

# StreamSets

A DataOps Platform



WHAT IS  
STREAMSETS ?



WHY  
STREAMSETS ?



FEATURES



DEMO



STRENGTHS



COMPARISON



KEY NOTES

# Agenda

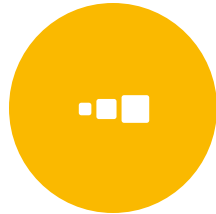
# What is StreamSets

- Lightweight, powerful design and execution engine that streams data in real time.
- Design data pipelines to route, process in data streams.
- Pipeline Stages:
  - Origin
  - Processing/Execution
  - Destination
- Data Collector processes data when it arrives at the origin and waits quietly when not needed.
- Real-time statistics about your data
- Inspect data as it passes through the pipeline
- The industry's first data operations platform for full life-cycle management of data in motion.

# Why StreamSets

- StreamSets easily loads data from multiple sources
  - RDBMS, Flat File, CRM , NoSQL DBs
- Supports streaming data to Multiple Destination
- Pipeline can be reused within other pipeline
- Can build platform agnostic data pipelines.
- Optimized for continuous ingestion with no data latency.
- Batch and streaming dataflows can be built with minimal coding.
- Manage Data Across Multiple Platforms
  - Single view across all data operations, on premises or in the cloud.
- Realtime Data Synchronization across multiple platforms.
- Being platform-agnostic and legacy-friendly, it provides flexibility to work with modern data.

# Features



Build Batch &  
Streaming  
Pipelines in Hours



Map and Monitor  
Runtime  
Performance

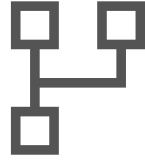


Protect Sensitive  
Data as it Arrives



Drag and Drop GUI  
facilitates Low  
Code Platform

# Demo



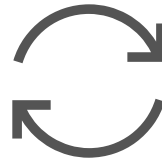
Hands on Data Pipeline  
Creation to Stream Bulk  
Data



Handling Complex  
Queries



Streaming data to  
multiple destination



Data Processing



Handling Data  
Modification



FAST



EASY TO USE



LOW CODE  
PLATFORM



FLEXIBLE TO  
HANDLE DATA  
DRIFTS



STABLE AND  
SUSTAINABLE DATA  
OPERATIONS



MANAGED DATA  
ACROSS MULTIPLE  
PLATFORMS



SURFACE ERRORS  
WITHOUT  
SEARCHING  
THROUGH LOG FILES



DRIFT DETECTION  
AND DRIFT  
ALERTING FOR A  
MULTI-CLOUD  
TOPOLOGY

# Strengths

# Similar Tools

Tools	Streamsets	AWS Glue	Apache Flink	AWS Data Pipeline
Focus	ETL, transformations, data prep	ETL, data catalog	Apache Flink is an open-source, unified stream-processing and batch-processing framework	Data transfer
Database replication	Full replication or change data capture Batch and Stream processing	Full table; incremental via change data capture through AWS Database Migration Service (DMS)	Full replication or change data capture Batch and stream processing	Full table; incremental replication via timestamp field
SaaS sources	Only Salesforce	None	Salesforce	None
add new data sources	Yes	Glue.	Yes	No
Connects to data warehouses? Data lakes?	Yes / Yes	Yes / Yes	Yes / Yes	Yes / Yes
Compliance, governance, and security certifications	HIPAA, GDPR	HIPAA, GDPR	GDPR	None



- StreamSets is an open source, enterprise-grade, continuous big data ingest infrastructure
- It accelerates time to analysis by bringing unprecedented transparency and processing to data in motion.
- It facilitates Realtime Data Synchronization across multiple platforms.
- Being platform-agnostic and legacy-friendly, it provides flexibility to work with modern data.

## Key Notes