StreamSets

Cloud Data Warehouse

StreamSets for Snowflake®

WHITE PAPER

Introduction

A data warehouse is a relational database that serves as a historical repository for integrating the information and data that is needed by the business for analytical purposes, which may come from a variety of different sources. Data has become more diverse and complex than ever. Transactional data is continually generated, but for many companies the most interesting insights come from incorporating machine-generated data. Data from business applications, web servers, devices, and sensors can be used to attain valuable business insights.

As analytics moves to the cloud, a fast and scalable cloud data warehouse becomes essential. Snowflake Elastic Data Warehouse makes it possible to stop worrying about the infrastructure you need to capture, integrate, and analyze data, and instead focus on how to extract insights from the data.

The Challenge

While the easiest-to-use cloud platform relieves your infrastructure burden, users often struggle to move data from their existing systems into a cloud data warehouse. The challenges that come with traditional data warehousing solutions are:

- **Delayed Analytics:** Reliance on batch loads increases latency, meaning analytic results are not real-time.
- **Complexity:** Architectural complexity grows geometrically as enterprises connect their one-stop-shop data warehouse to a diversity of fit-for-purpose systems such as relational databases (RDBMS), message queues (e.g. Apache Kafka) and NoSQL stores. Capabilities to handle structured, unstructured, semi-structured and binary data add on to the complexity for data ingestion.
- **Blindness:** Traditional data integration tooling, hand-coding or system-specific frameworks are poorly instrumented and cannot provide end-to-end runtime metrics. This creates operational blindness - an inability to track and detect problems in data flow logic, infrastructure systems or the data itself. Sometimes data problems are only discovered by the analytics consumer when they notice anomalous results.
- **Change (Data Drift):** The enterprise architecture is a living thing with an accelerating rate of change due to more systems receiving more frequent upgrades; higher volume, velocity and variety of data; and more users with their own evolving requirements. This constant change leads to data drift: the unexpected, unannounced and unending changes to data structure, infrastructure, and semantics. Data drift, when not addressed, pollutes data and breaks analytic workflows.

The implications of these challenges are poor developer productivity, lost analytic opportunities, and increased operational risk. Overruns, delays, and failures of data-driven projects result from a lack of available data engineers able to bridge multiple data movement technologies, from hand-coding to low-level platform-specific frameworks.

To maximize the value of cloud data warehouses such as Snowflake you must design a system that delivers all the data into the platform; easily, securely, and continuously.

Solution

StreamSets for Snowflake delivers streaming, batch, and change data capture (CDC) for the Snowflake cloud data warehouse. StreamSets helps organizations build a DataOps practice that manages delivery and performance of data pipelines on premises, across public clouds, and with managed services such as the Snowflake.

StreamSets helps you get the most from your cloud data warehouse:

- **Develop high performance ingest capabilities.** StreamSets provides a visual UI for designing data ingestion pipelines, covering a wide variety of on-premises and cloud data sources and including comprehensive monitoring, alerting, automation, and handling of data drift. As a result, data engineers become more productive and gain continuous visibility and control over data availability, integrity, and protection.

  They use StreamSets to design and run batch and streaming pipelines using a drag-and-drop environment that minimizes coding and facilitates collaboration. It can also detect and handle data drift, such as added fields or changed data types that occurs without notice when data sources are upgraded.

- **Deliver Quality data continuously.** Perform rich transformations with the numerous pre-built or custom data transformations on data-in-motion using StreamSets. This enables delivery of clean and consumption-ready data to data warehouse users.

- **Enhance operational agility and control.** With StreamSets, users can centralize deployment, management, and monitoring of their data pipelines, spanning workloads and requirements, to can remain collaborative and agile when working with cloud services. They can manage hundreds of complex data flow topologies from a central point, getting end-to-end visibility into their data movement. This helps modern organizations develop a DataOps practice and continuously monitor the health, delivery, and security of critical analytics pipelines.

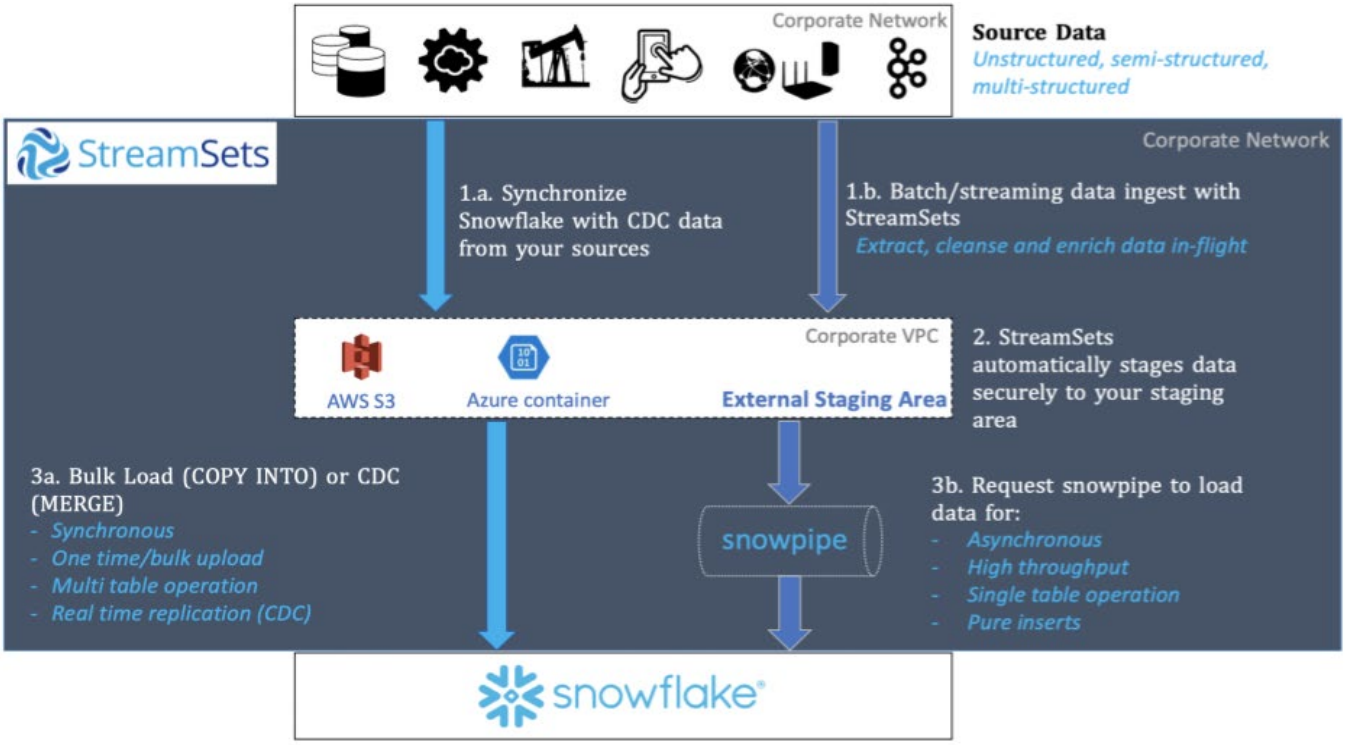
- **Automate bulk uploads and multi-table updates.** By automatically creating multiple tables, StreamSets makes ingestion into a cloud data warehouse resilient to shifting changes in the table structure. Users can focus on getting data into the cloud data warehouse and stop

worrying about drifting schema and structure. StreamSets users will see increased and consistent performance for both synchronous and asynchronous workloads.

- **Near real-time replication of data.** StreamSets provides continuous, automated, and cost-effective service that loads all your data without manual effort. It also enables Change Data Capture (CDC) for efficient synchronization of RDBMS origins (such as Oracle, PostgreSQL, and MySQL) with Snowflake, enabling near real-time analytics to end users.
- **Data protection.** StreamSets governs all data flows with policy-based auto-detection and protection of sensitive data before it lands in the Snowflake data warehouse, adding a layer of compliance with data privacy and regulations. StreamSets can detect PII and regulated data based on a large number of standard or custom templates and can apply many obfuscation and routing actions to the data.

Architecture

Here's how StreamSets ingests data and pushes to Snowflake:



1. StreamSets connects to a myriad of sources with out-of-the-box connectors. Users can create data pipelines to pull data in a variety of ways:
    1) **Batch** for one time/bulk loads from a source
    2) **Streaming** for continuously pulling near real time data
    3) **CDC** for capturing and replicating transactional data

2. As data is read from the sources, StreamSets automatically stages this data into your External Staging pools on S3 or Azure.

3. StreamSets then continuously pushes data into Snowflake in a variety of ways:
    1) **Synchronous** loads into Snowflake using COPY INTO or MERGE (thereby activating the warehouse). Use this option for:
        - Initial seeding of data into Snowflake/bulk uploads
        - Multi-table operations
        - Real-time replication with CDC origins

    2) **Asynchronous** loading into Snowflake using snowpipe. Snowpipe enables loading data from files as soon as they're available in the external staging area. This means data will be made available to users within minutes rather than waiting to COPY larger batches.
        - High throughput
        - Applicable for single table operations
        - Inserts only

Here's a snapshot of a dataflow pipeline in execution mode writing to Snowflake:

Conclusion

StreamSets amplifies the power of cloud data warehouses such as Snowflake by simplifying and automating the process of getting both structured and unstructured data into the cloud data warehouse platform, so that analytics experts, data engineers, SQL developers, enterprise architects and other users can concentrate on how to best use that data. Additionally, it lets you easily tie Snowflake into larger more complex data architectures.

To learn more including videos, solution briefs, and use cases, visit our website at streamsets.com/partners/snowflake.

ABOUT STREAMSETS

StreamSets transforms how enterprises flow big and fast data from myriad sources into data centers and cloud analytics platforms. Its DataOps platform helps companies build and operate continuous data flow topologies, combining award winning open source data movement software with a cloud-native Control Hub. Enterprises use StreamSets to enable cloud analytics, data lakes, Apache Kafka, IoT, and cybersecurity.

Founded by Girish Pancha, former chief product officer of Informatica, and Arvind Prabhakar, a former engineering leader at Cloudera, StreamSets is backed by top-tier Silicon Valley venture capital firms, including Battery Ventures, New Enterprise Associates (NEA), and Accel Partners. For more information, visit streamsets.com.

BENEFITS

DataOps

- One platform for data flow standardization in your organization
- Handle structured, unstructured, semi-structured or binary data seamlessly
- Monitoring, alerting, automation, and Data Drift handling
- Automatic offset management, delivery guarantees and failover handling

Change Data Capture

- Design replication with change data capture (CDC) from major RDBMS including Oracle, MS SQL Server, PostgreSQL
- Easy setup and automatically keep Snowflake in sync with source schemas
- Real-time replication for full database schemas

High Performance Ingest

- High throughput asynchronous ingest with snowpipe
- High throughput synchronous ingest for multi-table operations
- Multi-threaded for scale up and multi-node ingest for scale out

Data Privacy

- Policy driven protection and governance enforcement of data in motion