Big Data for Data Warehousing

Introduction

A Data Warehouse (DW) is a central repository of information created by integrating data from multiple disparate sources. DWs store current as well as historical data that can be culled to create trending reports for senior management, such as annual and quarterly comparisons. The data stored in the warehouse are uploaded from operational systems such as sales, customer relationship management, and marketing systems. Data may pass through an operational data store for additional transformations before they are used in the DW for reporting. With the advent of big data and distributed computing, DWs will not only become more dynamic in data processing, but the disciplines of data science and statistics will help enterprises leverage their data more than ever before.

Traditional Data Warehouse Architecture

Monolithic Servers

The majority of DWs run on monolithic-style server systems. In general, a data warehouse solution has three key monolithic servers: the database (DB) server, the extraction, transformation, and loading (ELT) server and the Business Intelligence (BI) server. The DB server contains the database of the data warehouse, the data marts, and the data staging area. This server's purpose is to be the data repository that the other two servers will read from and/or feed data into. The ELT server's purpose is to house the extract, transform, and load systems that will be responsible for collecting data and changing it into a more usable form to produce reports or decision making systems. Lastly, the BI server houses the reporting and dashboard components the business uses to gain insight into the data found on the DB server.

This kind of system architecture usually calls for powerful machines that can accommodate multiple users accessing the systems and achieve low-latency response times to queries made upon the data. This system architecture works well for queries being made on the data for which the data structures are known. Usually, this kind of system runs trend analyses and business reporting on a certain subset of data or lines of business. Additionally, this system architecture tends to produce a lot of network traffic, since the ETL server must pull data from the DB server in order to process the request and load the results back into the DB server.
Known structures of data can easily be related to one another. Data that is known to most of a company's personnel and that is tightly coupled to other aspects of the company makes it an invaluable source of information for critical decisions.

Data Warehouses (DWs) are purpose-built to answer pressing business questions quickly, making them the most efficient place to find information on the status of the enterprise. By adding custom-tailored business intelligence dashboards, data in a DW can be interpreted at a glance.

Well-known technologies most professionals understand, like RDBMS and DWs, have been around for several decades and have small learning curves. Almost anyone with a basic knowledge of these systems can perform analysis with relative ease. Even recent college graduates have a general understanding of these technologies and don’t require extensive training to be able to contribute to projects or solve data problems.

Data Modeling Practice

In traditional data warehouse systems, the data will typically be stored in a Relational Database Management System (RDBMS). Storing data in this kind of an environment takes planning and effort, as the data must be structured in such a way that it is usable by the business and flexible enough to be able to manage changes. However, if the structure becomes too flexible, it becomes hard to find the needed data. It may take too long to answer queries and thus become unusable for answering business questions. Such a scenario may require significant time and effort to be put into remodeling data structures.

In order to address these issues, additional data warehousing modeling disciplines were created: the Inmon Data Model (William Inmon) and the Kimball Data Model (Ralph Kimball). Inmon’s philosophy recommends first building a large, centralized, enterprise-wide data warehouse, followed by several satellite databases to serve the analytical needs of departments (henceforth known as ‘data marts’). Thus, his approach was termed ‘Top Down.’ Kimball’s philosophy advocates first building several data marts that serve the analytical needs of departments, and then ‘virtually’ integrating these data marts for consistency through an Information Bus. Thus, his approach was termed ‘Bottom Up.’ Mr. Kimball believes in the efficacy of various data marts that store information in dimensional models to quickly address the needs of various departments and various areas of the enterprise data.

Besides the differences in philosophy, the Inmon and Kimball models also differ in how they structure data. The Inmon model is relational-model (third normal form: 3NF) whereas the Kimball model is a multi-dimensional model (star-schema and snowflakes). Inmon argues that once the data is in a relational model, it will attain the enterprise-wide consistency that makes it easier to spin-off data-marts in dimensional-models. Kimball argues that users can understand, analyze, aggregate, and explore data-inconsistencies more easily if the data is structured in a dimensional-model.

ETL Workflow

In DWs, ETL processes are used to collect and convert data to a format that is useful for business decision making and reporting.

An ETL process consists of the following steps:

1. Data Profiling
2. Staging
3. Transform/Load

Data profiling is the first and most important step in creating an ETL process in traditional DWs. Data profiling software analyzes the data to determine its correct structure, content, and quality. Data profiling uses rules to determine what constitutes acceptable or valid data, also known as metadata. The user gathers the metadata that is available for an application as a starting point then processes it against the rules to find differences, which generally occur due to the existence of incorrect data. This step helps find the correct structure and data attributes that will be modeled in the staging and warehouse/mart areas of the data warehouse. This step also determines which data attributes are needed in the data warehouse and which can be left behind on the source system.

In the staging step, the data that was determined to be essential to the data profiling step is extracted and loaded in raw form into the initial data processing area of the data warehouse. This step is useful because data can be extracted quickly from the source systems, ensuring data warehousing and data processes don’t bog down operational systems.

The transform/load step is where data is aggregated and formatted in ways that are useful for answering business queries. This step is where most of the business logic is applied to the data extracted from staging. The data is prepared within staging to be inserted into the warehouse.

Advantages of Traditional Data Warehouses

- Known structures of data can easily be related to one another. Data that is known to most of a company’s personnel and that is tightly coupled to other aspects of the company makes it an invaluable source of information for critical decisions.

- DWs are purpose-built to answer pressing business questions quickly, making them the most efficient place to find information on the status of the enterprise. By adding custom-tailored business intelligence dashboards, data in a DW can be interpreted at a glance.

- Well-known technologies most professionals understand, like RDBMS and DWs, have been around for several decades and have small learning curves. Almost anyone with a basic knowledge of these systems can perform analysis with relative ease. Even recent college graduates have a general understanding of these technologies and don’t require extensive training to be able to contribute to projects or solve data problems.
Limitations of Traditional Data Warehouses

- There is a general inflexibility in integrating new data models. When new sources of information are identified, it can take tremendous effort to integrate these new data models into a highly-structured and tightly-coupled system like a DW.

- Similar to the first limitation listed above, it takes a great deal of effort to implement changes to a DW in order to answer new questions. When an enterprise’s priorities change or markets change, the business will want a system that can change just as quickly to answer new questions. But having to modify or build new data structures, dashboards, and reports is a large undertaking because the new designs ideally should be compatible with the old in order to provide a holistic view of enterprise data.

- DWs are ill-suited for data discovery. Businesses increasingly rely on data to inform decisions; however, the problem with DWs is that much of the data they contain may end up being left behind in the upload portion of the ETL process due to size constraints or the business not being interested in certain data at the time of transfer. An incomplete data set mined for a specific purpose will be ill-suited to answer any other question, meaning that businesses trying to use a DW for data discovery will likely draw incorrect conclusions about patterns and trends.

- The downtime for a complete hardware upgrade can be massive. Due to the monolithic architecture of DW systems, if ever an upgrade of an entire server is needed, an enterprise runs the risk of downtime while the upgrade is in progress. The time and effort spent migrating a current system architecture to a new server makes upgrading even more costly.

Big Data Warehouse (BDW) Architecture

Distributed Architecture

Implementing a DW using big data technologies takes an enterprise away from monolithic architecture and into distributed architectures, in which several servers or components of hardware work together to accomplish a task or process. Big data technologies are the equivalent of a file system as opposed to a database. This gives big data technologies some compelling advantages like unlimited computing power and storage space, automatic data redundancy, automatic load balancing, or the ability to store an unlimited amount of data in any format.

Depending on how a BDW is implemented, an enterprise can save money on hardware by outsourcing the data center and by leveraging cloud computing services to host and process data. Since most big data technologies are based on open source projects, enterprises can save money on licenses and other fees.

Use Cases

Big data warehouse to traditional data mart

The data warehouse is implemented using big data technologies which give an enterprise the ability to process large amounts of information faster than a traditional DW could, along with the ability to house unstructured data for data discovery. The data marts are implemented using traditional data warehouse technologies that offer a more structured data environment with low-latency querying to answer a particular business question. This solution’s flexibility alleviates the most common concerns about DWs, (data processing or space limitations) as more nodes can easily be added without having to bring the entire server down.
**Big data repository to traditional DW**

In this solution, an enterprise can run its current ETL processes on a traditional DW that houses the data attributes important to the business. Additionally, an enterprise will create ELT processes that collect all data from source transactional systems. Not having to retroactively load data or run heavy data mining queries will give an enterprise the flexibility to answer new questions as they arise and find new trends and patterns without slowing the source system. Once an enterprise establishes the new data model that will be required in the traditional DW, it can collect the new data quickly and easily, in a way that will be transparent to other departments.

With this solution, an enterprise could also use the big data repository as a backup site for the traditional DW in case of a catastrophic system failure. Additionally, if an enterprise has processes that transfer data from a big data repository to the DW table structures, no retroactive loads will be required upon recovery.

**Augmented big data processing to DW or DM**

By augmenting data processing from a monolithic structure to a distributed system, an enterprise can use the distributed system as a data aggregation/integration point for data that might not be available in the traditional DW. This solution is typically used when an enterprise has a very processor intensive function that takes too long to complete or because an enterprise needs to run calculations on raw data that is impractical to store in a DW system. By leveraging big data technologies that offer greater computing power and storage space, an enterprise can process data quickly and efficiently.
ELT Workflow

In traditional DW environments, ETL is the term usually associated with the process of data acquisition and processing. With big data technologies, the order of execution is usually different compared to classic DW processes, wherein an ELT process has the load step occurring before the transform step. This is due to the fact that big data technologies are distributed file systems that can ingest data of any kind and format. Thus, an enterprise can use big data technologies to extract raw data from source systems, load it to a big data implementation program, and later transform it into any required format or structure. This is especially useful when mining new trends and behavior patterns from data; because one is managing raw, unstructured data, the same piece of information can be used to gain insight into multiple scenarios or can be used to answer other questions that the business may ask. When dealing with structured data, the ELT flow can still apply, only data profiling and staging are done within the same distributed system.

Advantages of Big Data Warehouses

Big data technologies have been around since 2007, but only recently have they been deployed outside Silicon Valley. Here are some advantages that helped make big data technologies a viable analytics platform:

- Limitless storage space and computing power to process terabytes of data. Scaling out a big data implementation is as easy as adding new servers to a cluster. This is something monolithic servers and databases cannot accomplish due to their physical limitations. Since big data technologies can be implemented on cheap commodity servers, the investment to create a cluster is minimal as compared to building a system on a larger, higher-end server. Also, since these technologies are available for free through open source projects, licensing costs can be kept low compared to the cost of database licenses. As an example, one could create a cluster of 10 servers that have 100 TB of storage and 40 processing cores that would cost as much as a single server that has less storage capacity and fewer processing cores.

- Effortless integration of any and all data types and models. Because big data is implemented on a file system and not a DBMS, an enterprise is able to ingest any and all formats and structures of data without incurring upfront overhead costs of integration.

- Flexibility to run purpose-built DW business decision systems, trend analysis and predictive analytics processes. Since a BDW can handle structured, semi-structured, and unstructured data, it can output very loosely coupled data that can be changed, manipulated and combined in any way to discover new insights on data.

- No downtime for forklift upgrades. Unlike systems using a more monolithic architecture, big data is able to accommodate entire sets of servers being taken offline for upgrade without incurring downtime or data loss. Since big data technologies replicate their data across difference nodes in the cluster, the risk of losing data due to hardware changes is minimal.

Limitation of Big Data Warehouses

All technologies have limitations, and big data brings with it some interesting challenges during implementation of a BDW:

- Unknown data structures make it harder to inter-relate data. Because a BDW can store unstructured data, it is difficult to know how this data will relate to the known data structures in an enterprise. One must also find out what the actual structure of the data is before one can start designing a data model to house this data in a formalized way.

- Personnel will need to be trained on how to use new technologies. Because these technologies are relatively new and introduce new programming models (map, reduce), few professionals know how to properly leverage these technologies. Thus, an enterprise will have to train their personnel to make them proficient in big data technologies. This in turn introduces a few challenges such as:

  - Rapid technological change. Because of their open source nature, big data technologies are undergo faster iteration cycles than traditional proprietary technologies, and new solutions are introduced into the market quickly. An example of this is that Hadoop will soon go to phase 2.0, addressing the single point of failure Namenode and implementing the concept of a Namenode Federation to address the single point of failure issue.

  - Personnel must know a wide range of technologies and how they can be integrated. An example of this is how the Hadoop project has spawned an entire ecosystem. Just knowing and training for Hadoop is not enough; personnel must also know how to implement and integrate frameworks and systems like Avro, Cassandra, Chukwa, Hbase, Hive, Mahout, Pig, Impala, Sqoop, Oozie and ZooKeeper.
Conclusion

As data becomes a more valuable asset and companies become more data centric, technologies like Hadoop are brought in to become the next data and analytics platform to drive business decisions to improve products and services. DW systems will need to adapt to leverage these technologies as data volumes keep increasing and demands for shortening processing times of these data volumes become more pressing.

In the future, it will be possible that DW will disappear entirely and BDW architectures will replace them in all enterprises across many industries. Because of this, it is important for companies to understand and implement BDW now so they are ready interface with different organizations and agencies to consume and process large data volumes. In addition, data transmission standards and data structure best practices can change over time, and a BDW will be able to ingest this data with no upfront overhead and the companies will be ready to work with them in an efficient and timely manner.

Big Data represent a paradigm shift which companies must be ready to embrace. Just like the paradigm shift from structure programming languages to object orientated programming, and from mainframe/terminal environment to client server environment, if companies do not learn from and adapt to the Big Data paradigm shift, they will likely lose competitiveness in the marketplace and miss opportunities in the future.

About the Author

Daniel Serrano is a Data Engineer with more than 10 years experience. He holds a Bachelor’s degree in Computer Science from the University of Puerto Rico Bayamón and a Master of Software Engineering from Polytechnic University of Puerto Rico. He has worked in the Telecom industry with Verizon Wireless, Insurance through Medical Card Systems Inc. and Hi Tech / Entertainment industry at Netflix. Currently at TCS he is responsible for architecting systems that scale out for customers across the US that have need solutions for ETL offloading, Data Warehouse offloading and massive data set processing.
About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India’s largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com

Contact

For more information about TCS’ Big Data Solutions, email us at nextgen.solutions@tcs.com