International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V71I1P106

Volume 71 Issue 1, 40-44, January 2023

Original Article

ETL (Extract, Transform, Load) Best Practices

Dhamotharan Seenivasan

Project Lead-Systems, Mphasis, Texas, USA

Received: 08 December 2022 Revised: 11 January 2023 Accepted: 21 January 2023 Published: 31 January 2023

Abstract *- This article provides an overview of the key principles and techniques for effectively extracting, transforming, and loading data from various sources into a target system. It covers topics such as data quality checks, testing, performance optimization, and security. The article aims to provide readers with a comprehensive understanding of the best practices for ETL to improve the efficiency, accuracy, and reliability of their data pipeline and provides actionable advice for implementing ETL processes successfully.*

Keywords *- Data warehouse, ETL jobs, Extract Transform and Load (ETL), ETL performance, ETL optimization.*

1. Introduction

ETL jobs, or Extract, Transform, and Load jobs, are critical components of data processing. These jobs are responsible for extracting data from various sources, transforming the data to make it consistent and usable, and then loading the data into a destination system, like a data warehouse or a data lake.

The importance of ETL jobs lies in their ability to enable organizations to make sense of their data by bringing it together from different sources and preparing it for analysis. Without ETL, data would remain in silos, scattered across different systems and in different formats, making it difficult to gain insights and make data-driven decisions. ETL jobs allow organizations to consolidate, clean, and transform their data, making it accessible and useful for reporting and analysis.

2. Literature Review

The Extract, Transform, Load (ETL) process is a critical component of data management, enabling data integration from various sources for use in business intelligence and analytics. As such, it is essential to implement best practices to ensure that the ETL process is efficient, accurate, and secure.

Many studies have been conducted on best practices for ETL. A study by Azeroual et al. (2019) found that data quality control, data validation, and data cleaning are essential for the success of the ETL process. They also emphasized the importance of error handling and logging to ensure data integrity.

A study by Kimball et al. (2008) found that data mapping and normalization are important for ensuring the consistency and completeness of data. They also highlighted the importance of handling null and missing data to prevent errors in the ETL process.

A study by Inmon et al. (2011) found that data warehousing and integration are important for efficiently loading data. They also emphasized the importance of strategies for handling large volumes of data to ensure the performance and scalability of the ETL process.

A study by Mrunalini et al. (2009) found that data encryption and access control are essential for ensuring data security in the ETL process. They also highlighted the importance of compliance with data privacy regulations such as GDPR (General Data Protection Regulation).

Lastly, monitoring and maintenance have been found to be important for ensuring the long-term success of the ETL process. A literature review by Kimbal et al. (2008) found that regular monitoring and maintenance of ETL systems can prevent errors and improve performance. They also emphasized the importance of keeping the ETL system updated with the latest technologies and best practices.

ETL best practices are essential for ensuring the efficiency, accuracy, and security of the ETL process. These include data quality control, data transformation techniques, data loading strategies, error handling and logging, performance optimization, security and data privacy, and monitoring and maintenance. It is important to keep track of ETL's latest developments and trends and adapt the best practices accordingly.

1. 3. Data Quality

3.1. Data Validation

Data validation is the process of verifying that the data conforms to a set of rules or constraints, such as data type, format, and range. This step helps identify and reject invalid data before loading it into the destination system. By validating the data, it can ensure that the data is in the correct format, is complete and accurate and is ready for a transformation.

3.2. Data Cleansing

Data Cleansing is the process of identifying and correcting inconsistencies and errors in the data, such as duplicate records, missing values, and outliers. This step helps ensure that the data is accurate and consistent and meets the needs of the destination system. This also includes removing duplicates and missing data and also handling NULL/Not-NULL values.

3.3. Exception and Error Handling

This refers to the process of detecting and handling errors that occur during the ETL process. This can include logging and reporting errors and automating error-handling procedures, such as retrying failed tasks or redirecting invalid data to a separate error dataset.

3.4. Auditing

Auditing is the process of tracking and documenting data changes. Auditing is useful for identifying where errors occurred and for compliance and governance purposes.

4. Job Design

4.1. Incremental Loads

Instead of loading all the data every time, incremental loads only load the new or changed data. This can improve performance and reduce the time it takes to load the data.

4.2. Storing Historical Data

Keeping a historical data record can be useful for auditing, compliance, and change tracking. It also allows you to go back to previous versions of the data in case you need to rerun a calculation or analysis.

4.3. Partitioning

Partitioning the data can improve the performance of ETL jobs by breaking up large datasets into smaller, more manageable pieces. This makes it faster to search and query the data, which can improve the overall performance of the ETL job.

4.4. Indexing

Creating indexes on the data can also improve the performance of ETL jobs by making it faster to search and query the data. An index creates a separate data structure that stores a mapping of the data, making it faster to find specific rows or values within the dataset.

4.5. Caching

Caching intermediate results can also speed up ETL jobs with multiple steps that must be repeated. Caching the results of a previous step means that the ETL job doesn't have to repeat the same calculations or transformations again.

4.6. Reusability

Reusing common functions, scripts, and objects can make ETL jobs more maintainable and reduce development time.

4.7. Automation

Automating ETL jobs can save time and reduce errors. Scheduling jobs to run at specific times can help to ensure that the data is up-to-date and that the load on the system is evenly distributed.

4.8. Parallel Processing

ETL jobs can be parallelized to run on multiple cores or multiple machines, which can greatly improve the performance of the job. Parallel processing allows ETL jobs to take advantage of all available resources and complete the job faster.

4.9. Materialized Views

Materialized views can be used to pre-aggregate data, store summary data, or pre-join tables. This allows to optimize of the performance when querying the data.

4.10. Optimizing Query Performance

By adding proper indexes, rewriting queries, tuning parameters, and configuration, you can improve the performance of the database and the ETL jobs that rely on it.

4.11. Resource Management

Have the right hardware resources and properly configure them to your running load. It can help to ensure the ETL job runs smoothly and efficiently.

Using these strategies, it is possible to optimize the performance of ETL jobs and improve their efficiency, leading to faster processing times and more accurate data.

5. Error Handling

Several techniques can be used to detect and handle errors in ETL jobs:

5.1. Error Detection

Error detection can be done by using data validation rules, data profiling, and other techniques that are used to identify inconsistencies and errors in the data.

5.2. Error Logging

Logging errors can be made to provide detailed information about what went wrong and where in the ETL process the error occurred. This can include logging error messages, stack traces, and other relevant information to help diagnose and fix the problem.

5.3. Error Notification

Error notifications can be set up to alert administrators or other stakeholders when an error occurs. This can include sending email or SMS notifications or posting messages to a chat service or other communication tool.

5.4. Error Handling

Error handling refers to the process of dealing with errors that occur during the ETL process. This can include retrying failed tasks, redirecting invalid data to a separate error dataset, or rolling back changes to the data.

5.5. Error Recovery

Error recovery refers to the process of returning to a stable state after an error occurs. This can include restoring data from a backup, recreating an index, or reloading data from the source.

5.6. Testing

Thoroughly testing the ETL job can help identify any issues early on. This can include testing the data flow, testing the data validation, and testing error scenarios.

5.7. Root-Cause Analysis

When an error occurs, it is important to investigate and understand the root cause of the problem so that it can be fixed and prevented from happening again.

6. Security

Securing data during ETL processing is crucial to protect sensitive information and maintain the data's confidentiality, integrity, and availability.

6.1. Confidentiality

During the ETL process, data is often extracted from multiple sources and consolidated into a central location, such as a data warehouse or a data lake. This increases the risk of unauthorized access to sensitive information. It is important to secure data access and ensure that only authorized users can access the data.

6.2. Integrity

Data integrity refers to the accuracy and completeness of the data. During the ETL process, data can be transformed and consolidated. It is important to ensure that the data is not tampered with or modified by unauthorized users during this process.

6.3. Availability

Ensuring data availability is important to ensure that the data is accessible to authorized users when needed. Protecting the ETL systems is important to protect against disasters like power outages or network failures.

6.4. Role-Based Access Control

Assigning roles and permissions to users and processes can help to limit access to sensitive data and ensure that authorized individuals only access it.

6.5. Auditing and Monitoring

Auditing and monitoring the ETL process can help to detect and respond to security incidents. This includes tracking and logging user access, data transformation and data movement.

6.6. Compliance

Many industries have regulations and compliance requirements that must be met regarding data security. It is important to ensure that the ETL process complies with these regulations to avoid penalties and fines.

6.7. Authentication and Authorization

Authentication is the process of verifying a user's identity, and authorization is granting access to resources based on the authenticated user's privileges. This will include application ids and passwords.

6.8. Encryption

Encryption is the process of converting plaintext data into ciphertext to protect it from unauthorized access. This can include encrypting data at rest (when it is stored on disk) and in transit (when it is sent over a network). Various encryption methods and algorithms, such as AES, RSA, or Blowfish, can be used.

6.9. Data Masking

Data masking is the process of hiding sensitive data while still making it usable for testing and development. This can include replacing sensitive data with fictitious data or applying a mathematical algorithm to obscure the data.

6.10. Data Tokenization

It is a method of replacing sensitive data with a non-sensitive equivalent, referred to as a token. The token can be used in place of the original data for the purposes of testing, analysis, and compliance without compromising security.

2. 7. Monitoring and Maintenance

Several strategies can be used to monitor the status and performance of ETL jobs:

7.1. Dashboards

Dashboards can provide real-time visibility into the status and performance of ETL jobs. They can display information such as the number of records processed, the job duration, and the task status.

7.2. Job Scheduling

Monitoring job scheduling can help to ensure that ETL jobs are running as expected and on schedule. It also helps to identify delays or bottlenecks in the pipeline and take corrective action.

7.3. Metrics

Tracking and logging metrics such as memory usage, CPU usage, disk I/O, network usage and other performance-related parameters can help to identify bottlenecks and optimize the performance of the ETL job.

7.4. Logging

Tracking detailed logging can help to identify and diagnose problems that occurred during the ETL process. This can include tracking error messages, stack traces, and data samples.

7.5. Alerts

Setting up alerts for specific conditions, such as job failures or delays, can help to notify the appropriate stakeholders when a problem occurs quickly.

7.6. Error Reporting

Generating error reports that provide summary information about the errors that have occurred, including the number of errors, the types of errors, and the data that was affected can help to understand the overall error rates and trends

7.7. Documenting

Keeping detailed documentation on the ETL job's design, data flow, data transformation, and validation rules can help understand the job and how it operates. This will be helpful when it comes to maintenance, troubleshooting and updating the job.

7.8. Version Control

Using version control systems (such as Git) to store the ETL jobs and related scripts can help to track changes and roll back to previous versions if necessary.

7.9. Planning for Downtime

It is important to plan for downtime when updating the ETL job so that the data pipeline is not disrupted and the process can be completed in a timely manner.

8. Conclusion

The article has highlighted best practices for data integration, transformation, loading, security, ETL automation, monitoring and auditing.

Some of the key takeaways from the article include the importance of data mapping and validation during the integration process, the need for data quality control and governance during the transformation process, the importance of data partitioning and indexing for efficient data loading, the importance of scheduling, job management, and error handling for ETL automation. Organizations should thoroughly review their current ETL processes to identify areas for improvement and implement best practices where appropriate.

In summary, ETL jobs are an important part of data processing, and by following best practices and techniques, organizations can improve the performance, scalability, and reliability of their ETL jobs, leading to more accurate and reliable data.

References

[1] Website, 2022. [Online]. Available: https://www.precisely.com/blog/big-data/etl-best-practices
[2] Website, 2022. [Online]. Available: https://portable.io/learn/etl-best-practices
[3] Website, 2022. [Online]. Available: https://medium.com/@dhamotharranvs/improving-performance-of-the-etl-jobs-4e141e4b566e
[4] Website, 2022. [Online]. Available: https://www.tutorialspoint.com/etl_testing/etl_testing_best_practices.htm
[5] Website, 2022. [Online]. Available: https://www.datachannel.co/blogs/etl-best-practices
[6] Website, 2022. [Online]. Available: https://www.codemag.com/Article/1803051/Better-Extract-Transform-Load-ETL-Practices-in-Data-Warehousing-Part-2-of-2
[7] Website, 2022. [Online]. Available: https://blog.aspiresys.com/digital/big-data-analytics/etl-design-process-best-practices/
[8] Website, 2022. [Online]. Available: https://sushantjha8.medium.com/etl-best-practice-33e7e4e92a29
[9] Website, 2022. [Online]. Available: https://www.element61.be/en/competence/best-practice-etl-architecture
[10] Website, 2022. [Online]. Available: https://www.timmitchell.net/etl-best-practices/
[11] Website, 2022. [Online]. Available: https://hevodata.com/learn/etl-best-practices/
[12] Website, 2022. [Online]. Available: https://www.integrate.io/blog/best-practices-for-etl-architecture/
[13] Website, 2022. [Online]. Available: https://www.geeksforgeeks.org/etl-process-in-data-warehouse/
[14] Website, 2022. [Online]. Available: https://www.integrate.io/blog/etl-data-warehousing-explained-etl-tool-basics/
[15] Website, 2022. [Online]. Available: https://www.guru99.com/etl-extract-load-process.html
[16] Website, 2022. [Online]. Available: https://sushantjha8.medium.com/etl-best-practice-33e7e4e92a29 (FAKE, LOOK AT 8)
[17] Website, 2022. [Online]. Available: https://www.phdata.io/blog/best-practices-data-activation-reverse-etl-on-snowflake/
[18] Website, 2022. [Online]. Available: https://etl-tools.info/informatica/design-best-practices.html
[19] Website, 2022. [Online]. Available: https://www.integrate.ai/blog/5-best-practices-for-etl-pipelines
[20] Website, 2022. [Online]. Available: https://nix-united.com/blog/what-is-etl-process-overview-tools-and-best-practices/
[21] Website, 2022. [Online]. Available: https://www.cleo.com/blog/knowledge-base-etl-integration
[22] Website, 2022. [Online]. Available: https://www.ibm.com/docs/en/cognos-analytics/10.2.2?topic=preparation-etl-scenarios-best-practices
[23] Website, 2022. [Online]. Available: https://www.keboola.com/blog/etl-process-overview
[24] Website, 2022. [Online]. Available: https://flatlogic.com/blog/etl-extract-transform-load-best-practices-etl-process-and-lifehacks/

[25] Website, 2022. [Online]. Available: https://www.developer.com/database/best-practices-etl-development-for-data-warehouse-projects/

[26] Website, 2022. [Online]. Available: https://docs.oracle.com/cd/E35287_01/fusionapps.7964/e14849/daccustomizingobjects.htm

[27] Kimball Ralph, *The Data Warehouse Lifecycle Toolkit*, 3rd Edition, John Wiley & Sons, 2008.

[28] Mitesh Athwani, "A Novel Approach to Version XML Data Warehouse," *SSRG International Journal of Computer Science and Engineering*, vol. 8, no. 9, pp. 5-11, 2021. *Crossref*, https://doi.org/10.14445/23488387/IJCSE-V8I9P102

[29] W. H. Inmon, and Krish Krishnan, *Building the Unstructured Data Warehouse: Architecture, Analysis, and Design*, Technics Publications, 2011.

[30] Joseph George, and Dr. M.K Jeyakumar, "A Comparative Analysis of Data Integration and Business Intelligence Tools with an Emphasis on Healthcare Data," *International Journal of Engineering Trends and Technology*, vol. 68, no. 9, pp. 5-9, 2020. *Crossref*, https://doi.org/10.14445/22315381/IJETT-V68I9P202

[31] M. Mrunalini, T. V. S. Kumar, and K. R. Kanth, "Simulating Secure Data Extraction in Extraction Transformation Loading (ETL) Processes," *2009 Third UKSim European Symposium on Computer Modeling and Simulation*, pp. 142-147, 2009. *Crossref*, https://doi.org/10.1109/EMS.2009.111

[32] Azeroual Otmane, Gunter Saake, and Mohammad Abuosba, "ETL Best Practices for Data Quality Checks in RIS Databases," *Informatics*, vol. 6, no. 1, 2019. *Crossref*, https://doi.org/10.3390/informatics6010010