snowflake®

MOVING FROM ON-PREMISES ETL TO CLOUD-DRIVEN ELT

Best practices for maximizing the value and efficiency of your data pipelines

CHAMPION GUIDES

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Yesterday's data pipelines were designed to accommodate predictable, slow-moving, and easily categorized data from on-premises business applications. They rely on extract, transform, and load (ETL) processes to capture data from various sources, transform it into a useful format, and load it into a target destination such as a data warehouse. These legacy pipelines work fine for structured data sources from enterprise applications, but they are no longer adequate for the diversity of data types and ingestion styles that characterize the modern data landscape.

Today's modern pipelines are designed to extract and load the data first and then transform the data once it reaches its intended destination—a cycle known as ELT. Modern ELT systems move transformation workloads to the cloud, enabling much greater scalability and elasticity. In traditional on-premises environments, ETL jobs contend for resources with other workloads running on the same infrastructure. With ELT, data can be loaded in its raw form and then transformed in multiple ways once it is clear how the data will be used.

With an ELT pipeline, you can load many types of raw data into a cloud-based repository such as a cloud data platform. The platform improves the speed at which you can ingest, transform, and share data across your organization. This allows you to run resource-intensive transformation workloads to the cloud, where you can maximize the processing power and capacity of scalable cloud resources.

As you will see in the pages that follow, ELT is a good choice in the following situations:

- **When you have massive data requirements:** ELT can process large quantities of structured and unstructured data quickly in the cloud.
- **For analytics experimentation:** ELT maximizes options as analysts and data scientists explore the potential of their data, transforming it as needed for specific projects.
- **For low-latency data pipelines:** ELT transfers data immediately, which can be valuable for low-latency analytics and near real-time use cases.

Read on to learn how you can maximize the value of your data pipelines by using the right type of transformation method for each situation and workload.

UNDERSTANDING BASIC TERMS AND CONCEPTS

ETL is a software integration process that involves **extracting** data from various sources, **transforming** the data in a staging server, and **loading** the data into a target destination such as a data warehouse, data lake, or cloud data platform. In traditional data warehouses, the data is mapped to fit a relational data model. It may be cleansed, enriched, and transformed into a common format before being loaded into the target database.

Structuring and transforming the data ensures fast and efficient analysis with SQL-based business intelligence (BI) tools, but it limits how the data can be used, since some artifacts of the raw data are lost in the transformation process. In most ETL workflows, data is captured from the source databases and staged to a data warehouse. A staging server executes the transformation logic, which may include filtering, masking, enrichment, mapping, deduplication, and integration of data from multiple sources.

Data engineers create data pipelines to orchestrate the movement of batch data uploads as well as to stream data continuously. These pipelines extract data from applications, devices, and event streams. ETL pipelines transform data into a business-ready form as part of the basic ETL workflow. That's fine when the business requirements are clear. However, for some of today's popular workloads, such as machine learning and data science, the data format requirements aren't always known up front. For example, data scientists may prefer to maintain data in a raw (or less processed) state and then convert into various formats to accommodate different types of models, prediction engines, and analytics scenarios.

THE RISE OF ELT

Traditional ETL operations use a separate processing engine, often running on dedicated compute servers. The database is modeled to accommodate data in specific, predefined formats before that data is loaded, according to the downstream business requirements. For example, data may be sorted, summarized, or parameterized for rapid display through dashboards, or rolled up to populate monthly financial reports.

These ETL procedures may work well for structured data sources from enterprise applications, such as enterprise resource planning (ERP), supply chain management (SCM), and customer relationship management (CRM) systems. However, these legacy ETL pipelines can't easily accommodate newer data formats in massive volume, such as machine-generated data from Internet of Things (IoT) systems, streaming data from social media networks, weblog data from internet websites, and mobile usage data from SaaS apps. They do a good job ingesting structured data and batch data, but they are too rigid to collect and ingest schemaless and semi-structured data.

To accommodate large amounts of data in newer forms and enable more-timely analytics, modern data pipelines are designed to extract and load the data first, and then transform it once it reaches its destination. In the transformation stage, the data is standardized, cleansed, mapped, and combined with data from other sources. These newer ELT data pipelines leverage the power of cloud data warehouses and cloud data platforms that can cost-effectively store and process immense amounts of data.

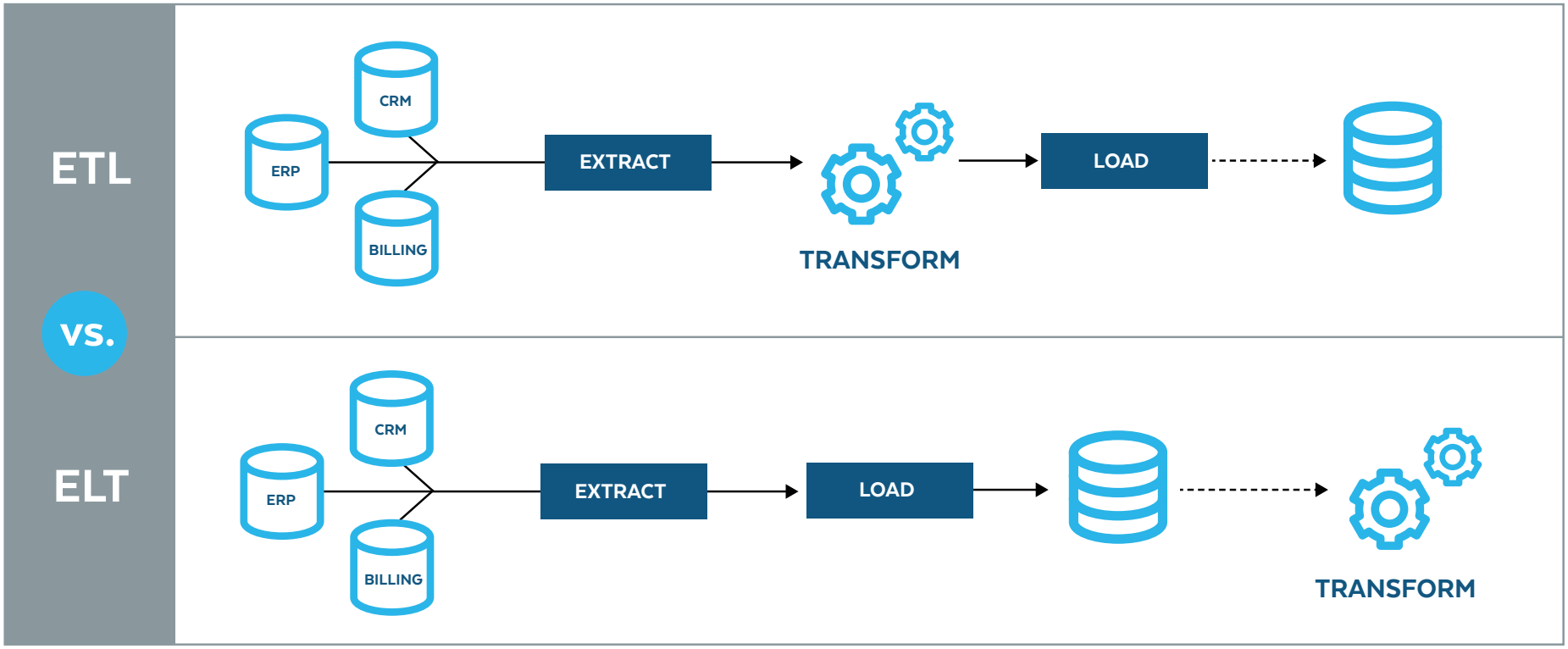
In addition to structured relational data, ELT pipelines can ingest unstructured, semi-structured, and raw data types and load it all into a cloud data platform or data lake. There is no need for data staging. The data can be maintained in its raw state to facilitate experimentation and rapid iteration.

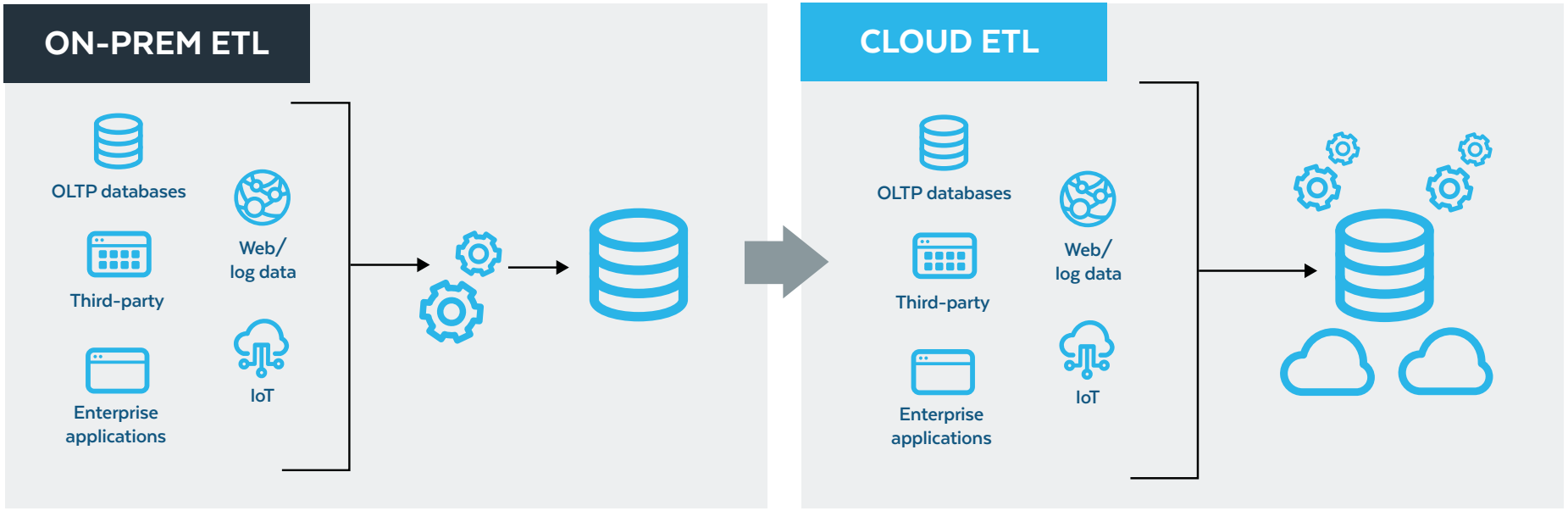
ADVANTAGES TO THE ELT APPROACH

Power:
Cloud-based repositories offer near-limitless storage capabilities backed by scalable compute servers, so you can keep up with a growing volume of data.

Scope:
ELT pipelines let you ingest all types of data as soon as the data becomes available, with no requirement to transform the data into a specific format.

Flexibility:
Transforms only the data that is required for moment-to-moment analytics requirements, allowing multiple teams to transform the data they need for reports, dashboards, data science models, and other tasks, thus maximizing options for putting the data to work.


ETL
CRM
ERP
BILLING
EXTRACT
TRANSFORM
LOAD
vs.
ELT
CRM
ERP
BILLING
EXTRACT
LOAD
TRANSFORM


ON-PREM ETL
OLTP databases
Web/
log data
Third-party
IoT
Enterprise
applications
CLOUD ETL
OLTP databases
Web/
log data
Third-party
IoT
Enterprise
applications

ESTABLISHING A VERSATILE DATA MANAGEMENT STRATEGY

ETL is a viable option when the data is predictable, manageable, and has a regular update interval. These batch ingestion processes are commonly used for application data that doesn't need to be constantly refreshed. For example, retail point-of-sale data may need to be updated in a data warehouse at the end of each day to accommodate daily revenue reports. Customer data in a CRM system may need to be uploaded to a call center dashboard once per hour to reflect current sales and service transactions. Electricity usage data collected from smart meters may need to be refreshed every 15 minutes to support time-of-use billing programs.

However, the situation can quickly get complicated. With ETL systems and legacy data architectures, data from discrete systems becomes siloed in many different places. For example, each type of data might land in a unique system that's designed and modeled for particular needs. This results in many different repositories, which can quickly turn into a maintenance nightmare. On premises or in the cloud, each production application creates its own data silo: marketing data in a marketing automation system, sales data in a CRM system, finance data in an ERP system, inventory data in a warehouse management system, and so on. Each of these apps may depend on specialized ETL tools and unique software procedures to collect data from the production systems and transform that data for analysis.

Given today's business and analytics needs, you should consolidate all your data in one location — as a single source of truth that has been architected for universal access by various workgroups, applications, and tools.

WHEN SHOULD YOU CONSIDER ELT?

ETL technology remains popular for situations in which data is moved from one system to another in batch mode. However, most legacy ETL solutions cannot handle all types of data. They work fine for structured data from enterprise applications, but are not ideal for machine-generated data from IoT systems, streaming data from social media feeds, JSON event data, and weblog data from internet and mobile apps.

To determine which approach to use, remember these basic guidelines:

- **ETL processes work well for relational data that must maintain a tabular structure.**
- **ELT is a better method for semi-structured data that must be maintained in its raw or native form until specific analytics use cases are devised.**

Other considerations include how much data you will be processing and how quickly it needs to be prepared for downstream analysis. Transformation processes require many compute cycles. With ELT, automatic scalability instantly provisions the necessary resources to support each operation. Using an ELT process allows you to use the boundless resources of the cloud to process and transform data quickly and efficiently. It also minimizes data movement, since you can process the data where it resides rather than moving it to an independent server or storage mechanism.

Determine where you will run your processing engine, what infrastructure resources are available, and what performance you will require. Do you have scalability or concurrency issues, such as limited server capacity? Whether data is produced by online transaction processing (OLTP) systems, website interactions, SaaS apps, equipment sensors, or social media streams, data engineers must develop data pipelines to capture that data, ingest it into a data repository, and make it accessible to the business community. In many cases, data pipeline operations are enhanced by leveraging the processing power of target databases in the cloud.

DATA PIPELINE CHOICES

Use ETL when:

- The total volume of data to be processed is relatively small
- The source and target databases require different data types
- You are primarily processing structured data

Use ELT when:

- You have a large volume of data to process
- The source and target databases are of the same type
- The data is semi-structured or unstructured

SNOWFLAKE CUSTOMER CASE STUDY

Organization: Paciolan is a leader in ticketing, fundraising, marketing, analytics, and technology solutions powering more than 500 live-entertainment organizations that sell over 120 million tickets per year.

Problem: To convert semi-structured data into relational data, Paciolan wrote proprietary ETL code that parsed and normalized the data. Fifty thousand records could turn into 1 million rows in an on-premises data warehouse. The ETL process, comprising close to 100 GB of data daily, took 30–60 minutes to complete. Limited resources prevented analysts from effectively summarizing and rolling up data.

Solution: Paciolan now stores its semi-structured JSON data as a VARIANT data type in the Snowflake platform. It uses Snowflake as both a data lake and data warehouse via Data Vault, an architectural approach that includes a specific data model design pattern to support a modern, agile enterprise data warehouse.

Results: The ETL process, which took as long as an hour to complete in the legacy data warehouse, now takes just a few minutes with the Snowflake data pipeline. Developers can use simple Python scripts to insert statements dynamically.

Benefits

- **Separation of storage and compute provides performance stability and cost visibility**
- **Instant elasticity enables nearly unlimited compute power for virtually any number of users**
- **Support for storing semi-structured data as variant data types enables richer data insights**

"We compared before-and-after numbers and found that with Snowflake there was a 90% reduction in the code used for the ETL process. That's a big win for us."

Ashkan Khoshcheshmi
Principal Software Engineer
Paciolan

PROCESSING DATA WITH SNOWFLAKE

The Snowflake platform includes flexible, scalable data pipeline capabilities as part of the basic service. You can ingest raw data directly into Snowflake, so you don't need to create a pipeline to transform data into a different format. Snowflake performs these transformations automatically, minimizing storage and compute costs.

Snowflake also simplifies data management by eliminating data silos: You don't have to maintain multiple copies of your data for multiple downstream applications. It maintains the original shape of the raw data but also transparently applies highly optimized storage techniques so that analytics and data transformations can perform exceptionally well.

Most importantly, Snowflake was designed to take full advantage of the cloud's unique attributes. It is based on a multi-cluster shared data architecture that separates compute and storage resources to accommodate data transformations at scale. Each type of resource can be scaled independently to accommodate the specific needs of each application.

Snowflake's platform is built around a robust processing engine that is designed to handle all types of workloads without degrading performance, such as ingesting data via a data engineering pipeline while simultaneously training a machine learning model to use that same data. Its scalable pipeline service can continuously ingest data without affecting the performance of these other workloads. Data engineers can decide how much computing power to allot to each data ingestion process or allow the system to scale automatically.

Snowflake also allows data engineers to build data pipelines with a wide choice of languages and integration tools for managing the ingestion stream. It can support a range of popular data ingestion styles, including batch integration and streaming integration with Apache Kafka. In addition, Snowflake allows you to easily and efficiently ingest many types of data using standard SQL, the lingua franca of data processing.

CONCLUSION

Steady growth in the volume, variety, and velocity of data necessitates new types of data pipelines—and more-advanced, cloud-based data processing engines to capture the data and put it to work.

ETL processes are often handled by on-premises servers with fixed capacity, limited bandwidth, and a finite set of CPU cycles. Modern data integration workloads are enhanced by leveraging the processing power of cloud databases and cloud data platforms that can be scaled at will.

To take advantage of these cloud resources, a growing number of organizations are designing data pipelines that extract and load data to a cloud database, and then transform it once it reaches this destination, a cycle known as ELT. This approach is quicker than traditional ETL processes, because it leverages the power of modern data processing engines and cuts down on unnecessary data movement.

ELT processes push resource-intensive transformation workloads to the cloud for two primary reasons:

1. **To use the near-limitless resources of the cloud to process and transform data quickly and efficiently**
2. **To keep the data in a raw state until the business requirements are fully understood**

Whenever possible, use ELT instead of ETL to push resource-intensive transformation jobs to a cloud-based destination platform. This approach will simplify your data pipelines, minimize data movement, cut down on the number of data silos, and maximize options for how the data is ultimately used.

To learn more about Snowflake's data pipeline solutions, visit snowflake.com/workloads/data-engineering.

ABOUT SNOWFLAKE

Snowflake delivers the Data Cloud—a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, Snowflake delivers a single and seamless experience across multiple public clouds. Snowflake's platform is the engine that powers and provides access to the Data Cloud, creating a solution for data warehousing, data lakes, data engineering, data science, data application development, and data sharing. Join Snowflake customers, partners, and data providers already taking their businesses to new frontiers in the Data Cloud. **Snowflake.com**.