OPTIMIZING DATA WAREHOUSE WITH BIG DATA

JD Software Pvt Ltd., 23, Melandai 1st Cross Street,
West Tambaram, Chennai 600045, India
Tel: +91 44 22265767 Mobile: +91 9840902654

OPTIMIZING DATA WAREHOUSE WITH BIG DATA

JD Software Pvt Ltd

INTEGRATION STRATEGIES

Data integration refers to combining data from different source systems for usage by business users to study different behaviors of the business and its customers. In the early days of data integration, the data was limited to transactional systems and their applications. The limited data set provided the basis for creating decision support platforms that were used as analytic guides for making business decisions.

The growth of the volume of data and the data types over the last three decades, along with the advent of data warehousing, coupled with the advances in infrastructure and technologies to support the analysis and storage requirements for data, have changed the landscape of data integration forever.

Traditional data integration techniques have been focused on ETL, ELT, CDC and EAI types of architecture and associated programming models. In the world of big data, however, these techniques will need to either be modified to suit the size and processing complexity demands, including the formats of data that need to be processed. Big data processing needs to be implemented as a two-step process. The first step is a data-driven architecture that includes the analysis and design of data processing. The second step is the physical architecture implementation, which is discussed in the following sections.

DATA-DRIVEN INTEGRATION

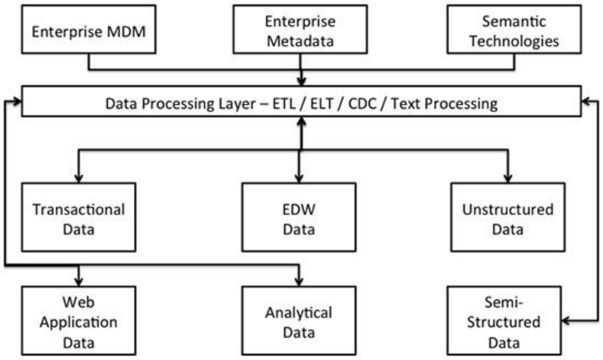
In this technique of building the next-generation data warehouse, all the data within the enterprise are categorized according to the data type, and depending on the nature of the data and its associated processing requirements, the data processing is completed using business rules encapsulated in processing logic and integrated into a series of program flows incorporating enterprise metadata, MDM, and semantic technologies such as taxonomies.

Data Warehousing and Analytics in the era of Big Data

Business reporting and analytics being generated from the data warehouse are now mission critical. There isn't a piece of information available (or not available) that your business analysts and analytic processes cannot consume to drive competitive advantage. New data types and external sources of information are needed to drive analytics. The volume of information is exploding and the need to make decisions on that information close to real-time is becoming more important.

Figure below shows the inbound data processing of different categories of data. This model segments each data type based on the format and structure of the data, and then processes the appropriate layers of processing rules within the ETL, ELT, CDC or text-processing techniques. Let us analyze the data integration architecture and its benefits.

DATA CLASSIFICATION



As shown in Figure above, there are broad classifications of data:

- *Transactional data.* The classic OLTP data belongs to this segment.
- *Web application data.* The data from Web applications that are developed by the organization can be added to this category. This data includes clickstream data, Web commerce data, and customer relationship and call center chat data.
- EDW data. This is the existing data from the data warehouse used by the organization currently. It can include all the different data warehouses and data marts in the organization where data is processed and stored for use by business users.
- *Analytical data.* This is data from analytical systems that are deployed currently in the organization. The data today is primarily based on EDW or transactional data.
- *Unstructured data.* Under this broad category, we can include:
  - Text: documents, notes, memos, contracts
  - Images: photos, diagrams, graphs
  - Videos: corporate and consumer videos associated with the organization
  - Social media: Facebook, Twitter, Instagram, LinkedIn, Forums, YouTube, community websites
  - Audio: call center conversations, broadcasts
  - Sensor data: includes data from sensors on any or all devices that are related to the organization's line of business. For example, smart meter data makes a business asset for an energy company, and truck and automotive sensors relate to logistics and shipping providers such as UPS and FedEx.
  - Weather data: used by both B2B and B2C businesses today to analyze the impact of weather on the business; has become a vital

    component of predictive analytics.

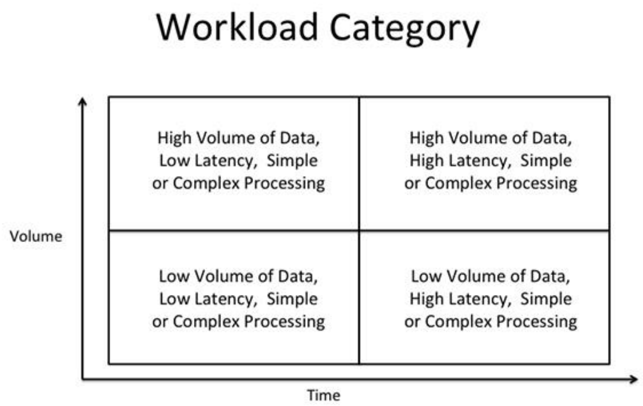
    - Scientific data: applies to medical, pharmaceutical, insurance, healthcare and financial services segments where a lot of the number-crunching type of computation is performed, including simulations and model generation.
    - Stock market data: used for processing financial data in many organizations to predict market trends, financial risk and actuarial computations.
- *Semi-structured data.* This includes emails, presentations, mathematical models and graphs, and geospatial data.

Architecture

With the different data types clearly identified and laid out, the data characteristics -- including the data type, the associated metadata, the key data elements that can be identified as master data elements, the complexities of the data, and the business users of the data from an ownership and stewardship perspective -- can be defined clearly.

Workload

The biggest need for processing big data is workload management, as discussed in earlier chapters.


Workload Category
High Volume of Data, Low Latency, Simple or Complex Processing
High Volume of Data, High Latency, Simple or Complex Processing
Volume
Low Volume of Data, Low Latency, Simple or Complex Processing
Low Volume of Data, High Latency, Simple or Complex Processing
Time

The data architecture and classification allow us to assign the appropriate infrastructure that can execute the workload demands of the categories of the data.

There are four broad categories of workload based on volume of data and the associated latencies that data can be assigned to (Figure 10.4). Depending on the type of category, the data can then be assigned to physical infrastructure layers for processing. This approach to workload management creates a dynamic scalability requirement for all parts of the data warehouse, which can be designed by efficiently harnessing the current and new infrastructure options. The key point to remember at this juncture is that the processing logic needs to be flexible to be implemented across the different physical infrastructure components, since the same data might be classified into different workloads depending on the urgency of processing.

The workload architecture will further identify the conditions of mixed workload management where the data from one category of workload will be added to processing along with another category of workload.

For example, processing high-volume, low-latency data with low-volume, high-latency data creates a diversified stress on the data-processing environment, where you normally would have processed one kind of data and its workload. Add to this complexity the user query and data loading happening at the same time or in relatively short intervals, and now the situation can get out of hand in quick succession and impact the overall performance. If the same infrastructure is processing big data and traditional data together with all of these complexities, the problem just compounds itself.

The goal of using the workload quadrant is to identify the complexities associated with the data processing and how to mitigate the associated risk in infrastructure design to create the next-generation data warehouse.

Analytics

Identifying and classifying analytical processing requirements for the entire set of data elements at play is a critical requirement in the design of the next-generation data warehouse platform. The underpinning for this requirement stems from the fact that you can create analytics at the data discovery level, which is very focused and driven by the business consumer and not aligned with the enterprise version of the truth, and you can equally create analytics after data acquisition in the data warehouse.

Figure 2, shows the analytics processing in the next-generation data warehouse platform. The key architecture integration layer here is the data integration layer, which is a combination of semantic, reporting and analytical technologies, which is based on the semantic knowledge framework, which is the foundation of next-generation analytics and business intelligence. This framework is discussed later in this chapter.

Finalizing the data architecture is the most time-consuming task that, once completed, will provide a strong foundation for the physical implementation. The physical implementation will be accomplished using technologies from the earlier discussions, including big data and RDBMS systems.

Physical component integration and architecture

The next-generation data warehouse will be deployed on a heterogeneous infrastructure and architectures that integrate both traditional structured data and big data into one scalable and performing environment. There are several options to deploy the physical architecture, with pros and cons for each option.

The primary challenges that will confront the physical architecture of the next-generation data warehouse platform include data loading, availability, data volume, storage performance, scalability, diverse and changing query demands against the data, and operational costs of maintaining the environment. The key challenges are outlined here and will be discussed with each architecture option.

Data loading

- With no definitive format or metadata or schema, the loading process for big data is simply acquiring the data and storing it as files. This task can be overwhelming when you want to process real-time feeds into the system, while processing the data as large or micro batch windows of processing. An appliance can be configured and tuned to address these rigors in the setup, as opposed to a pure-play implementation. The downside is that a custom architecture configuration may occur, but this can be managed.

- Continuous processing of data in the platform can create contention for resources over a period of time. This is especially true in the case of large documents, videos or images. If this requirement is a key architecture driver, an appliance can be suitable for this specificity, as the guessing game can be avoided in the configuration and setup process.
- MapReduce configuration and optimization can be daunting in large environments, and the appliance architecture provides you reference architecture setups to avoid this pitfall.

Data availability

- Data availability has been a challenge for any system that relates to processing and transforming data for use by end users, and big data is no exception. The benefit of Hadoop or NoSQL is to mitigate this risk and make data available for analysis immediately upon acquisition. The challenge is to load the data quickly as there is no pre-transformation required.
- Data availability depends on the specificity of metadata to the SerDe or Avro layers. If data can be adequately cataloged on acquisition, it can be available for analysis and discovery immediately.
- Since there is no update of data in the big data layers, reprocessing new data containing updates will create duplicate data, and this needs to be handled to minimize the impact on availability.

Data volumes

- Big data volumes can easily get out of control due to the intrinsic nature of the data. Care and attention needs to be paid to the growth of data upon each cycle of acquisition.
- Retention requirements for the data can vary depending on the nature of the data and the recency of the data and its relevance to the business:
    - Compliance requirements: Safe Harbor, SOX, HIPAA, GLBA and PCI regulations can impact data security and storage. If you are planning to use these data types, plan accordingly.
    - Legal mandates: There are several transactional data sets that were not stored online and were required by courts of law for discovery purposes in class-action lawsuits. The big data infrastructure can be used as the storage engine for this data type, but the data mandates certain compliance needs and additional security. This data volume can impact the overall performance, and if such data sets are being processed on the big data platform, the appliance configuration

can provide the administrators with tools and tips to zone the infrastructure to mark the data in its own area, minimizing both risk and performance impact.

- Data exploration and mining is a very common activity that is a driver for big data acquisition across organizations, and also produces large data sets as the output of processing. These data sets need to be maintained in the big data system by periodically sweeping and deleting intermediate data sets. This is an area that normally is ignored by organizations and can be a performance drain over a period of time.

Storage performance

- Disk performance is an important consideration when building big data systems, and the appliance model can provide a better focus on the storage class and tiering architecture. This will provide the starting kit for longer-term planning and growth management of the storage infrastructure.
- If a combination of in-memory, SSD and traditional storage architecture is planned for big data processing, the persistence and exchange of data across the different layers can be consuming both processing time and cycles. Care needs to be extended in this area, and the appliance architecture provides a reference for such complex storage requirements.

Operational costs

Calculating the operational cost for a data warehouse and its big data platform is a complex task that includes initial acquisition costs for infrastructure, plus labor costs for implementing the architecture, plus infrastructure and labor costs for ongoing maintenance, including external help commissioned from consultants and experts.

JD Software Pvt Ltd.,
23, Melandai 1st Cross Street,
West Tambaram,
Chennai 600045
India
Tel: +91 44 22265767
Mobile: +91 9840902654
www.jdsoft.in