International Journal of Research Publication and Reviews

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

A Study of Big Data in Data Warehouse

***Mohmmed Asharaff*[1], *Shanavas Moosafintavida*[2]**

[1]Research Scholar, Universidad Azteca -Azteca University,Chalco, Mexico
byjumannar@gmail.com
[2]Research Guide, Universidad Azteca -Azteca University,Chalco, Mexico
dr.shanavas@univ-azteca.com

ABSTRACT

One of the goals of Big Data systems is to allow for the analysis of data from many sources. Since data warehouses have been used for decades to accomplish the same aim, they might be utilized to analyze data stored in Big Data platforms as well. Many academics throughout the world have looked into the subject of adapting data warehouse data and schemata to changes in these needs as well as data sources. However, novel approaches must be created to support the evolution of data warehouses that evaluate data stored in Big Data systems.In this paper, I propose a data warehouse architecture that allows different types of analytical tasks, such as OLAP-like analysis, to be performed on big data loaded from multiple heterogeneous data sources with varying latency, and that can handle changes in data sources as well as evolving analysis requirements. The architecture's operation is heavily reliant on the information described in the article.

Keywords: Big Data,Data Warehouse,OLAP,Evolution

1. Introduction

Many years, data warehouses and OLAP approaches have been utilized to enable structured data analysis, and as a result, many solutions to established research problems have been developed in the context of traditional relational database environments. One of these issues is data warehouse evolution, which occurs as a result of changes in business requirements, data sources, or data warehouse architecture enhancements.Due to the growing volume and heterogeneity of data that needs to be processed and analyzed, Big Data technologies [1] that leverage distributed data storage systems and process data in parallel have recently emerged. The demand for analyzing data stored in such systems is growing, and data warehousing is one of the analytical alternatives. Several recent papers have outlined open Big Data issues and research directions. To address the challenge of Big Data evolution, an architecture that allows users to store and process structured and unstructured data at various levels of detail, analyze it using OLAP capabilities, and manage changes in needs and data expansion semi-automatically. The metadata described in the paper are used to operate the architecture components responsible for OLAP analysis and evolution handling.

2.Big Data Warehouse Architecture

Traditional data warehouse evolution framework [3] inspired the concept of the Big Data warehouse architecture. The data warehouse evolution framework was expanded to the setting of Big Data.

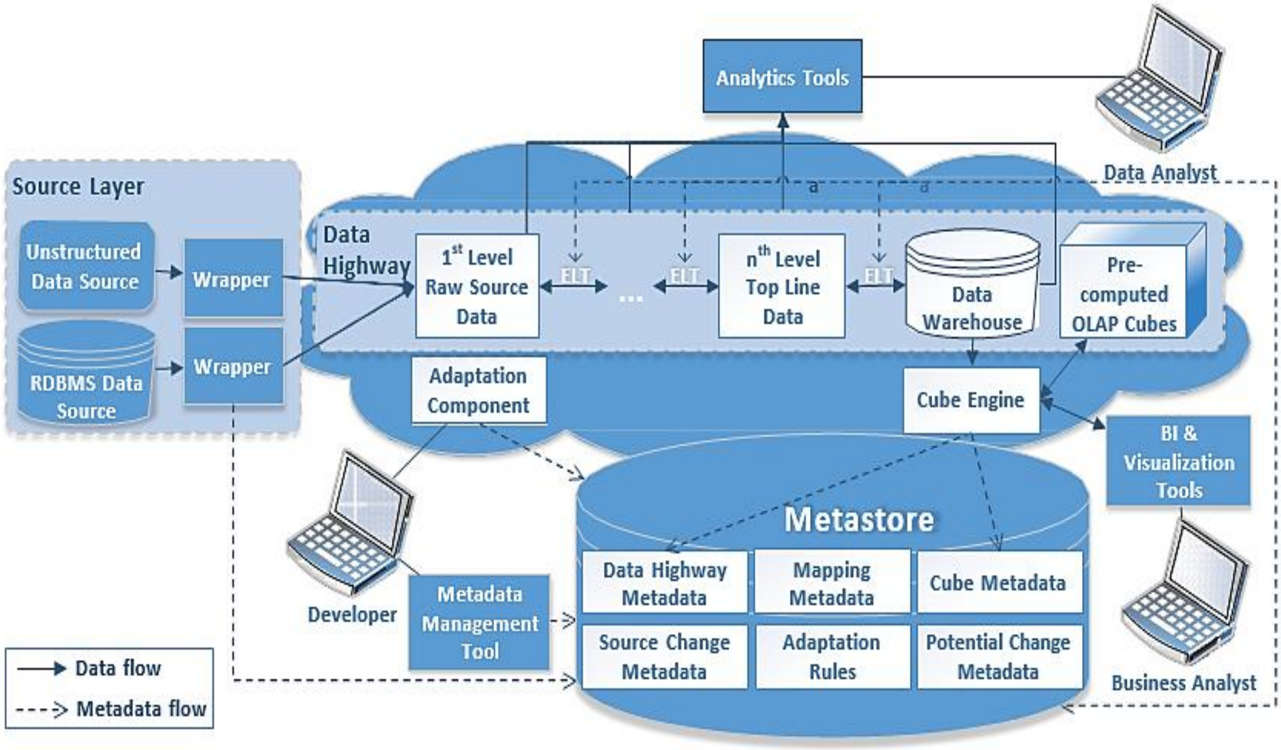

Fig 1: Big Datawarehouse architecture

2.1. Architecture Components

The suggested architecture is made up of various components (Fig: 1). Wrappers in the source layer acquire structured and unstructured data from data sources and load it into the system in its original format at varied rates. We embrace [4]'s proposal to construct a data highway with varying levels of latency. Starting with the raw source data, each subsequent level is derived from the preceding level data and is updated less frequently. Furthermore, data from numerous heterogeneous sources is gathered and then turned into a structured data warehouse schema at the latter level.ELT (Extract, Load, Transform) processes are responsible for data pre-processing in the proposed architecture since data is copied in their original format first and then processed subsequently. Advanced approaches such as data mining or sentiment analysis must be used to extract structured data from unstructured sources.

The adaption component is in charge of dealing with data source changes. For each change in a data source, the fundamental idea is to generate numerous alternative modifications in a data warehouse or other levels of the data highway, allowing a developer to choose the most relevant change to apply. Certain types of changes need the developer to provide additional data via the adaption component that cannot be discovered automatically.

We intend to support various types of analysis. Business analysts can use OLAP cubes to create dashboards, charts, and other reports, as well as perform OLAP operations utilizing business intelligence and visualization tools. The cube engine component pre-computes various dimensional combinations and aggregated measures for them since the volume of data stored in the data warehouse may be too enormous to offer adequate performance of data analysis queries. Apart from OLAP operations, data analysts might use existing analytics tools or create ad-hoc procedures to perform advanced analysis techniques (for example, data mining).

2.2. Metadata Management

The metastore, which includes six types of interconnected metadata required for the operation of other components of the architecture, is one of the most important components of the proposed architecture. The semantics and schema of Big Data stored at various levels of the system are described by data highway metadata. Cube metadata describes the schemata of pre-computed cubes and can be used not just for cube computation but also for query execution. The logics of ELT processes are defined by metadata mapping. They keep track of the connections between data from various sources and data highway items.The source change metadata records information regarding changes in data sources. This information can be gleaned through wrappers or while ELT procedures are running. Adaptation rules specify which adaptation options must be used for various sorts of changes. Finally, prospective change metadata collects proposed data warehouse schema changes.

A developer uses the metadata management tool to keep track of the data in the metastore. Furthermore, the metadata management tool enables the developer to make changes to the data highway and ELT operations to accommodate new or changing data requirements. The potential change metadata also keeps track of the history of chosen modifications made to propagate data source evolution, as well as changes made directly through the metadata management tool.

3.Related Work

Although various ways to handling data warehouse evolution in relational database environments have been presented, such as [5], [6,] they cannot be immediately applied to Big Data system adaption. Only a few Big Data systems, such as [7], explore the evolution element of Big Data; nevertheless, the given system's goal is not data analysis and it does not use a data warehouse.

Although there are various studies devoted to multidimensional Big Data analysis, such as [8], they do not address the issue of Big Data evolution. We identified only a few studies that address this issue. The authors of the work [9] explain a model enrichment process and propose iterative execution of the methodology's model design stage for building a Big Data analysis system. The approach proposed is complementary to ours. Slowly changing dimensions and fact table schema versions in metadata are supported by a data warehouse solution for Big Data analysis presented in the study [10]. The system does not process changes in Big Data sources, unlike our plan.The paper [11] describes an architecture that uses Big Data technology for OLAP analytics at LinkedIn. When a new dimension is introduced to the cube, the authors discuss the topic of cube evolution. Manual cube schema redefinition and data recalculation are used to deal with such modifications.

4.Conclusions and Future Work

The data warehouse architecture for facilitating Big Data analysis was proposed in this study. Our proposed architecture has the unique ability to react to changes in needs or data expansion automatically or semi-automatically.

Because the architecture is still not fully completed, we need building metadata models to define data highway schemata, data requirements, source data, and changes, as well as developing algorithms for automatic and semi-automatic change detection and treatment. We want to employ existing tools and technologies as well as develop original solutions to achieve the big data warehouse architecture.

References

- Thusoo, A., Sarma, J.S.., Jain, N., Shao, Z., Chakka, P., Zhang, N., Antony, S., Liu, H., Murthy, R.: Hive - a petabyte scale data warehouse using Hadoop. In: International Confer- ence on Data Engineering, pp. 996–1005. (2010)
- Ceravolo, P., Azzini, A., Angelini, M., Catarci, T., Cudré-Mauroux, P., Damiani, E., Mazak, A., Van Keulen, M., Jarrar, M., Santucci, G., Sattler, K., Scannapieco, M., Wimmer, M., Wrembel, R., Zaraket, F.: Big Data Semantics. J Data Semantics 7(2), 65-85 (2018)
- Solodovnikova, D.: Data Warehouse Evolution Framework. In: Spring Young Researchers Colloquium on Database and Information Systems SYRCoDIS, pp. 4. (2007)
- Kimball, R., Ross, M.: The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. 3rd edition. John Wiley & Sons, Inc. (2013)
- Golfarelli, M., Lechtenbörger, J., Rizzi, S., Vossen, G.: Schema versioning in data ware- houses: Enabling cross-version querying via schema augmentation. Data & Knowledge En- gineering, 59(2) 435-459 (2006)
- Ahmed, W., Zimányi, E., Wrembel, R.: A Logical Model for Multiversion Data Ware- houses. In: Intl. Conf. on Data Warehousing and Knowledge Discovery, pp. 23-34. (2014)
- Nadal, S., Romero, O., Abelló, A., Vassiliadis, P., Vansummeren, S.: An integration-ori- ented ontology to govern evolution in big data ecosystems. In: Workshops of the EDBT/ICDT 2017 Joint Conference. (2017)
- Song, J., Guo, C., Wang, Z., Zhang, Y., Yu, G., Pierson, J.: HaoLap: A Hadoop based OLAP system for big data. Journal of Systems and Software, 102, 167-181. (2015)
- Tardio, R., Mate, A., Trujillo, J.: An Iterative Methodology for Big Data Management, Anal- ysis and Visualization. In: International Conference on Big Data, pp. 545-550. (2015)
- Chen, S.: Cheetah: A High Performance, Custom Data Warehouse on Top of MapReduce. VLDB Endowment, 3(2), 1459-1468. (2010)
- Wu, L., Sumbaly, R., Riccomini, C., Koo, G., Kim, H.J., Kreps, J., Shah, S.: Avatara: OLAP for Web-scale Analytics Products. VLDB Endowment, 5(12), 1874-1877. (2012)