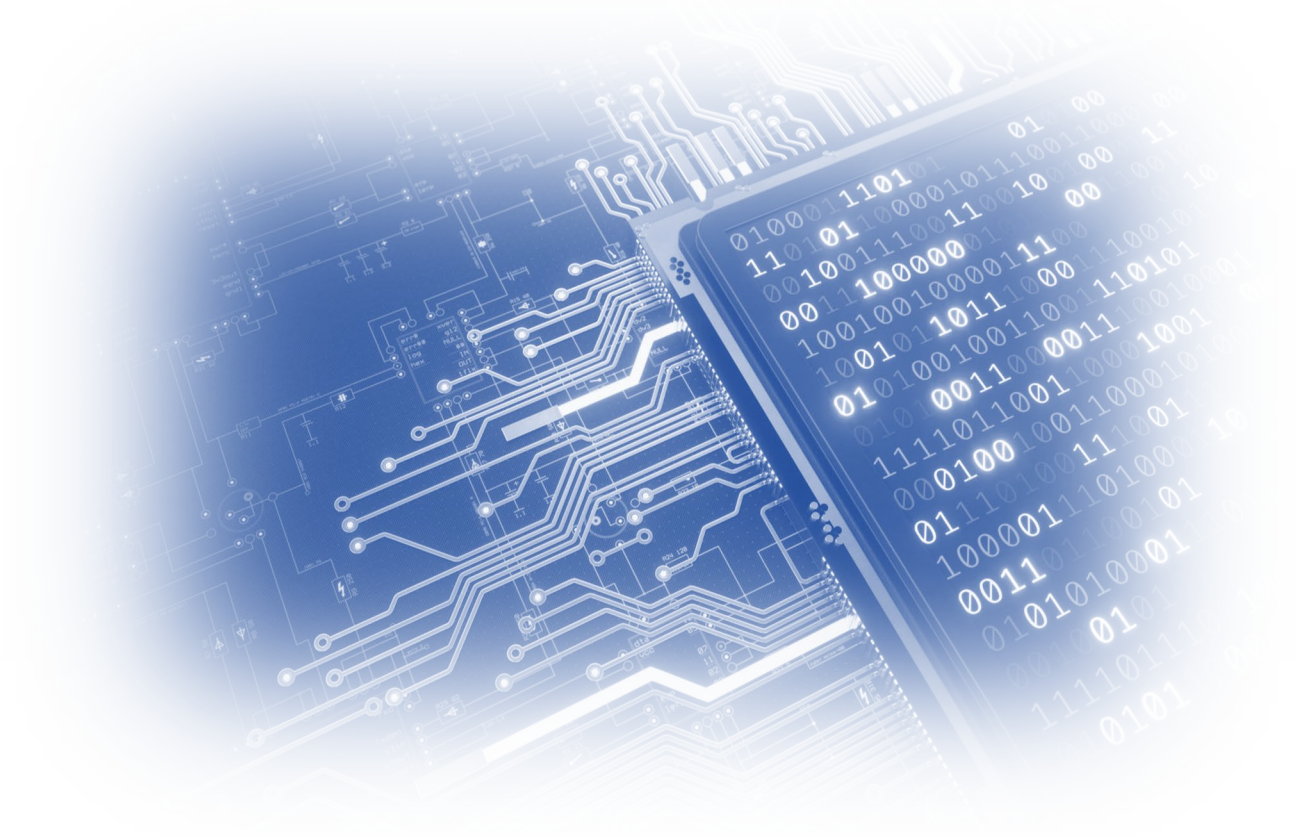
White Paper

Data Engineering a successful Data Migration project with ETL



Thiruvarasamurthy G

Software Architect

Table of Contents

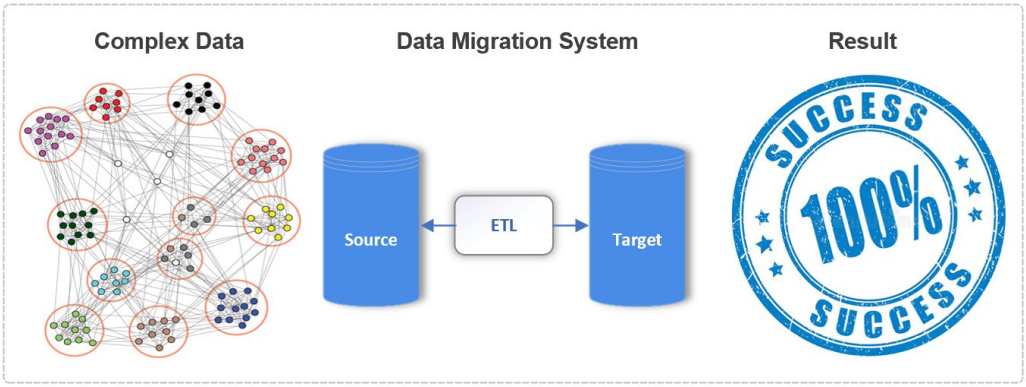
About Author

Thiruvarasamurthy G is a Software Architect, Full Stack Software Developer and Microsoft Certified Software Professional. He is passionate about technologies such as Cloud Computing, Software Design & Development, DevOps, Databases, BI Datawarehouse, Data Migrations, Application Securities & Ethical Hacking. He has experience with various MNC IT firms and runs a personal technical web blog www.thirufactory.com to share tech content.

Introduction

I worked as an architect on a significant data migration project for a corporate organization. The project involved handling confidential, high-volume, and incompatible data. With meticulous design and architecture utilizing ETL techniques, I completed the project with 100% success and without encountering any bugs. The exceptional quality of the deliverables left the business stakeholders delighted.

In this white paper, I will share my first-hand experience, providing an overview of data migration concepts as well as discussing the challenges faced during the project and the technical architecture employed. Let's start!



Overview of Data Migration

Data migration is a complex process that requires careful planning, attention to detail, and a thorough understanding of the data and systems involved. below are the short glances of aspects,

Definition

Data migration is the process of transferring data from one system or storage platform to another. This process can be complex and challenging, especially when dealing with large amounts of data. Data migration is often required when businesses need to upgrade their systems, consolidate data, or move to a new platform.

Key Concepts

	Key Concept	Description
1	Data Mapping	Identifying and mapping the data fields between source and target system.
2	Data Cleansing	Identifying and correcting any inconsistencies in the data.
3	Data Validation	Ensure the data migrated to the target system is accurate.
4	Data Testing	Test the migrated data in the target system to ensure that it works as expected.

Challenges

	Challenge	Description
1	Compatibility	Ensure that the data is compatible with the target system. This can involve converting data formats, dealing with differences in field sizes.
2	Data Security	It is essential to ensure that sensitive data is protected during the migration process, and prevent unauthorized access.
3	Data Loss	This can occur due to hardware or software failures, network issues, or other problems and ensure that data is not lost during the migration process.

Core Concepts

	Core	Description
1	Plan Ahead	Identifying the scope of the migration, detailed plan, and allocating resources.
2	Risk Assessment	Conduct a risk assessment to identify potential risks and develop contingency plans to address them.
3	Architecture and design	It is essential and important for a delivery.

Project Analysis

I initiated the project by conducting a comprehensive analysis that encompassed the data, source system, target system, and associated applications. The results of the analysis clearly highlighted the presence of significant risks in critical areas. Below, I have summarized requirement and some of these risks.

Requirements

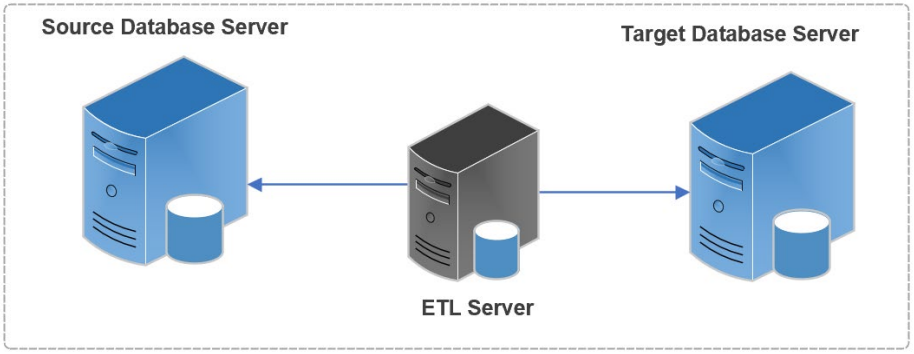
- The source and target systems are live and actively used by their respective applications and users. The data within these systems is constantly being updated and growing with fresh information on a daily basis.
- Retire the old systems and transition the users to the target system post data migration.
- The business owners expect the complete data to be included in the migration and also wish to rectify any erroneous data using new rules. Additionally, they desire to keep the source system unaltered.

Risks

- In the event of a data migration failure, regardless of the cause
    - The request for business downtime will be treated as a failure in the data migration process.
    - In the event of data changes that require a rollback, appropriate measures will be taken to revert the changes and hand over the system to the users.
    - The data migration process will be retried from the beginning.
- The source and target database systems have distinct and incompatible structures. The source schema follows an entity relationship-less type, while the target schema adheres to a snowflake type.
- Approximately 80% of the data from the source system is incompatible with the target system.
- The entire dataset needs to be migrated to the target system, necessitating data cleansing and correction at multiple levels, identifying bad data manually can be a perplexing process.
- Performing manual data validation and testing on a high-volume dataset can be a laborious task. Furthermore, the data migration process will introduce new rules for data cleansing.

Architecture and Technical Design

It is a critical phase of the project, and after conducting a thorough risk assessment, the decision was made to utilize ETL (Extract, Transform, Load) concepts for the project. SSIS technology has been finalized as the tool of choice. To achieve optimal outcomes, my focus was on creating a robust architecture and design. Now, let's proceed with a high-level overview of the approach.



Scope

- A single ETL solution is created for **data cleansing** and **data migration**.
- The both systems are live environments, each with their own dependent applications and users.
- In the event of a successful data migration, the source system is scheduled for complete decommissioning.
- A dedicated server has been procured exclusively for hosting SSIS and facilitating the data migration process.

Modular Design

- Modularization is the process of abstracting ETL processes into multiple reusable blocks.
- It reduces duplication work, makes testing activities easier.
- It establishes a standard that every process must follow and one of the ETL best practices.

OVAL Principle

- **Operations**: ETL steps and logics.
- **Volume**: Total volume of the data in production.
- **Applications**: Best suited for the task such as SQL, SSIS or Custom.
- **Location**: Hosting location of the ETL.

Data Load Design

- The source and target system are MS-SQL Server
- Data loading mechanism is below,
    - **Extract** data only from source system into ETL Server.
    - **Transform** the data in in-memory of ETL Server (no disk-based operation)
    - **Load** the data into Target system.

General

- **Restartability:** If any failures during the execution, it should restart from the point of failure point.
- **Logger:** ETL logs detailed information of each steps executions.
- **Transaction:** Implementation of native SQL transaction on the target system to sort out fail overs.
- **Performance:** ETL designed with efficient memory buffer utilization and parallel process.
- **Security:** No third-party and no inbound and outbound network connection except source and target systems.
- **Minimalistic Resource:** Architecture does not use the staging database (common in ETL design), DDL and DELETE database operations, 'Temp' tables, SYSTEM databases.

SSIS

SSIS (SQL Server Integration Services) is a powerful ETL (Extract, Transform, and Load) tool provided by Microsoft as part of its SQL Server database. It is a platform for building enterprise-level data integration and data transformation solutions, which enables businesses to transfer and transform data between different data sources. It processes large amounts of data quickly and efficiently and can also be used to perform complex data transformations perfectly.

SSIS is one of the ideal options for my design approaches, fulfilling all the necessary requirements.

Data Cleansing and Correction

As mentioned previously, approximately 80% of the source data is incompatible with the target system. Identifying and resolving these problematic data instances can be a challenging task, particularly when multiple data hierarchies are involved. To tackle this issue, SSIS packages have been developed with the capability to identify problematic data and capture it in separate files. These files are then validated by business stakeholders, who provide feedback to incorporate new rules into the ETL transformation process.

The ETL execution is repeated iteratively until all the data is rectified, aiming to achieve a state where 100% of the data meets the required standards.

Data Validation and Testing

Data validation and testing is intricate, particularly when dealing with substantial amounts of data. To facilitate this process, SQL scripts have been developed with the ability to validate both the source and target data. These scripts are incorporated into the SSIS package, allowing for seamless validation within the data migration pipeline.

Additionally, testing is conducted using the dependent applications of both the source and target systems. This comprehensive approach ensures that the data is validated from multiple perspectives.

Hardware and Deployment

Hardware: The hardware design of the ETL Server is relatively straightforward due to the architecture's focus on performing all operations as in-memory processes, minimizing disk operations. The server is equipped with 16 GB of

RAM and an Intel Xeon processor, ensuring sufficient resources for efficient execution. It is important to note that the performance of the SSIS packages is directly influenced by the available RAM size.

Deployment: A GUI based utility app has been created using .NET technology this application serves as a user-friendly interface primarily utilized by deployment members, allowing them to easily configure and manage the execution of the SSIS packages.

Result

On the final day of the data migration project, everything was in place and ready for execution. The ETL process was initiated, and thorough monitoring ensured that the execution progressed smoothly. The migration process, which involved millions of data rows, was completed within approximately 1.5 hours without encountering any issues or obstacles. The business stakeholders acknowledged the successful data migration.

The data migration project was executed flawlessly, without a single bug. As a significant achievement, the source legacy system was decommissioned successfully, marking a milestone, Great Time!

Conclusion

The Data Migration project and the ETL Architecture & Design hold a special place for me due to the numerous significant challenges they presented. I firmly believe that the success and timely delivery of such projects heavily rely on having the right architecture and design in place.

I would like to thank to my family, friends, supports for their solid encourages. Feel free to leave your comments or clarifications. If any correction is needed, I will do update on the subsequent version of white paper releases.

White Paper Details

Title: Data Engineering a successful Data Migration project with ETL
Initial Release Date: 29-May-2023
Latest Release Date: 29-May-2023
Latest Version: 1.0

References

The design and technical solution have been crafted by me.
Microsoft SSIS/ETL documentation.

Contact

Please do contact me at thirufactory@gmail.com or www.thirufactory.com/p/contact.html for any clarifications or feedbacks.

Disclaimer and Copyrights

The content provided in this white paper is intended solely for general information purposes, and is provided with the understanding that the authors and publishers are not herein engaged in rendering engineering or other professional advice or services. The practice of Engineering is driven by site-specific circumstances unique to each project. Consequently, any use of this information should be done only in consultation with a qualified and licensed professional who can take into account all relevant factors and desired outcomes. The information in these white papers was posted with reasonable care and attention. However, it is possible that some information in these white papers is incomplete, incorrect, or inapplicable to particular circumstances or conditions. We do not accept liability for direct or indirect losses resulting from using, relying or acting upon information in these white papers.

This paper does not use any sensitive or confidential information and only focus on the problem and overview of the technical solutions.