

Cloud ETL Offload with VAST

Optimize ETL Costs. Preserve Compute Flexibility.



Cloud analytics platforms make it easy to scale compute, but using premium cloud resources for routine data preparation and transformation drives up costs as data volumes and pipeline complexity grow. Traditional ETL architectures also introduce repeated data movement between storage, compute clusters, staging areas, and analytical systems.

The VAST AI Operating System (AI OS) provides a more efficient path. Data can land in VAST DataStore for transformation with DataEngine or stream directly into VAST DataBase for continuously updated analytics, all within the VAST AI OS. DataEngine transforms and orchestrates pipelines close to the data, reducing reliance on external compute and unnecessary data movement. Organizations can offload data-intensive ETL from external compute and accelerate the path from raw data to insight while preserving existing Spark, Trino, SQL, and Python workflows.

Modernize ETL with a Unified Data Foundation

Transform Data Where It Lives

Traditional ETL pipelines, from companies like Databricks and Snowflake, repeatedly move data between storage, compute clusters, and analytical systems. At scale, every additional hop consumes network bandwidth, compute cycles, and storage while adding another system that data teams must manage.

For data that lands in VAST DataStore, DataEngine can execute custom Python functions for data preparation and transformation against the shared data foundation. Bringing compute closer to the data reduces unnecessary movement and intermediate copies, while external compute remains available for workloads that benefit from it.

Lower the Cost of Data Transformation

ETL costs extend beyond compute. Moving data between storage and processing environments consumes network resources, creates intermediate copies, and increases the infrastructure required to manage each pipeline. By offloading appropriate transformations to the same unified platform where data lands, VAST reduces external compute consumption, repeated I/O, and staging overhead, improving the economics of ETL as data volumes grow.

Why VAST

Reduce Databricks and Snowflake ETL Spend:

Reduce external compute, data movement, and intermediate storage required for ETL.

Transform Data Where It Lands:

Process data on VAST instead of repeatedly moving it between storage and compute.

Simplify Databricks and Snowflake ETL Pipelines:

Reduce staging, orchestration, and separate infrastructure across the data pipeline.

Accelerate Data Availability:

Stream data directly into queryable tables or trigger transformations as data arrives.

Modernize Without Rewrites:

Preserve existing Spark, Trino, SQL, and Python code and workflows.

Move from Scheduled ETL to Event-Driven Pipelines

Traditional batch pipelines introduce delays between when data arrives and when it becomes available for analysis. VAST DataEngine combines the Kafka-API compatible VAST Event Broker with data triggers and custom Python functions, allowing processing to begin automatically as new data arrives.

Streaming data can flow directly into VAST DataBase, where topics are represented as queryable tables, or trigger transformation and enrichment as part of an event-driven pipeline. This enables continuously updated analytical datasets while reducing the separate infrastructure typically required to coordinate ingestion, streaming, transformation, and orchestration.

Load Analytics-Ready Data Directly into VAST DataBase

Data can be loaded into VAST DataBase following transformation or ingested directly from event streams, providing a native analytical layer for high-concurrency SQL, low-latency retrieval, and continuously updated analytics.

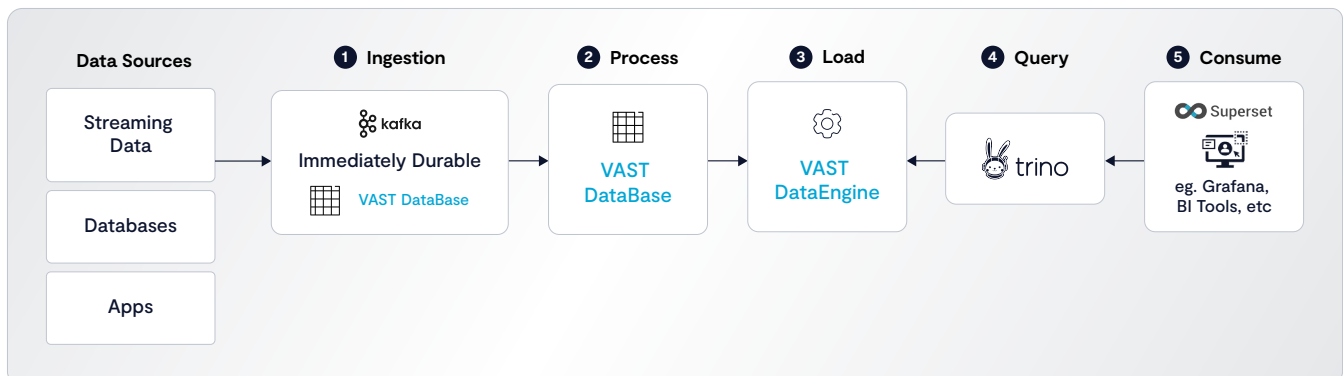
Storage, transformation, and analytics are integrated capabilities of the VAST AI Operating System, built on the same shared data foundation. This eliminates much of the staging, copying, and synchronization required when those functions reside in separate systems. Data moves through the analytical lifecycle on a single unified platform rather than between infrastructure silos.

Modernize Without Disrupting Existing Workflows

ETL offload doesn't require a rip-and-replace approach. Existing Spark, Trino, SQL, Python, and external compute environments can continue to operate against data on VAST, while DataEngine takes on appropriate transformation and pipeline workloads.

Organizations can modernize incrementally, shifting predictable or data-intensive processing onto VAST while preserving existing applications, code, skills, and workflows. Teams retain the flexibility to run each workload where it makes the most sense, adopting a more efficient architecture without disrupting the analytics environment already in place.

Improve Time to Insights with Real-Time Data Pipelines



About VAST Data

VAST Data is the AI Operating System company – powering the next generation of intelligent systems with a unified software infrastructure stack that was purpose-built to unlock the full potential of AI. The VAST AI OS consolidates foundational data and compute services and agentic execution into one scalable platform, enabling organizations to deploy and facilitate communication between AI agents, reason over real-time data, and automate complex workflows at global scale. Built on VAST's breakthrough DASE architecture – the world's first true parallel distributed system architecture that eliminates tradeoffs between performance, scale, simplicity, and resilience – VAST has transformed its modern infrastructure into a global fabric for reasoning AI. Learn more at vastdata.com and follow VAST Data on [LinkedIn](#), [YouTube](#) and [X](#).