CSC-475, CSC-484

BSCS Final Year Project

ADS DATA ANALYTICS

Submitted by

SOHAIL PERVAIZ 191302
ABDUL QADOOS 183312

PROJECT SUPERVISOR

Mr. Shaukat Ali Chaudhary

SUBMITTED TO

Mr. Shaukat Ali Chaudhary

In-Charge BSCS/BSSE/M.Sc. Projects

Pak-AIMS

The Institute of Management Sciences (Pak-AIMS)

Lahore, Pakistan

Session 2019 – 2023

ACKNOWLEDGEMENT

Up and above, everything, all appreciation to Allah Almighty, the compassionate and merciful, who enabled us to elucidate a drop from the existing ocean of knowledge, countless salutation be upon the Holy Prophet Muhammad (Salalah Alaihi Wasallam (S.A.W)(ﷺ)) the city of knowledge, who has guided our “Ummah” to seek knowledge from cradle to grave.

We truly acknowledge the cooperation and help make by Project Management Office, CS & IT department, Pak-AIMS (IMS), Lahore. They have been a constant source of guidance throughout the course of this project. It is quite delectable and to becoming to avail of this most propitious opportunity to articulate with utmost gratification, my profound and intense sense if indebtedness to my affectionate supervisor, Mr. Shaukat Ali Chaudhary, His proficient counseling, valuable suggestions, boundless forbearance, indefatigable help with anything, anywhere, anytime, consummate advice and thought-provoking instruction in piloting this project. Special thanks for his would always be due.

DEDICATION

I dedicate this work to ALLAH Almighty who glorified us with the knowledge and bravery to complete this responsibility with grace. Secondly, my loving and supportive family whose prayers, advice, and continuous support played a major role in achieving this goal. Thirdly, to my Supervisor Prof. Shaukat Ali Chaudhary a who always supported and guided me throughout the project. Finally, I would like to thank The Institute of Management Sciences (Pak-AIMS) Lahore, Pakistan and especially the Department of Computer Science for always believing in us and providing us the platforms to participate in engaging activities where we were able to boost our skills through practical and theoretical knowledge.

Sohail Pervaiz ____________________

Abdul Qadoos ____________________

ABSTRACT

Business intelligence (BI) is a technology-driven process for analyzing data and delivering actionable information that helps executives, managers and workers make informed business decisions. As part of the BI process, organizations collect data from internal IT systems and external sources, prepare it for analysis, run queries against the data and create data visualizations, BI dashboards and reports to make the analytics results available to business users for operational decision-making and strategic planning. My Project aim is to store ads data from OLTP to data warehouse & then perform data analysis using SQL queries & shown results in the form of dashboard using Business Intelligence Tool Tableau.

The project is related to Business Intelligence. Main purpose of project is to show up data in KPI’s, Dashboard & on the base of KPI’s perform business decision making. Firstly, I am taking Ads data from company located in Lahore in the form of excel files. I will load data from excel files to Staging Database (Staging tables) using SQL server integration services (SSIS) packages. After performing business rules using SQL queries, I will populate dimension & fact tables of data warehouse. After this I will connect Tableau (Business Intelligence tool) with SQL server database & will develop Dashboard of ads data. This dashboard will show complete picture of stake holder business & help them to perform their future decision of the base of provided dashboard.

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

Table of Contents

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

List of Figures

CHAPTER 1

Introduction

Business intelligence (BI) is a technology-driven process for analyzing data and delivering actionable information that helps executives, managers and workers make informed business decisions. As part of the BI process, organizations collect data from internal IT systems and external sources, prepare it for analysis, run queries against the data and create data visualizations, BI dashboards and reports to make the analytics results available to business users for operational decision-making and strategic planning. The Project aim is to store ads data from OLTP to data warehouse & then perform data analysis using SQL queries & shown results in the form of dashboard using Business Intelligence Tool Tableau.

Dashboard will show up the clear picture of stake holder business. Using Dashboard stake holder can perform decision making about their business. I will also provide some predictive analytics on dashboard as well. Tableau provide predictive analytics algorithm like Time Series algorithm. Using predictive analytics stake holder can see their future performance of data as well. That’s the overall purpose of my project.

1.1 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

1.2 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.

CHAPTER 2

Literature Review

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

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

Data Mart is a subset of data warehouse and is defined as body of historical data in electronic repository that does 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.

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.

2.0 Data Warehouse Concepts

Data warehousing is the process of collecting data to be stored in a managed database in which the data are subject-oriented and integrated, time variant, and nonvolatile for the support of decision-making. Data from the different operations of a corporation are reconciled and stored in a central repository (a data warehouse) from where analysts extract information that enables better decision making.

Data can then be aggregated or parsed, and sliced and diced as needed in order to provide information. There are two main authors that are known in the world of data warehouse design, their approaches to some areas of the data warehousing are different; William Inman and Ralph Kimball. The approach by Inman 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 Data Warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data used in support of decision-making processes. “Subject Oriented” means that a data warehouse focuses on the high-level entities of the business and the data are organized according to subject

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

“Time-variant” means warehouses provide access to a greater volume of more detailed information over a longer period and that the data are associated with a point in time 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 Finally,” nonvolatile” means that the data do not change.

According to 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. 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 and that of that Kimball’s conformed dimensions are de-normalized, whereas Inman uses a highly normalized central database model. Inman’s data marts store 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 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.

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 summarized. DWs usually do not keep as much detail as transaction-oriented systems.

2.2.1 The Data Warehouse Data Model

Inman 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.

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

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. 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.

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 nonfunctional (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.

2.1 Developing Data Warehouse

Explicit when it says 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;

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;

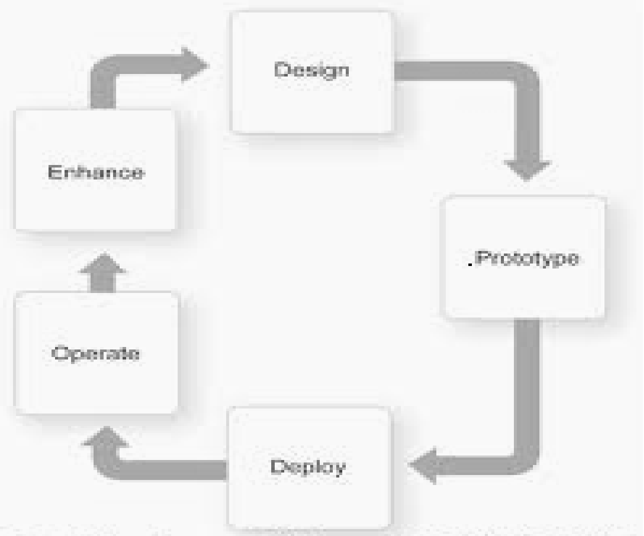


**Figure 1:	DW Development Lifecycle (DWLC) Model [1]**

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.

2.2 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 organization 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 . “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, classifies data mining (DM) as a BI technique includes all resources (DW, DM, hypertext analysis and web information) in the creation of a BI system; and finally, linking BI and the Internet, 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:

- 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.

2.3 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 Kimber (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. Organizations choose to have two separate information systems, one OLTP and One OLAP system

Poet stresses 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 was a bit different on the approach to the development of data warehouse system. He argues 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. differ from this approach by following a requirements-driven development lifecycle.

2.4 Data Warehouse and Business Intelligence High Level Architecture

From the Data warehouse institute did study on the success factor in implementing BI, systems in organizations and the role of data warehouse in this process.

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. 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. These are the OLTP system. The data from the source systems go through a process called data staging to the presentation servers 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 built 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, 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 Inman’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 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.

2.5 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 Inman 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 Inman. 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.

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 methodologies are consider being systems of explicit rules and produced, upon which research is base, and against which claim for knowledge are evaluate. 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 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. 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 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 sources of evidence. Third, even if all the relevant variables are quantitative, distinctive strategies will be needed for research design and for analysis.

We have decided to do a case study as it gives an in-depth study of a particular situation rather than a sweeping statistics survey. It is a method used to narrow down a very broad field of research topic. Case study provides more realistic responses than a purely statistics survey. They are more detail than the statistical method. Case study in this project will give us the opportunity to study the aim of the project using some past project in the same industry and compare it with aim of this project.

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.

As mention in the literature review, the two fathers of Data warehouse are Bill Inman 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.

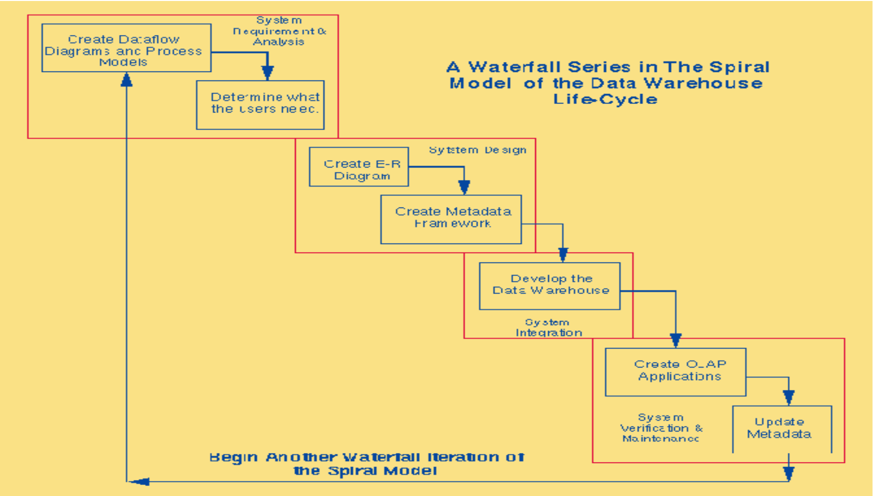

System Requirement & Analysis
Create Dataflow Diagrams and Process Models
Determine what the users need.
A Waterfall Series in The Spiral Model of the Data Warehouse Life-Cycle
Sytstem Design
Create E-R Diagram
Create Metadata Framework
Develop the Data Warehouse
System Integration
Create OLAP Applications
System Verification & Maintenance
Update Metadata
Begin Another Waterfall Iteration of the Spiral Model

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

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 center 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 contains 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”

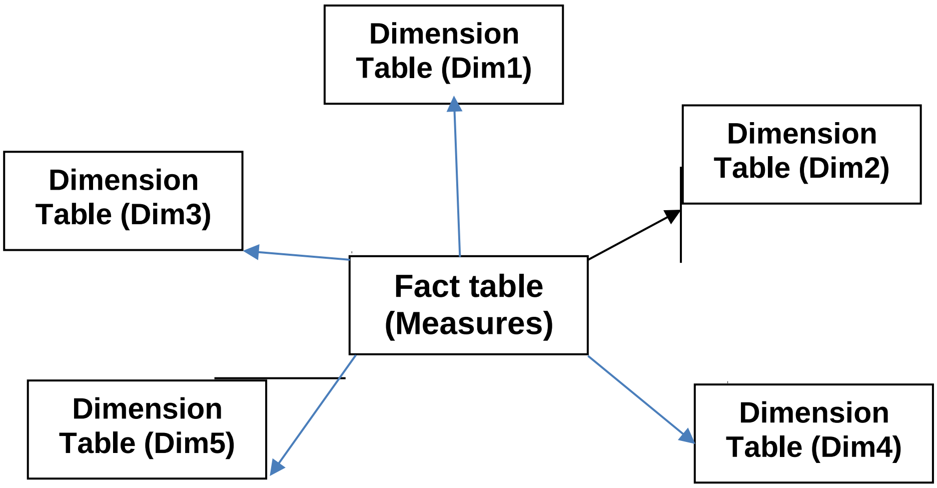


Figure 3: A Star Schema [3]

3.3 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.

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 Clickstream (ADS) 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. Organization 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 reveals 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 center 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 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.

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 analyses 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.

4.1.3 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.

- 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.

4.1.4 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
- all time.

4.1.5 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 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.

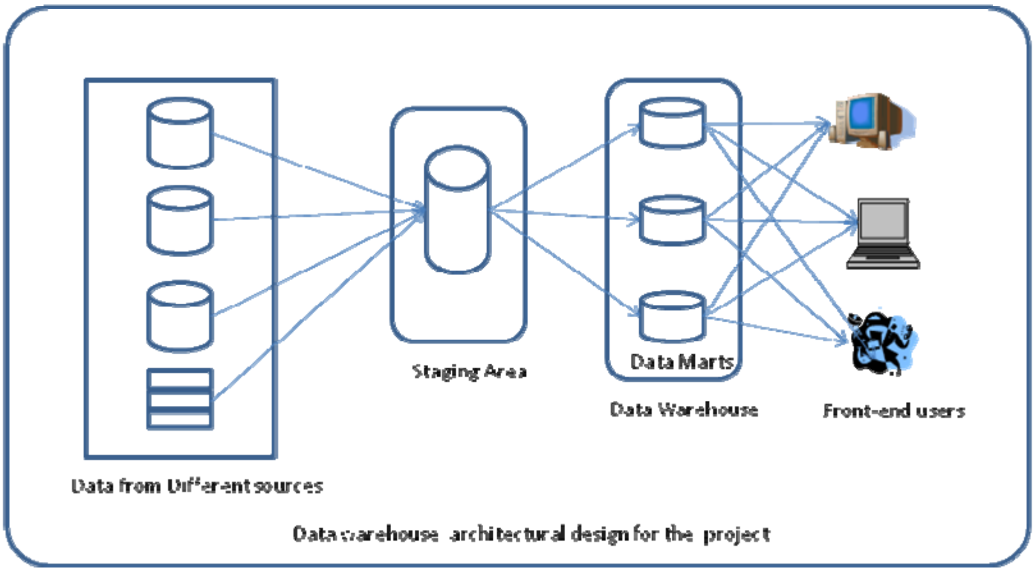

Staging Area
Data Marts
Data Warehouse
Front-end users
Data from Different sources
Data warehouse architectural design for the project

Figure 4: Data Warehouse Architecture Design of the Project[4]

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

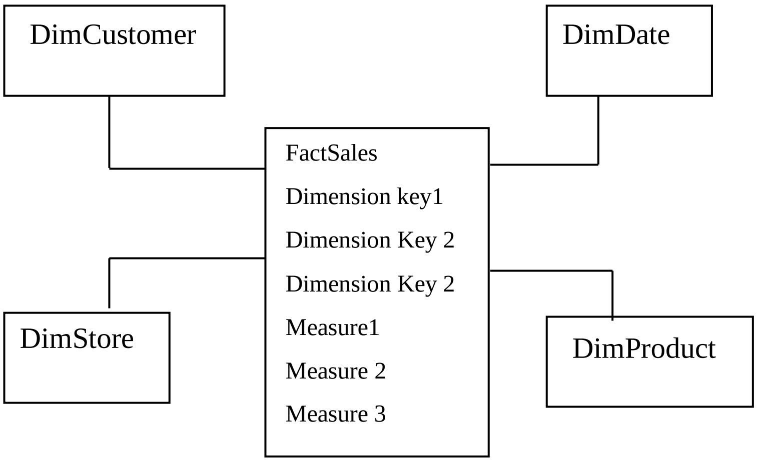


Figure 5: Star Schema for the Product Sales Data Mart [5]

The dimension and the fact table are as follow date, customer, product and store. The dimensions' structure would be discussed 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.

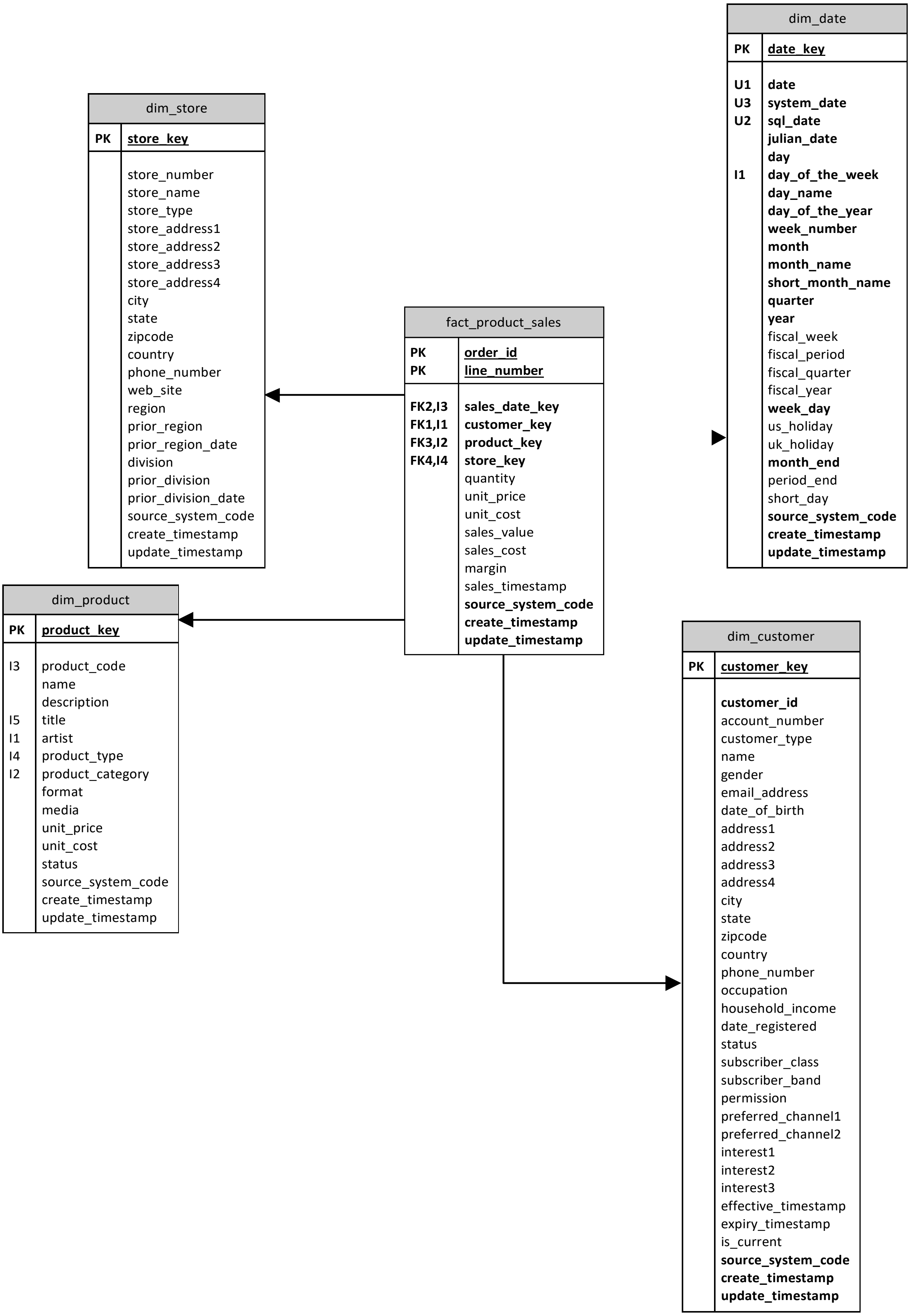


Figure 6: Logical Model of Product Sales Department [6]

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;

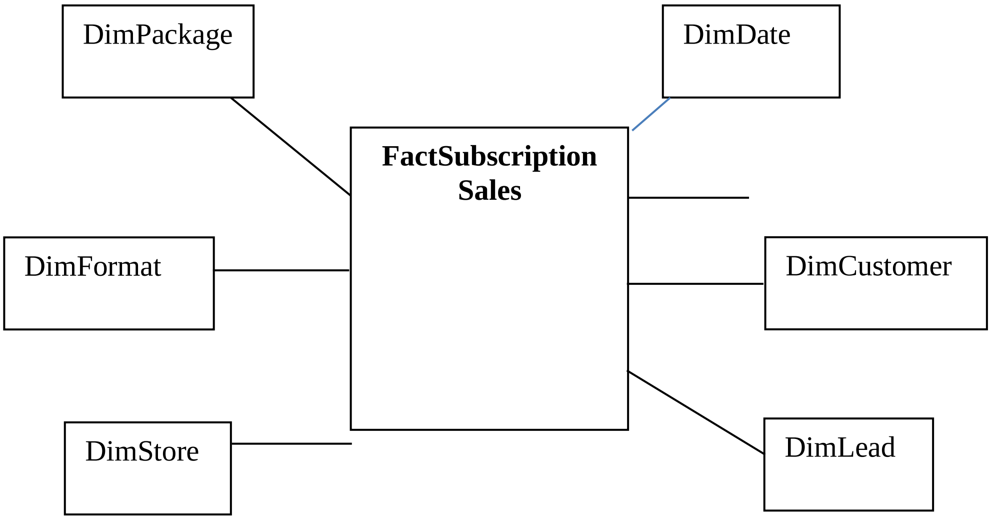


Figure 7: Star Schema for the Fact Subscription Sales [7]

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

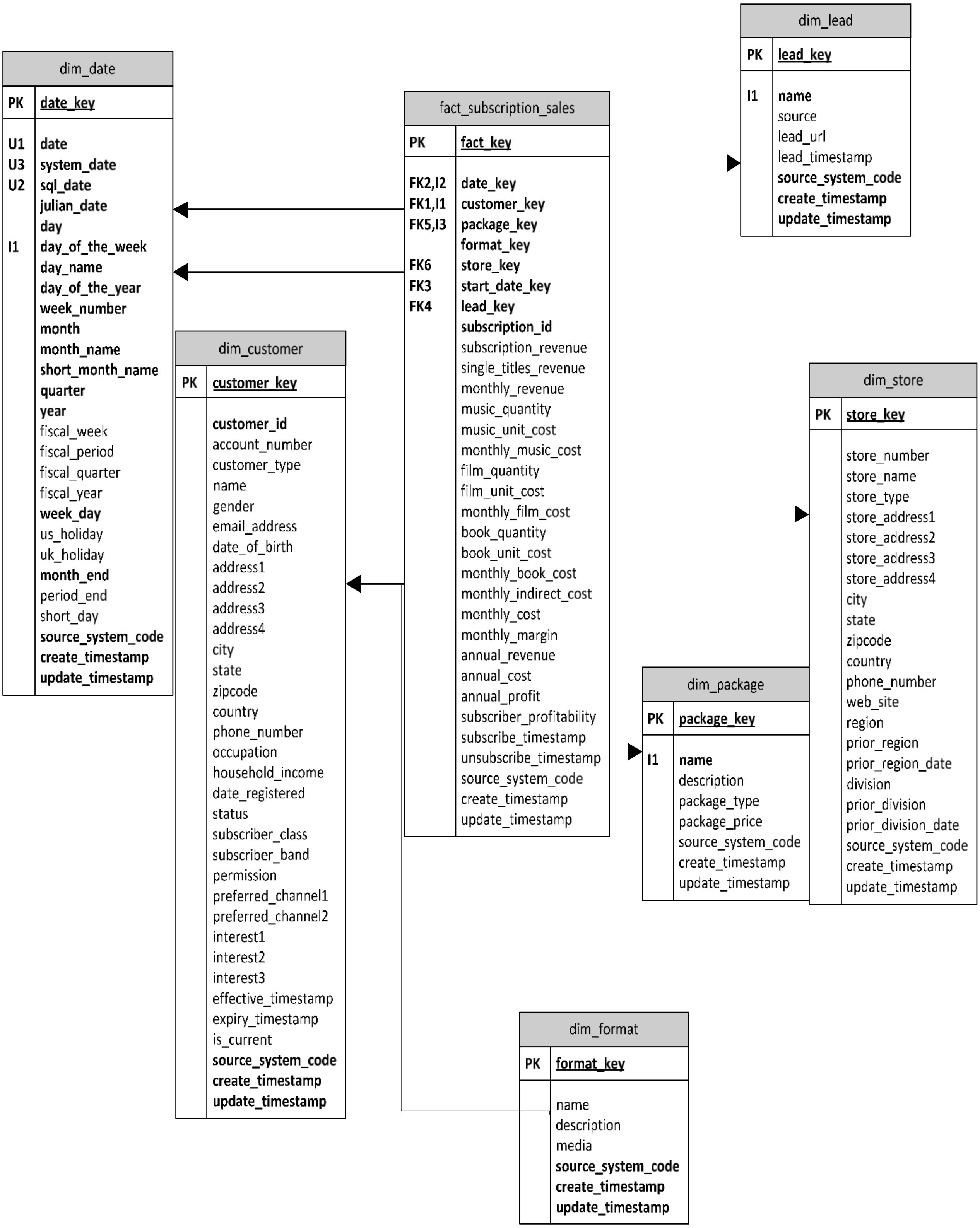


Figure 8: Logical Model of Fact Subscription Sales[8]

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

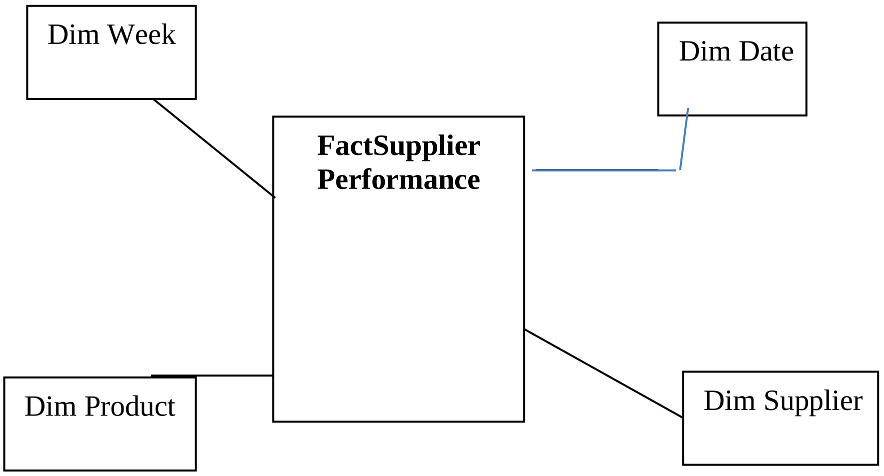


Figure 9: Star Schema of the Fact Supplier Performance[9]

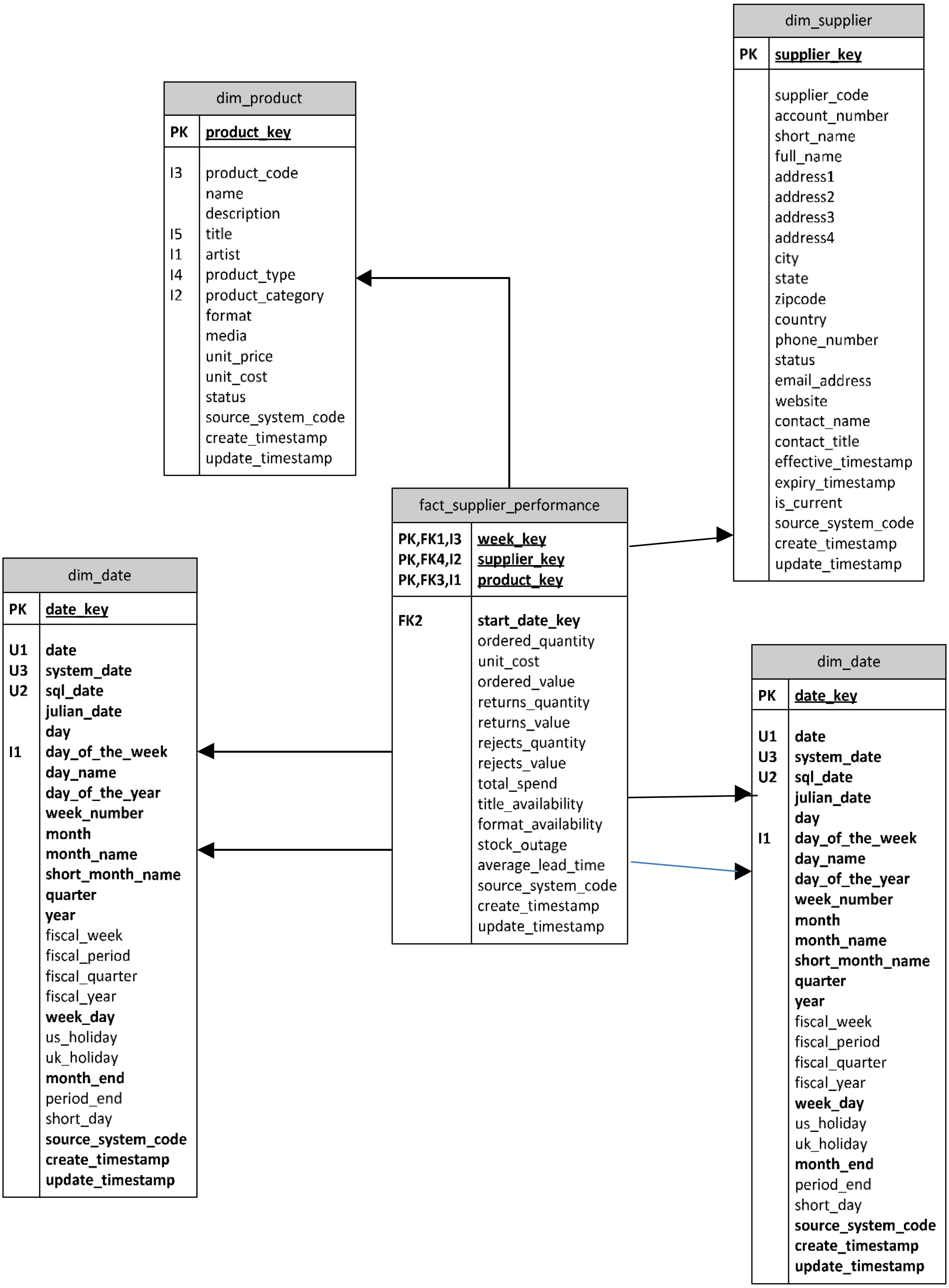


Figure 10: Logical Model of Fact Suppliers Performance [10]

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.

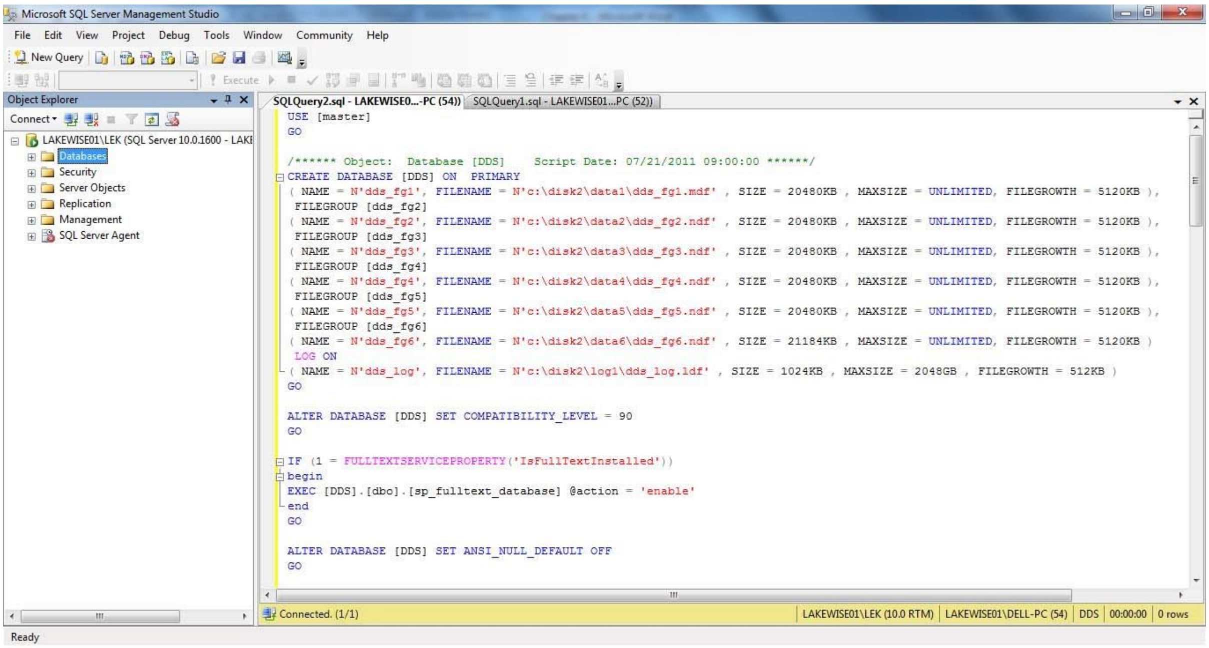
The main key column is a primary key in the dimension table, and it is a foreign key on the fact table. This is known in the database world as referential integrity. The main key in the dimensional table are usually the surrogate key, they are unique and not null, it uniquely identify the record in a 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.



Microsoft SQL Server Management Studio
File Edit View Project Debug Tools Window Community Help
New Query
Execute
Object Explorer
Connect
LAKEWISE01\LEK (SQL Server 10.0.1600 - LAKE
Databases
Security
Server Objects
Replication
Management
SQL Server Agent
SQLQuery2.sql - LAKEWISE0...-PC (54))
SQLQuery1.sql - LAKEWISE01...PC (52))
Connected. (1/1)
LAKEWISE01\LEK (10.0 RTM)
LAKEWISE01\DELL-PC (54)
DDS
00:00:00
0 rows
Ready

```sql
USE [master]
GO

/****** Object:  Database [DDS]    Script Date: 07/21/2011 09:00:00 ******/
CREATE DATABASE [DDS] ON  PRIMARY
( NAME = N'dds_fg1', FILENAME = N'c:\disk2\data1\dds_fg1.mdf' , SIZE = 20480KB , MAXSIZE = UNLIMITED, FILEGROWTH = 5120KB ),
 FILEGROUP [dds_fg2]
( NAME = N'dds_fg2', FILENAME = N'c:\disk2\data2\dds_fg2.ndf' , SIZE = 20480KB , MAXSIZE = UNLIMITED, FILEGROWTH = 5120KB ),
 FILEGROUP [dds_fg3]
( NAME = N'dds_fg3', FILENAME = N'c:\disk2\data3\dds_fg3.ndf' , SIZE = 20480KB , MAXSIZE = UNLIMITED, FILEGROWTH = 5120KB ),
 FILEGROUP [dds_fg4]
( NAME = N'dds_fg4', FILENAME = N'c:\disk2\data4\dds_fg4.ndf' , SIZE = 20480KB , MAXSIZE = UNLIMITED, FILEGROWTH = 5120KB ),
 FILEGROUP [dds_fg5]
( NAME = N'dds_fg5', FILENAME = N'c:\disk2\data5\dds_fg5.ndf' , SIZE = 20480KB , MAXSIZE = UNLIMITED, FILEGROWTH = 5120KB ),
 FILEGROUP [dds_fg6]
( NAME = N'dds_fg6', FILENAME = N'c:\disk2\data6\dds_fg6.ndf' , SIZE = 21184KB , MAXSIZE = UNLIMITED, FILEGROWTH = 5120KB )
 LOG ON
( NAME = N'dds_log', FILENAME = N'c:\disk2\log1\dds_log.ldf' , SIZE = 1024KB , MAXSIZE = 2048GB , FILEGROWTH = 512KB )
GO

ALTER DATABASE [DDS] SET COMPATIBILITY_LEVEL = 90
GO

IF (1 = FULLTEXTSERVICEPROPERTY('IsFullTextInstalled'))
begin
EXEC [DDS].[dbo].[sp_fulltext_database] @action = 'enable'
end
GO

ALTER DATABASE [DDS] SET ANSI_NULL_DEFAULT OFF
GO
```

Figure 11: Query View of SQL Server Management Studio 2008[11]

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;

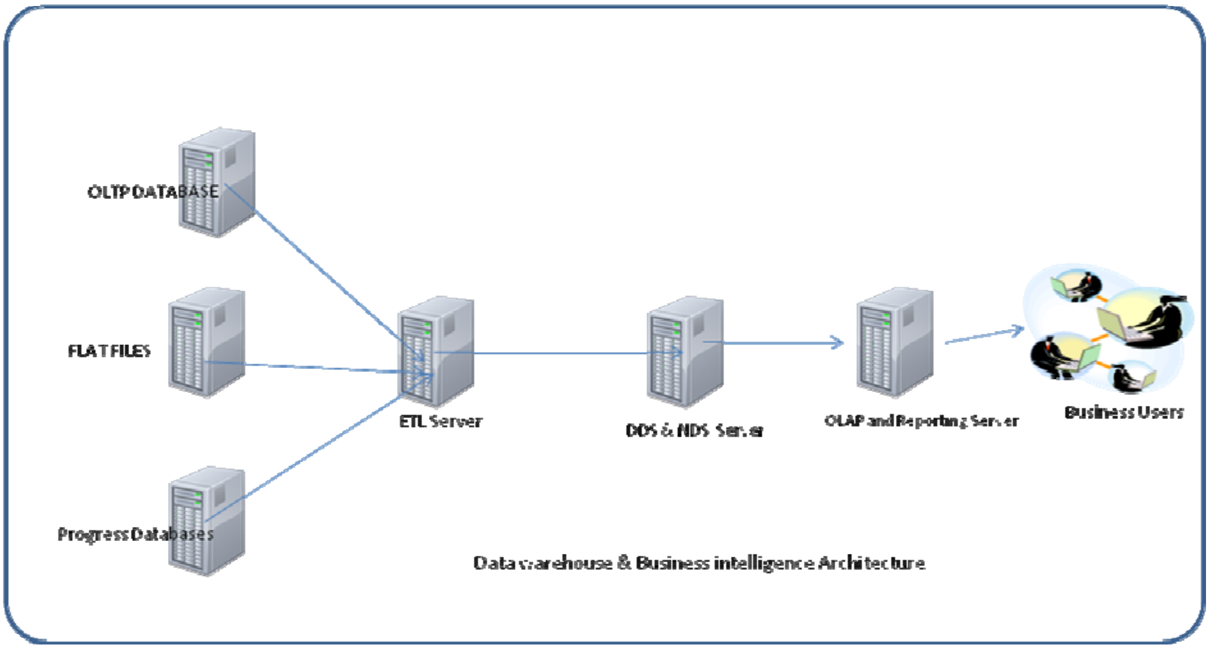


Figure 12: Data Warehouse and Business Intelligence Architecture[12]

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;

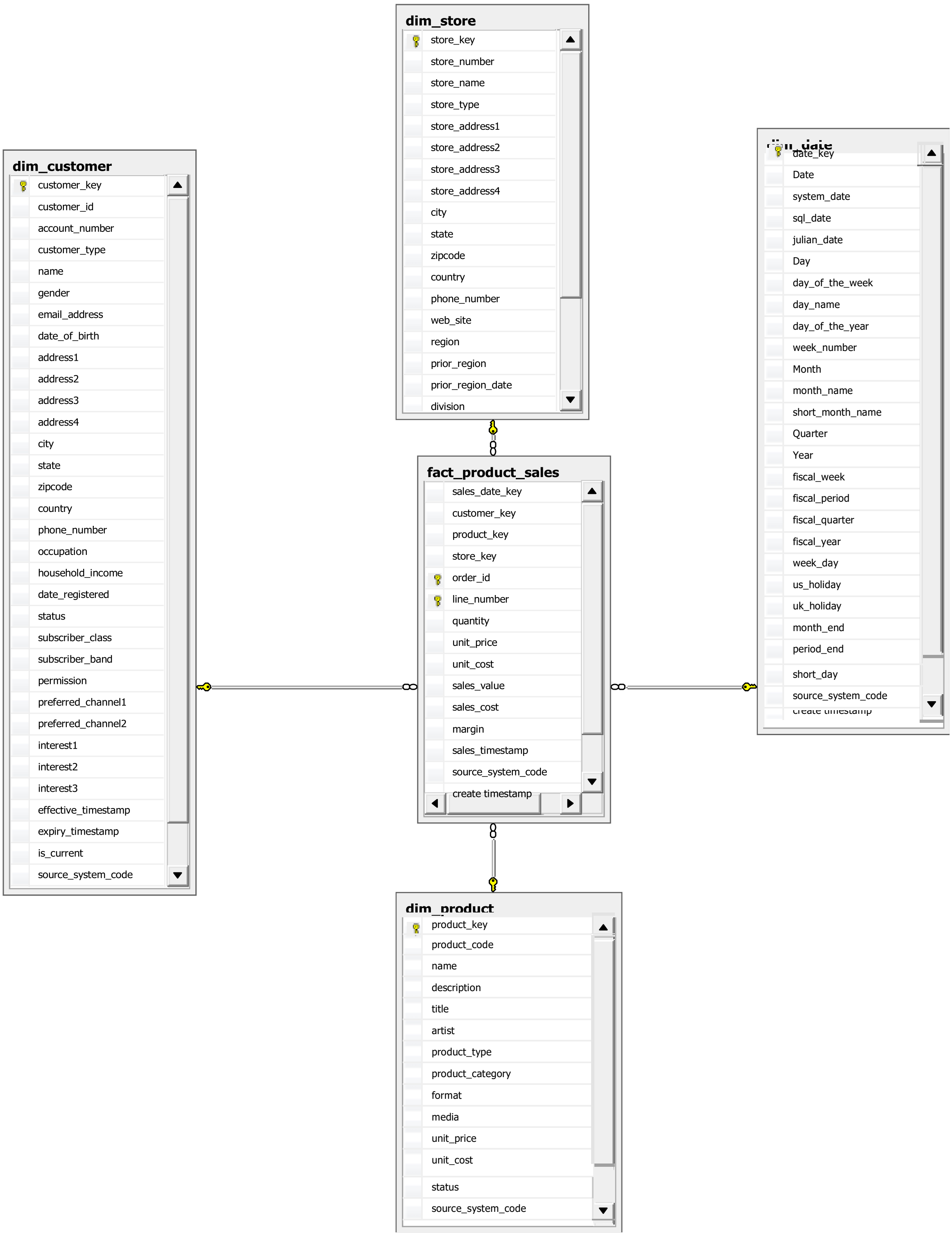


Figure 13: Physical Design of the Fact Product Sales Data Mart [13]

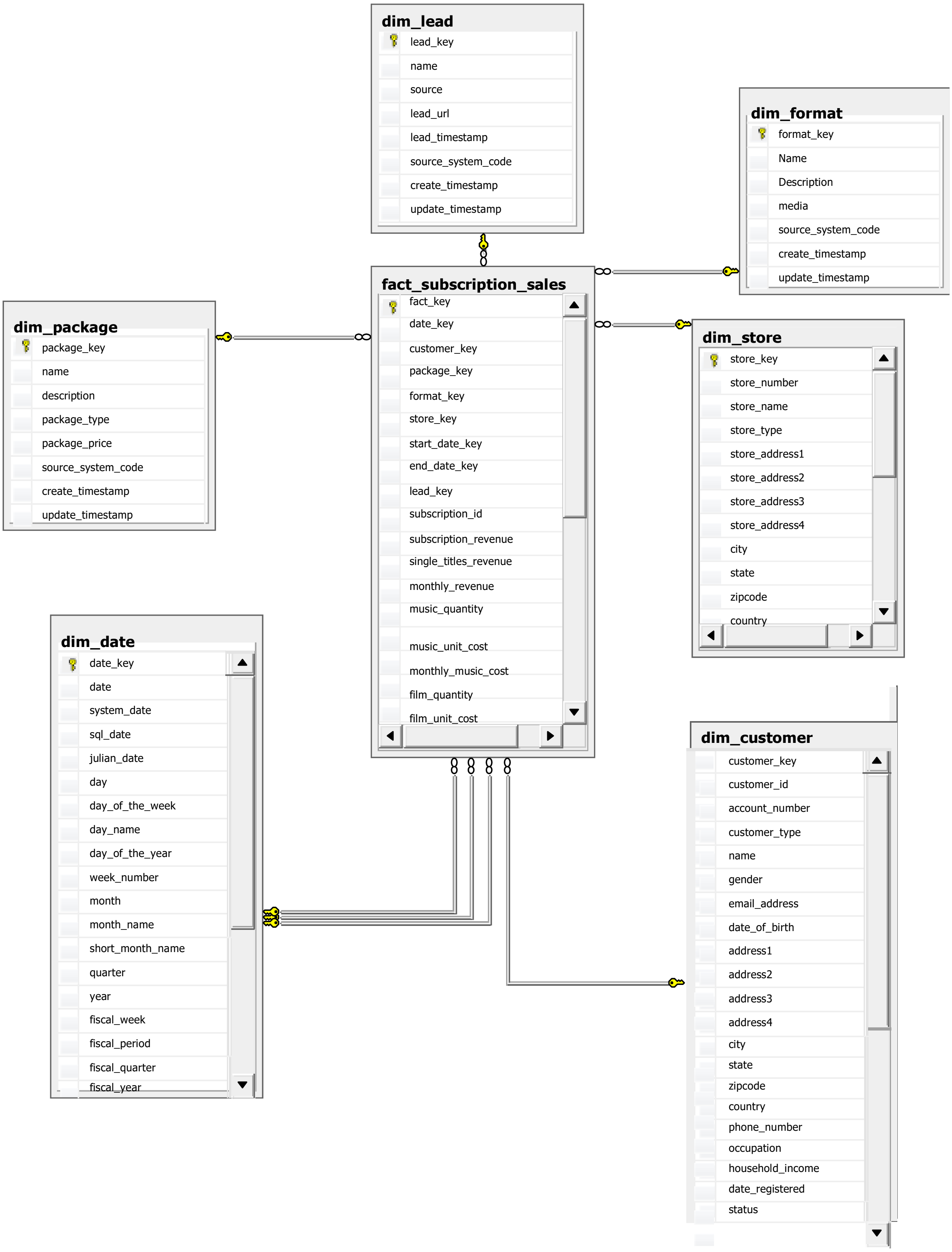


Figure 14: Physical Design of the Fact Subscription Sales Data Mart[14]

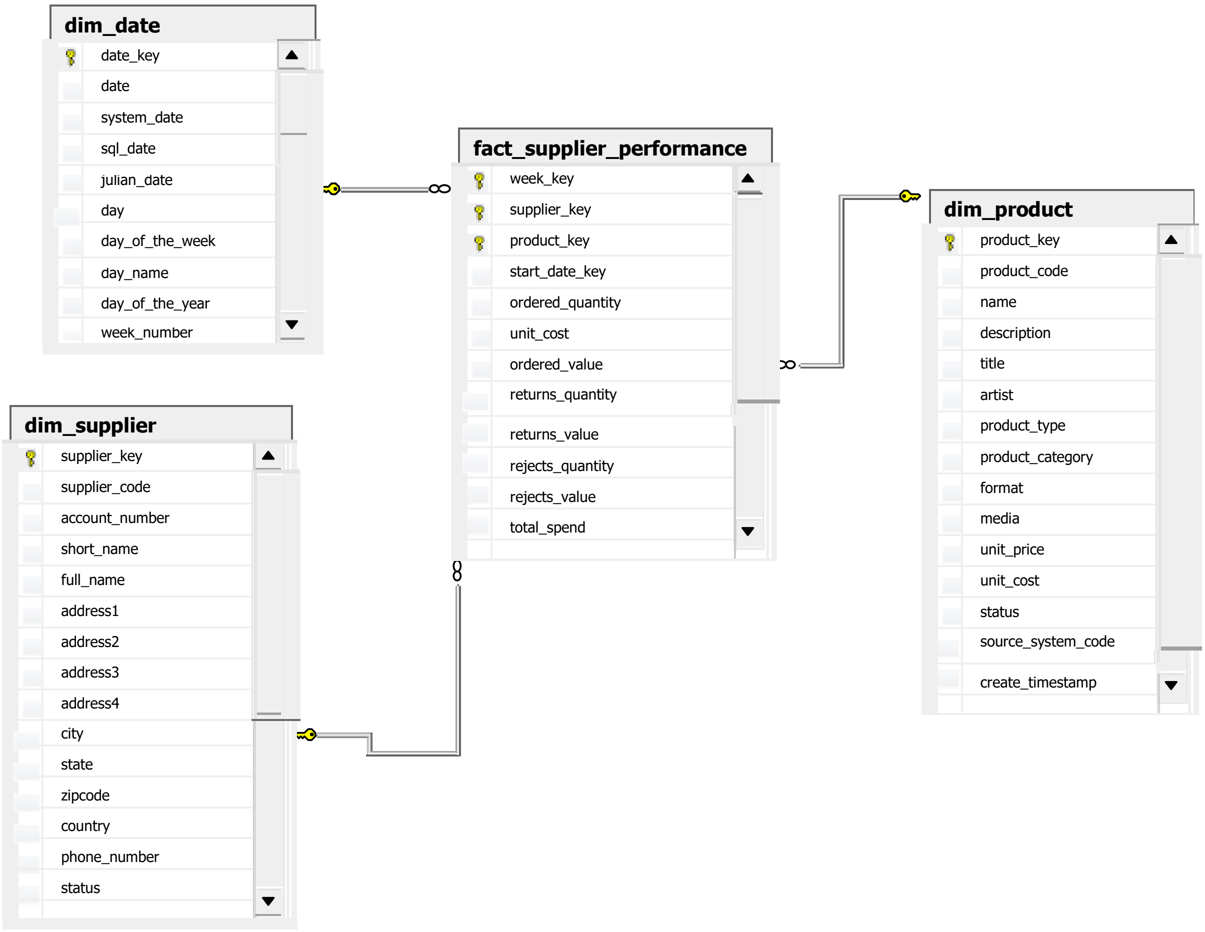


Figure 15: Physical Design of the Fact Supplier Performance Data Mart[15]

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

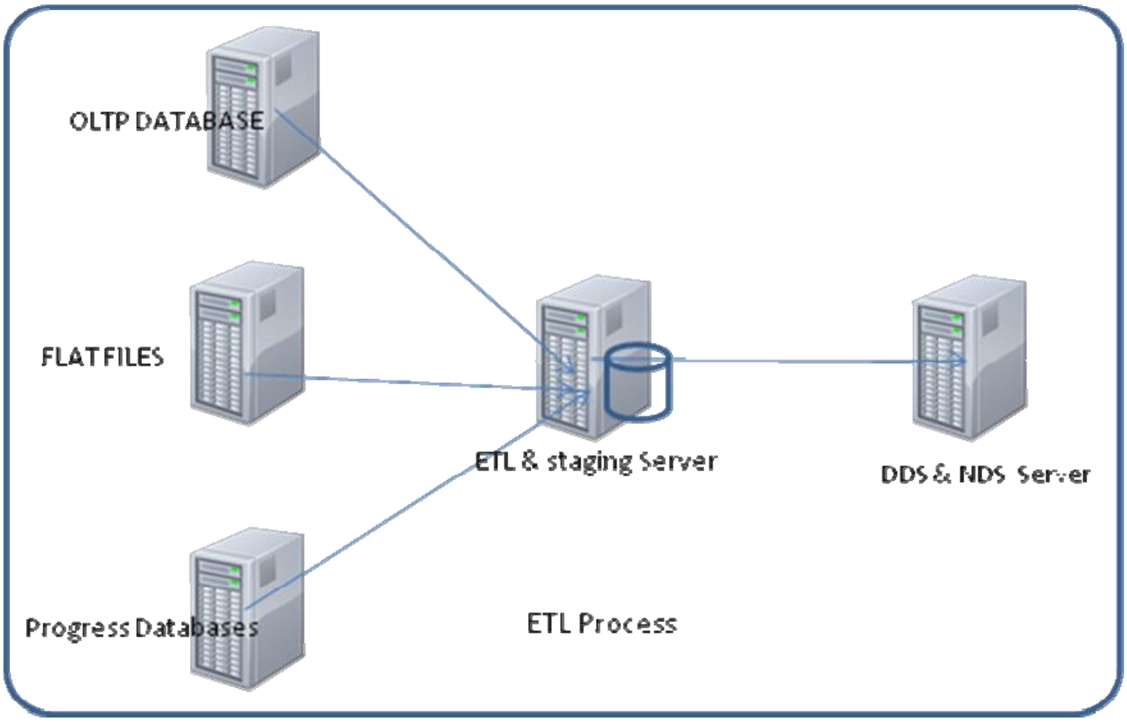


Figure 16: Extraction, Transformation, and Load (ETL) Architecture[16]

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.

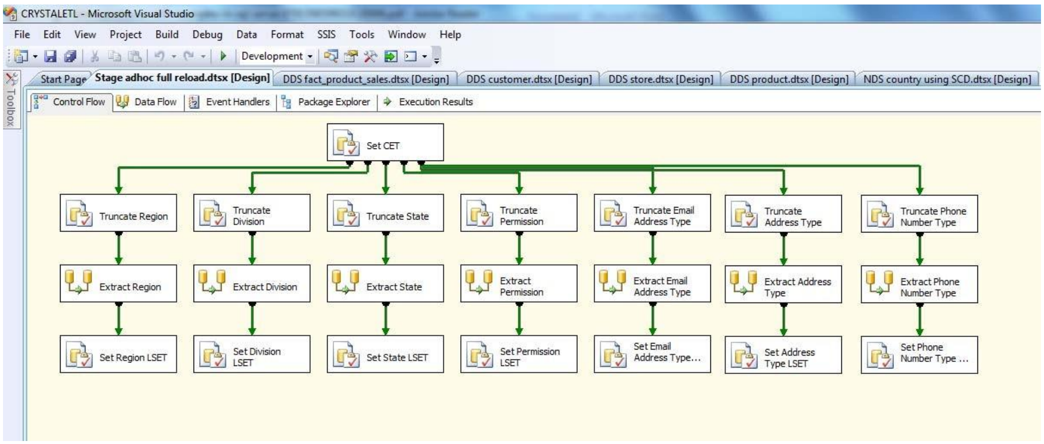

CRYSTALETL - Microsoft Visual Studio
File Edit View Project Build Debug Data Format SSIS Tools Window Help
Development
Toolbox
Start Page
Stage adhoc full reload.dtsx [Design]
DDS fact_product_sales.dtsx [Design]
DDS customer.dtsx [Design]
DDS store.dtsx [Design]
DDS product.dtsx [Design]
NDS country using SCD.dtsx [Design]
Control Flow
Data Flow
Event Handlers
Package Explorer
Execution Results
Set CET
Truncate Region
Truncate Division
Truncate State
Truncate Permission
Truncate Email Address Type
Truncate Address Type
Truncate Phone Number Type
Extract Region
Extract Division
Extract State
Extract Permission
Extract Email Address Type
Extract Address Type
Extract Phone Number Type
Set Region LSET
Set Division LSET
Set State LSET
Set Permission LSET
Set Email Address Type...
Set Address Type LSET
Set Phone Number Type ...

Figure 17: Stage Ad-hoc Full Load[17]

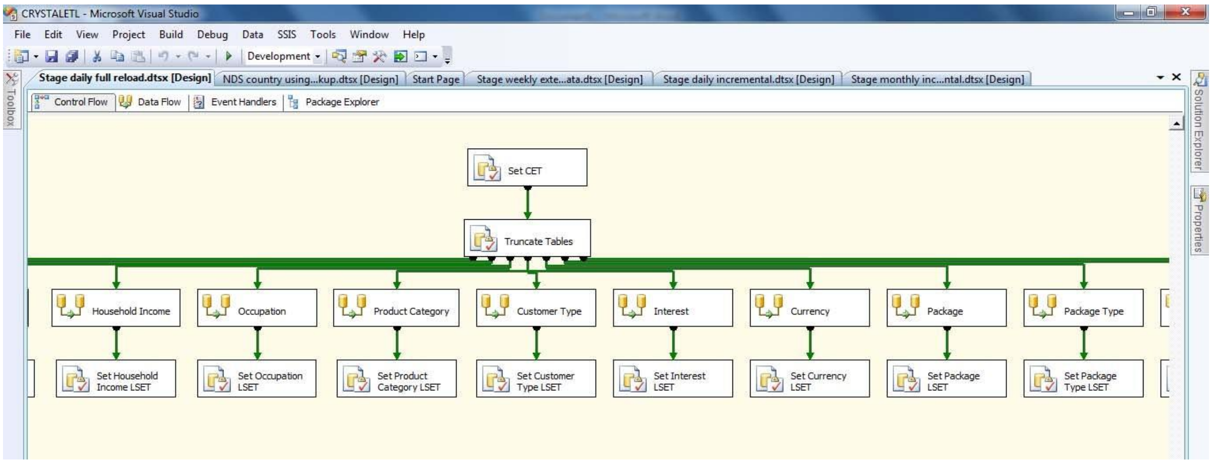

CRYSTALETL - Microsoft Visual Studio
File Edit View Project Build Debug Data SSIS Tools Window Help
Development
Stage daily full reload.dtsx [Design]
NDS country using...kup.dtsx [Design]
Start Page
Stage weekly exte...ata.dtsx [Design]
Stage daily incremental.dtsx [Design]
Stage monthly inc...ntal.dtsx [Design]
Control Flow
Data Flow
Event Handlers
Package Explorer
Toolbox
Solution Explorer
Properties
Set CET
Truncate Tables
Household Income
Occupation
Product Category
Customer Type
Interest
Currency
Package
Package Type
Set Household Income LSET
Set Occupation LSET
Set Product Category LSET
Set Customer Type LSET
Set Interest LSET
Set Currency LSET
Set Package LSET
Set Package Type LSET

Figure 18: Stage Daily Full Re-Load[18]

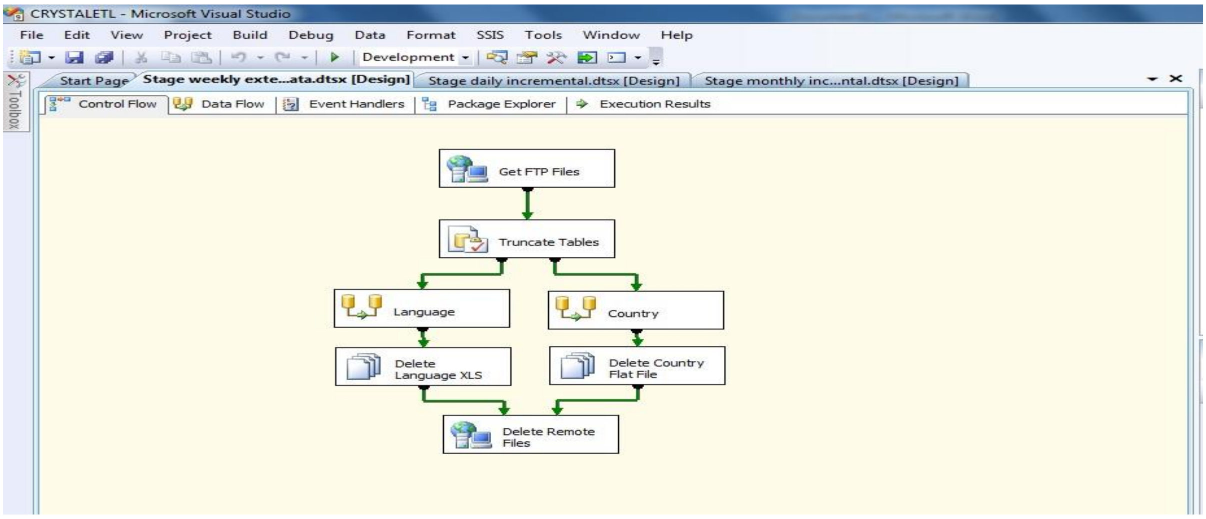

CRYSTALETL - Microsoft Visual Studio
File Edit View Project Build Debug Data Format SSIS Tools Window Help
Development
Start Page
Stage weekly exte...ata.dtsx [Design]
Stage daily incremental.dtsx [Design]
Stage monthly inc...ntal.dtsx [Design]
Control Flow
Data Flow
Event Handlers
Package Explorer
Execution Results
Toolbox
Get FTP Files
Truncate Tables
Language
Country
Delete Language XLS
Delete Country Flat File
Delete Remote Files

Figure 19: Stage External Data Load[19]

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;

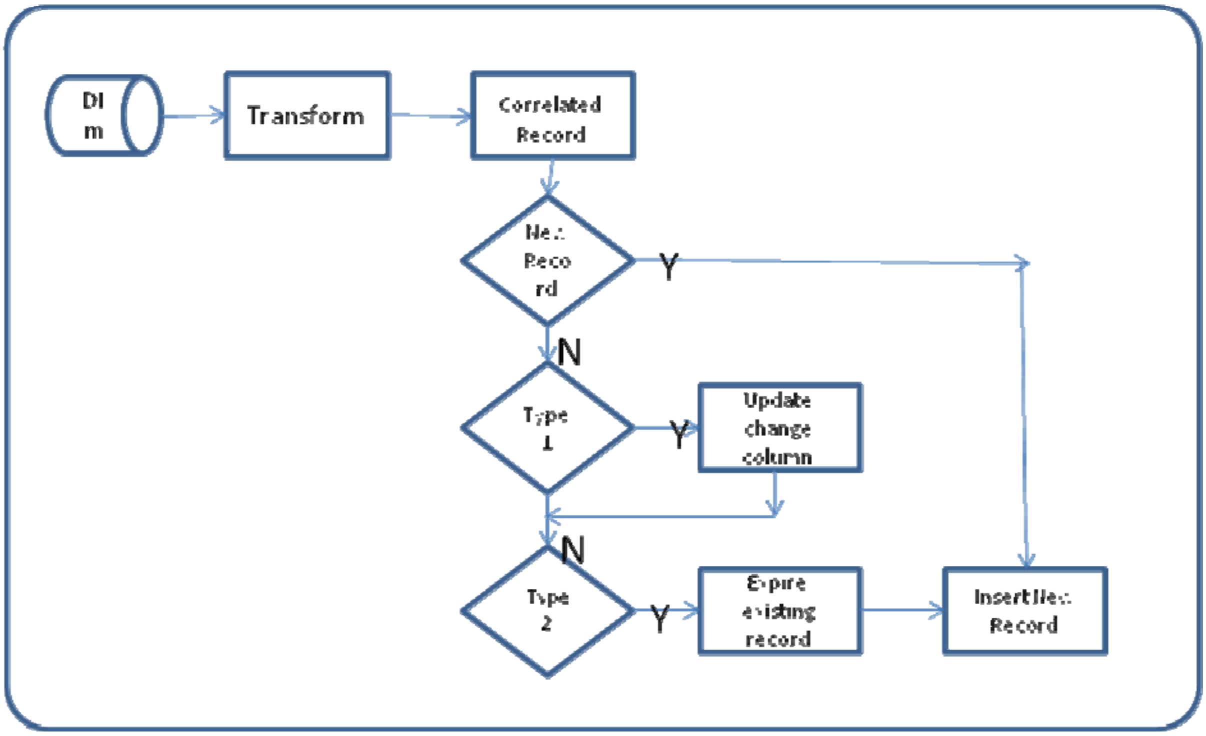
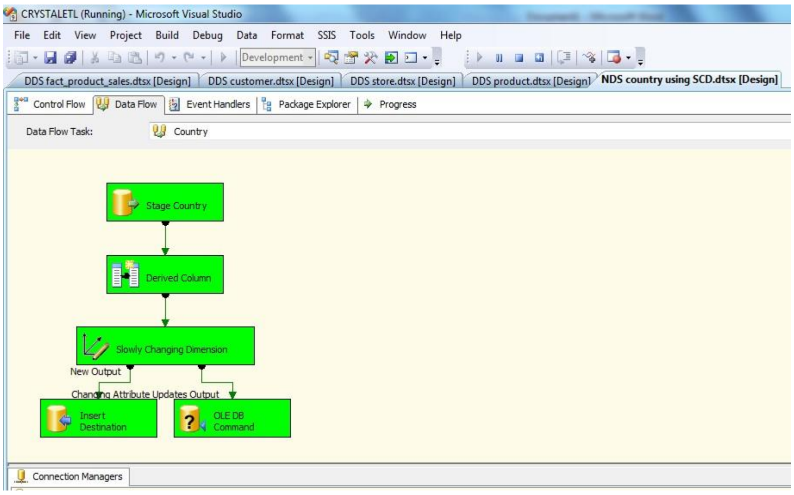


Figure 20: Flow Chart for loading the Dimension Tables[20]

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



Load the slowly changing dimension to preserve the history

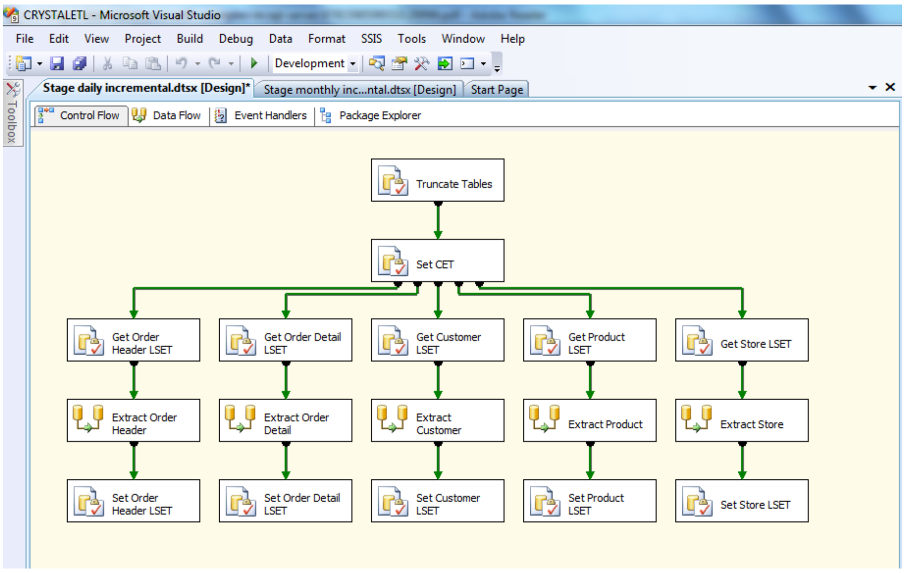


Figure 21: Daily Increment load of data into the data warehouse.[21]

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.

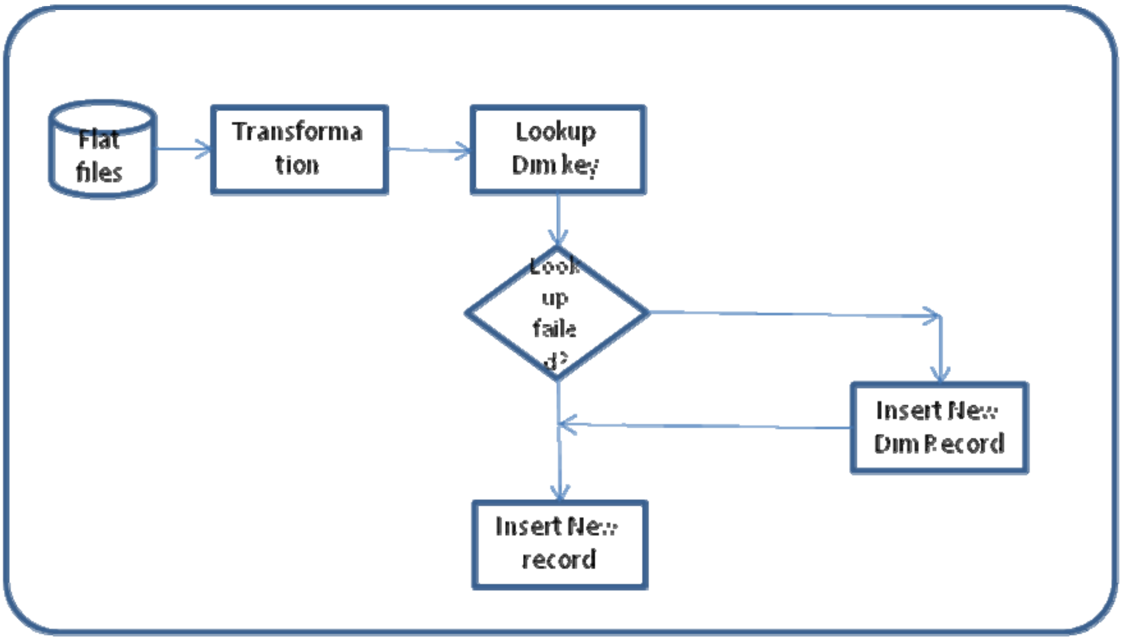


Figure 22: Flow Chart for loading the Dimension Tables [22]

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.

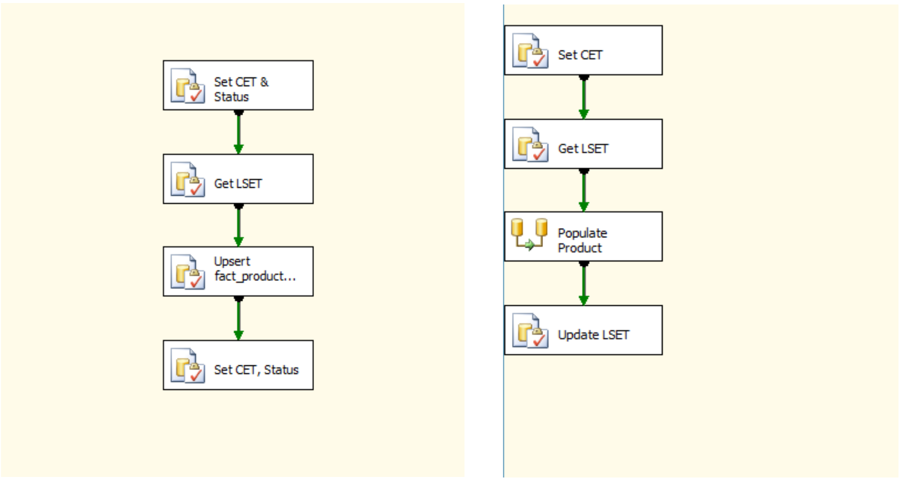
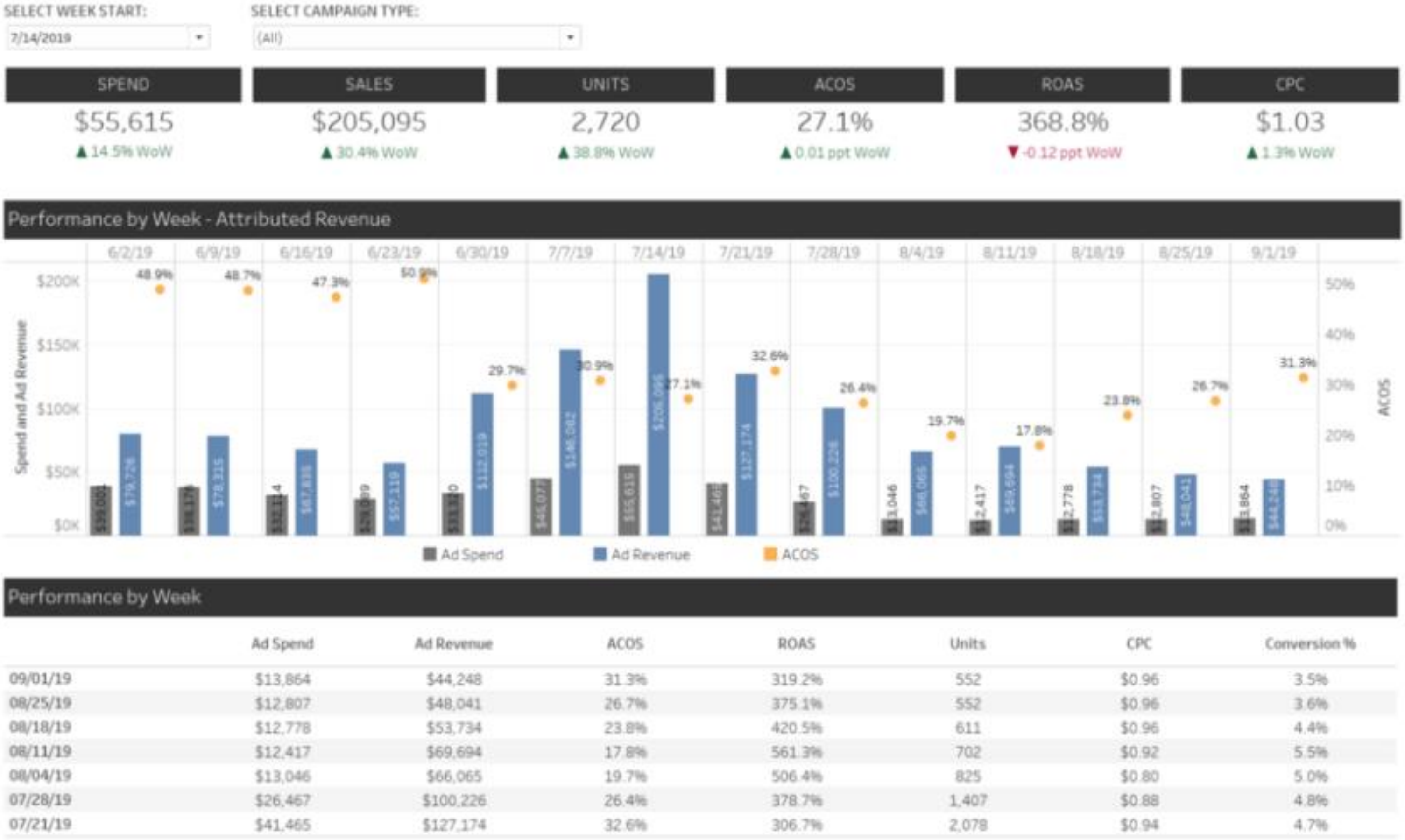


Figure 23: Fact Table Load[23]

4.3.4 End User Application

The system is designed with mass users in mind, it allow user to use familiarized tools such as Tableau 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.



SELECT WEEK START: 7/14/2019

SELECT CAMPAIGN TYPE: (All)

SPEND	SALES	UNITS	ACOS	ROAS	CPC
$55,615	$205,095	2,720	27.1%	368.8%	$1.03
▲ 14.5% WoW	▲ 30.4% WoW	▲ 38.8% WoW	▲ 0.01 ppt WoW	▼ -0.12 ppt WoW	▲ 1.3% WoW

Performance by Week - Attributed Revenue

Performance by Week

	Ad Spend	Ad Revenue	ACOS	ROAS	Units	CPC	Conversion %
09/01/19	$13,864	$44,248	31.3%	319.2%	552	$0.96	3.5%
08/25/19	$12,807	$48,041	26.7%	375.1%	552	$0.96	3.6%
08/18/19	$12,778	$53,734	23.8%	420.5%	611	$0.96	4.4%
08/11/19	$12,417	$69,694	17.8%	561.3%	702	$0.92	5.5%
08/04/19	$13,046	$66,065	19.7%	506.4%	825	$0.80	5.0%
07/28/19	$26,467	$100,226	26.4%	378.7%	1,407	$0.88	4.8%
07/21/19	$41,465	$127,174	32.6%	306.7%	2,078	$0.94	4.7%

Figure 24: End User Application Tool[24]

CHAPTER 5

5.0 Conclusion, Discussions and Limitations

The project focused on designing and implementation of DW and BI system for a Ads industry using Ads data 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 answer 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 References

[1] Review of Data Warehousing and Business Intelligence in different perspective

https://ijcsit.com/docs/Volume%205/vol5issue06/ijcsit20140506304.pdf

[2] DATABASE DESIGN – 2ND EDITION

https://opentextbc.ca/dbdesign01/chapter/chapter-5-data-modelling/

[3] Secondary Data Analysis: A Method of which the Time Has Come

http://qqml.net/papers/September_2014_Issue/336QQML_Journal_2014_Johnston_Sept_619-626.pdf

[4] ETL (Extract, Transform, and Load) Process in Data Warehouse

https://www.guru99.com/etl-extract-load-process.html

[5] Design of Data Warehouse and Business Intelligence System

https://www.diva-portal.org/smash/get/diva2:831050/FULLTEXT01.pdf