

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.

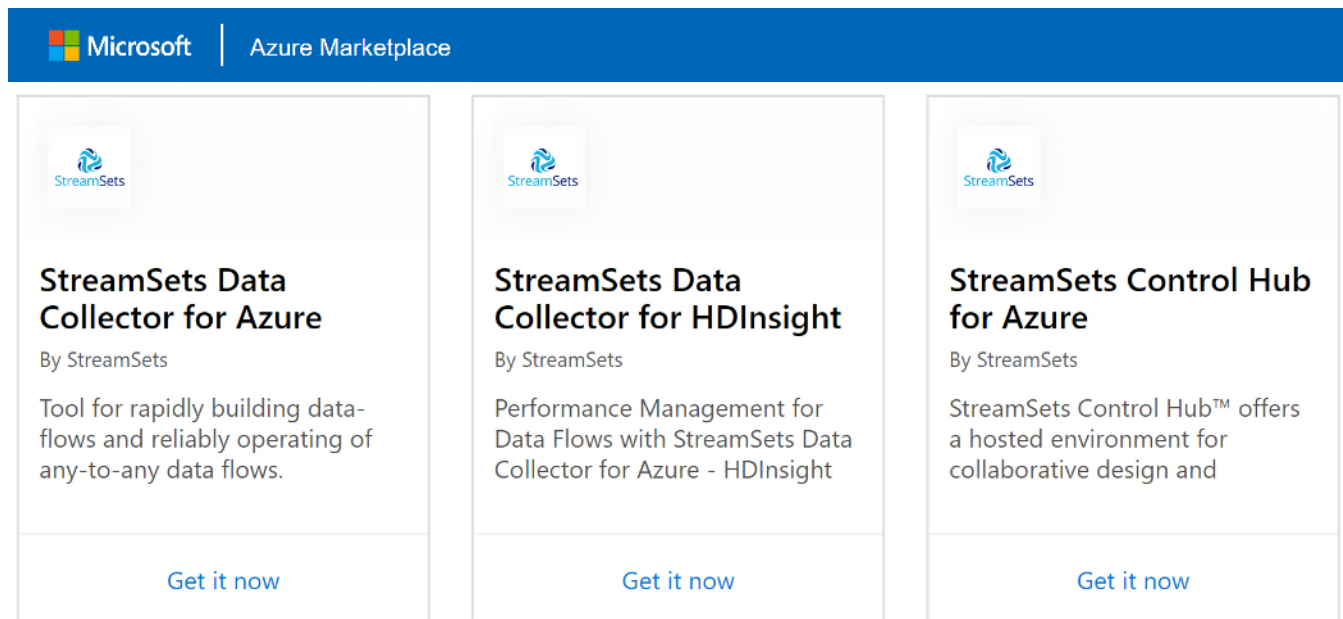
A Comprehensive DataOps Platform

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.



Microsoft | Azure Marketplace


StreamSets Data Collector for Azure
By StreamSets
Tool for rapidly building data-flows and reliably operating of any-to-any data flows.
[Get it now](#)


StreamSets Data Collector for HDInsight
By StreamSets
Performance Management for Data Flows with StreamSets Data Collector for Azure - HDInsight
[Get it now](#)


StreamSets Control Hub for Azure
By StreamSets
StreamSets Control Hub™ offers a hosted environment for collaborative design and
[Get it now](#)

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.

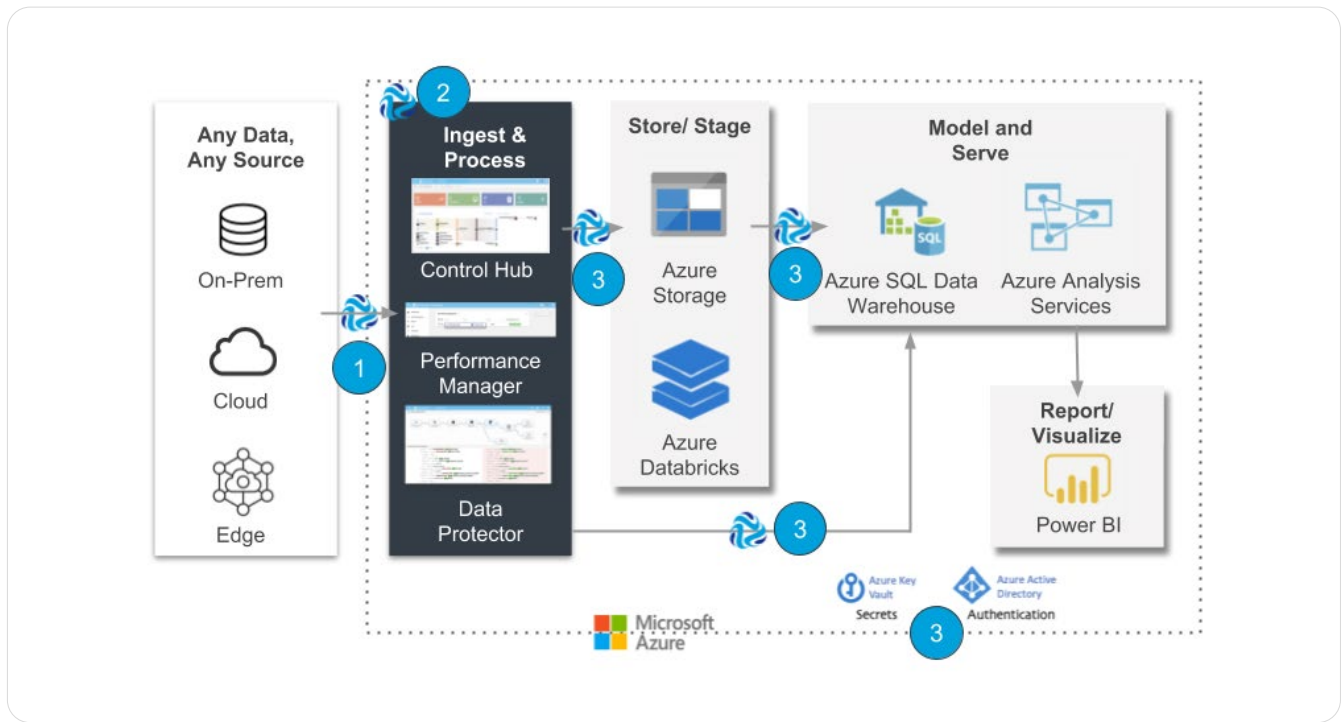
StreamSets Dataflow Performance Manager (DPM™)

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



1) StreamSets Data Collector

StreamSets Data Collector (SDC) is a lightweight design and execution engine that can live on-prem, in any cloud, or at the edge. With hundreds of connectors built, SDC's full-featured, open-source, in-memory data collectors can get a cloud DW the data it needs in no time with minimal coding.

2) StreamSets Enterprise

StreamSets Enterprise offers a comprehensive DataOps platform that includes StreamSets Control Hub, Performance Manager, and Data Protector. Build, monitor, secure, and continuously integrate and deploy data pipelines for cloud DWs using data and/ or infrastructure that lives on-prem, in the cloud, or at the edge.

3) Azure Integration

Move data to and from any Azure storage and processing service (Blob, ADLS Gen1, ADLS Gen2, or Databricks), or directly into Azure SQL Data Warehouse if your staging environment lives in another cloud or an on-prem database. With integration with [Azure Kubernetes services](#), an enterprise can automatically scale-out elastic workloads. Use Azure's Key Vault to unlock encrypted data and store keys for data that is masked by StreamSets Data Protector. All of StreamSets can be protected with Azure Active Directory for enterprise grade authentication.

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

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.