International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V71I3P105

Volume 71 Issue 3, 27-33, March 2023
© 2023 Seventh Sense Research Group®

Original Article

Improving the Performance of the ETL Jobs

Dhamotharan Seenivasan

Project Lead-Systems, Mphasis, Texas, USA

Corresponding Author : dhamotharranvs@gmail.com

Received: 30 January 2023 Revised: 09 March 2023 Accepted: 20 March 2023 Published: 31 March 2023

Abstract *- ETL (extract, transform, load) jobs are responsible for extracting data from a variety of sources, transforming it into a consistent format, and loading it into a target data store. The performance of ETL jobs can significantly impact the overall performance of an organization's data management system. Several factors can affect the performance of ETL jobs, including the volume of data being processed, the complexity of the transformation logic, and the efficiency of the extraction and load processes. In this article, we will discuss some techniques for improving the performance of ETL jobs.*

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

1. Introduction

ETL jobs play a vital role in data warehouses and data management systems. They are responsible for extracting data from multiple operational data source systems, cleaning, transforming as per business logic and loading it into a data warehouse. Data Warehouse will act as the source for reporting and data analytics environment.

In any organization, there will be a number of ETL jobs, and these jobs are scheduled to run at different times. Also, there will be a dependency between the ETL jobs. If the upstream job is poorly performing and taking longer than expected time to complete, it will impact all downstream jobs, which will cause data issues for that day. Increasing the performance of ETL jobs is crucial for improving the speed and accuracy of data processing. This article highlights simple steps that can help improve the performance of ETL jobs and make them more efficient.

2. Literature Review

The performance of ETL (extract, transform, load) jobs has been extensively studied by researchers and practitioners in the field of data management. A literature review of this topic reveals a number of key findings and best practices for improving the performance of ETL jobs, as presented by various authors in the field.

As noted by Kimball and Ross (2010), one common approach to improving ETL performance is the optimization of the data extraction process. This can be achieved through the use of optimized query structures, indexing strategies, and the minimization of data redundancies through data deduplication and aggregation.

The utilization of parallel processing and distributed computing systems have also been identified as a key strategy for improving ETL performance, as discussed by Hussain et al. (2013). Organizations can take advantage of modern computing systems' increased processing power and parallel processing capabilities by breaking down large data sets into smaller, manageable chunks.

In addition to these technical strategies, authors such as Gour V et al. (2012) have emphasized the importance of data warehousing and business intelligence tools in improving ETL performance. These tools can help organizations identify and address performance bottlenecks and provide valuable insights into the performance of their ETL jobs.

Effective data governance and management practices have also been identified as crucial for improving ETL performance, as noted by Korhonen et al. (2014). This includes implementing data quality controls, data lineage tracking, and metadata management.

Several case studies have demonstrated the effectiveness of these strategies in real-world environments. For example, a study conducted by Ranjan J (2009) found that implementing data warehousing and business intelligence tools, combined with optimized data extraction processes, resulted in significant performance improvements in the ETL jobs of a large financial services company.

The literature review highlights a number of strategies and best practices for improving the performance of ETL jobs, as presented by various authors in the field. These include the optimization of the data extraction process, the utilization of parallel processing and distributed computing systems, and the implementation of effective data governance and management practices. By taking a holistic approach to ETL performance improvement, organizations can achieve more efficient and effective data pipelines and improve their ability to manage and analyze large amounts of data.

3. Why ETL Performance is Important

ETL performance is important for several reasons. First, ETL jobs are typically used to load data into data warehouses or data marts. Data warehouses and data marts are used to support business intelligence activities such as reporting, analysis, and decision-making. If the ETL jobs that populate these systems run slowly, it can impact the timeliness and accuracy of the information available to decision-makers. Second, poor ETL performance can lead to data quality problems. For example, if an ETL job is extracting data from multiple sources and one of those sources changes frequently, the ETL job may not be able to keep up with the changes and load them into the target system accurately. This can result in incorrect or outdated information being stored in the target system. Finally, slow ETL performance can impact organizational productivity. For example, if an organization relies on information from a data warehouse or data mart for daily operations, and the ETL jobs that populate those systems are running slowly, it can lead to workflow disruptions and delays in completing tasks.

4. Optimize the Source Data

There are many factors that can impact the performance of an ETL job, but one of the most important is the optimization of the source data. Optimizing the source data ensures that the ETL process runs quickly and smoothly.

There are a few different ways to optimize the source data:

4.1. Remove Unnecessary Data

By identifying and removing data that is not needed for the ETL process, organizations can reduce the amount of data that needs to be processed and improve performance. Making sure that all the data is in a format that can be easily read and processed by the ETL job. This includes ensuring that all dates are in a consistent format, all text fields are properly formatted, and all numeric fields are in the proper range.

Remove any unnecessary data from the source files. This might include data that is not needed for the current ETL job or data that is no longer needed (such as old records that have been purged from the database). Identify and remove duplicate data. This can help reduce the amount of data that needs to be processed and improve performance.

4.2. Data Profiling

Data profiling is the process of analyzing data to understand its characteristics, structure, and content. It is an important step in identifying and removing unnecessary data and optimizing source data for better ETL performance. Data profiling tools such as Informatica data quality, Talend data quality, SAP data services, Trillium software, Microsoft SQL server data profiling tool and Data cleaner can be used to analyze data and provide information such as:

Identify the data types of the columns in the data set, such as text, integer, and date.

Identify any columns with null values, which may indicate missing or incomplete data.

Identify any data quality issues, such as invalid or duplicate data.

Analyze the distribution of data values in a column, such as the number of unique values or the most common values. Identify any relationships between columns in the data set, such as foreign keys or natural keys.

Analyze the amount of data in the data set, such as the number of rows or the size of the data set.

Data profiling can help organizations understand the characteristics of their data and identify any data that is not needed for the ETL process. This information can be used to optimize the data for ETL, improve data quality, and make data-driven decisions.

4.3. Data Archiving

Data archiving is the process of moving older data that is no longer needed for day-to-day operations to a separate storage location, where it can be preserved for future reference or compliance purposes. It is a strategy that organizations use to optimize source data for better ETL performance.

It is important to have a data archiving plan that includes:

Defining the criteria for archiving data.
Defining the archiving schedule
Defining the archiving process
Defining the archiving storage options
Defining data retention policies
Ensuring data is properly indexed and cataloged
Ensuring data is easily accessible when needed.

Data archiving is critical in managing and optimizing data for better ETL performance. It should be done regularly to ensure data quality, reduce cost, and ensure compliance.

4.4. Data Retention

Data retention is the practice of keeping data for a certain period of time, after which it is either deleted or archived. It is an important aspect of improving the performance of ETL jobs.

Many industries and government agencies have specific data retention requirements that organizations must comply with.

Data retention policies help organizations better manage the data they collect by identifying no longer needed data and either deleting or archiving it.

By regularly reviewing and deleting old data, organizations can reduce the risk of sensitive data falling into the wrong hands.

Data retention policies can help organizations improve their data quality by regularly reviewing and cleaning old data.

By regularly reviewing and deleting old data, organizations can improve the performance of their ETL jobs.

When creating a data retention policy, organizations should consider:

What data need to be retained and for how long
How will the data be saved and protected
How will the data be deleted or archived
How will the data be utilized and accessed
Compliance requirements

It is important to have a designated team to manage the data retention policies, regularly review and update them, and ensure that all stakeholders are aware of and adhere to them.

4.5. Data Quality

Data quality measures data's completeness, accuracy, consistency, and relevance. Poor data quality can lead to inaccurate business decisions, decreased productivity, and increased costs.

High-quality data ensures that decisions are based on accurate and relevant information.

High-quality data reduces the time and resources required to clean and process data, leading to increased productivity.

Poor data quality can lead to increased costs due to the need to clean and correct data.

High-quality data leads to better customer service and improved customer satisfaction.

It is important to review and monitor data quality regularly and to have a designated team to manage it. Data quality is a continuous process. It is important to regularly review and update data quality policies and procedures and ensure that all stakeholders are aware of and adhere to them.

5. Parallel Processing

Parallel processing is a method of executing multiple tasks simultaneously to increase a system's overall performance. By executing multiple tasks simultaneously, parallel processing can increase the overall throughput of the ETL process, allowing for more data to be processed in a shorter amount of time.

It allows for the addition of more resources (such as additional processors or machines) to the ETL process, improving scalability and allowing the process to handle increasing amounts of data. By executing multiple tasks simultaneously, parallel processing can reduce the overall latency of the ETL process, allowing for faster processing times. It can improve fault tolerance by allowing multiple tasks to be executed simultaneously, reducing the likelihood of a single point of failure. Parallel processing allows for the better use of resources by executing multiple tasks simultaneously instead of sequentially.

There are different ways to implement parallel processing in ETL jobs, such as:

5.1. Task Parallelism

Task parallelism is a technique used to divide a large ETL job into smaller, more manageable tasks, which can be executed in parallel. By executing tasks in parallel, the overall performance of the ETL job is improved, as the time required to complete the job is reduced. There are different ways to implement task parallelism in ETL jobs, such as:

5.1.1. Thread-based Parallelism

Tasks are executed in parallel using multiple threads within a single process.

5.1.2. Process-based Parallelism

Tasks are executed in parallel using multiple processes on a single machine.

5.1.3. Cluster-based Parallelism

Tasks are executed in parallel using multiple machines in a cluster.

Choosing the right parallelism strategy that best fits the data, the ETL process and the hardware resources is important. For example, thread-based parallelism would be a good choice if the data is small and the processing time is critical. If the data is big and the processing time is not critical, cluster-based parallelism would be a good choice.

5.2. Pipeline Parallelism

Pipeline parallelism is a technique that divides an ETL job into multiple stages, where each stage can be executed in parallel with the others. By executing stages in parallel, the overall performance of the ETL job is improved. There are different ways to implement pipeline parallelism in ETL jobs, such as:

5.2.1. Multi-threaded Pipeline

A pipeline is divided into multiple threads, where each thread executes a different pipeline stage.

5.2.2. Multi-process Pipeline

A pipeline is divided into multiple processes, where each process executes a different pipeline stage.

5.2.3 Multi-node Pipeline

A pipeline is divided into multiple nodes, where each node executes a different pipeline stage.

Choosing the right parallelism strategy that best fits the data, the ETL process and the hardware resources is important. For example, a multi-threaded pipeline would be a good choice if the data is small and the processing time is critical. A multi-node pipeline would be a good choice if the data is big and the processing time is not critical.

5.3. Cloud-based Parallel Processing

Utilizing cloud computing services, such as Amazon Elastic MapReduce or Google Cloud Dataflow, to run ETL jobs in parallel on a distributed cluster of machines. Cloud-based parallelism can be used to increase the performance of ETL jobs by leveraging the processing power and scalability of cloud-based resources. Cloud-based parallelism can be implemented using several different strategies, such as:

5.3.1. Cloud-based Data Partitioning

The data is partitioned into smaller chunks, and each chunk is processed in parallel on different cloud-based resources.

5.3.2. Cloud-based Distributed Data Processing

Data is processed in parallel on multiple cloud-based resources such as virtual machines or containers.

5.3.3. Cloud-based Serverless Computing

Serverless computing allows for the execution of ETL jobs in parallel without the need to manage the underlying infrastructure.

5.3.4. Cloud-based Multi-cloud Data Processing

Data is processed in parallel across multiple cloud platforms.

Many cloud providers offer a wide range of services and tools that can be easily integrated with ETL jobs, such as data storage, data processing, and machine learning services.

It is important to note that the cost should be considered when using cloud-based parallelism, and the best cloud provider should be chosen based on the requirements and the budget.

5.4. Data Partitioning

Data partitioning is a technique used to divide a large dataset into smaller, more manageable chunks, known as partitions.

There are several ways to partition data:

Data is partitioned based on a range of values, such as date or numerical values.

Data is partitioned by the hash function, which will map the data to a specific partition.

Data is partitioned based on a list of values, such as a list of countries or regions.

Data is partitioned into equal-sized partitions, with each partition receiving an equal number of rows.

Each partition is processed by a different processor, allowing for parallel data processing. This can significantly improve the performance of the ETL process, as it allows for faster data processing times, improved scalability, and reduced latency.

Choosing the right partitioning strategy that best fits the data and the ETL process is important. For example, if the data is time-series data, range partitioning would be a good choice. If the data is location-based, list partitioning would be a good choice. It is important to note that parallel processing can increase the complexity of the ETL process, and it is important to test and validate the performance improvements before applying them to production.

6. Caching

Caching is a technique used to store frequently accessed data in a temporary storage area, known as a cache, to improve the performance of data retrieval.

Caching can increase the performance of ETL jobs in the following ways:

The time required to retrieve the data is reduced, as the data can be accessed from the cache instead of having to be retrieved from the source.

The load on the source system is reduced, as the source system does not have to handle as many requests for the same data.

Caching allows for the efficient handling of a large number of requests for the same data, improving scalability and allowing the ETL process to handle increasing amounts of data.

Caching allows for storing a consistent version of the data, which can be used to ensure data consistency across multiple stages of the ETL process.

There are different ways to implement caching in ETL jobs, such as:

6.1. In-memory Caching

Data is stored in the memory of the ETL process, allowing for faster data retrieval times.

6.2. Disk-based Caching

Data is stored on disk, allowing for larger amounts of data to be cached.

6.3. Distributed Caching

Data is stored in a distributed cache, allowing for data to be cached across multiple machines.

Choosing the right caching strategy that best fits the data and the ETL process is important. For example, if the data is highly sensitive and the data processing time is not critical, disk-based caching would be a good choice; if the data processing time is critical in-memory caching would be a good choice.

7. Incremental Load

Incremental load is a technique that can improve ETL jobs' performance by only processing new or changed data rather than processing the entire dataset each time the ETL job is run. This can be achieved by using the following strategies:

7.1. Identifying New or Changed Data

This can be done using various methods like timestamps, version numbers, or change data capture (CDC) techniques.

7.2. Extracting Only New or Changed Data

Once new or changed data has been identified, it can be extracted and processed separately from the rest of the data.

7.3. Loading Only New or Changed Data

The extracted new or changed data can then be loaded into the target data store.

7.4. Updating the Metadata

The metadata should be updated to reflect that the data has been processed so the next time the ETL job runs, it can identify and process only new or changed data.

It is important to consider the specific requirements of the ETL job and the data sources and target data stores that are being used when implementing incremental load.

8. Monitoring

Monitoring is an essential aspect of ETL job performance management. It can help identify and resolve performance issues and ensure that the ETL jobs run efficiently and effectively. Here are a few ways that monitoring can improve the performance of ETL jobs:

8.1. Real-time Performance Monitoring

Monitoring the performance of the ETL jobs in real-time makes it possible to identify and resolve performance issues as they occur. This can help prevent performance bottlenecks and ensure that the ETL jobs run at optimal performance levels.

8.2. Root Cause Analysis

By monitoring the performance of the ETL jobs, it is possible to identify the root cause of performance issues. This can help to resolve performance issues more quickly and prevent them from recurring in the future.

8.3. Data Quality Monitoring

Monitoring the quality of the data being extracted, transformed, and loaded can help ensure that the ETL jobs produce accurate and reliable data. This can help improve the quality of the available data for analysis and reporting.

8.4. Job Scheduling and Execution Monitoring

By monitoring the scheduling and execution of the ETL jobs, it is possible to ensure that the jobs are running on schedule and do not conflict with other jobs or processes. This can help improve the ETL process's overall performance and ensure that the data is available in a timely manner.

8.5. Resource Utilization Monitoring

By monitoring the resources that are being used by the ETL jobs, it is possible to identify and resolve issues related to resource contention. This can help to improve the performance of the ETL jobs by ensuring that the jobs have the resources they need to run efficiently.

8.6. Alerts and Notifications

By setting up alerts and notifications, the ETL team can be informed in real time when a job is running too slow or when a job has failed, allowing them to resolve the issue and minimize downtime quickly.

Monitoring is a key element in ETL performance management, providing visibility into the ETL process, enabling the ETL team to quickly identify and resolve performance issues, improving the performance of the ETL jobs, and helping to ensure that the data is accurate and available in a timely manner.

9. Common Lookup Files

In advance, frequently referenced data can be unloaded and saved as a lookup file before the batch ETL process is started. This lookup file can be referenced in multiple ETL jobs without unloading each time from the database. Most of the dimension tables are unloaded and saved as lookup files.
Usually, the Lookup ETL process runs first, creating all lookup files necessary for the downstream jobs.

Lookup files can improve the performance of ETL jobs in several ways:

Reduces database load by caching data in memory
Increases query performance by narrowing down data sets
Enables reuse of pre-computed results
Avoid repetitive computations and I/O operations
Enables faster comparison of data by using indexing
Improves data validation through reference to a trusted source.

10. Disable & Enable Indexes

ETL tools will have components to enable and disable the indexes on the tables. To load data faster into the table, disable all indexes on the table, load the data, and finally enable the index.

Disabling and enabling indexes during ETL jobs can improve performance in the following ways:

Disabling indexes before bulk data loading can speed up the load process as it reduces overhead on the database during data insertion.

Enabling indexes after data load can improve query performance as indexes are used to locate data quickly.
Regularly rebuilding or reorganizing indexes can improve query performance by defragmenting and optimizing the index structure.

Dropping and recreating indexes can also improve query performance if the old index is fragmented or has become outdated.

11. Gather Statistics

The final step of the ETL job is gathering statistics on the target table. This is specifically important for tables that are reloaded from scratch. e.g., stage tables or temp tables. The outdated statistics degrade the query performance.

Gathering statistics after loading data in ETL jobs can improve performance in the following ways:

It helps the database optimizer make better decisions on query execution plans.

Increases accuracy of query optimizer's cost estimates, which can result in faster query execution.

Facilitates the database optimizer's ability to identify and make use of the most efficient access paths to the data.

Enables the database to determine the data distribution and use this information to optimize query processing.

Can identify and resolve skew data issues, which can slow down query performance.

It helps detect and eliminate suboptimal database indexes and join operations.

12. Schedule the Jobs After Peak Hours

Executing complex queries at peak times will lead to database server overload and restrict others from accessing the data.

Scheduling ETL jobs after peak hours can improve their performance in the following ways:

Reduces competition for resources such as CPU, memory, and I/O bandwidth with other processes running during peak hours.

Avoids slowdown or interruption due to higher load on the network, database, or storage systems during peak hours.
It helps ensure stable and consistent performance as there are fewer competing demands for resources.

Reduces the likelihood of data concurrency issues and conflicts, leading to more reliable and successful ETL job executions.

It can reduce overall processing time as ETL jobs can run more efficiently during off-peak hours.

13. Conclusion

The performance of ETL jobs can significantly impact the overall efficiency of an organization's data pipeline. Several strategies can be used to improve the performance of ETL jobs, including optimizing the data extraction process, reducing data redundancy, and utilizing parallel processing. Additionally, organizations can take advantage of tools and technologies, such as data warehouses and distributed computing systems, to manage and process large amounts of data more effectively.

It is important to remember that the specific techniques and tools used to improve ETL job performance will vary depending on the needs and resources of an organization. In some cases, a combination of strategies may be necessary to achieve the desired level of performance improvement.

Regardless of the approach taken, it is crucial to monitor the performance of ETL jobs regularly in order to identify any issues or areas for improvement. This can be accomplished through the use of performance metrics and monitoring tools, which can provide valuable insight into the performance of the data pipeline.

In conclusion, organizations can take several steps to improve the performance of their ETL jobs and achieve a more efficient and effective data pipeline. Whether through the optimization of data extraction processes, the reduction of data redundancy, or the utilization of advanced tools and technologies, organizations can improve their ability to manage and process the large volume of data in a timely and efficient manner.

References

[1] Ralph Kimball, and Margy Ross, *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence*, John Wiley & Sons, 2010. [Publisher Link]

[2] 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] [Google Scholar] [Publisher Link]

[3] Hameed Hussain et al., “A Survey on Resource Allocation in High Performance Distributed Computing Systems,” *Parallel Computing*, vol. 39, no. 11, pp. 709-736, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[4] Vishal Goar et al., “Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse,” *International Journal on Computer Science and Engineering*, vol. 2, no. 3, pp. 786-789, 2010. [Google Scholar] [Publisher Link]

[5] Vishal Goar et al., “Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse,” *International Journal on Computer Science and Engineering*, vol. 2, no. 3, pp. 786-789, 2010. [Google Scholar] [Publisher Link]

[6] Janne J. Korhonen et al., “Designing Data Governance Structure: An Organizational Perspective,” *GSTF Journal on Computing (JoC)*, vol. 2, no. 4, 2014. [Google Scholar] [Publisher Link]

[7] Baljit Singh, “Enterprise Reporting on SAP S/4HANA using Snowflake as Cloud Datawarehouse,” *International Journal of Computer Trends and Technology*, vol. 71, no. 1, pp. 28-39, 2023. [CrossRef] [Publisher Link]

[8] Jayanthi Ranjan, “Business Intelligence: Concepts, Components, Techniques and Benefits,” *Journal of Theoretical and Applied Information Technology*, vol. 9, no. 1, pp. 60-70, 2009. [Google Scholar] [Publisher Link]

[9] Vangipuram Radhakrishna, Vangipuram SravanKiran, and K. Ravikiran, “Automating ETL Process with Scripting Technology,” *Nirma University International Conference on Engineering*, pp. 1-4, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[10] Dhamotharan Seenivasan, “ETL (Extract, Transform, Load) Best Practices,” *International Journal of Computer Trends and Technology*, vol. 71, no. 1, pp. 40-44, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Kamal Kakish, and Theresa A. Kraft, “ETL Evolution for Real-Time Data Warehousing,” *In Proceedings of the Conference on Information Systems Applied Research*, vol. 2167, pp. 1508, 2012. [Google Scholar] [Publisher Link]

[12] Syed Muhammad Fawad Ali, and Robert Wrembel, “From Conceptual Design to Performance Optimization of ETL Workflows: Current State of Research and Open Problems,” *The VLDB Journal*, vol. 26, no. 6, pp. 777-780, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[13] [Online]. Available:https://www.integrate.io/blog/7-tips-improve-etl-performance/

[14] [Online]. Available: https://danischnider.wordpress.com/2017/07/23/10-tips-to-improve-etl-performance/

[15] [Online]. Available: https://medium.com/ziegert-group/etl-performance-improvement-c5a9bd65b6af

[16] [Online]. Available: https://blog.devart.com/how-to-optimize-sql-query.html

[17] [Online]. Available:http://www.ijmer.com/papers/(NCASG)%20-%202013/24.pdf

[18] [Online]. Available:https://dataintegrationinfo.com/improve-etl-performance/

[19] [Online]. Available: https://www.tridex.org/wp-content/uploads/Tridex-ETL.pdf

[20] [Online]. Available:https://www.researchgate.net/publication/341435560_Performance_Optimization_of_ETL_Process

[21] [Online]. Available:https://solutioncenter.apexsql.com/improve-the-performance-of-etl-process/

[22] [Online]. Available:https://elink.io/p/ways-to-optimize-the-performance-of-etl-process-9a0c9e9

[23] [Online]. Available:https://www.researchgate.net/publication/368300555_ETL_for_Data_Warehousing

[24] [Online]. Available:https://link.springer.com/article/10.1007/s00778-017-0477-2

[25] Dhamotharan Seenivasan, “Exploring Popular ETL Testing Techniques,” *International Journal of Computer Trends and Technology*, vol. 71, no. 2, pp. 32-39, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] [Online]. Available:https://www.disoln.org/search/label/Performance%20Tips

[27] [Online]. Available:https://www.timmitchell.net/etl-best-practices/

[28] [Online]. Available: https://medium.com/@data_analytics/etl-for-data-warehousing-1203dc346a4e