International Journal of Computer Applications Technology and Research
Volume 4– Issue 8, 584 - 591, 2015, ISSN: 2319–8656

Enhancing Data Staging as a Mechanism for Fast Data Access

Reagan Muriithi Gatimu
School of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture and Technology (JKUAT)
Nairobi, Kenya

Wilson Cheruiyot
School of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture and Technology (JKUAT)
Nairobi, Kenya

Michael Kimwele
School of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture and Technology (JKUAT)
Nairobi, Kenya

Abstract: Most organizations rely on data in their daily transactions and operations. This data is retrieved from different source systems in a distributed network hence it comes in varying data types and formats. The source data is prepared and cleaned by subjecting it to algorithms and functions before transferring it to the target systems which takes more time. Moreover, there is pressure from data users within the data warehouse for data to be availed quickly for them to make appropriate decisions and forecasts. This has not been the case due to immense data explosion in millions of transactions resulting from business processes of the organizations. The current legacy systems cannot handle large data levels due to processing capabilities and customizations. This approach has failed because there lacks clear procedures to decide which data to collect or exempt. It is with this concern that performance degradation should be addressed because organizations invest a lot of resources to establish a functioning data warehouse. Data staging is a technological innovation within data warehouses where data manipulations are carried out before transfer to target systems. It carries out data integration by harmonizing the staging functions, cleansing, verification, and archiving source data. Deterministic Prioritization Approach will be employed to enhance data staging, and to clearly prove this change Experiment design is needed to test scenarios in the study. Previous studies in this field have mainly focused in the data warehouses processes as a whole but less to the specifics of data staging area.

Keywords: Data Staging; Source System; Deterministic Prioritization; Data Warehouse

1. INTRODUCTION

The growing number of business transactions in any enterprise is directly proportional to growth of data size. This data comes from variant source systems and applications and needs to be organized in a workable state so that it remains relevant and meaningful to the users. Technological development has led to the rise of Data Warehouse (DW). (Inmon, 2002) defines a data warehouse as "collection of integrated, subject-oriented databases designated to support the decision making process". Both (Kimball and Inmon, 2002) agree that a DW has to be integrated, subject-oriented, nonvolatile and time variant. This concept of time-variance is so crucial and ultimate concern and sets the basis for this research since it focuses on improved data access. The foundations of a DW as explained by (Zineb, Esteban, Jose-Norberto, Juan, 2011) encompass integration of multiple different data sources. This allows the provision of complete and correct view of the enterprise operational data which is synthesized into a set of strategic indicators and measures that the users of the data can associate with.

DW has business intelligence implemented in three major processes used to prepare data to match user's needs. They are commonly referred to as ETL processes namely; Extraction, Transformation and Loading. Extraction process retrieves data as is from source systems before subjecting it to any manipulations. Transformation process also referred to as transportation phase is the operational base and the most intriguing of all. Business rules and functions are some of the operations applied to the extracted data. Loading process involves moving the desired data as determined by the users to the DW. It's important to note that the discussed flow of data is not as simple and smooth as it sounds and this is as a result of impeding performance issues raised by the following observed bottlenecks.

(El-Wessimy et al, 2013) shows the relevance of DW in decision making in today's environment. "The best decisions are made when all the relevant data is taken into consideration. Today, the biggest challenge in any organization is to achieve better performance with least cost, and to make better decisions than competitors. That is why data warehouses are widely used within the largest and most complex businesses in the world."

In this paper, SQL Server Integration Services tool is used to experimentally show the impact of prioritizing the data from sources as per the confidence levels and the distinctiveness of the data. The units of measure also focus on general performance of the whole ETL process after enhancement of the data staging area.

2. RELATED WORK

2.1 Extraction Stage

This is the initial stage of data migration to a data warehouse. (Kimball et al., 1998) informs that the extraction process consists of two phases, initial extraction, and changed data extraction. In the initial extraction, data from the different operational sources to be loaded into the DW is captured for the first time. This process is done only one time after building the DW to populate it with a huge amount of data from source systems. The next phase involves incremental extraction also referred to as changed data capture (CDC).

(Stephen, 2013) informs that "The staging tables usually get populated by some outside source, by either pulling or

pushing the data from the source systems. This process is usually an insert only process and therefore does not rely on statistics for its successful execution."

2.2 Transformation Stage

Once the data is extracted to the staging area, there are numerous potential transformations, such as cleansing the data (correcting misspellings, resolving domain conflicts, dealing with missing elements, or parsing into standard formats), combining data from multiple sources, reduplicating data, and assigning warehouse keys. These transformations are all precursors to loading the data into the data warehouse presentation area. Unfortunately, there is still considerable industry consternation about whether the data that supports or results from this process should be instantiated in physical normalized structures prior to loading into the presentation area for querying and reporting.

(Erhard and Hong, 2000) elaborate on activities within transformation phase towards clean data. These include data analysis that focus on meta-data and due to fewer integrity rules it cannot guarantee sufficient data quality of a source. Two approaches have been put across to assist in data analysis i.e. data profiling and data mining. Data profiling focuses on the instance analysis of individual attributes. It derives information such as the data type, length, value range, discrete values and their frequency, variance, uniqueness, occurrence of null values, typical string pattern providing an exact view of various quality aspects of the attribute. Data mining helps discover specific data patterns in large data sets, e.g., relationships holding between several attributes.

2.3 Loading Stage

Loading in the data warehouse environment usually takes the form of presenting the quality-assured dimensional tables to the bulk loading facilities of each data mart. The target data mart must then index the newly arrived data for query performance. When each data mart has been freshly loaded, indexed, supplied with appropriate aggregates, and further quality assured, the user community is notified that the new data has been published.

When it comes to moving data to DW (Stephen, 2013) informs "The biggest question for the staging area is – how do we keep the statistics up-to-date such that the statistics for a particular daily load are always available and reasonably accurate. This is actually more difficult than it sounds. If the partitions would only be analyzed in the first quarter of the month each night, going to every other night and eventually each week because of the 10% stale setting. This obviously leaves us with a problem…. In order to have the statistics available for the latest day which is loaded, the statistics would have to be gathered after the staging tables have been loaded but before the ETL process starts.

2.4 Data Staging

Data staging emerges as a new technological development with an attempt to handle the low performance issues noted above. Its location within the ETL process differs as (Kimball and Ross, 2002) state that data staging is available in the extraction and transformation phases of ETL framework. In some current systems a data stage exists as a location that interconnects Online Transaction Processing systems (OLTPs) to the Online Analytical processing systems (OLAPs).

2.5 Reasons for Data Staging Enhancement

Although data staging is not a new technology since it has been researched before, the focus has been shifted to designs and development of data staging frameworks. Little attention has been given to its operability and its significant role in speeding the ETL process.

In a production environment especially a busy organization that deals with large transactions in its daily operations, data flow to DW and storage repositories becomes an issue. Some may not be experiencing the performance problems initially but when their data levels grow they start getting intermittent performance. This should be handled early to have a maintained work path. It still remains a challenge on selection algorithms that can pre-determine the data needed at the target system. The proposed solution in data staging which forms the enhancement is to work on the pre-determining and prioritization mechanisms on the data to load.

(Aksoy et al, 2001) introduces a more workable approach to data staging concerns. They based their work on broadcast scheduling and data staging. According to them the key design considered for development of large scale on-demand broadcast server was the scheduling algorithm selection useful for selecting of items to be broadcast. However, this solution is based on assumptions that data will be available before hand which is not true due to its dynamism.

3. PROPOSED APPROACH IN DATA STAGING

Considering the improvements already observed from great works of other authors it's important to appreciate their efforts in finding gaps in existing systems. Relating to the recurring problems within the data staging, the researcher identifies the following dependent and independent variables that assist in choosing deterministic prioritization approach as a probable solution.

3.1 Dependent and Independent Variables

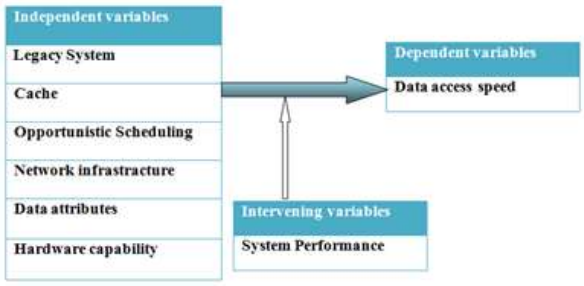

Figure 1. Dependent and Independent variables

Legacy systems use stacking, where a job is pushed into the pool of tasks and popped out of the stack sequentially when its time of execution arrives according to the queue structure. The order of execution is not highly dependent on pre-arranged structure but on the procedural mode which degrades performance.

The cache offer pre-fetching capability where data storage occurs temporarily for the most used data. The scheduled procedure looks firstly in the cache memory before checking on secondary storage locations to minimize the search time. The limitation lies on the cache size and amount of data to be maintained in cache at a time.

To achieve concurrent operation there needs to be selective algorithms to decide the priority of jobs from the source systems to the data staging area. With Opportunistic scheduling there is high probability of improving speed of data retrieval and access.

Remote connection to source systems affects the speed of retrieval and query execution is delayed by the time-lapse for

distributed systems. This impact on the nature of ordering results from query execution and thus optimization should be introduced to work with stored procedures and cache facility.
The data characteristics are defined by type and formats since it comes from disparate systems. Destination requirements must be matched before data is moved to the target systems. Poor data manipulation functions result in longer time processing the data slowing down the systems. The functions for manipulating flat files are different from the ones for relational tables and databases due to underlying data formats.

3.2 Deterministic Prioritization (DP)

After thorough considerations of the above variables and the research gaps identified, Deterministic Prioritization approach is put forward as a solution to the data access problem. The implementation of deterministic prioritization in the data staging area expounds the relationship of the other ETL processes. This approach will tend to manipulate data immediately it is collected and availed to the staging area. Less activity is experienced in the extraction phase but the actual data work area is within the staging area. With this approach appropriate data selection is coordinated to filter out unwanted "dirty" data and assigning priority to the important "clean" data that is moved to data warehouse or data marts. The fact remains that previous activities within staging area are important and hence the approach aims to improve on the order of execution to avoid redundancy and repetition of tasks. This concept is illustrated in the following diagram that shows the relationship among the ETL processes of a data warehouse by applying the DP rules.

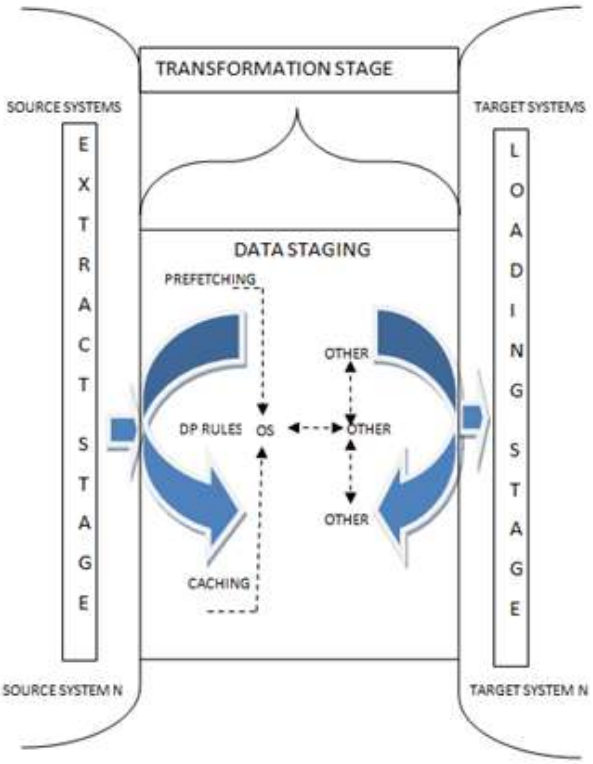

Figure 2. Proposed Conceptual Framework

This implementation is in two sections i.e. deterministic section and prioritization section as discussed below.

3.3 Deterministic Section

According to Macmillan dictionary, the term deterministic means "using or believing in the idea that everything is caused by another event or action and so you are not free to choose what you do". The deterministic approach is relying on data selections based on the confidence levels or behavior of data accesses in relation to previous selections. The key characteristic of determinism is that, the output remains the same for the same inputs. When applied to data staging, the selection of column values from a table is based on the distinctiveness of the data stored and frequency of previous accesses. The most distinct values (fewer NULLS) the higher the confidence levels to be loaded to the next stage in ETL. The data columns with high nullity are negated from the selection hence reducing the amount of load to the next stage.

3.4 Prioritization Section

According to Macmillan dictionary, the term prioritize means "to decide in what order you should do things, based on how important or urgent they are".
The main goal of implementing prioritization is to reduce the amount of active data being manipulated at a time by focusing on the minimum and meaningful details. Once the data to be loaded has been determined, then the order at which processing occurs is vital to reduce delays in data handling. The data selection from the raw tables needs to be prioritized by altering the query execution plan to give more priority on the data columns with high confidence levels. To implement this clustered indexes are introduced based on these distinct columns. These indexes alter the normal execution plan for the queries hence improving performance and efficiency. This execution plan is explained by (Grant, 2009) that "an execution plan is the result of query optimizer's attempt to calculate the most efficient way to implement the request represented by the T-SQL query..."
The query executer has an engine that optimizes query execution on its own and by altering the data selection time is reduced. Prioritization affects the logical aspects enforced by the business rules to maintain the dimensional model and giving way for faster way of retrieving data.

3.5 Included Improvements

Creating a new stable staging framework that is freely available to everyone and run across different hardware platforms (cross-platform) and supporting concurrent processing. This will magnify the core benefits of having intelligent data warehouses that are supportive to the top-level management systems majorly due to development of staging area.
(Stephen, 2013) elaborates their approach in Oracle environment." Most ETL applications use a staging area to stage source system data before loading it into the warehouse or marts. When implemented within an oracle environment a partitioning strategy is usually employed such that data that is not required any longer can be removed from the tables with the minimum amount of effort."
The proposed framework will have forecasting and prioritization mechanisms to decide which data is necessary before transfer begins hence saving on network services and bandwidth.
Current systems such as HANA databases have high processing capability which is meaningless if there is no proper scheduling of resources. This can result to lots of losses of resources not being manned properly. Hardware is static while data is dynamic and at some point the available hardware would not be sufficient to handle the data affecting on performance. Eventually, Scheduling plays a vital role in

the performance implications of any system. It represents the effect being sought is measurable to make comparison.

3.6 Prioritization by indexing specific columns

The newly distinct derived columns that were added to each staging table (external columns; hence do not affect the data from the source in any way to ensure consistency and integrity), are used to create indexes as shown below.

```
CREATE CLUSTERED INDEX IX_newStaging_Customer
    ON dbo.newStaging_Customer (STCustomerID)
GO
```

Derived column

Figure 3. Index Creation Sample

The created index named "IX_newStaging_Customer" is prioritizing the derived unique column named "STCustomerID" on the staging table named "newStaging_Customer". The created user defined index hints the order of execution of the Data Definition Language (DDL) queries submitted to the server hence overriding the server's query execution plan. The created index is also not affected in future in case of recreations of the source data from extraction stage since its an external derived column. Prioritization by distinct columns enhances the efficiency and performance of the query execution plan by the server during query search. This results in optimized selection costs while maintaining the quality of data to the data warehouse. The measure of improved efficiency is shown below for a selection query.

Figure 4. New Execution plan for Customer Table

3.7 Test scenario preparation in SSIS tool

The ETL processes of a data warehouse are demonstrated using the SSIS tool. The first scenario setup is for the current situation before enhancement and the second scenario setup is for the new situation after enhancement. The following is a demonstration of the scenario used in the experiments in run mode.

Figure 5. Test Scenario Setup.

4. RESULTS

The experiments are performed and collecting results of time variables in a Ms Excel file. This file is generated automatically from the scripts written in Visual Studio 2008 and C# programming language when the scenario is run. The experiment is run for fifteen times and for each cycle it records the time change for the different variables to the file. Finally, the comparison is made based on these results from both scenarios to note the impact of the change introduced as shown below.

Table 1. Before Enhancement ETL Processes results

TestRunNumber	prevExtractionTime	prevStagingTime	prevLoadingTime
1	63368.28	68717.21	92851.52
2	97940.12	79815.68	132989.3
3	52449.29	45874.92	62384.48
4	54358.59	87338.59	49553.42
5	49600.26	50350.14	45086.72
6	48285.48	54800.62	45312.21
7	52569.22	48670	48858.79
8	53581.86	51044.78	66801.6
9	88992.69	96697.22	60973.71
10	72884.88	56306.1	52252.09
11	58839.45	54556.57	59913
12	69740.99	58111.31	53562.38
13	59401.77	50840.68	51099.61
14	65250.34	55357.5	52560.51
15	56115.72	71712.82	58035.14

Table2. After Enhancement ETL Processes results

TestRunNumber	newExtractionTime	newStagingTime	newLoadingTime	newTotalTime
1	62261.37	50691.27	4824.541	117777.2
2	59639.95	53591.02	4299.265	117530.2
3	55766.46	54468.57	3652.78	113887.8
4	54012.87	53544.93	3552.135	111109.9
5	61811.67	54836.81	3340.999	119989.5
6	59963.43	57082.69	3374.887	120421
7	63116.5	69659.48	4147.719	136923.7
8	61813.26	61603.78	3455.976	126873
9	69365.94	54521.06	3623.27	127510.3
10	56561.51	52100.49	5690.811	114352.8
11	54106.91	59241.29	3701.506	117049.7
12	59291.2	86575.51	3493.712	213390.8
13	66104.49	89692.62	4130.86	159928
14	72926.03	56987.29	5984.634	135898
15	57962.95	50812.86	3926.081	112701.9

5. DISCUSSION

The following discussion is based on the comparison of the results obtained from the tests. Each of the ETL stage is compared separately for the two situations and graphically shown in the following figures to have a clear distinction of the two situations.

5.1 Comparison of Extraction stage

The following is an illustration of the individual comparison per stage of the ETL processes to bring out a clear view of the improvement made. Explanation for each comparison follows for every illustration. The negative sign indicates that it is in the reducing direction thus showing the enhancement has taken place by reducing the particular running time per stage.

Table 3. Extraction stage results analysis

Test	prevExtractionTime	newExtractionTime
Average	62891.92931	60980.30203
Change		-1911.627273
Change %		-3.039543061

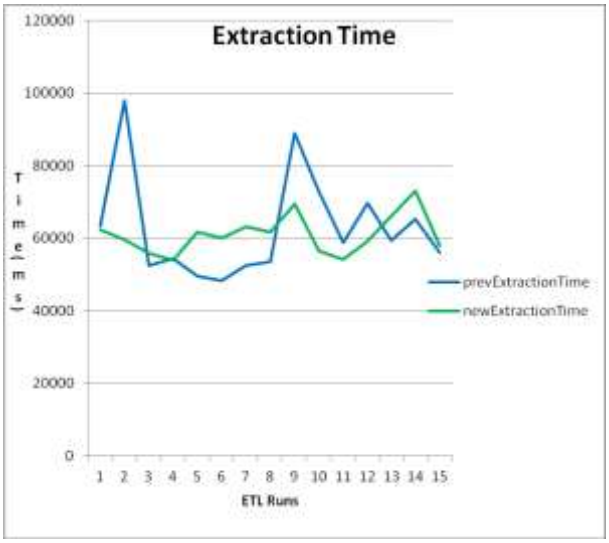

Figure 6. Extraction stage comparison

The above illustrations show the time taken for extraction using Deterministic Prioritization approach has reduced by 3.04% compared to the previous extraction time.

5.2 Comparison of Staging stage

Table 4. Staging stage results analysis

Test	prevStagingTime	newStagingTime
Average	62012.94289	60360.64505
Change		-1652.29784
Change %		-2.664440297

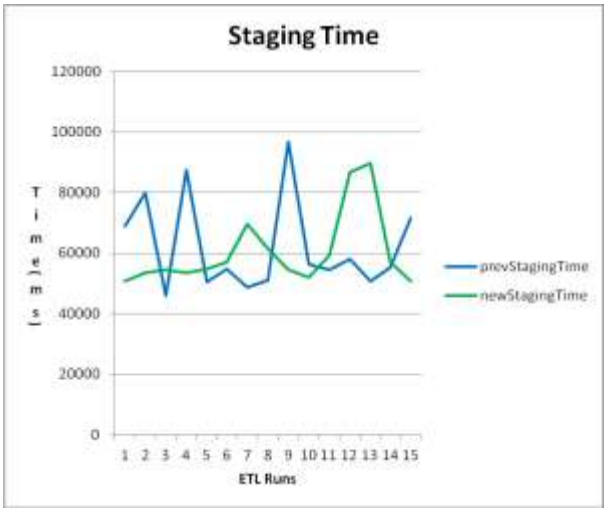

Figure 7. Staging Stage Comparison

The above illustrations show the time taken in staging area using Deterministic Prioritization approach has reduced by 2.66% compared to the previous extraction time.

5.3 Comparison of Loading stage

Table 5. Loading stage results analysis

Test	prevLoadingTime	newLoadingTime
Average	62148.96424	4079.94504
Change		-58069.0192
Change %		-93.43521635

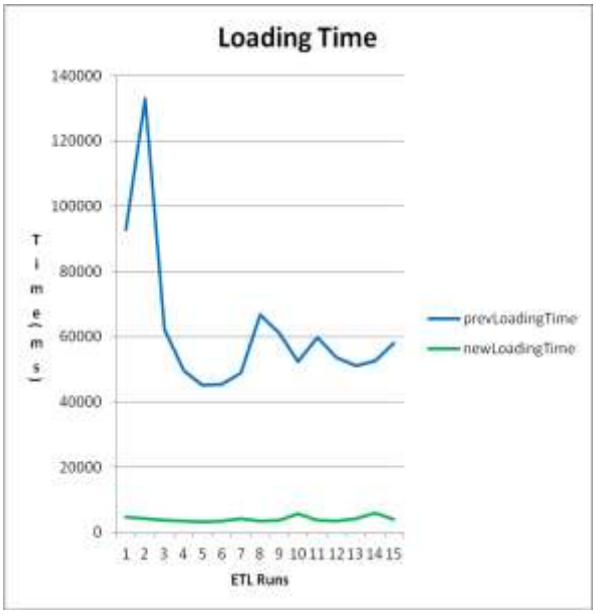

Figure 8. Loading Stage Comparison

The above illustrations show the time taken in loading stage using Deterministic Prioritization approach has immensely reduced by 93.44% compared to the previous loading time. This is a great change and is attributed to the fact that the number of operations talking place at this stage being minimal and less traffic to the destination targets of the data i.e. the data warehouses. However with increased traffic of data especially over the network, the loading time percentage change may further reduce. The loading process has benefited a lot from changes in previous stages.

6. SUMMARY OF FINDINGS

There has been a significant change for every stage of the ETL processes after implementation of deterministic prioritization approach. Moreover, the subsequent impact of this improvement has been depicted at the loading stage where the rate of transfer improved by a large margin of 93.44%. This shows that any change occurring in the staging area will definitely affect the entire process flow. However, if more time is taken in proper design of the data staging area, then there are high expectations for further improvements on performance beyond the ones achieved by the researcher.

Based on the limitation of resources and the specifications of the workstation used to run the tests, the performance levels are considered acceptable for an organization with fast growing data sets. However, the setup is limited by the resources at run time but with increased disk space as provided by the owners as data size grows, then the total data access time in the data warehouse will be highly improved. The choice of data prioritization mechanism using indexes for columns is supported by (Cecilia and Mihai, 2011), where they state the use of Indexes on database queries improves the performance of the whole system. Clustered indexes perform better than nonclustered indexes when the expected returned records are many and should be set for the most unique column of a table. This proposition goes in line with the research with the use of clustered indexes in the staging area mainly due to the need to fetch and process large data.

(Costel G et al, 2014) did a research on query execution and optimization in the MSSQL Server and put across the missing of indexes as a contributor to low performance of query execution. They inform that when a table misses indexes, the search engine has to parse through the entire table step by step to find the searched value. The resources spent on this process are enormous and considerably increases time to execute the queries.

(Grant, F. 2012) explains about execution plan management done by query optimizer. The database relational engine performs logical reads within the cache memory while the storage engine performs physical reads directly from disk. Improvements are highly realized mostly for data manipulation language statements since the engine needs to parse the query for correctness. The SQL server generates statistics against the indexes and sends them to the optimizer to determine the execution plan.

(El-Wessimy et al, 2013). Similarly did an enhancement in the data warehouse staging area by using different techniques (FIFO, MC,RR time and record rotation) targeting the loading phase. The tests ran captured the time taken to transfer data in each stage of the ETL process and suggest the most suitable technique. They did a comparison amongst all techniques and noted that FIFO performed better for less data set while Record Limit Based Round Robin was best for large data sets. Their research supported further reduction of overall time taken to deliver data from source to destination. The uniqueness of this study is the ability to handle large data sets from the beginning as well as newer inclusions of data from

the sources without any readjustments of the system structure setup.

7. CONCLUSION

The adoption of Deterministic Prioritization approach in the staging area has shown promising results and the users at the presentation level of the data warehouse are rest assured of fast data access and retrievals. They can timely make decisive conclusions and reports based on current data that is made available in a timely fashion similar to real time systems. They will also benefit to wide range of data availed since the ETL processes aim to denormalize the data comprehensively before its delivered to users. This forms an association that is deterministic in nature for future priority loads. The researcher was keen to avoid compromising data quality for high performance gains and this resulted in a more balanced system setup where the data still meet the qualitative assurance defined by the users' requirements.

8. RECOMENDATIONS AND FUTURE WORK

In view of the results and findings of the experiments undertaken in this research, the researcher recommends the incorporation of the deterministic prioritization approach in the design and development phase of the data staging frameworks. This is so because it is cross-platform to all database management systems that support the SQL language in the market today. The change gives room for further customization since it only happens at the design and before query execution. Notably for well formed queries the performance will be even better.

The data staging area is an area which has not been researched exhaustively and the impact of high resources should be considered as a next check on performance gain over cost. The setup experiments are carried based on same format of data sources and further studies should be carried out on variant data sources and using different staging framework other than the SSIS tool used here. The scenario in this test is demonstrated in a local workstation and it would be essential to note the performance levels in a distributed system with both sources and targets widely separated by networks.

9. REFERENCES

[1] Abbasi, H., Wolf, M., Eisenhauer, G., Klasky, S., Schwan, K., & Zheng, F. (2010). Datastager: scalable data staging services for petascale applications. *Cluster Computing*, *13*(3), 277-290.

[2] Akkaoui, Z. E., Munoz, E. Z. J.-N., and Trujillo .J. A. (2011). *Model-Driven Framework for ETL Process Development*. In Proceedings of the international workshop on Data Warehousing and OLAP. pp. 45–52 Glasgow, Scotland, UK.

[3] Aksoy, D., Franklin, M. J., & Zdonik, S. (2001). Data staging for on-demand broadcast. In *VLDB* (Vol. 1, pp. 571-580).

[4] B´ezivin, J. (2005). On the unification power of models.Software and System Modeling, 4(2):171–188

[5] Cecilia, C., Mihai, G. (2011). Increasing Database Performance using Indexes. *Database Systems Journal vol. II, no. 2/2011*. Economic Informatics Department, Academy of Economic Studies Bucharest, Romania.

[6] Costel, G.C., Marius, M. L., Valentina, L., Octavian, T. P.(2014). Query Optimization Techniques in Microsoft SQL Server. *Database Systems Journal vol. V, no. 2/2014*. University of Economic Studies, Bucharest, Romania.

[7] Da Silva, M.S.,Times, V.C., Kwakye, M.M. (2012). Journal of Information and Data Management.3 (3).

[8] Deterministic. (2009-2015). In Macmillan Dictionary. Macmillan Publishers Limited: Accessed from: www.macmillandictionary.com on 20th July 2015.

[9] Eckerson, W., & White, C. (2003). Evaluating ETL and data integration platforms. *Seattle: The DW Institute*.

[10] El-Wessimy, M., Mokhtar, H. M., & Hegazy, O. (2013). ENHANCEMENT TECHNIQUES FOR DATA WAREHOUSE STAGING AREA. *International Journal of Data Mining & Knowledge Management Process*, *3*(6).

[11] Erhard, R., and Hong, H.D. (2000). *Data Cleaning*: Problems and Current Approaches. Journal IEEE Data Eng. Bull.23 (4), 3-13.

[12] Grant, F.(2009).The Art of High Performance SQL Code: SQL Server Execution Plans. Simple-Talk Publishing. ISBN 978-1-906434-02-1

[13] Grant, F. (2012). SQL Server Execution Plans. Second Ed. ISBN: 978-1-906434-92-2. Simple Talk Publishing.

[14] El-Wessimy, M., Mokhtar, H. M., & Hegazy, O. (2013). ENHANCEMENT TECHNIQUES FOR DATA WAREHOUSE STAGING AREA. *International Journal of Data Mining & Knowledge Management Process*, *3*(6).

[15] Firestone, J. M. (1998). Dimensional modeling and ER modeling in the data warehouse. *White Paper No, Eight June*, *22*.

[16] Flinn, J., Sinnamohideen, S., Tolia, N., & Satyanarayanan, M. (2003). Data Staging on Untrusted Surrogates. In *FAST* (Vol. 3, pp. 15-28).

[17] Inmon, W. H. (2002). *Building the Data Warehouse*. Wiley.

[18] Inmon, W. H. (2005). *Building the data warehouse*. John wiley & sons.

[19] Kimball, R., & Caserta, J. (2004). *The data warehouse ETL toolkit*. John Wiley & Sons.

[20] Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (2008). The Data Warehouse Lifecycle Toolkit, 2nd ed. Practical Techniques for Building Data Warehouse and Business Intelligence Systems.

[21] Muller, P. A., Studer, P., Fondement, F., & Bézivin, J. (2005). Platform independent Web application modeling and development with Netsilon. *Software & Systems Modeling*, *4*(4), 424-442.

[22] Per-Åke, L., Cipri, C., Campbell, F., Eric, N. H., Mostafa, M., Michal, N., Vassilis, P., Susan, L. P., Srikumar, R., Remus, R., Mayukh, S.(2013).Enhancements to SQL Server Column Stores. ACM 978-1 -4503-2037-5/13/06. New York, USA.

[23] Prioritize. (2009-2015). In Macmillan Dictionary. Macmillan Publishers Limited: Accessed from: www.macmillandictionary.com on 20th July 2015.

[24] Ralph, K. and Margy, R. (2002). *The Data WarehouseToolkit*: The Complete Guide to Dimensional Modeling. Second Edition. Published by John Wiley and Sons, Inc. Canada.

[25] Russom, P. (2012). BI Experts: Big Data and Your Data Warehouse's Data Staging Area. *TDWI Best Practices Report, Fourth Quarter*. Retrieved from http://tdwi.org/articles/2012/07/10/big-data-staging-area.aspx

[26] SAP AG. (2002). Business Information Warehouse – Data Staging Retrieved from http://scn.sap.com/docs/DOC-8100.

[27] Stephen, B. (2013). Staging, Statistics & Common Sense: Oracle Statistics Maintenance

[28] Strategy in an ETL environment Retrieved from http://www.seethehippo.com/

[29] Vassiliadis, P. (2009). A survey of Extract–transform–Load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, *5*(3), 1-27.

[30] Zineb, A., Esteban, Z., Jose-Norberto.M., Juan, T. (2011). *A Model-Driven Framework for ETL Process Development*. In DOLAP 11 Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP Pages 45-52. ACM New York, NY, USA. ISBN: 978-1-4503-0963-9