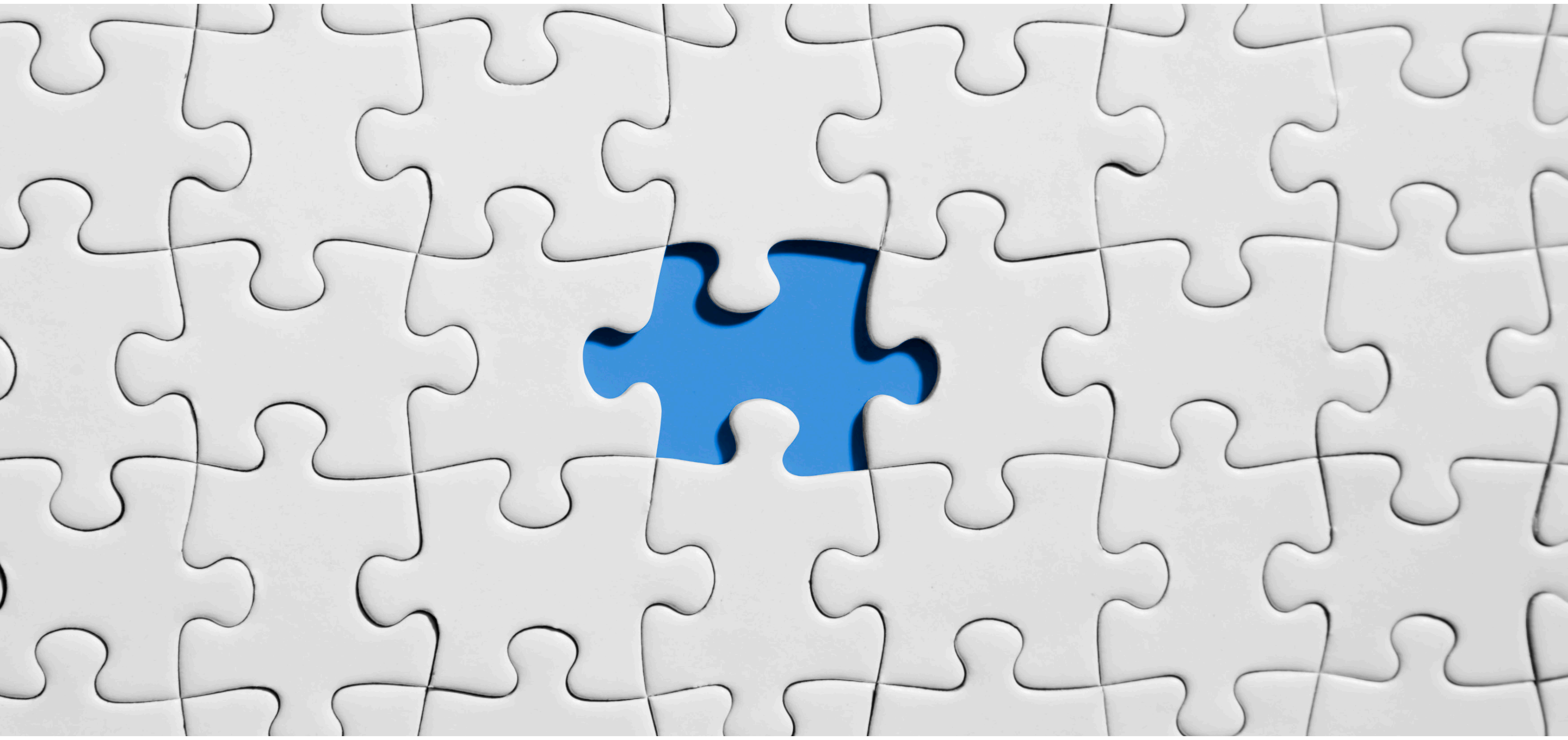
# IMPORTANCE OF DATA INTEGRATION IN THE DATA DECADE



Bhavana Sayiram

Customer eXperience Engineer | Central Corp NA

Dell Technologies

Bhavana_Sayiram@Dell.com

Knowledge Sharing Article

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

# Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the author. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

## Introduction

Data has become an intellectual part of our life. Data is usually huge & unstructured most of the time. According to the prediction made by International Data Corporation (IDC), by 2025, the global data growth will reach 163 zettabytes. Most of these data originated from social media, online marketing analytics, transactional data tracking, survey data etc. Data is the most valuable asset and using it in the right way is necessary to make intelligent business decisions, drive growth and improve profitability. Nevertheless, most of companies lack a centralized approach to data, with data siloes being one of the most common issues. the average business uses up to 13 different applications, therefore data is distributed amongst various systems which results in poor communication between departments, systems and processes. This leads in inability to meet both an organization's needs and business goals.

Data integration allows businesses to combine data existing in different sources to provide users with a real-time view of business performance. Data integration is the first step towards reforming the data into useful & meaningful information. It is essential to pursue Data Integration as a strategic function in order to support any businesses with advanced analytics processes or to create multi-dimensional views of customers.

How does data integration help an organization to meet their business goals?

- **Improves decision making capabilities** – Providing access to real-time data provided in simple format will help the organizations to be proactive, exploring opportunities and discovering potential bottlenecks before they happen.
- **Improves customer experience** - Siloed data sets wouldn't provide customers a complete view and thus impacting on sales which would affect revenue. Organizations can reach out to customers at right time with right message only when they have access to real-time customer data as well as historic data. That is how any organizations can improve customer experience, protect their business & increase revenue.
- **Simplifies operations** – Accessing real-time data is useful to improve processes, reducing costs and increasing production, across various business units
- **Increases productivity** - Productivity is significantly reduced if one needs to check multiple systems to gather required data. With automatic data integration, data from various sources is combined together into a single customer view & improves productivity.
- **Can help predict future** – Using historic data along with real-time data, one can use this information to forecast customer needs & requirement demands. This will help organizations to evaluate their products and services, while providing them with the ability to remain ahead in the market.

Most the organizations use a variety of systems and applications residing on various platforms such as cloud, social and mobile system and gathering data from all these sources is challenging. To overcome this difficulty, it is important to adopt a data integration strategy.

Before implementing a data integration strategy, first identify the importance of data to people and processes, to identify data silos among internal and external systems and to understand how the it is being processed and integrated.

Merging data from a legacy system into new system is a huge task as there are inconsistencies with formatting, duplicate data or naming conventions. So, it is important to establish a strategy which ensures reliable data is bought to new system; one such strategy is the Extraction, Transform, Load (ETL) process.

## ETL Process

A major role in data integration strategies, ETL enables businesses to gather data from various sources and combine it into a single, centralized location. ETL also makes it possible for different kinds of data to work together.

### Extraction

Since most business organizations manage data from a variety of sources and use a number of data analysis tools to produce business intelligence, the data must be able to travel freely between systems and applications they use. Before moving data to a new destination, it must first be extracted from its source. In this phase, structured and unstructured data is imported and consolidated into a single repository and raw data is extracted from different sources, i.e. cloud, existing databases, analytics tools, CRM systems, legacy systems, mobile devices and applications, etc. This process of extracting data can be done manually but it's time consuming and prone to errors. ETL tools help automate this process and create an efficient workflow.

The three kinds of extraction are Update notification, Incremental extraction and Full extraction

1. **Update notification** – The system notices when a record has been changed.
2. **Incremental extraction** – Deals with small changes in the data. The tool will be aware of its need to recognize new or changed information based on time and dates.
3. **Full extraction** – The data is extracted directly from the source system at once.

### Transform

Cleansing and consolidating of data prior to its analysis.

There are two approaches with data transformation

1. **Multi-stage data transformation** – Extracted data is moved to a staging area where transformations occur prior to loading the data into the warehouse.
2. **In-warehouse data transformation** – Data is extracted and loaded into the analytics warehouse and transformations are done there.

Recently, businesses tend to transform data within the warehouse rather than transforming it beforehand due to increased performance and scalability of the modern analytics database. Thus, it has become a default to use this approach in the transformation phase of ETL.

During this phase, rules are applied on data to ensure data quality, integrity and accessibility. The transformation phase consists of many sub processes which can be categorized as Basic Transformation and Advanced Transformation.

Basic transformations include Cleaning, Deduplication, Format revision, Key Structuring, etc.

Advanced transformations include Derivation, Filtering, Joining, Splitting, Data Validation, Summarization, Aggregation, Integration, etc.

### Load

This is the final phase of ETL process. The transformed data is loaded to a new destination. Transformed data can either be loaded all at once known or at a scheduled interval.

1. **Full load:** The entire transformed data is loaded to the destination which usually happens the first time a data is loaded into the warehouse.

2. **Incremental load:** Data is loaded at regular intervals. The last extract date is stored so that only records added after this date are loaded. Incremental loads are further categorized into two types based on the volume of data that's been loaded.

   - Streaming incremental load – For loading small data volumes
   - Batch incremental load – For loading large data volumes

## ETL Architecture

Before designing ETL architecture, consider these five points

**Understand the requirement of the organization**

Most used data source
How data is utilized and who will be using the data
How often the end users use the fresh data

**Segregate the data sources**

Production databases
Sales & marketing sources
Customer support sources
Operational sources

**Determine Data Extraction approach**

Sources of data extraction
Method to use for extracting data

**Build cleansing strategy**

Restructuring the data
Maintaining data type

**Manage ETL Process**

Scheduling ETL process
Monitoring the process
Recovery of ETL process in case of failure
ETL testing for data accuracy
Security measures to transfer data securely to a new destination

## ETL and Business Intelligence

Strategies used for data analytics and business units have greater access to vast number of data sources than ever before. ETL aids in transforming the huge quantity of data into business intelligence.

Consider the amount of data available in a business unit, the data collected from marketing, sales, logistics, and financial teams and the data generated by sensors used in the machineries of the business units in the facility. All of this data must be extracted, transformed, and loaded into a new location in order to perform any analysis over that. In such a scenario, ETL helps create business intelligence by:

- Delivering a unified view of the data which makes it easier to analyze, visualize, and make sense of large data sets.
- Providing historical context of the data along with the data collected from new sources, in turn providing a long-term view of both legacy and new data.
- Improving productivity and efficiency by automating the process of hand-coded data migration which allows developers and their team to spend more time on other innovative tasks.

## Modern Data Integration (MDI)

Business units face huge competition these days due to the advancement in technology. Innovations in technology such as Internet of Things (IoT) generate an enormous amount of data every day. Even with all its benefits, the ETL process becomes a bottleneck with massive data flowing in from multiple sources in multiple format. Principles of Modern Data Integration would resolve the bottlenecks created with ETL process.

Since much useful data has a short shelf life, modern data integration serves as a crucial means for businesses to overcome complexities that might occur in data analysis. All the business processes must be able to tap into such data at the earliest time possible. For that, organizations need to develop a data management strategy that focuses on generating data which is business ready. To do so, organizations need a new mindset where IT teams and other business units collaborate to improve the data strategies and all integration processes which meet dynamic and rapid business needs. In addition, modern data integration must ensure end user participation in numerous data management activities.

### Drivers of MDI

Complexities that come along with data hits all kinds of organizations and to solve them, the right technologies must be used. It's important for organizations to consider the impact on business results when deciding on a technology purchase.

Increasingly, organizations are proactively participating in making sure that the right technologies are implemented. Business units are recognizing the strategic value of data and understanding the need to embrace technologies that provide the most valuable information with continuous improvement and change.

Information and intelligence handled in real-time are the key for effective business processes throughout the organization. The need for the most up to date information is the primary factor driving MDI.

### Characteristics of MDI

Characteristics of MDI align with both technical and business needs of an organization.

- **Mandate to keep up with everything related to data –** MDI technologies, processes and users must be able to handle data twists and turns all the time.
- **Hybrid solution to address data integration flow –** Hybrid solution refers to the hybrid infrastructures of the organization (on-premises and cloud). This is the next step to remove silos among data, business processes and end users.
- **Familiar platform –** Flexible, adaptable, elastic and responsive platform with universal interoperability and connectivity to support any kind of data integration process
- **Support and access for all users –** Efforts to create business user access can be applied to simplify capabilities for technical users.
- **Reusable service-orientation –** Reusable services speed up the creation of more complex data integration processes and help IT to keep up with business demands.
- **Centralized management**

## Conclusion

Sophisticated, wide-reaching technologies matter greatly as the basis of a modern data integration platform. But the technology alone does not engender modern data integration. It's a mindset of business and technology teams working together to use data to improve business results and competitive edge. Ultimately, true differentiation centers on the value and benefits delivered to businesses and all users, by taking advantage of modern data integration. Metrics can be put into play to connect modern data integration capabilities and processes to business outcomes; from such metrics certain kinds of business value can be determined.

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” DELL TECHNOLOGIES MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell Technologies software described in this publication requires an applicable software license.