

Course 2: Production ML Systems

Module 2: Ingesting data for Cloud-based analytics and ML

Lesson Title: **Introduction**

Presenter: Val

Format: Talking Head

Video Name: T-PSML-0_2_I1_introduction



Ingesting data for Cloud-based
analytics and ML

Val Fontama

Course Progress

Architecting Production ML Systems

Ingesting data for Cloud-based analytics and ML

Designing Adaptable ML systems

Designing High
Performance ML Systems

Hybrid ML Systems

Agenda

Migration Overview

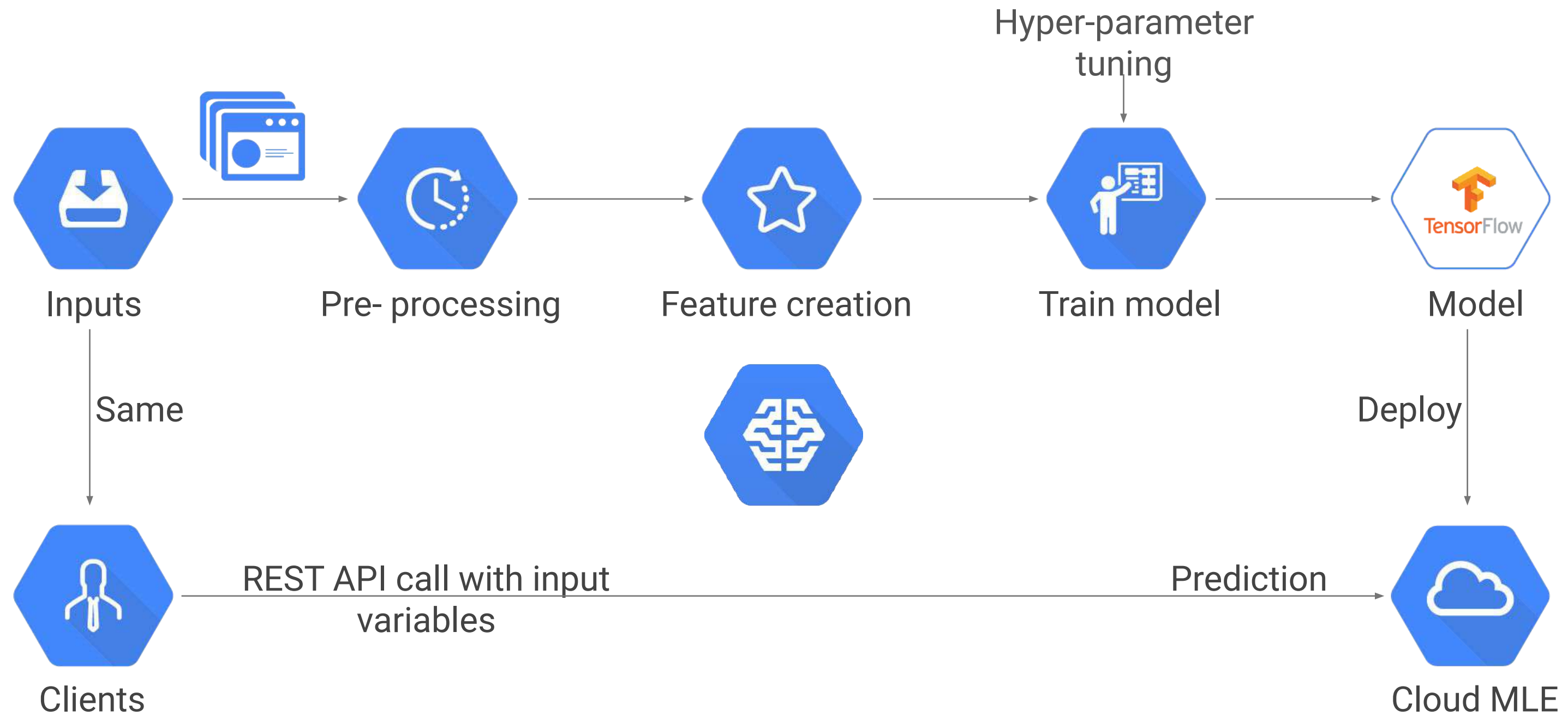
Data On-Premise

Large Datasets

Data on other clouds

Existing Databases

Your data must be on the cloud to benefit from ML Engine



Challenges to Get Data Onto the Cloud



So much
data



Too little
bandwidth

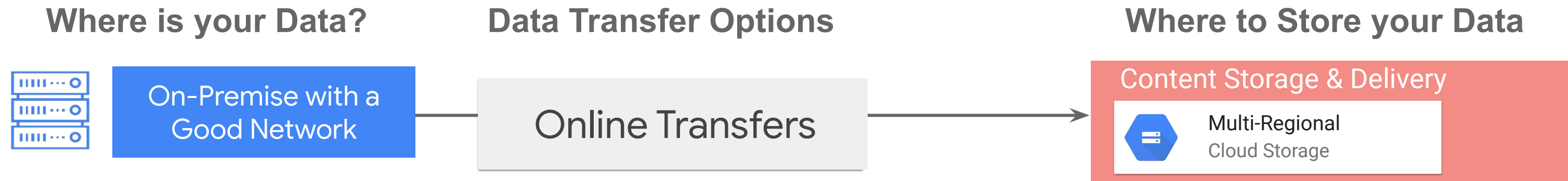


Checksumming,
encryption,
firewalls...

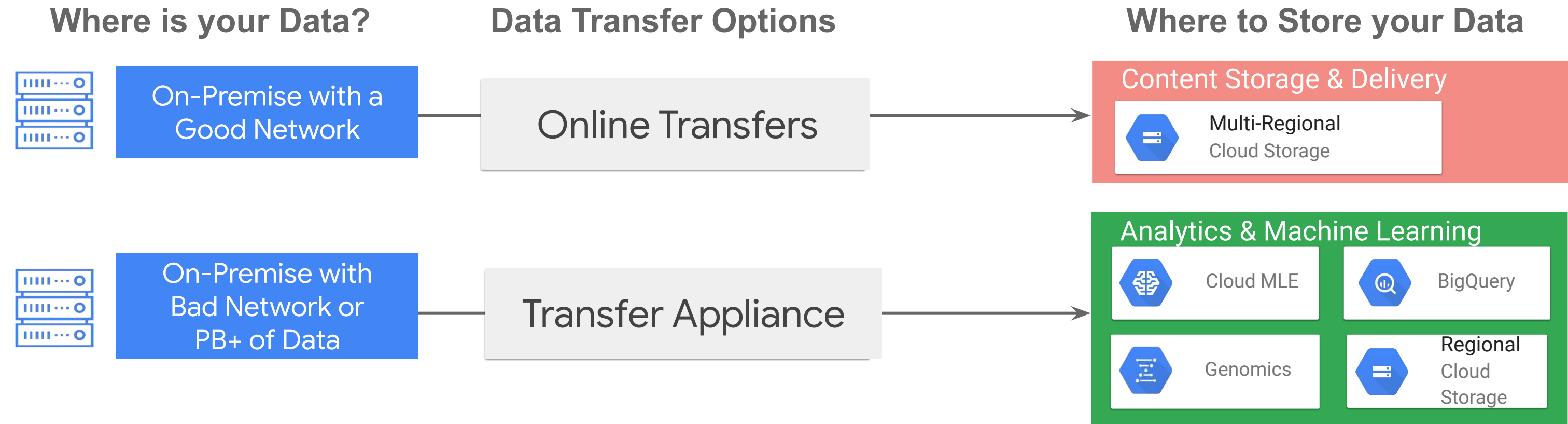


No time
and few
resources

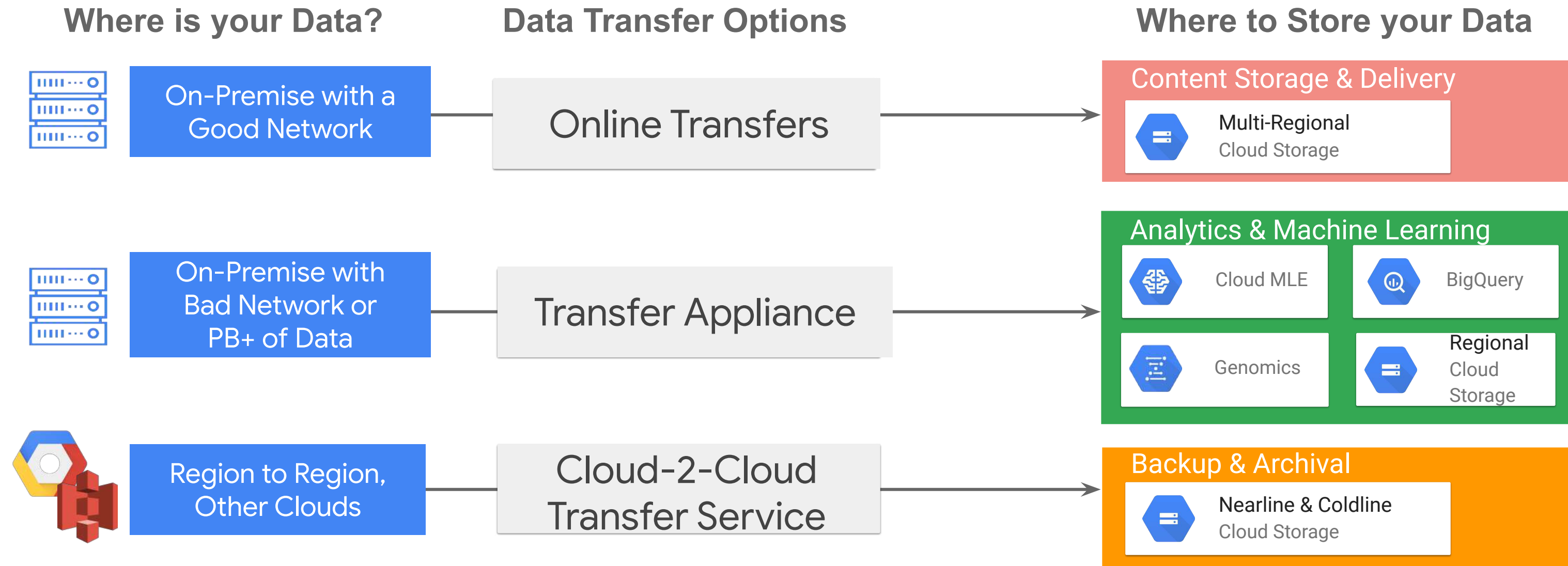
Migration methods: How to get your data on GCP



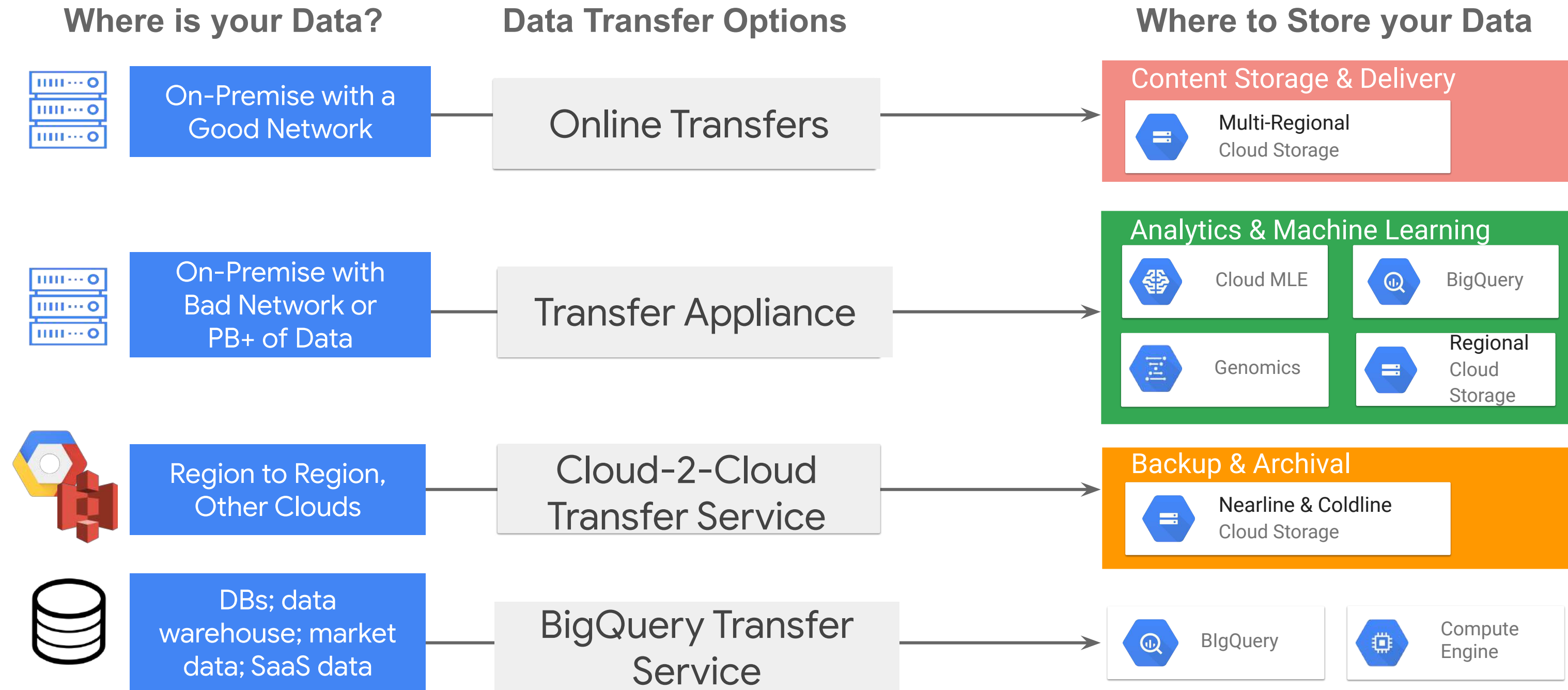
Migration methods: How to get your data on GCP



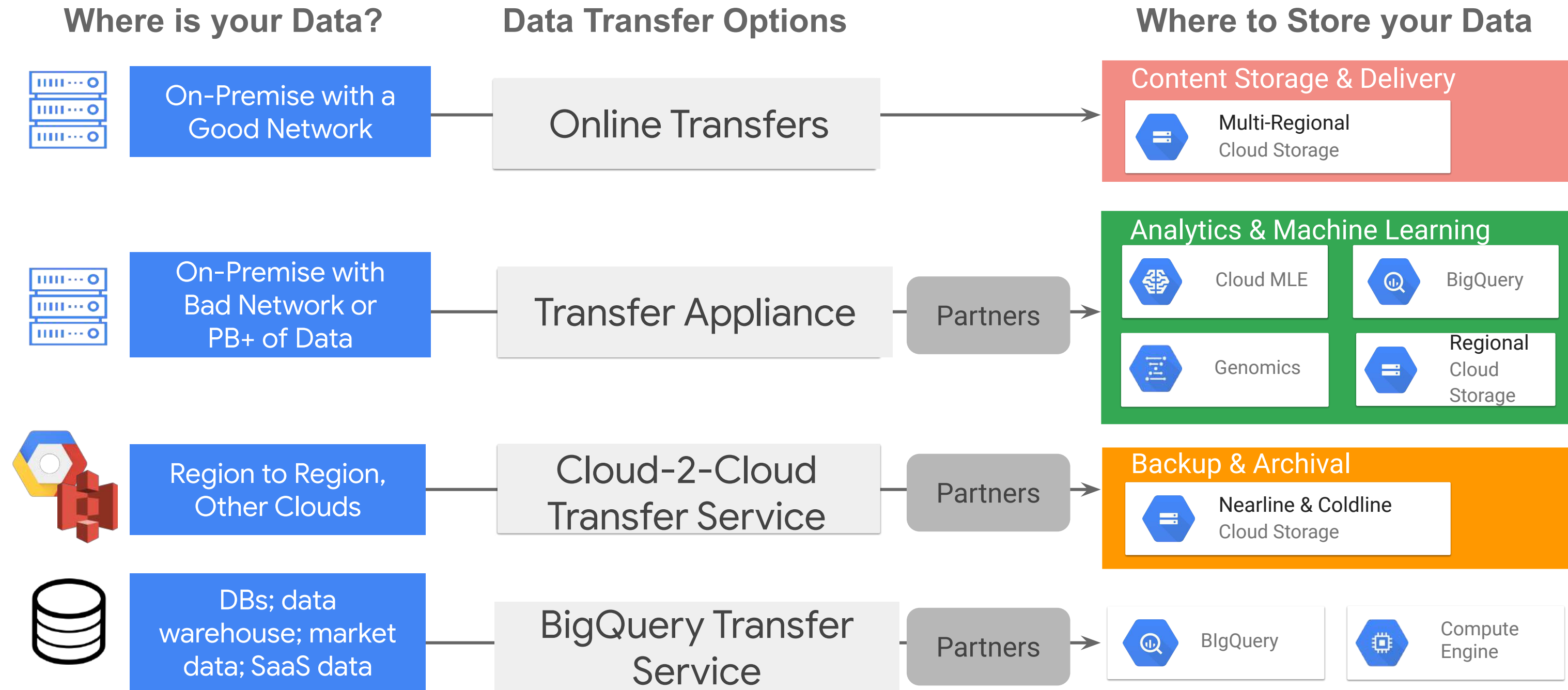
Migration methods: How to get your data on GCP



Migration methods: How to get your data on GCP



Migration methods: How to get your data on GCP



Course 2: Production ML Systems

Module 2: Ingesting data for Cloud-based analytics and ML

Lesson Title: **Data On-Premise**

Presenter: Val

Format: Talking Head

Video Name: T-PSML-0_2_I2_data_on-premise

Agenda

Migration Overview

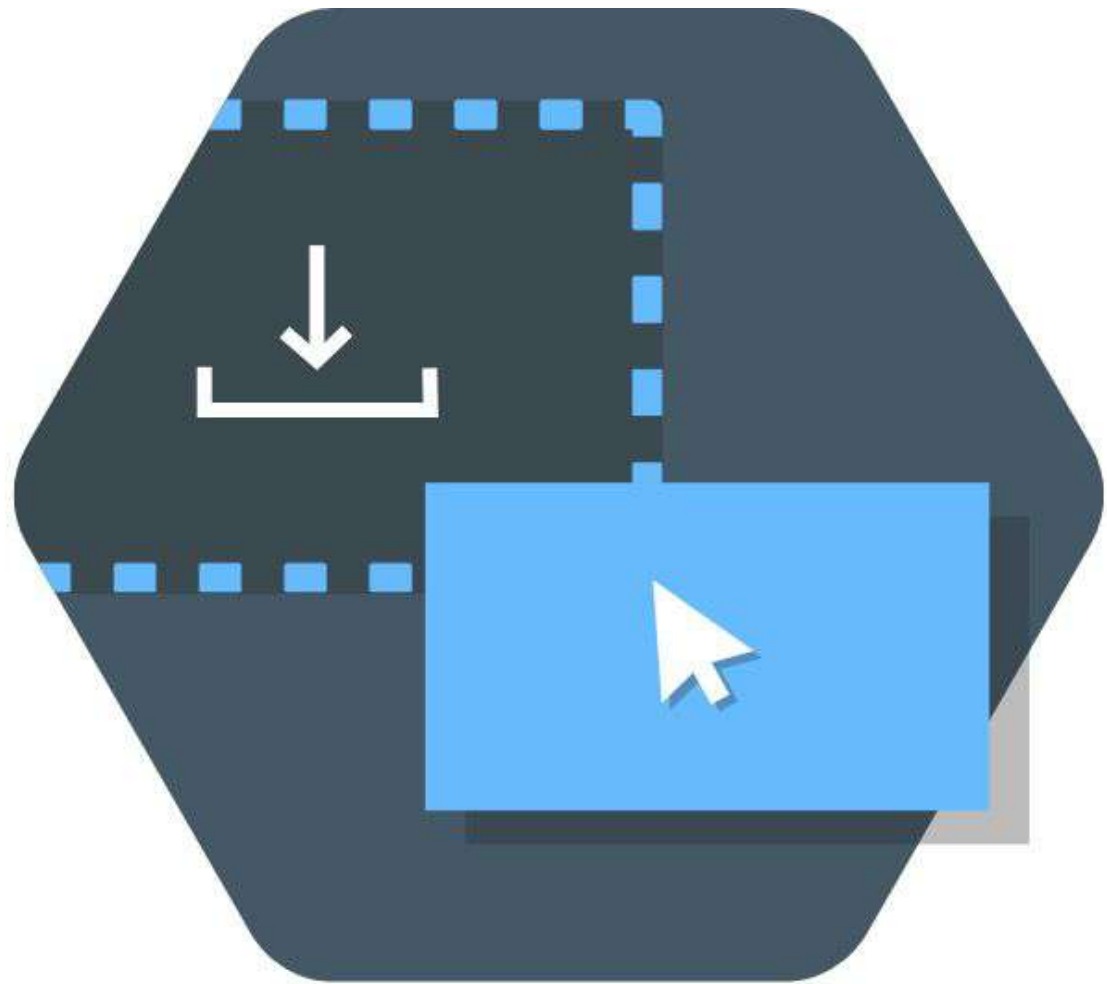
Data On-Premise

Large Datasets

Data on other clouds

Existing Databases

Use your network to
move data to Google
Cloud Storage



Use `gsutil` to move data
to Google Cloud Storage

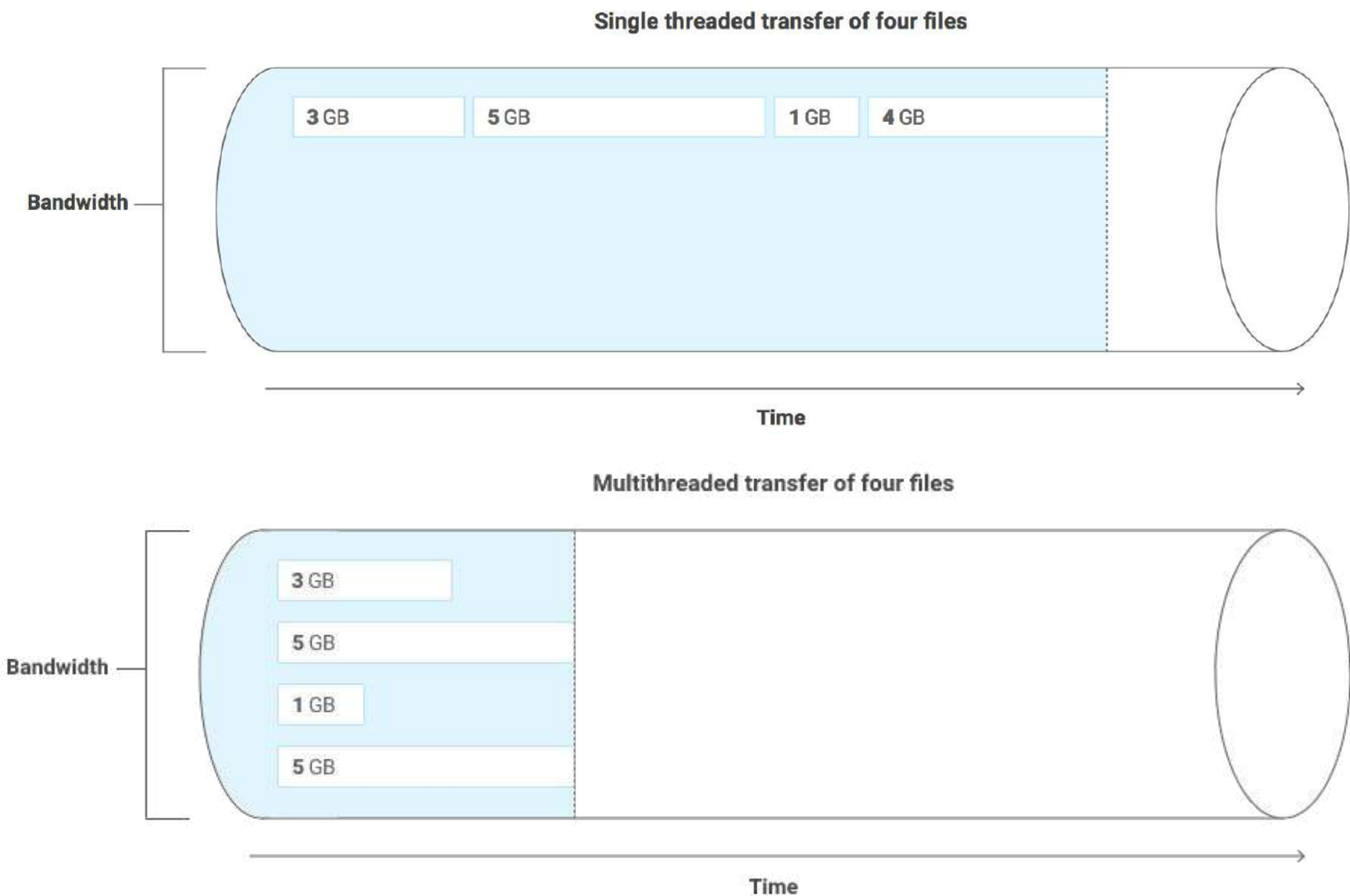
```
gsutil cp *.txt gs://my-bucket
```

Copy all text files in my
local directory to my Google
Cloud Storage Bucket

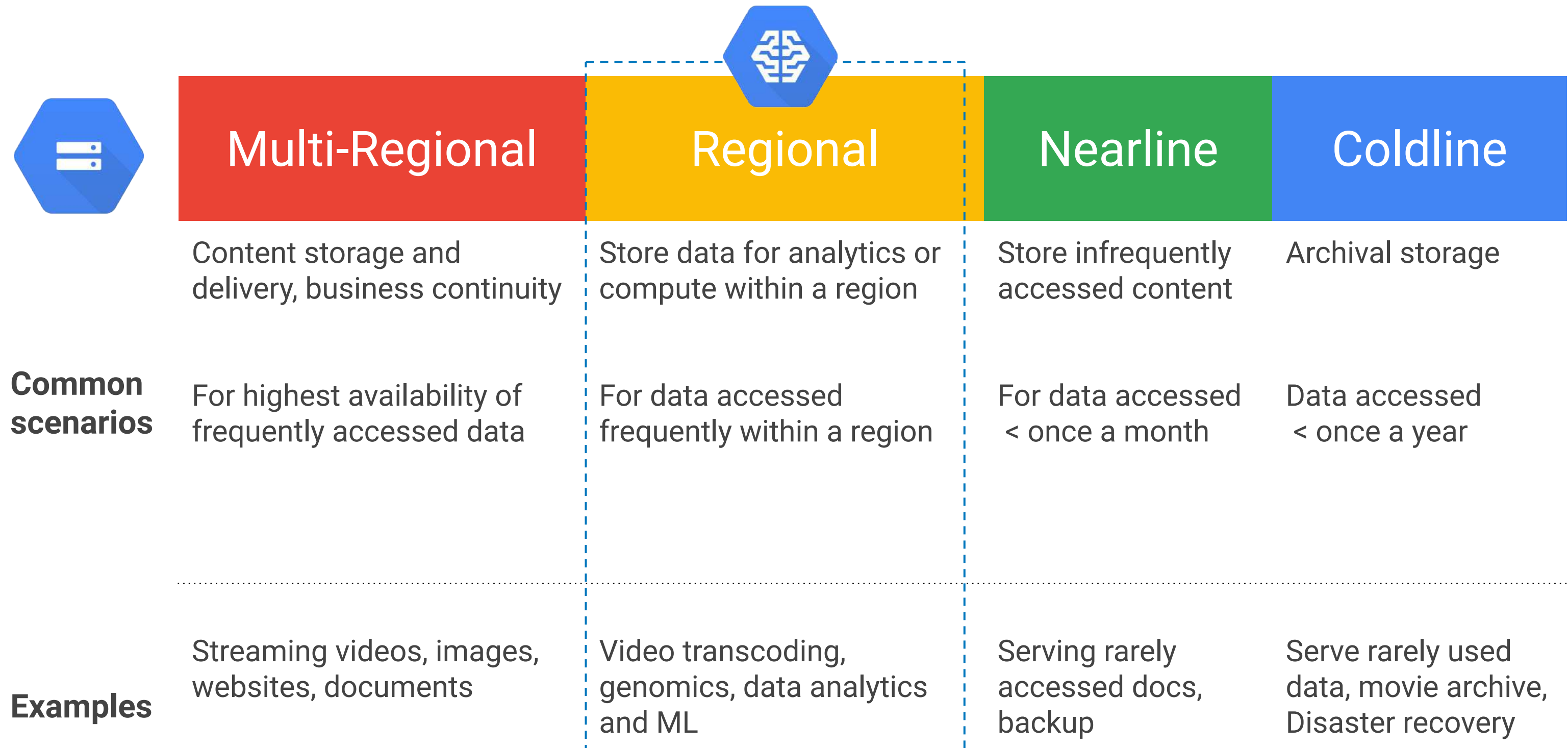
Use multithreaded transfers to Google Cloud Storage

```
gsutil -m cp -r  
[SOURCE_DIRECTORY]  
gs://[BUCKET_NAME]
```

Include **-m** to
enable
multi-threading



ML model data is typically stored regionally in Cloud Storage



Course 2: Production ML Systems

Module 2: Ingesting data for Cloud-based analytics and ML

Lesson Title: **Large Datasets**

Presenter: Val

Format: Talking Head

Video Name: T-PSML-0_2_I3_large_datasets

Agenda

Migration Overview

Data On-Premise

Large Datasets

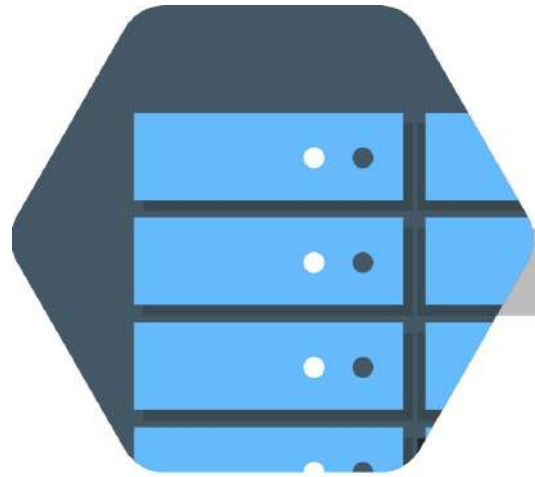
Data on other clouds

Existing Databases

Google Cloud Transfer Appliance



If online transfer would
take more than a week, use
transfer appliance



If you have
60TB+ data



If you have 1TB+
data and a 10 Mbps
network

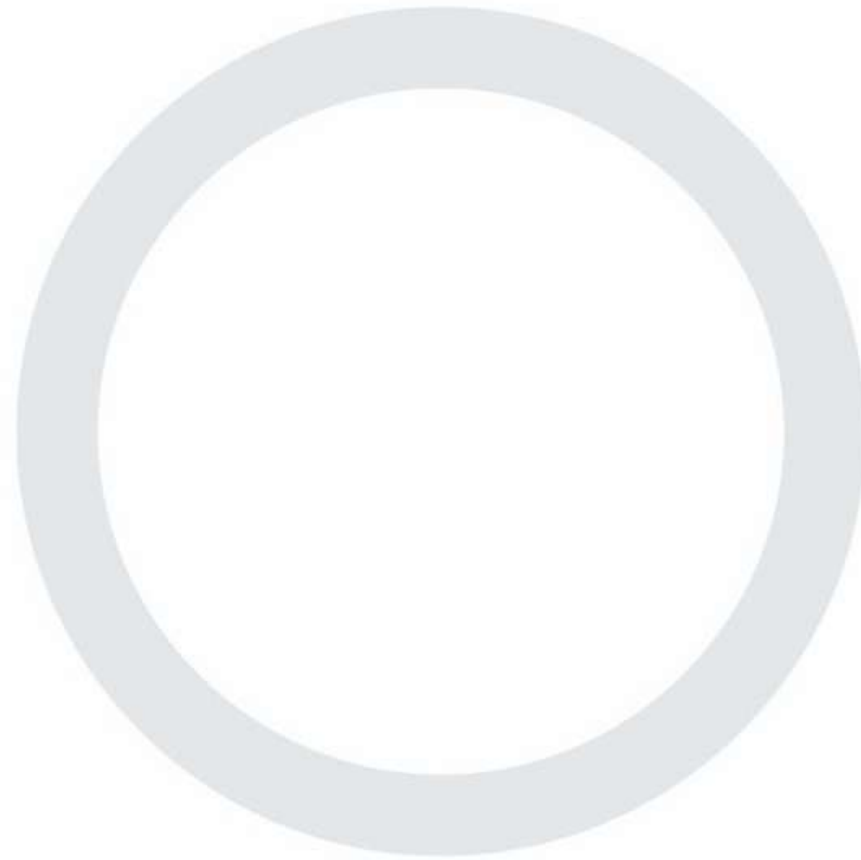
Networks Bottleneck at Big Data Scale

Total bandwidth

		1 Mbps	10 Mbps	100 Mbps	1 GB	10 GB	100 GB
Target Data	1 GB	3 hrs	18 mins	2 mins	11 secs	1 sec	.1 secs
	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 yrs	34 years	3 years	124 days	12 days	30 hrs
	10 PB	3,404 yrs	340 yrs	34 years	3 years	124 days	12 days
	100 PB	34,048 yrs	3,404 yrs	340 yrs	34 years	3 years	124 days

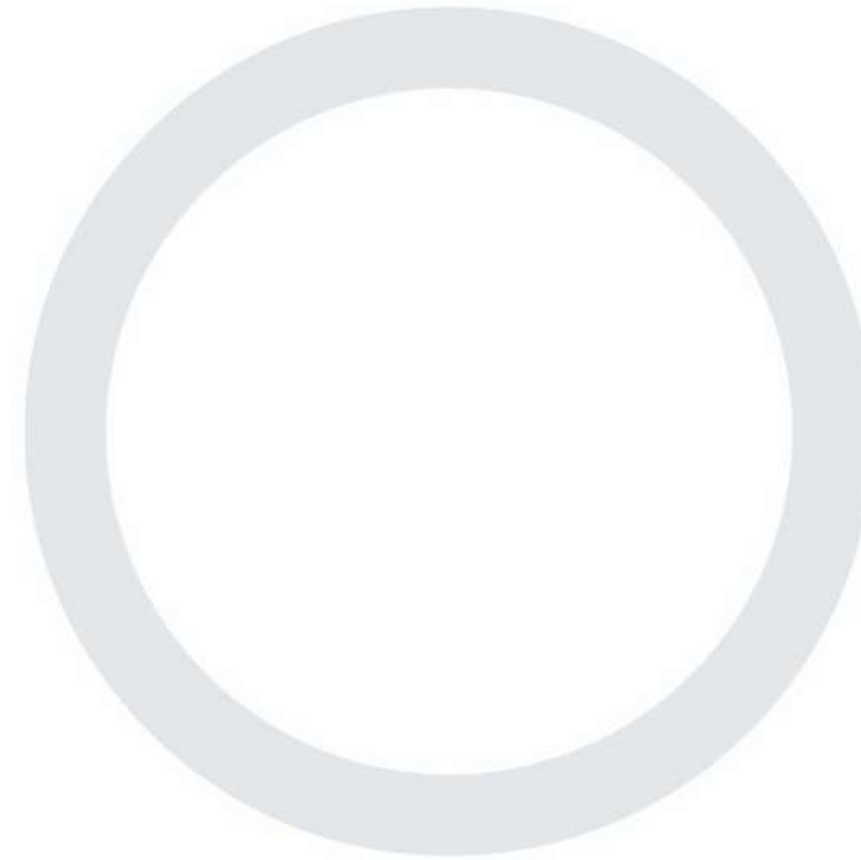
Transferring a PB to the Cloud

Typical Network
100Mbps



0 Days

Transfer Appliance



0 Days

Transfer Appliance greatly speeds data transfer rates

	1 Mbps	10 Mbps	100 Mbps	1 GB	10 GB	100 GB
1 GB	3 hrs	18 mins	2 mins	11 secs	1 sec	.1 secs
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	 16 days		30 hrs	3 hrs
1 PB	340 yrs	34 years	 43 days		12 days	30 hrs
10 PB	3,404 yrs	340 yrs	34 years	3 years	124 days	12 days
100 PB	34,048 yrs	3,404 yrs	340 yrs	34 years	3 years	124 days

Case Studies

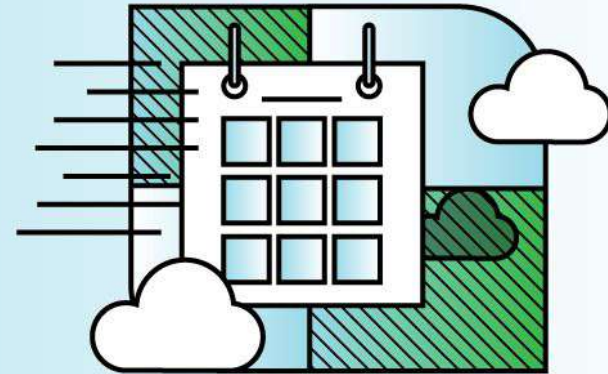
**Google Transfer
Appliance**

Case Studies

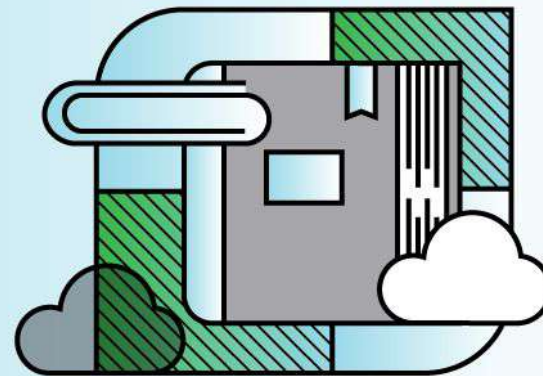
Evernote



In just
70 days over 3 petabytes
moved to the Cloud.



**5 billion notes
& 5 billion attachments**
moved to the Cloud.



Case Studies

Makani

Transfer Appliance
moves petabytes of data
for Makani to help them
find where the wind is
the most windy.



MAKANI

Course 2: Production ML Systems

Module 2: Ingesting data for Cloud-based analytics and ML

Lesson Title: **Data on Other Clouds**

Presenter: Val

Format: Talking Head

Video Name: T-PSML-0_2_I4_data_on_other_clouds

Agenda

Migration Overview

Data On-Premise

Large Datasets

Data on other clouds

Existing Databases

Cloud Storage Transfer Service



Create a transfer job

You can transfer data to your Cloud Storage bucket from a source you specify here. Required permissions: You must be a project owner and destination bucket owner, and you need read access to the source. [Learn more](#)

1 Select source

- ☒ Google Cloud Storage bucket
- ☐ Amazon S3 bucket
- ☐ List of object URLs

You must have read access to the source bucket.

Cloud Storage bucket

 bucket

 [Specify file filters](#)

2 Select destination

3 Configure transfer

Creating a transfer job grants a Cloud Storage Transfer Service account the necessary source, destination, and project permissions to complete the transfer. Your permissions will update to reflect this change.


Create a transfer job

You can transfer data to your Cloud Storage bucket from a source you specify here. Required permissions: You must be a project owner and destination bucket owner, and you need read access to the source. [Learn more](#)

☒ Select source 

☒ Select destination 

3 Configure transfer 

Schedule 

☒ Run now

☐ Run daily at 8:34:37 AM

Description

product-data-exports-dataprep-staging-8c69a357-1e33-4958-8f55-

Choose a unique description to help identify your transfer.

Create

Cancel

Creating a transfer job grants a Cloud Storage Transfer Service account the necessary source, destination, and project permissions to complete the transfer. Your permissions will update to reflect this change.

Course 2: Production ML Systems

Module 2: Ingesting data for Cloud-based analytics and ML

Lesson Title: **Existing Databases**

Presenter: Val

Format: Talking Head

Video Name: T-PSML-0_2_I5_existing_databases

Agenda

Migration Overview

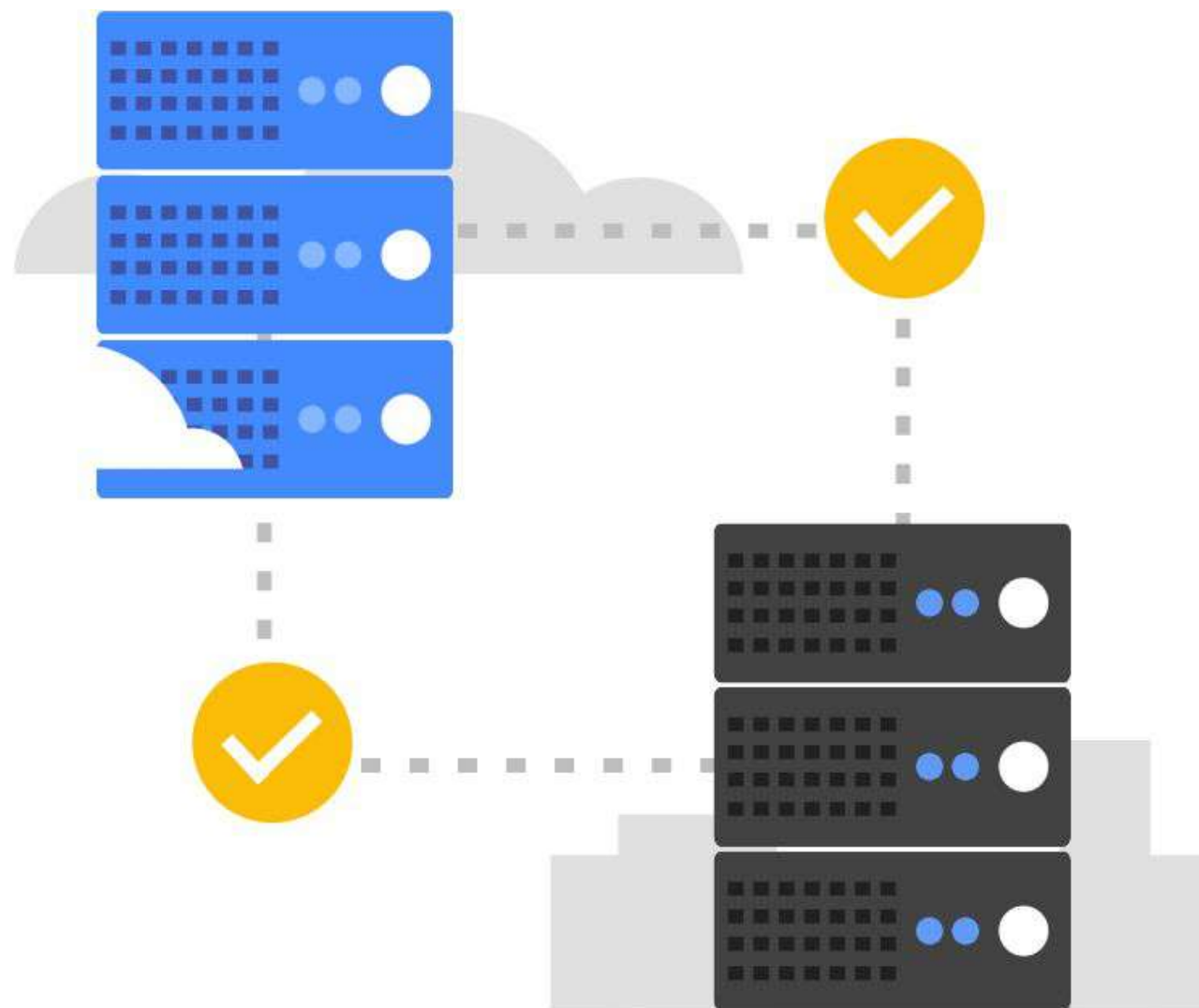
Data On-Premise

Large Datasets

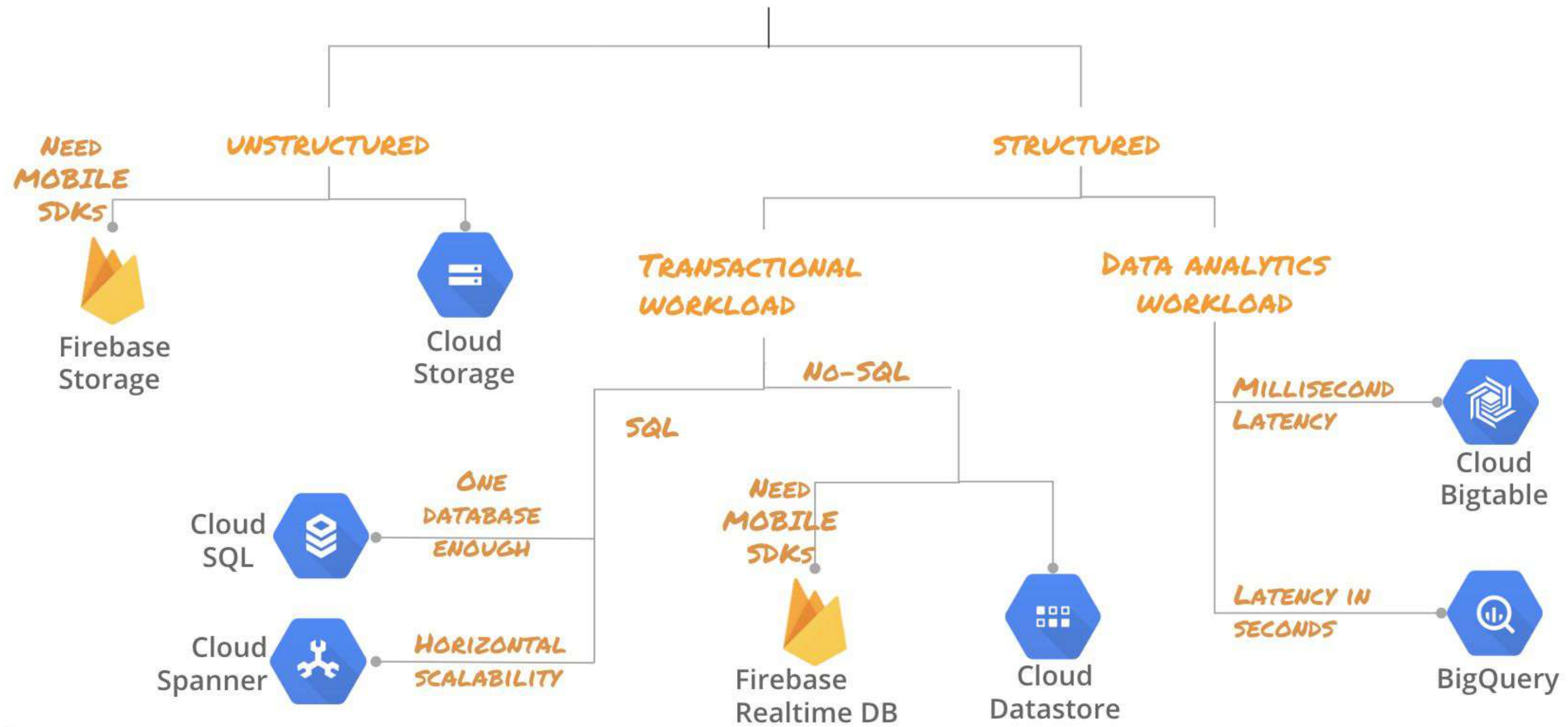
Data on other clouds

Existing Databases

Migrate your existing database to Google Cloud Platform

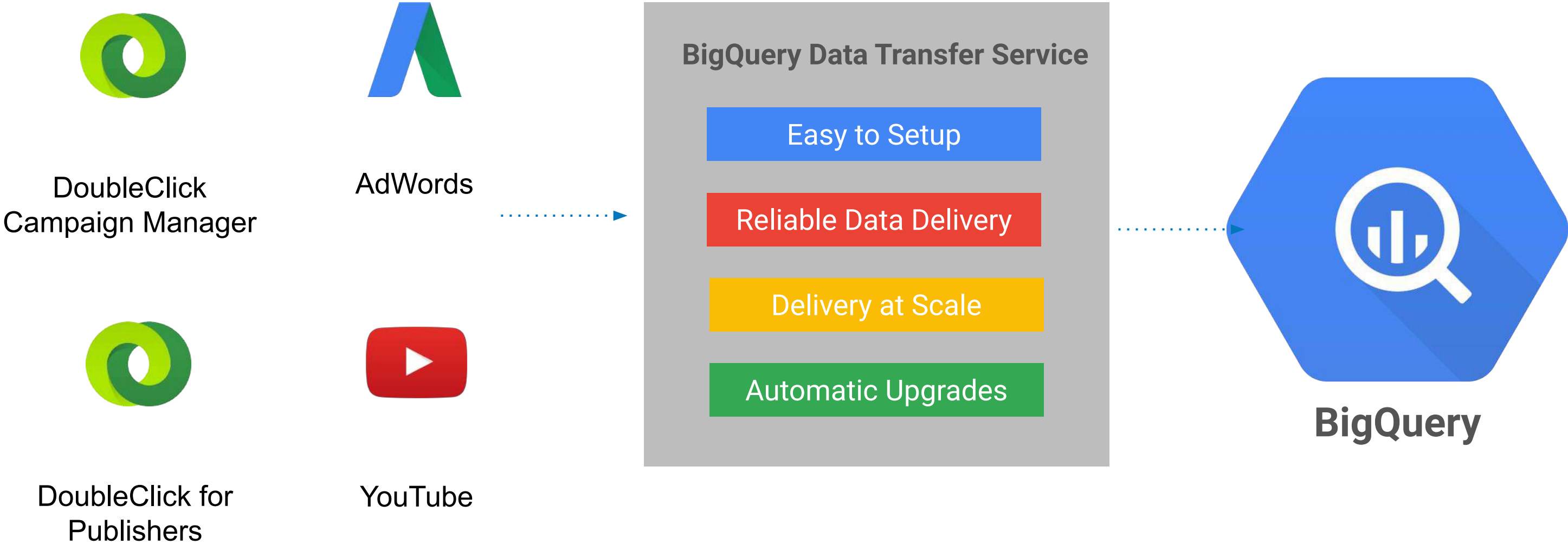


Choosing where your data should be stored

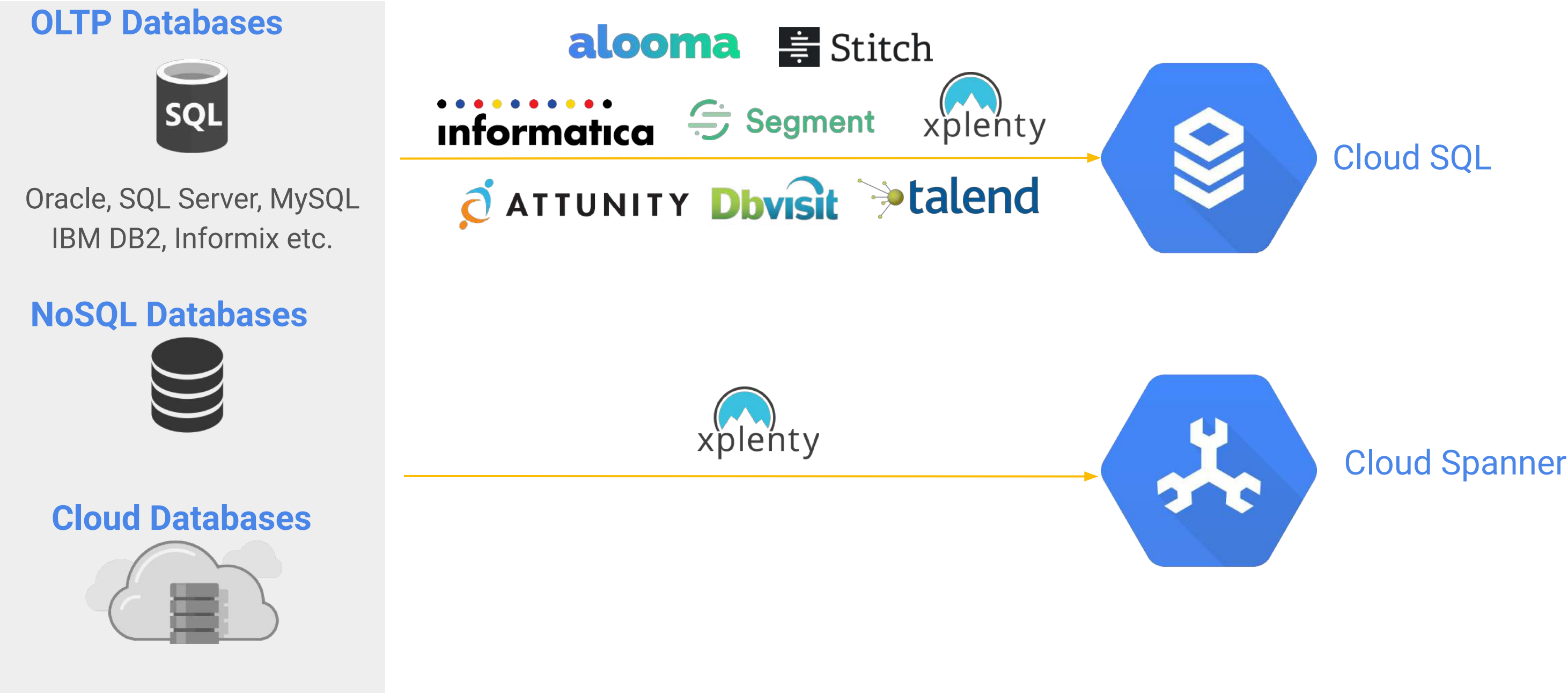


BigQuery Data Transfer Service

Fully managed data import service for Google BigQuery



Migrate Databases to Cloud SQL/Spanner/Bigtable



Migrate Hadoop and HDFS to Cloud Dataproc

On-prem HDFS



Apache Hadoop, Cloudera,
Hortonworks, MapR

Cloud Object Stores



S3, Azure Blob Storage



Cloud Dataproc

Course 2: Production ML Systems

Module 2: Ingesting data for Cloud-based analytics and ML

Lesson Title: **Demo: Load Data into BigQuery**

Presenter: Val

Format: Talking Head

Video Name: T-PSML-0_2_I6_demo:_load_data_into_bigquery

Demo

Loading Data into BigQuery

Ingesting Data into BigQuery

Types: [CSV, JSON, AVRO,
ORC, Parquet]



BigQuery

Course 2: Production ML Systems

Module 2: Ingesting data for Cloud-based analytics and ML

Lesson Title: **Demo: Automatic ETL Pipelines into GCP**

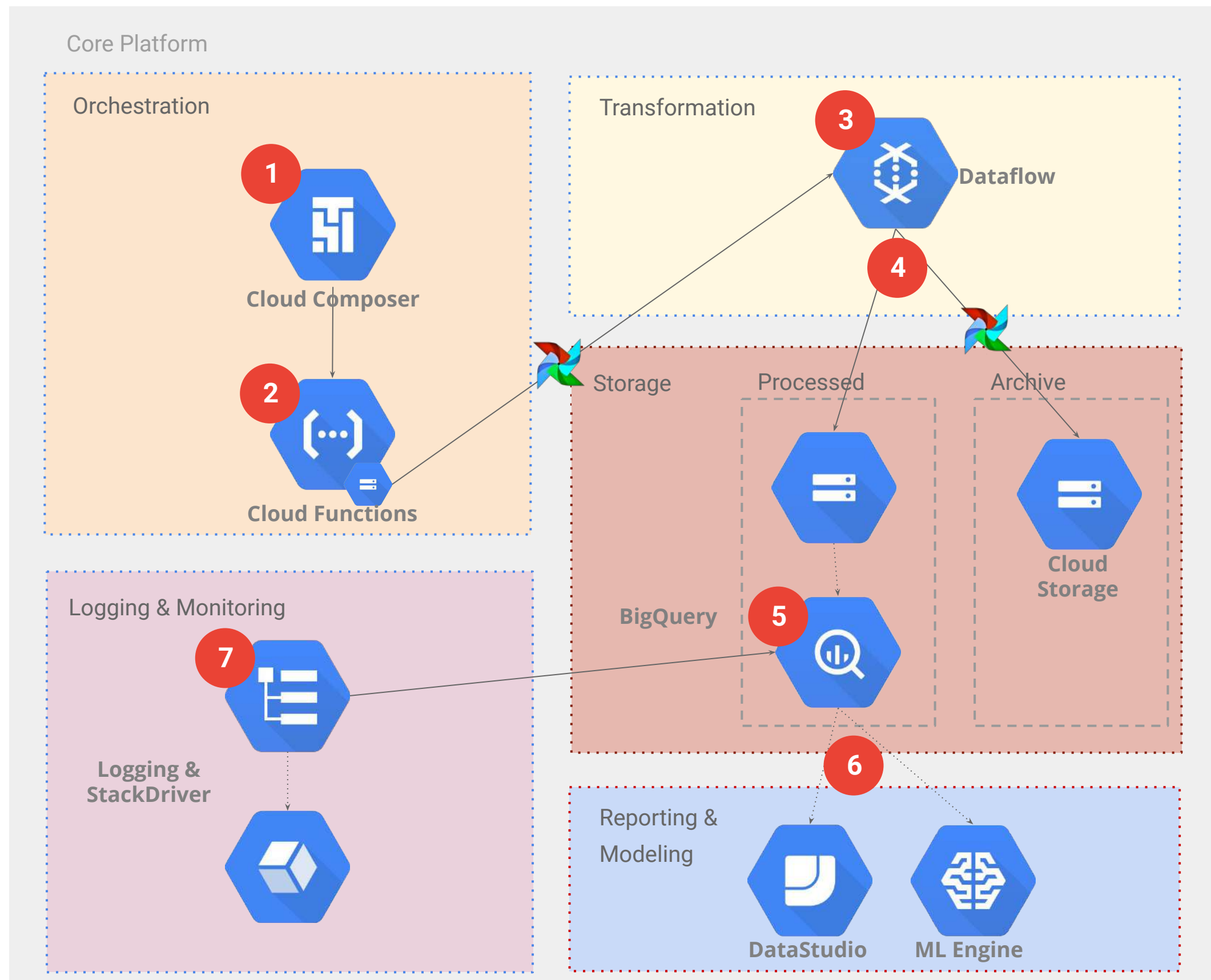
Presenter: Tony

Format: Talking Head

Video Name: T-PSML-0_2_I7_demo:_automatic_etl

Automatic ETL Pipelines into GCP

ETL Pattern 1: Push Solution Architecture



ETL Pattern 2: Pull Solution Architecture

