Data Warehouse Design Plan

Azamat Cherikchiev

Mary McDonald

CTU Online

**Table of Contents**

[Data Warehouse Requirements (Week 1) 3](#_Toc508200089)

[Introduction 3](#_Toc508200090)

[Company Profile 4](#_Toc508200091)

[Problem 4](#_Toc508200092)

[Data Acquisition 5](#_Toc508200093)

[Source data 6](#_Toc508200094)

[Data staging 6](#_Toc508200095)

[Data Storage 7](#_Toc508200096)

[Data warehouse database management systems 7](#_Toc508200097)

[Data Marts 8](#_Toc508200098)

[Metadata 8](#_Toc508200099)

[Information Delivery System 9](#_Toc508200100)

[Multidimensional Databases 9](#_Toc508200101)

[Data Mining 10](#_Toc508200102)

[OLAP (Online Analytical Processing) 11](#_Toc508200103)

[Report and Query 11](#_Toc508200104)

[Control and Management 12](#_Toc508200105)

[Design Requirements (Week 2) 13](#_Toc508200106)

[Introduction 13](#_Toc508200107)

[Conceptual data Model 14](#_Toc508200108)

[Logical data Model 15](#_Toc508200109)

[Physical data Model 15](#_Toc508200110)

[Load Data Analysis (Week 3) 16](#_Toc508200111)

[Introduction 16](#_Toc508200112)

[ETL Process Implementation 17](#_Toc508200113)

[Extraction 17](#_Toc508200114)

[Transforming 17](#_Toc508200115)

[Loading 19](#_Toc508200116)

[Meta Data Logging 19](#_Toc508200117)

[References 20](#_Toc508200118)

# **Data Warehouse Requirements (Week 1)**

## **Introduction**

This project presents the components of data warehouse in the context of the environment of a hotel enterprise. The aim of designing and implementing data warehouse in the hotel enterprise is so as to give the employees and other users of information at the levels of an enterprise a secure, consistent and integrated source of data in which they will be in a position to set and report the requirements of their enterprise efficiently unlike without having a data warehouse. The data warehouse project is intended to be implemented in the case organization that is the Palms Beach Hotel and Spa enterprise and this is discussed in this document as follow.

## **Company Profile**

The enterprise of the Palms Beach and Spa hotel is a business enterprise that offers fitness, recreational, food, and accommodation services to the clients. It is among the largest hotel enterprises in Turkey, Kuwait. It comprises of eight to ten building and provides the services of the hotel enterprises to more than eight thousands customers every day. In Kuwait, this hotel enterprise is located between the South and mid beaches hence it’s in noise free area. The guests of this hotel enterprise experiences easy space, free internet and breakfast services. They also experiences exceptional beaches, free airport transportation and parking services. It also has an inviting hammock, its soaring palm trees’ gardens and ground are clean and the rooms of accommodation are also clean (PBHS, 2017).

 The hotel enterprise has many divisions and it has a Spa, Gym and outdoor pools. This hotel has a primary priority for the awareness and wellness of the environment.It also have a website through which it communicates to its global clients. The business also allows its members to book for the hotel services online. Each sector of this hotel enterprise stores its own data, they only requested data from each department when the need come. They experience difficulty in sharing data within the enterprise between the departments(The Palms Beach Hotel and Spa. 2017).

## **Problem**

In the environment of this hotel enterprise, the design and implementation of the data warehouse aims at solving the multiple systems integration issues from different departments to a common data source. With diverse functions and Services carried out at this hotel enterprise and the available divisions, data about processes, employees, customers and other stakeholders resides in distinct databases in the hotel enterprise. Most information exists of itself and in a silo in, and the data doesn’t translate across the spectrum to enable this hotel enterprise to grow (The Palms Beach Hotel and Spa. 2017).

 The main issue of lacking a data warehouse is that users consume data in the actual reports form. Hence, most of the time the individuals who require information from the data they utilize have to wait for the reports that are based on the schedule of someone else. In addition, once the reports are obtained, the data is manipulated within applications like Microsoft Excel to fit their respective requirements, and most of the time this leads to miscommunication or error. As the public and private companies recover from the fraud of the corporate of previous years, it is of importance that correct presentation of reports is effected, and this calls for the implementation of the current procedures and methodologies like data warehouse. Instead of the managers having to wait for the delivery of the reports, it has to be made easy for them to be able to acquire the information when the need arises.

## **Data Acquisition**

The acquisition of data is the task of extracting data from the systems of operation and placing the data in the desired format for the application of the information that operates data warehouse. The model of data acquisition converts analog waveforms into the digital values for processing. It performs key changes, summarizations, conversions, considerations and structural changes required for transforming disparate data into the information that ca be used by the tools used to support decision making in an organization.It also provides the control statements and programs required to move the data into the data warehouse for multiple systems of operation (Matteo, and Stefano, 2009).The data acquisition performs the following functions;

* Accommodates changes of the source data definition
* Establishes defaults for the missing data
* Removes the data that is unwanted from the operational databases
* Converts data to the common definitions and names of the data.

### **Source data**

 The source data originates from the applications of operation. When data enters into data warehouse is transformed and cleaned up into the integrated format and structure. The process of transformation involves condensation, filtration, summarization and conversion of the data. Since the data contains components of history, the data warehouse have to be in a position of managing and holding large data volumes together with different structures of data over time for the same database(Syed, 2016).

### **Data staging**

It enables one to stage the data required by a task before the execution of the task and to move the data back to the archive after executing that task. The data stored to sources has to be extracted from the sources and then cleaned to fill gaps, inconsistencies are removed and integrated to merge the sources of heterogeneous in a common schema. ETL (Extraction, Transformation and Loading) tools merges the heterogeneous schemata, filter, validate, cleanse, load, transform and extract data source into the data warehouse. Data staging deals with the problems associated with the distributed information systems like management of inconsistent data and incompatible structures of the data. The data staging is everything between data warehouse and source systems. After the data is extracted from the source system it is cleansed, stored and transformed into the data staging area to be load into the data warehouse(Syed, 2016).

## **Data Storage**

The storage of data is how the data is stored and the manner in which the information is presented, manipulated and obtained. It has to conserve data accessibility, reliability and integrity. It ensures that large amounts large amounts of data are structured and stored based on the use or location and regular backup is performed, this enables the information to be available always. In the designing and implementation of the data warehouse in the Palms Beach & Spa hotel enterprise, data storage enables the queries asked by the end users to be effectively answered. It also covers some choices of developing a data warehouse which is established in a specific type of dimensional data mode1 (William, 2002).

### **Data warehouse database management systems**

The systems of managing database in the data warehouse are used to retrieve data from the database. These systems are applications designed to interact with the end user, other applications as well as the databases to analyze and capture data in data warehouse. The database management system is designed to allow the creation, update, querying, definition and the administration of the warehouse databases. The database management systems recommended to be employed in this project of implementing data warehouse in the Palms Beach & Spa hotel enterprise are such as Microsoft SQL server, MySQL, and many others (Rodolfo, 2017).

 The data warehouse is used mostly to facilitate analysis and reporting while the database is used to store the data. Database is a place where the data is stored, and for one to access or to analyze the stored data, there is a need to develop the system of managing the database. The data warehouse necessarily doesn’t need a system of managing the database. The aim of developing a data warehouse is for easy the user’s access to the data, and it is also it may be used in analyzing the data(William, 2002).

### **Data Marts**

The data mart it is the data aggregation that is stored in the main data warehouse. It comprises of a set of pieces of information that is relevant to particular area of the business, group of the users and corporate department. Dependent are data marts that are populated from a primary data warehouse. In data warehouse systems, data marts are useful in midsize to the large firms as they are able to deliver performance since they are smaller than the primary data warehouses. They also mark out information needed by particular category of users to solve queries, and are used as building blocks during the development of incremental of the data warehouse (Matteo, and Stefano, 2009).

### **Metadata**

This refers to the data about the data which describes the data warehouse, and is used in managing, maintaining, using and building the data warehouse. Metadata plays an important role in data warehousing as it identifies the features, usage, values and source of the data of the data warehouse, and explains the way the data can be processed and changed at each layer of the architecture. The repository of the metadata is connected closely to the data warehouse. The applications use the metadata repository in carrying out data analysis and staging tasks. The administrators of the system are interested in the internal metadata as it provides definition of the user profiles, constraints, physical and logical schemata, population policies, processes of transformation, and sources of the data. End users are interested on the external metadata (John, 2013).

 The management of the metadata is provided through the repository of the metadata and software accompanying. The management software of the metadata repository runs on the workstation and is used in mapping the source data to the targeted database, in generating codes for transformations of data, in transforming and integrating the data, and controlling the moving data to the data warehouse.

 As the interaction of the user with the data warehouse increases, the methods of reviewing the requested results for information may be expected to evolve from the manual Analysis of the trends and exceptions, to the initiation driven by an agent of the analysis based on the thresholds defined by the user. The thresholds’ definitions, the software agents’ configuration parameters that uses the thresholds, and information directory that indicates where the desired information sources can be obtained are all stored in the repository of the metadata(Matteo, and Stefano, 2009).

## **Information Delivery System**

The component of information delivery is useful in allowing the subscribing task for the information of the data warehouse and to have it delivered to various destinations as part of some user specific scheduling algorithms. The system of information delivery distributes data stored in the data warehouse to other warehouses and end users’ products like spreadsheets and local databases (Syed, 2016).

 The delivery of information can base on the time or completion of the external event. The rationale of the delivery systems’ component bases on the fact that after the installation of an operational and data warehouse, the users they aren’t required to be aware of its maintenance and location. The value of the data warehouse is maximized when the right information gets to the people who are in need of it and at the place they need it.

### **Multidimensional Databases**

They are database technologies that are implemented by the use of a familiar RDBMS. They are developed to avoid any negative issues placed on the data warehouse by the rational nature pf the data model. The multidimensional databases allows OLAP techniques that belongs to a category of the components of data warehousing grouped as the data mining, reporting, analysis, and query tools (Syed, 2016).

They provide faster response together with additional capabilities for sophisticated processing of the OLAP than RDBMSs, though the multidimensional databases have severe size limitations, and are difficult to integrate with other data warehouse architecture components. For the Palms Beach & Spa Hotel enterprise, I would recommend that the organization design, build and implement the data warehouse on a rational database such as MS SQL server, Oracle, or IBM DB2. The model of the data for the data warehouse has to be based on a dimensional design so as to facilitate scalability and integration, and to enhance performance of the analytical processing(Matteo, and Stefano, 2009).

### **Data Mining**

A modern technique of analyzing data is data mining which is used for statistical data analysis and discovery of the knowledge. Unusual patterns are detected by the statistical data analysis in data and apply mathematical and statistical techniques of modeling in explaining the patterns. The models of data are used in predicting and forecasting. A high complexity level exists in data stored and data interrelations in the data warehouse, which are not easy to discover without data mining. The process of data mining provides insights in the enterprise which can be difficult to discover with multidimensional analysis or reporting and query. It enables businesses to discover insights relating to the enterprise by providing them with answers to the questions that one might not have thought to ask(Matteo, and Stefano, 2009).

### **OLAP (Online Analytical Processing)**

 The main way of exploiting information in a data warehouse is the use of the OLAP that provides end users with the opportunity to explore and analyze date interactively on the basis of the multidimensional model. The users of OLAP are in a position to initiate a complex session of analysis in which every step is the result of the proceeding steps’ outcome. The GUI of the OLAP has to be effective, flexible and easy to use by the end users. Its session comprises of the path of navigation which corresponds to the process of analysis for facts as per distinct viewpoints and different detail levels. The path is then turned into a sequence of queries that are not directly issued but are expressed with reference to the previous query. The query results tend to be multidimensional and the tools of OLAP displays data in tables(Syed, 2016).

### **Report and Query**

The main purpose of the data warehouse to provide business users with the information that helps them in strategic decision making in the enterprise. The query and reporting process involves posing a question that has to be answered, retrieving the data that is considered as relevant from the data warehouse, transforming the retrieved data into the desired context, and displaying the data into the format that is readable. The capability of query and reporting comprises of selecting the data elements that are associated, summarizing those data elements, and grouping the data elements using some category, and presentation of the results.

 To execute this capability, the hotel enterprise will be required to implement the entity relationship model with a denormalized or normalized data structure. The definition of query is the process of taking business hypothesis or questions and translating them in the format of a query that be utilized by a specific decision support tool. The tools of query and reporting can be classified into tools of managed query and reporting tools. The production reporting tools enable enterprises to generate regular operational reports. The report writers tools are inexpensive desktop tools designed for use by the end-users. The managed query tools tend to shield end users from the database and SQL structures complexities by inserting the meta-layer between the database and the users. The analytical requirements of the users of data warehouse exceeds the built-in capabilities of the reporting and query tools(Rizzi and Golfarelli, 2012).

 At the time the query is executed, the query and reporting tool generates the desired language commands for accessing and retrieving the data requested that is then returned in an answer set. The analyst of the data then carries out the manipulations and calculations required on the answer set to achieve the desired results. Then the results are formatted in a display selected for end users’ easy understanding. The report is then delivered to the end users on the desired medium of output that may be visualized on a computer device, printed on paper or audibly presented (Syed, 2016).

## **Control and Management**

The data warehouse is larger than the related operational databases, in size it reaches terabytes depending on the amount of history required to be saved. In the real time, data warehouse isn’t synchronized to the associated to data of operation but are updated regularly when the application needs it. Almost all products of the data warehouse includes gateways to have access to the multiple organization sources of data transparently without having to rewrite the applications to utilize and interpret the data (Matteo, and Stefano, 2009).

 Furthermore, in a heterogeneous environment of the data warehouse, various databases tend to reside on the disparate systems, hence requiring tools of inter-networking. Therefore, there is a need for managing the environment of the data warehouse.

The task of managing the data warehouse involves priority and security management;checks of the data quality; monitoring and controlling updates from the multiple sources of data; updating and managing the metadata; reporting, controlling and editing the status and usage of the data warehouse; data purging, distributing, sub-setting and replicating the data; management and control of the storage, recovery and backup of the data warehouse (Rizzi and Golfarelli, 2012).

# **Design Requirements (Week 2)**

## **Introduction**

The models of logical, physical and conceptual have various differences in data modeling. As both data models contain relationships and entities, they have differences in the functions they are established to perform and the targeted audiences. The system analyst use logical and conceptual data models for data modeling needed and provided by the business angle system, and designer of the warehouse refines the first design to give physical model to be used in presentation of the physical structure of the database for construction of the data warehouse (Rodolfo, 2017).

## **Conceptual data Model**

The conceptual model is used to represent data of an organization. Its function is to indicate as many rules as possible about the interrelationships and the meaning of the data as required. The conceptual data modeling is done typically in parallel with other analysis of requirements and steps of structuring at the time of the system analysis. This conceptual data model is carried out throughout the process of data warehouse development. It is useful for analysis and planning in the data warehouse development(Rodolfo, 2017). In the Palms Beach & Spa hotel enterprise, the conceptual data model will have to contain about 10 to 20 entries and relevant relationships referred to as the group entries. The conceptual modeling of the data is the crucial step in the process of designing a data warehouse. It purely represents the reality of the enterprise’s data. The approaches of designing the conceptual data model that may be used by the hotel enterprise to design this step are the model of enhanced entity relationship (EER) or the object role modeling (ORM) (Varuni, 2016).



## **Logical data Model**

The logical model is the evolution of the model of conceptual data towards the technology of data managing like the relational data warehouses. It is the actual implementation of what was gathered in the conceptual data model. This model defines the relationship between the data structures and elements (John, 2013).



## **Physical data Model**

The physical entity relationship diagram represents the design blueprint of the relational database. This data model represents how the data has to be related and structured in a specific database management system. It is of importance for the enterprise to consider the restriction and convention of the database management system that the enterprise uses at the time of designing physical data model. The proper use of the type of data is required for the use of the words reserved and entity columns should not be used in naming the entities and columns. To the design of the data warehouse the designers can add foreign keys, primary keys and constraints (Varuni, 2016).

It is the representation of the data design that takes into account the constraints and facilities of the existing database management system. This model represents the way the models will be designed in the data warehouse. The physical data model indicates the structures of the table, primary key, tables’ relationships, constraints and foreign keys (Rodolfo, 2017).



# **Load Data Analysis** (Week 3)

## **Introduction**

In the process of ETL, the data from the database of OLTP is extracted,then it’s transformed to resemble the schema of the data warehouse, lastly the data loaded in the database of the data warehouse. Data is incorporated by various data warehouses from the systems that are not OLTP for example legacy systems, and spreadsheets. The data has to be extracted from the source systems, transformed and finally the data is loaded in the desired database. The Microsoft® SQL Server™ provides enhancements to the available performance and capabilities, and establishes new features, which makes deployment, maintenance and development functions of the ETL process easier and simpler (Edward, 2014).

## **ETL Process Implementation**

### **Extraction**

Having previously designed the operational database as the primary source of the data, data is extracted from the source, then stored temporary into the buffer area. Some of the Services performed on the data include scrubbing, cleansing and reconciling, fixing errors of data entry. Then the data is transformed in a standard that is normalized. After the data has been transformed, it is then indexed and loaded in the data warehouse. This phase is where creation of new tables, addition of new columns, dropping of tables and discarding of columns takes place. The ETL staging database’s SELECT queries are executed against the source of OLTP system, and then imported into the ETL database tables (Boon, and Ezam, 2014).

### **Transforming**

In the case study hotel enterprise, the Transact-SQL technique is proposed to be used in implementing the ETL functional phases. Transformation step aims at preparing the data for loading in the data warehouse. It is performed once the data is extracted from the source systems. Transformation will be carried out in the staging database of ETL in data from various sources of data, the data will be formatted and organized in required format. The transformations and queries of SQL required for the establishment and data loading in the data warehouse are carried out in this step (Youssef, 2012).

**Dropping Tables**

Since decision making in this particular hotel enterprise dwell on the trends associated with booking rate, behavior and history of the enterprise, and not on the assets and items tables, the asset tables may be discarded from the data warehouse as shown below

DROP TABLE assets;

DROP TABLE item;

 **Merging Tables**

The department, service and room tables lack relevant parameters for extracting important patterns and discovering performance. Therefore, the tables will be merged together with the “transcript\_fact\_table,” table, as shown;

SELECT co\_name FROM course, room, transcript WHERE tr\_id = n

AND str\_booking/day = se\_booking/day AND tr\_se\_num = se\_num AND se\_code = co\_code;

ALTER TABLE transcript fact table ADD co\_room TEXT;

DROP TABLE department, service, room;

In addition, table “Services” is merged with table “bookingServices” and a new table is produced called “bookedServices”.

SELECT act\_name FROM Services, bookingServices

WHERE book\_act\_id = act\_id;

**New Columns**

At the time of transformation new columns may be added. To enhance the quality of the services offered by the hotel enterprise, tr\_serviceDifficulty will be added to table “transcript\_fact\_table.”

 ALTER TABLE transcript\_fact\_table ADD tr\_serviceDifficulty TEXT;

Furthermore, a Boolean column will be added to table “booked” named re\_paidOnDueDate ALTER TABLE booked (re\_paidOnDueDate);

 **Removing Columns**

During the process of transformation, the columns that will be seen as unnecessary will be removed as shown below;

ALTER TABLE Booked REMOVE re\_dueDate

 REMOVE re\_dateOfPayment;

ALTER TABLE Services REMOVE ac\_supervisor;

ALTER TABLE Guest REMOVE st\_phone, st\_email;

### **Loading**

After extracting the data from various sources, then combining and transforming it, the function of data loading takes place. Loading comprises of insertion of records into fact tables and data warehouse dimensions. The function of data loading into the data warehouse tables takes place in this phase of the ETL process (Boon, and Ezam, 2014).

### **Meta Data Logging**

Meta data is sourced from various ETL process operations. The Meta data logging implementation in this hotel enterprise will depend on how particular tasks are performed. The tasks developed using Transact-SQL technique, Meta data will be captured with the statements of this technique in the task processes. Recording and capturing of the Meta data that documents the functional tasks and areas of the ETL operation are performed in this phase. It includes identification of data that moves via the ETL system as well as the ETL tasks efficiency (Edward, 2014).

# **References**

Boon K., and Ezam S., (2014). Design and implementation of data warehouse with data model using survey-based services data. Retrieved from: <http://ieeexplore.ieee.org/document/6927748/>

Edward M., (2014). Design and Implementation of an Enterprise Data Warehouse. Retrieved from:<https://epublications.marquette.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1118&context=theses_open>

John W., (2013). DATA WAREHOUSING FUNDAMENTALS. <https://anuradhasrinivas.files.wordpress.com/2013/03/data-warehousing-fundamentals-by-paulraj-ponniah.pdf>

Matteo G., and Stefano R., (2009). Data Warehouse Design: Modern Principles and Methodologies. Retrieved at: https://cs09lects.files.wordpress.com/2012/12/book-data-warehouse-design-golfarelli-\_-rizzi.pdf

Rizzi and Golfarelli, (2012). Introduction to data warehouse. Retrieved from: http://cdn.ttgtmedia.com/searchDataManagement/downloads/Data\_Warehouse\_Design.pdf

Rodolfo R., (2017). Relational Databases for Data Warehousing. Retrieved from: <https://www.logisticsonline.com/doc/relational-databases-for-data-warehousing-0001>

Syed A., (2016). Componenta of a data warehouse. Retrieved from: <http://tdan.com/components-of-a-data-warehouse/4213>

The Palms Beach Hotel and Spa. (2017). The Palms Beach and Spa Hotel Profile. Available at: <http://www.thepalms-kuwait.com/Default.aspx?pageId=0>

Varuni M., (2016). Data warehouse: Data Modeling. Retrieved from: <https://www.umsl.edu/~sauterv/analysis/Fall2010Papers/varuni/>

William M., (2002). Choosing DBMS for Data Warehouse. Retrieved from: <http://assets.teradata.com/resourceCenter/downloads/WhitePapers/sybaseWP3.pdf>

Youssef B., (2012). A Data Warehouse Design for a Typical University Information System. Journal of Computer Science & Research (JCSCR) - ISSN 2227-328X. Retrieved from: https://arxiv.org/ftp/arxiv/papers/1212/1212.2071.pdf