business areas in the organization is build on.

The data in the data mart or data warehouse is stored as star schemas consisting of FACT and DIMENSION tables. This is different from the entity relational diagram (ERD) used in traditional systems.

There is a difference between the data warehouse and business intelligence architecture as advocated by the two known scholars in the industry, (Inmon, 1993) advocates the use of data-driven method. This means that the Decision Support System process begins with data and ends with requirements. In contrast to Inmon's approach, advocate the use of requirements-driven methods. The data warehouse starts with the project planning to determine the readiness of the organization for a data warehouse and to set the staff requirement for the data warehouse team. A clear understanding of business requirements is the most important success factor and Kimball, (1996) state that this process of requirements collection differs substantially from data-driven requirements analysis. The business requirements establish the foundation for the three parallel tracks focused on technology, data and end user applications.

2.7 Data Warehouse Design Concepts

The design of the database depends on the approaches of the father of data warehouse developers. The two-design processes are referred to as Top-down process, as described by

Bill Inmon and Bottom-up as described by Ralph Kimball. These are explained in detail below.

2.7.1 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.

2.7.2 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).

CHAPTER 3

3.0 Methodology

This chapter consists of the approach we have taken to undertake the designing the data warehouse and business intelligence system. Due to the fact that the project is based on the business requirement, the implementation will be based on three major phases which are analysis, design and development. According to Nachamais et al (1996), methodologies are consider being systems of explicit rules and produced, upon which research is base, and against which claim for knowledge are evaluate. Eldabi et al (2002) argue that the conducting of any type of research should be governed by a well-defined research methodology based on scientific principles.

Data will be collected from various sources within the organization. The collected data would validate the existing requirement analysis being carried out by the company already. Case study will be used to analyze the objective of the research. All precaution has been taken not to go out of the scope of the project and not to go over what the company is accepting as the best practice for the industry. 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.

3.1 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.

3.1.1 Secondary Research

According to Alan & Emma (2007) Secondary analysis is the analysis of data by researchers who will probably not have been involved in the collection of those data, for purposes that in all likelihood were not envisage by those responsible for the data collection. In this research, the company has data collected through the OLTP (Online Transaction Process) and would be use in the thesis.

Why do we need to use secondary research for this project? 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.

3.1.2 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.

3.1.3 Case Study

According to (Yin, 2003); 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. (Yin, 2003)

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.

3.2 System Design

The main focus of this phase is to translate the systems requirements into a set of specifications through deriving logical and physical data models or data marts for the data warehouse. The specifications are then used to generate other component such as data warehouse extractors and transformation, data integration tools and so on. Decisions making by managers can be intricate influenced by both political, economically and other technical reasons, But poorly made decision can ruin the whole essence of the analysis. Finally, processes are identified to connect the data sources, the data warehouse, and the end user access tools together. Burton (2010)

As mention in the literature review, the two fathers of Data warehouse are Bill Inmon and Ralph Kimball. They have different approach to the design of the data warehouse. For the purpose of this project, we adopted the Ralph Kimball Method. The data from each department is treated as a data mart and we can design from start to finish a complete data warehouse and business intelligence for chosen departments as mentioned in the scope of the project.

Software engineering development methodologies can contribute immensely to the development of DW can constitute well-established strategies and techniques for the development process. We have employed the “Spiral model” The Spiral Model is a sequence of the corresponding waterfall models which corresponds to a risk oriented iterative enhancement, and recognizes that requirements are not always available and clear when the system is first implemented. Since designing and building a data warehouse is an iterative process, the spiral method is one of the development methodologies of choice. This is to ensure that any business requirements not clear at the beginning of the analysis stage can be often re-visited. The diagram below shows one waterfall series in a recommended spiral model of a data warehouse life-cycle.

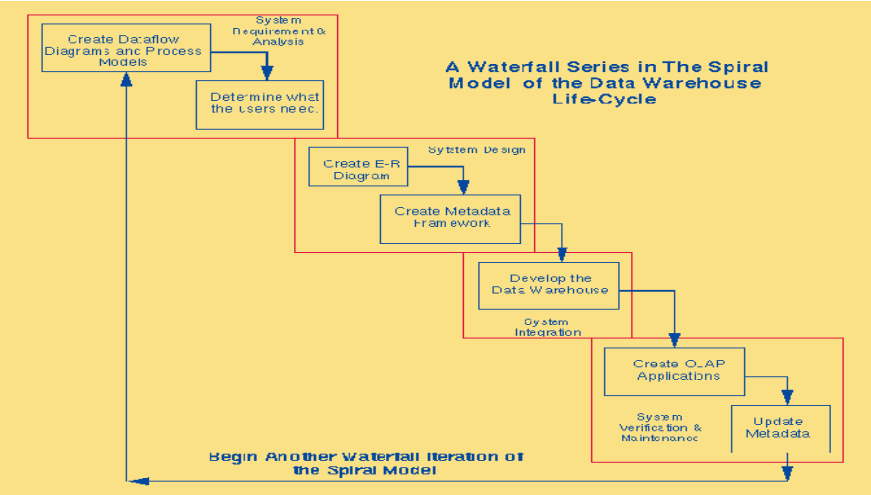


Figure 2: Spiral Model of the Data Warehouse Life-cycle [6]

We start our data mart design by specifying the measure, the measure are the foundation and feedback information that the decision makers require. We reconcile these requirements with what is available in the source system (OLTP). For the purpose of this project, we used the star schema for the data warehouse design. The star schema is a relational database schema used to hold measures and dimensions in a data mart. The measures are stored in a fact table and the dimensions are stored in dimension tables. For each data mart, there is only one measure surrounded by the dimension tables, hence the name star schema.

The centre of the star is formed by the fact table. The fact table has a column or the measure and the column for each dimension containing the foreign key for a member of that dimensions. The key for this table is formed by concatenate all of the foreign key fields. The primary key for the fact table is usually referred to as composite key. It contain the measures, hence the name “Fact”

The dimensions are stored in dimension tables. The dimension table has a column for the unique identifier of a member of the dimension, usually an integer of a short character value. It has another column for a description. In this project to follow the naming convention we are going to name the dimension tables based on the information they contained and prefix with “Dim”

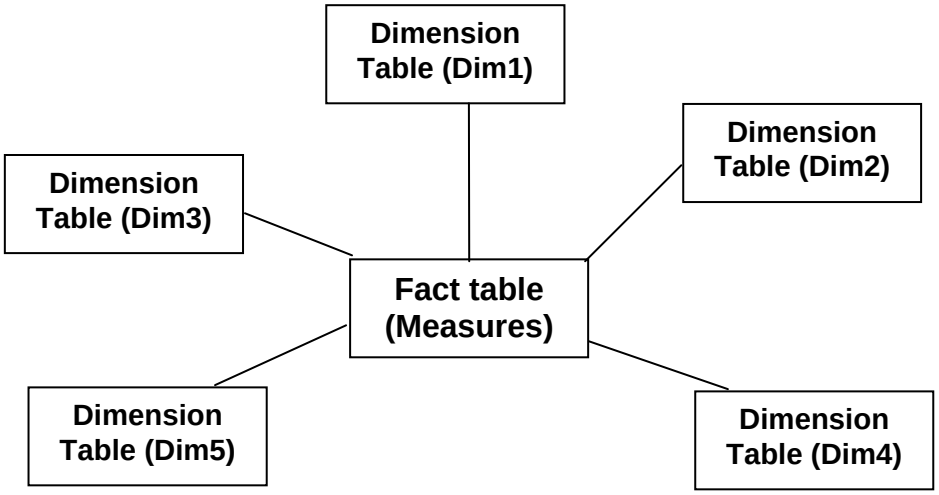


Figure 3: A Star Schema

3.3 System Development and Validation

The system development is the actual implementation of the analysis and design carried out. In this phase of the project, we designed the data warehouse (Fact and dimension tables), the ETL (Extract, Transform and Load) and the front end application for the purpose of this project.

Validation process involved the confirmation by examination and provision of objective that an information system has been implemented correctly and conforms to the need of the user and intended use.

The main focus of this phase is developing procedures to validate the data that has been extracted and moved data in a form that can then be loaded into the warehouse. Finally, the data must be analyzed to determine whether or not certain elements should be cleansed prior to putting it into the warehouse. Burton (2010)

3.4 System Verification & Maintenance

The best methods for verifying the data in the warehouse is to prepare reports on the data in the warehouse and compare it to figures based on the data subsequent to putting into the

warehouse which are perceived to be correct. It is seldom believed that users verify the data because they are quite familiar with the detailed type of data they are after. Lastly, Maintenance is essential at each and every stage of the life-cycle. Primarily this entails documentation of processes, applications and most significantly, metadata.

CHAPTER 4

4.0 Implementation

This chapter focuses on the process of implementing the data warehouse and business intelligence for Crystal Entertainment. This chapter looks at the system analysis, system design and finally system development.

4.1 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

4.1.1 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.

4.1.2 Uses of Crystal Music Company Data

The data available within the company are so vast that the company can make the best use of the data to project ahead. Due to the dynamism in technology, the music industry has suffered a bit as music is downloaded through the internet and it has led to many store closure. The crystal company needs to make use of the available data to determine where the major revenue to the company is coming from and establish any of the POS that are not profitable to the company and probably to close them down or merge. The company also needs to understand the kind of music that people love to listen at and the supplier that are profitable as well. Within the scope of this project, we looked at developing data mart for some of the department that is the core of the business. They are the Sales & Marketing, supply & logistic departments.

4.1.3 Functional Requirement

In general, requirements are partitioned into **functional requirements** and **non-functional requirements**.

Functional requirements are associated with specific functions, tasks or behaviors the system must support, it can be in any format but has to be in line with the business requirements.

At this stage we design a web based questionnaire and sent the link to the used on anonymously based and we also ask to interview some of the key identified user that also agreed to tell us more about what they do, how they do it and in what way the project can be of help to them. We sent the questionnaire to some of the managers and users in each department of the company.

After the collation of the questionnaire and interview with staff anonymously, we came up with this requirement. As this is an academic project, we decided to list out the requirements but we did not intend to achieve all due to time and resources required. The identified functional requirements are stated below; the function requirement can be review at different stages of the project in order to cater for new discoveries during the project design.

Some of the data used in this project are secondary data as some of the analysis has been carried out and going through all the process again would be too big for this project in the context of academic. We adopted some of the functional requirements from existing analysis while we validated the functionality through our questionnaire.

- The users need to be able to access the data from different company's application in one single location and in a format that can be easily manipulated.
- The user should be able to analyze the product sales over time geographically. This will be based on the actual sales on monthly basis.
- The business should be able to calculate the profit margin on monthly sales for subscribed customer by different segments
- The user will like to analyse the sales of their product by geographical, store, and different point of sales.
- The business will like to determine the cost of sales and profit from the subscribed customer on an annual basis by package, demography and by store.
- Business user will like to classify customer based on their performance and loyalty
- Business user will be able to determine the performance of the supplier based on the product.
- The system will be able to display the figures and charts and will be able to print. The system should be able to render the report in different format that are useful to the users.

4.1.4 Non Functional Requirement

- The system should be based on windows authentication so that user does not require to log on to the application many times. Single sign-on

- The front end application should be web enabled and no installation is required on users system
- The front end application should be integrated into the existing portal like Share-Point and intranet
- User level permission is required in order to protect the integrity of the data and restrict user's accessibility to data
- The system should perform very well at all times and should be easy to recover after system down time.
- The system should be able to keep up to-date information at all time.

4.1.5 User Requirement

- Ability to generate report with little effort
- Ability to get the aggregate report and drill down for further details
- Ability to download data from data warehouse and use it for further analysis
- System reliability at all time.

4.1.6 System Requirement

For the purpose of this project, we looked at using Microsoft Windows Operating system. The main application, we will be using SQL server 2008 R2 which is the latest product in Microsoft technology that support the enterprise data warehouse and business intelligence. The application has the relational database management system that is capable of storing all the data required for the data warehouse. It has the functionality that can extract data from different sources and consolidate it into one single location for better analysis. This is known as the integration services (SSIS)

It has the module that can be used to develop a cube for each of the data marts we have chosen to analyze in this project. Cubes are pre-calculated aggregates and can be store in the OLAP database. The cube can be analyzing further using other front end application. This is known as analysis services. (SSAS)

The front end application for this project would be the Reporting services (SSRS); it can be used to design reports according to user's request and can also design different gauge and dash board. It has the capability to design different charts and graphs. (SSRS) We used the developer edition which can be upgraded to enterprise edition in the future.

To run the above application, the operating system would be from Windows 2003 and above, minimum of 2 GB memory, including the data 100GB of hard drive space is required. The speed of processor could be from 3.0 MHz duo core. The other system unit components are required to support the operating system and the application for this project.

4.2 System Design

One of the main aims of the data warehouse is to extract data from different OLTP or flat files sources and consolidate them in a single repository for easy access and make best of use of the data. The two processes of data warehouse are data load and access. The design of the system was very robust in order for the aim to be achieved. The loading of the data warehouse was done through the use of ETL (Extract, Transformation and Load) process.

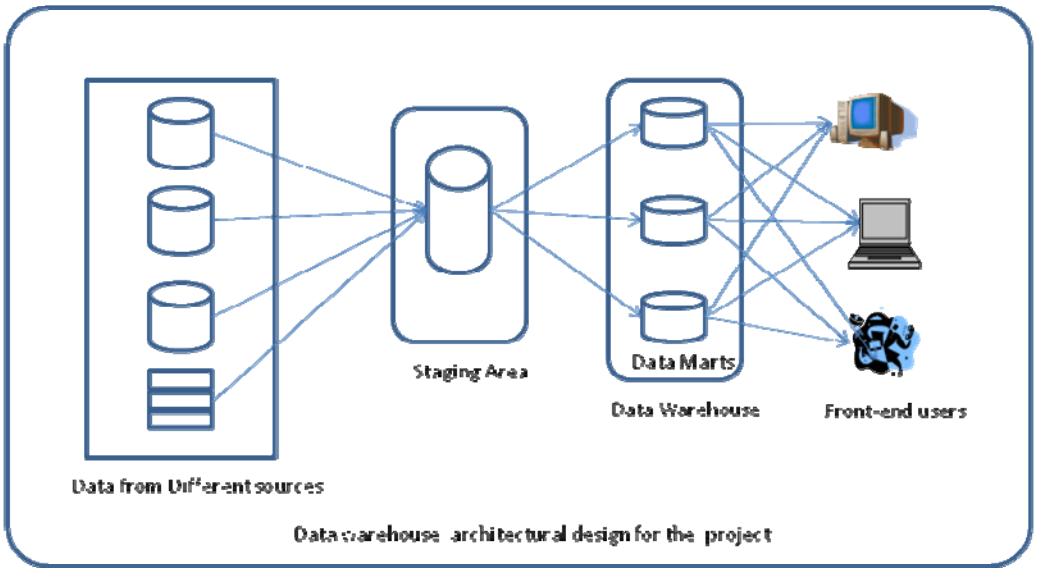


Figure 4: Data Warehouse Architecture Design of the Project

Above is the architectural design for the data warehouse and business intelligence we chose for this project. It may not necessarily follow the industry standard; it is a way of learning as this is academic research.

The design of the databases started with the principle and theories of database design and the rule that support business need. To start the process of data warehouse design this comprises of the data mart. We started the process with the logical design.

4.2.1 Logical Models

The logical model is a representation of the data in a way that can be presented to the business as well as serve as a road map for the physical implantation. The main elements of a logical model are entities, attributes, and relationships. We started the design of the data marts through the fact and dimension tables. All database design start with logical design.

4.2.2 Facts and Dimensions Tables

Fact table contains the measurements associated with a specific business process. A record in a fact table is a measurement and a measurement event can always produce a fact table record. These events usually have numeric measurements that quantify the magnitude of the events. These numbers are called facts; they are also referring to as measure in the analysis services.

Dimensions are the foundation of the dimensional model, describing the objects of the business such as customer, suppliers, subscriber and other dimension table to be used in this design of the data mart.

According to Ralph Kimball the dimension serves as the nouns of the DW/BI system. They describe the surrounding measurement events. The business processes (facts) are the action of the business in which the dimension participates. Each dimension table links to all the business processes in which it participates

For the purpose of this project, we looked at the design of the three data marts within the organization and they are product sales, subscription sales and supplier's performance. According to Ralph Kimball, data marts represent a unit or departmental process within an organization. Data mart is the collection of fact table and its dimensions table. Using the bottom up data warehouse design, combination of the data mart would form the data warehouse. The design of these data marts will be combination of NDS and DDS architecture.

Starting with the sales department, the analogy for the sales events are where, who and what, the actors are the product, customer and store. The facts are based on the users' requirement as specify in the functional requirements. According to Ralph Kimball, it is important to declare the grain of the fact table. Grain is the smallest Unit of occurrence of the business events in which the events is measured.

Sales Data Mart

The business event is the fact table row and stated below

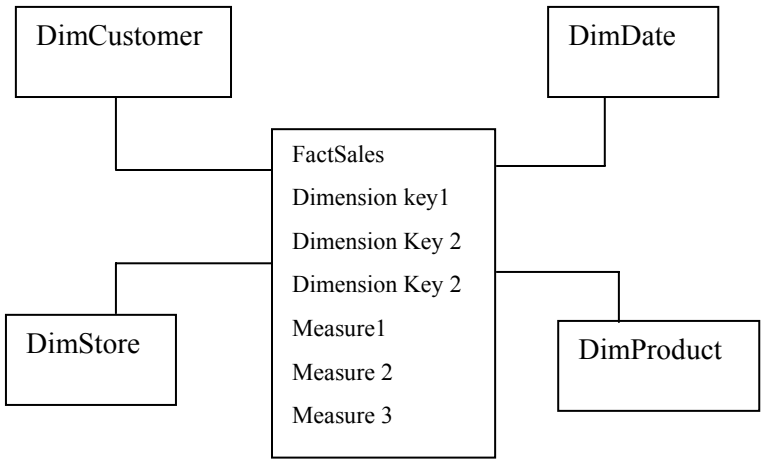


Figure 5: Star Schema for the Product Sales Data Mart

The dimension and the fact table are as follow date, customer, product and store. The dimensions structure would be discuss and designed later in this chapter. Some of the measures are derived from the source while other are calculated based on the available information within the source data. The first 4 fields in the table represent the key that link the fact table to the dimension table. Next step is to determine which column combination will uniquely identify a fact table row. This is important as it is required for both logical and physical design in order to determine the primary key. We will need to design the dimension table for the fact table to complete the data mart. A dimension table is a table that contains various attributes explaining the dimension key in the fact table.

The link between the fact and dimension table is through the referential integrity. The dimension table has the primary key while the fact table has the foreign key. The referential

integrity is a concept of establishing a parent-child relationship between two tables with the purpose of ensuring that every row in the child table has a corresponding parent entry in the parent table. We can enforce the referential integrity as either hard referential integrity or soft referential integrity. The former is enforced at the table level while the later can be enforcing through the ETL.

The dimension tables contain various attributes explaining the condition of entities involved in the business event. They are known as dimensional attributes. To complete the data mart for the sales department above, we need to add the dimension tables. The data warehouse is mostly design to cater for the historical data, while talking about the dimension table, we are looking at the SCD (Slowly Changing Dimension) these are the values that change in the life time of the dimension table, and there is need to keep the historical value to help us in analyzing the data better.

We considered the Type 2; it adds the new value as additional column rather than overwrite the existing data. Type 2 can help in preserving the table structure. Whenever the data changes, new data would be added as a new column in order to preserve the history. This type of slowly changing dimension enables data analysis using historical data.

As mentioned earlier, dimension table is what is used to explain the attributes of the fact table in a data mart. Combinations of the fact and dimension table form the data mart. For the Product sales data mart, the dimension tables are Date, store, customer and products. The product dimension describes the details being sold by the company. Because we are dealing with a music store, the product detail is related to the industry. Crystal entertainment sells music.

The star schema in figure 5 above formulates the logical design of the data marts shown below. It shows the structure of the entire dimension and fact table. It uses the arrow to identify the referential integrity.

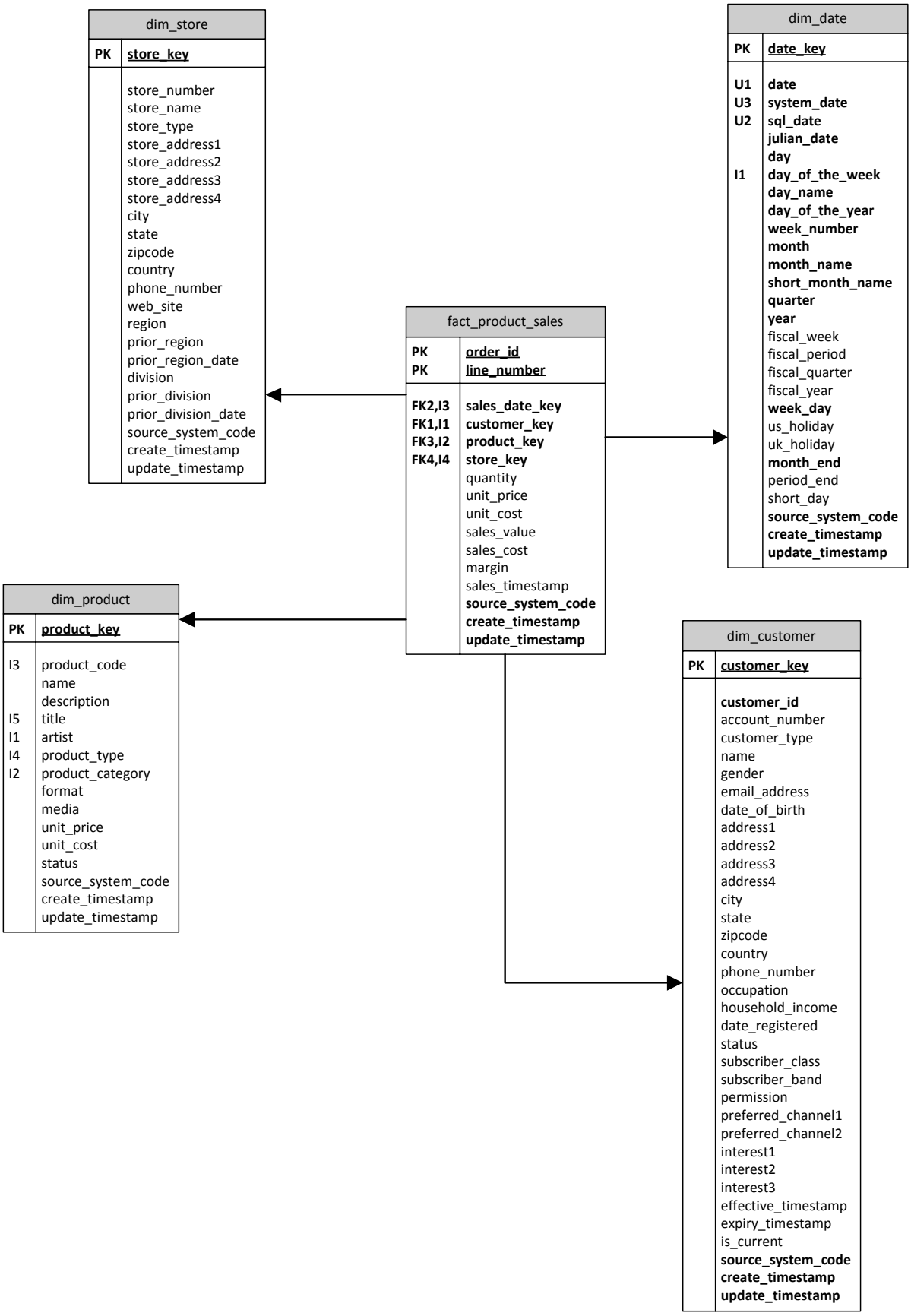


Figure 6: Logical Model of Product Sales Department

Subscription Sales Data Mart

As part of the user's requirement, we thought it necessary to design the data structure for the subscription sales data mart. As part of the business requirements is the ability to analyses subscription sales based on region and stores, analyze subscription profitability and classification of the subscribers. For the design of this data mart, we used the star schema and add as much granularity as possible in order to foster better analysis.

The diagram representation of the of the subscription sales is as follows;

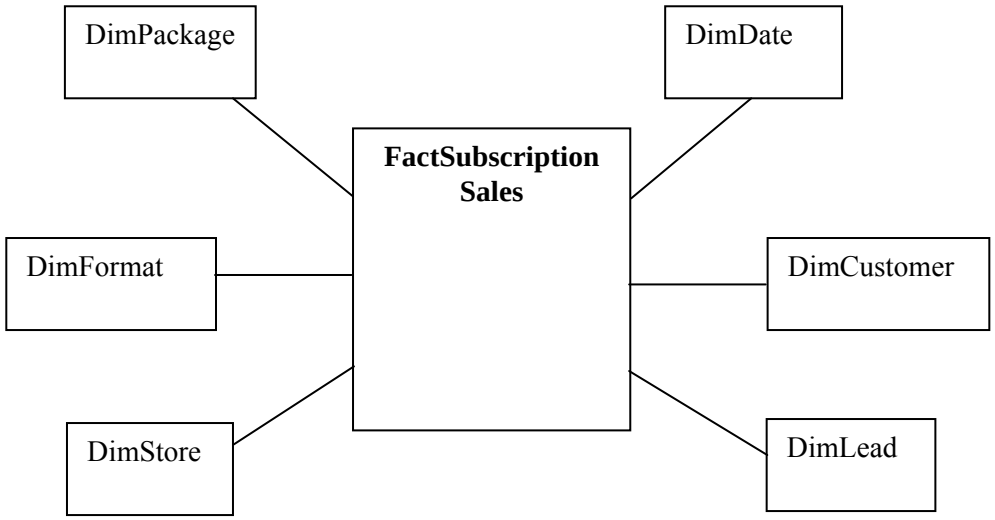


Figure 7: Star Schema for the Fact Subscription Sales

The logical design of the schema shows the structure and the relationship between the fact and the dimension tables. The structure of fact table for the data mart is as follows; the dimension tables to the fact table are as follow Date, customer package, store and lead. Some of the dimension has been defined and design and can be re-used by the fact table or measure.

The DimPackage describes the different packages the company has that the customers can subscribe for. It include the price and as well. The DimDate structure has been describe above for the fact sales table and can be re-used in the fact subscription star scheme. DimFormat describes the different medium that the music or the book can be.

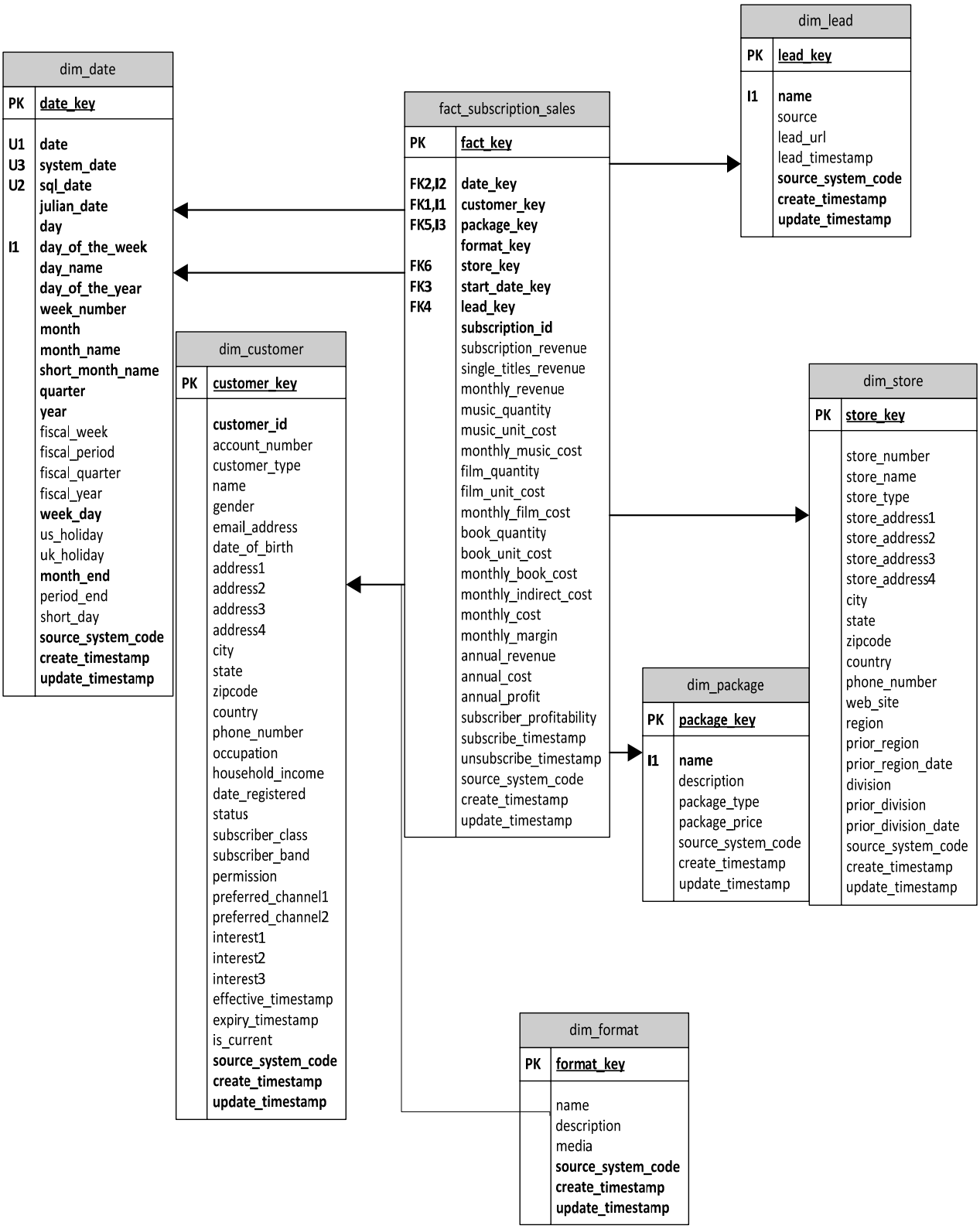


Figure 8: Logical Model of Fact Subscription Sales

Supplier's Performance Data Mart

Why is the data mart required and how important is it? The purpose of this data mart is to support the user to analyze "supplier performance," which is the weighted average of the total spent, costs, value of returns, value of rejects, title and format availability, stock outages, lead time, and promptness.

The diagram below represents the dimensional start schema for the supplier performance data mart

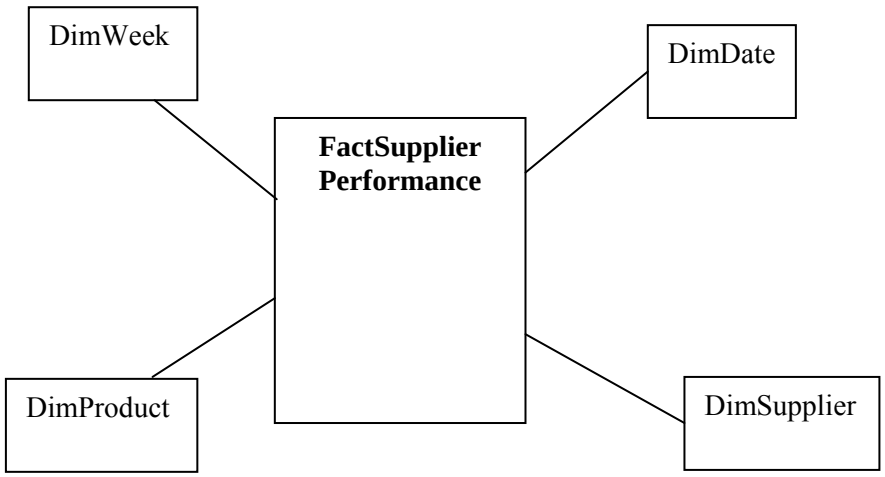


Figure 9: Star Schema of the Fact Supplier Performance

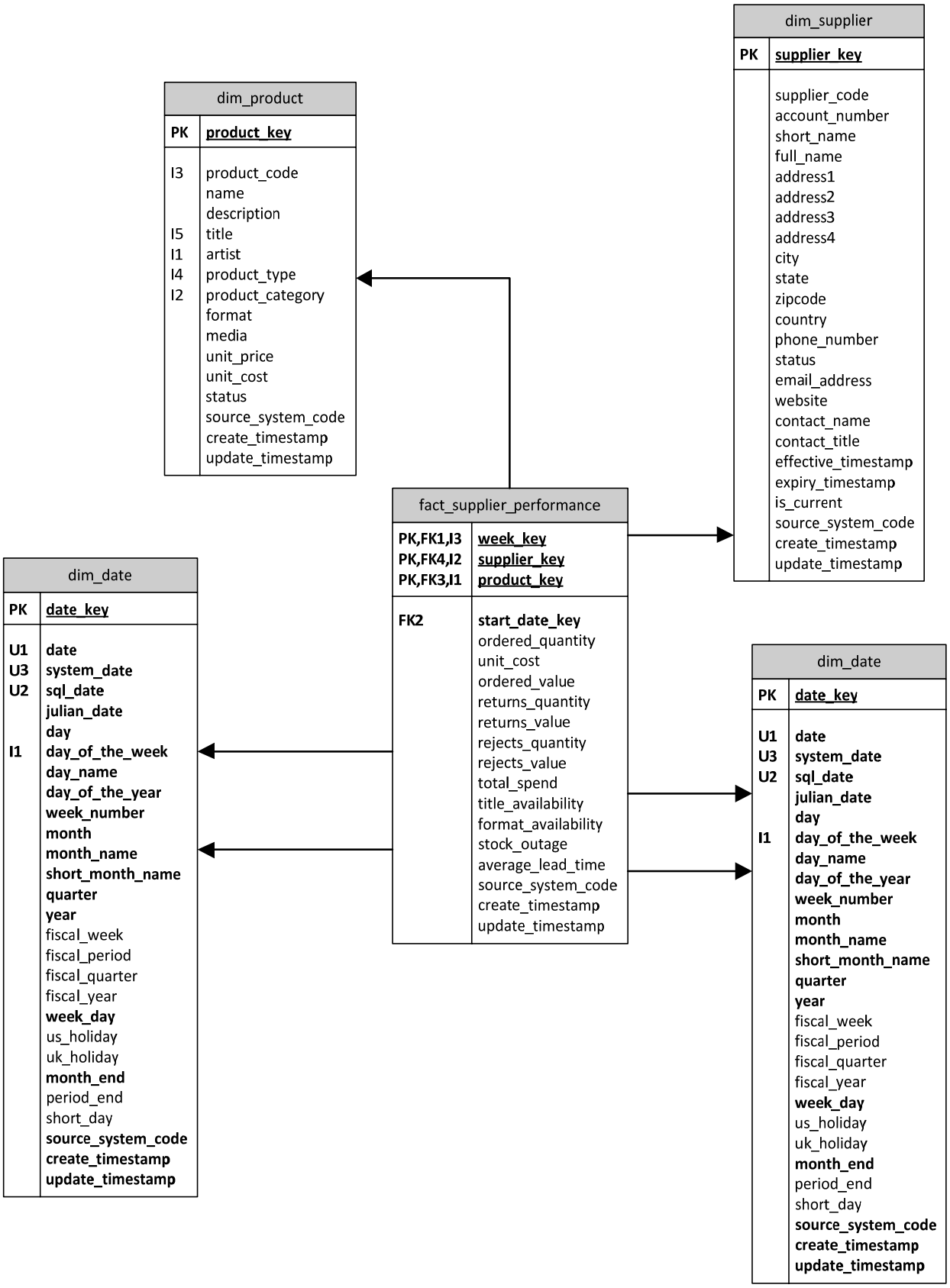


Figure 10: Logical Model of Fact Suppliers Performance

In this project, Data hierarchy was very important because of the relationship business wise between the data. Hierarchies are meaningful, standard way to group the data within a dimension so you can begin with the big picture and drill down to lower levels to investigate anomalies. According to Kimball, hierarchies are the main paths for summarizing the data. Data hierarchy is an arrangement of data consisting of sets and subsets such that every subset of a set is of lower rank than the set. In the context of the data warehouse, it can be used to provide paths that can be use to roll up and drill down when analyzing the data. The data hierarchy is applicable to the dimension table and it allows for organization of data. We talk briefly about the hierarchy as it is related to the way report layout can be organized.

4.3 System Development

The system development stage can now be embarked upon after the actual understanding of the expectation of the business users has been captured.

4.3.1 Design of the Physical Database

Having designed the logical model of both the fact and dimensional tables, it is now time to design the actual database on the SQL server database management system. The script for the design of the database and the table would be part of figure 11. The actual design of the database is carried out by executing scripts in the management studio of SQL server 2008.

The main_key column is a primary key in the dimension table, and it is a foreign key on the fact table. This is known in the database world as referential integrity. The main key in the dimensional table are usually the surrogate key, they are unique and not null, it uniquely identify the record in a dimension tables. We made use of the surrogate because the data to each of the dimensional table are from different sources and there is need to have a unique key to identify the record. This is where the referential integrity is important.

Referential integrity is a concept of establishing a parent-child relationship between two tables, with the purpose of ensuring that every row in the child table has a corresponding parent entry in the parent table. Designing the actual database in the SQL server database management system, we will enforce the referential integrity.

Data warehouse is optimized for data retrieval and it is very important that users are able to run their reports as quickly as possible. In the data storage, it is good to have a database structure and right index. What is index? Indexes are the pointers to the record stored in a database. In the concept of the data warehouse, indexes are important and the help in the loading and data retrieval of the data warehouse. Indexing can significantly improve the query and loading performance of data warehousing.

In the Dimensional Data Store, we have fact tables and we have dimension tables. They require different indexing and primary keys. We will discuss dimension tables first and then the fact tables. Each dimension table has a surrogate key column. This is an identity (1,1) column, and the values are unique. We made this surrogate key column the primary key of the dimension table. We also used this surrogate key column the clustered index of the dimension table. The reason for doing this is because in the DDS, the dimension tables are joined to the fact table on the surrogate key column. By making the surrogate key a clustered primary key, we will get good query performance because in SQL Server a clustered index determines the order of how the rows are physically stored on disk. Because of this, we can have only one clustered index on each table.

For the fact table, we used two approaches to the indexes; they are creating a surrogate key this is usually by creating an identity column in a table or uses a combination of the keys which is called a degenerate key. In the project we have chosen the second option because it can uniquely identify the record in a table.

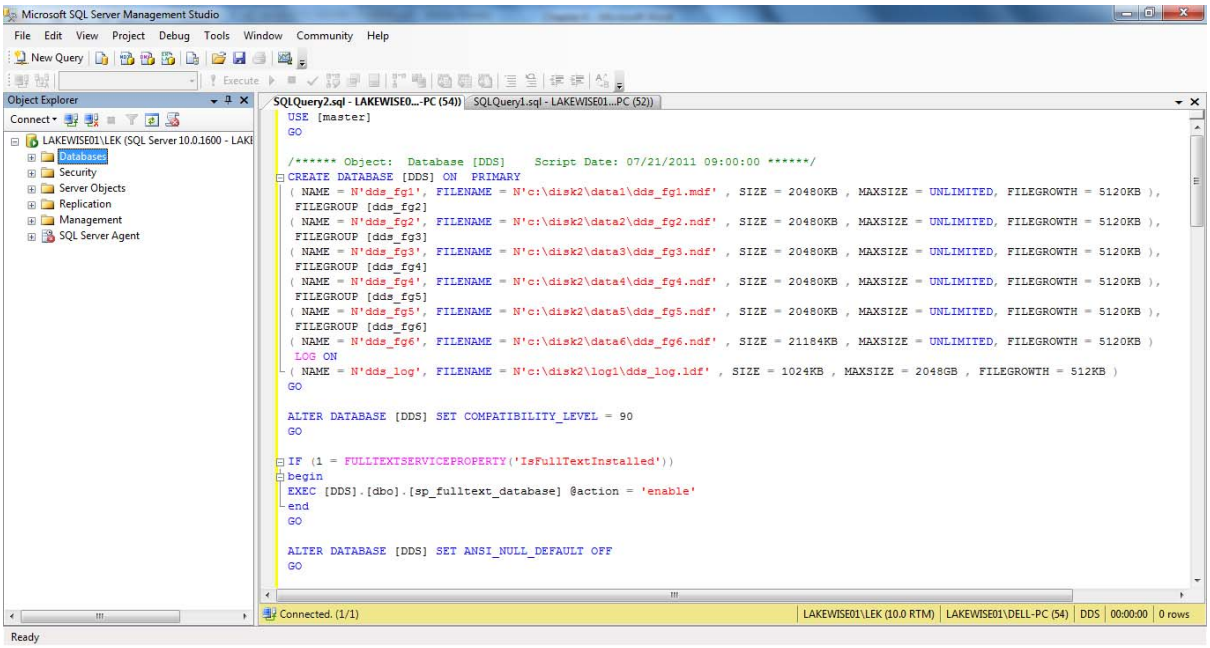


Figure 11: Query View of SQL Server Management Studio 2008

For this project, we have decided to use 3 instances of SQL server to diagrammatically represent a physical box of SQL server as shown below;

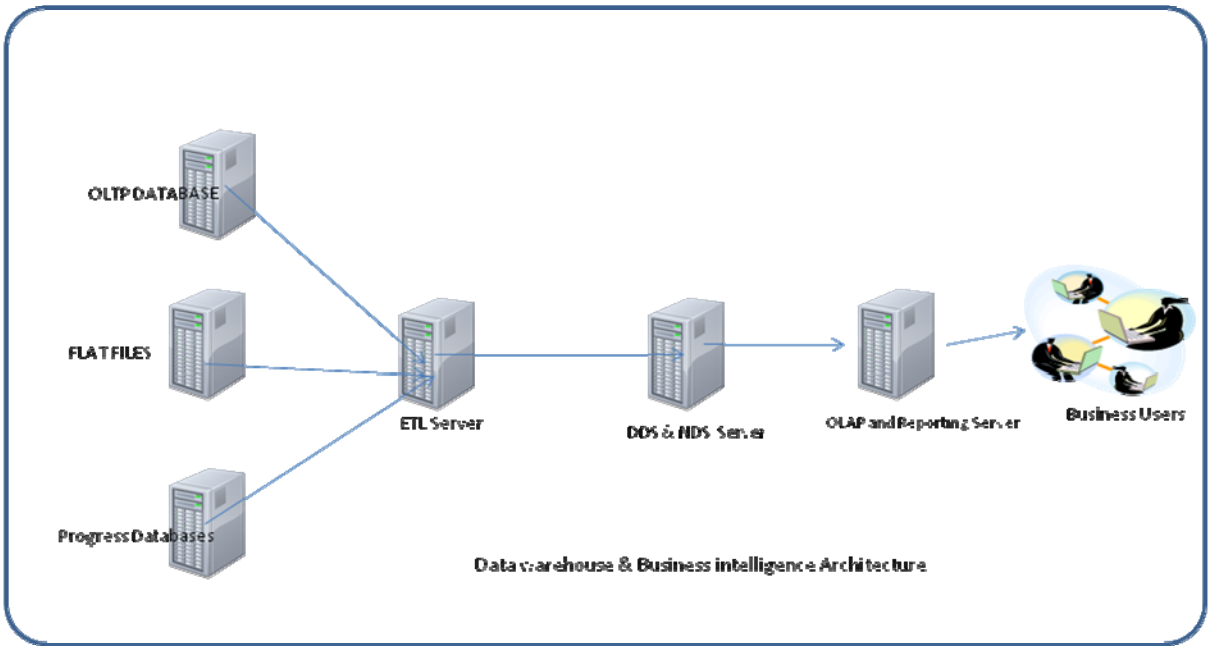


Figure 12: Data Warehouse and Business Intelligence Architecture

From the architecture above, the database engine can be installed on the Dimension and Normal data store as an instance. This is where the database and tables would reside after the

ETL process would have extracted the data from the different sources. The physical design of the data marts are as follows;

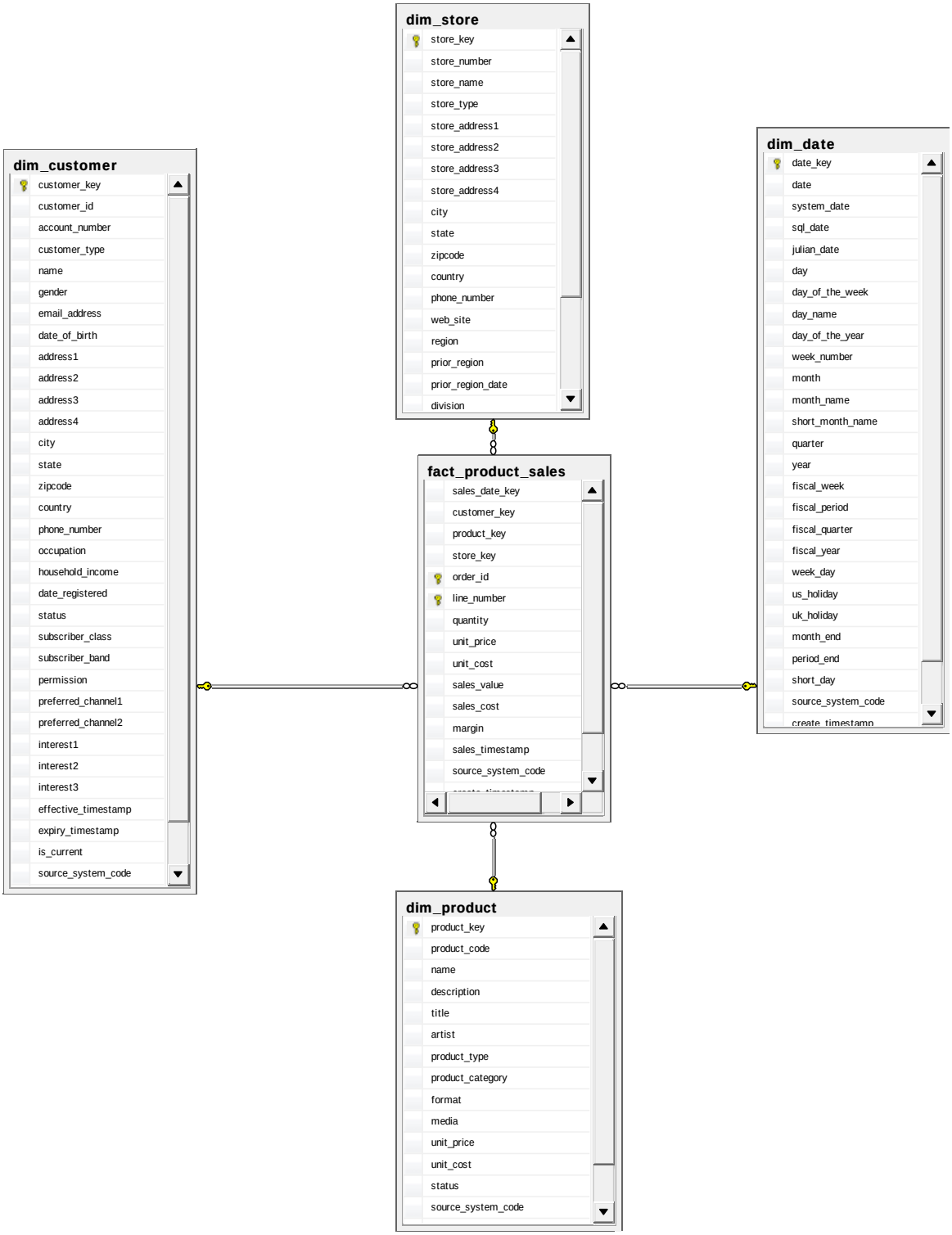


Figure 13: Physical Design of the Fact Product Sales Data Mart

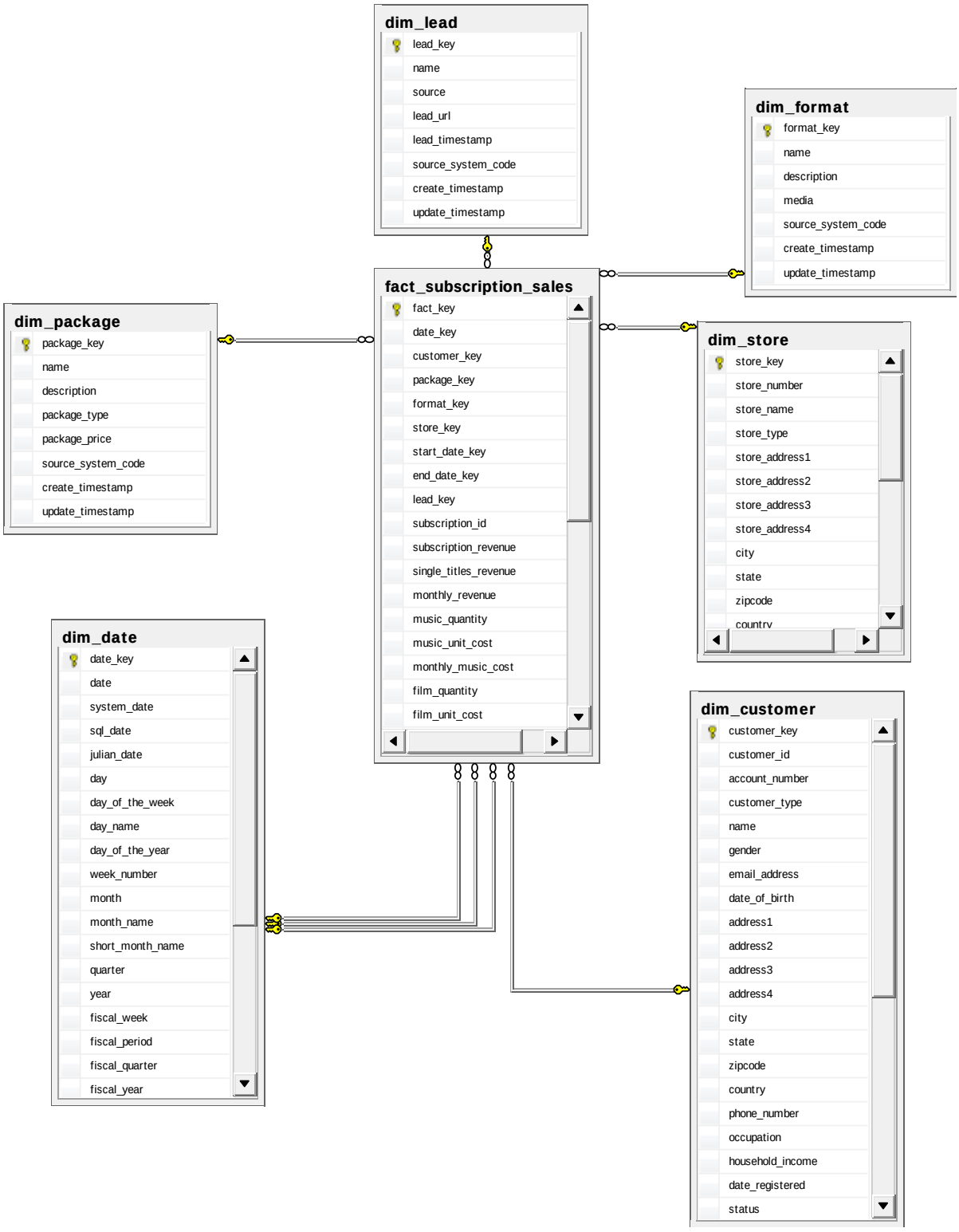


Figure 14: Physical Design of the Fact Subscription Sales Data Mart

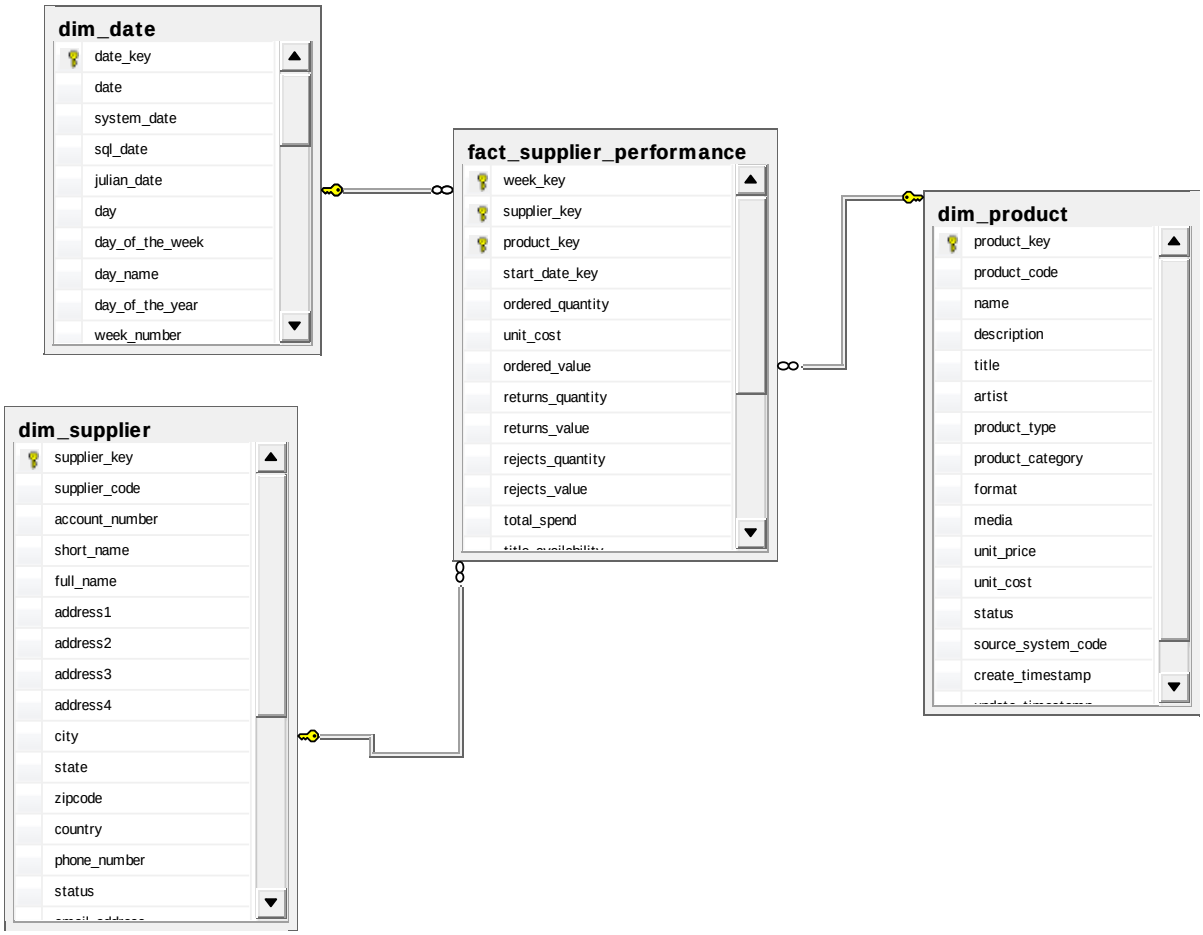


Figure 15: Physical Design of the Fact Supplier Performance Data Mart

4.3.2 Design of the ETL Process

This is one of the crucial processes of the data warehouse once the design has been completed. We can extract the data from the sources and load them into the Normal Data store and Dimensional Data store. It is one of the processes that needs to be carefully done so that the right data can be extracted.

ETL stands for Extract, Transform and Load. As mentioned in the beginning of the project that the company has different OLTP databases to extract from including the flat files in excel format. It is the process of retrieving and transforming data from the source system and putting it into the data warehouse. With the scope of this project, we used SQL Server Integration services (SSIS) to design an ETL process in order to load the data into the data store.

Before we go into the design of the ETL, we have decided to go with the architecture where the ETL will pull the data from the sources and push it to the staging for transformation before finally push and load it into data warehouse. The Diagram below explains the architecture

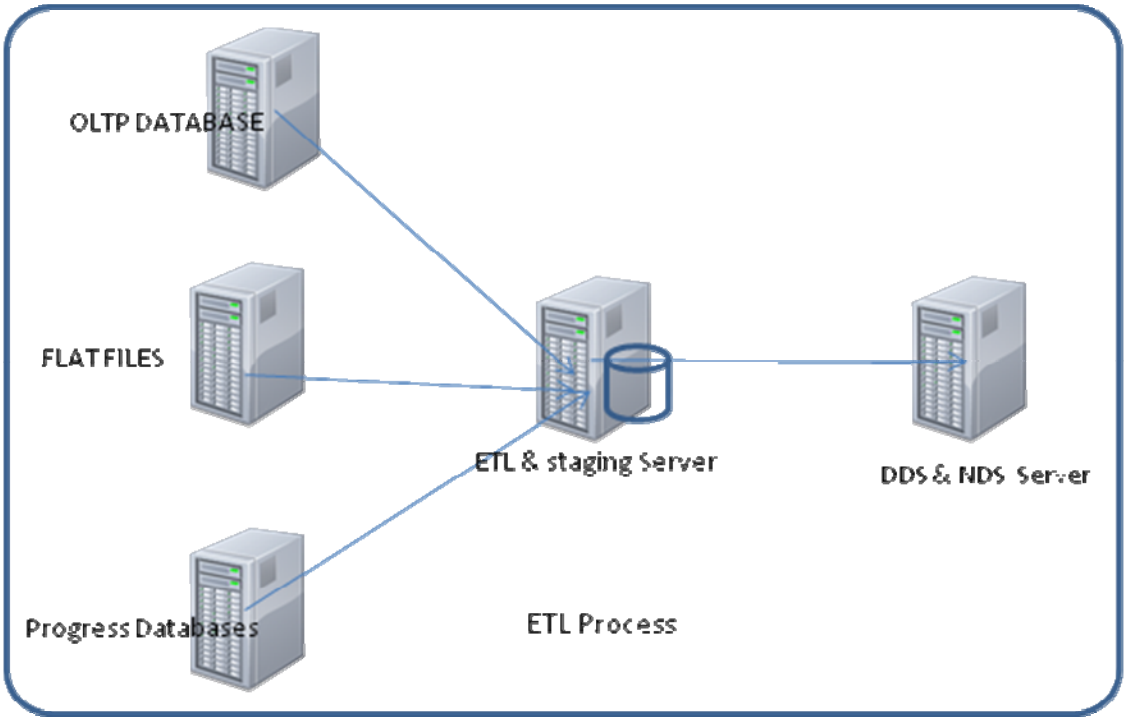


Figure 16: Extraction, Transformation, and Load (ETL) Architecture

Now that we have designed the ETL architecture, it is now the time to design the actual ETL process. The actual ETL design for the data load from sources to the staging database was design using SSIS, the diagram below represent the three loading process for the staging, They are Stage Ad-hoc Full Load, Stage Weekly external data and Stage daily full Re-Load.

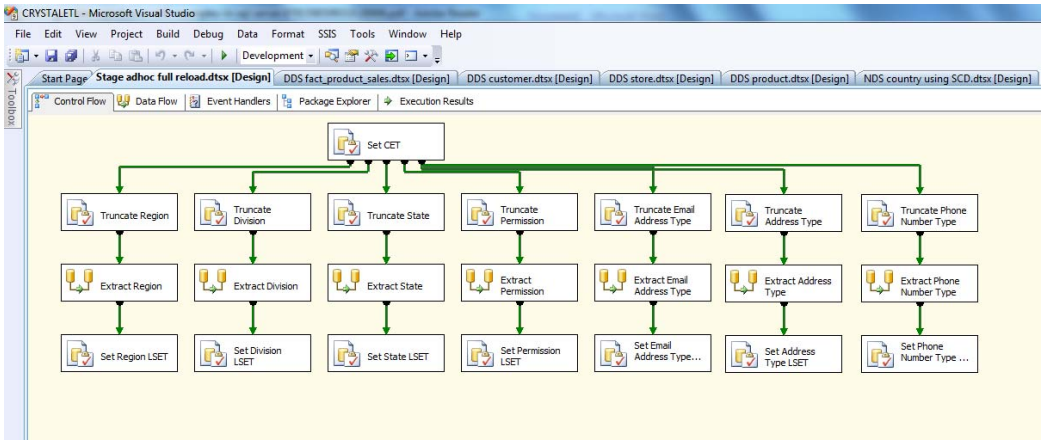


Figure 17: Stage Ad-hoc Full Load

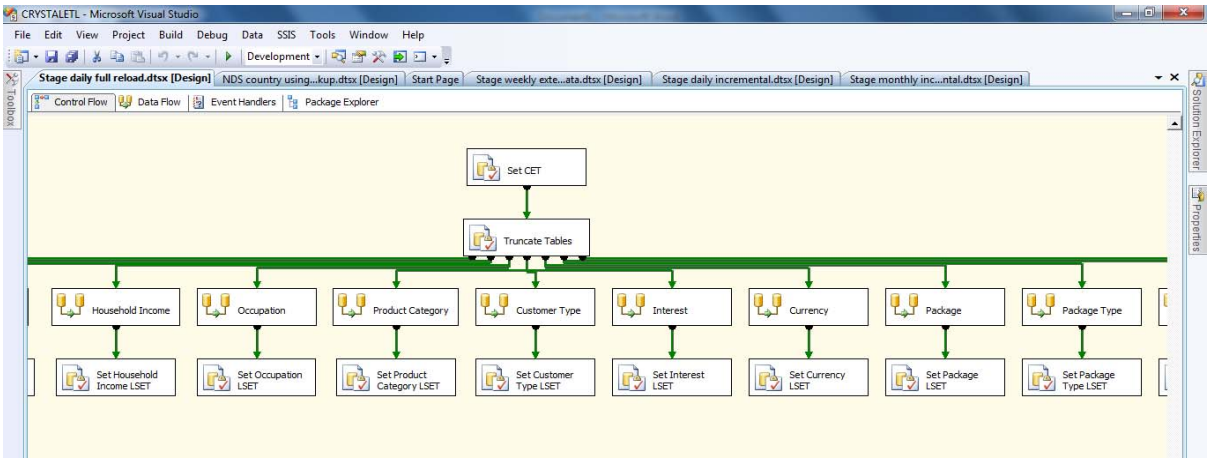


Figure 18: Stage Daily Full Re-Load

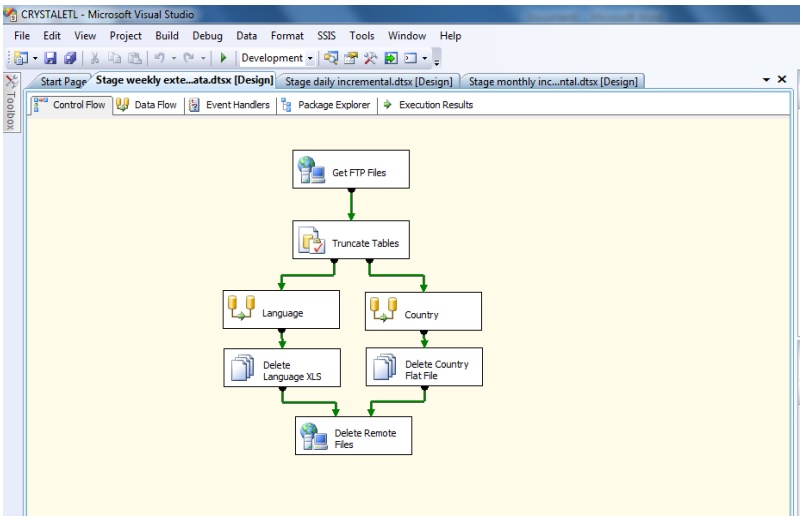


Figure 19: Stage External Data Load

4.3.3 Loading of the Data Warehouse

After the design of the fact and dimension table, it was then we loaded the tables that made up the data mart for each of the department we intend to analyze. While carrying out the logical design, we took a look at the relationship between the fact table and the dimension table. The two tables are linked together using the referential integrity.

After the extraction of data from the source system, we populated the normalized and the dimensional data store with the data we have extracted to staging databases.

Loading the stage database - The source data is loaded into the stage data base, the aim of the staging is to load the data without much transformation. The staging database is almost similar to that of the source system.

Checking the data quality when it is loaded from the source into the Normal Data store or operational data store is one of the ways to enhance data quality in DWH load. The check is based on the define business rule. Once the data quality and business rules has been applied to the data, we were able to load both the dimension and the fact table as it is required.

Loading of the dimension tables involve loading form Normalized Data store to the dimension data store. Renormalizations do to takes place in the DDS including the slowly changing dimension (SCD).

The loading of the dimension can be described using the flow chart;

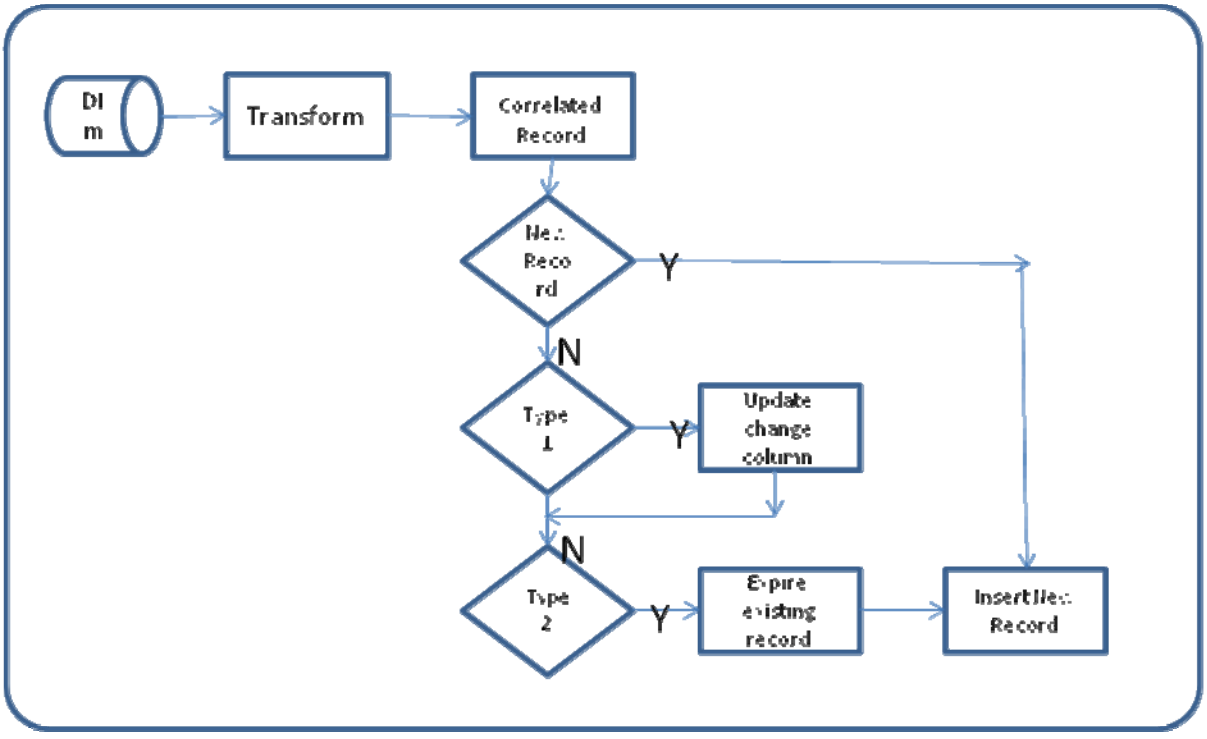
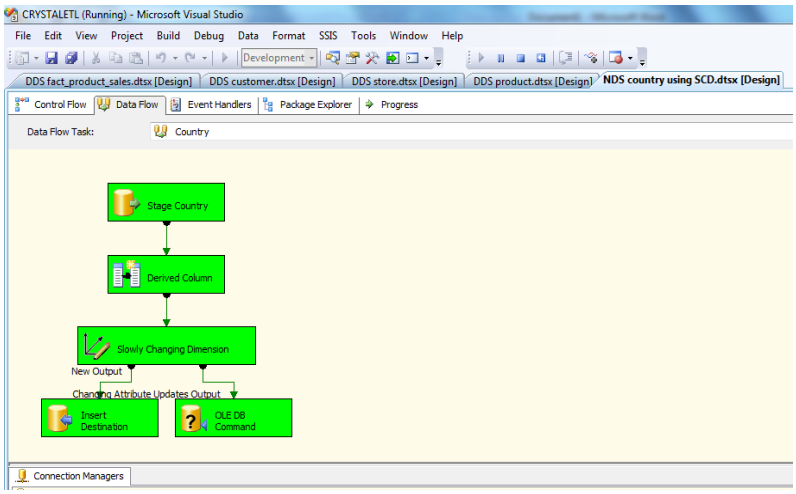


Figure 20: Flow Chart for loading the Dimension Tables

The following ETL design loads the data into the dimensional table;



Load the slowly changing dimension to preserve the history

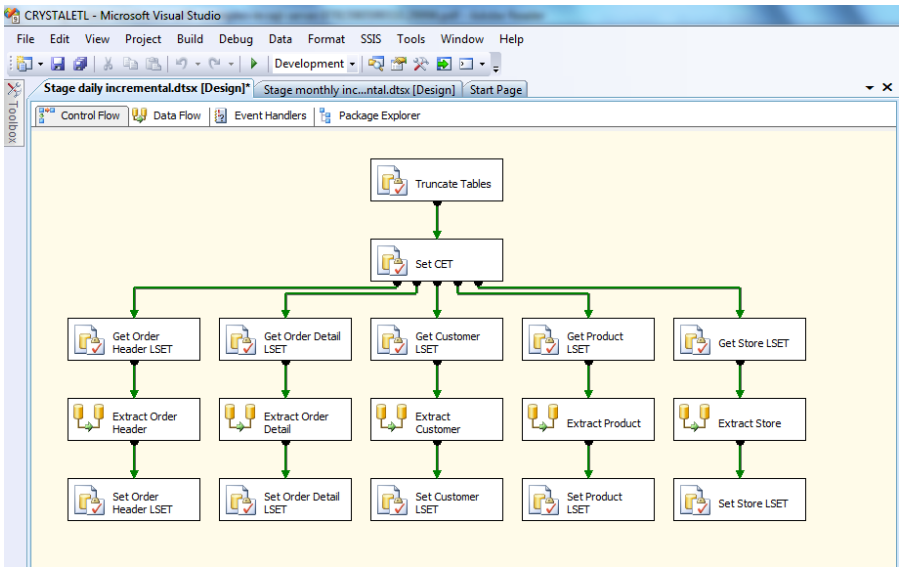


Figure 21: Daily Increment load of data into the data warehouse.

After successful loading of the dimension tables, it is now time to load the fact table which is the last step in the process of data warehouse loading. Data are loaded from the Normalized Data store and operational data store as the situation demands.

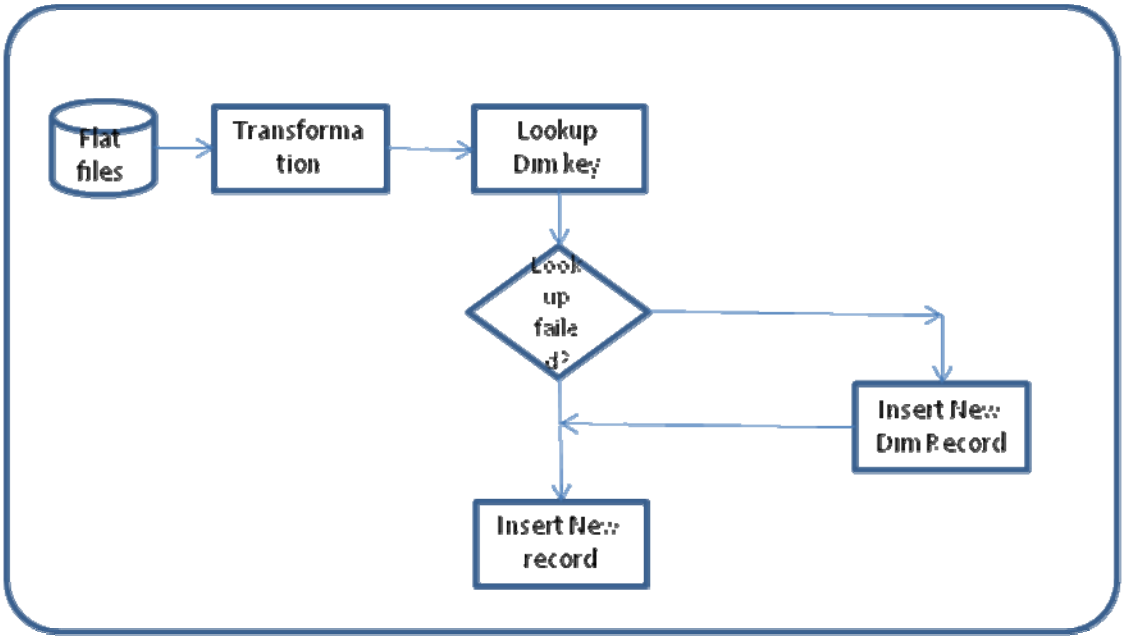


Figure 22: Flow Chart for loading the Dimension Tables

The data warehouse has been loaded with the right data of higher quality. It is now the time to make use of the data stored in the data warehouse. Business intelligence is all about make best use of the available data in order to make better decision about the company. This now takes us to the next; the loading of the DWH was carried out using the ETL tools. For this project, we used SSIS which is part of the SQL server application that we use for the database repository.

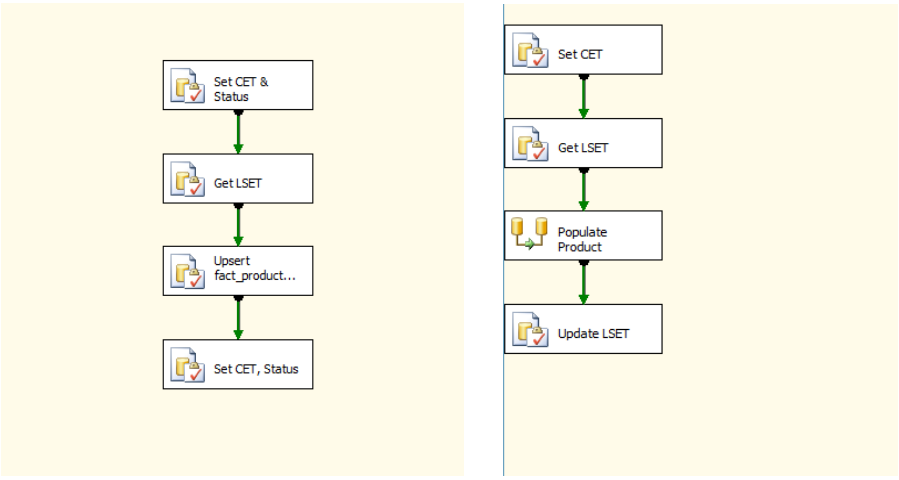
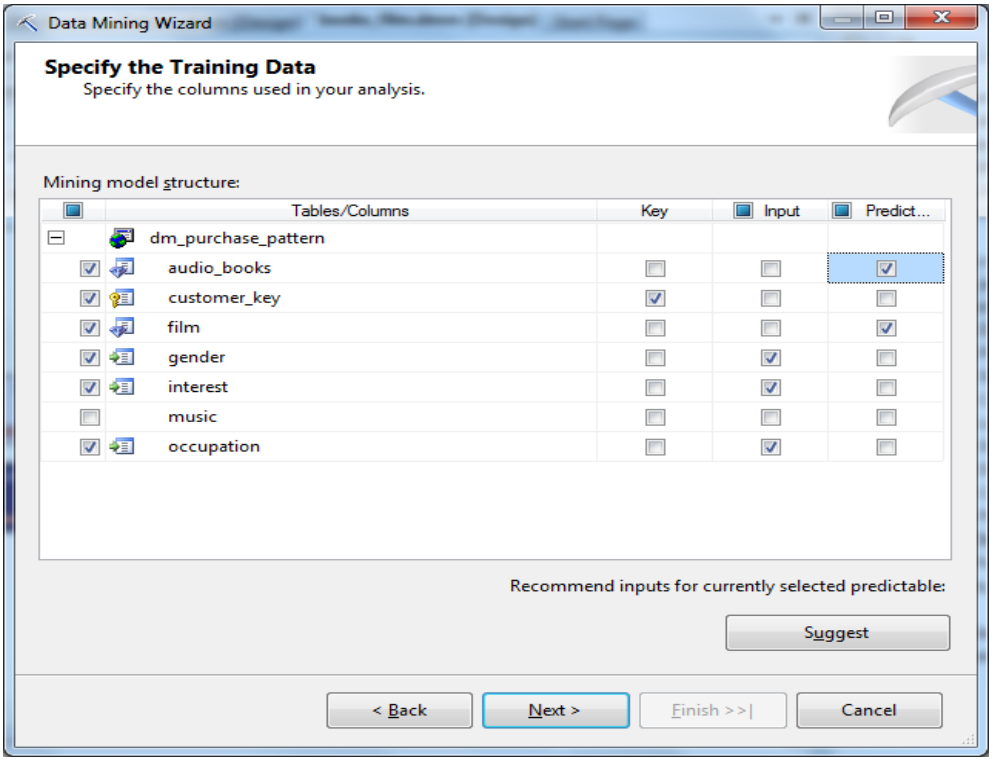


Figure 23: Fact Table Load

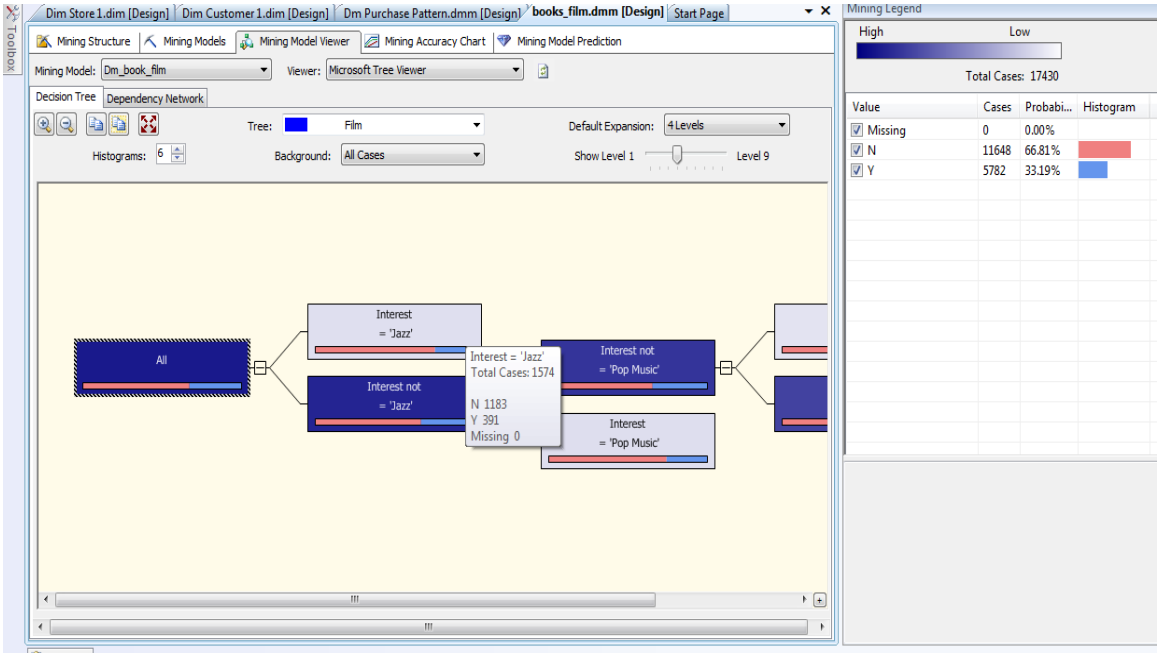
4.3.4 Data Mining

The aim is to see how we can predict the future of the business or show relationship between demography and action with the data we have today. After consulting series of references, we come up with this definition for data mining. This definition is from Kimball, he described data mining as a process of data exploration with the intent to find patterns or relationship that can be made useful to the business organization.

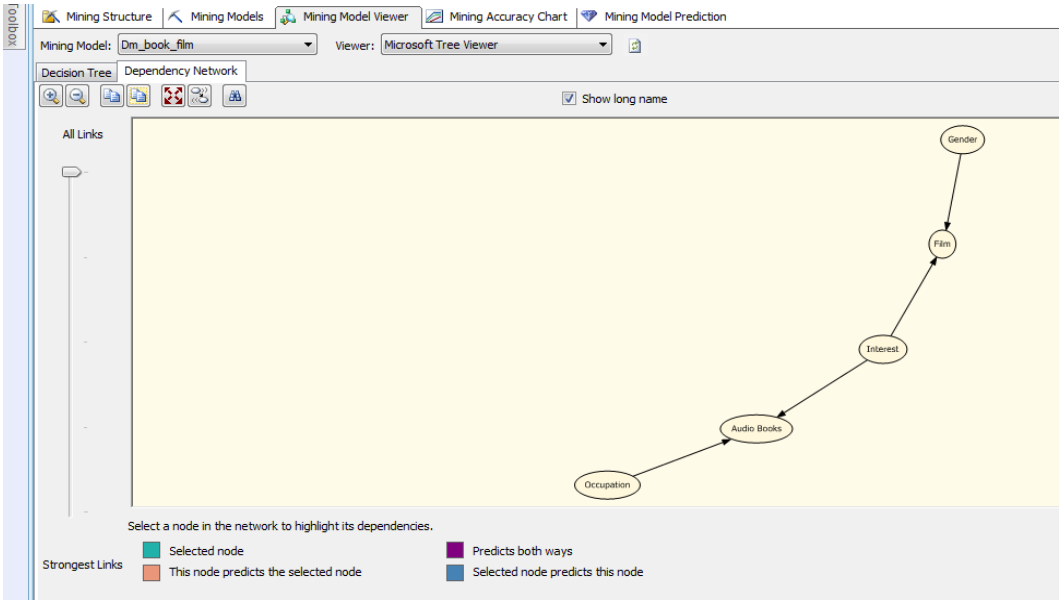
We intend to find relationship between audio books & film purchases and demographic attributes. We created a table to look at the purchasing pattern, the table look at the purchases of film and audio with the demographic attributes like interest, gender and occupation. We selected the table that contains the relevant information that we intend to use in the data mining process.



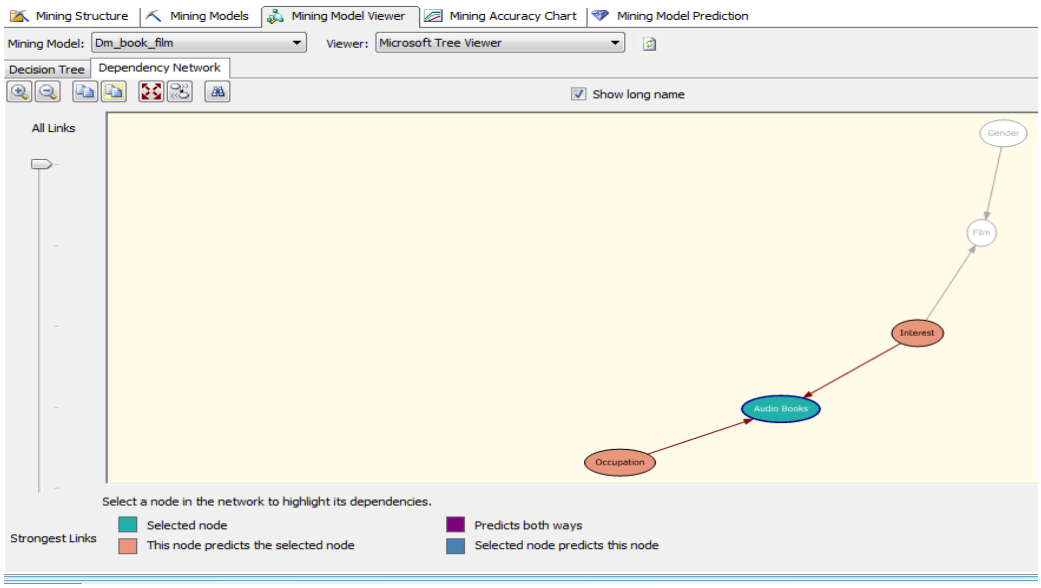
In order to use data mining, there is need for training of the data. We specified the columns to be used in the data analysis, the following colums were used. As for the inputs we selected the following demography column which are gender, interest, occupation and for the predictions audio books and film. Using the decision tree algorithm, it shows the number of people that would buy each of the items and those that would not be interested and what they are interested. The decision tree below shows the no of people interetsted in film and audio book.



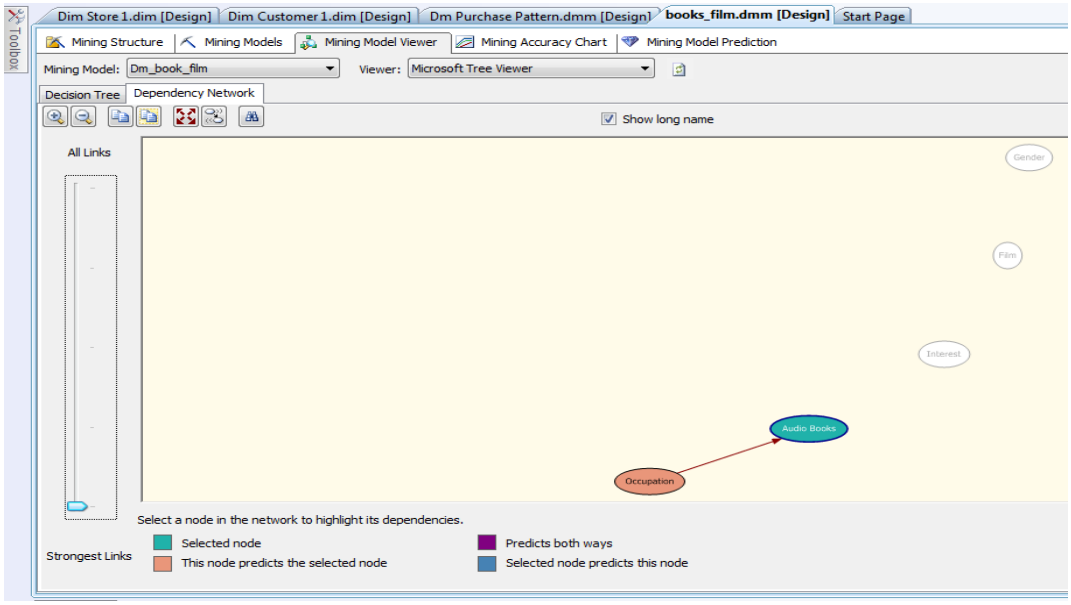
Further prediction was carried out using the dependency network, it is possible to view the relationship between demography and buying of audio books and film. The dependency looks as below, there is a link between the demography and the buying of book as film.



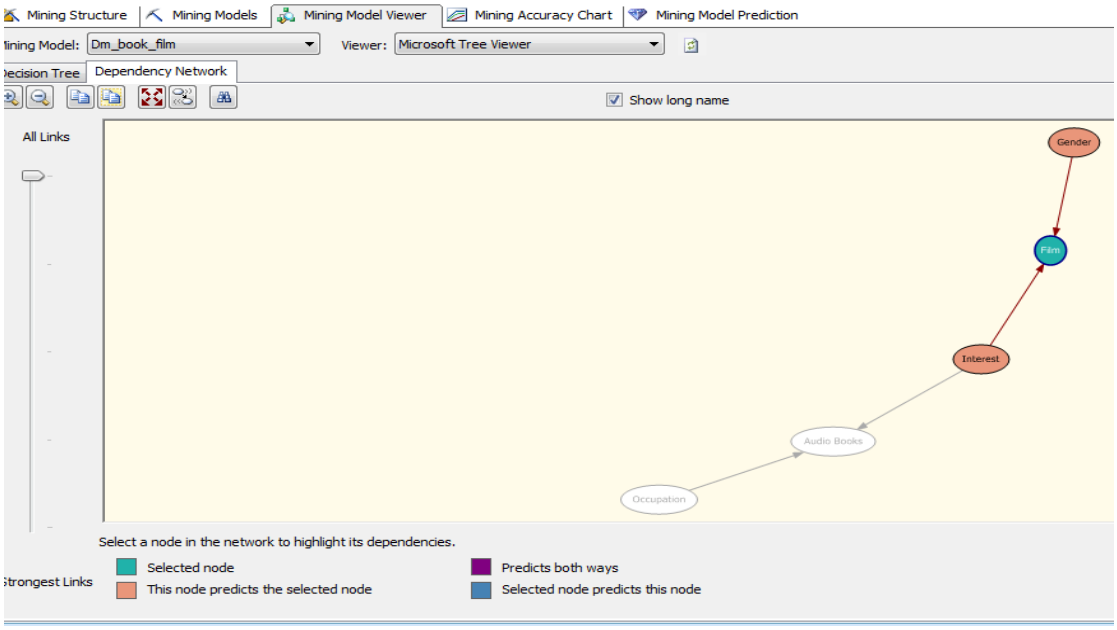
Selecting the each of the two nodes independently show the link between the demography and buying of either books or film, moving the slide down would show the strongest link.



When we selected the Audio books node, all links is strong between the occupation and interest, moving the slide down shows the strongetst between occupation and audio_book. The occupation of the customer would be a great link to sell book. The prediction shows that customer are likely to by books based on their occupation rather than interest.



As for the film, dependency network shows the link of the film as gender and interest. As we move the slide down to see how strong the link, it shows that the relationship between film and interest is stronger than the relationship between film and gender. Moving the slides to identify the strongest link, all the links were broken which shows tha the relationship are not strong when compared to that of the audio. Books. The two figures below show the situation in both cases.



What is business intelligence? It is the delivery of accurate, useful information to the appropriate decision makers within the necessary time frame to support effective decision making. Business intelligence is not simply fact and figures on printed reports. Rows upon rows of numbers showing detailed sales figures or production numbers may be extremely accurate, but they are not business intelligence until they are put into a format that can be understood by the decision makers who need them. In this project, we need to make use of the data stored in the data warehouse to present the decision makers with quality and accurate reports. We used the reporting and analysis services of SQL server to develop different type of report and data mining for the purpose of this project.

Using the analysis services, we browsed the data in multi dimensional to show that the user can make use of the data in the data warehouse and make a better decision on behalf of the company. The cube allows users to interact with the data in order to make view the result in different ways. This diagram below is showing customer's margin. It shows how much the company is making from a product by customer.

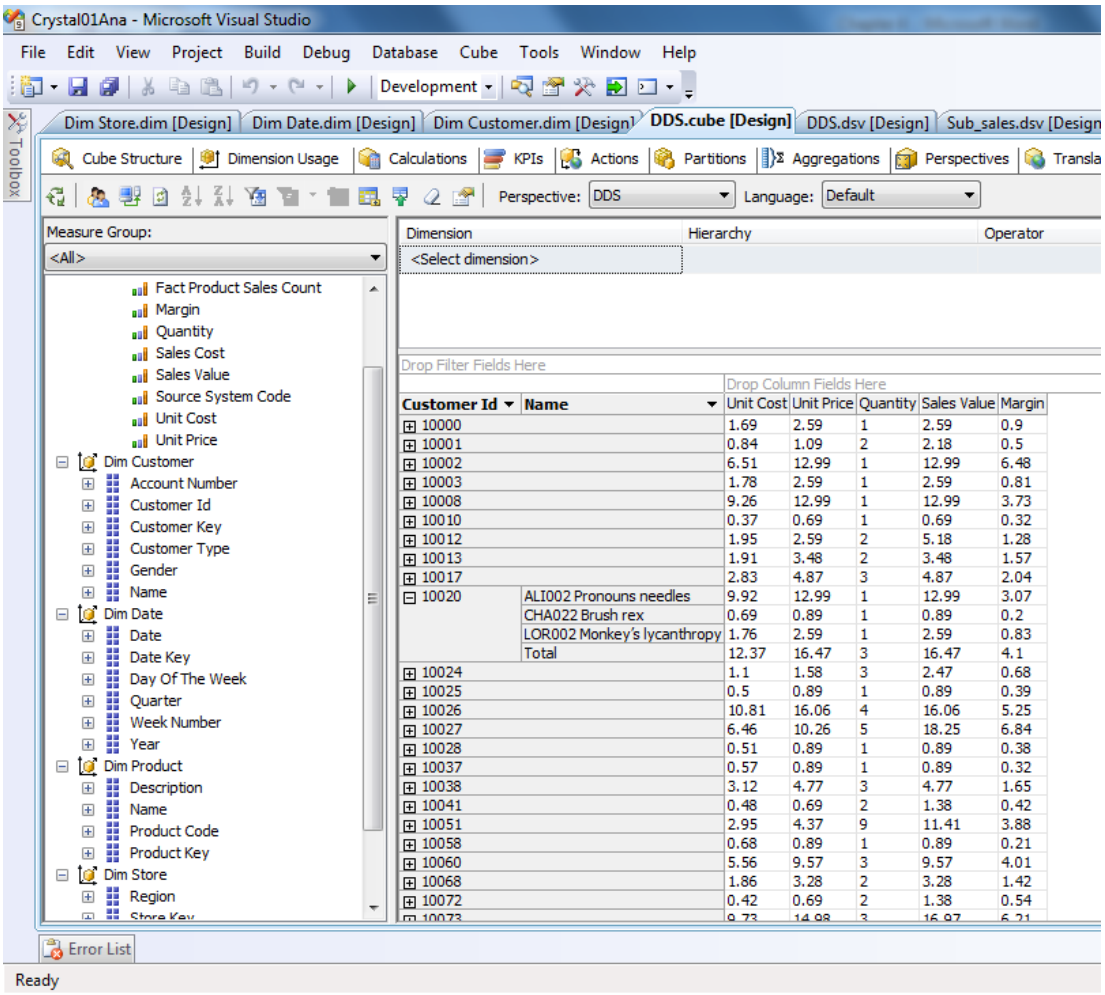


Figure 24: Analysis of Sales Report by Customer

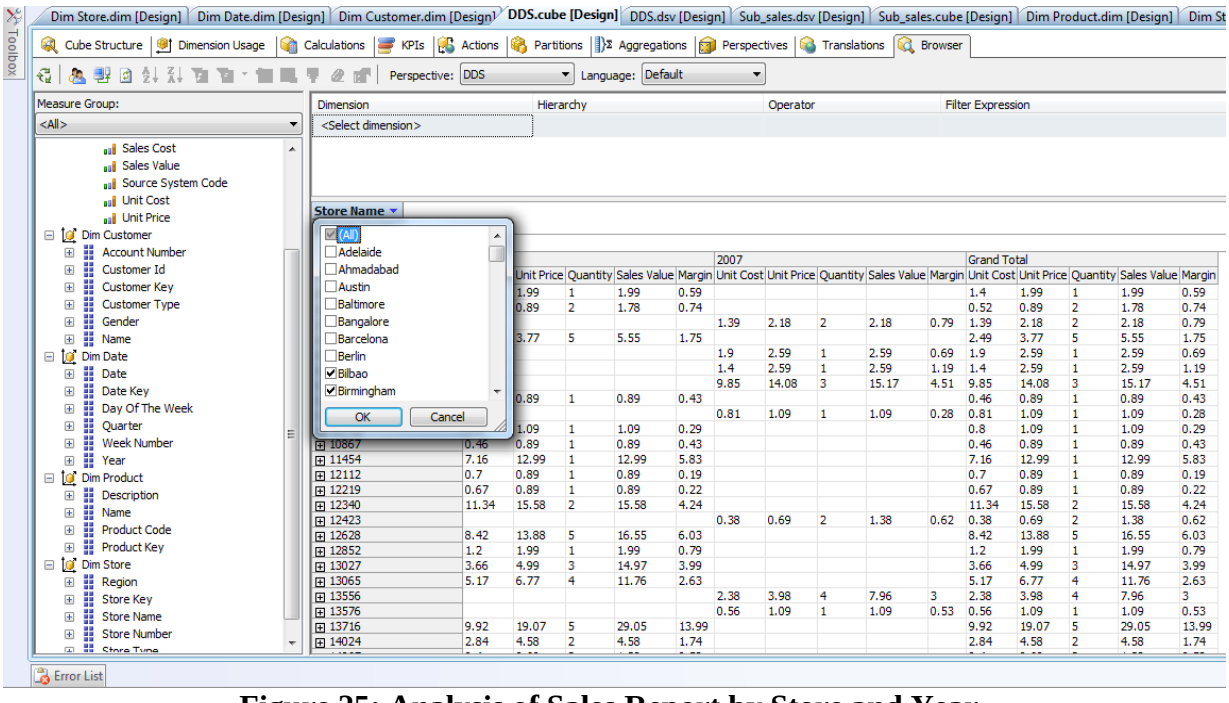


Figure 25: Analysis of Sales Report by Store and Year

The figure above shows the margin made from each customer and by product in a selected store Birmingham and Bilbao.

The end user can interact with the data to get different information required to take further business decision. In another interaction, the user can determine which store is performing and which are non-performing.

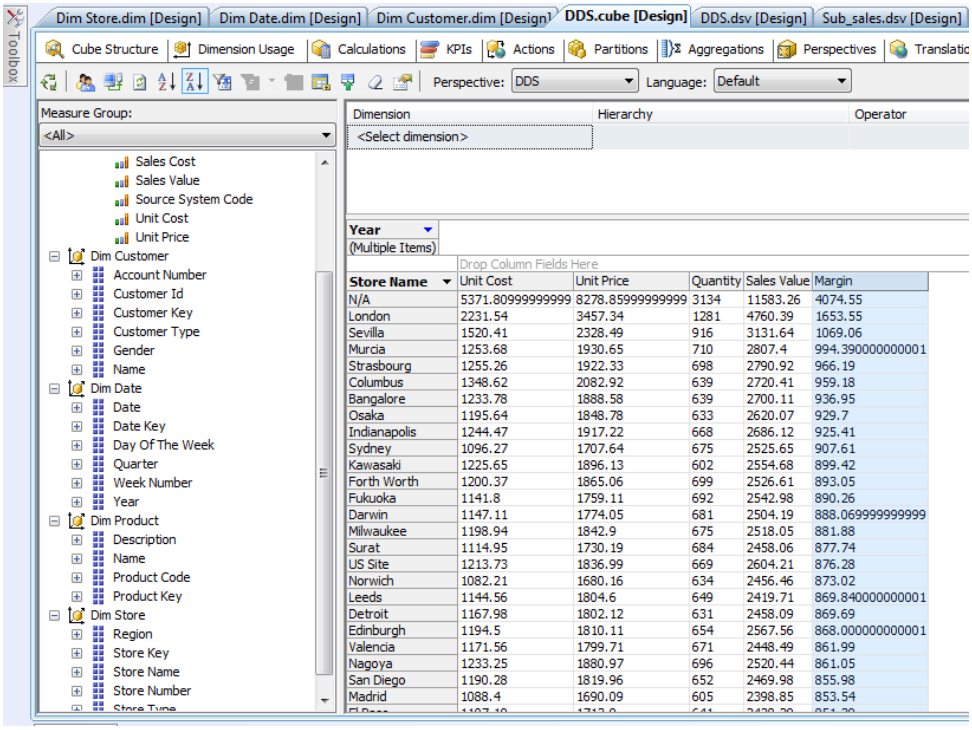


Figure 26: Analysis of Sales Margin by Store

From the analysis above, the London store has the highest sales margin; the store can be categorized as one of the leading store for the company.

In the previous section, we explained how we can load the data from different sources to the data warehouse, transform the data and carry out some analysis. Now we need to deliver the business intelligence to the end users and the executives that would make use of the information available. Creating the information is just the beginning, delivering the business intelligence to the end user is as important as creating the information.

Based on the business requirement, we developed some of the report which are standard report and are interactive which can help the user in day to day activity. The reports are;

Customer list: this would give the users the full details of their customer with a click of a button.

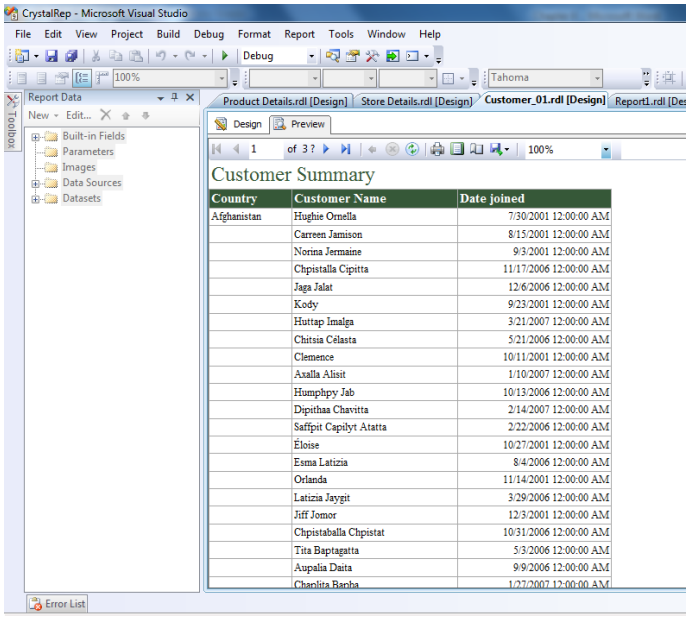


Figure 27: Customer List by Store

Price List: There would be a standard price list across the organization, any changes made to the price would reflect immediate once the ETL process has been process to load the data warehouse with up to-date data.

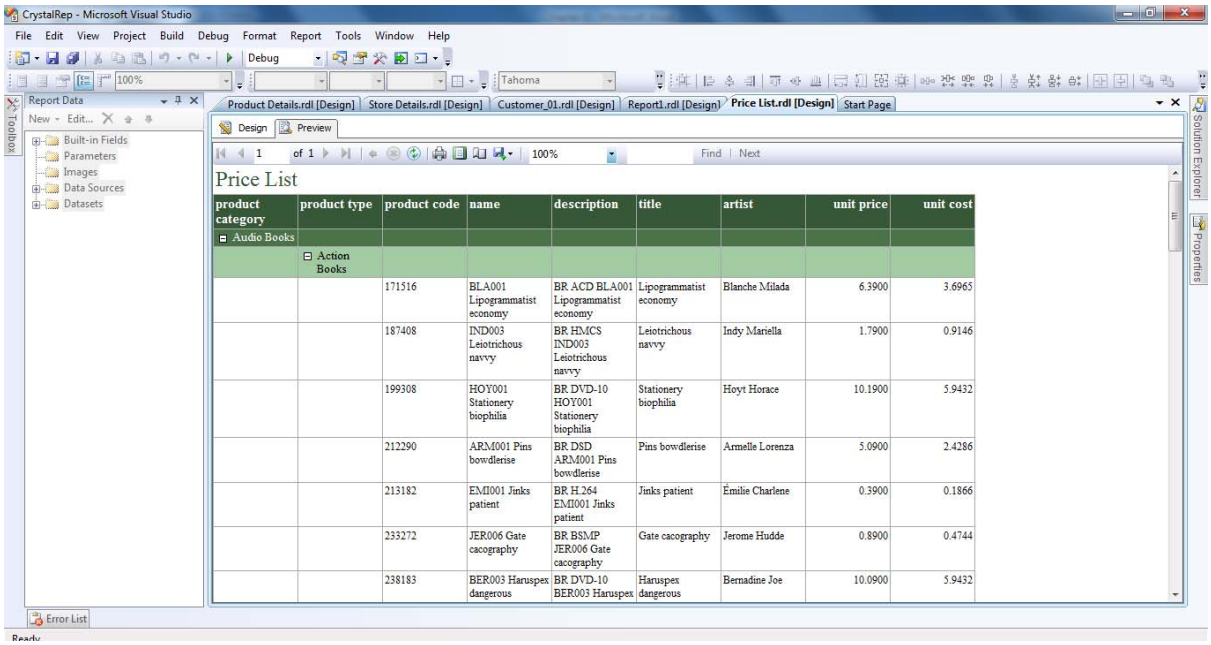


Figure 28: Price List by Product Category

Product Detail: The report presents the users with the list of product and the category according to the business definition. The user knows all the products that the company is selling. Here is the diagram showing the e sample report. When you drop down the list you get further product type under each category.

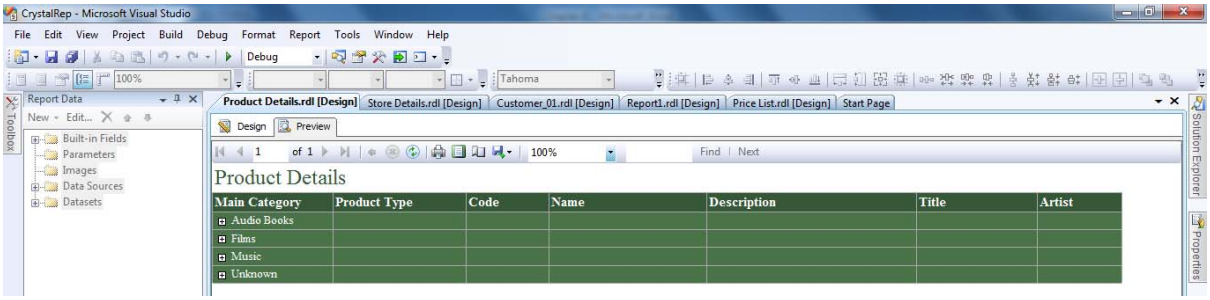


Figure 29: Description of Product Detail -1

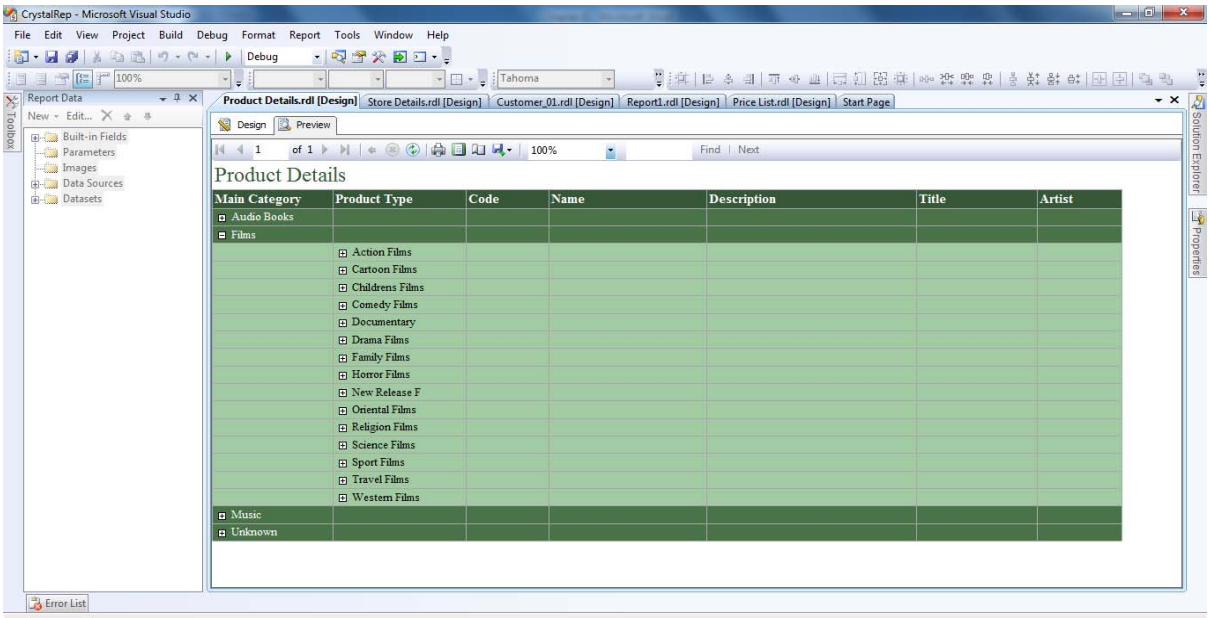


Figure 30: Description of Product Detail -2

Store List: This report consists of all the stores that the company is operating business from; the user can refer customers to the nearest business operating office.

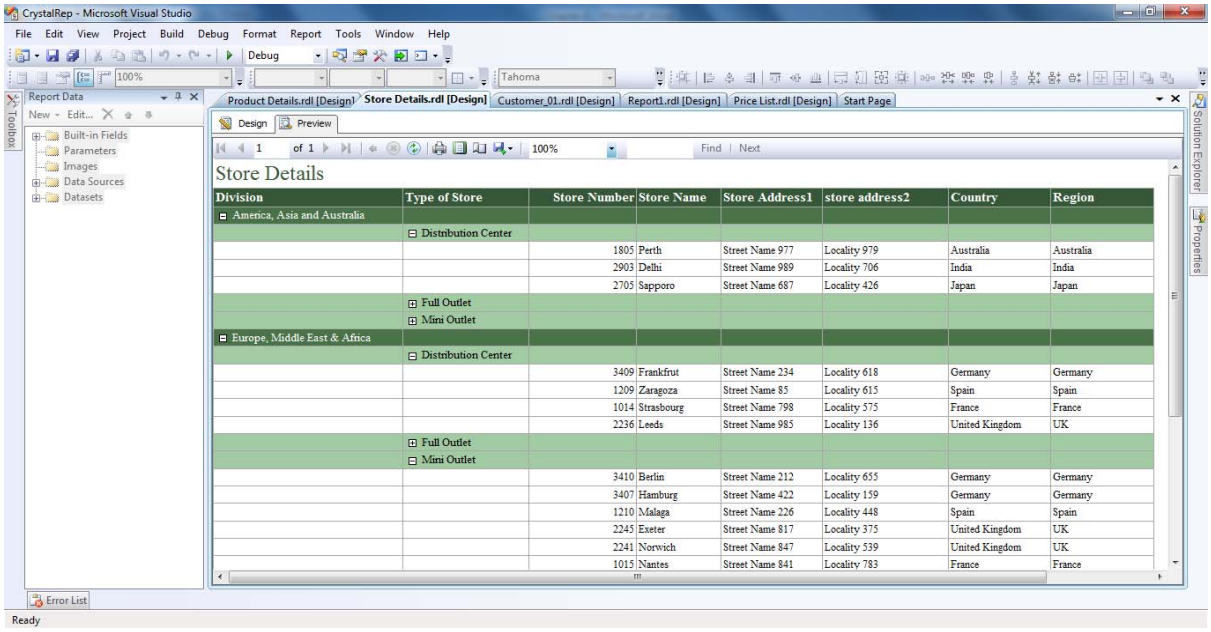


Figure 31: Store List Details

Weekly Sales Summary: The decision maker's use can see the details of the sales on weekly basin. From the report, the user select by year, quarter and business division. The diagram below shows the report;

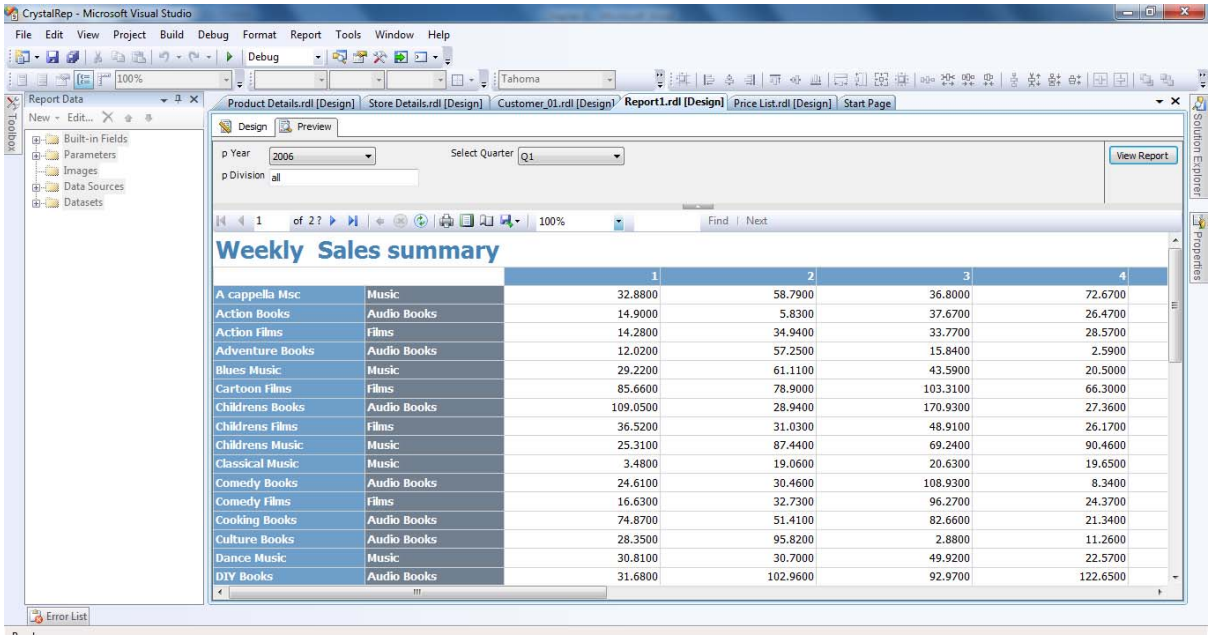


Figure 32: Weekly Sales Summary Report

4.3.5 End User Application

The system is designed with mass users in mind, it allow user to use familiarized tools such as Microsoft Excel to connect to the data warehouse and other data analysis system in order to make a better decision; which was one of the user requirements gathered when complying the business needs from the engaging the business users.

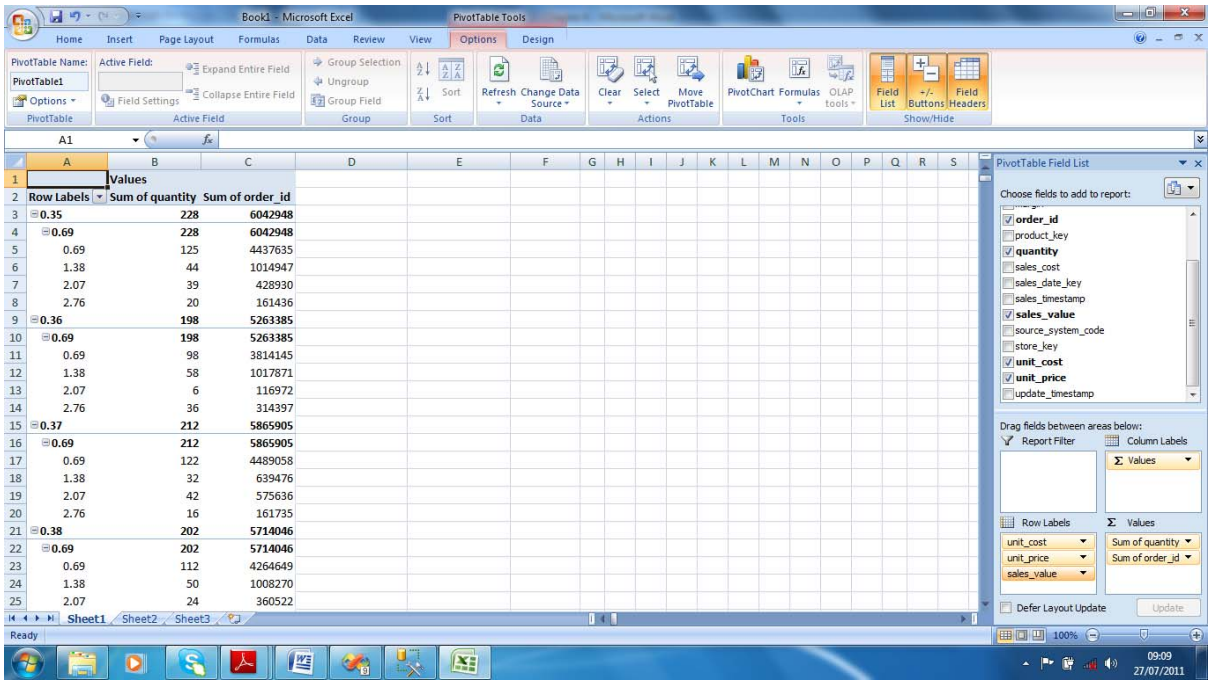


Figure 33: End User Application Tool

CHAPTER 5

5.0 Conclusion, Discussions and Limitations

The project focused on designing and implementation of DW and BI system for a retail industry using Crystal Music as case study. Effort were been made to consider all the principles of Data warehouse and business intelligence during the course of this project. We have shown how data can be integrated from different sources to a single repository called data warehouse which is used for delivering business intelligence to the end users and executives. We have succeeded in designing how data can be made available to business for day to day activities of their businesses.

We have developed data analysis template that user can interact with to get an immediate answers to the business question. We have been able to develop standard report for the business users. The reports can be generated with click of a button.

5.1 Limitations

In the course of this project, we encountered series of problem related to design and scope. We were able to put too many into a small hole without compromising the quality of the project. Due to the time and resources required to carry out further analysis, we have limited our design and analysis of data to customer and product sales area. The topic is very wide in the real world, in the context of academic, we have been able to show our understanding in this area and future enterprise project is possible from where we stopped.

5.2 End-User Survey

This end-user survey in Appendix A was conducted as part of Requirement Analysis phase. The online survey questionnaire link can be found in was given as http://www.snapsurveys.com/swh/surveylogin.asp?k=131115933830. The main goal of the questionnaire is to help us to understand the user expectation for the DW solution. It is also going to help us to understand the historical data that the organization gather, we then use this information to relate requirements with actual design. It consisted of 17 questions: 7

questions were text based and the rest 10 had answer -options. The options include Strongly Agree, Tend to agree, neither agree or disagree, Tend to disagree, and Strongly Agree. These answers were associated with the following values respectively 1, 2, 3, 4, 5. Figure 34 below gives the user interface and Appendix 1 gives the details questions.

The survey questionnaire served as supplements to the industry standards in which we followed and face–to face contact we had with managers and business user of the DW Systems; the input from the managers and business users and industry standards are more appropriate in the actual research work as we were proposing BI systems to Crystal Entertainments.

The response acquired from the online survey are hereby analyzed in the spreadsheet attached to this work; the analysis and conclusion of the survey helped to formulate the business case/need of Crystal Entertainment Inc.

Question Number	Strongly Agree (%)	Tends to Agree (%)	Neither Agree nor Disagree (%)	Tends to Disagree (%)	Strongly Disagree (%)
Q01	77.80%	22.20%	0.00%	0.00%	0.00%
Q02	77.80%	22.20%	0.00%	0.00%	0.00%
Q03	44.40%	50.00%	5.60%	0.00%	0.00%
Q04	58.80%	17.60%	23.50%	0.00%	0.00%
Q05	38.90%	44.40%	16.70%	0.00%	0.00%
Q06	50.00%	38.90%	11.10%	0.00%	0.00%
Q07	50.00%	38.90%	11.10%	0.00%	0.00%
Q08	61.10%	33.30%	5.60%	0.00%	0.00%
Q09	61.10%	38.90%	0.00%	0.00%	0.00%
Q10	82.40%	11.80%	5.90%	0.00%	0.00%

Table 1: Analysis of End-User Survey Report

From the Table 1; the respondent of the survey has proved that level of Data correctness in their DW is an important factor in the ability to make informed decision. 83% of the respondents support the view that since a DW is data driven, it is important that Data imported into the DW should be validated and is as close as possible to the real life situation. Trusted Data makes the world of Business Intelligence rosy. 78% of the respondents equally suggested that the DW should be able to integrate data from various sources is as important as

the close synchronization among functional area units within the business, so that it can help managers make decision on the overall working together of the units and how they affect each other. The response to question 8 also reiterates how significant the real world events is closely related in the data, about 62% of respondents points out. Furthermore, when data in the DW becomes larger, Able to handle increase in the number of simultaneous queries without impacting system performances. As they don't want a system that slows down their ability to make swift decision, as one of the demands placed on the system is to be able to provide data speedily with minimal impact of other users of the system. Other answers to the survey can be found under the spreadsheet Survey Results - End User.xls attached to the thesis

Question Number	Strongly Agree (%)	Tends to Agree (%)	Neither Agree nor Disagree (%)	Tends to Disagree (%)	Strongly Disagree (%)
Q10	82.40%	11.80%	5.90%	0.00%	0.00%
Q01	77.80%	22.20%	0.00%	0.00%	0.00%
Q02	77.80%	22.20%	0.00%	0.00%	0.00%
Q09	61.10%	38.90%	0.00%	0.00%	0.00%
Q08	61.10%	33.30%	5.60%	0.00%	0.00%
Q04	58.80%	17.60%	23.50%	0.00%	0.00%
Q06	50.00%	38.90%	11.10%	0.00%	0.00%
Q07	50.00%	38.90%	11.10%	0.00%	0.00%
Q03	44.40%	50.00%	5.60%	0.00%	0.00%
Q05	38.90%	44.40%	16.70%	0.00%	0.00%

Table 2: Users' Priority Table of End-User Survey Report

Table 2 priorities the responses by question number; the respondent strongly agrees that the level of data correctness in DW is most important element in the users' requirement analysis and the do not strongly agree that they need to be involved in the design (and implementation) of the system is as important. Instead, they tend to agree that they like to be involved in the design (and implementation) of any system that will impact their job or make them do the jobs better by giving it a whooping 45% vote. They also tend to agree most that they should be able to access the data more easily and quickly, and able to extract singular reports that can make them make decision quicker, it got the highest tend to agree of 50% of their support.

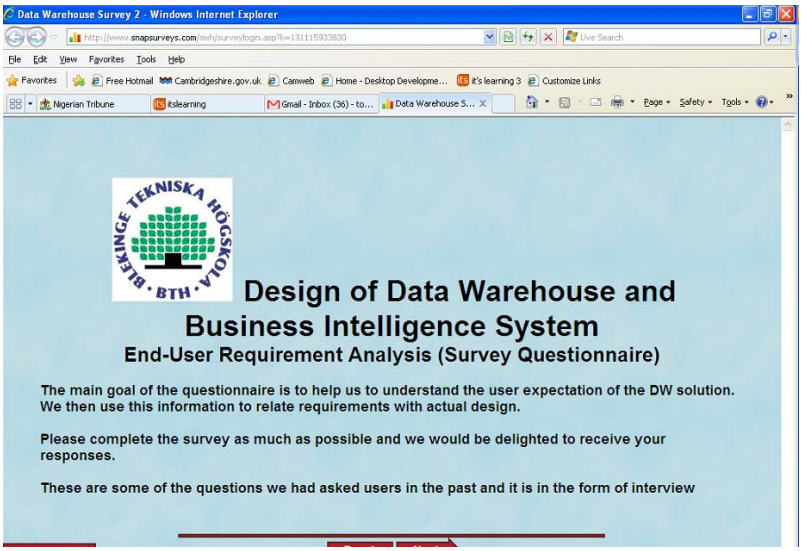


Figure 34: End-User Requirement Analysis Screenshot

In putting together this question to capture the business requirement – which we have concentrated on the decision makers as a scope in this project, this helps us to limit our scope in the design to such business area. The entire business is to be taken into consideration; Distribution Channels, Sales & Marketing, Accounts & Finance, Productions/Operations, Customer Support business units as information from each unit is needed in the forecast for other units.

Hence, our interview was directed at Managers and decision makers, to ask them question about their jobs, their objectives, and their challenges, and what they would like DW/BI system do for them what they have not been able to do with any legacy application (if any). It was carried out on an individual basis where the respondent would answer Strongly Disagree to Strongly Agree choices.

5.3 Data Warehouse Evaluation

At the end of the design, we thought it might be necessary to ask questions concerning general perception of the success of the data warehouse system. Appendix 2 detailed the questions that help to evaluate the design of the DW/BI System. Here is a link to the DW Online evaluation question http://www.snapsurveys.com/swh/surveylogin.asp?k=131057257997. The general perception of the measuring of success in the DW implementation is such that decision makers notice the ease of use offered by the system and are happy to use the data warehouse for reporting business intelligence. Furthermore, the evaluation results should also

prove that another factor that was quite important to the decision makers was the easy access to in-depth information that can be generated by the system such as sales summary etc

Figure 35: Data warehouse Evaluation Screenshot

This evaluation report is quite key to support the ongoing research in business intelligence that Data Warehouse system are a tool in the hand of Managers/Owners of Business in making insightful and informed decision. Hence, the response is just an industry acceptable standard for the use of DW.

5.4 Conclusion

We have been able to demonstrate that the data warehouse and business intelligence can be implemented by any organization and that the project should have a clear aim of supporting both business users and executives. We have based the design on three major units within the organization to develop a data mart. It is possible to design an enterprise wide data warehouse from the data mart. The principle we had followed can be adopted for any retail company. The project is scalable industry wise. In the future, possible integration of the business intelligence window into the collaboration tools is possible.

Businesses have stored operational data for long, and they have continued to amass huge amounts of data at a massive rate as the operational databases becomes more valuable, it helps grow the rate at which business succeed. We have by this thesis shown that the data warehouse collect, consolidate, organize, and summarize this structured and unstructured data so that this intelligent data can be used to inform business decisions.

We have come to see that the key to data warehousing is data design. A good data design creates a workable data warehouse. The industry users know what data they need and how they want to use it. We focus on the users to determine what data is needed, locate sources for the data, and organize the data in a dimensional model that represents the business needs. Data analysis and reporting are not enough to leverage the benefits of business intelligence in such a dynamic industry. The required, needed and right information should also find a way to get to the right people in ample time. But importantly, the information must be acted upon to activate the benefits. It is expedient that business intelligence should be part of the business process, within the management sector. BI must not only help to understand the past, but also work to find new opportunities and emerging trends in the future!

6.0 References

1. Alan, B, Emma B, (2007), 2nd Edition Business research methods, Oxford, Oxford Press pp.126 – 148.
2. Balaji, P. & Alexander T. (2003). On the Use of Optimization for Data Mining: Theoretical Interactions and ECRM Opportunities. Management Sciences, New York, INFORMS 479(10), 1327 – 1343
3. Ballard Chuck, et al (1998*)* Data Modeling Techniques for Data Warehousing IBM International Technical Support Organization; Pgs 36-37
4. Başaran, Beril P (2005)**,** a Comparison of Data Warehouse Design Models, the Graduate School of Natural and Applied Sciences, Atilim University
5. Başaran, Beril P (2007), Developing a Data Warehouse for a University Decision Support System, The Graduate School of Natural and Applied Sciences, Atilim University
6. Burton, P et al (2010), Meta Data: The Key to Data Warehouse Design (A Systems Engineering Approach) ENSE623 Project Institute of System Research
7. Chaffey, D. (2002). E-business and E-commerce management. New York Prentice Hall, p 330-370
8. Chan S. S., (1999). The Impact of Technology on Users and the Workplace. *New Directions for Institutional Research*. Volume 1999, Issue 103. pp 3 – 21
9. Cho V. and Ngai E.W.T. (2003). Data Mining for Selection of Insurance Sales Agents. *Expert Systems*. Vol. 20, No. 3.
10. Cooper, H. (1998). Synthesizing Research: A Guide for Literature Reviews
11. Demarest, M.,(2008) "Data Warehouse Prototyping: Reducing Risk, Securing Commitment and Improving Project Governance"
12. Descombe, M. (2000), Good the Research Guide; for small scale-social research projects. Buckingham, Philadelphia., Open University Press
13. Drucker, P (1996) "The information executives truly need" Harvard Business Review, January – February, pp, 54-62
14. Eckerson. W., (2003) Smart companies in the 21st century: The secrets of creating successful BI solutions. (In A report of the Data Warehouse Institute
15. Eldabi, T., et al (2002), Quantitative and qualitative decision making methods in simulation modeling. Management Decision, Vol.40(1) p. 64-73

16. Fox R., (2004) Moving from data to information OCLC Systems and Services: *International Digital Library Perspectives* Volume 20 Number 3 pp 96-101
17. Giovinazzo, W (2002), 'Internet-Enabled Business Intelligence', Prentice Hall.
18. Goldstein, I., Benbasat, and Mead, M. (1987). The case research strategy in studies of Information systems. MIS Quarterly, pp86-95
19. Hackathorn, R (1999), 'Farming the Web for Systematic Business Intelligence', Proceedings of the Fifth ACM SIGKDD, San Diego, CA.
20. Han, J and Kamber , M (2001), Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann
21. http://blog.maia-intelligence.com/2010/01/18/analytics-in-media-industry
22. http://www.nos.org/htm/sad1.htm
23. Inmon W.H., (1993) *Building the Data Warehouse*, A Wiley QED publication, John Wiley and Sons, Inc. New York 123-133
24. Kakinda M. F., (2000). Introduction to Social Research.
25. Kimball R. and Ross M., (1996). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*.
26. Kimball R. and Ross M., (2002) *the Data Warehouse Toolkit: Second Edition, the Complete Guide to Dimensional Modeling*.
27. Krishna P. R. and Kumar S.D., (2001) A Fuzzy Approach to Build an Intelligent Data Warehouse. *Journal of Intelligent and Fuzzy Systems*
28. Krumenaker, M. and Bhattacharya, J (2003), User Implementation and Revision of Business Rules without Hard Coding: Macro-Generated SAS Code, Proceedings of the 16th Annual Northeast SAS User Group Conference.
29. Larson, B (2006) Delivering Business Intelligence with Microsoft SQL Server 2005, McGraw-Hill/Osborne
30. Lechtenborger, Jens (2001) Data Warehouse Schema Design, Dept. of Information Systems University of Munster Germany
31. Liautaud, B (2000) E-Business Intelligence: Turning Information into Knowledge into Profit, McGraw-Hill.
32. Ma C., et al, (2000) Data Warehousing, Technology Assessment and Management. *Industrial Management and Data Systems* Volume 100 No. 3 pp. 125 – 135
33. Mistry, R and Misner, S (2010) Introducing Microsoft SQL Server 2008 R2; Microsoft Press, Redmond

34. Mullins, Stuart (2007) Data Warehouse Data Model Design http://it.toolbox.com/blogs/dw-cents/data-warehouse-data-model-design-18699
35. O'Leary D.E. (1999). REAL-D: A Schema for Data Warehouses. *Journal of Information Systems* vol, 13. No, I pp. 49-62.
36. Orr, Ken (2000) Data Warehousing Technology, The Ken Orr Institute; revised edition
37. Scoggins, J (1999), 'A Practitioner's View of Techniques Used in Data Warehousing for Sifting through Data to Provide Information', Proceedings Of The Eight International Conference On Information and Knowledge Management, Kansas City, MI.
38. Stark, Dave (2005) Business Intelligence Plays a Major Role in the Entertainment Industry, http://www.b-eye-network.com/view/1952
39. Tapscott, Don (2008); Business Intelligence for the Retail Industry: Actionable Insights for Business Decision Maker, New Paradigm Learning Corporation
40. Peter, C Bryan et al (2005) Key Account Management in financial services, London, Kogan Page publisher.
41. Poe, V., et al (1997). Building a Data Warehouse for Decision Support 2nd ed., Prentice Hall
42. Rainardi, V. (2008), "Building a Data Warehouse", Apress, Springer-Verlag New York, Inc, USA
43. Yin, Robert K (2003) Applications of case study research, applied social research method series Volume 34. Pp 3-4
44. Zeng, Y et al (2003) Enterprise Integration with Advanced Information Technologies: ERP & data warehousing. Information Management and Computer Security.

Appendix 1: End-User Requirement Analysis (Survey Questionnaire)

1. How do you get the right data to generate your report in order to carry out business activities? i.e. Sources of your Data

2. How do you authenticate the source of the data?

3. What kind of report do you think can help you to make better decision in the day to day activities?

4. How do you know if you have the right data to work with?

5. How do you know when you have done a great job?

6. How long does it take you to prepare your daily report and what do you see as a problem and how can it be resolved in your view

7. What sort of data or data do you required to work with – Push or pull.

S/No	Question	Strongly Agree	Tends to Agree	Neither Agree nor Disagree	Tends to Disagree	Strongly Disagree
1	Data warehouse should support various data sources as possible and have the ability to integrate more data sources with the reporting, modeling, analysis strategy, in the near future.					
2	Close synchronization among functional area units within the business, so that it can help managers make decision on the overall working together of the units and how they affect each other.					
3	Users should be able to access the data more easily and quickly, and able to extract singular reports that can make them make decision quicker.					
4	Organization should be able to identify what part of the business is bringing more revenue by the structured queries and reporting generated by DW.					
5	I will like to be involved in the design (and implementation) of any system that will impact my job					
6	Senior managers must be able to generated One report that can describe the state of the business and its units per time.					
7	DW can adjust to new business processes and easily support future and new requirements.					
8	Data values in DW correctly represent the real world objects and events being described.					
9	Able to handle increase in the complexity and number of simultaneous queries without impacting system performances.					
10	Level of Data correctness in DW should be at its highest possible					

Appendix 2: Data Warehouse Evaluation

S/N	Factors/Bench Marks	Strongly Agree	Tends to Agree	Neither Agree nor Disagree	Tends to Disagree	Strongly Disagree
1	The Data warehousing system implemented is a success					
2	The Data warehousing system does not satisfy my requirements					
3	The Data warehousing system is easy to use					
4	It is easy to retrieve data from the system and understand such data.					
5	Using the system make my job easier for me.					
6	The report/output in line with the business requirements					
7	You have access to timely information when you need it					
8	The system provide sufficient information					
9	You are satisfied with the system accuracy					

Appendix 3: Answers to Appendix 1

Question 1
The data are normally generated by the IT department and based on the user collated data within each unit.
It is the responsibility of every unit in the organization to input data into the DW - The Sales, Marketing, IT, Distribution Channels teams and so on. Data can be gotten from CSV files, Images, email, XML files and so on
Most of the data are from the IT department and individual unit that has up to date data
Data are generated from all unit of the business which is then uploaded into the DW
Question 2
The data is being authenticated by comparing the data with the available immediate information
Against past activities such as sales, finance etc activities
We authenticate the source of the data through our request to the IT
Can verify where the data originated from.
Question 3
Daily activity report and sales report by different criteria would help in the better decision making. The snapshot would also help
Reports that can be generated easily from various fact tables
Daily stock availability, customer status, previous daily sales, price list, stores list.
Weekly Sales Summary report, end of month financial analysis, price list report, product list report, store list
Question 4
We have a specific format for different report and we expect to get the report according to the specs. The right data would depend on the source and input from colleague
It must be authenticated from its originated source
Comparing the data with what we had in the previous day and add the current transaction to it.
you get the desired report you want
Question 5
When I am able to present the right report at the right time to the user within the SAL
When any kind of report can be generated ranging from type of clientele, accounting reports, sale forecast etc
Whenever I am able to deliver the right reports to my manager that would aid his decision making
Make decision intelligently with the report you generate
Question 6
It takes longer time to get all the reports ready as we need format the report as required which makes it more difficult to meet up
If I know what I want, it is quicker, the problem would be knowing what kind of report would help my decision
Due to the manual process and formatting the report, better part of the day is being used to prepare the report
At the moment it is quite horrendous, we want to be able to generate report swiftly and accurately
Question 7
A Push data will be good as it makes the data available as need and pull process can be will help in search for further data.
Push
A push data is good as we always have it as a subscribed data
Push