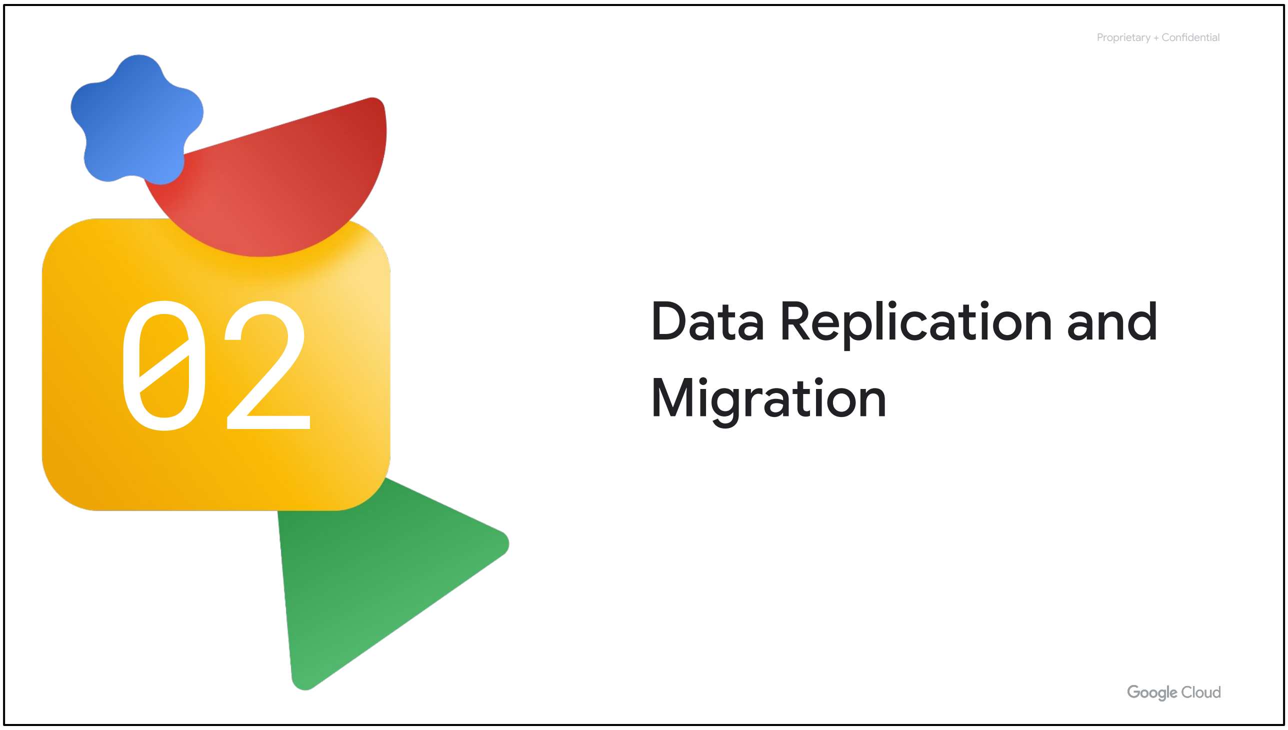
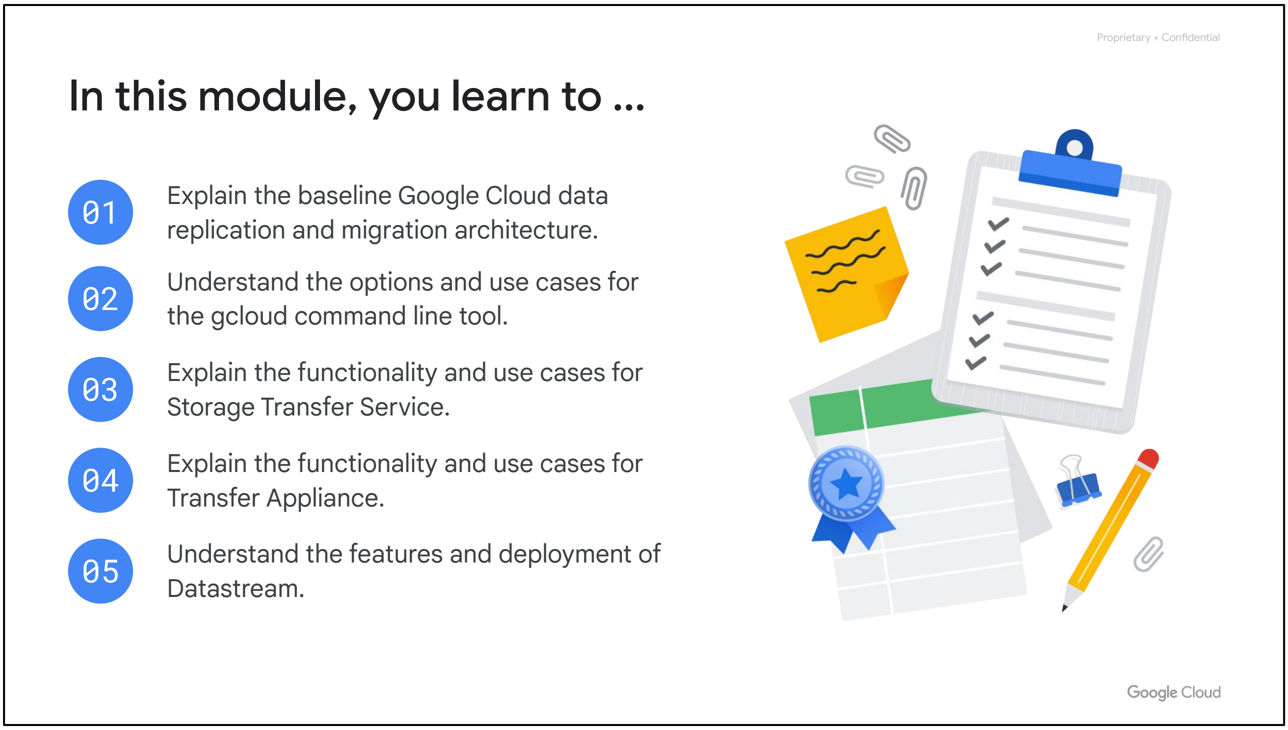


02

# Data Replication and Migration



# In this module, you learn to ...

01 Explain the baseline Google Cloud data replication and migration architecture.

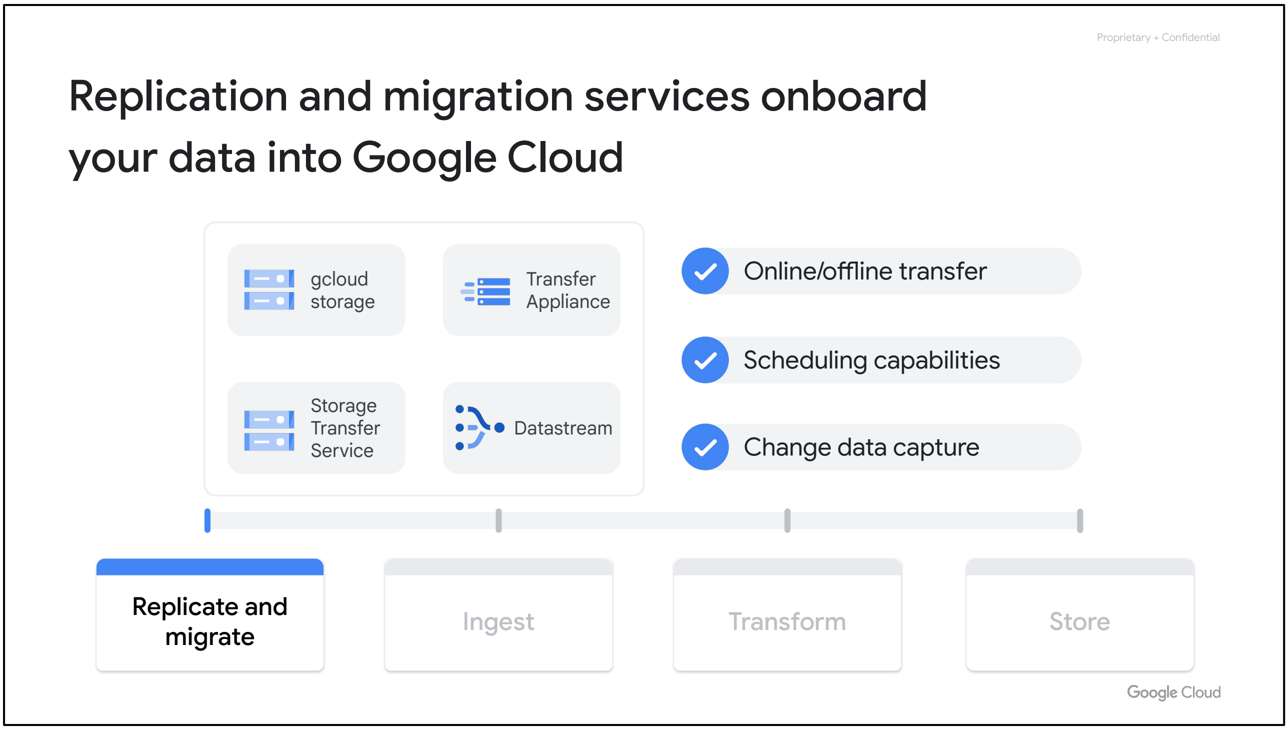
02 Understand the options and use cases for the gcloud command line tool.

03 Explain the functionality and use cases for Storage Transfer Service.

04 Explain the functionality and use cases for Transfer Appliance.

05 Understand the features and deployment of Datastream.

In this module, first, you review the baseline Google Cloud data replication and migration architecture.  Second, you will cover the options and use cases for the gcloud command line tool.  Then you will review the functionality and use cases for the Storage Transfer Service. The next topic addresses the functionality and use cases for the Transfer Appliance. Finally, you will look at the features and deployment of Datastream.



## Replication and migration services onboard your data into Google Cloud

- ✓ Online/offline transfer
- ✓ Scheduling capabilities
- ✓ Change data capture

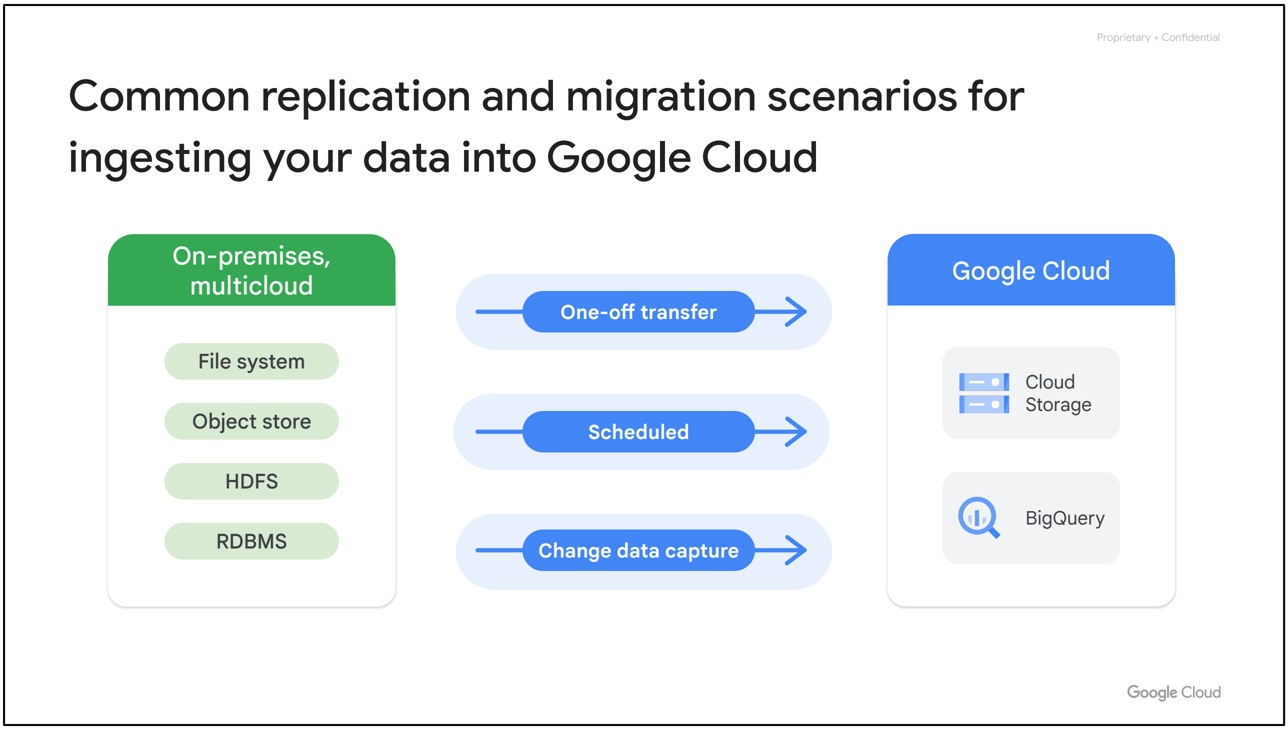
Google Cloud

The replicate and migrate stage of a data pipeline focuses on the tools and options to bring data from external or internal systems into Google Cloud for further refinement.

Google Cloud provides a comprehensive suite of tools to migrate and replicate your data.

Start replicating and migrating data by using tools like the  “gcloud storage” command, Transfer Appliance, Storage Transfer Service, or Datastream.

You can then transform the data as needed before finally storing it within Google Cloud.



Proprietary + Confidential

# Common replication and migration scenarios for ingesting your data into Google Cloud

On-premises, multicloud

File system

Object store

HDFS

RDBMS

One-off transfer

Scheduled

Change data capture

Google Cloud

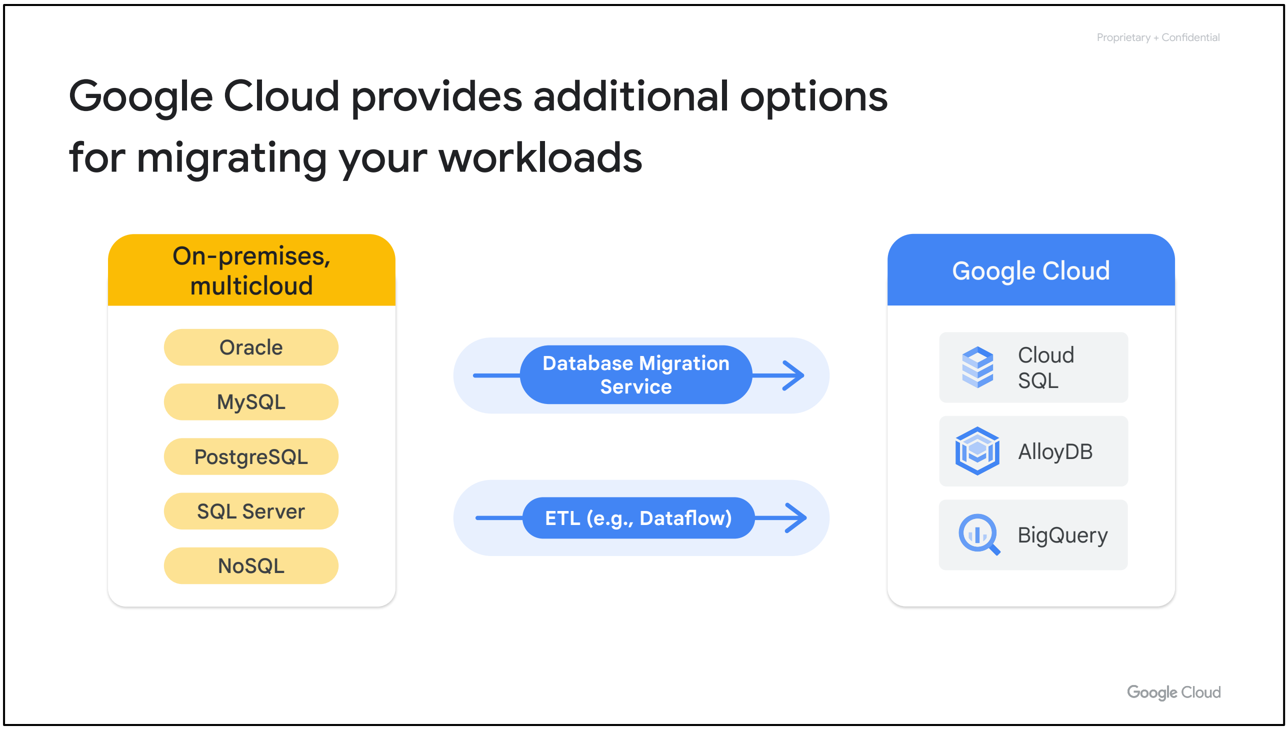
Cloud Storage

BigQuery

Google Cloud

Data can originate from on-premises or multicloud environments, including file systems, object stores, HDFS, and Relational Databases.

Google Cloud offers options for one-off transfers, scheduled replications, and change data capture, ultimately landing data in Cloud Storage or BigQuery.



Proprietary + Confidential

## Google Cloud provides additional options for migrating your workloads

On-premises, multicloud

- Oracle
- MySQL
- PostgreSQL
- SQL Server
- NoSQL

Database Migration Service

ETL (e.g., Dataflow)

Google Cloud

- Cloud SQL
- AlloyDB
- BigQuery

Google Cloud

Google Cloud provides additional workload migration with options for various database types.

Leverage Database Migration Service for seamless transitions from Oracle, MySQL, PostgreSQL, and SQL Server. For other data formats or complex migrations, use ETL tools like Dataflow with a wide range of templates that handle NoSQL or non-relational databases.

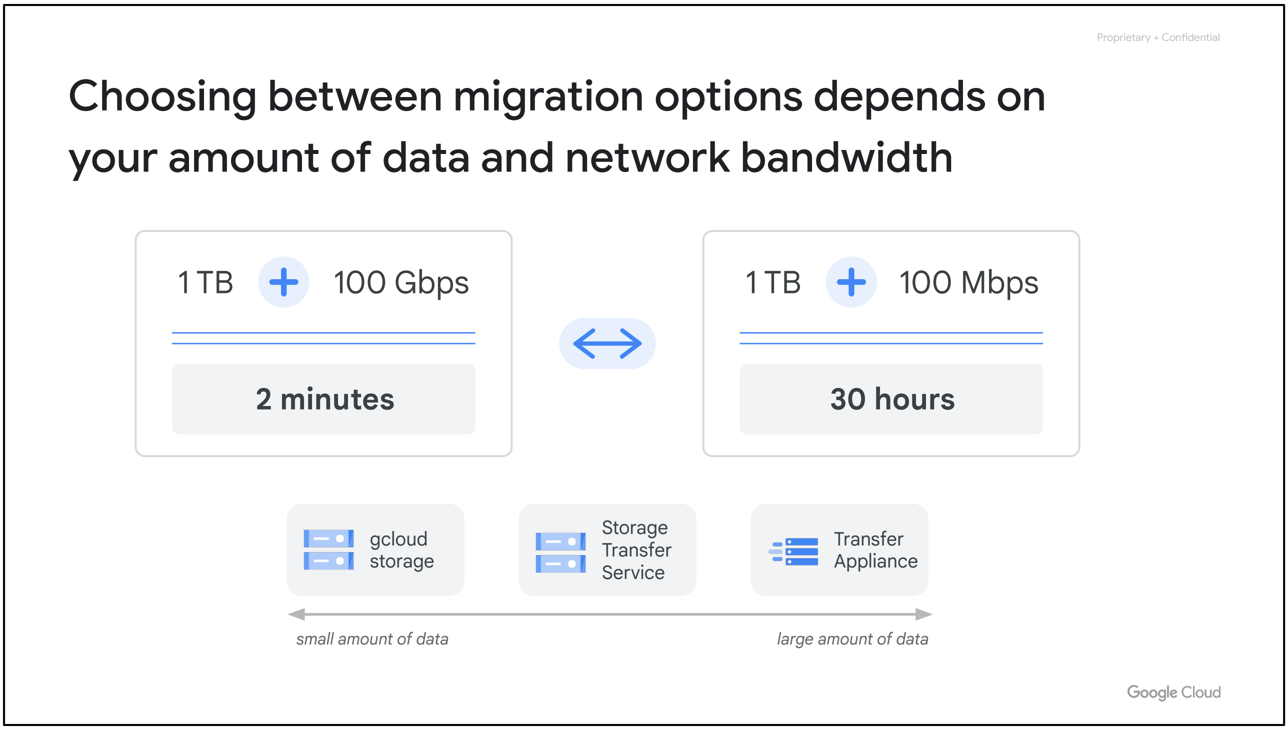
Your target destination can be Cloud SQL, AlloyDB, or BigQuery depending on your needs.

**References:**

DB Migration Service:
https://cloud.google.com/database-migration/docs/supported-databases

**Instructor notes:**

- service isn't covered by this course, just mentioned here to give a complete picture



Proprietary + Confidential

## Choosing between migration options depends on your amount of data and network bandwidth

1 TB + 100 Gbps

**2 minutes**

↔

1 TB + 100 Mbps

**30 hours**

gcloud storage

Storage Transfer Service

Transfer Appliance

*small amount of data*

*large amount of data*

Google Cloud

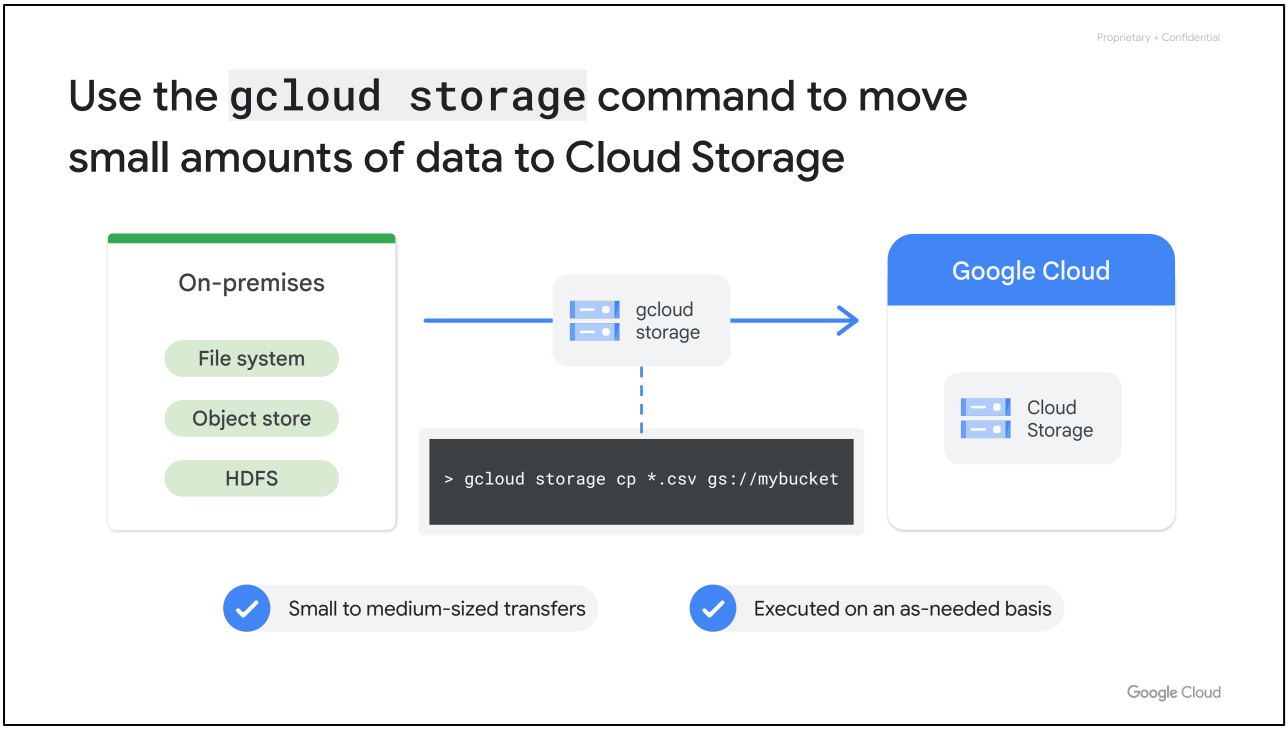
The ease of migrating data depends heavily on data size and network bandwidth.

With 1 Terabyte of data, a 100 Giga bits per second network takes about 2 minutes to transfer, while the same size on a 100 Mega bits per second network takes 30 hours.

The "gcloud storage" command or Storage Transfer Service are suitable for smaller datasets. For larger datasets, consider Transfer Appliance for faster offline transfer.

**References:**

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#online_versus_offline_transfer



Proprietary + Confidential

## Use the `gcloud storage` command to move small amounts of data to Cloud Storage

On-premises

File system

Object store

HDFS

gcloud storage

```
> gcloud storage cp *.csv gs://mybucket
```

Google Cloud

Cloud Storage

✓ Small to medium-sized transfers

✓ Executed on an as-needed basis

Google Cloud

You can use the “gcloud storage” command to transfer small to medium-sized datasets to Cloud Storage.

The data can originate from various on-premise sources like file systems, object stores, or HDFS.

The “cp” command, shown in the code snippet, facilitates these ad-hoc transfers directly to Cloud Storage.

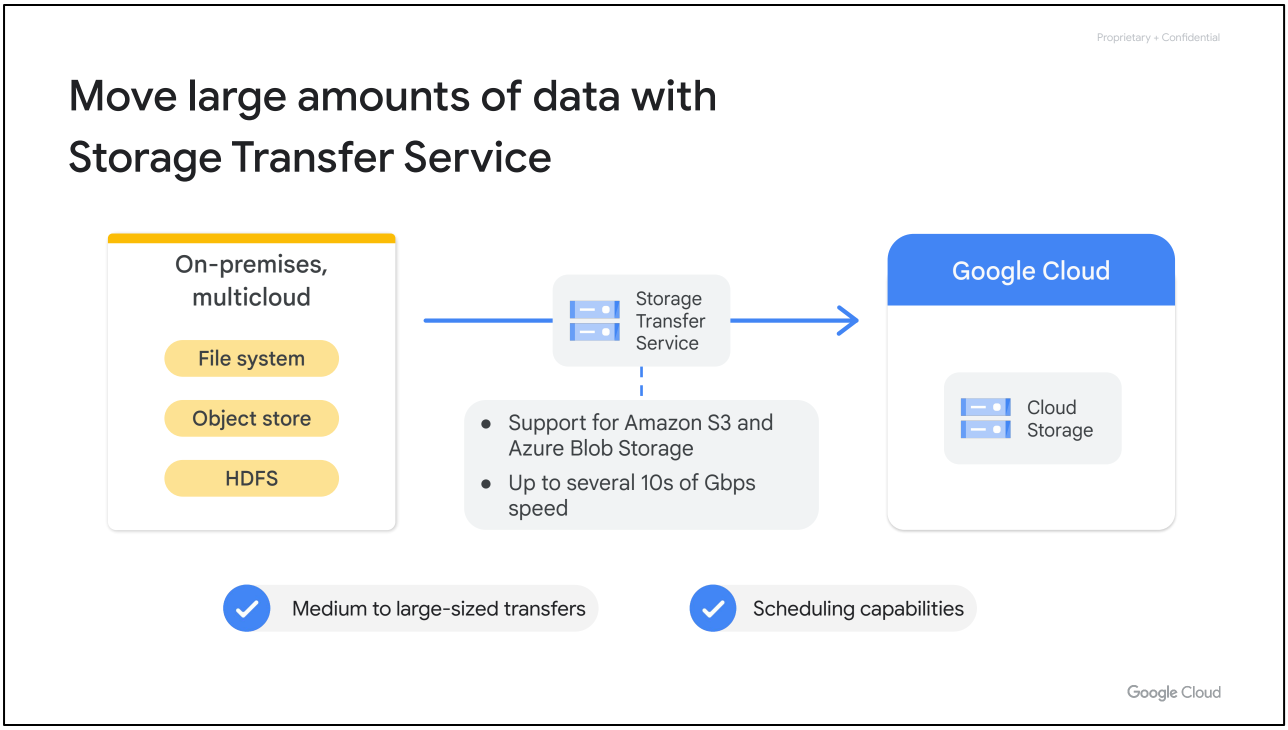
**References:**

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#gcloud_storage_command_for_smaller_transfers_of_on-premises_data

**Instructor notes:**

- we can also use it to transfer data from other cloud providers, but Storage Transfer Service is recommended for it
- The gcloud storage command is especially useful in the following scenarios:
  - Your transfers need to be executed on an as-needed basis, or during

- command-line sessions by your users.
- You're transferring only a few files or very large files, or both.
- You're consuming the output of a program (streaming output to Cloud Storage).
- You need to watch a directory with a moderate number of files and sync any updates with very low latencies.



Proprietary + Confidential

## Move large amounts of data with Storage Transfer Service

On-premises, multicloud

File system

Object store

HDFS

Storage Transfer Service

- Support for Amazon S3 and Azure Blob Storage
- Up to several 10s of Gbps speed

Google Cloud

Cloud Storage

✓ Medium to large-sized transfers

✓ Scheduling capabilities

Google Cloud

To move larger datasets, consider using Storage Transfer Service.

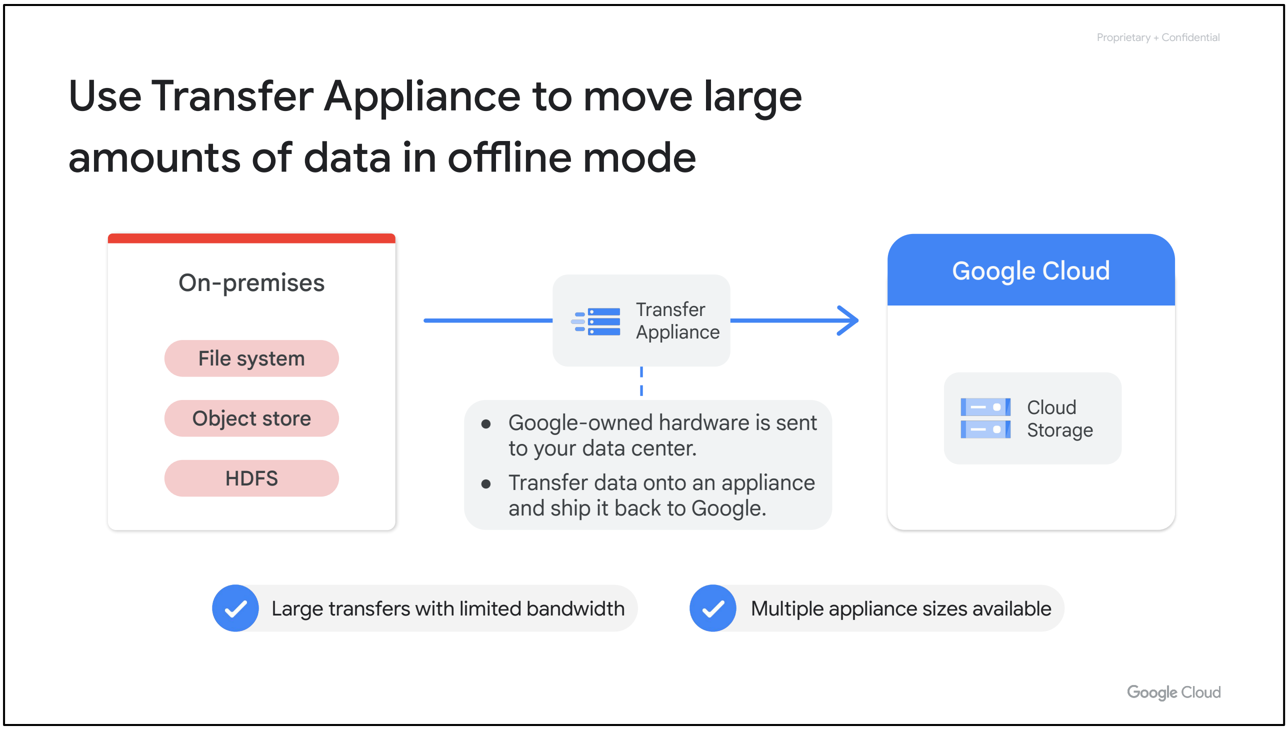
Storage Transfer Service efficiently moves large datasets from on-premises, multicloud file systems, object stores (including Amazon S3 and Azure Blob Storage), and HDFS into Cloud Storage.

It boasts high transfer speeds (up to tens of Giga bits per second) and supports scheduled transfers for convenient data migration.

**References:**

https://cloud.google.com/storage-transfer/docs/sources-and-sinks
https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#storage-transfer-service-for-large-transfers-of-on-premises-data

Storage transfer service scaling limitations (several 10s of Gbps):
https://cloud.google.com/storage-transfer/docs/known-limitations-transfer#scaling



Proprietary + Confidential

## Use Transfer Appliance to move large amounts of data in offline mode

On-premises

- File system
- Object store
- HDFS

Transfer Appliance

- Google-owned hardware is sent to your data center.
- Transfer data onto an appliance and ship it back to Google.

Google Cloud

Cloud Storage

✓ Large transfers with limited bandwidth

✓ Multiple appliance sizes available

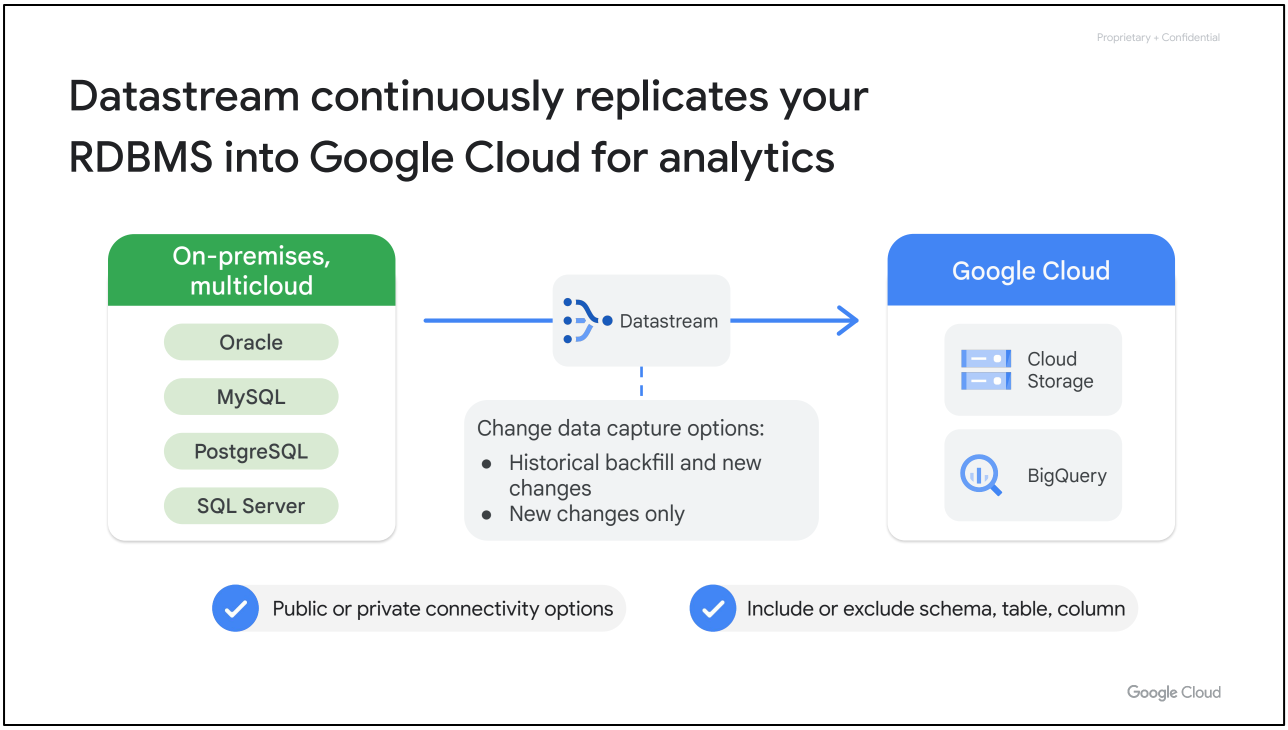
Transfer Appliance is Google's solution for moving massive datasets offline.

Google provides the hardware, you transfer your data onto it, then ship it back.

It is ideal for scenarios with limited bandwidth or very large transfers, and it comes in multiple sizes to suit your needs.

**References:**

https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#transfer_appliance_for_larger_transfers



Proprietary + Confidential

## Datastream continuously replicates your RDBMS into Google Cloud for analytics

Change data capture options:

- Historical backfill and new changes
- New changes only

✓ Public or private connectivity options

✓ Include or exclude schema, table, column

Google Cloud

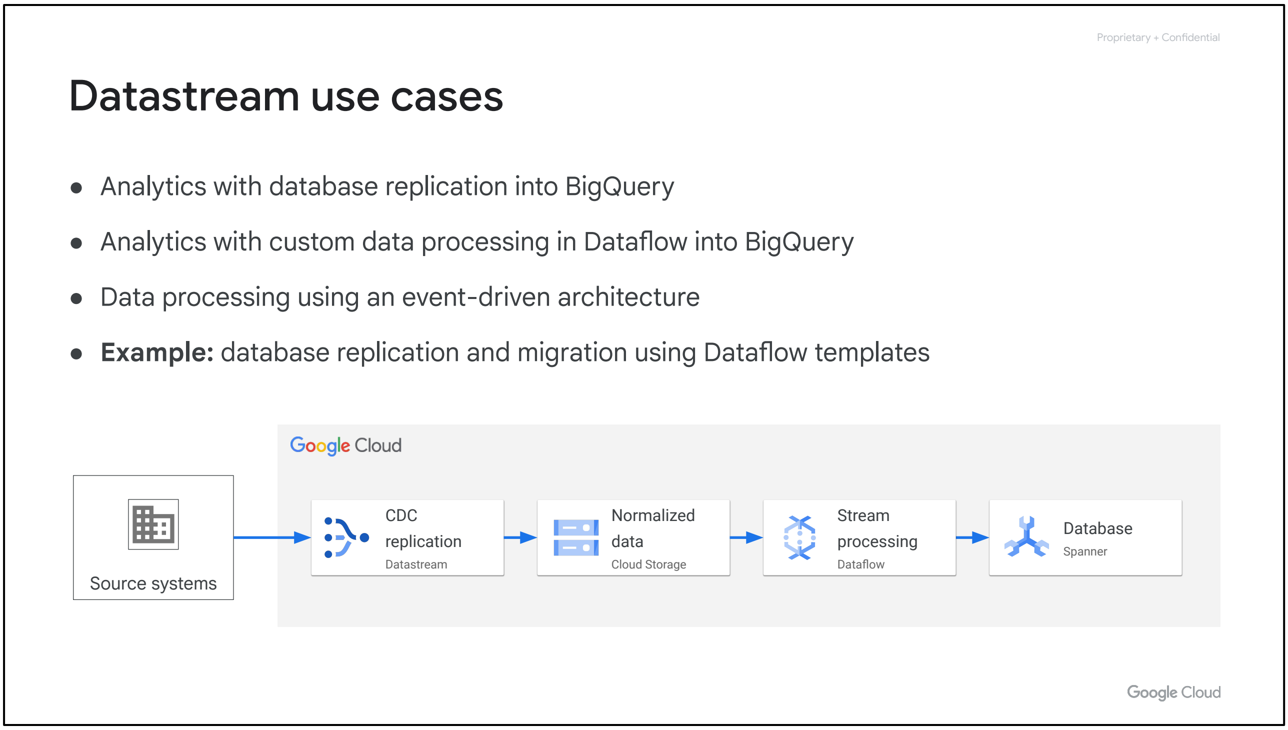
Datastream enables continuous replication of your on-premises or multicloud relational databases such as Oracle, MySQL, PostgreSQL, or SQL Server into Google Cloud.

Datastream offers change data capture options for historical backfill or allows you to just propagate new changes, with data landing in Cloud Storage or BigQuery for analytics.

You have flexibility in connectivity options and can selectively replicate data at the schema, table, or column level.

**References:**

https://cloud.google.com/datastream/docs/sources
https://cloud.google.com/datastream/docs/network-connectivity-options



Proprietary + Confidential

# Datastream use cases

- Analytics with database replication into BigQuery
- Analytics with custom data processing in Dataflow into BigQuery
- Data processing using an event-driven architecture
- **Example:** database replication and migration using Dataflow templates

Google Cloud

Datastream enables real-time data replication from source systems for various use cases.

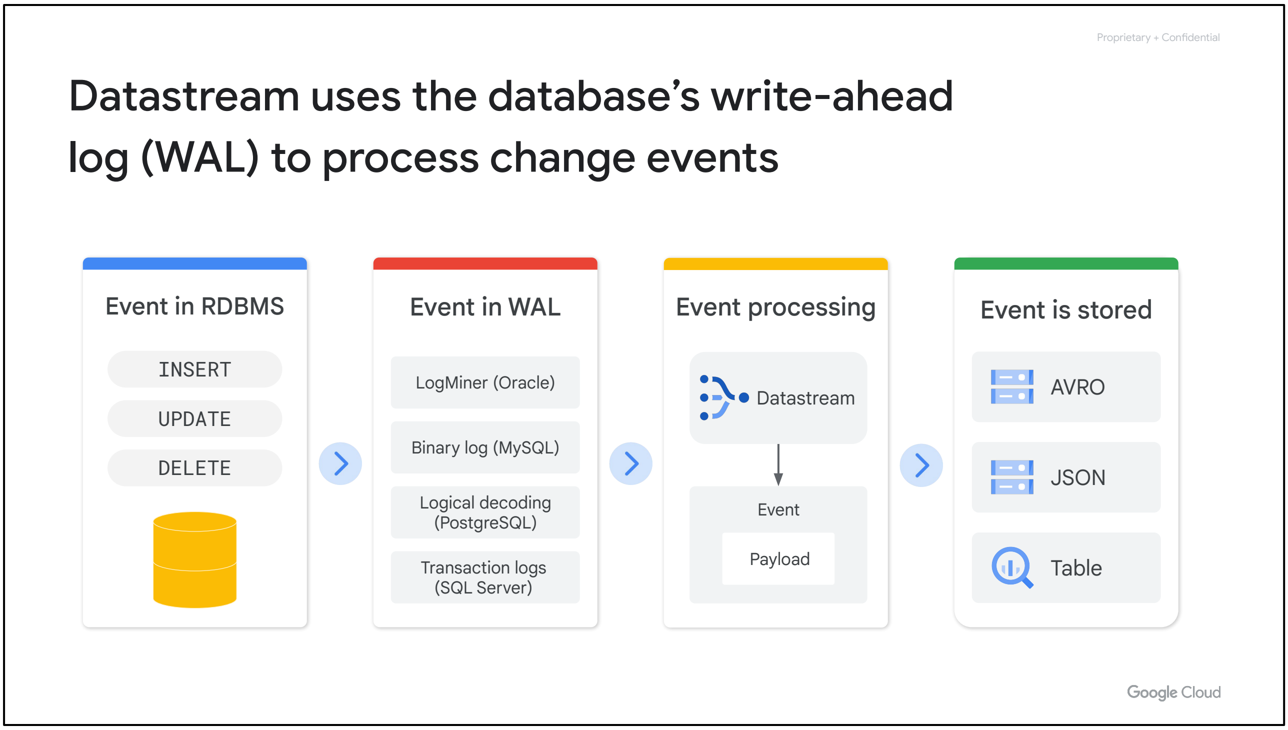
It supports direct replication into BigQuery for analytics, allows custom data processing in Dataflow before loading into BigQuery, and facilitates event-driven architectures.

Additionally, Datastream can be used with Dataflow templates for seamless database replication and migration tasks, making it a versatile tool for integrating data into Google Cloud.

**Instructor notes:**

- Use Datastream:
  - fully managed solution to replicate data from transactional databases into BigQuery
  - with Dataflow for a customizable solution to replicate data from transactional databases into BigQuery (Datastream > Cloud Storage > Dataflow > BigQuery)
  - to ingest changes from multiple sources into object stores like Google Cloud Storage or, in the future, messaging queues such as Pub/Sub or Kafka for building event-driven architectures

  - with pre-built Dataflow templates to continuously replicate and synchronize database data into Cloud SQL for PostgreSQL or Cloud Spanner to power low-downtime database migration or hybrid-cloud configuration
- Dataflow templates for Datastream:
  - https://cloud.google.com/dataflow/docs/guides/templates/provided/datastream-to-bigquery
  - https://cloud.google.com/dataflow/docs/guides/templates/provided/datastream-to-cloud-spanner



Proprietary + Confidential

## Datastream uses the database's write-ahead log (WAL) to process change events

Event in RDBMS

INSERT

UPDATE

DELETE

Event in WAL

LogMiner (Oracle)

Binary log (MySQL)

Logical decoding (PostgreSQL)

Transaction logs (SQL Server)

Event processing

Datastream

Event

Payload

Event is stored

AVRO

JSON

Table

Google Cloud

Datastream taps into the source database's write-ahead log (WAL) to capture and process changes for propagation downstream.

Datastream supports reading the logging mechanisms for the specific source database such as LogMiner for Oracle, binary log for MySQL, PostgreSQL's logical decoding, and transaction logs from SQL Server.

These change events such as inserts, updates, and deletes are then processed by Datastream and transformed into structured formats like AVRO or JSON, ready for storage in Google Cloud, typically in BigQuery tables, enabling near real-time data replication for analytics and other use cases.

**Instructor notes:**

Step 2: Event in WAL:

- In a transactional database, the set of operations that the database is going to perform is usually written to a write-ahead log (WAL) before any operations are executed on the storage engine
  - https://cloud.google.com/datastream/docs/behavior-overview#transactional_databases
- For MySQL, Datastream processes the MySQL binary log to extract change

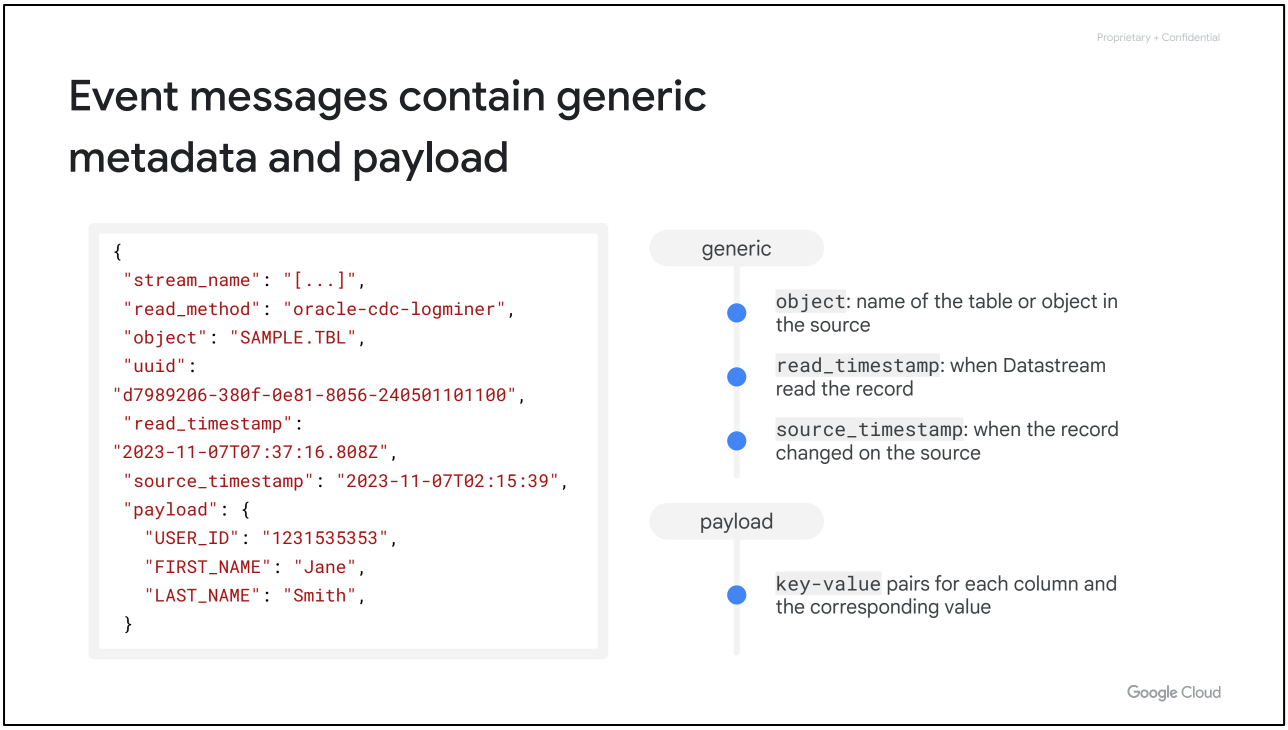
- events.
- For Oracle, Datastream uses LogMiner and supplemental logging settings to extract data from Oracle's redo logs.
- For PostgreSQL and AlloyDB for PostgreSQL, Datastream relies on PostgreSQL's logical decoding feature. Logical decoding exposes all changes committed to the database and allows consuming and processing these changes.
- For SQL Server, Datastream tracks data manipulation language (DML) changes using transaction logs.
    - https://cloud.google.com/datastream/docs/faq#general-source

Phase 3: Event is processed:

- event-sourcing: every change to a state of an application is captured in an event object
- Events represent every data manipulation language (DML) change for a given object.
- Objects represent a sub-portion of a stream. For instance, a database stream has a data object for every table being streamed.

Step 4: Event is stored:

- Cloud Storage: Datastream supports two output formats: Avro and JSON
    - https://cloud.google.com/datastream/docs/create-a-stream#cloud-storage-destination
- BigQuery: https://cloud.google.com/datastream/docs/create-a-stream#bigquery-destination



# Event messages contain generic metadata and payload

```
{
 "stream_name": "[...]",
 "read_method": "oracle-cdc-logminer",
 "object": "SAMPLE.TBL",
 "uuid":
"d7989206-380f-0e81-8056-240501101100",
 "read_timestamp":
"2023-11-07T07:37:16.808Z",
 "source_timestamp": "2023-11-07T02:15:39",
 "payload": {
   "USER_ID": "1231535353",
   "FIRST_NAME": "Jane",
   "LAST_NAME": "Smith",
 }
```

generic

- `object`: name of the table or object in the source
- `read_timestamp`: when Datastream read the record
- `source_timestamp`: when the record changed on the source

payload

- `key-value` pairs for each column and the corresponding value

Datastream event messages contain two main sections: generic metadata and payload.

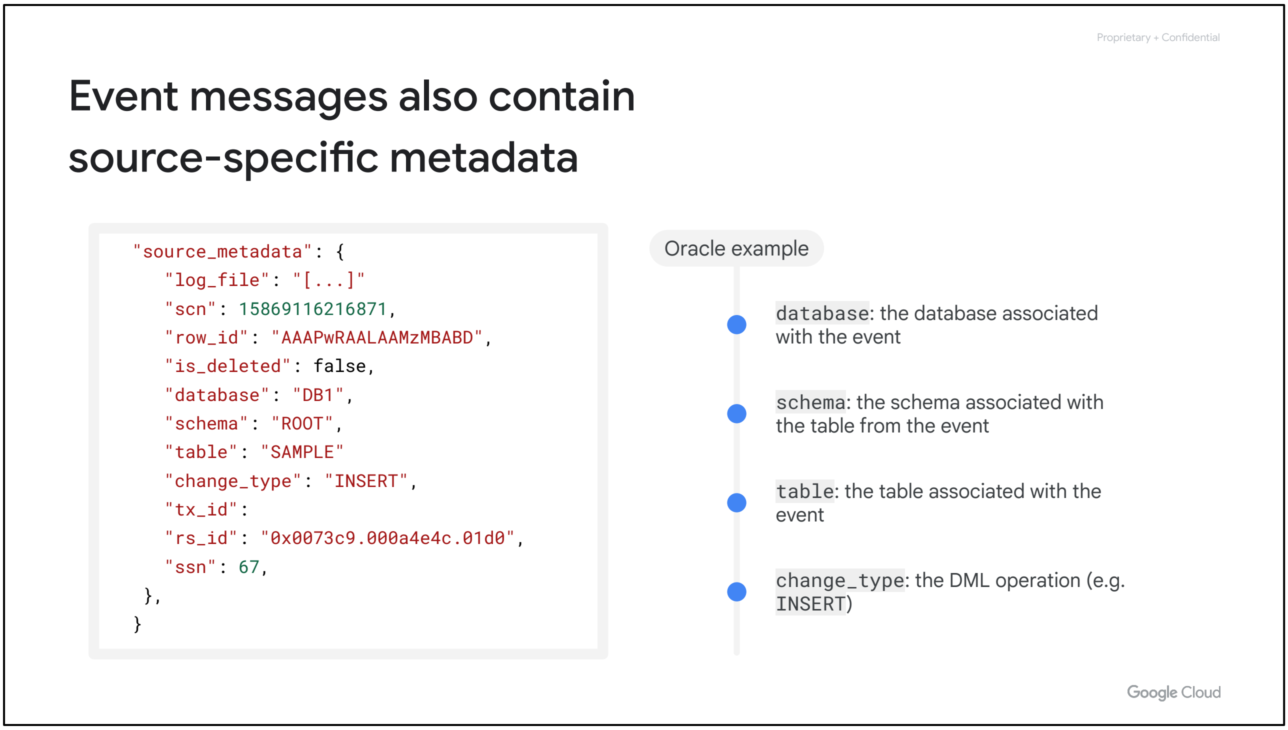
Metadata provides context about the data, like source table, timestamps, and related information.

Payload contains the actual data changes in a key-value format, reflecting column names and their corresponding values.

This structure allows for efficient and organized data replication and tracking of changes.

**References:**

https://cloud.google.com/datastream/docs/events-and-streams#genericmetadata
https://cloud.google.com/datastream/docs/events-and-streams#insert_t0



Proprietary + Confidential

## Event messages also contain source-specific metadata

```
"source_metadata": {
   "log_file": "[...]"
   "scn": 15869116216871,
   "row_id": "AAAPwRAALAAMzMBABD",
   "is_deleted": false,
   "database": "DB1",
   "schema": "ROOT",
   "table": "SAMPLE"
   "change_type": "INSERT",
   "tx_id":
   "rs_id": "0x0073c9.000a4e4c.01d0",
   "ssn": 67,
 },
}
```

Oracle example

- `database`: the database associated with the event
- `schema`: the schema associated with the table from the event
- `table`: the table associated with the event
- `change_type`: the DML operation (e.g. `INSERT`)

Google Cloud

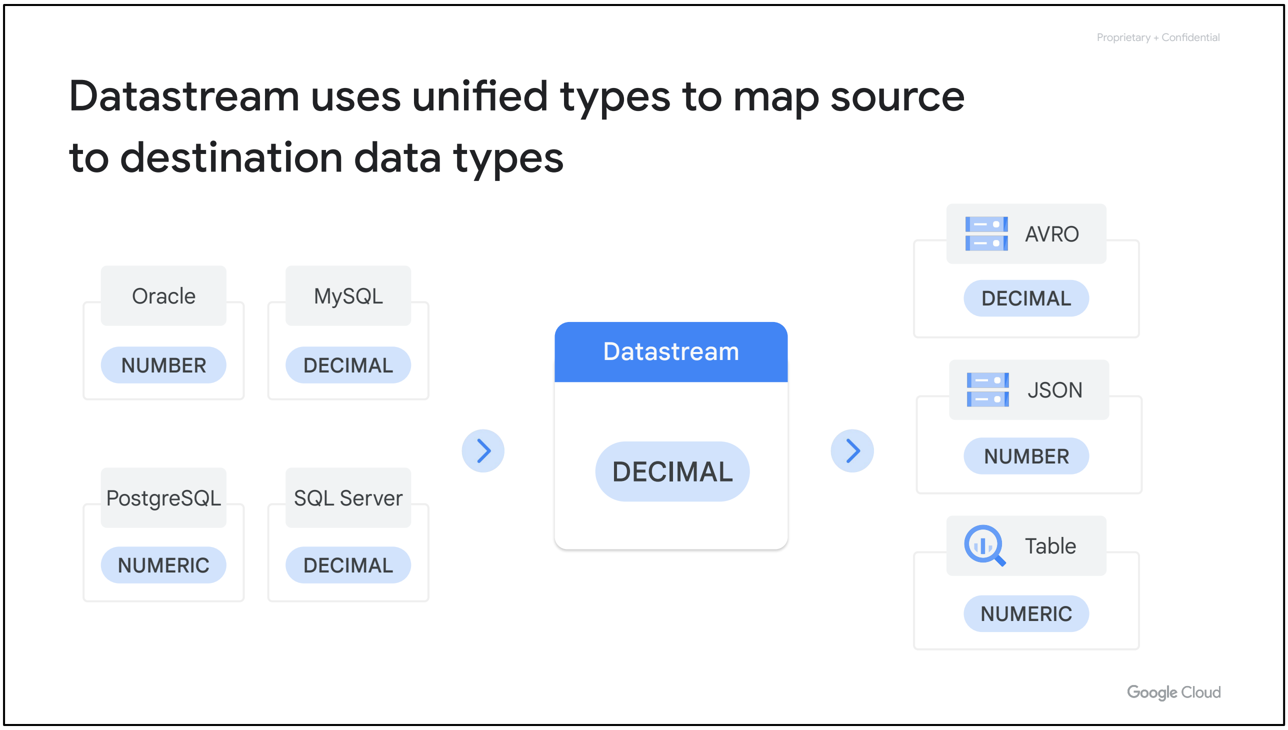
Datastream event messages also include source-specific metadata in addition to generic metadata and payload.

This metadata provides context about the data's origin within the source system, including details like the database name, schema, table, change type (such as INSERT), and other system-specific identifiers.

This additional information helps track data lineage and understand the context of changes replicated from the source database.

**References:**

https://cloud.google.com/datastream/docs/events-and-streams#sourcespecificmetadata



Proprietary + Confidential

# Datastream uses unified types to map source to destination data types

Oracle
NUMBER

MySQL
DECIMAL

PostgreSQL
NUMERIC

SQL Server
DECIMAL

Datastream
DECIMAL

AVRO
DECIMAL

JSON
NUMBER

Table
NUMERIC

Google Cloud

Datastream simplifies data replication by using unified data types to map between different source and destination databases.

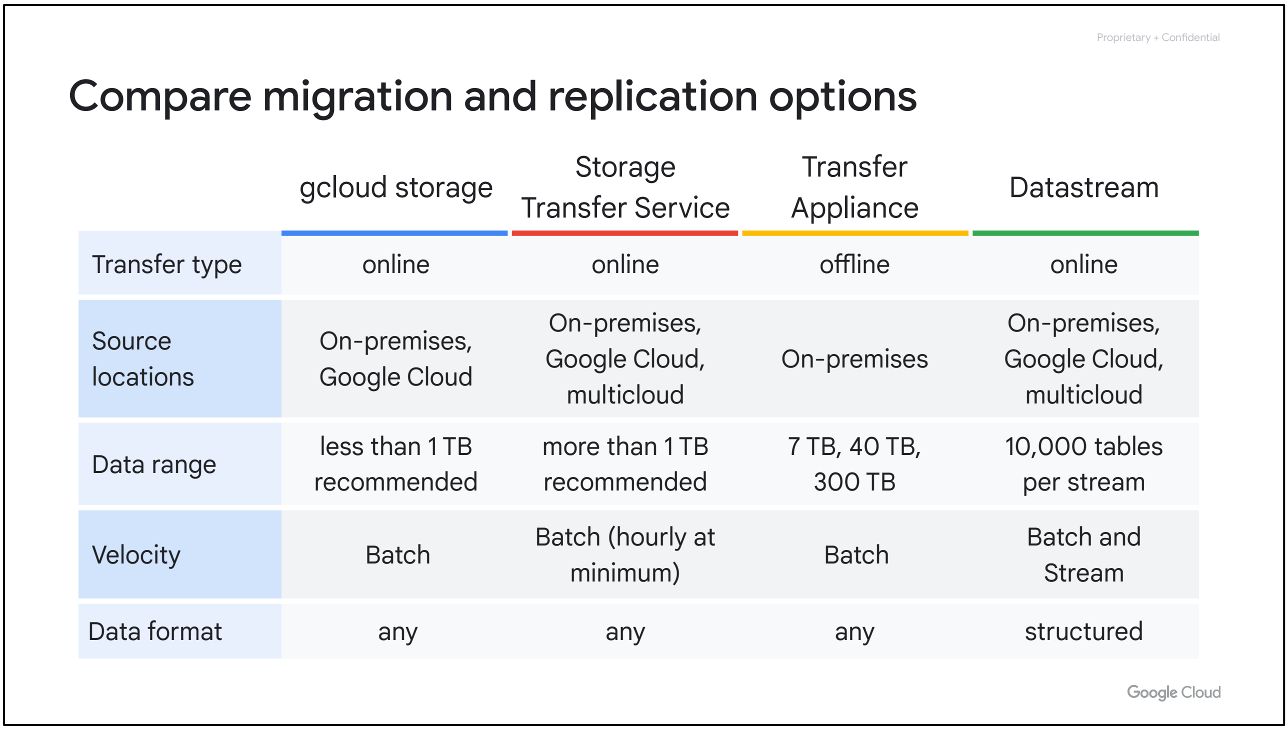
This means that regardless of whether your source data is in Oracle as number, MySQL as decimal, PostgreSQL as numeric, or SQL Server as decimal, Datastream will consistently represent it as decimal during replication.

When this data lands in Google Cloud, it can be further transformed into format-specific data types in different file types or destinations, such as Avro as decimal, JSON as number, or stored natively in BigQuery tables as numeric.

This ensures data type consistency and compatibility across different database systems, streamlining the data replication process.

**References:**

https://cloud.google.com/datastream/docs/unified-types
https://cloud.google.com/datastream/docs/destination-bigquery#map-data-types



## Compare migration and replication options

| | gcloud storage | Storage Transfer Service | Transfer Appliance | Datastream |
|---|---|---|---|---|
| Transfer type | online | online | offline | online |
| Source locations | On-premises, Google Cloud | On-premises, Google Cloud, multicloud | On-premises | On-premises, Google Cloud, multicloud |
| Data range | less than 1 TB recommended | more than 1 TB recommended | 7 TB, 40 TB, 300 TB | 10,000 tables per stream |
| Velocity | Batch | Batch (hourly at minimum) | Batch | Batch and Stream |
| Data format | any | any | any | structured |

Google Cloud

In summary, Google Cloud offers several data migration and replication options. The “gcloud storage” command is suitable for smaller, online transfers. Storage Transfer Service handles larger online transfers efficiently. Transfer Appliance is ideal for massive offline data migrations. And Datastream provides continuous, online replication of structured data, supporting both batch and streaming velocities.

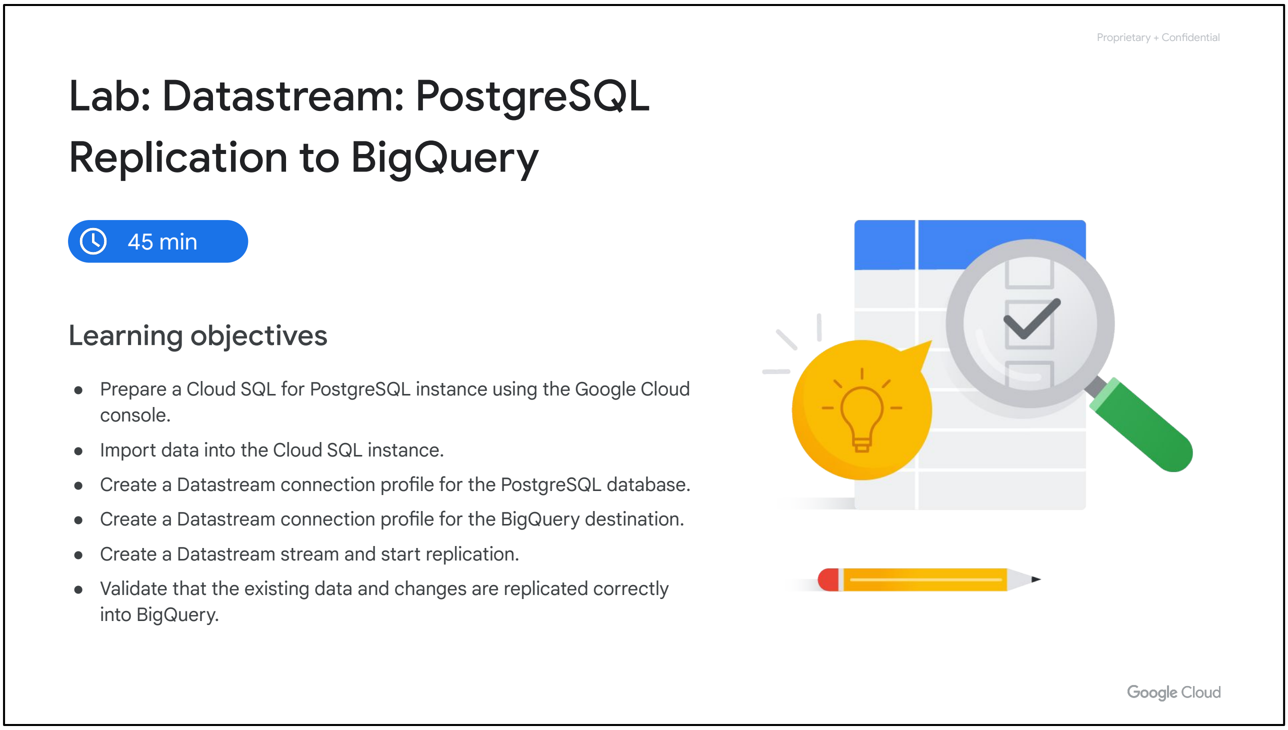
Choose the option that best fits your data size, transfer type, and data availability requirements.

**References:**

gcloud storage vs. Storage Transfer service limits:
https://cloud.google.com/storage-transfer/docs/transfer-options#move_data_between_buckets

Storage transfer service known limitations (no sub-hourly support):
https://cloud.google.com/storage-transfer/docs/known-limitations-transfer#real-time

Transfer appliance capacities:
https://cloud.google.com/transfer-appliance/docs/4.0/specifications#ta-weights



# Lab: Datastream: PostgreSQL Replication to BigQuery

45 min

## Learning objectives

- Prepare a Cloud SQL for PostgreSQL instance using the Google Cloud console.
- Import data into the Cloud SQL instance.
- Create a Datastream connection profile for the PostgreSQL database.
- Create a Datastream connection profile for the BigQuery destination.
- Create a Datastream stream and start replication.
- Validate that the existing data and changes are replicated correctly into BigQuery.

In this lab, you use Datastream to replicate data from PostgreSQL to BigQuery. You prepare and load a Cloud SQL for PostgreSQL instance. You create Datastream connection profiles for the source and target. You then create a Datastream processing stream and start replication. Finally, you validate the replication in BigQuery.

Lab URL: https://www.cloudskillsboost.google/catalog_lab/5777

Google Cloud