INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 9, Issue 14, (2019) E-ISSN: 2222-6990

Data Warehouse Design and Implementation Based on Star Schema vs. Snowflake Schema

K. I. Mohammed

Department of Computing, Universiti Pendidikan Sultan Idris, Tanjong Malim
Perak, Malaysia
Email: Khalid_ib81@yahoo.com

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v9-i14/6502 DOI:10.6007/IJARBSS/v9-i14/6502

Published Date: 23 October 2019

Abstract

The data warehouses are considered modern ancient techniques, since the early days for the relational databases, the idea of the keeping a historical data for reference when it needed has been originated, and the idea was primitive to create archives for the historical data to save these data, despite of the usage of a special techniques for the recovery of these data from the different storage modes. This research applied of structured databases for a trading company operating across the continents, has a set of branches each one has its own stores and showrooms, and the company branch's group of sections with specific activities, such as stores management, showrooms management, accounting management, contracts and other departments. It also assumes that the company center exported software to manage databases for all branches to ensure the safety performance, standardization of processors and prevent the possible errors and bottlenecks problems. Also the research provides this methods the best requirements have been used for the applied of the data warehouse (DW), the information that managed by such an applied must be with high accuracy. It must be emphasized to ensure compatibility information and hedge its security, in schemes domain, been applied to a comparison between the two schemes (Star and Snowflake Schemas) with the concepts of multidimensional database. It turns out that Star Schema is better than Snowflake Schema in (Query complexity, Query performance, Foreign Key Joins),And finally it has been concluded that Star Schema center fact and change, while Snowflake Schema center fact and not change.

Keywords: Data Warehouses, OLAP Operation, ETL, DSS, Data Quality.

Introduction

A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions. The data warehouse contains granular corporate data. Data in the data warehouse is able to be used for many different purposes, including sitting and waiting for future requirements which are unknown today (Inmon, 2005). Data warehouse provides the primary support for Decision Support Systems (DSS) and Business Intelligence (BI) systems. Data warehouse, combined with On-Line Analytical

Processing (OLAP) operations, has become and more popular in Decision Support Systems and Business Intelligence systems. The most popular data model of Data warehouse is multidimensional model, which consists of a group of dimension tables and one fact table according to the functional requirements (Kimball, Reeves, Ross, & Thornthwaite, 1998). The purpose of a data warehouse is to ensure the appropriate data is available to the appropriate end user at the appropriate time (Chau, Cao, Anson, & Zhang, 2003). Data warehouses are based on multidimensional modeling. Using On-Line Analytical Processing tools, decision makers navigate through and analyze multidimensional data (Prat, Comyn-Wattiau, & Akoka, 2011).

Data warehouse uses a data model that is based on multidimensional data model. This model is also known as a data cube which allows data to be modeled and viewed in multiple dimensions (Singhal, 2007). And the schema of a data warehouse lies on two kinds of elements: facts and dimensions. Facts are used to memorize measures about situations or events. Dimensions are used to analyze these measures, particularly through aggregations operations(counting, summation, average, etc.) (Bhansali, 2009; J. Wang, 2009). Data Quality (DQ) is the crucial factor in data warehouse creation and data integration. The data warehouse must fail and cause a great economic loss and decision fault without insight analysis of data problems (Yu, Xiao-yi, Zhen, & Guo-quan, 2009). The quality of data is often evaluated to determine usability and to establish the processes necessary for improving data quality. Data quality may be measured objectively or subjectively. Data quality is a state of completeness, validity, consistency, timeliness and accuracy that make data appropriate for a specific use (Manjunath, Hegadi, & Ravikumar, 2011).

Data quality has been defined as the fraction of performance over expectancy, or as the loss imparted to society from the time a product is shipped (Besterfield, Besterfield-Michna, Besterfield, & Besterfield-Sacre, n.d.)]. The believe was the best definition is the one found in (Orr, 1998; Tayi & Ballou, 1998; R. Y. Wang & Strong, 1996): data quality is defined as "fitness for use". The nature of this definition directly implies that the concept of data quality is relative. For example, data semantics is different for each distinct user. The main purpose of data quality is about horrific data - data which is missing or incorrect or invalid in some perspective. A large term is that, data quality is attained when business uses data that is comprehensive, understandable, and consistent, indulging the main data quality magnitude is the first step to data quality perfection which is a method and able to understand in an effective and efficient manner, data has to satisfy a set of quality criteria. Data gratifying the quality criterion is said to be of high quality (Manjunath et al., 2011). This paper is divided into seven sections. Section 1 introduction, Definition of Data Warehouse and The Quality of Data Warehouse. Section 2 presents related work, Section 3 presents Data Warehouse Creation and the main idea is that a Data warehouse database gathers data from an overseas trading company databases. Section 4 describes Data Warehouse Design For this study, we suppose a hypothetical company with many branches around the world, each branch has so many stores and showrooms scattered within the branch location. Each branch has a database to manage branch information. Section 5 describes our evaluation Study of Quality Criteria for DW, which covers aspects related both to quality and performance of our approach, and the obtained results, and work on compare between star schema and snowflake schema. Section 6 provides conclusions. Finally, Section 7 describes open issues and our planned future work.

Relat Work

In this section we will review related work in Data Warehouse Design and Implementation Based on Quality Requirements. We will start with the former. The paper introduced by (Vassiliadis, Bouzeghoub, & Quix, 2000), The proposed approach covers the full lifecycle of the data warehouse, and allows capturing the interrelationships between different quality factors and helps the interested user to organize them in order to fulfill specific quality goals. Furthermore, they prove how the quality management of the data warehouse can guide the process of data warehouse evolution, by tracking the interrelationships between the components of the data warehouse. Finally, they presented a case study, as a proof of concept for the proposed methodology. The paper introduced by (Santoso & Gunadi, 2007), this paper describes a study which explores modeling of the dynamic parts of the data warehouse. This metamodel enables data warehouse management, design and evolution based on a high level conceptual perspective, which can be linked to the actual structural and physical aspects of the data warehouse architecture. Moreover, this metamodel is capable of modeling complex activities, their interrelationships, the relationship of activities with data sources and execution details. The paper introduced by (AbuAli & Abu-Addose, 2010), The aim of this paper is to discover the main critical success factors(CSF) that leaded to an efficient implementation of DW in different organizations, by comparing two organizations namely: First American Corporation (FAC) and Whirlpool to come up with a more general (CSF) to guide other organizations in implementing DW efficiently. The result from this study showed that FAC Corporation had greater returns from data warehousing than Whirlpool. After that and based on them extensive study of these organizations and other related resource according to CSFs, they categorized these (CSF) into five main categories to help other organization in implementing DW efficiently and avoiding data warehouse killers, based on these factors. The paper introduced by (Manjunath & Hegadi, 2013), The proposed model evaluates the data quality of decision databases and evaluates the model at different dimensions like accuracy derivation integrity, consistency, timeliness, completeness, validity, precision and interpretability, on various data sets after migration. The proposed data quality assessment model evaluates the data at different dimensionsto give confidence for the end users to rely on their businesses. Author extended to classify various data setswhich are suitable for decision making. The results reveal the proposed model is performing an average of 12.8 percent ofimprovement in evaluation criteria dimensions with respect to the selected case study.

Data Warehouse Creation

The main idea is that a Data warehouse database gathers data from an overseas trading company databases. For each branch of the supposed company we have a database consisting of the following schemas:

- Contracting schema consists a contract and contractor date.
- Stores schema managing storing information.
- Showrooms schema to manage showrooms information for any branch of the supposed company.
- At the top of the above schemas, an accounting schema was installed which manages all accounting operations for any branch or the while company.

All information is stored into fully relational tables according to the known third normal form. The data integrity is maintained by using a foreign keys relationship between related tables,

non-null constraints, check constraints, and oracle database triggers are used for the same purpose. Many indexes are created to be used by oracle optimizer to minimize DML and query response time. Security constraints are maintained using oracle privileges. Oracle OLAP policy is taken in consideration.

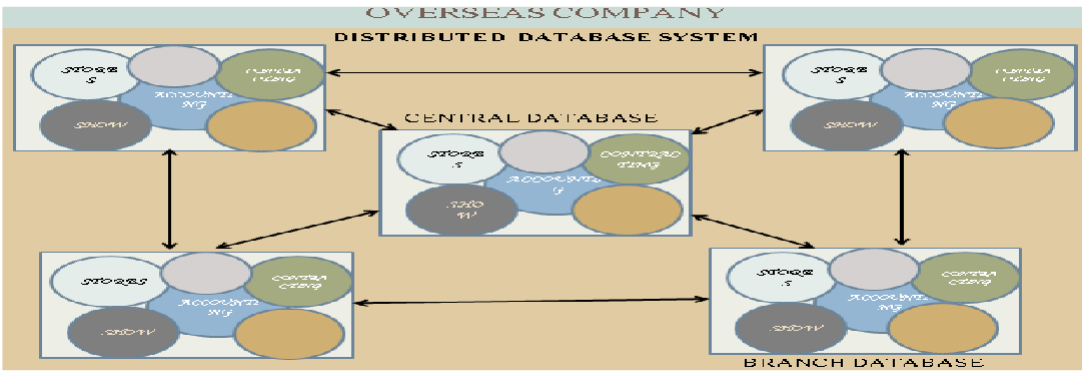
Data Warehouse Design

As mentioned above a warehouse home is installed on the same machine. The data warehouse is stored on a separate oracle tablespaces and configured to use the above relational online tables as a source data. So the mentioned schemas are treated as data locations. Oracle warehouse builder is a java program, which are used warehouse managements. The locations of data sources are:

1. Accounting schema.
2. Stores schema.
3. Contracting schema.
4. Showrooms schema.

For this study, we suppose a hypothetical company with many branches around the world, each branch has so many stores and showrooms scattered within the branch location. Each branch has adatabase to manage branch information. Within each supposed branch database there are the following schemas which work according to OLAP policies and maintain securities and data integrity. The schemas are: Accounting schema, Contracting schema, Stores schema and showrooms schema. All branches databases are connected to each other over WAN.

Figure 1. Distributed database for hypothetical company.



Overseas company, as a base (or a source) for warehouse database. This paper supposes that each node belongs to one company branch, so each branch controls its own data. The main office of the

company, controls the while data also with the central node. The warehouse database could be at the central node, as the company needs. We suppose that all nodes use the same programs, which are applied the database(s). Within each node, each activity is represented by a database schema, i.e. stores, showrooms, contracting, and other schemas. The core of all schemas is the accounting schema. According to jobs load, each schema could be installed on a separate database or on the same database. All related databases around company branches are cooperated within the same WAN.

Study of Quality Criteria for DW

In this study, we will carry out some of the criteria, and these criteria are:

Data Warehouse Using Snowflake Schema

Using oracle warehouse policies, each database has the following snow flaking modules:

- Sales module.
- Supplying module.

Sales Module

It consist of the following relational tables

Table 1

Explain the relational table

Sales_sh	Fact table	Showrooms
showrooms	Dimensional table	Showrooms
Items	Dimensional table	Accounting
Currencies	Dimensional table	Accounting
Customers	Dimensional table	Showrooms
Locations	Dimensional table	Accounting

The following diagram depicts the relations between the above dimensional and fact tables

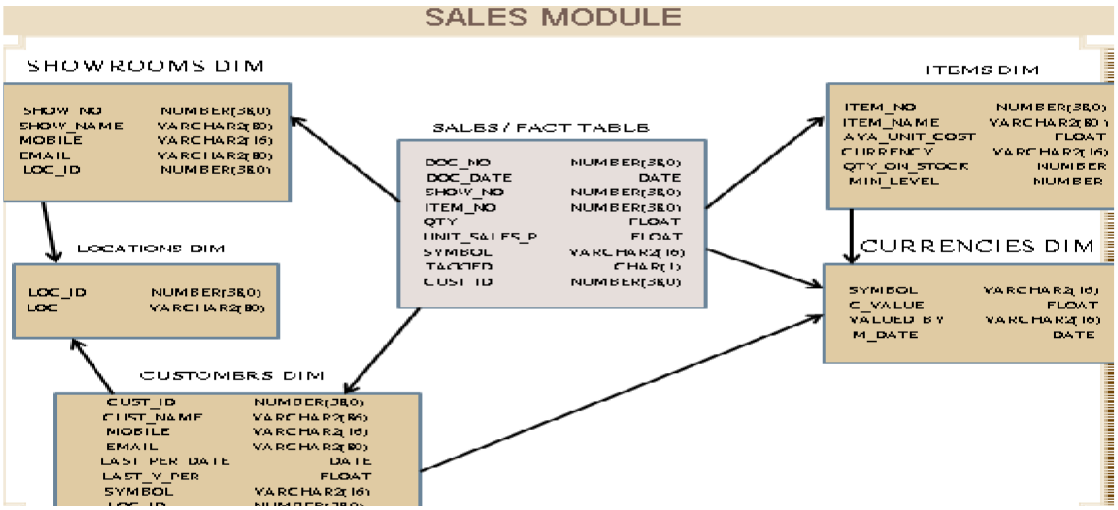


Figure 2. Sales module.

Figure 2 above, represents all entities within the sales module. Any entity is designed using the Third Normal Form (3NF) rule, so it has a primary key. The most important tools used to implement integrity and validations are oracle constraints. After supplying data to the above module, and transferring it to oracle warehouse design center, retrieving data (557,441 rows) from fact table sales

which are shown in the following figure 3, which mentions the detailed information for each single sales within each showroom, and location. It consists of: Voucher (doc) no. And date, the sold item, sold quantity and price. This data is available at the corresponding node (branch) and the center. Of course, the same data would be transferred to warehouse database for historical purpose.

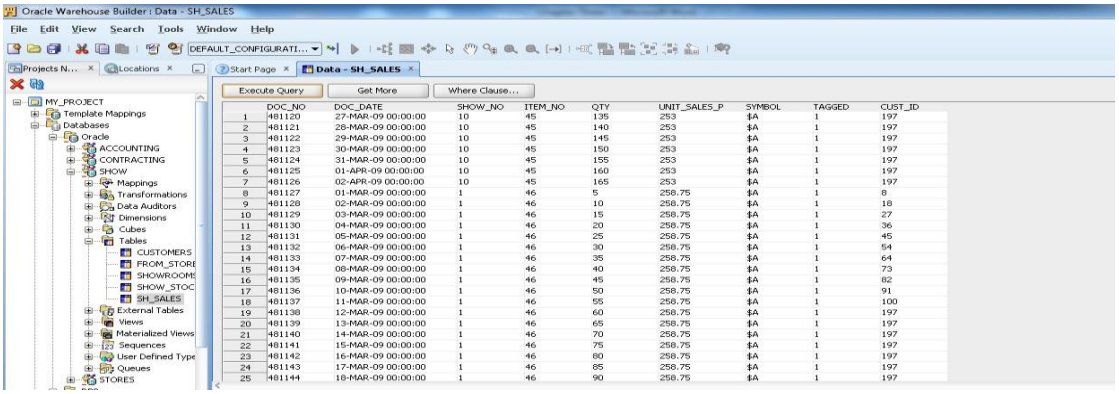


Figure 3. Sales data.

Customers dimensional table consists of some personal date about the customer, like mobile, email, and location which is useful to contact him. Also it indicates the date of last purchase, and the total amount purchased for last year. This data is available for the corresponding node and the center; also it refers to the warehouse database. (See figure 4 customer data) The dimensional table of customers would is shown below.

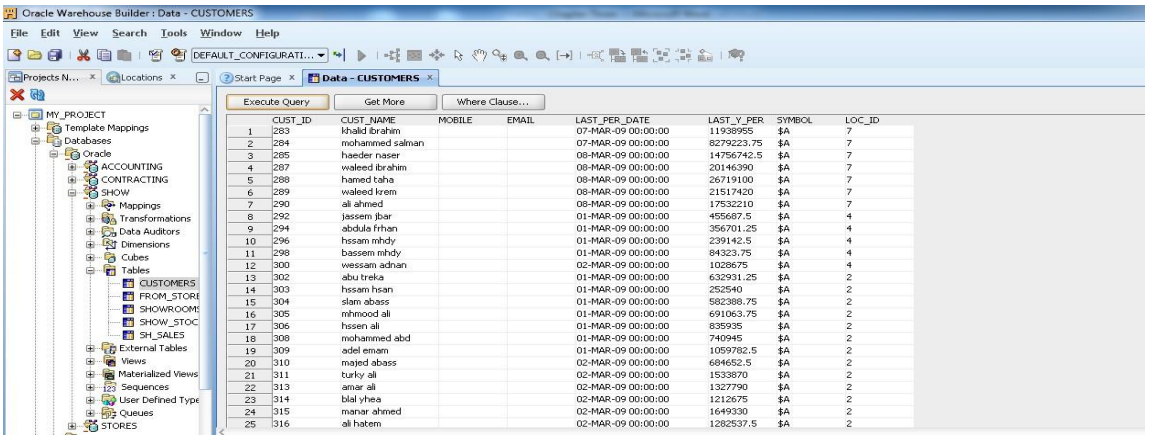


Figure 4. Customer's data.

Supplying Module

Supplying the company by materials according to company usual contracts is managed by this module according to snowflake design. It consists of the following relational tables.

Table 2
Within supplying module

STR_RECIEVING	Fact table	Stores
Contracts	Dimensional table	contracting
Items	Dimensional table	Accounting
Currencies	Dimensional table	Accounting
Stores	Dimensional table	Stores
Locations	Dimensional table	Accounting
Daily_headers	Dimensional table	Accounting

The following diagram Fig 5 depicts the relations between the above dimensional and fact tables. They are obey 3NF rule, so they have their primary key constraints, and constrained to each other using foreign keys constraints. The fact table STR_RECIEVING consists of all charges information received at company stores (contented by stores table owned by stores schema), according to the contracts (contented by contracts table owned by contracting schema). Daily headers dimensional table represent the accounting information for each contract. Using oracle triggers when new record inserted into the STR_RECIEVING fact table, some other accounting data would be created into details table row related (through foreign key) to Daily headers dimensional table. Also any charge value could be converted to the wanted currency using the data maintained by currencies dimensional table owned by accounting schema.

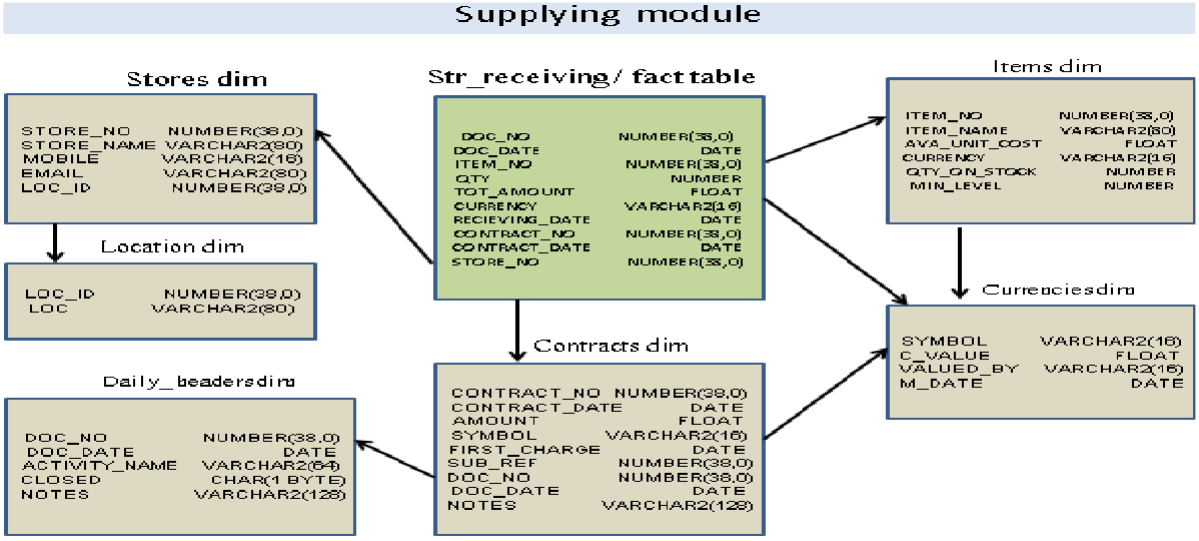


Figure 5. Supplying module.

For security reasons, direct excess to object fact table is not allowed, un imaginary view is created (named str_recieving_v) , then all users are allowed to generate a DML (data manipulation language instructions) on this view. A certain piece of code (oracle trigger) is written to manipulate data, according to server policies (data integrity and consistency) as user supplies data to the imaginary view. After supplying data to the above module, and

transferring it to oracle warehouse design center, retrieving data (415,511rows) from fact table str_recieving as shown in the following figure.

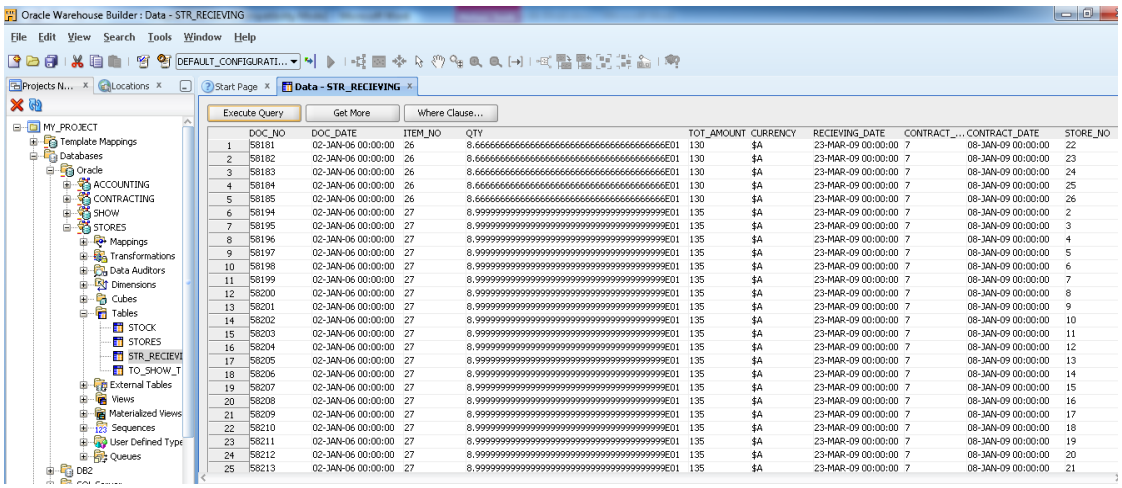


Figure 6. Received charges on str_recieving fact table.

During charges insertion, a background process (oracle trigger) should update the stock dimension data to reflect the latest information about quantities in stock at each node and the center. The stock data contains the quantity balance, quantity in and out for the current year. It's available at the corresponding node and the center, at the online database, and at warehouse database for previous years. Stock data could be like figure 7 as viewed by oracle SQL developer.

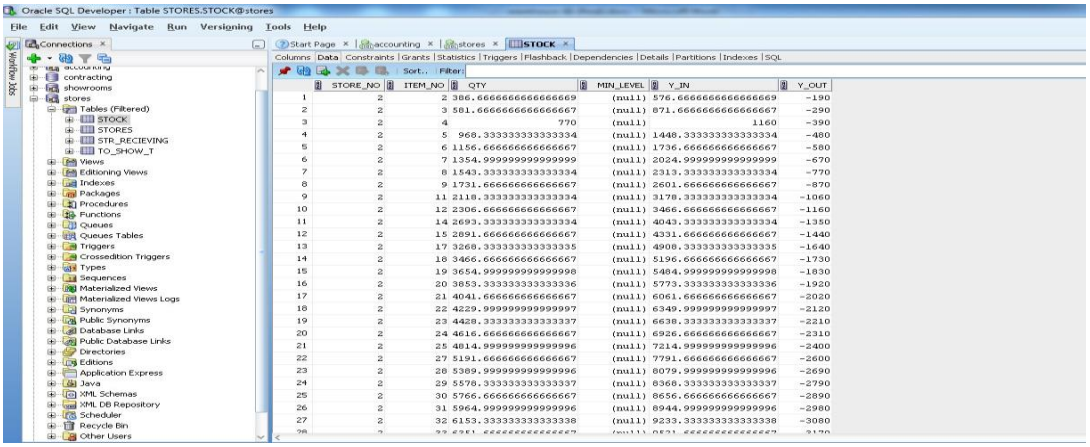


Figure 7. Stock dimensional table data.

Data Warehouse Using Star Schema

As a study case using warehouse star schema, we have:

- Stocktaking module.
- Accounting module.

Stocktaking Module

Its manage for the current stock within any store for the company, its data is determined by the following table.

Table 3
Cooperated within stocktaking module

Stock	Fact table	Stores
Items	Dimensional table	Accounting
Stores	Dimensional table	Stores
Currencies	Dimensional table	Accounting
showrooms	Dimensional table	Showrooms
Locations	Dimensional table	Accounting
contracts	Dimensional table	Contracting
Str_recieving	Fact table	Stores
To_show_t	Fact table	Stores

The stock fact table stands for the actual stock balances within each store belongs to each branch, and the whole company at the center. The following diagram depicts the relations between the below dimensions.

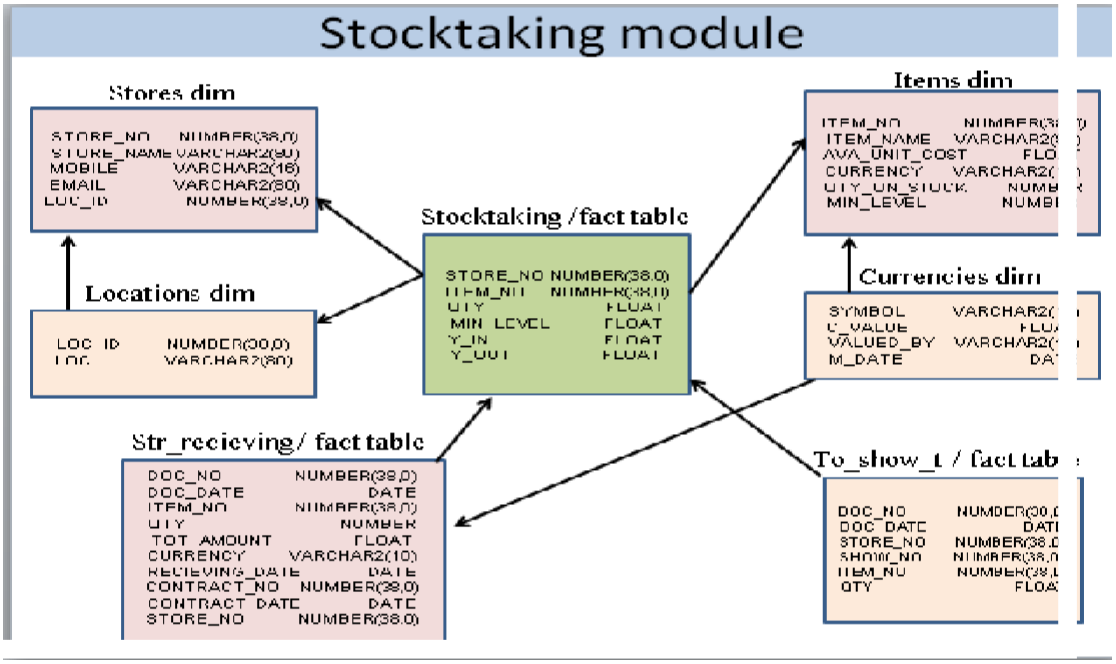
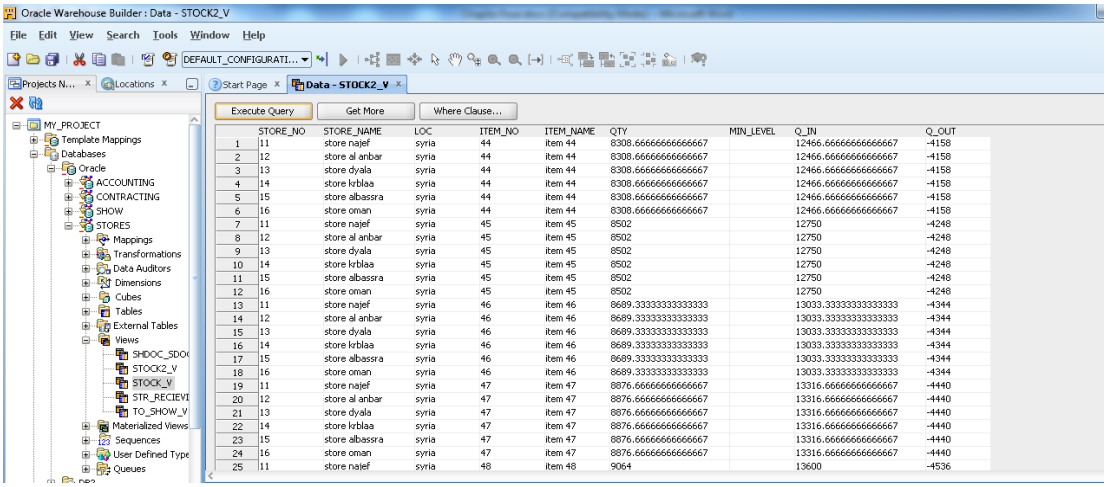


Figure 8. Stocktaking module as a warehouse star schema.

DML (data manipulation language instructions) is done on stock fact table through oracle triggers which are the most trusted programs to maintain the highest level of integrity and security, so the imaginary view (named stock_v) was created, users are allowed to supply data to that view, then the server would process the supplied data using oracle trigger. Querying the renormalized stock fact table within the star schema module, using oracle design center is depicted as below (no. of rows on stock table within our study case is 15,150). This figure 9 query execution is allowed for all users (public).



	STORE_NO	STORE_NAME	LOC	ITEM_NO	ITEM_NAME	QTY	MIN_LEVEL	Q_IN	Q_OUT
1	11	store najef	syria	44	item 44	8308.66666666666667		12466.66666666666667	-4158
2	12	store al anbar	syria	44	item 44	8308.66666666666667		12466.66666666666667	-4158
3	13	store dyala	syria	44	item 44	8308.66666666666667		12466.66666666666667	-4158
4	14	store krblaa	syria	44	item 44	8308.66666666666667		12466.66666666666667	-4158
5	15	store albassra	syria	44	item 44	8308.66666666666667		12466.66666666666667	-4158
6	16	store oman	syria	44	item 44	8308.66666666666667		12466.66666666666667	-4158
7	11	store najef	syria	45	item 45	8502		12750	-4248
8	12	store al anbar	syria	45	item 45	8502		12750	-4248
9	13	store dyala	syria	45	item 45	8502		12750	-4248
10	14	store krblaa	syria	45	item 45	8502		12750	-4248
11	15	store albassra	syria	45	item 45	8502		12750	-4248
12	16	store oman	syria	45	item 45	8502		12750	-4248
13	11	store najef	syria	46	item 46	8689.33333333333333		13033.33333333333333	-4344
14	12	store al anbar	syria	46	item 46	8689.33333333333333		13033.33333333333333	-4344
15	13	store dyala	syria	46	item 46	8689.33333333333333		13033.33333333333333	-4344
16	14	store krblaa	syria	46	item 46	8689.33333333333333		13033.33333333333333	-4344
17	15	store albassra	syria	46	item 46	8689.33333333333333		13033.33333333333333	-4344
18	16	store oman	syria	46	item 46	8689.33333333333333		13033.33333333333333	-4344
19	11	store najef	syria	47	item 47	8876.66666666666667		13316.66666666666667	-4440
20	12	store al anbar	syria	47	item 47	8876.66666666666667		13316.66666666666667	-4440
21	13	store dyala	syria	47	item 47	8876.66666666666667		13316.66666666666667	-4440
22	14	store krblaa	syria	47	item 47	8876.66666666666667		13316.66666666666667	-4440
23	15	store albassra	syria	47	item 47	8876.66666666666667		13316.66666666666667	-4440
24	16	store oman	syria	47	item 47	8876.66666666666667		13316.66666666666667	-4440
25	11	store najef	syria	48	item 48	9064		13600	-4536

Figure 9. Stocktaking on oracle warehouse design center.

Accounting module

One of the most importance aspects of accounting functions is the calculations of the daily cash within each showroom belongs to the company. The daily totals for each branch and the grand total could be calculated. Timely based cash could be accumulated later on demand.

Table 4

The tables needed for this activity

Daily cash	**Fact table**	**Accounting**
Show_sales	**Fact table**	**Accounting**
Showrooms	**Dimensional table**	**Showroom**
Currencies_tab	**Dimensional table**	**Accounting**
Locations	**Dimensional table**	**Accounting**
Customers	**Dimensional table**	**showrooms**

The daily cash is a view used to reflect the actual cash with each showroom on daily base.

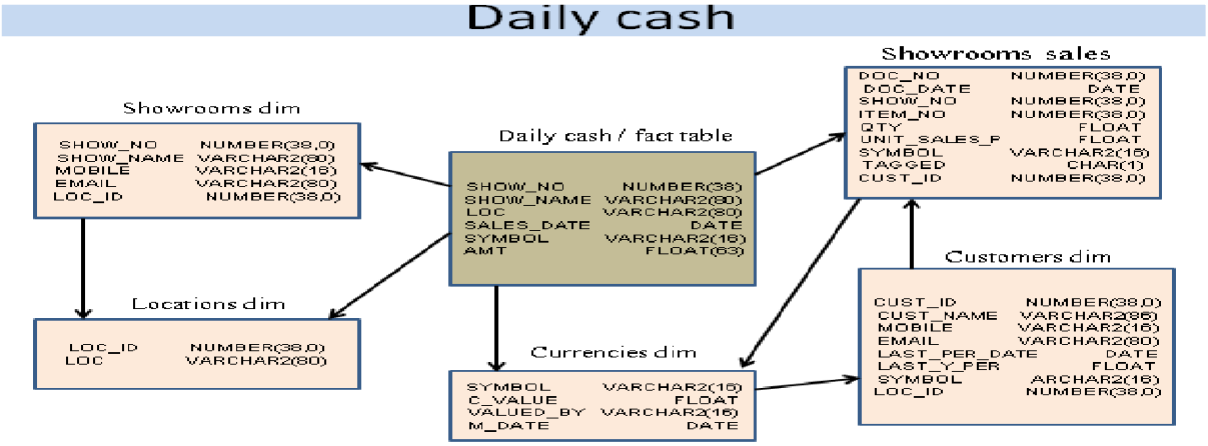


Figure 10. Daily cash using warehouse star schema.

Using inner SQL joins, one could retrieve data about daily cash as follows.

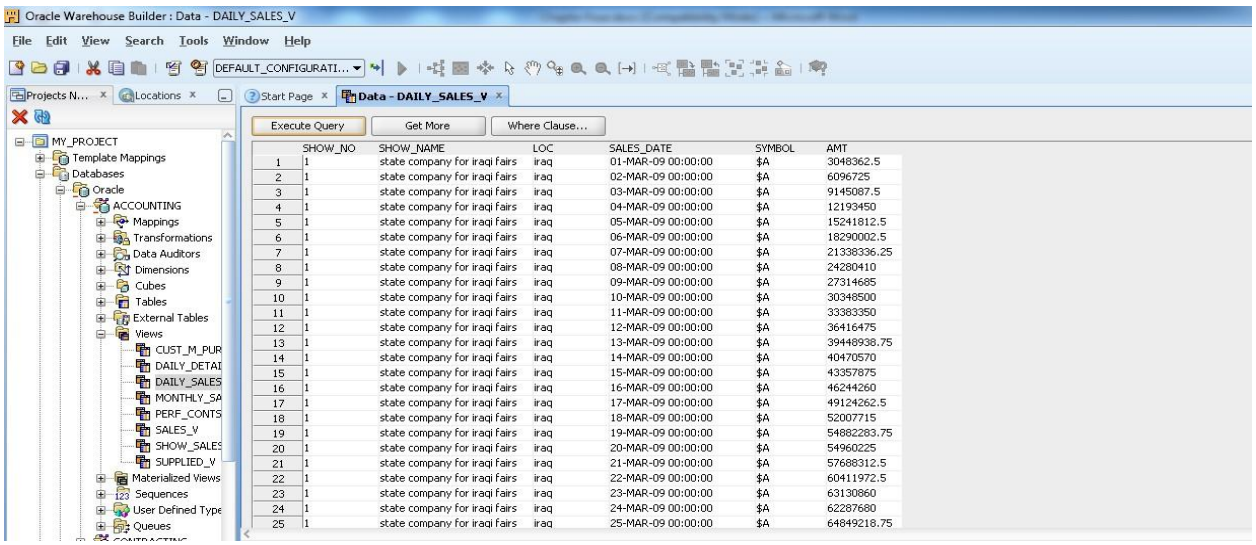


Figure 11. Grand daily cash as depicted by Oracle warehouse design center.

Conclusions

The following expected conclusions have been drawn:

1. Reduce the query response time and Data Manipulation Language and using many indexes which are created to be used by oracle optimizer.
2. Star Schema is best of them Snowflake Schema the following points are reached:

- Query complexity: Star Schema the query is very simple and easy to understand, while Snowflake Schema is more complex query due to multiple foreign key which joins between dimension tables .
- Query performance: Star Schema High performance. Database engine can optimize and boost the query performance based on predictable framework, while Snowflake Schema is more foreign key joins; therefore, longer execution time of query in compare with star schema.
- Foreign Key Joins: Star Schema Fewer Joins, while Snowflake Schema has higher

number of joins.
And finally it has been concluded that Star Schema center fact and change, while Snowflake Schema center fact and not change.

Future Works

1. Using any other criteria in development implementation of the proposed system.
2. Using statistical methods to implement other criteria of Data Warehouse.
3. Applying algorithm Metadata and comparing between bitmap index and b-tree index.
4. Applying this work for a real organization not prototype warehouse.
5. Take advantage of the above standards in improving the performance of the use of the data warehouse and institutions according to their environment.

References

AbuAli, A. N., & Abu-Addose, H. Y. (2010). Data warehouse critical success factors. *European Journal of Scientific Research*, *42*(2), 326–335.

Besterfield, D. H., Besterfield-Michna, C., Besterfield, G. H., & Besterfield-Sacre, M. (n.d.) (2009). Total quality management. 1995. Prentice-Hall Inc, Englewood Cliffs.

Bhansali, N. (2009). *Strategic data warehousing: achieving alignment with business*. Auerbach Publications.

Chau, K. W., Cao, Y., Anson, M., & Zhang, J. (2003). Application of data warehouse and decision support system in construction management. *Automation in Construction*, *12*(2), 213–224.

Inmon, H. W. (2005). *Building the data warehouse, Fourth Edition Published by Wiley Publishing, Inc., Indianapolis, Indiana*. John wiley & sons.

Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*. John Wiley & Sons.

Manjunath, T. N., & Hegadi, R. S. (2013). Data Quality Assessment Model for Data Migration Business Enterprise. *International Journal of Engineering and Technology (IJET)*, *5*(1).

Manjunath, T. N., Hegadi, R. S., & Ravikumar, G. K. (2011). Analysis of data quality aspects in datawarehouse systems. *International Journal of Computer Science and Information Technologies*, *2*(1), 477–485.

Orr, K. (1998). Data quality and systems theory. *Communications of the ACM*, *41*(2), 66–71.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2011). Combining objects with rules to represent aggregation knowledge in data warehouse and OLAP systems. *Data & Knowledge Engineering*, *70*(8), 732–752.

Santoso, L. W., & Gunadi, K. (2007). A proposal of data quality for data warehouses environment.
Jurnal Informatika, *7*(2), 143–148.

Singhal, A. (2007). *Data warehousing and data mining techniques for cyber security* (Vol. 31). Springer Science & Business Media.

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, *41*(2), 54–57.

Vassiliadis, P., Bouzeghoub, M., & Quix, C. (2000). Towards quality-oriented data warehouse usage and evolution. *Information Systems*, *25*(2), 89–115.

Wang, J. (2009). Encyclopedia of Data Warehousing and Mining, Second Edition. *Published by Information Science Reference. United States of America*, *I A-Data P*.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Yu, H., Xiao-yi, Z., Zhen, Y., & Guo-quan, J. (2009). A universal data cleaning framework based on user model. In *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on* (Vol. 2, pp. 200–202). IEEE.