**What is DataStream (A GCP Service)**

**Google Cloud Datastream** is a fully managed, serverless data streaming service designed to facilitate real-time data movement and synchronization across various systems. It provides efficient data migration and replication capabilities, including support for **Change Data Capture (CDC)**, making it an ideal solution for transferring data seamlessly across diverse platforms. Datastream enables organizations to replicate data from operational databases, storage systems, and applications to data warehouses, data lakes, and other downstream services with minimal effort and without the need for complex infrastructure management. It supports data migration from on-premises systems and multi-cloud environments to Google Cloud analytics services such as BigQuery, ensuring smooth integration within the cloud ecosystem. As part of Google Cloud Platform's (GCP) suite of services, Datastream offers reliable, scalable, and real-time data replication, making it a powerful tool for businesses seeking to modernize their data infrastructure or enable advanced analytics.

**Change Data Capture (CDC)**

**Change Data Capture** is a method for tracking and capturing changes applied to data in a source system, ensuring that these changes are reflected in a target system (sink).

For example, imagine a real-time survey collecting data on the prices of items in supermarkets within a specific area. If someone mistakenly enters the wrong price for an item, they can recall and update the price instantly. With **CDC**, these changes are automatically tracked, captured, and replicated to the target system. This ensures that the analytics or reporting system always has the most accurate and updated data.

DataStream achieves this by using **publication and replication mechanisms** based on unique identifiers for each record, ensuring efficient tracking of changes.

**Replication and Publication Explained**

Publication and replication are the two key steps used in systems like DataStream to move and synchronize data between a source and a target system.

1. Publication
   * What it means: Think of publication as the process of identifying and preparing the data that needs to be shared.
   * How it works: The source system, like a PostgreSQL database, creates a "list" or "package" of data changes, such as updates, new records, or deleted records. These changes are organized based on unique identifiers (e.g., IDs of rows in a table).
   * Example: Imagine a survey application tracking item prices in supermarkets. If a price is updated, the system publishes this update, saying, "Hey, the price for Item ID 123 has changed!"
2. **Replication**
   * What it means: Replication is the actual process of copying the published data to the target system.
   * How it works: The changes published by the source system are sent (or streamed) to the target system, like BigQuery or another analytics database. The target system then applies these changes, so it always reflects the latest data from the source.
   * Example: Once the updated price for Item ID 123 is published, the replication process ensures this new price appears in the analytics dashboard or reporting database.

By combining publication and replication, DataStream ensures the efficient and reliable transfer of changes between systems, preserving data consistency and integrity.

**Why Do You Need DataStream?**

DataStream plays a critical role in modern data workflows due to the following reasons:

1. **Real-Time Data Movement:** It ensures timely and efficient data replication, enabling real-time updates from source systems to target systems like BigQuery or Cloud Storage.
2. **Support for Analytics:** By migrating data from OLTP systems to OLAP systems, DataStream facilitates analytics-driven decision-making.
3. **Change Data Capture (CDC):** It tracks and captures all changes made to the source data, ensuring that the target system remains up to date without requiring full data reloads.
4. **Publication and Replication:** Its publication and replication mechanisms ensure efficient, precise, and secure data movement between systems, even for complex datasets.
5. **Flexibility Across Platforms:** DataStream supports a variety of data sources, including on-premises and multi-cloud systems, making it versatile for different organizational needs.
6. **Simplified Data Management:** It reduces the complexity of managing ETL (Extract, Transform, Load) pipelines by automating data synchronization between sources and sinks.

**Importance of DataStream**

1. **Real-Time Data Synchronization:** Captures and streams real-time changes (inserts, updates, deletes) using Change Data Capture (CDC). This ensures systems like databases, data lakes, or analytics platforms stay updated without delays.
2. **Seamless Data Migration:** Transfers data from on-premises systems or multi-cloud environments to Google Cloud services like BigQuery or Cloud Storage.
3. **Serverless and Fully Managed**: Requires no infrastructure setup or maintenance, automatically scales resources based on workload, and handles failovers for reliability.
4. **Integration with Google Cloud Ecosystem:** Works seamlessly with services like BigQuery (data analytics), Pub/Sub (event-driven processing), and Dataflow (data transformations).
5. **Support for Multi-Cloud and Hybrid Environments:** Connects to diverse data sources, including on-premises databases and other cloud platforms.
6. **Automatic Handling of Schema Changes**: Adapts to changes in the source schema (e.g., adding new fields) without interrupting the data flow.
7. **Real-Time Analytics Enablement:** Feeds real-time data to analytics platforms like BigQuery, enabling up-to-date dashboards and reports.
8. **Data Transformation and Filtering:** Allows filtering or transforming data at the source before streaming it to destinations.
9. **Cost-Effectiveness**: Offers pay-as-you-go pricing and optimizes resource allocation.
10. **Scalability and Reliability**: Automatically adjusts resources to handle high data loads and ensures continuous operation with built-in failovers.

**Architecture**

A blue and white logo

Description automatically generated

**Building Solution using Google CLI and Google Cloud Console**

Logged into your Google Cloud Console

Click on the Cloud Shell icon

Activate your

Open your Google Cloud

Create a Postgress db on Google Cloud SQL using the CLI command

The DATASTREAM\_IPS depends on the region you set for you project and this is the region you are going to select for resources created in this Tutorial

Region is us-west1

POSTGRES\_INSTANCE=postgres-db

DATASTREAM\_IPS=35.247.10.221,35.233.208.195,34.82.253.59,35.247.95.52,34.82.254.46

gcloud sql instances create ${POSTGRES\_INSTANCE} \

--database-version=POSTGRES\_14 \

--cpu=2 --memory=10GB \

--authorized-networks=${DATASTREAM\_IPS} \

--region=us-west1 \

--root-password pwd \

--database-flags=cloudsql.logical\_decoding=on