DATA VIRTUALIZATION FOR DECISION MAKING IN BIG DATA

Manoj Muniswamaiah, Tilak Agerwala and Charles Tappert

Seidenberg School of CSIS, Pace University, White Plains, New York

ABSTRACT

Data analytics and Business Intelligence (BI) are essential components of decision support technologies that gather and analyze data for faster and better strategic and operational decision making in an organization. Data analytics emphasizes on algorithms to control the relationship between data offering insights. The major difference between BI and analytics is that analytics has predictive competence which helps in making future predictions whereas Business Intelligence helps in informed decision-making built on the analysis of past data. Business Intelligence solutions are among the most valued data management tools whose main objective is to enable interactive access to real-time data, manipulation of data and provide business organizations with appropriate analysis. Business Intelligence solutions leverage software and services to collect and transform raw data into useful information that enable more informed and quality business decisions regarding customers, market competitors, internal operations and so on. Data needs to be integrated from disparate sources in order to derive valuable insights. Extract-Transform-Load (ETL), which are traditionally employed by organizations help in extracting data from different sources, transforming and aggregating and finally loading large volume of data into warehouses. Recently Data virtualization has been used to speed up the data integration process. Data virtualization and ETL often serve unique and complementary purposes in performing complex, multi-pass data transformation and cleansing operations, and bulk loading the data into a target data store. In this paper we provide an overview of Data virtualization technique used for Data analytics and BI.

KEYWORDS

Data Analytics, Business Intelligence, Big data, Data Virtualization, ETL and Data Integration.

1. INTRODUCTION

Success of an organization depends upon their business strategies and decision-making process. These decisions are heavily dependent on the collection and analysis of data been gathered. In the 1990s, during which business intelligence and analytical development were just blooming, data generated through various legacy systems were mostly structured. The decision support systems were based on these data collected that were stored in relational databases. These databases also supported queries, online analytical processing and reporting on the enterprise-specific data. Besides reporting functionalities, additional data mining techniques such as clustering, regression analysis, anomaly detection and classifications was also supported.

In the early 2000s the raise of internet helped HTTP-based web search engines such as Google, Yahoo and e-commerce companies such as Amazon to deliver their businesses online and interact with their users directly. Companies began collecting user specific data through server logs and cookies in order to better understand user behaviors and identify new business opportunities. Web based analytics tools such Google and Adobe analytics was developed to determine the user clickstream data logs that hinted the user's behavior across web pages and help in better key conversion metrics of a website. These tools provide key insights through user path analysis, user

DOI: 10.5121/ijsea.2019.10503

behavioural analysis, search terms that drives traffic and pages that are visited most. With these understanding, better web site design, personalization and recommendation engine can be built to enhance marketing plans and increase customer acquisition. In recent times, the use of IoT ("Internet of Things") such as mobile and sensor devices have increased in usage and provides an opportunity for analytics based on location-aware and user-centric operations [1].

Today, the environment in which businesses operate is constantly changing and getting more complex. With the advent of smart phones, IoT devices, health care devices, real-time media, huge volumes of big data get generated from them. This data hold key insights that would help the organization in shaping their business decisions and predictive analysis. Besides, organizations are also required to analyze these real-time data and quickly adapt with better tactical and operational decisions in order to meet market, customer and technology demands. Hence, data integration process which extracts, transforms and loads the generated data into data warehouses becomes a crucial operation for all business intelligence systems. Extracting data from multiple databases which are either in cloud or on-premise, is an integral part of data integration process. Data from these repositories can be extracted in many ways using either push or pull techniques. Often, the same business entity have different semantic value in which case it needs to be reconciled and correlated. Extracted data must undergo cleansing and transformation process without which it is highly impossible to consolidate the data for deriving key insights. By harnessing the power of business intelligence and analytical techniques, various applications such as e-commerce, healthcare and security can tap into their databases and uncover potential analytical insights and predict behavior to improve their business performance. In this paper we are focused on Data analytics and Business Intelligence process using Data virtualization for data integration [2].

The traditional data integration approach of extracting data from multiple sources into a central target data warehouse after cleansing, demoralization and transformation for analytical reporting is called Extract-Transform-Load. ETL process moves bulks of large data sets from operational and transactional systems to enterprise data warehouses which runs on high end parallel computational hardware systems. As the data sources evolved to become more heterogeneous, it becomes a challenge to rely on the ETL for data integration as these were designed to handle structured data. Organizations are embracing new technologies such as Data virtualization for adapting to the new data sources and formats.

Data virtualization is the latest data integration technique where data is no longer transferred from data sources. Instead it creates a single abstract layer between data producers and data consumers for accessing data originating from many different applications and distributed across different storage locations. Through data virtualization, users, business intelligence tools can virtually connect to variety of data sources, aggregate and combine data to form virtual views which later can be exposed as data services or APIs to support multiple formats such as SQL. This abstraction provides better agility and has shorter data integration life cycle. Data virtualization can be a vital part of data integration and business intelligence solutions. In this research we examine data virtualization technique impact for analytics and business intelligence and contrast it with traditional data warehouse process [3]

2. BACKGROUND

Data analytics and Business Intelligence drive organizations decision making process and helps in improving their stand in the competitive business world. Data discovery is the process of collecting data from various databases in silos and unifying them into a single source that can be easily and instantly evaluated. It is an iterative business-oriented process which involves identification of hidden patterns by utilizing various advanced techniques such as artificial

intelligence, heat maps, pivot tables, graphs that answers almost all business queries and helps organizations in meeting their goals. Thus, Data discovery consolidates all the business information and enables organizations to make improved data driven decision making.

Data cleansing is the next important stage of data integration. Data cleansing involves methods to identify corrupted, incomplete or duplicate data sets and correct or remove these discrepancies. This ensures the quality, accuracy, completeness, consistency and uniformity of the data discovered is preserved.

Data cleansed is then transformed so that it can be converted to more appropriate format to support various end to end processes. Data transformation normalizes tables and resolves any data type and unit differences in the data. Other activities such as data mapping, discretization, aggregation, remove null or duplicate data etc. are also a part of data transformation.

It is common to have large databases and understanding the relations between tables and data, is the first crucial step of data correlation. Having to integrate data across heterogeneous data stores often results in semantic integration problem where the same business entity can have different meaning in a different semantic context. In such cases, it becomes very important to interrelate information from each semantic context so that detail and context of the data involved is maintained. Data can be correlated based on filtering, joins or aggregation. Business analysts can examine their hypothesis on the data available after data correlation stage by applying statistical, logical techniques. The analyzed data can be visualized using various graphical tools for describing and illustrating purposes.

Figure 1: Phases of Data Integration

ETL is a batch process which involves migrating data from their operational source data stores to staging area. Extraction, Transformation and Loading are the essential components of data integration process. During extraction, data from all possible sources are collected in different formats. Connections are established to all the relevant source data stores for selecting and collecting necessary data required for analytical processing. Several business rules and validations are then applied to the extracted data during transformation to ensure high quality data. This involves eliminating duplicates, rejecting, reformatting, sorting, normalization and many other processes depending on the data in order to meet the requirement of the target data warehouses. Finally, the transformed data is then loaded to target data warehouses either in a sequence or through batch process. Data warehouse off loads data analysis related work from source data stores, provides an integrated and consistent view of the integrated data. Data warehouse also supports materialized views of the tables, indexing on columns and creation of star and snowflake schemas which groups data into fact table containing business related information [4].

Data virtualization does data cleansing, transformation, association and correlation from source data stores evading any in-between physical data movement. Its main objective is to provide a single source of truth to all the users without requiring them to know any technical details or location of the underlying data in real-time. At each stage, distinct virtual views are created using virtual tables and each step in-turn uses virtual views from the preceding view to create their own virtual view. Any request for the data by the reporting tools are executed against these virtual

views, which in turn transform the request into series of queries needed to access data. It uses connectors such as JDBC or ODBC to access the source data. The relationship of different tables, attributes and constraints metadata are stored in data source catalogs which is used by Data virtualization. Virtual tables are the result set of queries which acts like a regular table but are not stored physically and hence the name, virtual tables. When a virtual table query is executed, the query itself is broken down into sub queries. Each of these sub queries are reformatted so that they can be executed on their respective underlying data stores where the data resides while still behaving as a single data model from the top. Multiple views would be defined at various level of abstraction and these virtual views can be reused which increases data quality and reduces data latency. When a query is executed data is returned to these views for normalization before producing the result [5].

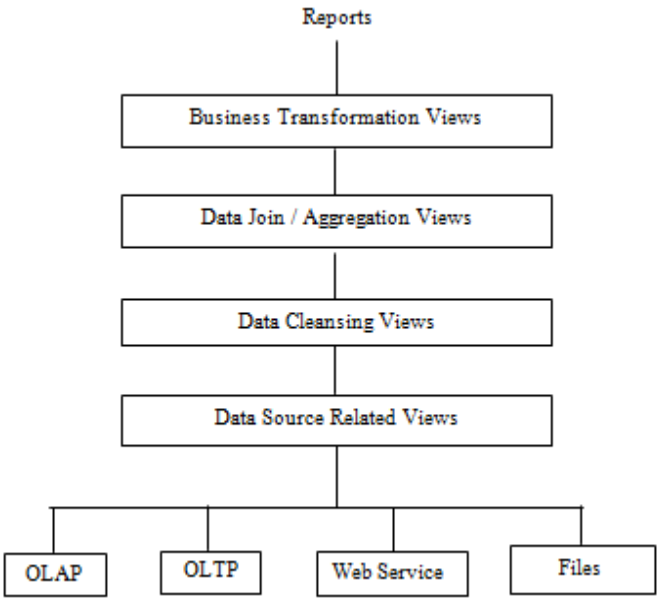

Figure 2: Data Virtualization

Since data is fetched from disparate data sources and views are defined for various data integration stages, the relevance of having a cost-based query optimizer becomes important which reduces the query latency and improves performance. Figure 2 shows the various views of the Data virtualization process which extracts, cleans and transforms the data fetched from different sources.

3. DATA VIRTUALIZATION OVERVIEW

Data virtualization is the modern data integration technology which plays a very crucial role in today's Business intelligence analytics of an organization. It facilitates extraction of valuable business insights by creating an abstracted, encapsulated virtual layer between data consumers and the variety of data sources that are distributed across different locations or reside in the cloud. It provides a unified view of data by integrating all the underlying data sources. The abstraction layer exposes only accessible data to the users without requiring them to know or move the data from their physical location. This ensures that the users meet the data governance policies as well increase their access rate to real-time data.

Data virtualization enables easy unified overview of an enterprise data gathered from multiple distributed data sources and manipulation thus eliminating data duplication and compression across different databases. This also helps in reducing infrastructure cost needed to maintain disparate databases to a great extent. This reduces silos that are created by having multiple applications for multiple data repositories within an organization. It does not require to perform ETL process but instead virtually connects different databases to provide virtual views and publish them. By doing so, data virtualization ensures that data access violations are minimized by exposing only part of data users are interested in, masking sensitive financial or personal data. Besides, data is made readily available for analysis and reporting in real-time. Data virtualization is an integral part of data management which extracts greater value from data sources.

Data virtualization delivers data as a service to interested data consumers on demand. It also enables data transformation through user interface and eliminates the need for replication since data is not moved from the sources physically. Lesser the data movement, lesser the management of multiple inconsistent copies of outdated data. Virtualized abstraction layer hides the storage structure and technology from the users, allowing them to access structured and unstructured data and focus on the required tasks. It also makes it easy for users to experiment with new ideas, speed up prototyping and use the test result according to their requirements. It brings agility to business decisions as it provides an adaptable data environment. Customized virtual views are constructed as the request for new data flows thus enabling users to react to any new change instantly. The infrastructure cost is also reduced as the administrators are exempt from operational and data storage costs. This in turn lowers power consumption required to maintain all the hardware and maximizes energy efficiency. Data integration from cloud sources and on-premise databases is also made easy when organizations adopt Data virtualization. It helps in improving services of existing or new products providing speed-to-market value.

Data virtualization enables logical data warehouse functionalities which federates queries across data warehouses and provides data access using different protocols. Organizations requires access to complete, consistent data both in historical and real-time in order to run their business in a data-driven market. They are heavily dependent on data in order to undertake any strategic and important business initiatives. Ability to have faster access and greater visibility of their enterprise data is what enables swift and profitable decisions leading organizations to adopt different technologies which are designed for special requirements. Having a unified view, allows organizations to use data more efficiently and gain faster insights from more accurate, updated data and provide better business intelligence. Traditional ETL process needs to handle bulk and outdated data from previous operation. It is a scalability issue as it requires advanced and more powerful applications when dealing with data at zetta or petabytes scale. Data virtualization on other hand streamlines the standard ETL process by providing a platform that integrates data from different databases on demand and creating a single point of access to its consumers thus improving scalability [6].

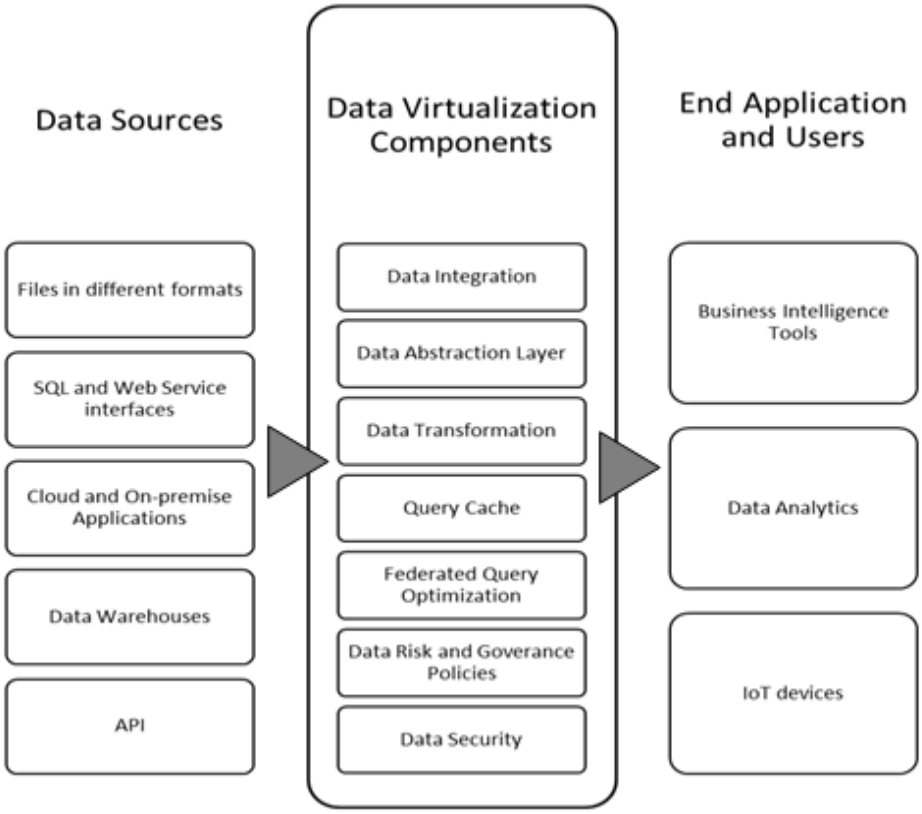

Figure 3: Data Virtualization Overview

Figure 3 shows how Data virtualization integrates different data sources which are soiled and helps in the BI and analytics decision-making process.

4. DATA VIRTUALIZATION TECHNIQUES

Data analytics and Business Intelligence help in improving and grasping insights hidden in their data that are discovered from various techniques such as data integration, data mining and statistical analysis. Most of these technologies depend upon relational databases, data warehouse, BPM, ETL and OLAP cubes for analysing historic and current data patterns. Popular algorithms have been incorporated into data mining systems including K-means, Apriori, AdaBoost, KNN, Support Vector Machine and PageRank which helps in clustering, classification and association analysis. Data analytics aims at cleansing and transforming raw data into meaningful format that helps is future business predictions. Having to extract knowledge by analysing various patterns of the data requires immense effort. Hence, Data analytics continues to be an active area of research for Data science and statistical analysis community. The commercial databases show efficiency in query processing and having high level query interface.

Most firms today use Business Intelligence of some form and it plays a key role in performance of the organization. A variety of modern Business Intelligence tools are now available, and it involves cost. Organizations need to understand the full breakdown of total cost of ownership when adopting any BI solution. It depends on what is already installed, and the hardware required to upgrade the data warehouse. Software cost involves license purchase and subscription to various Business Intelligence packages. Implementation cost increases when organizations

require more complex and customized solutions. Operations and maintenance cost include vendor support for adjusting security settings, data backup, initial staff training and annual software and hardware maintenance costs [7].

ETL and Data virtualization are both data integration tools. While ETL facilitates historical accumulation of data, Data virtualization is more agile, flexible, dynamic and cost. Data virtualization dynamically computes an optimal way to fetch data from different dissimilar sources and achieve necessary joins and transformations and presents the results to the users. This allows organizations to remain focused on fine-tuning their business initiatives without having to know about the location of the data. Data virtualization does not move data from the sources rather it delegates the queries to the source data stores. When requests are issued, the data sources are queried in real time and results are returned. Also, they are more agile with the data models where new data stores can be added easily and help in the rapid iteration of project development life cycle. Data virtualization defers costly commitment to ETL process and accelerates the dialog between business users and IT to reduce the risk of an ETL process and help in developing efficient data marts [8].

Data virtualization is the most suitable choice for Business Intelligence and analytics when structured and unstructured data from dissimilar sources needs to be combined and queried quickly. This is very important for business decisions on inventory levels and portfolio risk analysis. It also helps in eliminating data duplications and privacy risk concerns regarding the data being accessed. The data required for analytics needs to be transformed, undergo cleansing and enriched before it can be used which are done through virtual tables.

Business Intelligence and analytics that traditionally use ETL can be enhanced to include semi-structured and unstructured sources. It can help identify areas for improvement, track performance, predict success of new initiatives and identify market trends that increase revenue. It can be used to pull data from social media to analyse user behaviour patterns and build recommendation systems. Mobile applications that access corporate data requires a virtualization that separates these applications from the underlying data sources. Mobile applications can access corporate data through REST web services and Data virtualization can adequately help in accomplishing this.

Data virtualization helps in building Business Intelligence system from either the existing data warehouse or from disparate data sources virtually. It also helps in pulling data from various components such as CRM and provides an integrated view of data for data analysts and data scientists. This provides flexibility and time-to-value for any business decisions been made. Information governance policies can also be implemented to bring in compliance with industry regulations [9].

Table 1: Key Characteristics of BI, Technologies and Research [9]

KEY CHARACTERISTICS OF BI	TECHNOLOGIES	RESEARCH
1. Structured data 2. Relational database and data warehouse 3. ETL and OLAP cubes 4. Dashboards and reporting **5.** Data mining	1. Cloud Relational Databases 2. Cloud data warehouse 3. Cloud based ETL 4. BPM 5. Clustering 6. Classification 7. Regression analysis 8. Anomaly detection 9. Deep learning 10. Sequencing and Genetic algorithms	1. In-memory analytics 2. Parallel processing 3. Cloud computing 4. Statistical machine 5. Learning 6. Mining IoT data 7. Temporal mining 8. Spatial mining 9. Columnar data stores

Table 2: Data Virtualization and ETL categories [9]

BI & ANALYTICS CATEGORY	DATA VIRTUALIZATION	ETL
1. Time to value	Could be implemented quickly	Takes longer time
2. Requirements	Requirements can evolve	Requirements needs to be well defined before implementation
3. Data cleansing	Generally single pass	Generally multi pass
4. Application use	Tactical decision making based on operational data	Heavy analytical BI and analytics
5. Data formats	Can handle both structured and unstructured data	Mostly limited to structured data
6. Data availability	Available in near real time	Data is available at the end of load operation
7. Data Volume	Depends on the view capabilities	Can process large amount of data

5. CONCLUSION

Data virtualization is powerful data integration platform that reduces complexity of data management systems and provides single consolidated, integrated view of the data. It helps resolve the issue of data silos which are created by multiple applications. Data virtualization abstracts the users from the underlying data sources and allows for real-time data access and brings agility to decision making process. It eliminates the need for replication as data is not

moved physically from the source. Virtualization maximizes data utilization in many ways that can benefit any business. It is also infrastructure agnostic which reduces project life cycle time.

REFERENCES

[1] Chen, Hsinchun, Roger H.L. Chiang, and Veda C. Storey (2012),"Business Intelligence and Analytics: From Big Data to BigImpact,"*Management Information Systems Quarterly*, 36 (4),1165–88

[2] http://web.mit.edu/smadnick/www/wp/2013-10.pdf

[3] https://www.tibco.com/sites/tibco/files/resources/wp-ten-things-data-virtualization-final.pdf

[4] http://www.northtexasdama.org/wp-content/uploads/2017/03/1_Data-Virtualization.pdf

[5] http://www.datavirtualizationblog.com/data-movement-killed-the-bi-star/

[6] https://www.astera.com/type/blog/data-virtualization-technology-overview/

[7] https://globaljournals.org/GJCST_Volume17/3-Emerging-Virtualization-Technology.pdf

[8] https://www.astera.com/type/blog/data-virtualization-technology-overview/

[9] Muniswamaiah, Manoj & Agerwala, Tilak & Tappert, Charles. (2019). Data Virtualization for Analytics and Business Intelligence in Big Data. 297-302. 10.5121/csit.2019.90925.