AWS Reference Architecture Guide for StreamSets

Using StreamSets DataOps Platform to Integrate Data from PostgreSQL to AWS S3 and Redshift: A Reference Architecture

This document describes the reference architecture for integrating data from a database to Amazon Web Services (AWS) data analytics stack utilizing the StreamSets DataOps Platform, including the StreamSets Data Collector and Transformer engines, as the data integration platform.

It assumes a certain level of technical expertise but aims to deliver a high-level understanding of successfully deploying in AWS.

Business Use Case Examples

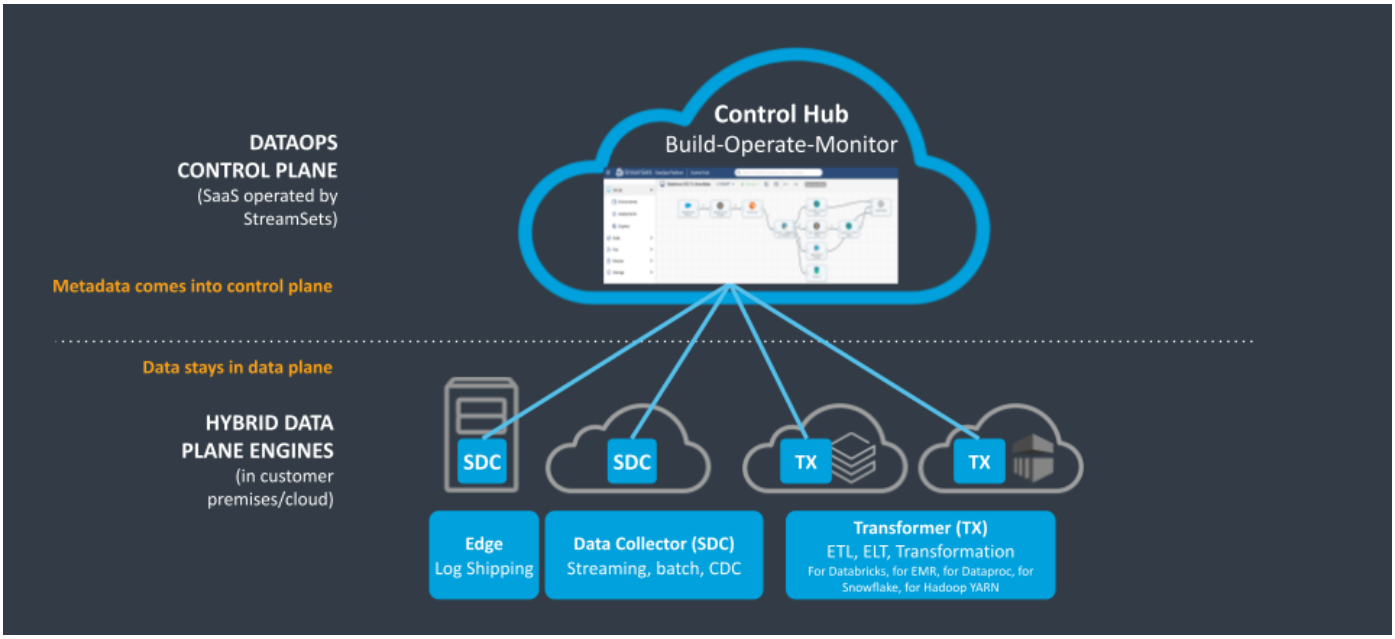
Many business use cases would fit this pattern of integrating data from a database, such as PostgreSQL to S3 and Redshift on AWS. Here are some potential applications:

Financial Services:

A large, multinational bank stores customer profile data such as demographics, account types, and total asset value in an PostgreSQL database. The bank wants to use this customer profile data, combined with data on web behavior already in AWS S3, to improve their personalized offers to customers. By taking core customer profile data from their PostgreSQL database and integrating it to S3, the bank will be able to consolidate a single view of their customers in their AWS analytics stack. They can use Sagemaker to make predictions on what services each customer cohort would be interested in purchasing using this secure data. Also, by deploying everything in their Amazon VPC, including StreamSets engines, they can ensure all data movement is highly secure.

Life sciences:

A pharmaceutical company integrates clinical research data from multiple databases from different lab sites around the world into a single Redshift database in real-time with StreamSets. The data is moved by StreamSets via change data capture. This ensures that as soon as a lab logs results from a clinical study, the pharmaceutical company's data scientists can access and analyze the results immediately, accelerating the new drug discovery process. The data scientists can harness the power of Sagemaker for in-depth analysis and Quicksight to display their findings for publication.



Engines and Deployment - How to Set Up StreamSets DataOps Platform on AWS

The Control Plane for StreamSets is a cloud-native application where all your engines, pipelines, and jobs can be created, scheduled, managed and monitored. However, data from the data pipelines does not enter the Control Plane; it remains within the engines. So, if you set up your Data Collector and Transformer engines in AWS, that data remains secure and separate within AWS.

Here is how to start setting up the StreamSets DataOps Platform to run in your AWS environment:

1. You can quickly deploy the StreamSets DataOps Platform directly from AWS Marketplace, or you can provision an EC2 instance and configure it as needed. Once configured, the StreamSets DataOps Platform automatically provisions the resources needed to run engines in AWS, so you don't have to worry about installing prerequisites.

2. StreamSets Transformer for Spark must be deployed where it can submit Spark jobs to your cluster manager, so deploying both in the same cloud environment makes a lot of sense. You can use Amazon Elastic MapReduce (EMR) clusters to run Transformer for

TIPS AND TRICKS

- You are responsible for all costs from AWS incurred by the resources provisioned by the StreamSets DataOps Platform. You are strongly advised against directly modifying the provisioned resources in AWS. Doing so may cause unexpected errors.
- Add the region and purpose to your StreamSets Data Collector engine labels to more easily manage these resources; e.g. Production, West or Development, East.

Spark. Amazon EMR is a managed cluster platform that can run big data frameworks, including Apache Spark, which Transformer uses to power up pipelines. You can choose an existing Spark cluster or set up Transformer to provision clusters to run pipelines. This second choice can be more cost effective because Transformer can terminate a cluster after pipelines stop, helping to ensure that you only pay for what you use.

3. Credential management can be done a few different ways securely. StreamSets can use instance profile credentials to authenticate automatically with AWS when engines are run on an Amazon Elastic Compute Cloud (EC2) instance. Alternatively, if your EC2 instance doesn't have an instance profile, or you are testing your pipelines locally, Amazon Secrets Manager is fully supported and can be used to store Amazon Access Keys.

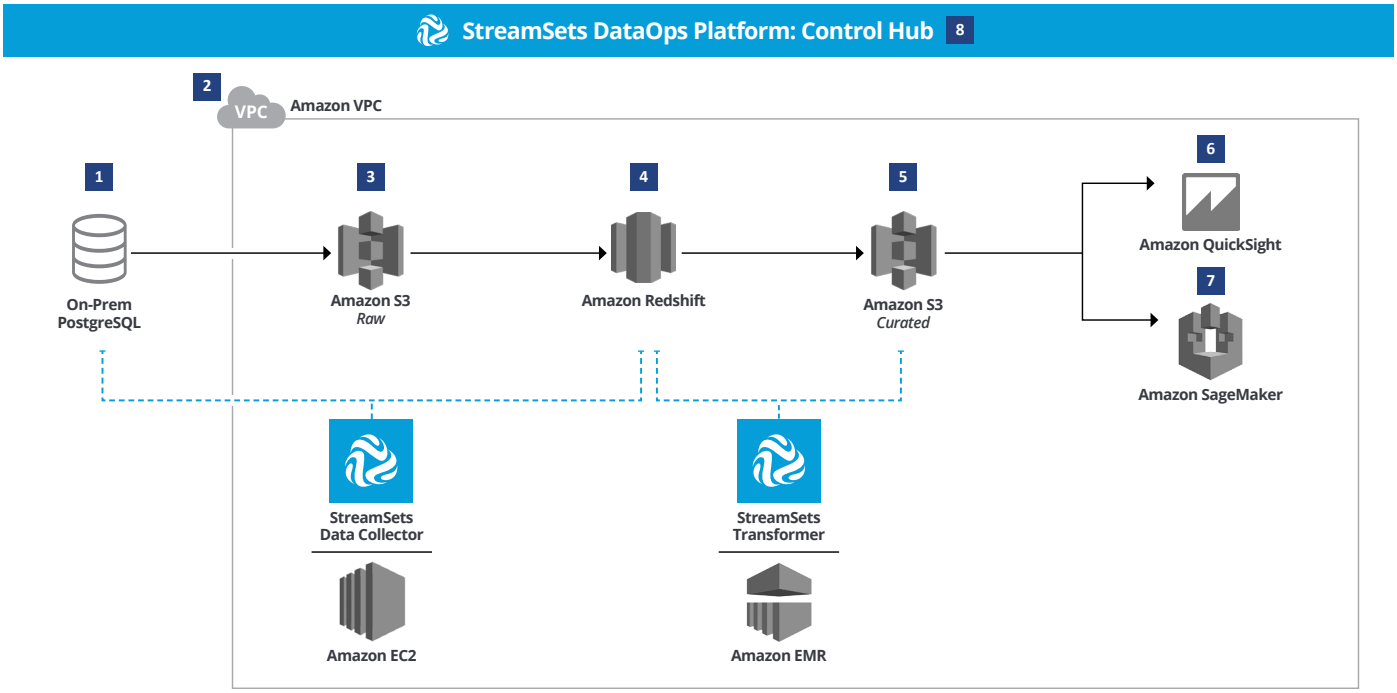
Reference Architecture:

Integrating data from PostgreSQL database to AWS S3 and Redshift

Overview: the key data source is a PostgreSQL database. StreamSets DataOps Platform's Data Collector engine, which is deployed on a VPC in AWS, moves the data from PostgreSQL into AWS S3 as a staging area and then subsequently to AWS Redshift. Redshift supports high availability and high volume analytical workloads. The StreamSets Transformer engine is used to further cleanse and curate the data utilizing AWS EMR. The cleansed and curated data is staged in S3 and then available for analytics and data science using tools such as Quicksight and Sagemaker.

1. The key data source, a PostgreSQL database, runs on premises within a private infrastructure.

2. AWS Virtual Private Cloud (VPC) is a private network within AWS that allows connected resources to communicate with each other. In the most straightforward AWS implementation of StreamSets Data Collector, engines for Data Collector pipelines



should be run inside the VPC on an EC2 instance.

3. StreamSets Data Collector is used to load change data capture (CDC) data from the PostgreSQL database into an Amazon S3 bucket via a data pipeline. Pipelines that use CDC will detect CRUD operations like insert, update or delete and pass those changes to a destination. StreamSets Transformer doesn't support CDC, but could connect an PostgreSQL database to S3. You would use StreamSets Transformer in this pattern if you needed higher performance or were operating on a larger scale. After the data is moved into S3 from the database it is available across the VPC.

4. This intermediate copy of the data from S3 can be moved to Amazon Redshift using the same StreamSets Data Collector pipeline in step 3. When an event occurs like data landing in S3, a JDBC executor can be triggered within a single pipeline to copy the data into Redshift. Redshift is selected for its ability to perform real-time or near real-time operations.

5. In this step, data is cleaned, aggregated, and batched for downstream analysis with a Streamsets Transformer for Spark pipeline. Transformer pipelines can be run on Spark deployed on an (EMR) cluster. Scale up or down depending on the amount of compute necessary to transform the data for the next steps. Curated data lands in Amazon S3 from the result of the operations of the StreamSets Transformer. Landing data in object storage after transformation is recommended for durability and optimizing cost savings. It is recommended that you keep raw and curated data in separate buckets for clarity.

6. Including the visualization layer in Amazon VPC allows for immediate connectivity. Amazon Quicksight can create visualizations to help get more out of the data hosted in S3.

Component	Tool	Description
Storage	Amazon S3	An object storage service offered by Amazon.
Data Integration	StreamSets DataOps Platform with Data Collector engine and Transformer for Spark engine	Cloud-native data integration platform. Data Collector and Transformer for Spark are execution engines within the platform.
Compute	Amazon EMR Amazon EC2	Amazon EMR is a managed cluster platform for running big data frameworks like Spark. Amazon EC2 are virtual computers in the cloud.
Data Warehouse	Amazon Redshift	Cloud data warehouse that is able to handle large scale data sets and operations.
Visualization and Analytics	Amazon Quicksight Amazon Sagemaker	Advanced analytics & BI and a machine learning & AI platform respectively.

Where to learn more about using StreamSets with AWS

Get Amazon's free tier to begin creating cloud-native pipelines with StreamSets on AWS today. Deploy StreamSets Data Collector or StreamSets Transformer in minutes from the AWS Marketplace.

Explore more data integration patterns in the Data Engineers Handbook and the Multi-cloud Matters White Paper.

To go deeper with StreamSets, join StreamSets Academy for instructor-led or self-paced video training, tutorials, and more. You'll also find resources, sample pipelines, and ideas in our community.