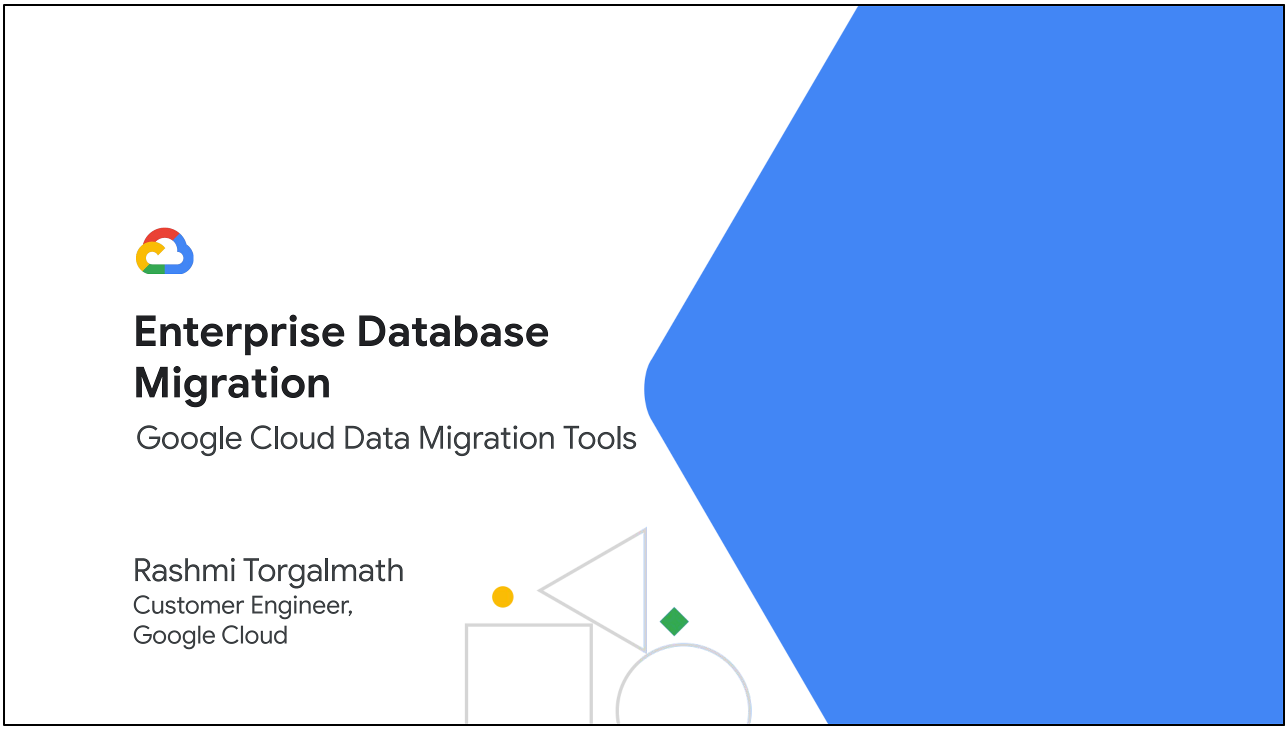
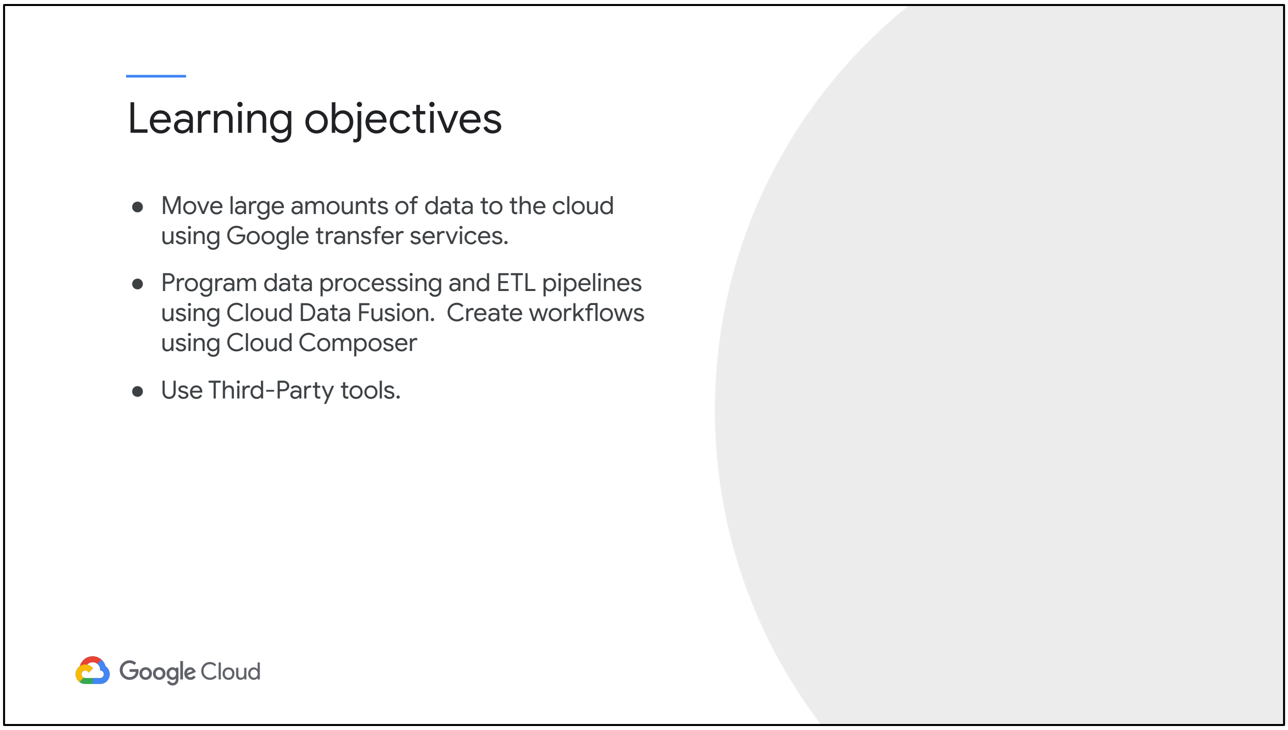


Enterprise Database Migration

Google Cloud Data Migration Tools

Rashmi Torgalmath
Customer Engineer,
Google Cloud

Hello everyone, Rashmi here. If you just have a few GBs of data in a SQL backup file that you want to restore in a database in Google Cloud, you can probably skip this module. But life is not always that simple. If you have TBs or even PBs of data, getting it in the cloud is more complicated. If you want to alter the data before loading it into the new database, that can be complicated. Or you may be moving from one type of database to another. That requires some work as well. Google and other third parties have tools to help when things get difficult.



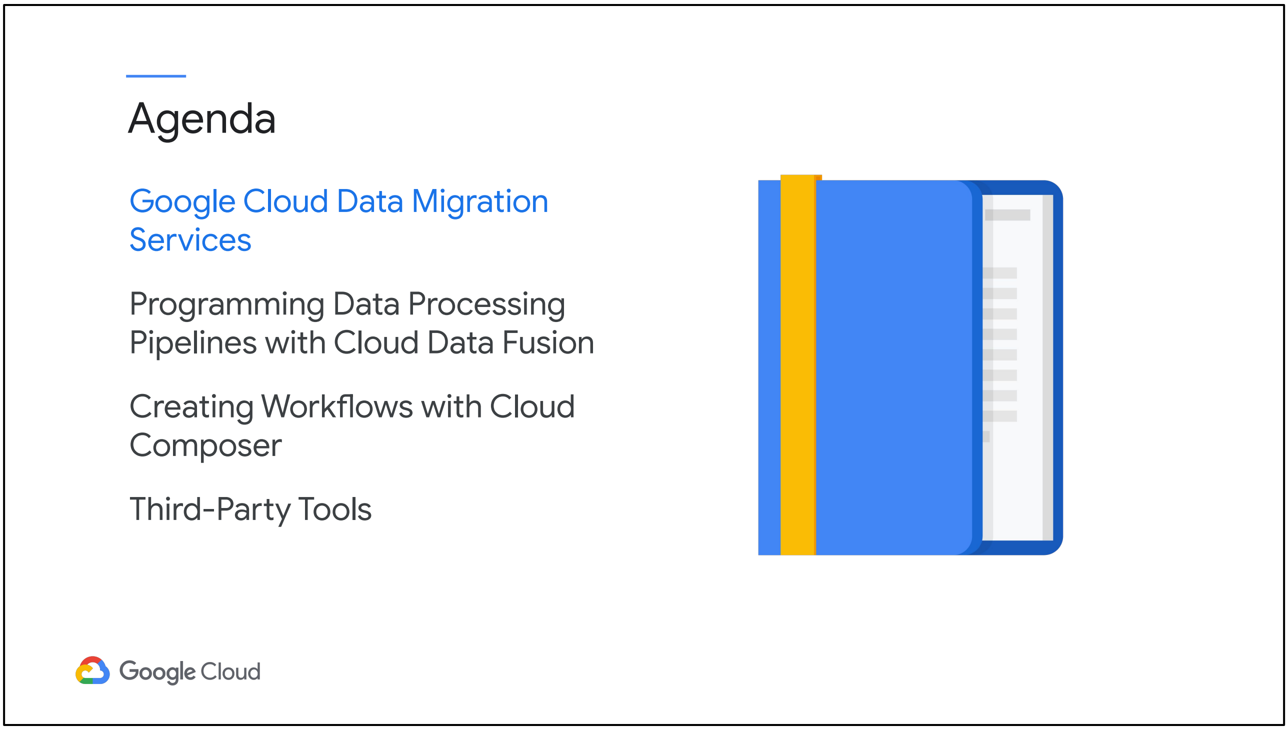
Learning objectives

- Move large amounts of data to the cloud using Google transfer services.
- Program data processing and ETL pipelines using Cloud Data Fusion. Create workflows using Cloud Composer
- Use Third-Party tools.

In this module, you learn to move large amounts of data to the cloud using Google transfer services.

You also learn to program data processing and ETL pipelines using Cloud Data Fusion and workflows using Cloud Composer.

You also automate a database migration using a third-party tool called Striim.



Agenda

Google Cloud Data Migration Services

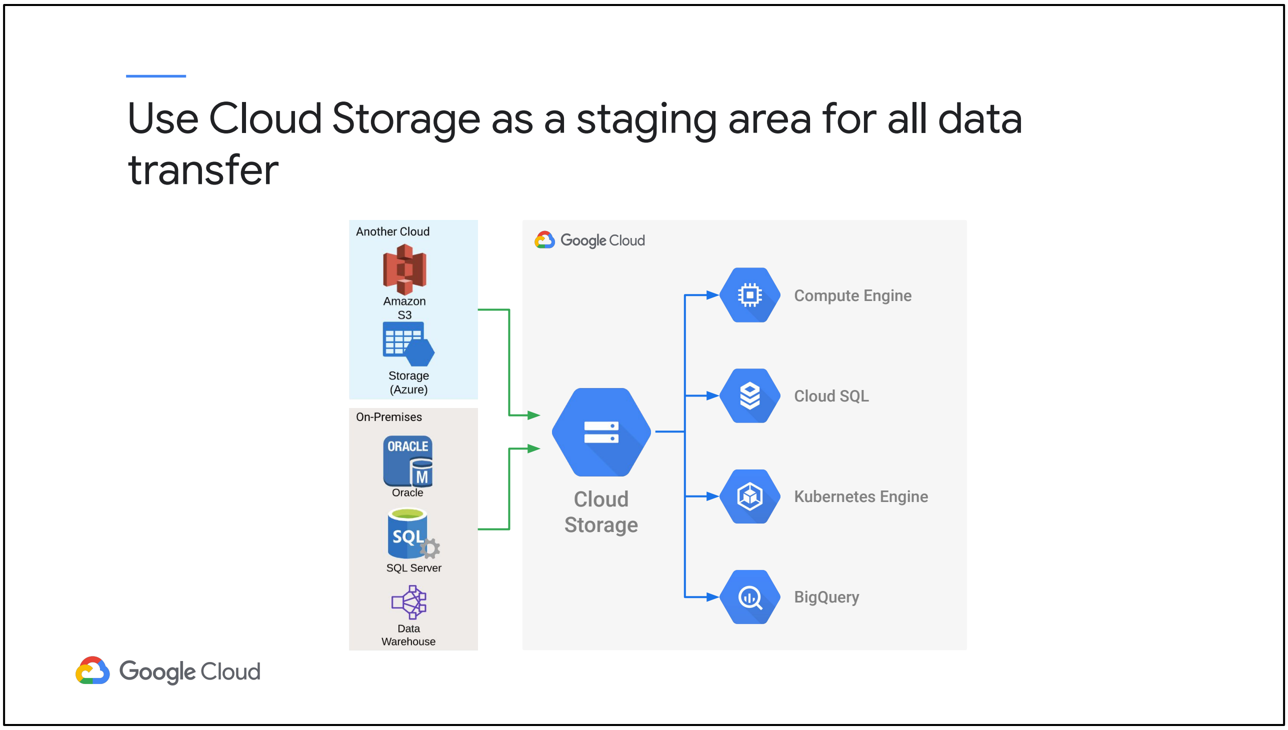
Programming Data Processing Pipelines with Cloud Data Fusion

Creating Workflows with Cloud Composer

Third-Party Tools

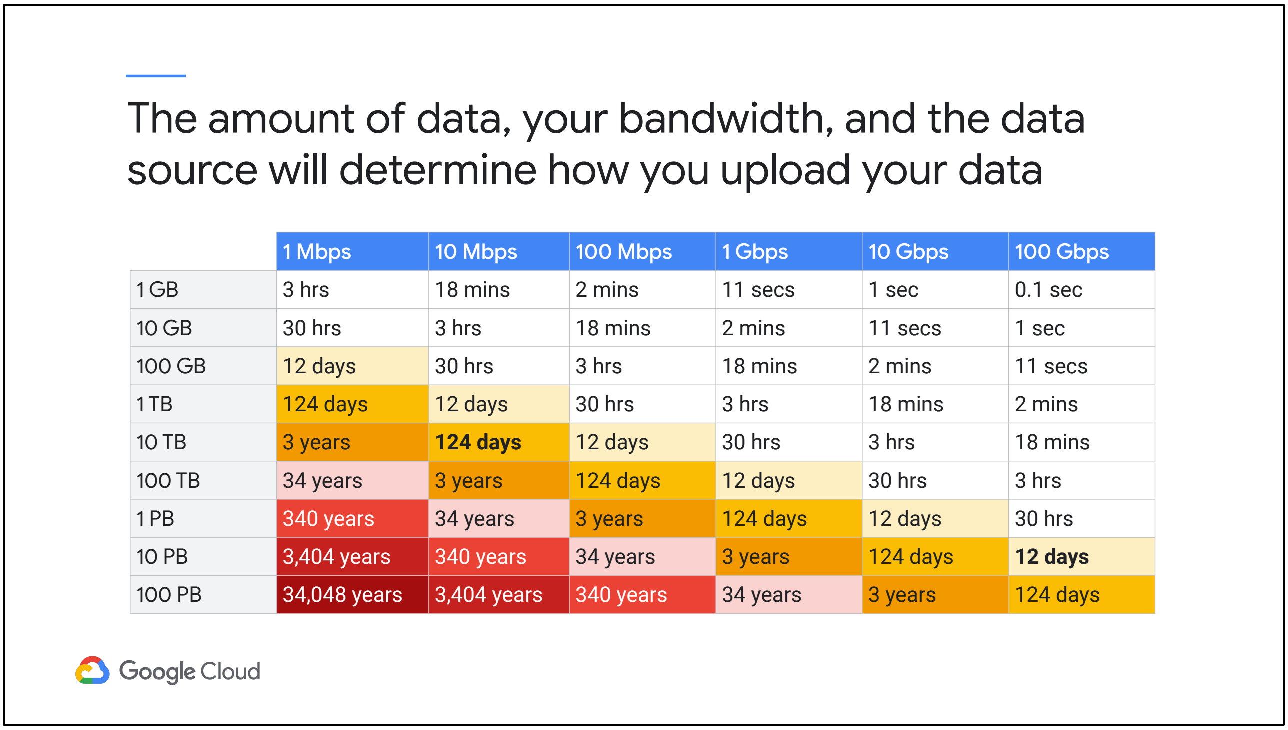
Google Cloud

Let's start with data migration.



Use Cloud Storage as a staging area for all data transfer

When you have a small amount of data, you can load it directly from your environment into the target database. For large amounts of data, you should use Cloud Storage as a staging area. After you get the data into Cloud Storage, you can then move it around at Google speed. Also, many Google Cloud tools are designed to work with Cloud Storage. For example, if you want to restore a database into Cloud SQL, it asks you to specify the location of the backup in a Cloud Storage bucket.



The amount of data, your bandwidth, and the data source will determine how you upload your data

	1 Mbps	10 Mbps	100 Mbps	1 Gbps	10 Gbps	100 Gbps
1 GB	3 hrs	18 mins	2 mins	11 secs	1 sec	0.1 sec
10 GB	30 hrs	3 hrs	18 mins	2 mins	11 secs	1 sec
100 GB	12 days	30 hrs	3 hrs	18 mins	2 mins	11 secs
1 TB	124 days	12 days	30 hrs	3 hrs	18 mins	2 mins
10 TB	3 years	**124 days**	12 days	30 hrs	3 hrs	18 mins
100 TB	34 years	3 years	124 days	12 days	30 hrs	3 hrs
1 PB	340 years	34 years	3 years	124 days	12 days	30 hrs
10 PB	3,404 years	340 years	34 years	3 years	124 days	**12 days**
100 PB	34,048 years	3,404 years	340 years	34 years	3 years	124 days

Google Cloud

Transferring data into Google Cloud is not a trivial operation. Multiple factors must be considered, including cost, time, offline versus online transfer options, transfer tools and technologies, and security and privacy.

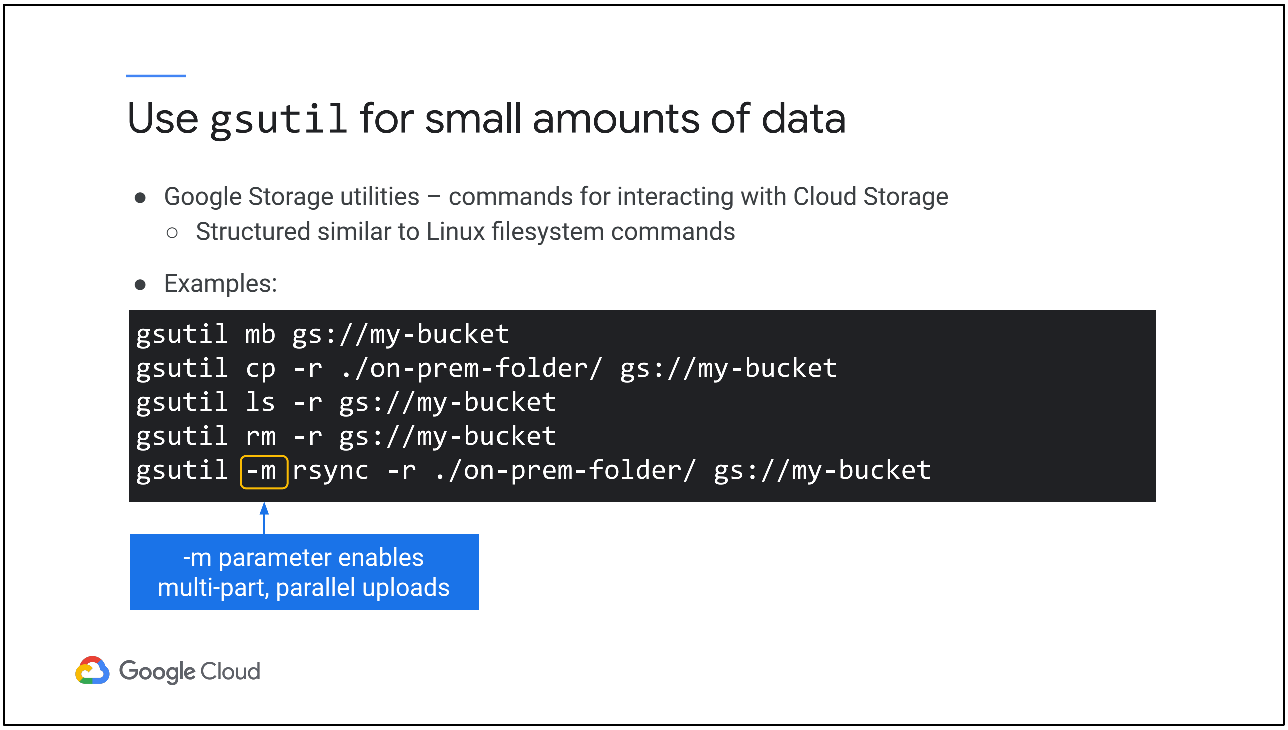
Although transfer into Google Cloud is free, there will be costs with the storage of the data, possibly hardware costs, and possibly egress costs if transferring from another cloud provider.

If you have huge datasets, the time required for transfer across a network may be unrealistic. Even if it is realistic, the effects on your organization's infrastructure may be damaging while the transfer is taking place. This needs to be considered. The table above shows the challenge of moving large data sets.

You have to decide whether to transfer data over the network or use a hardware appliance. Several tools are available to support network transfer when only low bandwidth is available, such as Storage Transfer Service. Google also offers a hardware solution known as Transfer Appliance. Here Google ships you hardware that you fill with data from your data center and ship back, where it is transferred to Cloud Storage. The data is encrypted until you choose to decrypt it.

Company policies might prevent transferring data over a public internet. In that case, Direct Peering or Cloud Interconnect are possible solutions.

There are other considerations, such as protecting the data at rest (authorization and access to the source and destination storage system), protecting data while in transit, and protecting access to the transfer product. The general approach is to have the data encrypted and access gained only via access keys.



Use `gsutil` for small amounts of data

- Google Storage utilities – commands for interacting with Cloud Storage
    - Structured similar to Linux filesystem commands
- Examples:

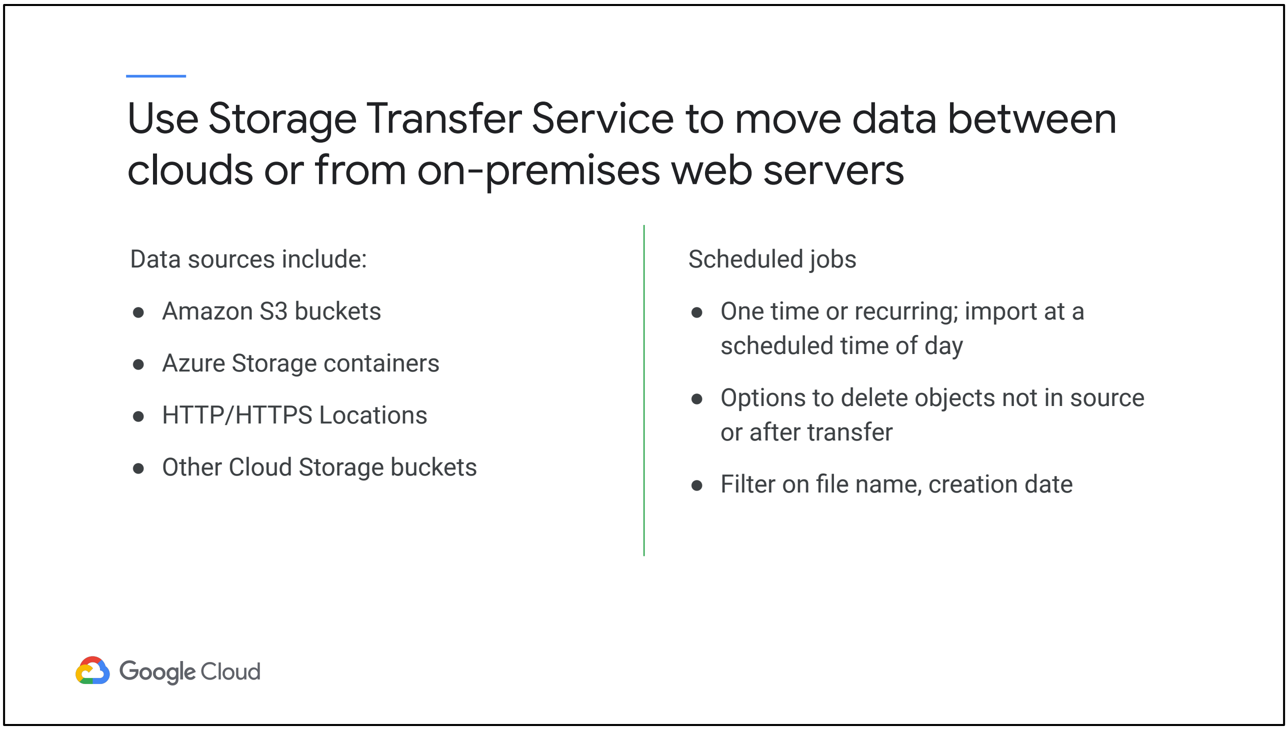
```
gsutil mb gs://my-bucket
gsutil cp -r ./on-prem-folder/ gs://my-bucket
gsutil ls -r gs://my-bucket
gsutil rm -r gs://my-bucket
gsutil -m rsync -r ./on-prem-folder/ gs://my-bucket
```

-m parameter enables multi-part, parallel uploads

Google Cloud

The `gcloud` CLI is not used for interacting with Cloud Storage. Use `gsutil` for that. `gsutil` commands are designed to look as much as possible like Linux file system commands.

To create a bucket, use the `mb` command.
To copy files, use `gsutil cp`.
To see files in a bucket, use `ls`; to delete files, use `rm`.
To synchronize a folder and a bucket, use the `rsync` command.
When you have very large files or a tremendous number of files, use the `-m` parameter to enable multi-part, parallel uploads.



Use Storage Transfer Service to move data between clouds or from on-premises web servers

Data sources include:

- Amazon S3 buckets
- Azure Storage containers
- HTTP/HTTPS Locations
- Other Cloud Storage buckets

Scheduled jobs

- One time or recurring; import at a scheduled time of day
- Options to delete objects not in source or after transfer
- Filter on file name, creation date

Google Cloud

- Storage Transfer Service is a product that enables you to:

  Move or back up data to a Cloud Storage bucket either from other cloud storage providers or from your on-premises storage.
  Move data from one Cloud Storage bucket to another, so that it is available to different groups of users or applications.
  Periodically move data as part of a data processing pipeline or analytical workflow.

- Storage Transfer Service provides options that make data transfers and synchronization easier. For example, you can:

  Schedule one-time transfer operations or recurring transfer operations.
  Delete existing objects in the destination bucket if they don’t have a corresponding object in the source.
  Delete data source objects after transferring them.
  Schedule periodic synchronization from a data source to a data sink with advanced filters based on file creation dates, file-name filters, and the times of day you prefer to import data.

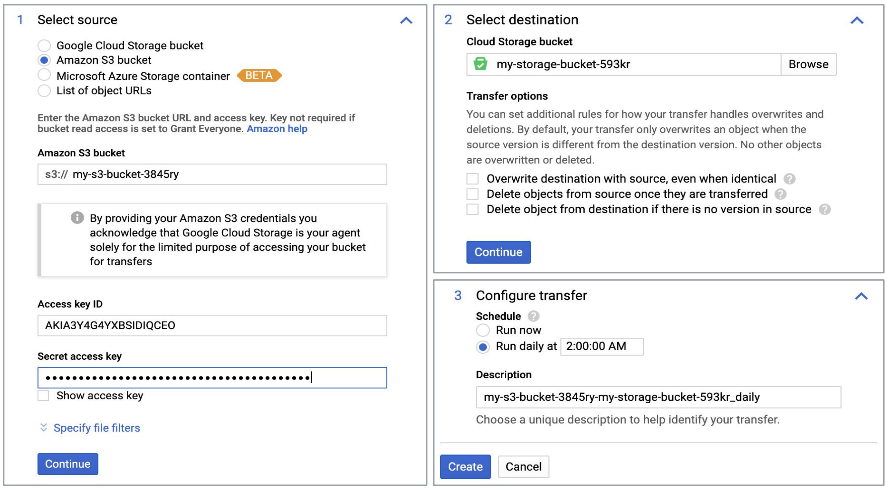
  The gsutil utility also allows transfer of data between Cloud Storage and other locations.

In addition, Google has Transfer service for on-premises data.

To help decide which tool to use, consider the following:

Transferring from another cloud storage provider: use Storage Transfer Service.
Transferring less than 1 TB from on-premises: use gsutil.
Transferring more than 1 TB from on-premises: use Transfer Service for on-premises data (beta).

Configuring the Storage Transfer Service



1 Select source

- ○ Google Cloud Storage bucket
- ● Amazon S3 bucket
- ○ Microsoft Azure Storage container BETA
- ○ List of object URLs

Enter the Amazon S3 bucket URL and access key. Key not required if bucket read access is set to Grant Everyone. Amazon help

Amazon S3 bucket: s3:// my-s3-bucket-3845ry

By providing your Amazon S3 credentials you acknowledge that Google Cloud Storage is your agent solely for the limited purpose of accessing your bucket for transfers

Access key ID: AKIA3Y4G4YXBSIDIQCEO

Secret access key: ••

□ Show access key

Specify file filters

Continue

2 Select destination

Cloud Storage bucket: my-storage-bucket-593kr Browse

Transfer options

You can set additional rules for how your transfer handles overwrites and deletions. By default, your transfer only overwrites an object when the source version is different from the destination version. No other objects are overwritten or deleted.

- □ Overwrite destination with source, even when identical
- □ Delete objects from source once they are transferred
- □ Delete object from destination if there is no version in source

Continue

3 Configure transfer

Schedule

- ○ Run now
- ● Run daily at 2:00:00 AM

Description: my-s3-bucket-3845ry-my-storage-bucket-593kr_daily

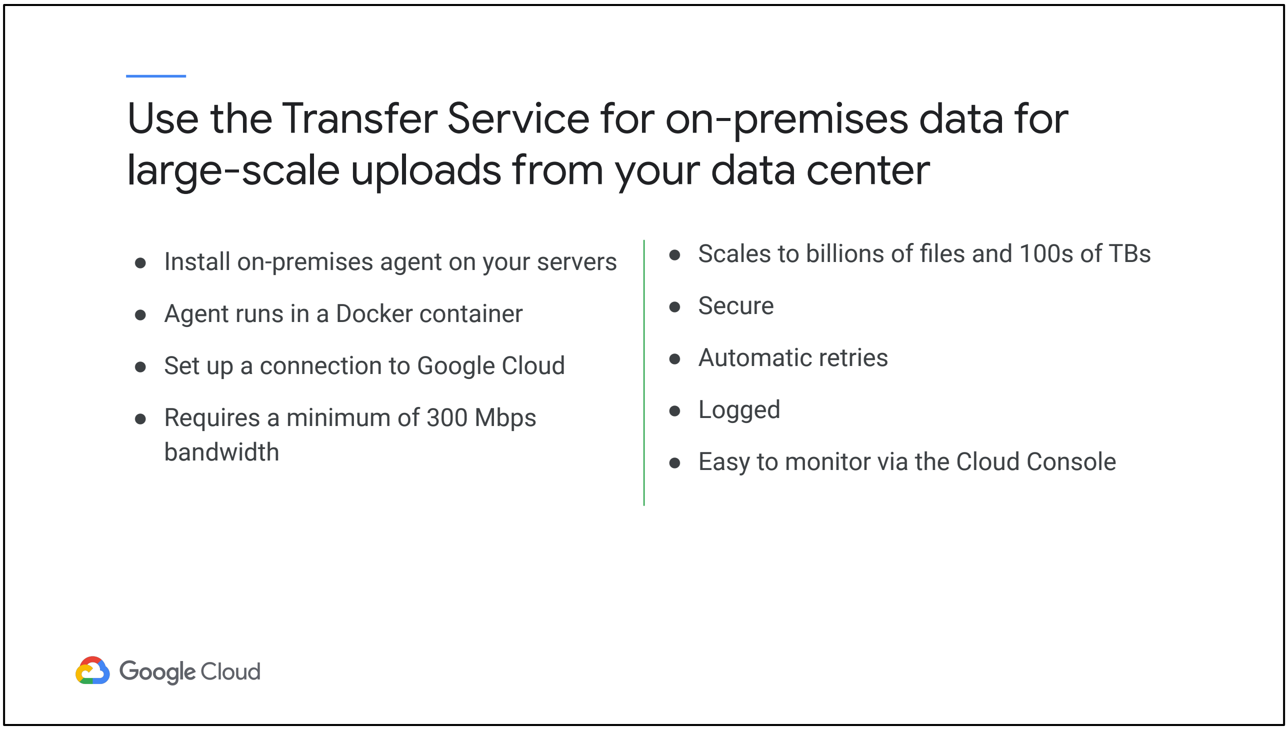
Choose a unique description to help identify your transfer.

Create Cancel

Google Cloud

Configuration and data transfer using Storage Transfer Service is easy.

- First, specify the source. The source can be another Cloud Storage bucket, an S3 bucket, Azure Storage, or an on-premises server that Transfer Service can access.
- Second, specify the bucket in which you want to transfer the data.
- Lastly, specify when you want to run the job. The job can be run immediately or scheduled for a later time. You can also schedule recurring jobs.



Use the Transfer Service for on-premises data for large-scale uploads from your data center

- Install on-premises agent on your servers
- Agent runs in a Docker container
- Set up a connection to Google Cloud
- Requires a minimum of 300 Mbps bandwidth

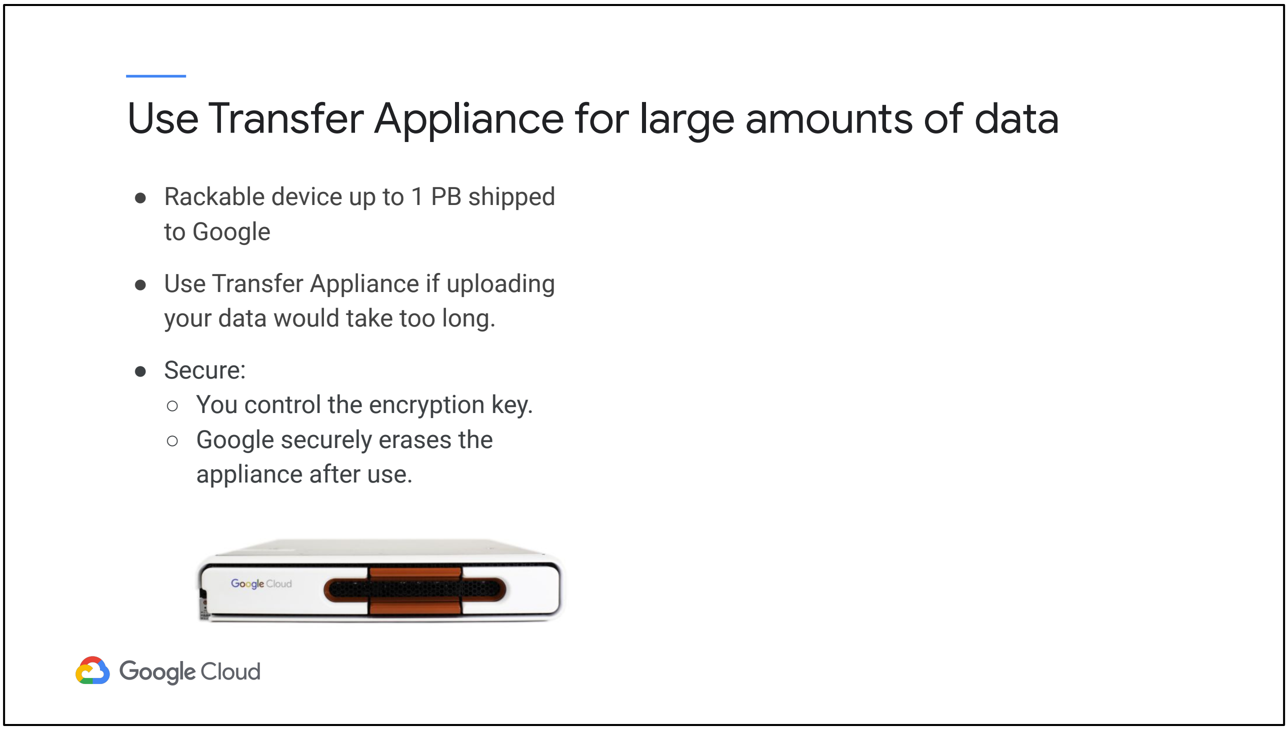
- Scales to billions of files and 100s of TBs
- Secure
- Automatic retries
- Logged
- Easy to monitor via the Cloud Console

Google Cloud

The Transfer Service for on-premises data allows large-scale online data transfers from on-premises storage to Cloud Storage. With this service, data validation, encryption, error retries, and fault tolerance are built in. On-premises software is installed—it comes as a Docker container—and then via the Cloud Console, directories to be transferred to Cloud Storage are selected. Once data transfer begins, the service will parallelize the transfer across many agents. Via the Cloud Console, a user can view detailed transfer logs as well as the creation, management, and monitoring of transfer jobs.

To use the Transfer Service for on-premises data, a Posix-compliant source and a network connection of at least 300Mbps are required. Also, a Docker-supported Linux server that can access the data to be transferred is required with ports 80 and 443 open for outbound connections.

The use case is for on-premises transfer of data greater in size than 1 TB.

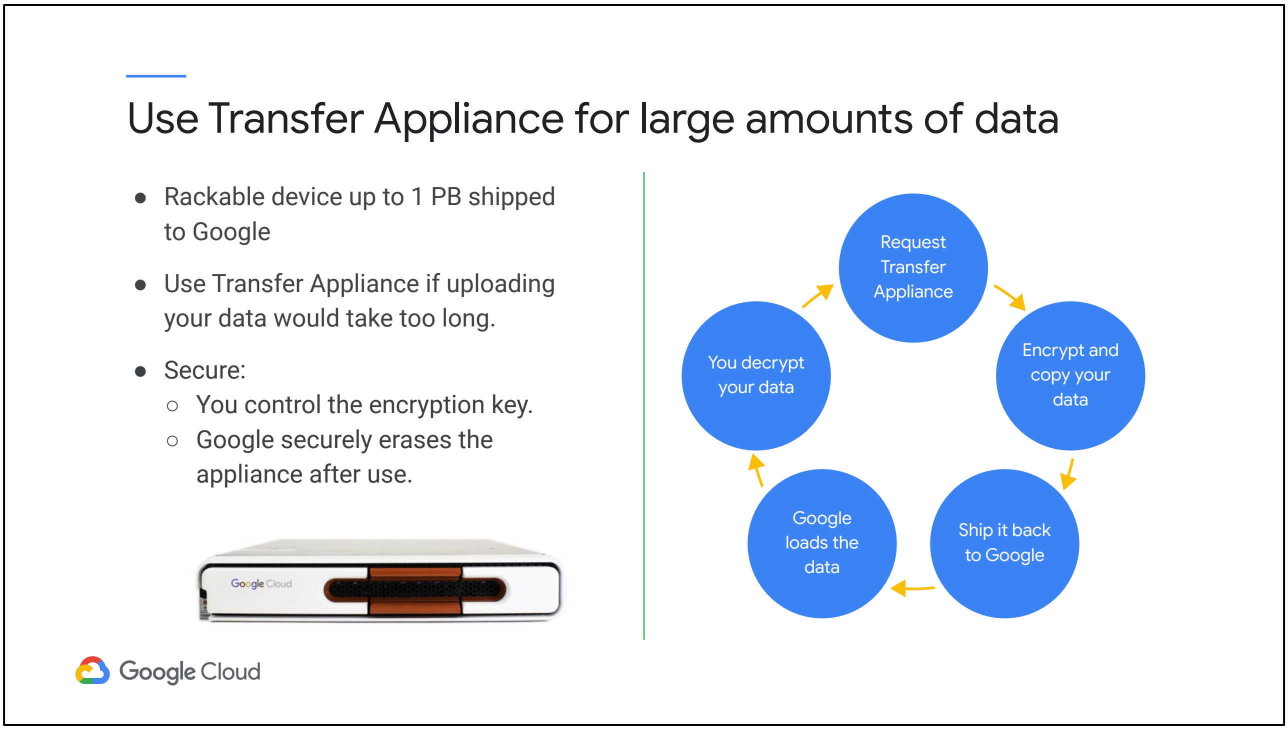


Use Transfer Appliance for large amounts of data

- Rackable device up to 1 PB shipped to Google
- Use Transfer Appliance if uploading your data would take too long.
- Secure:
  - You control the encryption key.
  - Google securely erases the appliance after use.

Transfer Appliance is a secure, rackable, high-capacity storage server that you set up in your data center. You fill it with data and ship it to an ingest location, where the data is uploaded to Cloud Storage. Data is encrypted automatically and remains safe until decrypted. Two sizes of appliance are available: 100 TB and 480 TB.

Transfer Appliance is easily mounted in rackspace in a data center and can be mounted as Network Attached Storage (NAS). A simple user interface is provided to guide users through local data capture, and the Cloud Console is used to decrypt and ingest data.



Use Transfer Appliance for large amounts of data

- Rackable device up to 1 PB shipped to Google
- Use Transfer Appliance if uploading your data would take too long.
- Secure:
  - You control the encryption key.
  - Google securely erases the appliance after use.

Google Cloud

Request Transfer Appliance

Encrypt and copy your data

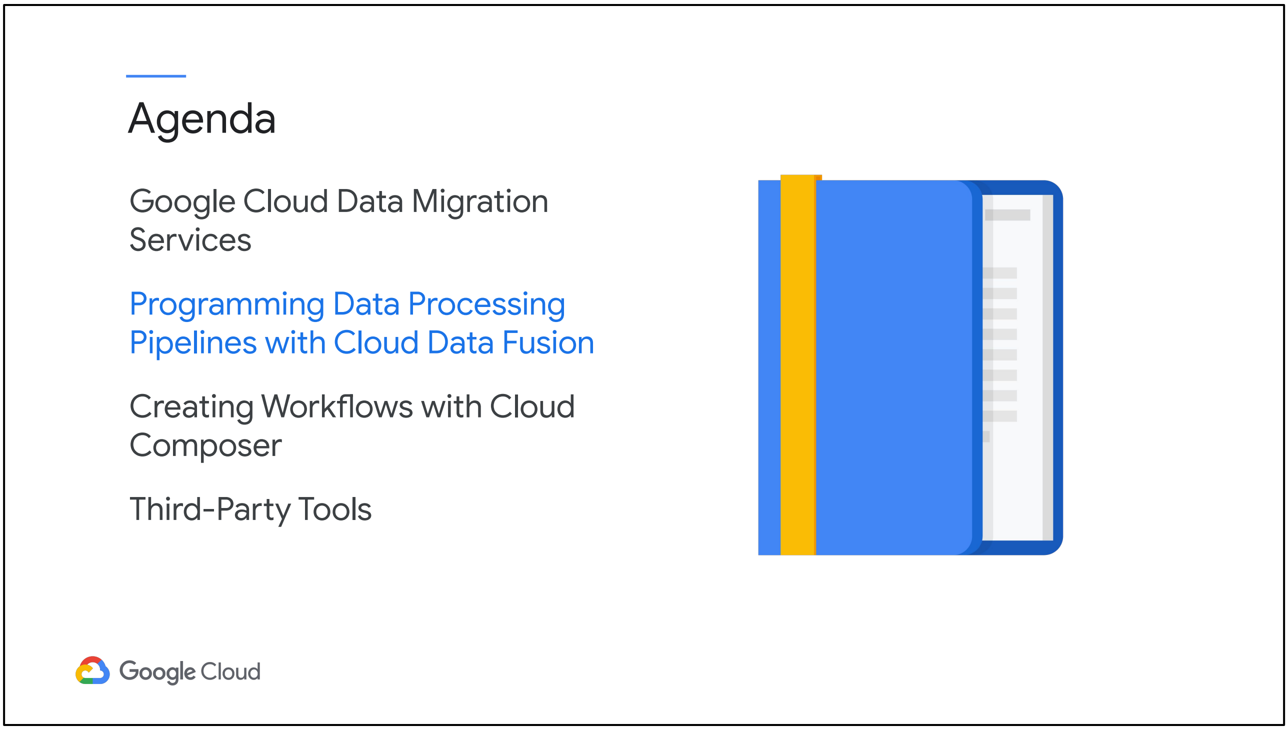
Ship it back to Google

Google loads the data

You decrypt your data

Google Cloud

The process for using Transfer Appliance is that you request an appliance, and it is shipped in a tamper-evident case. Data is transferred to the appliance. The appliance is shipped back to Google, data is loaded to Cloud Storage, and you are notified that it is available. Google uses tamper-evident seals on shipping cases to and from the data ingest site. Data is encrypted to AES256 standards at the moment of capture. After the transfer is complete, the appliance is erased per NIST-800-88 standards.



Agenda

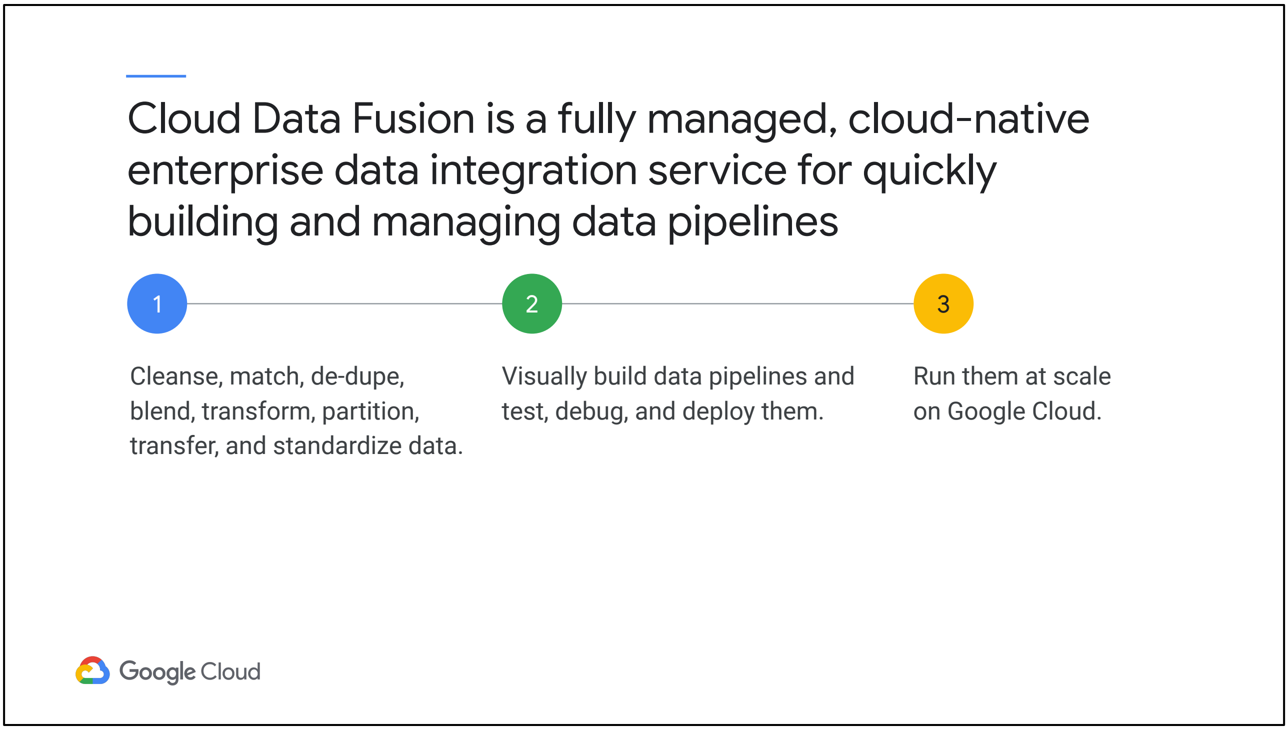
Google Cloud Data Migration Services

Programming Data Processing Pipelines with Cloud Data Fusion

Creating Workflows with Cloud Composer

Third-Party Tools

Data processing pipelines allow you to grab data from a source, alter it, and then save it to a sink (or target). This is sometimes called an ETL job (Extract, Transform, and Load). You can write a data processing pipeline in your favorite programming language, but there are also tools that can help make it easier. Cloud Data Fusion is one of those tools provided by Google Cloud.



Cloud Data Fusion is a fully managed, cloud-native enterprise data integration service for quickly building and managing data pipelines

1 Cleanse, match, de-dupe, blend, transform, partition, transfer, and standardize data.

2 Visually build data pipelines and test, debug, and deploy them.

3 Run them at scale on Google Cloud.

Cloud Data Fusion is a fully managed, cloud-native enterprise data integration service for quickly building and managing data pipelines.

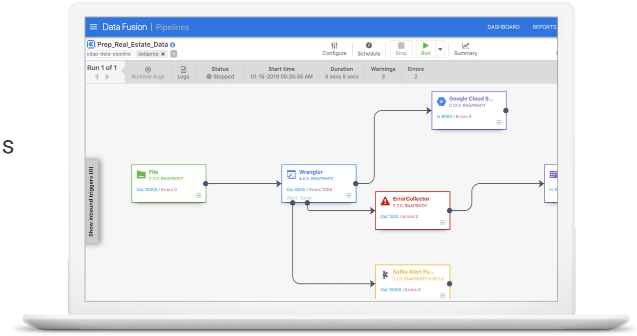
1. You can use it to cleanse, match, de-dupe, blend, transform, or partition data. It will automate the execution of the job and allow you to monitor job progress.

2. A visual user interface is available for drag-and-drop building of pipelines. You can also quickly test and debug pipelines with a small subset of the data. When you deploy pipelines, Cloud Data Fusion automatically provisions Google Cloud infrastructure to actually run the job.

3. There is deep integration with Google Cloud, so you can run your data processing jobs at scale. Cloud Data Fusion's integration with Google Cloud simplifies data security and ensures that data is immediately available for analysis. Whether you're curating a data lake with Cloud Storage and Dataproc, moving data into BigQuery for data warehousing, or transforming data to export it to a relational store like Cloud Spanner, Cloud Data Fusion's integration makes development and iteration fast and easy.

Build data pipelines with a friendly UI

- Rich graphical interface
- 100+ plugins: connectors, transforms, and actions
- Code-free visual transformations
- Test and debug pipelines
- Pre-built pipelines
- Developer SDK



Run 1 of 1

Status	Start time	Duration	Warnings	Errors
Stopped	01-15-2019 05:35:35 AM	3 mins 5 secs	3	2

Google Cloud

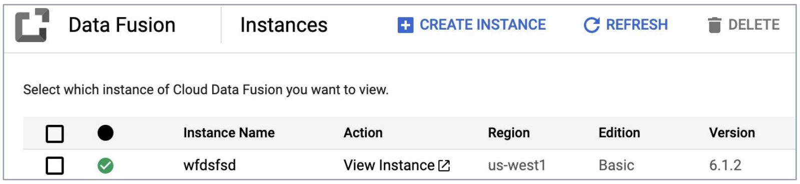
You build data pipelines with a friendly UI. The rich graphical interface allows for drag-and-drop visualization of pipelines. There are over a hundred built-in plugins, connectors, transforms, actions, and support for many legacy data sources.

The UI allows you to write pipeline without coding; you can test and debug pipelines, and pre-built pipelines are available to get you started.

For those who want to write code or for use cases where significant customization is required, there is also a developer SDK.

Cloud Data Fusion instances are managed environments for building pipelines

- Basic edition for development
  - $1.80 per hour
- Enterprise edition for production and streaming pipelines
  - $4.20 per hour
- Pipelines run on a Dataproc cluster that you are also charged for.



Select which instance of Cloud Data Fusion you want to view.

□	●	Instance Name	Action	Region	Edition	Version
□	✓	wfdsfsd	View Instance	us-west1	Basic	6.1.2

Google Cloud

Cloud Data Fusion instances are completely managed environments for building pipelines running in Google Cloud.

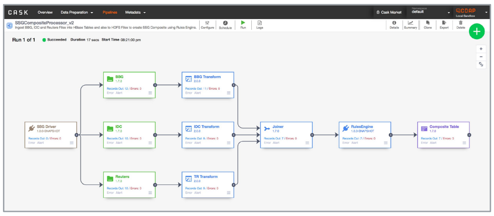
There are two editions you can choose from when creating the Dataflow environment. Basic edition is recommended for development and costs $1.80 per hour.

Enterprise edition is recommended for production and streaming pipelines. It costs $4.20 per hour.

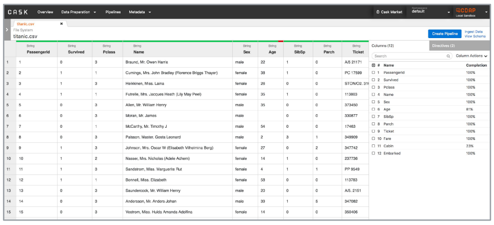
Additionally, pipelines run on a Dataproc cluster, which you are also charged for.

Cloud Data Fusion is based on the open source CDAP data analytics platform

- Pipelines provide an interface for building ETL jobs:
  - Connect to a source.
  - Transform source data.
  - Write to a sink.
- Wranglers provide a visual interface for transforming data.



Ingest BBG, IDC and Reuters Files into HBase Tables and also to HDFS Files to create SSG Composite using Rules Engine.



titanic.csv

	String PassengerId	String Survived	String Pclass	String Name	String Sex	String Age	String SibSp	String Parch	String Ticket
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 31
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450
6	6	0	3	Moran, Mr. James	male		0	0	330877
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549
12	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783
13	13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151
14	14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406

Columns (12)

☐	#	Name	Completion
☐	1	PassengerId	100%
☐	2	Survived	100%
☐	3	Pclass	100%
☐	4	Name	100%
☐	5	Sex	100%
☐	6	Age	81%
☐	7	SibSp	100%
☐	8	Parch	100%
☐	9	Ticket	100%
☐	10	Fare	100%
☐	11	Cabin	23%
☐	12	Embarked	100%

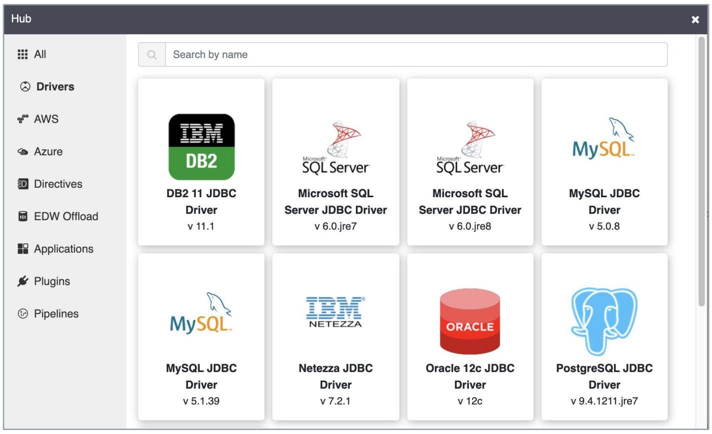
Google Cloud

Cloud Data Fusion is based on an open-source project call CDAP. This is important because an investment in learning Cloud Data Fusion is not only useful in Google Cloud, but will also be useful in on-premises and hybrid environments.

Pipelines provide an interface for building ETL jobs. You first connect to a source, then transform the source data, and finally write to a sink.

Wranglers provide a visual interface for specific data transformations. These are built-in, drag-and-drop objects for manipulating the data.

Cloud Data Fusion Hub provides access to drivers, plugins, and pre-configured pipelines

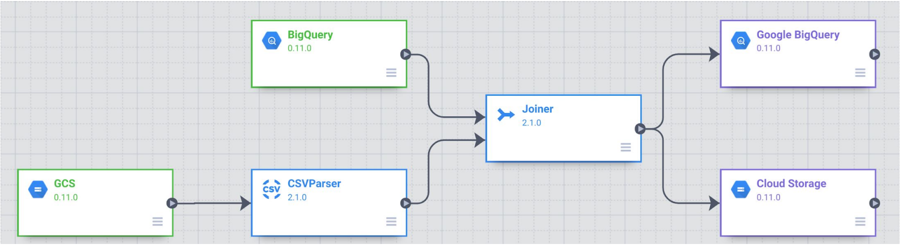


Google Cloud

Cloud Data Fusion Hub provides access to drivers, plugins, and pre-configured pipelines. This is available from your Cloud Data Fusion instance when you start it in your Google Cloud project.

Pipelines represent a series of stages arranged in a directed acyclic graph (DAG)

- Stages “nodes” in the pipeline graph can be of different types.
- Allows non-linear pipelines:
  - Can fork where output from a node can be sent to two or more stages.
  - Two or more forked nodes can merge at a transform or a sink node.



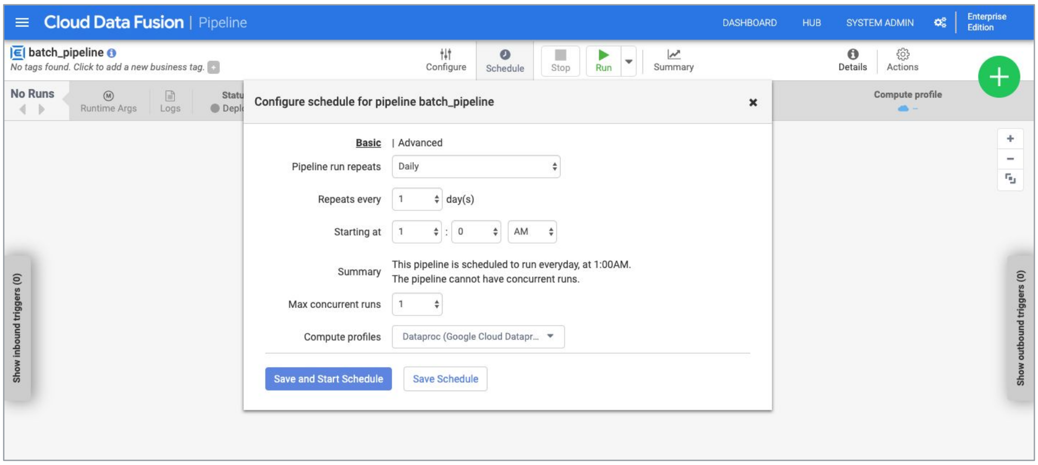
Google Cloud

Pipelines represent a series of stages arranged in a directed acyclic graph (DAG).

Stages “nodes” in the pipeline graph can be of different types.

Non-linear pipelines are supported: the nodes can fork where output from a node can be sent to two or more stages, and then two or more forked nodes can merge at a transform or a sink node.

You can schedule batch pipelines



Configure schedule for pipeline batch_pipeline

Pipeline run repeats: Daily

Repeats every: 1 day(s)

Starting at: 1 : 0 AM

Summary: This pipeline is scheduled to run everyday, at 1:00AM.
The pipeline cannot have concurrent runs.

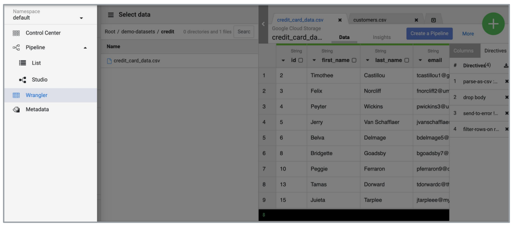
Max concurrent runs: 1

Compute profiles: Dataproc (Google Cloud Datapr...

There is also a scheduler where you can set up batch, recurring jobs at an appropriate interval.

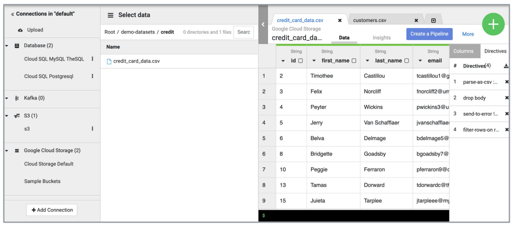
Data analysts can explore datasets in Wrangler

Code-free, visual environment for transforming data in data pipelines



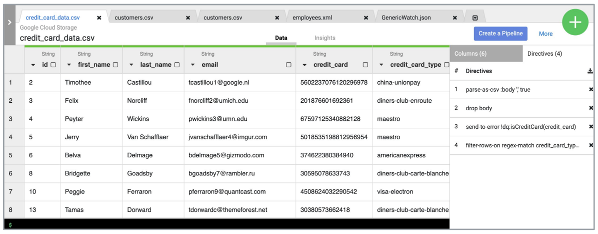
	String id	String first_name	String last_name	String email
1	2	Timothee	Castillou	tcastillou1@g
2	3	Felix	Norcliff	fnorcliff2@ur
3	4	Peyter	Wickins	pwickins3@u
4	5	Jerry	Van Schafflaer	jvanschafflae
5	6	Belva	Delmage	bdelmage5@
6	8	Bridgette	Goadsby	bgoadsby7@
7	10	Peggie	Ferraron	pferraron9@c
8	13	Tamas	Dorward	tdorwardc@tl
9	15	Juieta	Tarplee	jtarpleee@my

#	Directives(4)
1	parse-as-csv :...
2	drop body
3	send-to-error !...
4	filter-rows-on r...



	String id	String first_name	String last_name	String email
1	2	Timothee	Castillou	tcastillou1@g
2	3	Felix	Norcliff	fnorcliff2@ur
3	4	Peyter	Wickins	pwickins3@u
4	5	Jerry	Van Schafflaer	jvanschafflae
5	6	Belva	Delmage	bdelmage5@
6	8	Bridgette	Goadsby	bgoadsby7@
7	10	Peggie	Ferraron	pferraron9@c
8	13	Tamas	Dorward	tdorwardc@tl
9	15	Juieta	Tarplee	jtarpleee@my

#	Directives(4)
1	parse-as-csv :...
2	drop body
3	send-to-error !...
4	filter-rows-on r...

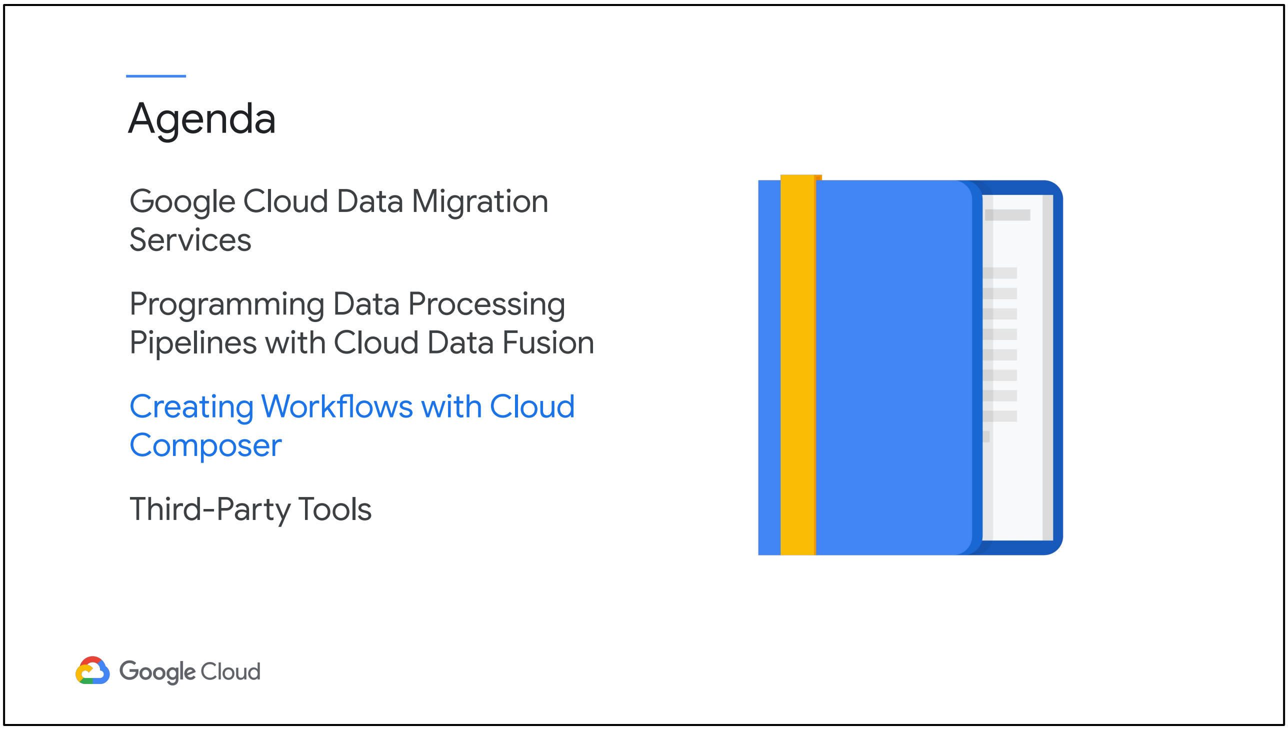


	String id	String first_name	String last_name	String email	String credit_card	String credit_card_type
1	2	Timothee	Castillou	tcastillou1@google.nl	5602237076120296978	china-unionpay
2	3	Felix	Norcliff	fnorcliff2@umich.edu	201876601692361	diners-club-enroute
3	4	Peyter	Wickins	pwickins3@umn.edu	67597125340882128	maestro
4	5	Jerry	Van Schafflaer	jvanschafflaer4@imgur.com	5018535198812956954	maestro
5	6	Belva	Delmage	bdelmage5@gizmodo.com	374622380384940	americanexpress
6	8	Bridgette	Goadsby	bgoadsby7@rambler.ru	30595078633743	diners-club-carte-blanche
7	10	Peggie	Ferraron	pferraron9@quantcast.com	4508624032290542	visa-electron
8	13	Tamas	Dorward	tdorwardc@themeforest.net	30380573662418	diners-club-carte-blanche

#	Directives
1	parse-as-csv :body ',' true
2	drop body
3	send-to-error !dq:isCreditCard(credit_card)
4	filter-rows-on regex-match credit_card_typ...

Google Cloud

Data analysts can explore datasets in Wrangler and preview the results of their transformation. This code-free visual environment is especially useful for data analysts who are often less comfortable programming.



Agenda

Google Cloud Data Migration Services

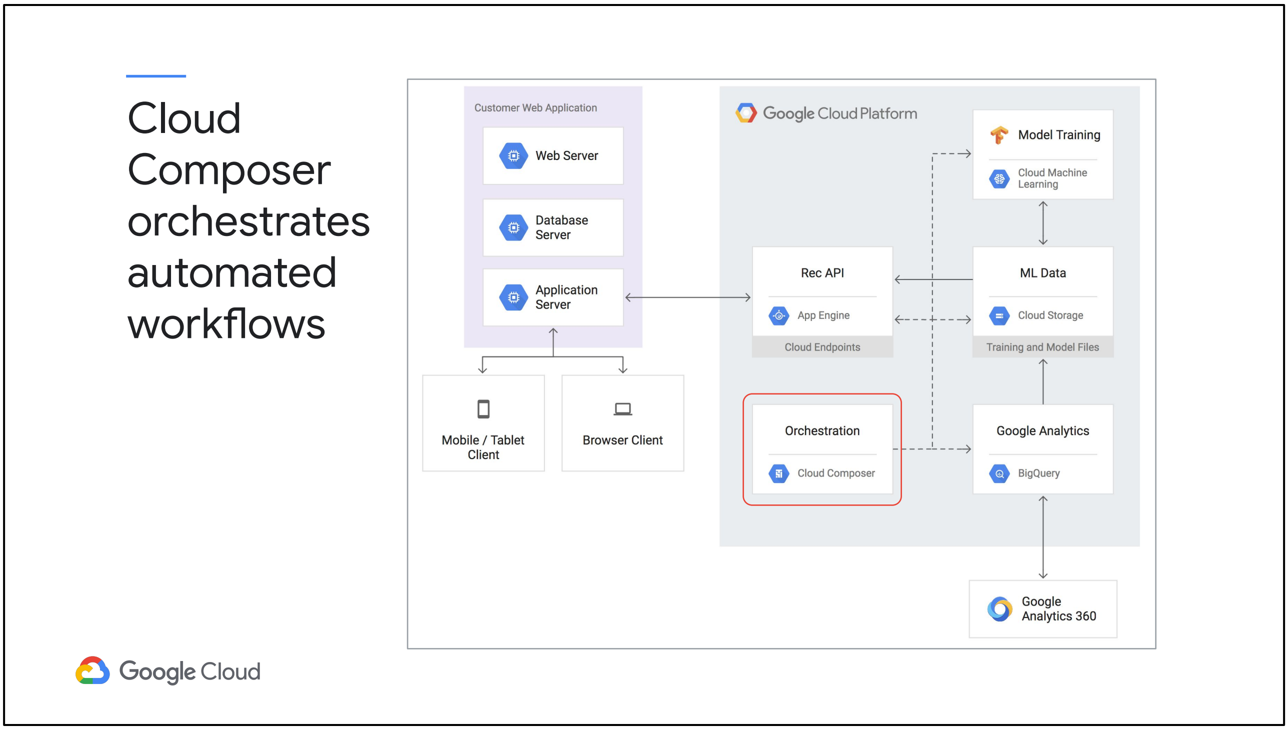
Programming Data Processing Pipelines with Cloud Data Fusion

Creating Workflows with Cloud Composer

Third-Party Tools

Google Cloud

Cloud Data Fusion helps you build data processing jobs. Sometimes you want to coordinate larger workflows, which may include an ETL pipeline as just one of the steps. This is where Cloud Composer comes in.



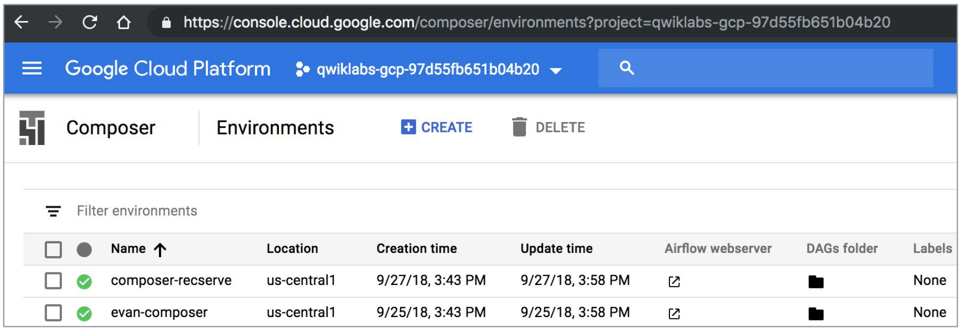
Cloud Composer orchestrates automated workflows

This sort of workflow can be programmed, managed, and executed using Cloud Composer.

Cloud Composer help with the orchestration of various steps.

Cloud Composer is a managed Apache Airflow environment

- Open source
- Create workflows using Python



https://console.cloud.google.com/composer/environments?project=qwiklabs-gcp-97d55fb651b04b20

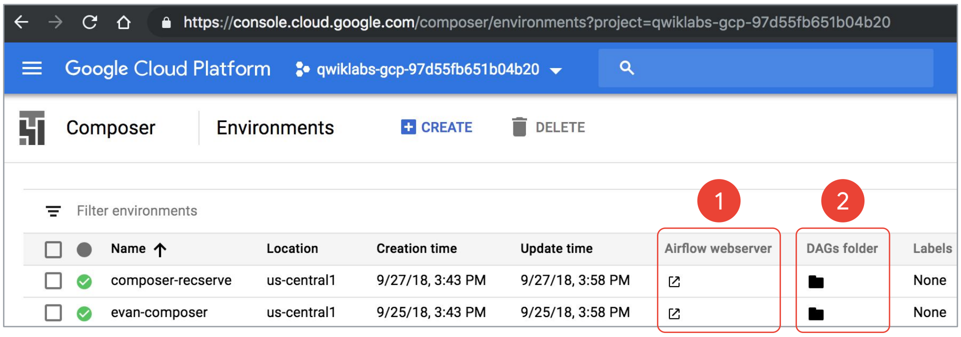
□	●	Name ↑	Location	Creation time	Update time	Airflow webserver	DAGs folder	Labels
□	✓	composer-recserve	us-central1	9/27/18, 3:43 PM	9/27/18, 3:58 PM	↗	■	None
□	✓	evan-composer	us-central1	9/25/18, 3:43 PM	9/25/18, 3:58 PM	↗	■	None

Google Cloud

Cloud Composer is really just a fully managed environment running Apache Airflow. Airflow is an open-source workflow engine. You create your workflows with some simple Python code.

You can find more information about Airflow at the URL airflow.apache.org. As with Cloud Data Fusion, because Airflow is open-source and will run anywhere, your investment in learning it is not restricted to Google Cloud.

Each Airflow environment has a separate web server and folder in Cloud Storage for pipeline DAGs



https://console.cloud.google.com/composer/environments?project=qwiklabs-gcp-97d55fb651b04b20
Google Cloud Platform
qwiklabs-gcp-97d55fb651b04b20
Composer
Environments
CREATE
DELETE
Filter environments
1
2

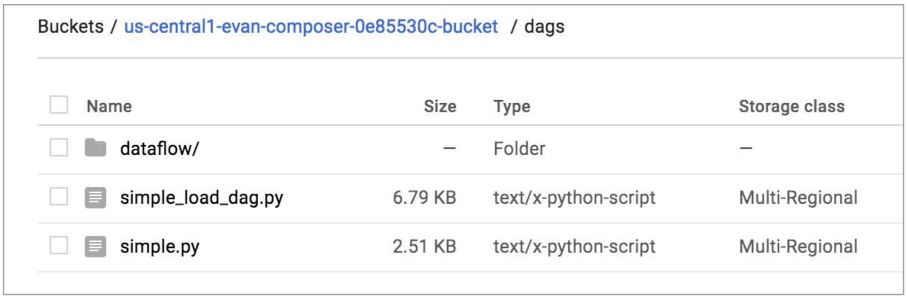
□	●	Name ↑	Location	Creation time	Update time	Airflow webserver	DAGs folder	Labels
□	✓	composer-recserve	us-central1	9/27/18, 3:43 PM	9/27/18, 3:58 PM	↗	📁	None
□	✓	evan-composer	us-central1	9/25/18, 3:43 PM	9/25/18, 3:58 PM	↗	📁	None

Google Cloud

Each Airflow environment hosts a separate web server for access to the Airflow UI. There is also a folder created in Google Cloud Storage where you place the Python code files for your pipelines.

The DAGs folder is simply a Cloud Storage bucket where you will load your pipeline code

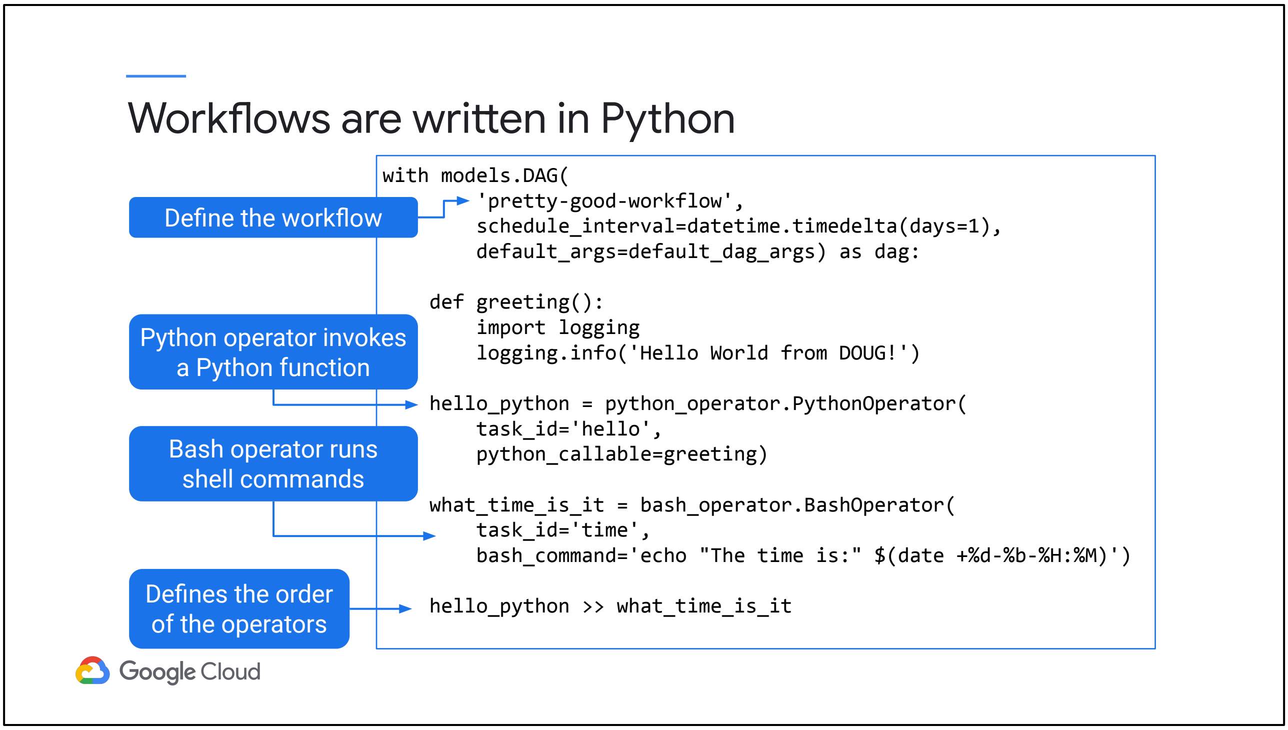
Any workflows added to the folder are automatically scheduled



Buckets / us-central1-evan-composer-0e85530c-bucket / dags

□	Name	Size	Type	Storage class
□	dataflow/	—	Folder	—
□	simple_load_dag.py	6.79 KB	text/x-python-script	Multi-Regional
□	simple.py	2.51 KB	text/x-python-script	Multi-Regional

The DAGs folder is simply a Cloud Storage bucket where you load your pipeline code. To run a workflow, just save its code file into the Cloud Storage bucket. Airflow detects the new file and runs it on the schedule specified in the file.



Workflows are written in Python

Define the workflow

Python operator invokes a Python function

Bash operator runs shell commands

Defines the order of the operators

```
with models.DAG(
        'pretty-good-workflow',
        schedule_interval=datetime.timedelta(days=1),
        default_args=default_dag_args) as dag:

    def greeting():
        import logging
        logging.info('Hello World from DOUG!')

    hello_python = python_operator.PythonOperator(
        task_id='hello',
        python_callable=greeting)

    what_time_is_it = bash_operator.BashOperator(
        task_id='time',
        bash_command='echo "The time is:" $(date +%d-%b-%H:%M)')

    hello_python >> what_time_is_it
```

Google Cloud

Here is an example of a simple workflow. At the top, the workflow is defined.

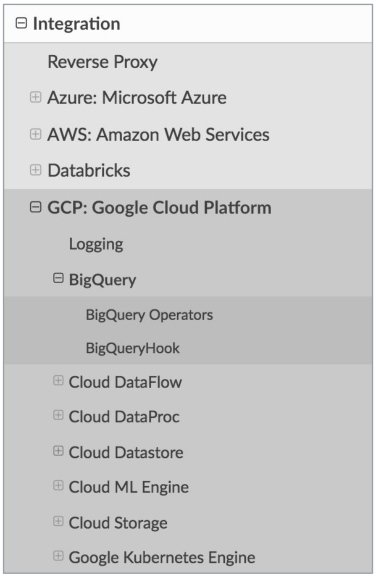
- Inside the workflow are a series of steps which make up the DAG (or Directed Acyclic Graph). Each step is implemented as an operator, and there are different types of operators.

- In this example, the first operator is a PythonOperator; it is used to invoke the Python greeting() function above it.

- The second operator is a BashOperator; it is used to run a shell command or script.
  If you can run Python functions and Shell scripts, you can do almost anything you need to in Google Cloud. Remember, every Google Cloud resource can be created with a shell script. There are many other types of operators as well.

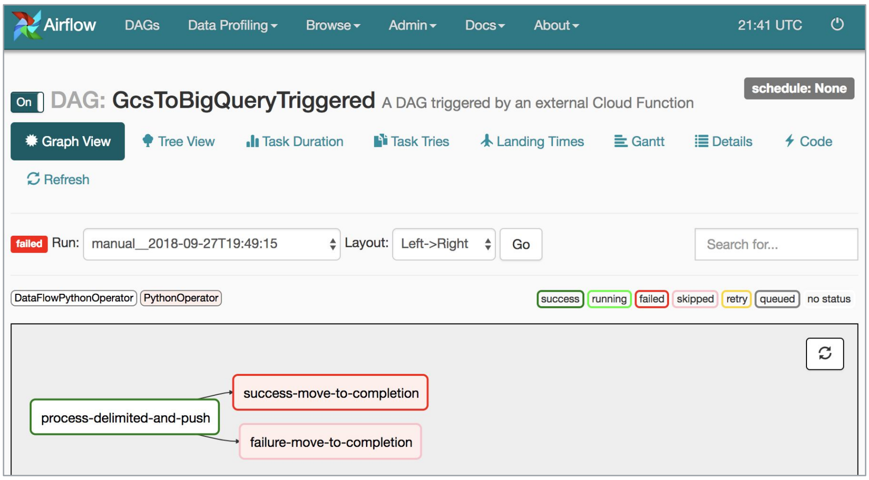
- At the very bottom, after the workflow is defined, the order of the operators is specified. In this example, the operator "hello_python" runs first, followed by the "what_time_is_it" operator.

Airflow provides many operators to orchestrate Google Cloud and other services



Airflow provides many operators to orchestrate Google Cloud and other services.

Airflow console allows you to monitor your workflows

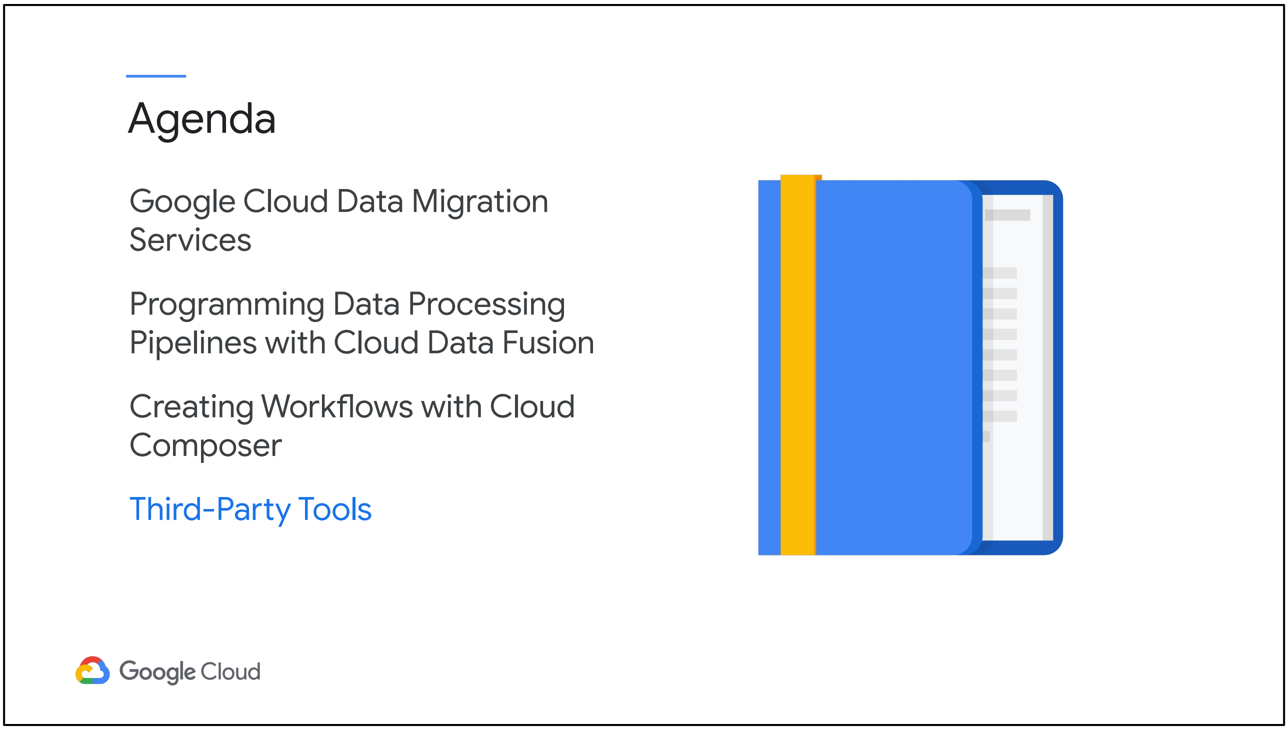


DAG: GcsToBigQueryTriggered A DAG triggered by an external Cloud Function

schedule: None

failed Run: manual__2018-09-27T19:49:15 Layout: Left->Right

The Airflow website allows you to monitor your workflows. You can see the history, how long each step took, whether there were errors, and so on. The console is available as part of the Cloud Composer environment and is automatically created by the service.



Agenda

Google Cloud Data Migration Services

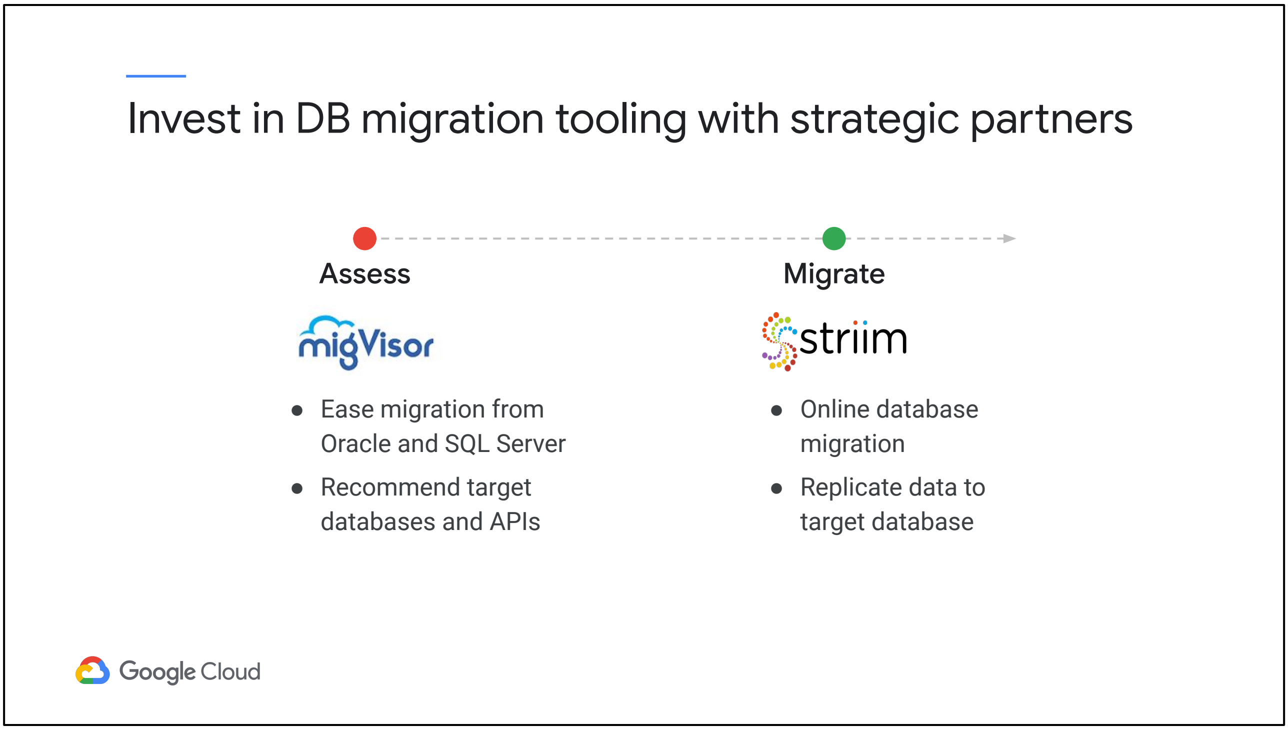
Programming Data Processing Pipelines with Cloud Data Fusion

Creating Workflows with Cloud Composer

Third-Party Tools

Google Cloud

Google recommends leveraging third-party tools and hiring strategic partners to help with your database migration projects. Let's talk about a couple of those tools now.



Invest in DB migration tooling with strategic partners

Assess

migVisor

- Ease migration from Oracle and SQL Server
- Recommend target databases and APIs

Migrate

striim

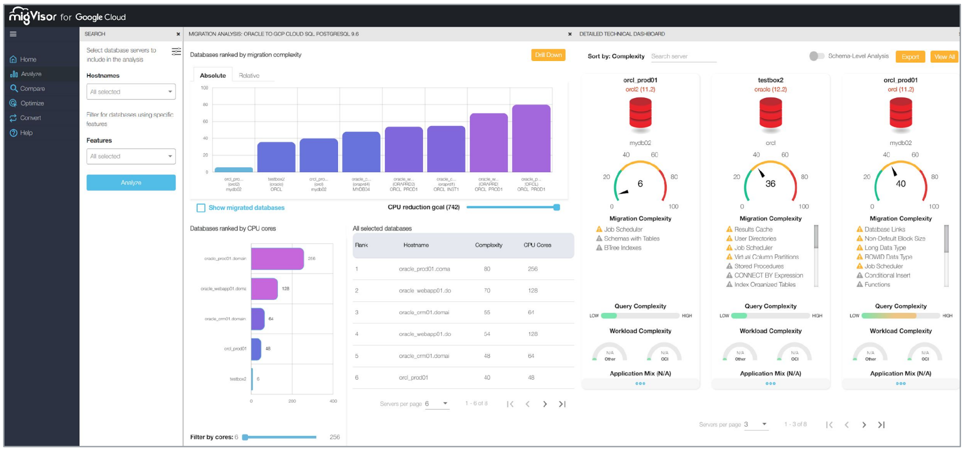
- Online database migration
- Replicate data to target database

Google Cloud

migVisor is an automated assessment tool. Recall that earlier in the course, you learned about Google's Implementation methodology which consists of four steps: assess, plan, deploy, optimize. migVisor helps with those first two steps. It helps find dependencies and dependents and makes recommendations for which Google databases and APIs you should pick as targets.

Striim, however, will help with the Deployment step. Striim allows you to do online database migrations and it handles the transferring and synchronization of the databases.

migVisor analyzes source database configuration, attributes, schema objects, and proprietary features



Select database servers to include in the analysis

Hostnames

All selected

Filter for databases using specific features

Features

All selected

Databases ranked by migration complexity

☐ Show migrated databases

CPU reduction goal (742)

Databases ranked by CPU cores

Filter by cores: 6 256

All selected databases

Rank	Hostname	Complexity	CPU Cores
1	oracle_prod01.coma	80	256
2	oracle_webapp01.do	70	128
3	oracle_crm01.domai	55	64
4	oracle_webapp01.do	54	128
5	oracle_crm01.domai	48	64
6	orcl_prod01	40	48

Servers per page 6 1 - 6 of 8

Sort by: Complexity Search server

Schema-Level Analysis

Migration Complexity

- Job Scheduler
- Schemas with Tables
- BTree Indexes

Query Complexity

Workload Complexity

Application Mix (N/A)

Migration Complexity

- Results Cache
- User Directories
- Job Scheduler
- Virtual Column Partitions
- Stored Procedures
- CONNECT BY Expression
- Index Organized Tables

Query Complexity

Workload Complexity

Application Mix (N/A)

Migration Complexity

- Database Links
- Non-Default Block Size
- Long Data Type
- ROWID Data Type
- Job Scheduler
- Conditional Insert
- Functions

Query Complexity

Workload Complexity

Application Mix (N/A)

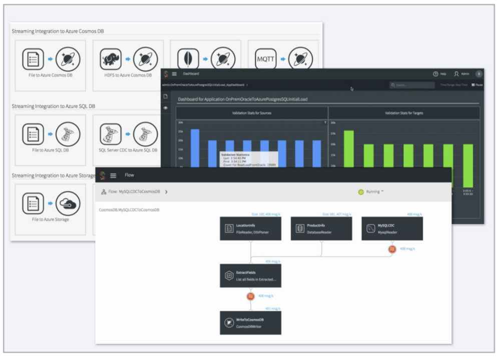
Servers per page 3 1 - 3 of 8

Google Cloud

migVisor's automated tool analyzes source database configuration, attributes, schema objects, and proprietary features. This helps build that initial plan required for a successful migration project.

Striim provides an easy-to-use interface for transferring data between databases

- Stream data between targets and sources.
- Transform data within data pipelines using SQL.
- Captures changes in real-time.
- Allows for data transfer between different database types.

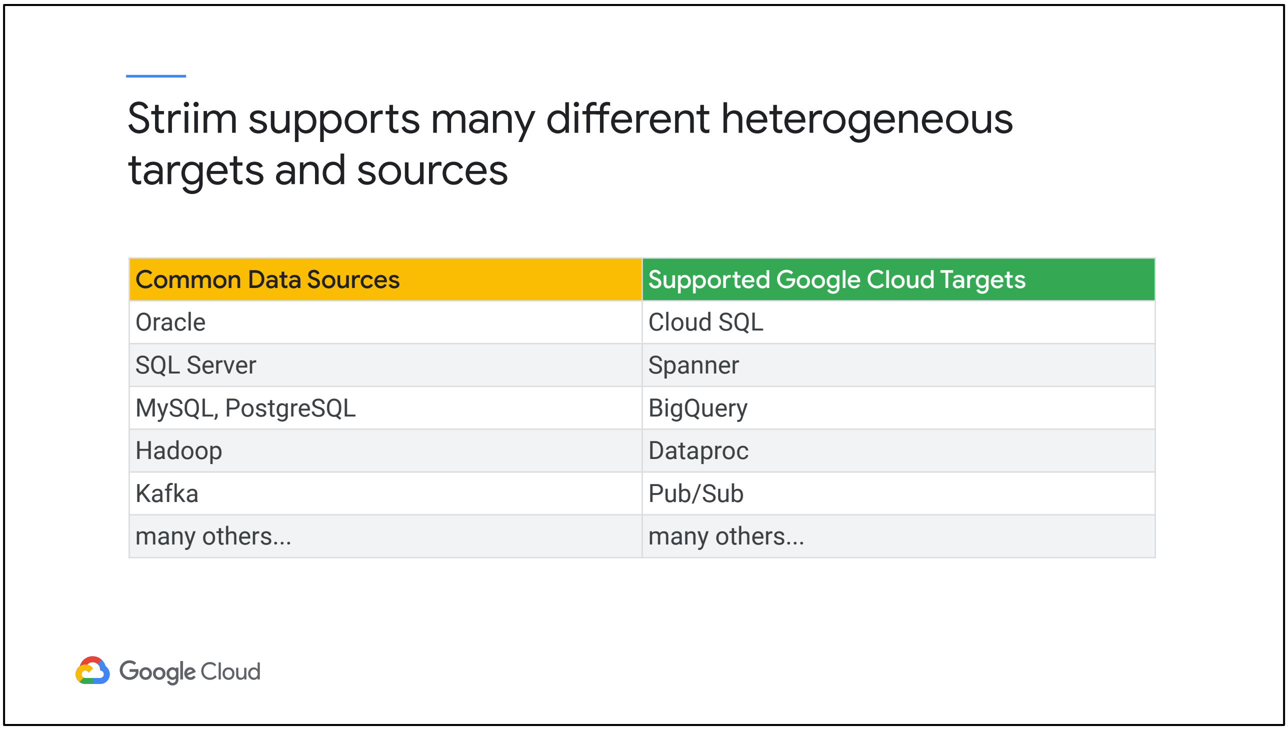


Streaming Integration to Azure Cosmos DB
File to Azure Cosmos DB
HDFS to Azure Cosmos DB
MQTT
Streaming Integration to Azure SQL DB
File to Azure SQL DB
SQL Server CDC to Azure SQL DB
Streaming Integration to Azure Storage
File to Azure Storage
Dashboard
Help
Admin
Dashboard for Application OnPremOracleToAzurePostgresSQLInitialLoad
Validation Stats for Sources
Validation Stats for Targets
Flow
Flow: MySQLCDCToCosmosDB
Running
CosmosDB.MySQLCDCToCosmosDB
LocationInfo
FileReader, DSVParser
ProductInfo
DatabaseReader
MySQLCDC
MysqlReader
408 msg/s
ExtractFields
List all fields in Extracted...
408 msg/s
WriteToCosmosDB
CosmosDBWriter

Google Cloud

Striim provides an easy-to-use interface for transferring data between databases. For detailed information, go to Striim's website at www.striim.com.

- Striim allows you to stream data between a target and a source. Many targets and sources are supported.
- You can also do data transformations within the pipelines using SQL.
- Typically, an initial data transfer is done to move the bulk of the historical data. Then, Striim captures data changes on the source in real time and synchronizes them with the new target database. During the synchronization period, you can migrate clients to the new database. Eventually, all of the old databases will have no clients, and it can be retired.
- Using Striim, you can even transfer data between different database types.

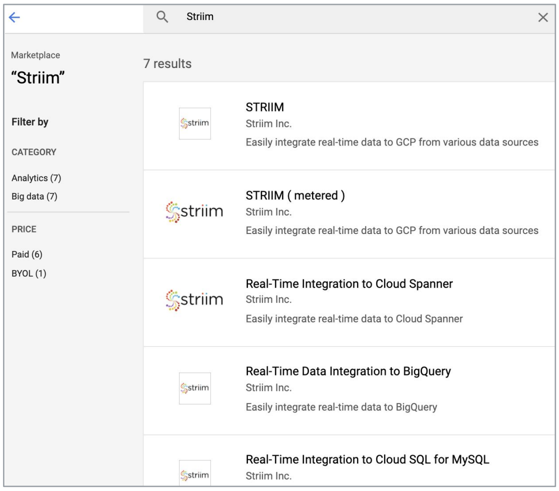


Striim supports many different heterogeneous targets and sources

Common Data Sources	Supported Google Cloud Targets
Oracle	Cloud SQL
SQL Server	Spanner
MySQL, PostgreSQL	BigQuery
Hadoop	Dataproc
Kafka	Pub/Sub
many others...	many others...

Striim supports many different heterogeneous targets and sources. The table here lists some common targets and sources. Notice that it is not limited to only relational databases. It can be used with big data, data warehousing, and streaming analytics targets, as well as sources.

Striim is available from Google Cloud Marketplace



Striim

Marketplace

"Striim"

Filter by

CATEGORY

Analytics (7)

Big data (7)

PRICE

Paid (6)

BYOL (1)

7 results

STRIIM
Striim Inc.
Easily integrate real-time data to GCP from various data sources

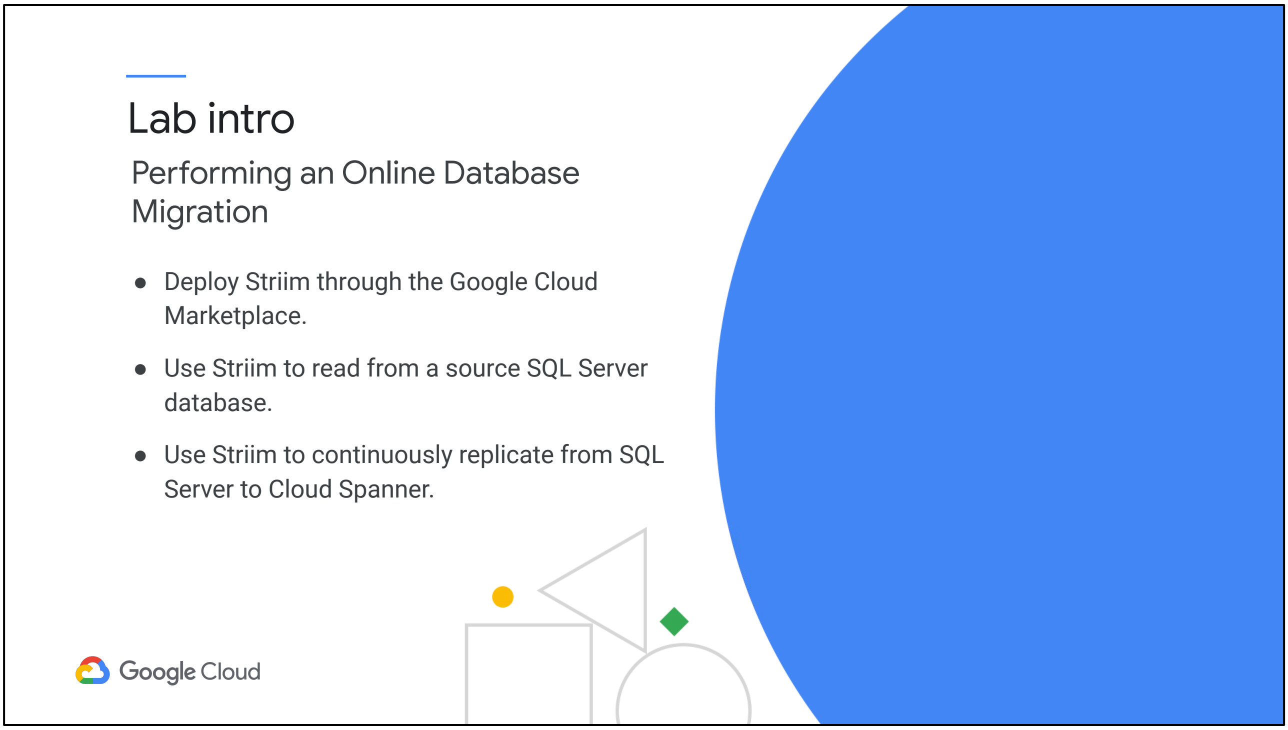
STRIIM (metered)
Striim Inc.
Easily integrate real-time data to GCP from various data sources

Real-Time Integration to Cloud Spanner
Striim Inc.
Easily integrate real-time data to Cloud Spanner

Real-Time Data Integration to BigQuery
Striim Inc.
Easily integrate real-time data to BigQuery

Real-Time Integration to Cloud SQL for MySQL
Striim Inc.

Striim is easily installed in Google Cloud from Marketplace. After the server is up, a simple configuration is run.



Lab intro

Performing an Online Database Migration

- Deploy Striim through the Google Cloud Marketplace.
- Use Striim to read from a source SQL Server database.
- Use Striim to continuously replicate from SQL Server to Cloud Spanner.

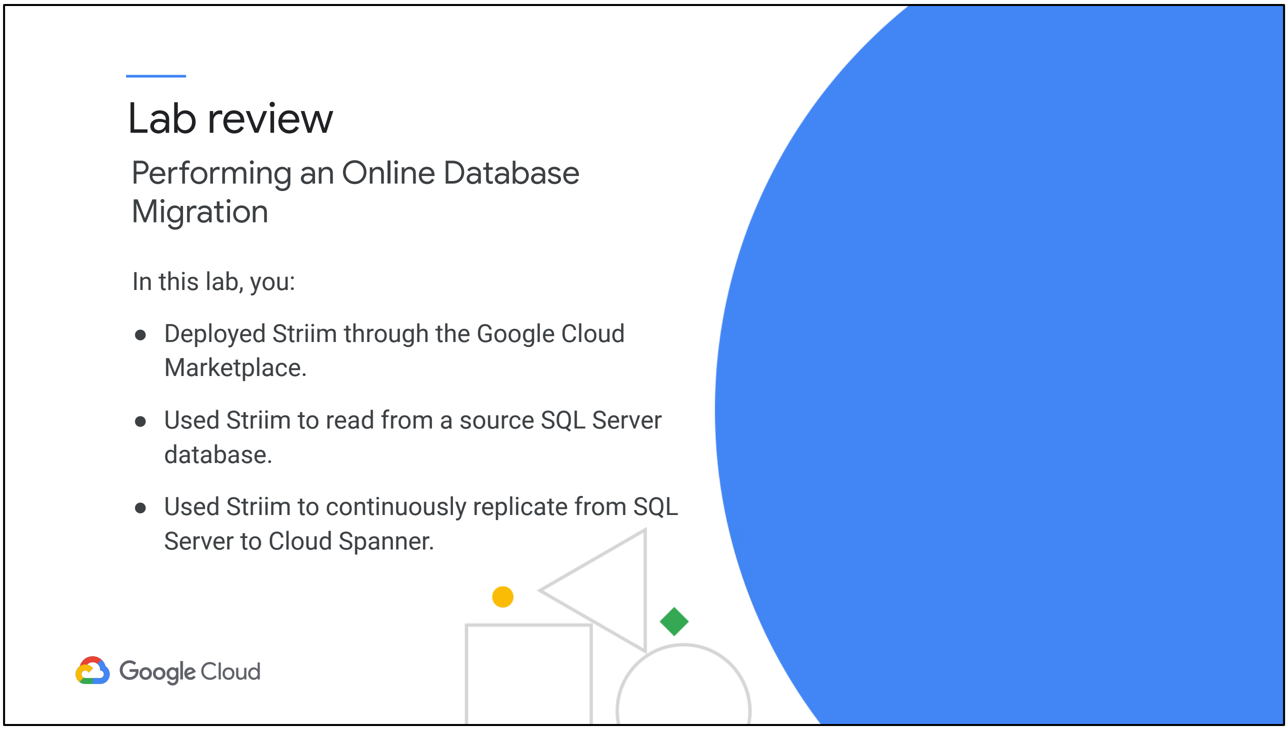
Google Cloud

In this lab, you use Striim to perform an automated data migration from Cloud SQL to Spanner. What you learn here applies to any target or source database.

In this lab, you will deploy Striim through the Google Cloud Marketplace and set it up.

Then you will use Striim to read from a SQL Server database running in Cloud SQL and perform an initial batch migration.

Lastly, you will use Striim to continuously replicate from Cloud SQL to Cloud Spanner until you are ready to make the switch from the old database to the new one.



Lab review

Performing an Online Database Migration

In this lab, you:

- Deployed Striim through the Google Cloud Marketplace.
- Used Striim to read from a source SQL Server database.
- Used Striim to continuously replicate from SQL Server to Cloud Spanner.

Google Cloud

In this lab, you deployed Striim through the Google Cloud Marketplace.

Then you used Striim to read from a source SQL Server database, and continuously replicate data from SQL Server to Cloud Spanner.

Striim is an automated tool that supports homogenous and heterogenous data migrations. It supports many different targets and sources. It can greatly simplify your database migration projects and make them more automated and reliable.

Module review

In this module, you learned how to move large amounts of data to the cloud using Google transfer services. You also, learned to program data processing and ETL pipelines using Cloud Data Fusion and Cloud Composer. Finally, you used Striim to automate the migration of an Enterprise SQL Server database into Google Spanner.

Google Cloud