WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Study on the Performance Criteria of a Datawarehouse in A Company: Analysis of ETL Models and Perspectives

Kasonga Badibanga Maxime.H

University of Kananga

SUMMARY

This article discusses the performance of a data warehouse as well as the analysis of the generation of ETL (Extract-Transform-Load) operators to feed a data store from a relational data source. First, we add new rules to those proposed by the authors, these rules deal with the combination of ETL operators. Secondly, we propose a hybrid partitioner of the Vassiliadis approach based on model transformation for the generation of ETL operations necessary for loading a data warehouse.

This approach offers the possibility to the designer to define certain conditions necessary for loading.

Keywords: ETL, Extraction, Transformation, Loading, data sources, data warehouse, data store, RR, operator.

GENERAL INTRODUCTION

The evolution of new information and communication technologies has caused various adjustments in leadership agendas. The more the latter intersects with different scientific disciplines which change their nature, and generate new fields of investigation, it is necessary to move towards the decision-making system.

Decision-making systems for processing and valorizing data are now well established in companies. They are particularly enriched by presenting powerful analytical, prospecting and data optimization methods.

In this sense, decision-making is a real and essential problem that concerns business managers and all levels of society. It always results from a more complex process: the data to be taken into account are ever more voluminous and the issues are so important (human, financial) that the IT tool has become strategic.

The performance of a data warehouse is essential for any company that relies on data to make strategic decisions. A high-performance data warehouse must be able to store, manage and analyze large amounts of data quickly and efficiently. It must also be capable of providing real-time reporting and analysis, in order to offer the company valuable information for decision-making.

Optimal data warehouse performance allows the company to maximize its resources, reduce data management costs, and remain competitive in the market. The performance of a data warehouse is therefore a key factor in the success of a data-driven business.

It is with knowledge of these issues that we come to make our contribution by devoting ourselves to this study.

0.1. Context

The “context” of a study involving the design of a computer application refers to the environment in which the application will be used, as well as the needs and expectations of end users.

This may include information about the target market, existing technologies, constraints and opportunities, and any other external factors relevant to the design of the application. Context provides a frame of reference necessary for understanding project requirements and objectives, and helps guide and justify design decisions.

The context of our study is in the field of data management and business intelligence. It focuses on evaluating the performance of data warehouses using ETL methods, a key element for organizations looking to analyze large amounts of data to make strategic decisions.

More specifically, this dissertation aims to analyze the ETL model to ensure the performance of different data warehouse solutions, taking into account criteria such as the speed of processing masses of data, reliability and ease of use.

The context would also include information on the growing importance of data warehouses in businesses, as well as the challenges of managing and analyzing data at scale. He will also discuss the evolution of data warehouse technologies and practices, as well as the importance of choosing the right solution based on the specific needs of the organization.

Finally, context could also include information on current market trends, regulatory requirements, and end-user expectations regarding data warehouse performance and functionality.

The choice and interest of the topic of a scientific study refers to the decision to select a specific area to study, as well as the importance and relevance of the topic to the research. This involves justifying why this topic was chosen, highlighting its originality, its potential to contribute to existing knowledge in the field and its relevance to the scientific community. In other words, it is a question of demonstrating why the chosen subject is worth studying and how it can bring significant added value to the discipline concerned.

The choice and interest of this subject lies in the growing need for companies to make decisions based on reliable and relevant data.

A data warehouse is an essential element of IT infrastructure for storing and analyzing data. Benchmarking the performance of existing data warehouses can help identify best practices and the most efficient solutions for managing and analyzing large amounts of data. This study is therefore of major interest to businesses and IT professionals seeking to improve the efficiency and reliability of their data management systems.

Furthermore, with the continued emergence of new technologies and new approaches to data warehouses, such a study could provide important insights for decision-makers seeking to invest in these infrastructures.

0.2. STATE OF THE ART

The state of the question refers to the synthesis of already existing knowledge on the research subject. This synthesis consists of taking stock of previous research, existing theories, work already carried out and the results obtained.

It allows us to situate the scientific context in which the study takes place, to justify its relevance and to determine the problem to be resolved.

The literature review on previous research and existing theories on the performance criteria of a data warehouse as well as ETL model operators, several approaches have been proposed. Among which we can cite:

- Ghozzi.F, in his doctoral thesis entitled “design and manipulation of constrained dimensional databases” at Paul Sabatier University Toulouse 3, November 2004, the author presents an approach based on web technology semantics to facilitate the process of selecting relevant information from data sources to transform this data and load it into a data warehouse.

M. Ben Taher and H.Ben Abdallah, in their article on “the semi-automatic approach for ETL procedure generation”, November 2010, the authors propose a second approach for constructing the ontology, this approach is broken down in three steps and ensures the manual extraction of the semantics of the elements of the data source as well as the data warehouse.

- Trésor ANONGA in his DEA thesis entitled “Comparative study on the performance of a Data Mart in a banking company” at the University of Liège in 2010, the author underlines the ability to offer high performance, effective management of queries and integration of data from various sources;

- Jérôme Dormant in his thesis entitled “optimization and performance evaluation to help with the design and administration of complex data warehouses”, at the University of MONTPELLIER in 2015, the author proposes a solution to level of optimization and evolution of the performance of databases in general and data warehouses in particular, by applying the principles proposed initially to the context of complex data structured in the form of XML documents.

- Antoine ROBERT in his thesis entitled “Comparative study of the design approach of multidimensional models used for data warehouses”, at the MOULOUD MAMMERI DE TIZI University – OUZOU/Algeria in 2016, in this research, the author pursues the objective to study different approaches to designing multidimensional models used for data warehouses and to compare the models constructed from the different approaches.

On our part, the objective is to design an algorithm which processes a mass of data passing through an ETL process following the ETL process modeling approach processing data of big data dimension according to the MapReduce paradigm which is based on the approach by Vassiladis by combining two partitioners to have a partitioner called “Hybrid Partitioner” which will ensure the performance of Data warehouse.

The problem of a study involving the design of a computer application refers to the central question or challenge that the study seeks to resolve.

As in most African countries, businesses operate in a disorganized and fragmented manner due to the multiple pressures in processing mass data, which causes loss and latency in the execution of certain queries.

The Democratic Republic of Congo is a vast, very landlocked country. Despite its progress, the problem of data warehouse performance criteria in different companies remains a challenge for the decision-making system at different levels of the managerial pyramid.

However, the resolution of the problem linked to the effective exploitation of mass analytical data which can thus allow decision-makers to make a good decision, leads us to a study on the study of data werahouse performance criteria: ETL model analysis and perspective which proves essential.

Indeed, to be concrete in carrying out this study, the main question of said study is to know which ETL model to implement that can ensure the performance of a data warehouse?

From this main question, other sub-questions arise, namely:

- How to organize the data so that it makes sense in its processing?

A hypothesis aims to propose a plausible explanation or a possible answer to the problem of the study.

In the first hypothesis, the use of a Vassiliadis modeling approach following the ETL process in a MapReduce model, can improve the performance of a data warehouse by reducing data processing time would be an adequate solution to this anomaly.

While at the second hypothesis, efficient memory management, using techniques such as implementing a hybrid partitioner algorithm can significantly influence performance, by speeding up the fetching, transferring and loading of data while reducing the load on the system.

The study on the performance criterion of a data warehouse adopts a qualitative epistemological stance in order to understand in depth the issues linked to the performance of data warehouses. This approach is based on in-depth questioning around current practices, the challenges encountered and the needs of organizations in terms of managing their data.

The preferred type of reasoning is that of induction, making it possible to identify trends, correlations and explanatory factors from a typical example.

The objective of a study is what researchers aim to accomplish or discover through their research, including describing a phenomenon, understanding its causes, evaluating the effectiveness of an intervention or comparing different populations, among other things. The objectives of a study must be clearly defined to guide the research and establish criteria for evaluating the results.

The aim of our study is to evaluate the performance criterion of a data warehouse by comparing different available options.

The general objective of a study is the main goal that the researcher or research team seeks to achieve by carrying out this study. This is a general, concise statement that describes the overall result the study aims to produce. The overall objective helps provide clear direction for the research and determines the parameters and scope of the study. It can be formulated based on the research questions, hypotheses, specific objectives, organizational needs or any other relevant aspect of the study.

The general objective of this study is to design an algorithm which processes a mass of data passing through an ETL process following the ETL process modeling approach processing data of big data dimension according to the MapReduce paradigm which is based on the approach by Vassiladis by combining two partitioners to have a partitioner called “Hybrid Partitioner” which will ensure the performance of Data warehouse.

To properly conduct our research, we have subdivided this work into four main chapters apart from the general introduction and the general Conclusion detailed as follows:

The first part of this article will be based on the fundamental notions of the data warehouse, which will allow us to define the fundamental concepts of the data warehouse, as well as specific notions such as performance analysis, quality of service in the data warehouse;

The second will be the subject of a proposal for a new approach, which aims to propose a new algorithm following the approach studied which will be considered as a palliative solution in our case study.

I. FUNDAMENTAL NOTIONS ABOUT THE DATAWERHOUSE

A crucial stage of our study because, it aims to deepen our knowledge and understanding of the different key aspects related to this area, we will explore several fundamental concepts of the data warehouse, as well as specific notions such as performance analysis[1].

First, we will define the data warehouse and examine its main concepts. The data warehouse represents an essential IT solution for managing and analyzing large volumes of data from different sources[2].

We will study the different characteristics and functions of a data warehouse, the key elements that make it up and the advantages it offers in terms of decision-making.

Next, we will discuss the analysis of performance criteria in the data warehouse. This analysis focuses on evaluating and improving the performance of a data warehouse, particularly in terms of response time, data throughput and resource utilization. We will explore different techniques and strategies that optimize data warehouse performance, taking into account data modeling and system architecture.

[1] R.K. &. M. Ross, *Data Warehouse: A Practical Guide to Dimensional Modeling*, 2nd ed., 2002, p. 25

[2] B. FYAMA, *Treatise on methods and techniques in Computer Sciences*, Paris. Ed.Eryol, 2018, p. 1

I.1. Data warehouse definition and concepts

The Data Warehouse is a central concept in the field of information management. Among the many terms and notions linked to this discipline, certain key concepts are worth mentioning to fully understand the functioning and importance of a data warehouse.

First of all, it is essential to understand what a data warehouse is. It is a database specifically designed to store and analyze large amounts of data from multiple sources.

The main objective is to centralize this data in the same place, in order to facilitate their subsequent use for analyzes and decision-making.

Centralizing data in a data warehouse makes it possible to create a unified and coherent view of all of the company's information. This means that the different data sources, whether operational databases, flat files or Excel files, are integrated and harmonized within the data warehouse. Thus, users can access reliable and quality information for their analyzes and reports.

I.2. Data warehouse concepts

The key concept of the data warehouse is the notion of star or snowflake schema. It is a specific database structure used to organize data dimensions and facts.

Dimensions represent different characteristics of the data, while facts correspond to the quantitative measures that we want to analyze[3]. This schema format makes it easier to navigate and analyze data in the data warehouse.

One of the key concepts of the data warehouse is data extraction, transformation and loading (ETL). This process brings together all the operations necessary to extract data from sources, transform it into a format suitable for the data warehouse, then load it into the latter[4]. ETL is an essential step to guarantee the quality and integrity of data within the data warehouse5.

Dimensional modeling helps organize data using dimensions, which represent meaningful aspects of the data (e.g., time, location, products, etc.), and facts, which correspond to the quantitative measures that the we want to analyze (sales, revenue, etc.).

Dimensional modeling facilitates data navigation and analysis, while ensuring optimal system performance, data transformation and loading (ETL), which is a necessary process to feed the Data Warehouse with data from different sources[6].

ETL involves extracting raw data, transforming it into an appropriate format, and then loading it into the Data Warehouse. This process is essential to ensure data quality and integrity, as well as to maintain system efficiency.

Finally, it is worth emphasizing the importance of security and user management in the context of a Data Warehouse. Data security is paramount, both from a confidentiality and integrity perspective, and it is essential to have appropriate access control mechanisms in place to ensure secure access to Data Warehouse data[7].

I.3. Performance analysis in data warehouses

Performance analysis in IT data management involves evaluating and monitoring the performance of a computer system, such as a server, network or application. This assessment involves collecting data on response times, capacity, resource utilization, errors, bottlenecks, and other key metrics[8].

Performance analysis plays a crucial role in maintaining the availability and reliability of IT systems, identifying potential problems before they become critical and helping to make informed decisions about system improvements. . It can also help with capacity planning and cost optimization by avoiding unnecessary oversizing of resources.

II. MODELING AN APPROACH FOLLOWING THE ETL PROCESS

This chapter aims to propose a new approach within the framework based on our research.

With this in mind, we will start with a description of the typical existing algorithm, highlighting its characteristics and its mode of operation.

[3] POLLETO.M., *business intelligence*, Ed. Organization, Paris, 2012.

[4] Inmon. B., *Building the data warehouse, Wiley publishing*, 3edition, 2002, p23.

[5] Inmon.B., *building the data warehouse*, Dunod, Paris, 1991

[6] Kimball R., Ross M., *The Data warehouse Toolkit*, 2ème edition John Wiley & Sons Inc., New York, 2002.

[7] HAMER, *Susan., IT security: principles and methods for use by CIOs, CISOs and administrators*, Ed. Dunod, 2018.

[8] AUDIBERT .L, *Databases: From modeling to SQL*, Paris, 2009

Finally, we will present our proposal for a new approach to improve the processing capacity of big data, as well as the speed and efficiency of data performed in ETL.

II.1. Description of the typical existing algorithm

The description of the existing typical algorithm usually refers to the detailed documentation of an algorithm already implemented or used by other developers or systems.

This may include explanations of how the algorithm works, expected inputs and outputs, specific steps or operations performed, data structures used, and the advantages and limitations of the algorithm.

This description allows other developers to understand and use the existing algorithm for their own purposes.

"The description of a typical algorithm consists of an ordered sequence of instructions, each of which represents a logical step for solving a specific problem.[9]"

II.2. Proposal for a new approach

A proposal for a new approach is a suggestion or recommendation aimed at adopting a method or strategy different from that already considered to solve a problem or meet a specific need.

This proposal suggests the use of new technologies, the adoption of alternative practices or methodologies, or a substantial modification of the way in which the solution is designed or implemented.

The objective of this proposal is to improve the efficiency, performance, or quality of the solution, taking into account the constraints and specific requirements of the project.

This approach takes into account factors that influence data storage and processing capacity, as well as the speed and efficiency of data queries.

Starting from the aforementioned example in the second chapter including that of the sale of goods and taking into account certain anomalies noted in the partitioner Hashing by field and Roud Robin of Vassiliadis' approach, we propose another partitioner called "hybrid partitioner" assumed as a stopgap solution in mass data processing in ETL.

This partitioner is unique to our research in that it combines two partitioners including Hashing by field and Roud Robin. The first phase consists of forming the provisional partitions by Hashing and the second phase consists of assigning the final partition for each of these using Roud Robin (RR) by group.

Its diagram following the aforementioned example is presented as follows:

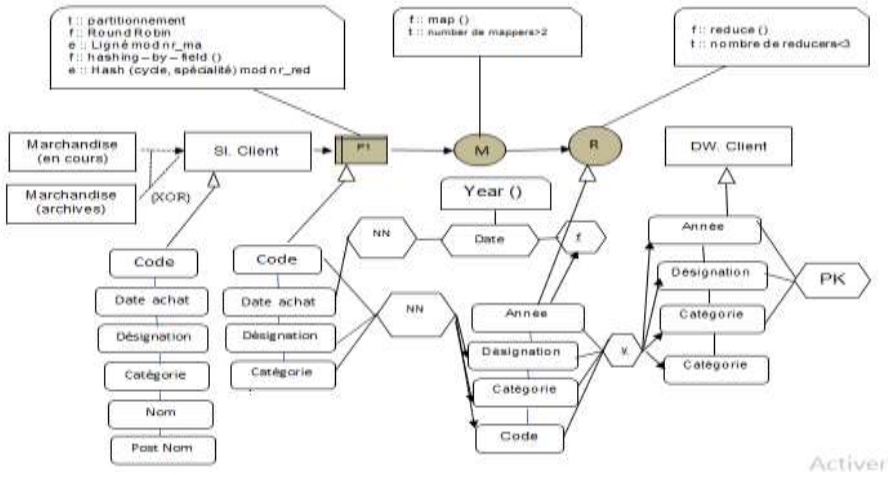

Figure: example of the merchandise sales establishment in the hybrid partitioner.

Hybrid Partitioner Algorithm

Beginning

i=0;

[9] Brassard, G., & Bratley, P., *Algorithmics: Theory and Practice*, 1996. P. 23

```
T: record;
As long as not end (source file);
Do
    Read (source file,T);
    i= Hash(T.attr1, T.attr2,…)modnb_part;
    i= i mod nb_part;
    Write (Pi,T);
    I=i+1;
Do ;
```

END.

Our proposal for a new approach offers new perspectives for improving the storage and processing capacity of data in data warehouses. This approach aims to optimize the speed and efficiency of data queries, while taking into account the particularities of the ETL model.

However, this approach remains to be tested in a real environment in order to validate its impact on the overall performance of the system.

GENERAL CONCLUSION AND OUTLOOK

Modeling is a means that facilitates understanding of the operating details of ETL and then makes it possible to control its complexity and anticipate possible problems and risks before implementing the ETL tool.

Vassilliadis' approach being one of the most interesting in this field, faced with the appearance of data warehouses, decision-making systems known as large consumers of data require a re-study to evaluate the performance of the data warehouse on them. In this context, we proposed an ETL modeling approach based on the MapReduce model.

This approach reveals three new concepts in the ETL model, namely the partitioning of source data, transformations in the Map phase and finally the fusion and aggregation of data in the Reduce phase.

We addressed the issue of data warehouse in ETL modeling which deserves to be experimented on massive data to evaluate its impact as well as the effectiveness of the MapReduce model in improving performance.

To improve the performance of ETL in the data warehouse, it is also interesting to evaluate what NoSQL structures can bring to ETL.

BIBLIOGRAPHICAL REFERENCES

1. AUDIBERT .L, Databases: From modeling to SQL, Paris, 2009.

2. B. FYAMA, Treatise on methods and techniques in Computer Sciences, Paris. Ed.Eryol, 2018.

3. Brassard, G., & Bratley, P., Algorithmics: Theory and Practice, 1996.

4. HAMER, Susan., IT security: principles and methods for use by CIOs, CISOs and administrators, Ed. Dunod, 2018

5. Inmon. B., Building the data warehouse, Wiley publishing, 3edition, 2002.

6. Inmon.B., building the data warehouse, Dunod, Paris, 1991.

7. Kimball R., Ross M., The Data warehouse Toolkit, 2nd edition John Wiley & Sons Inc., New York, 2002.

8. POLLETO.M., business intelligence, Ed. Organization, Paris, 2012.

9. R.K. &. M. Ross, Data Warehouse: A Practical Guide to Dimensional Modeling, 2nd ed., 2002.