RESEARCH REVIEW JOURNALS
Volume-04
Issue-06
June-2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com[UGC Listed Journal]

The Advantage of Data Warehousing Technology in Big Data and Data Mining

[1]Jignesh P. Shah and [2]Dr. A.J. Patel

[1]*Research Scholar, Department of Statistics, Gujarat University, Ahmedabad, Gujarat (India)*
[2]*Head & Associate Professor, R. H. Patel Arts and Commerce College, Ahmedabad, Gujarat (India)*

ARTICLE DETAILS

Article History
Published Online: 04 June 2019

Keywords
BDW, OLAP, PAP, DWH, computing, Data, Technology, Data Governance, Data profiling

ABSTRACT

A data warehouse (DW or DWH), is also known as an enterprise data warehouse (EDW). Data warehousing is a system used for storing, reporting and data analysis technique that coverts raw data to a simplified required formatted data. DWs are the techniques where it tries to extract more value from larger data sets and tries to maintain central repositories of integrated data from one or more than one different sources. They store current and past data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons. The data stored in the warehouse is uploaded from the operational systems. The data may pass through an operational data store for additional operations before it is used in the data warehousing for reporting. The types of system consist data mart, online analytical processing (OLAP), online transaction processing (OLTP) and predictive analysis processing (PAP). A Big Data Warehouse (BDW) is an architecture for data management and organization that utilizes both traditional data warehouse architectures and modern Big Data technologies, with the goal of providing rapid analysis across a broad range of information types. This paper addresses some of the research challenges in Big Data Warehousing systems, proposing a vision that looks into: i) the integration of new business processes and data sources; ii) the proper way to achieve this integration; iii) the management of these complex data systems and the enhancement of their performance.

1. Introduction

In the present decade organizations Data Warehouse (DWH) Technology is a database system which is designed for analytical analysis instead of transactional work. Here DWH is also used for solving decision related problems, because of highly unstructured and semi-structured situations. Data mining is a process of analyzing data patterns whereas Data Warehousing is the process of pooling all relevant data together. Thus, Data mining is considered as a process of extracting data from large data sets. Actually, decision making is made at all levels of Management. Even though decision making at top levels remains always crucial and complex. Present Executive Decision Maker is more interested in forecasting the future, future opportunities, as well as problems which are going to incur in near future. So that well in advance they can prepare to take the opportunity or to deal with the problem situation. Thus, to solve any type of problem the decision making is required, and to make decisions Information is required which is made available by the technology called Data Warehousing (DWH) Technology. The fundamental use of data, which can be refereed as a separate discipline, as that from operational use. The operational databases have been basically designed to meet mission critical requirements of the On-Line Transaction Processing and Batch Processing. The strategic data usage is basically requiring On-Line Query Processing or Batch-intelligence gathering for decision support. In business systems efficiency is no longer a success major factor, it has been replaced by flexibility and responsiveness. In addition to this the emergence of communication factor is to be added to this system. Because of the enormous development in communication infrastructure the transactions have been processed in real-time mode of business operations. The Organization which has come to known the power of information is getting the stronger competitive edge over their rivals and the solution for this is Data Warehousing (DW)Technology.

In surfacing of Data Warehousing the information value and its usage has been newly recognized. Data warehouses works as means for the strategic data utilization. A Data Warehouse works as an integrated platform having the integrated data of refined quality to which an executive decision maker gets support by making usage of Decision Support System (DSS) or EIS like applications. By making consolidation, conversion, transformation and integration of operational data and by provision of a consistent view, a Data Warehouse has been enhanced the productivity of decision maker. A senior decision maker is flavored very well by making usage of EIS applications, which allows him the data as per their working style. EIS does have the ability to report regarding impulsive drill-down functionality and provides high performance, and highly user-friendly as compared to analyst. Briefly, Data Warehousing is a blend of technologies having orientation towards effective integration of operational databases in the environment that enable the usage of data for strategic purpose.

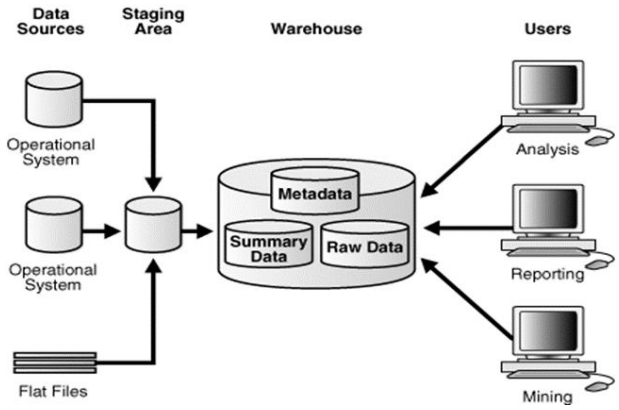
2. Why Data Warehousing (DWH) And Data Mining?

Data Warehouse is database system which is designed for analytical analysis instead of transactional work. Data Mining is the process of analyzing data patterns. For Big data, Data is stored periodically and it is analyzed regularly in Data mining process. Data Warehouses are not any physical products, instead it is the process of extracting and storing data to allow easier reporting whereas Data mining is used to recognize the pattern of data and it is a kind of logic that identifies the

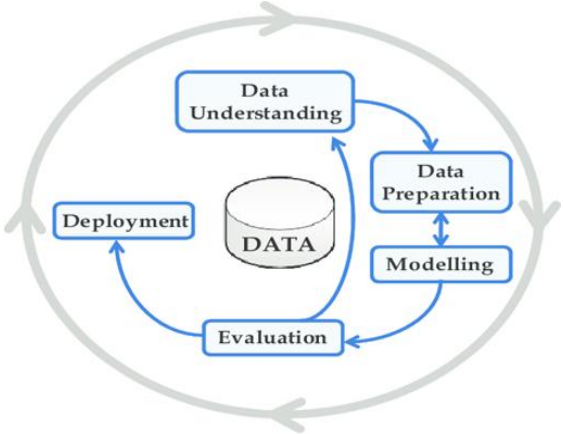
patterns. The Data Warehouses plays vital role of cornerstone, by providing the ability to the organizations to perform information processing effectively. It is a combination of construction of Information Systems, in which the user with present and historical information can be used for decision making. In other words, the data accessing becomes harder, which needed by Decision Maker. Data Warehousing provides strategic business opportunities by allowing customers and vendors access to corporate data while maintaining required security measures. In order to survive and succeed in today's stronger competitive global environment, the business people & users do need the answers because of following reasons:

- Decisions are required to be made fast and correctly, by using available data.
- The Transactional data generated, which gets doubled every one and half years that cause slow-response time, and inability to extract the required data contents.
- Due to competition there is an increase in information value.
- Business experts may be people who are non-computer.

Structures Of Data warehouse



Structures Of Data Mining



3. Why Big Data warehouse (BDW)?

A Big Data Warehouse is an architecture for data management and organization that utilizes both traditional data warehouse architectures and modern Big Data technologies, with the goal of providing rapid analysis across a broad range of information types. While analytics can certainly be run exclusively on Big Data repositories or on enterprise data repositories. It is the combination of the two types of repositories into a *unified* data architecture that distinguishes a *Big Data warehouse.*

Forrester defines the Big Data warehouse as: "A specialized, cohesive set of data repositories and platforms used to support a broad variety of analytics running on-premises, in the cloud, or in a hybrid environment. BDW leverages both traditional and new technologies such as Hadoop, columnar and row-based data warehouses, ETL and streaming, and elastic in-memory and storage frameworks."

Key elements of the Big Data Warehouse

A Big Data warehouse architecture typically encompasses the following elements:

- **A data repository:** These include repositories for both Big Data and enterprise, structured data. A Big Data warehouse typically draws from multiple data repositories, including traditional relational databases that is house structured, enterprise data; columnar data stores tailored enterprise data aggregation and Big Data stores that handle both unstructured and structured data in massive volumes.
- **Compute/processing:** Fundamental processing can happen at multiple levels in a Big Data warehouse architecture. Fast-turn analytical processing can also happen at a higher layer, such as using the Spark engine on Big Data. Machine learning analytics can also be applied at a higher level in the stack.
- **Data management capabilities:** The data management capabilities necessary for an effective Big Data warehouse include:
  - Data integration (tying systems together).
  - Data quality (ensuring a level of cleanliness or correctness of information).
  - Data transformation (ensuring consistency of data format).
  - Data security.
  - Data governance (ensuring compliance with appropriate policy and regulatory rules).
  - Data pattern analysis.
- **Interactive analytics.** Interactive analytical capabilities include in-memory analytics, ad hoc interactions, or the ability for analysts to do self-service analytics on the underlying data.
- **Advanced analytics.** In addition to traditional data analysis techniques, organizations can also add advanced analytical engines to data managed by the Big Data warehouse architecture. This includes predictive analytics, graph analytics, and spatial analytics.
- **A variety of data environments.** Big Data warehouses is typically a variety of data environments that often combining on-premises databases, cloud data stores, and hybrid environments that have already been integrated. While it is possible for some organizations to have all on-premises environments or all cloud environments, this is increasingly unusual.

Big Data warehouse general architecture

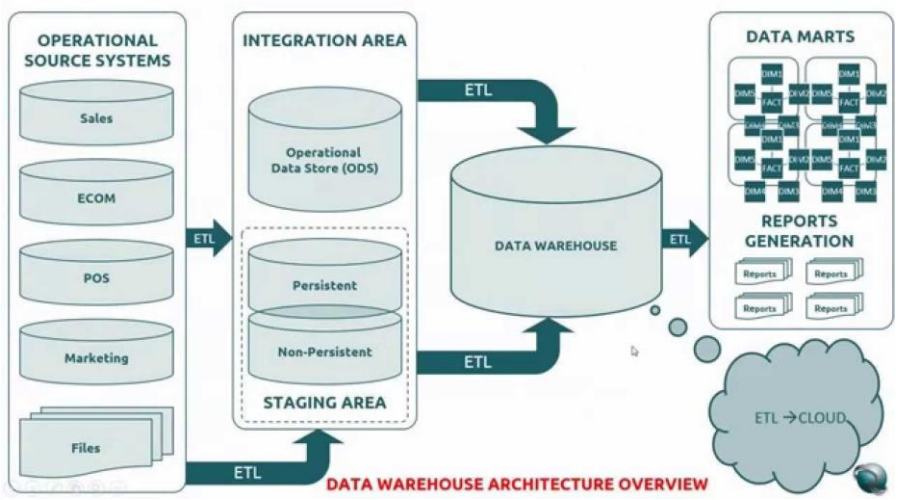


Figure: Generic Big Data warehouse architecture.

I. Key issues to keep in mind

- **Ease of integration.** By definition, a Big Data warehouse requires the integration of a wide variety of data repositories, processing capabilities, and analytical capabilities. Thoroughly investigating the ease of integration of major components of the Big Data warehouse will be key to initial deployment success and also the ongoing success of the architecture.
- **Extensibility.** Rapid innovation in data management, data storage, and analytics happens simultaneously. Ensuring that the architecture can be easily extended to incorporate emerging technologies will be important to ensure the ongoing relevance of the overall data architecture.
- **Transposition.** It is difficult to create data pipelines that cross the different elements of the data warehouse.

II Technological Causes

Data Warehouses are designed in such a way to over-come the problem of incompatibility of Informational and Operational Transactional Systems. The EDP and MIS are designed to satisfy the frequently, incompatible requirements of the business. But IT infrastructure is changing at a rapid speed and its capabilities are increased, as evidenced as follows:

- The price of computer processing speed MIPS (Million Instructions per second) is constantly declining, while the microprocessor power is getting doubled very fast.
- At the same time the prices of digital storage media is decreasing at a rapid speed.
- Bandwidth of network is increasing, with decrease in the price of high bandwidth.

Because of shift in computing paradigm the Distributed Client or Server computing is emerged in organizations and hence more and more critical applications were being placed on this new environment. The early developed Client/Server solutions were found to be scarce to solve today's business problems. This shortage was found in case of runtime and new development aspects of the Client-Server computing. It was not problem to the design or architecture of client-severs, but it has additionally introduced new demands on the environment itself.

Object Orientation:

- The need for development in object-oriented analysis design and programming, which was became the requirement of the time.
- It can be said that the implicit or explicit acceptance of object orientation by IT practitioners, which coupled with the emergence of object-oriented standards, and availability of related developments and run-time tools, results, which needs to be accepted by the client or server model.

Middle Ware:

- This is the layer in client or server architecture which transforms two-tier client or server computing model into a more complicated client-middle ware-server model.

Storage and Handling of multifaceted data:

- The system ability to handle, store, manipulate a variety type of data for instance Video, text, images, special & time series data in new-object-oriented database systems, as if they were traditional data types.
- High performance commercial parallel computing and largest database (VLDB)Processing.

4. Why Model Shift?

There is an enhanced movement in computing paradigm. The traditional view of computer users was updating in access to a computer through communication network. The main importance was to access to known computer program lying on known or a remote system. This concept was mainly like host-based or master-slave computing. The user makes use of computers to solve the problems. Thus, there is an assumed understanding that the services requested by a user may not be available on a single system but may be distributed across a network. The users make use of their computer as doorstep to have access to this distributed computing power. This computing power is roughly classified

as workgroup computing and Local Area Network or Wide Area Network environments.

The main purpose is laid on shared reusable resources, the attention given to the internet information highway, and that of more and more use of global networking services, such as internet and the World Wide Web (WWW). These are evidences which made the shift in computer paradigm. Thus, the primary shift is prompted by Data Warehousing and its related technologies. Due to increased need of right sizing, the process of business processing reengineering, and the result of this there is an introduction of work group computing, and a prominent role played by decisions support systems. The users/executives which are information workers now a days works within paperless office. So as to achieve these objectives the technologies like client or server computing, object-oriented technology, multimedia technology, distributed computing, usage of artificial intelligence and expert system, communications infrastructure are made use of comprehensively. Informational data can be pulled out from operational data resources. As operational data is in splinted form and inconsistent it cannot be used for decision support. Designing of Data Warehouses is done so as to provide an architecture which will make corporate data accessible and can be utilized by knowledge workers and by decision makers also.

Thus, the Data Warehouses with Big Data are significantly differing from operational systems in following respect:

- Supports large number (Big) of shoot transactions.
- Based is processed on daily based transactions.
- Organized to achieve best performance.
- Supported to a large number of concurrent users.
- Works on primary key direct record access.

BDW Modelling and Implementation

In the vision proposed in this paper, BDW Modelling and Implementation must be guided by appropriate methodological guidelines, and not by adhoc or use case-driven approaches, identifying the suitable data model to integrate the new data in the BDW, ensuring efficient query processing, mixed complex workloads, and an adequate decision support environment (see Fig. 1).

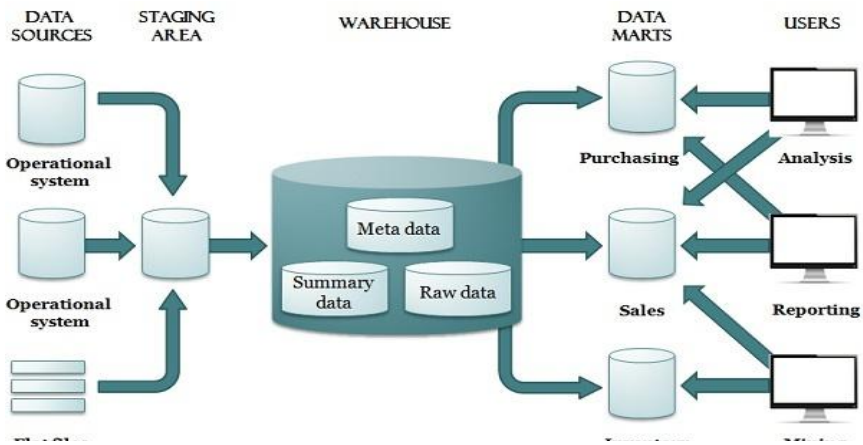


Fig:1(Data Warehouse with BWD architecture)

Different data modeling approaches can be followed for designing and implementing BDWs, such as completely flat (denormalized) data tables, star schema models, or hybrid approaches. Hence, different design patterns, which can optimize query processing, and for the new businessprocessesanddata,thedatamodellingconstructsneedto beinferredusingthe information available from the BDW Entities Resolution. Here, the characteristics of the data and the data modelling constructs are mapped, identifying:

- Analytical objects (such as sales, inventory management, purchases, among others).
- Complementary analytical objects.
- Descriptiveandanalyticalattributes,wheredescriptiveat tributesaddameaningtotheanalyticalattributes.
- Materialized objects (views) that increase efficiency for complex and long- running queries.

Using above data modelling points, a Design Patterns Knowledge Base (DPKB) is used to derive a data model and to later implement it using semi-automated procedures, adding the new physical structures and data to the BDW. This knowledge base stores information about the data modeling design patterns, its performance to the characteristics of the data. Hence, a new data model can be derived, suggesting its implementation by following a specific design pattern. Afterwards, i) if the volume of data increases, ii) if the data distribution changes, iii) if performance in query processing is not satisfactory, then the BDW Management component can recommend changes to the data model the adoption of different design patterns.

5. Conclusion

Data warehouse technology is made for satisfying the informational need of knowledge workers, executive decision makers in making strategic decision and it also support executive decision making at all the stages of management. In this paper, all these challenges were instantiated with research areas. Approaches from research areas like Entities Resolution, Data Profiling and Data Tagging can be applied to provide information in the proper way to achieve this integration, based on appropriate data models, with the attempt to provide a data model in a semi-automated way, based on a Design Patterns Knowledge Base (DPKB). This type of contribution will help organizations that deal with huge

amounts of data arriving from several sources and will help them to manage these complex data systems and to enhance their performance, reducing the time needed for tasks such as the BDW management and modeling, allowing their users to focus on retrieving value from data. Moreover, the capability to deal with other contexts, like events and streaming processing, automating the analytical capabilities of a BDW, is another way to enhance the BDW and its value.

References

1. Jignesh P Shah: The strategic role of Data Warehousing Technology in Decision Making: RESEARCH HUB – International Multidisciplinary Research Journal (RHIMRJ), Volume-2, Issue-1, January-2015 ISSN: 2349-7637.
2. Madden, S.: From Databases to Big Data. IEEE Internet Computing. 3(2012).
3. Dumbill, E.: Making Sense of Big Data. Big Data. 1, 1–2(2013).
4. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 35, 137–144(2015).
5. Philip Chen, C.L., Zhang, C.-Y.: Data-intensive Applications, Challenges, Techniques andTechnologies:ASurveyonBigData.InformationSciences.275,314–347(2014).
6. Costa, C., Santos, M.Y.: Big Data: State-of-the-art Concepts, Techniques, Technologies, Modeling Approaches and Research Challenges. IAENG International Journal of Computer Science. 44, 285–301(2017).
7. NBD-PWG: NIST Big Data Interoperability Framework,(2015).
8. Krishnan, K.: Data Warehousing in the Age of Big Data. Elsevier(2013).
9. WiederholdGio Database Design, PersonEducation
10. Jankiraman Foundation Of AI, BPBPublications
11. Bayross Ivan SQL, PL/SQL BPBpublication
12. Shrobe, Exploring AI, Survey Talks from the National Conferences on AICA, AAAI1988.
13. http://data-warehouses.net/

