DESIGN, IMPLEMENTATION, AND ASSESSMENT OF INNOVATIVE DATA WAREHOUSING; EXTRACT, TRANSFORMATION, AND LOAD(ETL); AND ONLINE ANALYTICAL PROCESSING(OLAP) IN BI

Ramesh Venkatakrishnan

Final Year Doctoral Student, Colorado Technical University, Colorado, USA

ABSTRACT

The effectiveness of a Business Intelligence System is hugely dependent on these three fundamental components, 1) Data Acquisition (ETL), 2) Data Storage (Data Warehouse), and 3) Data Analytics (OLAP). The predominant challenges with these fundamental components are Data Volume, Data Variety, Data Integration, Complex Analytics, Constant Business changes, Lack of skill sets, Compliance, Security, Data Quality, and Computing requirements. There is no comprehensive documentation that talks about guidelines for ETL, Data Warehouse and OLAP to include the recent trends such as Data Latency (to provide real-time data), BI flexibility (to accommodate changes with the explosion of data) and Self-Service BI. This research paper attempts to fill this gap by analyzing existing scholarly articles in the last three to five years to compile guidelines for effective design, implementation, and assessment of DW, ETL, and OLAP in BI.

KEYWORDS

Business Intelligence, ETL, DW, OLAP, design implementation and assessment

1. INTRODUCTION

"Business intelligence (BI) is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance" [1]. Data Acquisition, Data Storage, and Data Analytics are the primary components of a BI system, and for it to be successful, practical design and implementation of these three components are critical. Source systems such as OLTP records business events. ETL stands for Extract, Transact, Load. It is a set of tools and processes to extract the data from source systems to load into a data warehouse after transformation. Data warehouse (DW) is the core component of a BI system where data from various sources are centrally stored. OLAP stands for Online Analytical Processing and is responsible for analyzing the data warehouse data to aggregate and present it in a multi-dimensional format to answer "forecasts." The effectiveness of a BI System is hugely dependent on these three fundamental components Data Acquisition (ETL), Data Storage (Data Warehouse), and Data Analytics (OLAP). There is no comprehensive documentation that talks about guidelines for ETL, Data Warehouse and OLAP to include the recent trends such as Data Latency (to provide real-time data), BI flexibility (to accommodate changes with the explosion of data) and Self-Service BI.

The purpose of the article is to analyze existing scholarly articles in the last three to five years to compile guidelines for effective design, implementation, and assessment of DW, ETL, and

OLAP in BI. The drivers for the need for this "updated" guidelines are agility, the next generation of data, cloud computing, real-time data, situation awareness, and self-service. The new design, implementation, and assessment guidelines of DW, ETL, and OLAP would help decision-makers and BI IT practitioners in proactively avoiding the "known" pitfalls. This approach also helps BI practitioners prepare their systems to be flexible to accommodate the data explosion, and to move from an IT-Lead BI to IT-enabled BI. This paper is organized into three main sections. The first section of this paper describes the challenges associated with all three fundamental components (Data Acquisition, Data Storage, and Data Analytics) of BI. The second section provides guidelines for design and implementation. The third section provides an assessment methodology for ongoing effectiveness.

2. RELATED WORK

The limitation of traditional computing techniques is a significant limitation in dealing with large volumes of data [2]. Santos, Silva, and Belo (2014) highlighted the dealing of vast data sets in a short time slot for data warehouse systems as a challenge due to the need for massive computational resources and storage requirements [3]. Bousty, Krit, Elasikiri, Dani, Karimi, Bendaoud, and Kabrane (2018) have recommended continuous updates of server and storage processing capacities to tackle the new data requirement [4]. Their research also recommends utilizing cloud-based solutions to keep the processing capabilities elastic to reduce additional investments. Vo, Thomas, Cho, De, and Choi (2018) recommends three new features for the next generation BI, Operational BI (Near Real-time), Situational BI (Real-time), and Self Service BI (reduced dependency of IT Staff) [5]. Santos et al. (2014), in their research, recommended a grid environment-based scheduling solution for small to medium ETL processing needs [3]. Extensive ETL processing continues to remain a challenge. Separation of processing for OLTP and OLAP, the use of specialized databases such as NoSQL, Columnar, In-Memory-DB, and Hybrid-DB appears to be common recommendations from most of the studies.

3. THE MAIN CHALLENGES FOR DATA ACQUISITION, DATA STORAGE, AND DATA ANALYSIS

The main challenges for Data Acquisition, Data Storage, and Data Analysis are Data Volume, Data Variety, Data Integration, Complex Analytics, Constant Business changes, Skill sets, Compliance, Security, Data Quality, and Computing. The following paragraphs present to explain these challenges specific to all these three (ETL, DW, and OLAP) fundamental components.

3.1. ETL

ETL specific challenges are in deciding relevant and non-relevant data for extract, minimizing the performance overhead on the source system, flexibility with the data conversion, data clean-up, and balancing the real-time and batch loading.

- The extraction process has the potential to cause performance impacts on the source system [6].
- Source data systems typically are of different types. Standard formats are relational databases, flat files, non-relational databases (IMS, NoSQL). The extraction process should be able to read these diverse data types, and the transformation process should be able to apply complex rules in converting these to a single standard format for loading.

- The loading process typically happens as incremental batch jobs. These batch jobs get scheduled with minimum intervals to facilitate near-real-time load or with no intervals to support real-time loading. Completion of one scheduled before the start of another schedule is critical and mandates the implementation of audit trails to store all data changes associated with the batch job.
- The space efficiency is nearly zero (Guo, Yuan, Sun, & Yue, 2015) for traditional ETL due to the redundant storage of the data for the staging phase. Frequent re-loading takes place along with the application logic changes causing more data redundancy [7].
- Lack of a standard conceptual model to represent ETL processes is a real problem [8]. This problem makes ETL processes less flexible, and especially with every application logic change and the addition of new data sources, it results in a significant amount of work and reworks.

3.2. Data Warehouse

Data Warehouse specific challenges are mainly related to storing a massive amount of data. Data integration, duplication, cleansing, and timing of the data often add the complexities in maintaining the data quality. Suboptimal decisions are often due to these data quality issues. Complexities with volume (emergence of IoT), velocity (real-time), and variety (disparate source systems) of data puts a significant dependency on data quality. Khoso (2016), has estimated the growth of the total amount of data in use to shoot up to 44 Zettabytes in 2020 [9]. If we take healthcare as an example, volume (electronic health records, patient data like biometrics, variety (doctor notes, images), and velocity (IoT, wearables) have to ensure data quality to help with patient health, detection of diseases and cost-efficiency. The capability of the DW system to support real-time and on-demand workloads are quintessential. However, availability and performance issues post a significant barrier to supporting real-time workload [10]. No simplified DW scalability solution exists to tackle the explosion of data growth. Reporting on compliance (like GDPR) and security needs a tremendous amount of changes in the data warehouse architecture. Complexities exist with the Query-response service level agreement (SLA) adherence for all sizes of data ranging from single-digit terabyte-sized database to petabyte sized database.

3.3. OLAP

OLAP specific challenges include sophisticated analytics and different data types. The limitations of the underlying DBMS and compatibility put additional complexity to advanced analytics. Storing of analytic data inside the data warehouse or in a separate analytic database is a crucial design decision [10]. Computing requirements for advanced analytics are high, and the ability to accurately estimating the compute requirements is still a difficult task to answer. Variety and volume of the data make the data visualization and the dashboards to be highly dependent on IT staff.

4. DESIGN AND IMPLEMENTATION GUIDELINES FOR ETL, DW, AND OLAP

The core of the design principle of ETL, DW, and OLAP is to maintain its life cycle as a continuous loop covering conceptual, logical, physical, and starting again with the conceptual. Preference of a requirements-driven (demand-driven) approach over data-driven (supply-driven) approach for defining data warehouse requirements. Jukic and Nicholas (2010) have proposed the following requirements-driven framework [11]:

1) Data warehouse requirements collection and definition with end-user business analytical needs in the picture, 2) ETL design and development, 3) front-end application design and development, and 4) use/maintenance and growth with a feedback loop to step one.

ETL design should facilitate the flexibility (batch mode vs. real-time), metadata repository (source data and target data relationship map), and data integration (proactively fit data variations). Execution Sequence (which to run first?) and Execution Schedule (When to run?) and Recovery plan (fallback solution) are the three fundamental principles to be included [8]. "schema on read" and "schema on write" design needs to be balanced. If the source data types are constant, schema on write would be an ideal choice. If the source data types are unknown and continually changing, "schema on read" would be acceptable. ELT (Extraction, Load, and Transform) instead of ETL can be more useful for "schema on read" use cases.

Going with the ETL Tool instead of custom-built routines is highly recommended for most of the use cases [12]. The downside for TOOL based approach versus custom-built approach is mostly with the license cost and sometimes with the lack of flexibility (depending upon the business needs). In addition to cost, the other attributes that need to look at with an ETL tool are real-time refresh capabilities, metadata management, performance optimization capabilities, simplified upgrades (avoid forklift upgrades), and broad compatibility. Staging, DW Database, Data Marts, Operational Data Store [13], and Aggregation layer [14] form the conceptual architecture of a well-designed DW environment. Metadata repository with Technical (table structure) metadata and Business (business rules) metadata, as per Lam (2004), can help with cost optimization and faster implementation [15].

A real-time data warehouse is a key trend and transforms how a business run. Change Data capture should be part of the solution as it cares only about the changed data at the source and propagates the data changes to the target in an efficient manner. Data reduction techniques such as data virtualization can avoid data redundancy and optimize storage needs indirectly, helping with the performance and data integrity. A virtual layer to read the disparate source data from one place as virtual tables (utilizing heterogeneous database link / ODBC) could provide additional efficiency. Data virtualization also minimizes the complexities of the data staging layer.

Data warehouse appliance (Netezza, Teradata, Oracle Exadata), and Cloud computing (IAAS and SAAS) are the standard solutions for an accelerated start. "Clouds provide hefty infrastructure quickly, cheaply, and flexibly" [10]. Redshift offering from Amazon Web Services, Microsoft Azure's SQL DW service, and similar offerings from IBM and Google Cloud has evolved in the last few years and presents a viable alternative for traditional and on-premises implementations. Including Data Lake, NoSQL, and Hadoop as an extension of the traditional relational data warehouse and data mart databases as an answer to Big Data Analytics. Flexible deployments (including mobile), self-service IT, agile approach, high level of automation at all levels, analytics sandbox, are the norms of a modern data warehousing.

OLAP design should facilitate multiple designs to support departmental level needs. Ma, Chou, and Yen (2000) emphasizes the use of OLAP, Artificial Intelligence, and Visualization interfaces as determining a factor for the success of data warehousing [16]. Aggregate capability, analysis and deriving capability, fast online access, and interactive querying capability are the four primary characteristics of OLAP [16]. Data warehouse needs to be optimized to support both reporting and advanced analytics to avoid storing the analytical data in another database. The in-database analytics feature of the database combined with ANSI-SQL analytical capabilities gets heavily utilized with this approach of storing analytical data in the data warehouse database. The crux of the OLAP design is to eliminate IT dependency and making it self-service BI. Russom

(2009) recommends using Service Oriented Architectures or Web Services as data warehouse options [10].

5. ASSESSMENT METHODOLOGY FOR ETL, DW, AND OLAP

Data Quality, System Quality, User Experience, and Cost-Efficiency characterizes an assessment model for ETL, DW, and OLAP. The standard measures of data quality are accuracy, consistency, comprehensiveness, and completeness of the data [17]. The standard measures of system quality are scalability, flexibility, reliability, response time, and integration [17]. The space efficiency metric determines the "data redundancy" and other sub-optimal data storage methods. The adoption of "data reduction" techniques (data virtualization, ELT) maximizes storage efficiency. This metric can be monitored and reported periodically to fine-tune the transformation process (ETL).

Service Level Agreement guidelines for every layer, such as:

1. Data load time window: The documentation of Data latency requirements. (example: The loading of Claims data within 2 hours of source addition)
2. Query performance: Documented expectation on query retrieval time. Specific examples would be,
    a. 2 seconds of response in ODS 90% of the time,
    b. 1-2 minutes of response in data marts,
    c. 1-2 hours of response time for canned queries against EDW
    d. agreed-upon response times for ad-hoc analysis
    e. operational and disaster recovery
        i. Recovery Time Objective (example: 24 hours)
        ii. Recovery point objective (example: 2 seconds)

Any deviation in achieving SLAs would trigger the fine-tuning of ETL, DW, and OLAP processes for continuous improvement opportunities. Metrics such as cost per terabyte (with storage, software, hardware, and people cost integrated) would not only help internal department level chargeback but also for evaluating newer hardware/technology changes. Data Archive and Purge policy in place to keep the data size in check and to incorporate storage tiering, cold, warm, and hot for performance and cost perspectives based on data usage (most used for hot, moderately used for warm, and least used for cold).

6. CONCLUSIONS

In this paper, we looked at the main challenges of the three fundamental components of BI, Data Acquisition, Data Storage, and Data Analytics. Our analysis highlighted agility, the next generation of data, real-time data, situation awareness, and self-service as the drivers for the need for updated guidelines. We listed the compilation of design and implementation guidelines for ETL, DW, and OLAP. We also described the assessment methodology. The new design, implementation, and assessment guidelines of DW, ETL, and OLAP would help decision-makers and BI IT practitioners in proactively avoiding the "known" pitfalls.

ACKNOWLEDGMENTS

The author would like to thank the Editorial Secretary and the reviewers at IJDMS for the prompt correspondence and their recommendations in improving this paper.

REFERENCES

[1] Gartner. (2018). Gartner IT Glossary.

[2] Sonia Ordonez Salinas and Alba Consuelo Nieto Lemus. (2017). Data Warehouse and Big Data Integration. *International Journal of Computer Science and Information Technology (IJCSIT)*, 9(2). doi:10.5121/ijcsit.2017.9201

[3] Vasco Santos, Rui Silva, and Orlando Belo. (2014). Towards a low-cost ETL System. *International Journal of Database Management Systems (IJDMS)*, 6(2). doi:10.5121/ijdms.2014.6205

[4] Hicham El Bousty, Salah-ddine Krit, Mohammed Elaskiri, Hassan Dani, Khaoula Karimi, Kaoutar Bendaoud, and Mustapha Kabrane. (2018). Investigating Business Intelligence in the era of Big Data: Concepts, benefits, and challenges. *ICEMIS*. doi:10.1145/3234698.3234723

[5] Quoc Duy Vo, Jaya Thomas, Shinyoung Cho, Pradipta De, and Bong Jun Choi. (2018). Next-Generation Business Intelligence and Analytics. Association for Computing Machinery (ACM). doi: 10.1145/3278252.3278292

[6] Dhanda, P., & Sharma, N. (2016). Extract transform load data with ETL tools. *International Journal of Advanced Research in Computer Science*, 7(3)

[7] Guo, S., Yuan, Z., Sun, A., & Yue, Q. (2015). A new ETL approach based on data virtualization. *Journal of Computer Science and Technology*, 30(2), 311-323.

[8] Anand, N., & Kumar, M. (2013). Modeling and Optimization of Extraction-Transformation-Loading (ETL) processes in Data Warehouse: An overview. *The international conference for computing, communication, and networking technologies (ICCCNT).*

[9] Khoso, M. (2016) How much data is produced every day? North Eastern University.

[10] Russom, P. (2009). *TDWI best practices report Next-generation data warehouse platforms*.

[11] Jukic, N. and Nicholas, J. (2010). A framework for collecting and defining requirements for data warehousing projects. *Journal of Computing and Information Technology*, 2010(4), 377-384.

[12] Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2015). *The Kimball Group Reader*. Hoboken, NJ: John Wiley & Sons.

[13] Sacu, C. (2010). DWCMM: *The Data Warehouse Capability Maturity Model* (master's thesis). Utrecht University, The Netherlands.

[14] IBM (2012). Best practices physical database design for data warehouse databases.

[15] Lam, S. (2004). Methodology to minimize cost and time for new data warehouse implementation. *A thesis submitted to the faculty of graduate studies in partial fulfillment of the requirements for the degree of master of science*, University of Calgary, Calgary, Alberta.

[16] Ma, C., Chou, D. C., & Yen, D. C. (2000). Data warehousing, technology assessment, and management. *Industrial Management & Data Systems*, 100(3), 125-135.

[17] Wixom, B. H., & Watson, H. J. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly*, 25(1), 17-41.

AUTHOR

Ramesh Venkatakrishnan is a seasoned database professional with 20+ years of extensive experience in all aspects of Advanced Oracle database administration. He is currently pursuing his Doctor of Computer Science (Big Data Analytics) at CTU.