SOLUTION BRIEF

Building a Modern Data Warehouse with Microsoft Azure and StreamSets

Classic enterprise data warehouses (EDWs) have been a critical piece of every enterprise data strategy since the 1990s. These traditional EDWs are struggling to keep up with today's demands and need to be modernized with Microsoft Azure's cloud services to reduce costs, increase scale to meet SLAs, and unlock business value from modern data streams. From patient stratification to pay as you go car insurance, every industry needs to rethink their EDW strategy as new data systems, new use cases, and new data consumers continue to emerge.

The journey to modernizing a data warehouse using Azure has many paths including;

1) Refactoring a resource-intensive piece of the EDW in the cloud to meet SLAs
2) Rehosting the entire on-prem EDW in the cloud
3) Building a new modern data warehouse (DW) for emerging use cases

Wherever you are on your cloud DW journey, StreamSets can help remove friction from the process by offering a comprehensive DataOps platform that unlocks secure stream data processing, fully integrated with Azure's cloud services.

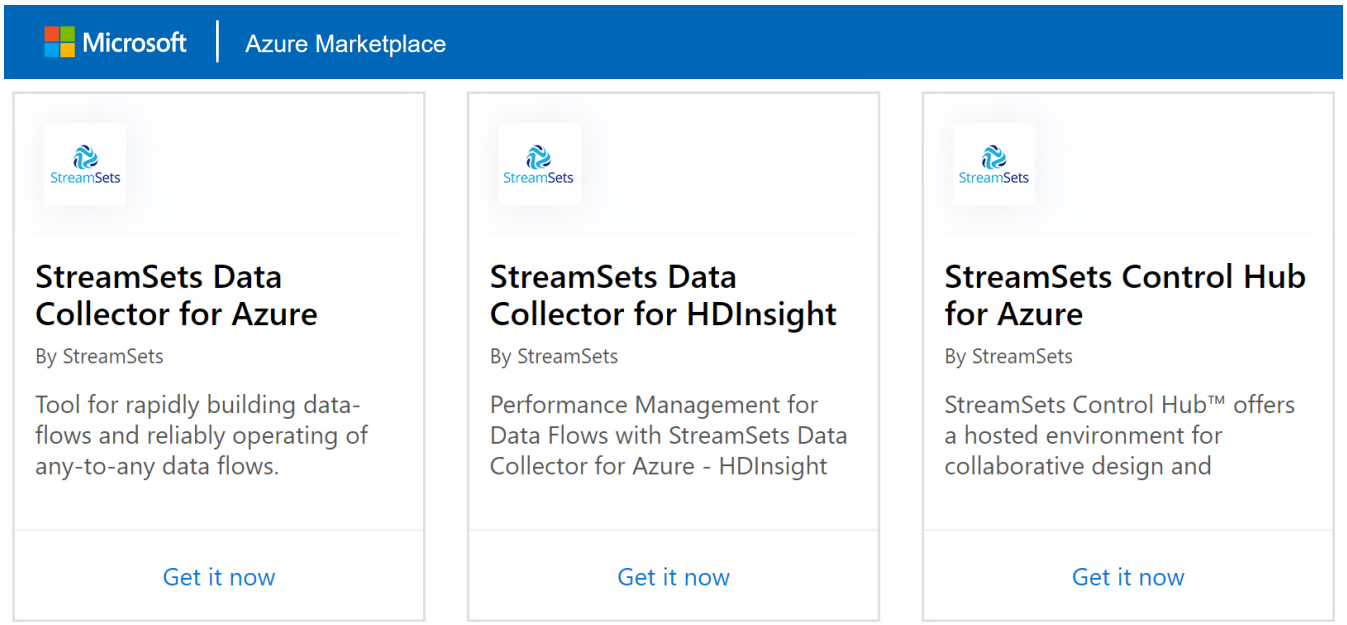
A Comprehensive DataOps Platform

With StreamSets' DataOps platform, enterprises can manage the entire data integration lifecycle from a single pane of glass, StreamSets Control Hub, without stitching together countless cloud services for an enterprise-grade pipeline. Out of the box alerting, change data capture capabilities, pipeline repository for easy code re-use, interactive UI, and access to hundreds of pre-built origins and destinations allow enterprises to quickly set up a data integration engine that will meet current and future needs; even when data or infrastructure drifts apart.

> **Data Drift** - The unpredictable, unannounced and unending mutation of data characteristics caused by the operation, maintenance, and modernization of the systems that produce the data.

Complete Azure Integration

StreamSets' DataOps platform is fully integrated with Azure's ingestion, processing, storage, machine learning, SQL, security, and authentication managed services so you don't have to worry about compatibility when Azure standards are already set. Deploying the full StreamSets enterprise capabilities, or just using StreamSets Data Collectors in the Azure Marketplace, provides enterprises all of the data integration capabilities they need to refactor, rehost, or build a modern DW on Azure.



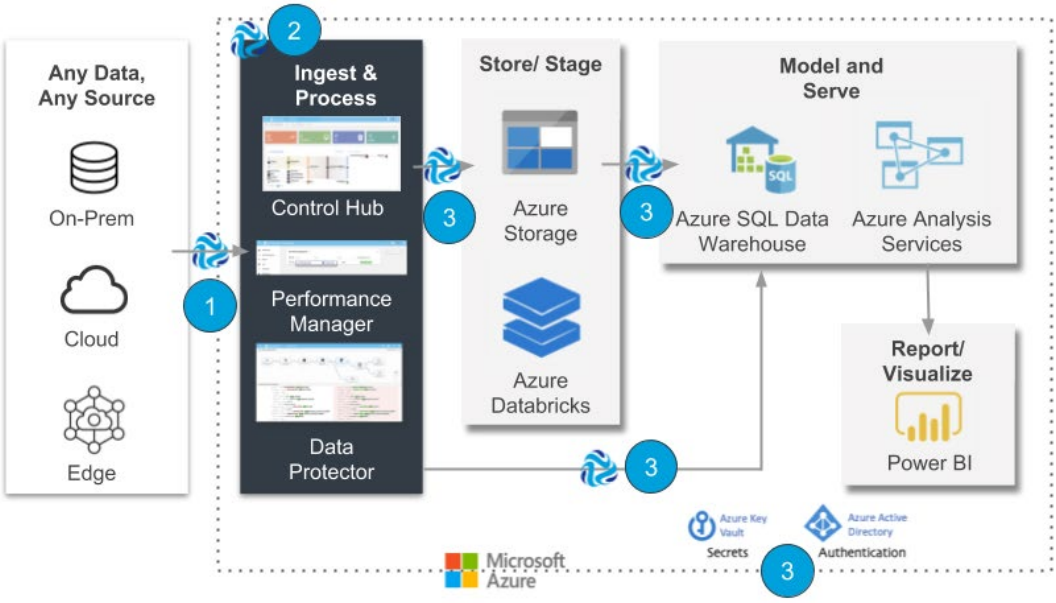
Autonomously Scale and Mask Data Pipelines

StreamSets allows enterprises to autonomously scale processing while detecting and masking sensitive data fields to meet all SLA and compliance or regulatory requirements. If sensitive information (e.g. PII data) gets mixed up in any data stream, enterprises can rest easy knowing StreamSets will detect and obfuscate sensitive dimension as it streams into your Cloud DW. The keys to unmasking the data can then be stored in Azure Key Vault for future use.

StreamSets Dataflow Performance Manager (DPM™)
Analyzes metadata from Data Collectors to enforce Data SLAs for data availability and accuracy.

StreamSets Data Protector™
Comply with data privacy regulations by automatically detecting and protecting sensitive data, before it is stored or shared.

Modern Cloud Data Warehouse on Azure with StreamSets



StreamSets Data Collector

StreamSets Data Collector (SDC) is an easy-to-use data pipeline engine for streaming, CDC and batch ingest from any source to any destination. With StreamSets, you spend your time building data pipelines, enabling self-service, and innovating, and minimize the time you spend maintaining, rewriting and fixing pipelines. Only StreamSets makes it both easy to get started building pipelines quickly with intent-driven design, and also easy to extend for complex enterprise needs. Only StreamSets' smart data pipelines eliminate 80% of the time you spend on break-fix and maintenance, and lets you port your pipelines to new data platforms without rewrites.

StreamSets Transformer

StreamSets Transformer is a data pipeline engine designed for any developer or data engineer to build ETL and ML pipelines that execute on Spark. Using an intent-driven visual design tool, users can create pipelines for performing ETL and machine learning operations. It allows everyone to harness the power of Apache Spark by eliminating the need for Scala and PySpark coding. Transformer pipelines provide unparalleled visibility into the execution of Spark applications for easy trouble-shooting, reducing the time to design and operate pipelines on Spark for developers of all skill levels. And Transformer not only supports all the major Spark distributions, its pipelines are portable across platforms, letting you redeploy without rewriting.

What's Next

Enterprises need to think through their cloud DW journey and decide exactly where they want to start and end. From building a single pipeline using StreamSets Data Collector in the Azure marketplace, to rehosting your legacy system, to building a new cloud DW from scratch; StreamSets can help reduce the data integration friction when modernizing your EDW with Azure's cloud services.

UNIQUE STREAMSETS FEATURES

- Data and infrastructure drift detection and resolution
- Continuously develop, test, and deploy pipelines out of the box
- 100s of pre-built connectors with a hosted repository for net-new pipelines
- Automatically mask sensitive data
- Autoscale to meet any SLA with Kubernetes
- Monitor pipelines and alert on multiple pipeline parameters
- Out of the box processing for streaming, batch, and changing data
- Record level transformations
- Deploy on any cloud (in any region), on-prem, on the edge, or across hybrid environments

ABOUT STREAMSETS

At StreamSets, our mission is to make data engineering teams wildly successful. Only StreamSets offers a platform dedicated to building the smart data pipelines needed to power DataOps across hybrid and multi-cloud architectures. That's why the largest companies in the world trust StreamSets to power millions of data pipelines for modern business intelligence, data science, and AI/ML. With StreamSets, data engineers spend less time fixing and more time doing.

To learn more, visit www.streamsets.com and follow us on LinkedIn.