# DATA VIRTUALIZATION FOR ANALYTICS AND BUSINESS INTELLIGENCE 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) is essential for 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 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 available. Business Intelligence solutions gather and examine current, actionable data with the determination of providing insights into refining business operations. Data needs to be integrated from disparate sources in order to derive insights. Traditionally organizations employ data warehouses and ETL process to obtain integrated data. Recently Data virtualization has been used to speed up the data integration process. Data virtualization and ETL are often complementary technologies 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 the decision making process. These decisions are based on the collection and analysis of data been gathered. During the early stages of business intelligence and analytical development in 1990s the data collected and processed was mostly structured, collected from legacy systems and stored in relational databases which also supported online analytical processing and reporting on the enterprise-specific data. In addition to reporting functionalities data mining techniques such as clustering, regression analysis, anomaly detection and classifications was also supported. In the early 2000s the raise of internet helped web search and e-commerce companies such as Google and Amazon to present their business online to the users and interact with them. Companies began collecting user specific data through logs and cookies in order to understand user behaviors and improve business. Web and text analytics was developed to determine user browsing and purchasing patterns. Web site design, personalization and recommendation engine can be built using web analytics. 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 person-centric operations [1].

Data needs to be collected from disparate data sources and processed as a part of data integration process. Analysis and correlation of data provides key insights into the business decisions and also helps in predictive analysis. Organization sales, human resources and marketing department gather and analyze data obtained from data integration process. They have multiple databases

Natarajan Meghanathan et al. (Eds) : CCSEA, CLOUD, SIPRO, AIFU, SEA, DKMP, NCOM - 2019
pp. 297-302, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90925

which are in cloud and on-premise, extracting data from all these repositories and sources is a part of data integration process. Data from these repositories can be extracted in many different ways using either push or pull techniques. Also, same business entity could have different semantic value which needs to be reconciled and correlated. Extracted data also undergoes cleansing and transformation process before deriving key insights from it. Business Intelligence and analytical techniques also provide business-centric methodologies which can be applied to various applications such as e-commerce, healthcare and security. 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 moving data from source to target database after cleaning, demoralization and transformation is called Extract-Transform-Load. The target data store is called data warehouse which runs on high end parallel computational hardware systems. Data virtualization is the new technique which does not move the data from the source data stores. Instead all the cleansing and transformation of data is done through virtual tables. It provides better agility and has shorter data integration life cycle. 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

Organizations employ Data analytics and Business Intelligence for decision making process. Data discovery helps in integration of all the available data across the organization. It is common to have large databases and understanding the relations between tables and data is the first step of data integration. Data discovery is the process of collecting data from various databases in silos and consolidating it into a single source that can be easily and instantly evaluated. Having heterogeneous data stores results in different semantic definition of the same business entity which needs to be cleansed to remove discrepancies. Once data is cleansed it requires transformations to normalize tables and resolve any data type and unit differences in the data. Data can be correlated based on filtering, joins or aggregation. Business analysts can examine their hypothesis on the data available after data correlation stage. The analyzed data can be visualized using various tools for presentation.


Figure 1: Phases of Data Integration

ETL process involves moving the data from source data stores to staging area and later it undergoes cleansing and transformation before been loaded in to target data warehouse. 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. Each of these stages are distinct using virtual tables, each step uses data from preceding one by using virtual tables. 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 query which acts like a regular table, it is virtual since data is not physically stored and data is brought from underlying table when virtual table query is executed. Multiple views would be defined at various level of abstraction and when a

query is executed data moves through these views before producing the result [5].

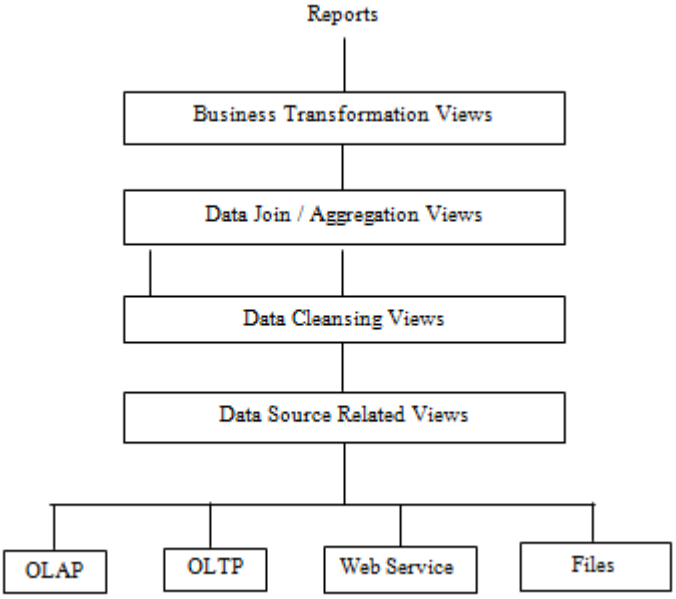


Figure 2: Data Virtualization

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 facilitates extraction of business insights by creating an abstraction layer and providing a unified view of data which are in traditional repositories or in cloud. The abstraction layer exposes only accessible data to the users without requiring them to know about the physical location of the data and it ensures that the users meet the data governance policies. Data virtualization enables easy data gathering and manipulation by reducing data duplication and compression across different databases. This also helps in infrastructure cost savings. It does not require to perform ETL process but instead virtually connects different databases to provide virtual views and publish them. This makes data readily available for analysis and reporting. This reduces silos across different data repositories with in an organization. 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 users. It also enables data transformation through user interface and eliminates the need for replication since data is not moved from the sources physically. Abstraction layer hides the storage structure and technology from the users allowing them to focus on the required tasks. It makes it easy for users to use the data according to their requirements. It brings agility to business decisions as data is readily available. The infrastructure cost is also reduced as the administrator are exempt from operational cost and data duplication. 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. Business requires real-time and historical data to make decisions leading organizations to adopt different technologies which are designed for special requirements. Having abstraction layer enables collective benefit of their functionalities. Traditional ETL process needs to handle bulk and outdated data from previous operation which is streamlined using Data virtualization [6].

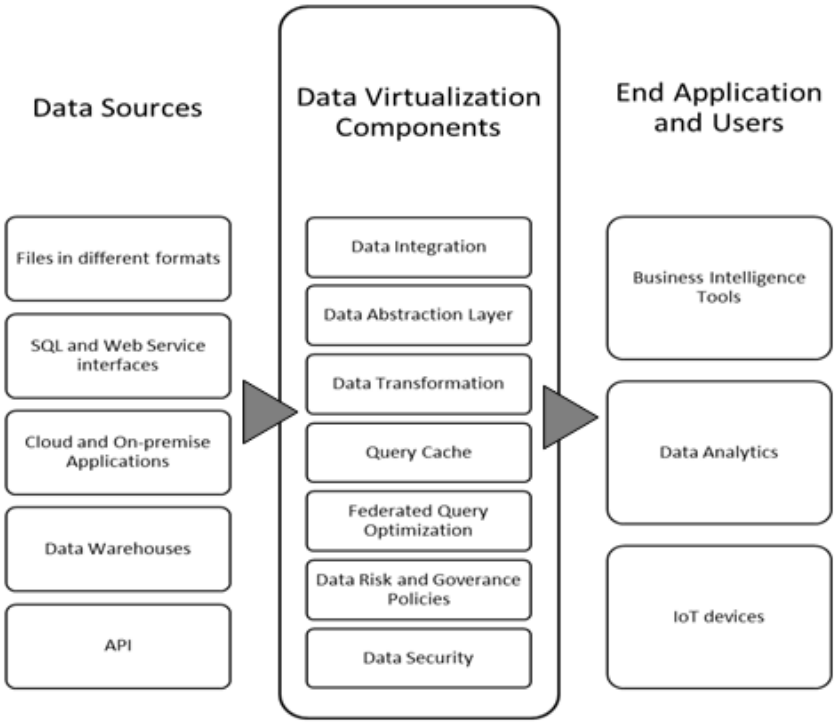


Figure 3: Data Virtualization Overview

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 gaining insight that is been discovered from data integration, data mining and statistical analysis. Most of these technologies depend upon relational databases, data warehouse, BPM, ETL and OLAP cubes. Popular algorithms have been incorporated into data mining systems including K-means, Apriori, AdaBoost, KNN, Support Vector Machine and Page Rank which helps in clustering, classification and association analysis. Data analytics continues to be an active area of research due to 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 involves cost. It depends on what is already installed and the hardware required to upgrade the data warehouse. Software cost involves subscription to various Business Intelligence packages and implementation cost includes initial training and annual software and hardware maintenance costs [7].

ETL and Data virtualization are both data integration tools. Data virtualization is more agile and cost efficient than ETL. Data virtualization computes optimal way to fetch data from different sources and achieve necessary joins and transformations and presents the results to the users without knowing 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 a decent 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 extend to include unstructured sources. 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 [9].

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.

Table 1: Key Characteristics of BI, Technologies and Research

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

Table 2: Data Virtualization and ETL categories

| 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 reduces complexity of data management systems and also 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. Finally it is 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] https://www.sciencedirect.com/topics/computer-science/data-virtualization-layer