snowflake®

A GUIDE TO MODERN DATA PIPELINES

CHAMPION
GUIDES

TABLE OF CONTENTS

INTRODUCTION

Only robust end-to-end data pipelines will properly equip organizations to source, collect, manage, analyze, and effectively use crucial data to generate new market opportunities and deliver cost-saving business processes.

Traditional data pipelines are rigid, brittle, and difficult to change, and they do not support the constantly evolving data needs of today's organizations.

They present many challenges, by:

- Taking significant time and cost to design and build
- Comprising multiple tools that are not compatible and require unnecessary integration
- Requiring that only qualified IT professionals, who have skills in short supply, build data pipelines, thereby creating work bottlenecks
- Introducing avoidable latency, causing delayed data extraction or transport through the pipeline
- Ignoring the demands of streaming data, handling batch-only data loading
- Being rigid, making it difficult to change and manage over time
- Relying on schema-dependent data loading processes

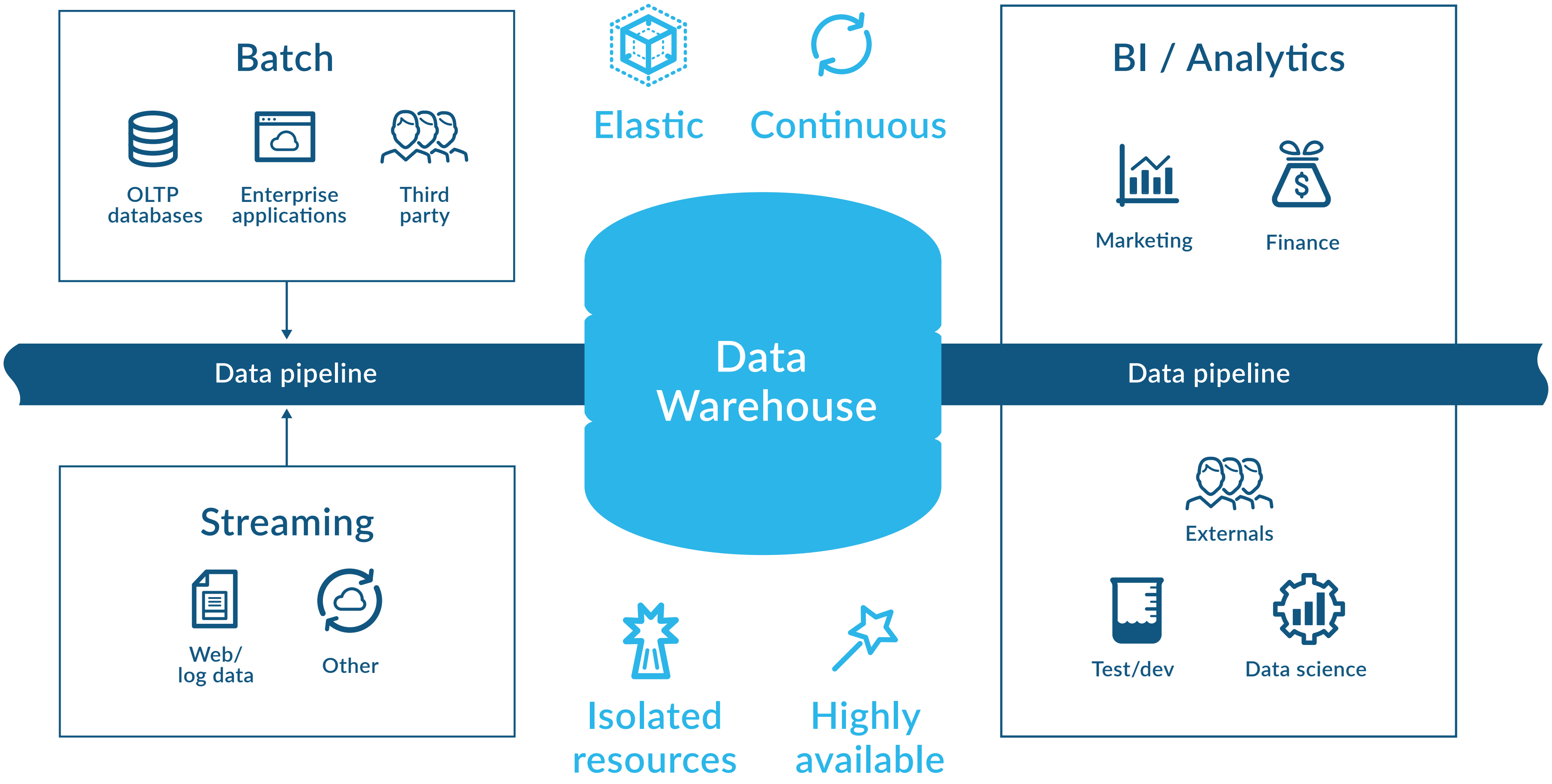
Modern data pipelines make it faster and easier to extract information from the data you collect. They start with extracting raw data from a number of different sources. The data is then collected and transported to a common place, typically a data repository in the cloud. From there, the data undergoes various transformations until it's usable for analytics and produces business value. The data is then loaded into a data warehouse, where it is easily managed and accessed by data science workloads, automated actions, and other such computing jobs.

To address these issues, the best data pipelines have these five characteristics:

- Continuous and extensible data processing
- The elasticity and agility of the cloud
- Isolated and independent resources for data processing
- Democratized data access and self-service management
- High availability and disaster recovery

These characteristics enable organizations to leverage their data quickly, accurately, and efficiently to make quicker and better business decisions.

Modern Data Pipeline Architecture



FIVE CHARACTERISTICS OF A MODERN DATA PIPELINE

CONTINUOUS AND EXTENSIBLE DATA PROCESSING

The scourge of stale data

Traditionally, organizations extract and ingest data in prescheduled batches, typically once every hour or every night. But these batch-oriented extract, transform, and load (ETL) operations result in data that is hours or days old, which substantially reduces the value of data analytics and creates missed opportunities. Marketing campaigns that rely on even day-old data could reduce their effectiveness. For example, an online retailer may fail to capture data that reveals the short-term buying spree of a certain type of product based on a celebrity discussing, using, or wearing the product.

Many traditional ETL processes need a dedicated service window. However, they often conflict with existing service windows or, worse, the windows are nonexistent. In addition, for global organizations active 24 hours a day, a nightly window required for batch data processing is no longer realistic. As data volume and complexity rise, it becomes difficult for ETL processes to finish within a dedicated service window, causing ongoing performance issues and delaying key insights.

Immediately actionable insights

On the flip side, modern data pipelines continuously and incrementally perform the processes involved in loading data into a data warehouse, transforming it into a usable form, and then analyzing it. The business can still use analytics tools such as Tableau, Looker, or Microsoft's PowerBI to run queries and reports, and the results will be current and immediately actionable.

Continuous processing decreases latency at every step along the way and enables users and systems to use data from a few minutes ago, instead of a day ago. With this kind of continuous processing, security applications can target and resolve external threats on a more timely basis, for example.

Modern data pipelines can also incorporate and leverage custom code that is written outside the platform. Using APIs and pipelining tools, they can stitch together a data flow using outside code seamlessly, avoiding complicated processes and maintaining the low latency.

In terms of decision-making, managers and analysts can make better business decisions based on more-current data. Successful data-driven decision-making relies on the relevance of the data being analyzed, and just a few hours of lag time can make a major difference for organizations that must operate in real time. For example, fleet management and logistics companies need to correct dangerous driving behaviors and diagnose hazardous vehicle conditions before they cause an accident or an expensive breakdown.

.

THE ELASTICITY AND AGILITY OF THE CLOUD

The cost of rigidity

Legacy data pipelines are run on premises using commodity hardware that is expensive to buy and manage. These solutions tend to be rigid and unable to scale easily. As a result, their creation and management involves significant upfront budgeting for peak-usage scenarios. When multiple workloads run concurrently, the competition for resources

increases and performance degrades, hamstringing the business with delayed insights and soaring costs during peak times.

Flexibility equals growth

Modern data pipelines offer the instant elasticity of the cloud and a significantly lower cost structure by automatically scaling back compute resources as necessary. They can provide immediate and agile provisioning when data sets and workloads grow. These pipelines can also simplify access to common shared data, and they enable businesses to quickly deploy their entire pipelines without the limits of a hardware setup. The ability to dedicate independent compute resources to ETL workloads enables them to handle complex transformations without impacting the performance of other workloads.

With an elastic and agile data pipeline, businesses can better handle spikes and growth. For example, a seasonal business that experiences a sales spike during the holiday season can add storage and processing capacity within seconds, not days or weeks. In addition, elastic pipelines are primed to handle compliance requests and audits more quickly. For example, the European Union's General Data Protection Regulation (GDPR), which stipulates how organizations can collect, store, and transmit the personal data of EU residents, might require a business to run reports that demonstrate compliance. The flexibility of the cloud enables businesses to perform analytics without impacting sales or customer service.

ISOLATED AND INDEPENDENT RESOURCES FOR DATA PROCESSING

Sluggish performance

A common challenge with traditional ETL processes occurs when workloads compete for resources. Running multiple workloads in parallel on the same set of resources impacts the performance and response time of each, increasing the time from data collection to insight. In addition, most platforms, whether in the cloud or on premises, use an older "shared nothing" architecture. This architecture tightly couples storage, compute, and database services. The tight coupling hampers the ability of the database administrator to elastically scale the database to store or analyze more data or to support more concurrent users.

Gaining speed and value

But imagine an architecture in which compute resources are separated into multiple independent clusters. In addition, the size and number of those clusters can grow and shrink instantly and nearly infinitely depending on the current load. All the while, each cluster has access to the same shared data set that they jointly process, transform, and analyze. Such an architecture has become crucial and cost-effective for today's organizations, thanks to cloud computing.

A modern data pipeline that features an elastic multi-cluster, shared data architecture makes it possible to allocate multiple and independent isolated clusters for processing, data loading, transformation, and analytics while sharing the same data concurrently without resource contention. Each cluster can read and write data with full transactional consistency, and its size and resources are based on the performance required for the workload at that time. In addition, users can load data while it's being transformed and analyzed in other clusters, without impacting performance. Each workload has its own dedicated resources. Modern data pipelines also rely on loosely coupled components that physically separate but logically integrate storage, compute, and services such as metadata and user management. Because each of the components is separate, they expand and contract independently of each other.

Such an architecture frees organizations from painful trade-offs, such as not being able to load and process additional data sets due to the rigid capacity limits of their traditional data pipelines or running the risk of violating SLAs when accommodating more use cases. Or, they had to run the risk of violating the business SLAs when accommodating more use cases. An elastic, multi-cluster, shared data architecture also makes processing times predictable because occasional spikes in data volume or load can be covered by instantly and elastically adding more resources.

DEMOCRATIZED DATA AND SELF-SERVICE MANAGEMENT

The inefficiency of ETL

With traditional solutions, the only way for multiple business applications to pull from centralized data is to invest in tools that extract data from data marts, transform it into the proper format for querying, and then load it into individual databases. ETL processes typically require a large set of external tools for extraction and ingestion. It often takes months for a team of experienced data engineers to set up such a process and integrate the tools, which creates bottlenecks. In addition, it requires yet more time to set up the process required for ongoing maintenance.

Organizations often have to discontinue important analytics projects because they don't have in-house expertise to create the data pipelines or the data is stale by the time the pipelines run. Much time is also lost conceptualizing how the data pipelines should look. In addition, the pipelines are unable to handle and process all types of data, whether structured, semi-structured, or unstructured.

Increased access, better insights

Modern data pipelines democratize data by increasing users' access to data and making it easier to conceptualize, create, and maintain data pipelines. They also provide the ability to manage all types of data, including semi-structured and unstructured data. With true elasticity and workload isolation, and advanced tools such as zero-copy cloning, users can more easily massage data to meet their needs. Businesses can use simple tools, such as SQL, to implement parts of a pipeline. This makes the creation, management, and monitoring of the data processes largely self-service, helping businesses to directly investigate the data and discover insights, decreasing decision-making time, and increasing business value.

HIGH AVAILABILITY AND DISASTER RECOVERY

The impact of downtime

If an internet outage occurs due to network issues, natural disasters, or viruses, the financial impact of downtime can be significant. Corporate and government mandates also require the durability and availability of data, and proven backup plans are necessary for compliance. However, fully restoring data and systems is time-consuming and leads to the potential of lost revenue.

A more secure way

A modern data pipeline supported by a highly available cloud-built environment provides quick recovery of data, no matter where the data is or who the cloud provider is. If a disaster occurs in one region or with one cloud provider, organizations can immediately access and control the data they have replicated in a different region or with a different cloud provider.

A COMPETITIVE ADVANTAGE

The massive enterprise shift to cloud-built software services combined with ETL and data pipelines offers the potential for organizations to greatly improve and simplify their data processing.

Companies that currently rely on batch ETL processing can begin implementing new continuous processing methodologies without disrupting their current processes. Instead of costly rip-and-replace, the implementation can be incremental and evolutionary, starting with certain types of data or areas of the business.

As you review the myriad data pipeline options available, consider that great data pipelines enable your business to gain a competitive advantage by making better, faster decisions. Just make sure your data pipeline provides continuous data processing; is elastic and agile; uses isolated, independent processing resources; increases data access; and is easy to set up and maintain.

Learn more about Snowflake Data Pipelines.

Your data pipeline checklist

- □ **Continuous and extensible data processing**
- □ **The elasticity and agility of the cloud**
- □ **Isolated and independent resources for data processing**
- □ **Democratized data access and self-service management**
- □ **High availability and disaster recovery**

snowflake®

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.**