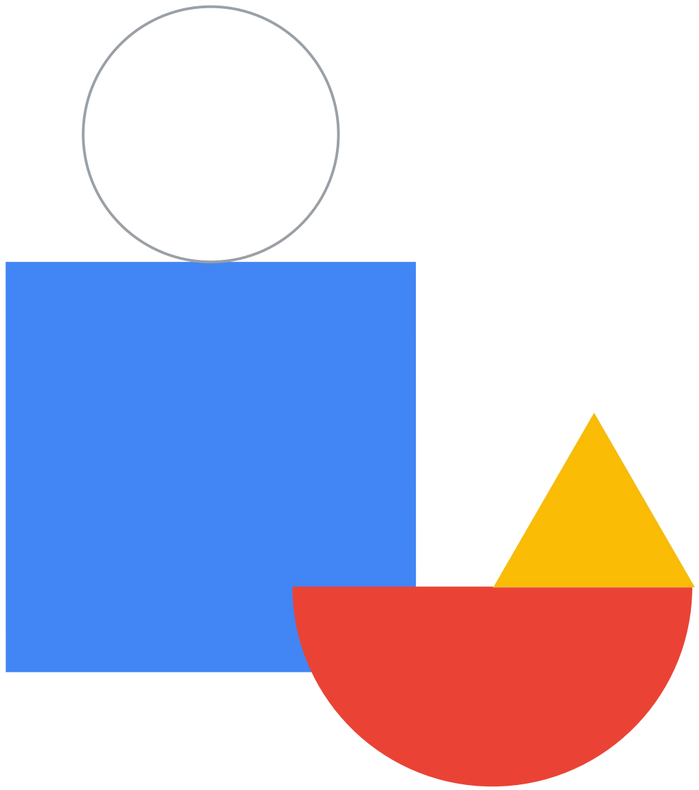
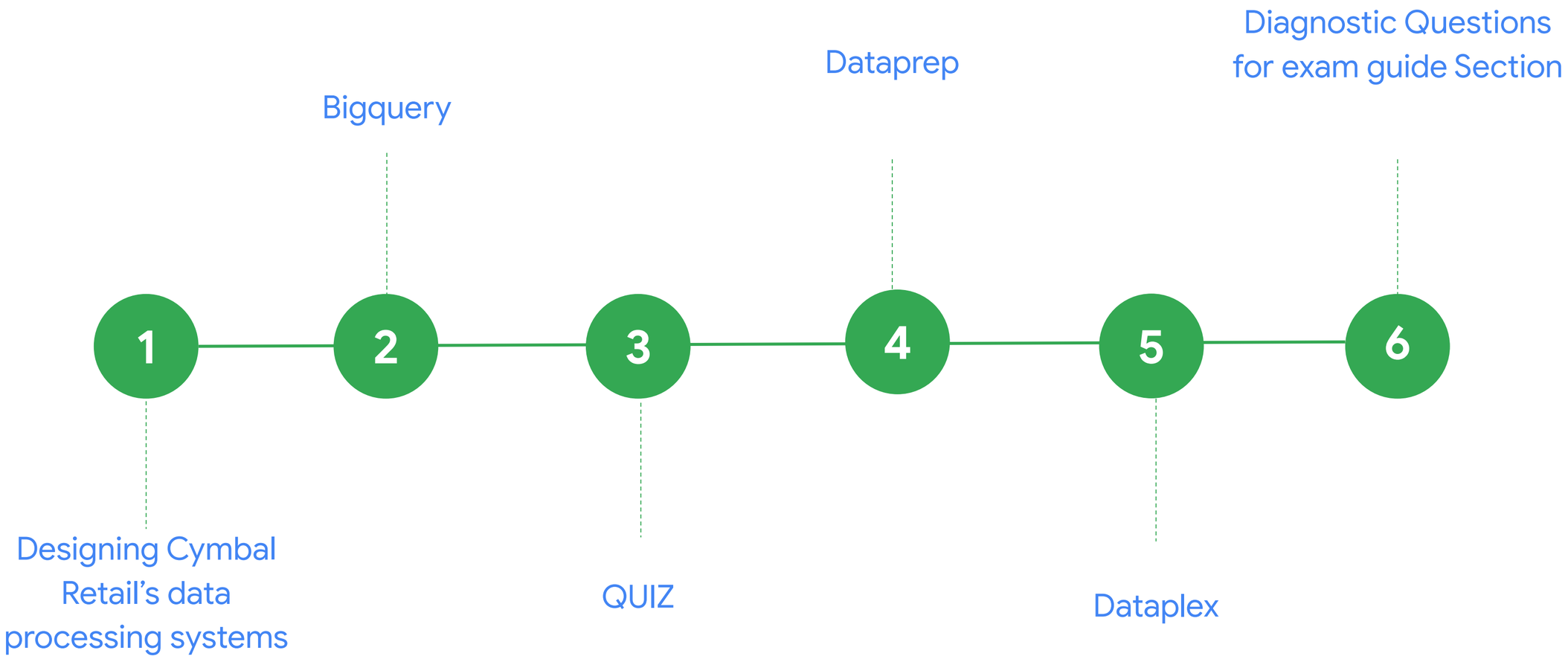
Google Cloud



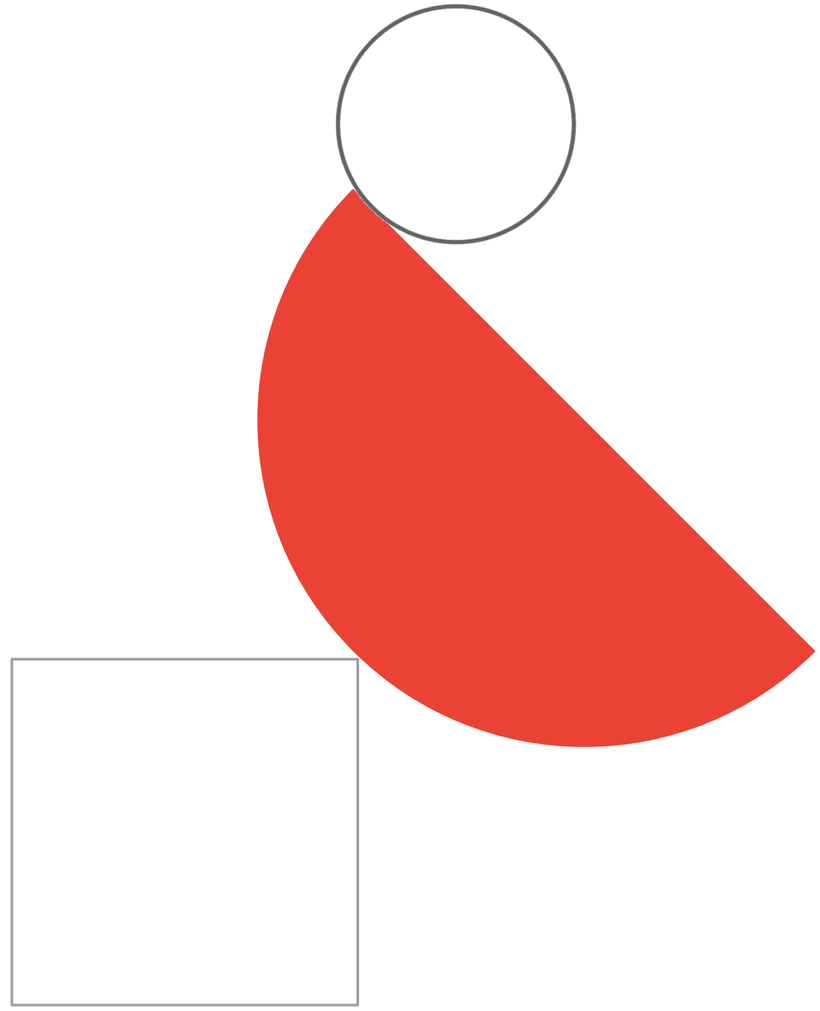
Preparing for Your Professional Data Engineer Journey

Module 1: Designing Data Processing Systems

Week 2 agenda



Designing Cymbal Retail's data processing systems

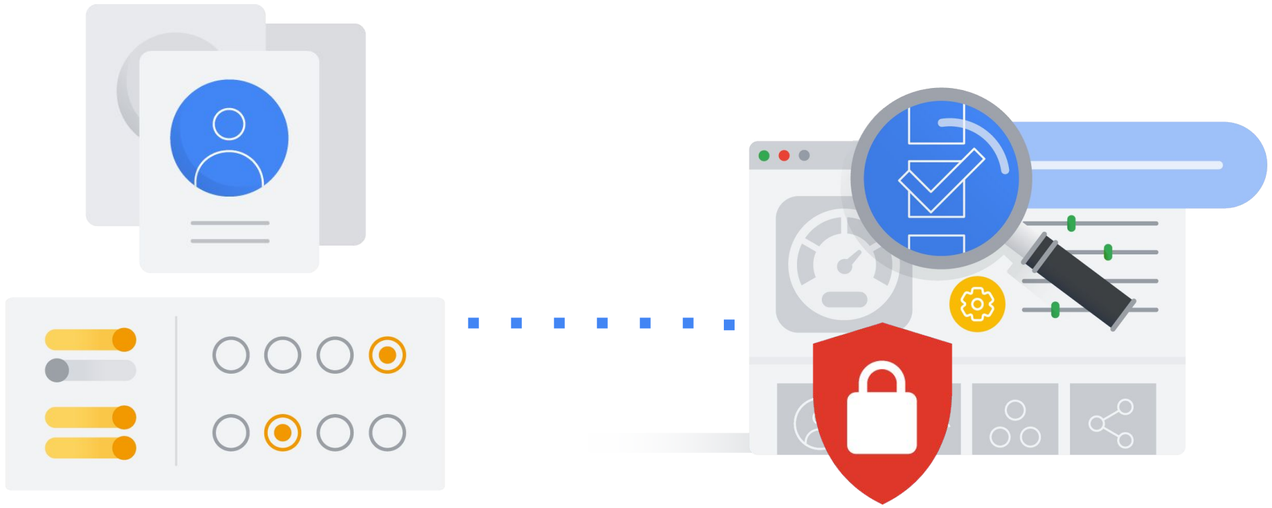


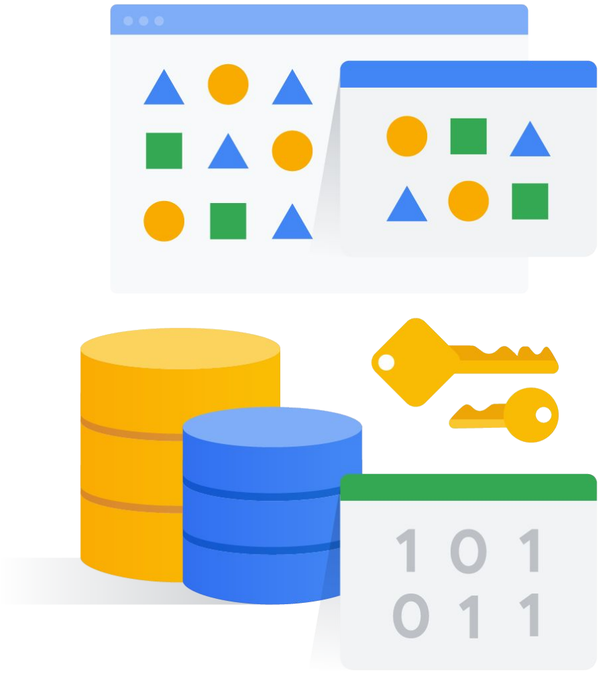
Migrating Cymbal's data to Google Cloud

You need to help migrate Cymbal Retail's existing data and data processing systems to Google Cloud.

Managing and securing data

- Uniform approach to managing and securing data
- Controlled access at a granular level





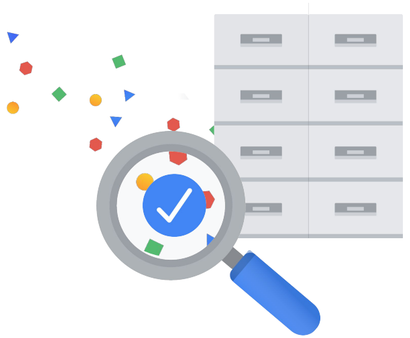
Encrypting and redacting data

- You need to understand and deploy options for data encryption.
- You need to employ tools to redact information to meet requirements.

Moving data from external sources to Google Cloud

Some data movement may occur over days.

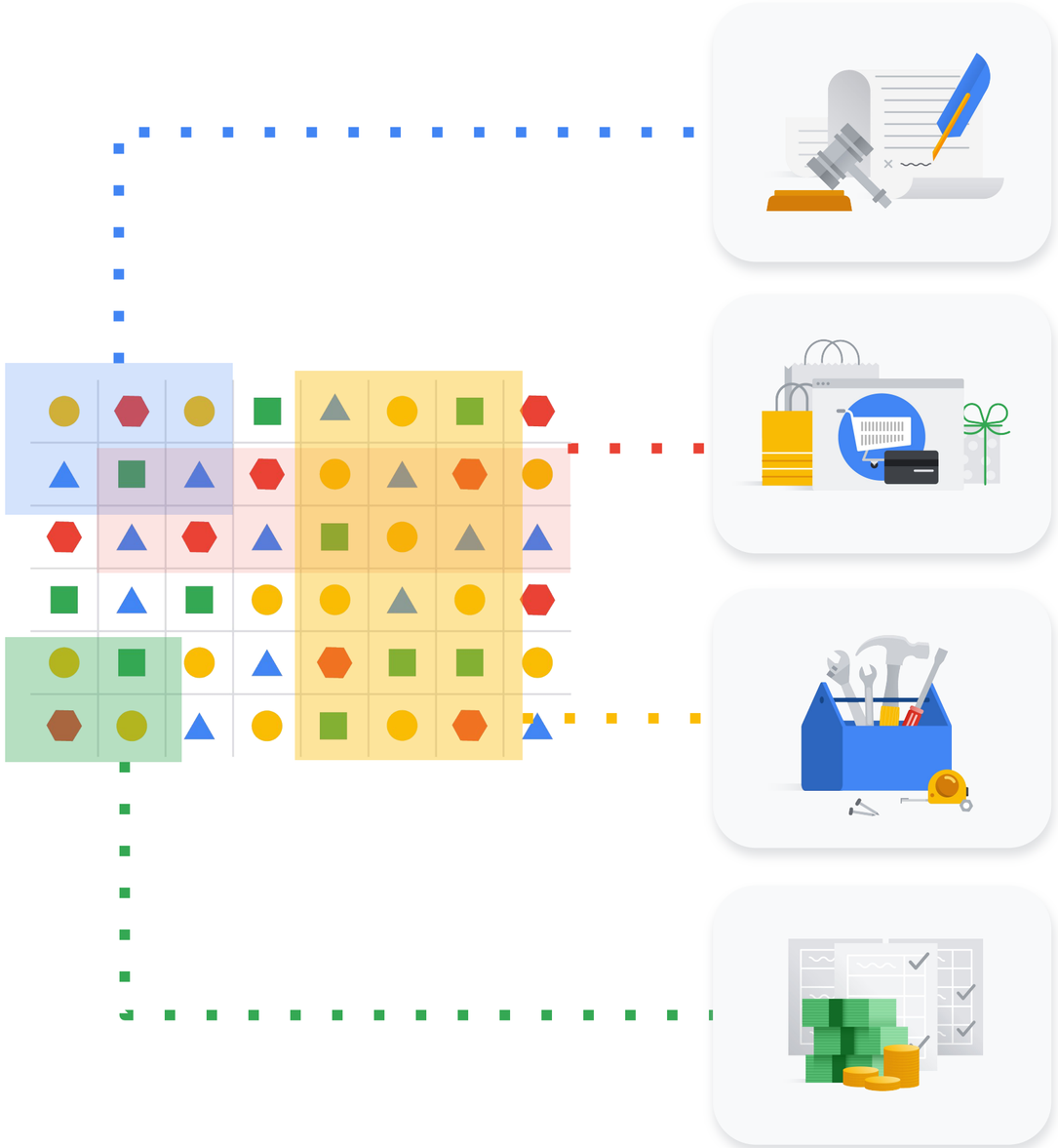
Some data needs to be moved immediately.



Data needs to be assessed for quality and cataloged for easy discoverability.

Ensuring fined-grained access to data

Access to data should be on a need-to-know basis.



Bigquery

What is BigQuery ?

Google Cloud's **enterprise data warehouse** for analytics

Built-in **ML and GIS**
Unique!

Fully managed and **serverless**
Unique!

Gigabyte- to **petabyte-scale** storage and SQL queries

Encrypted, durable, and highly available

Real-time analytics on streaming data
Unique!

BigQuery features

BigQuery

01 **Storage *plus* analytics**

A place to store petabytes of data.
(Petabyte = 11,000 4k movies)

A place to analyze data.
(Machine learning, geospatial analysis, and business intelligence)

02 **Serverless**

Free from provisioning resources or managing servers but focus on SQL queries.

03 **Flexible pay-as-you-go pricing model**

Pay for what your query processes, or use a flat-rate option.

04 **Data encryption at rest by default**

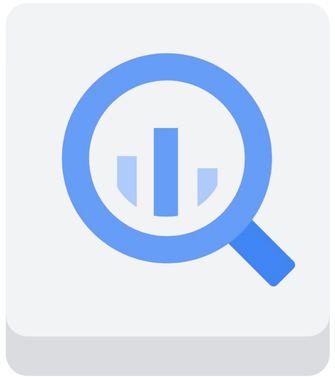
Encrypt data stored on a disk.

05 **Built-in machine learning**

Write ML models directly in BigQuery using SQL.

BigQuery has many capabilities that make it an ideal data warehouse

- Interactive SQL queries over large datasets (petabytes) in seconds
- Serverless and no-ops, including ad hoc queries
- Ecosystem of visualization and reporting tools
- Ecosystem of ETL and data processing tools
- Up-to-the-minute data
- Machine learning
- Security and collaboration



BigQuery

BigQuery charges for data processed + storage

BigQuery job types:

- **Query** - charged by bytes processed
- **Load Data** - free
- **Extract** - free
- **Copy** - free

This query will process 702.6MB when run. 

Note: storing data in BigQuery is a separate cost

BigQuery: Storage Pricing

Storage pricing is the cost to store data that you load into BigQuery. You pay for *active storage* and *long-term* storage.

- **Active storage** includes any table or table partition that has been modified in the last 90 days.
- **Long-term storage** includes any table or table partition that has not been modified for 90 consecutive days. The **price of storage for that table automatically drops by approximately 50%**. There is no difference in performance, durability, or availability between active and long-term storage.

The first 10 GB of storage per month is free.

US (multi-region)

Operation	Pricing	Details
Active storage	$0.02 per GB	The first 10 GB is free each month.
Long-term storage	$0.01 per GB	The first 10 GB is free each month.

BigQuery: Controlling access to datasets

You can grant access at the following BigQuery resource levels:

- organization or Google Cloud project level
- dataset level
- table or view level
  a. Authorized Views
- You can also restrict access to data on more granular level by using the following methods:
  a. column-level access control
  b. dynamic data masking (aka “some columns may be hidden, depending on privileges”)
     i. Works together with column-level security.
     ii. no need to modify existing queries by excluding the columns that the user cannot access
  c. row-level security (aka “some rows may be hidden, depending on privileges”)
     i. One table can have multiple row-level access policies. Row-level access policies can coexist on a table with column-level security as well as dataset-level, table-level, and project-level access controls.

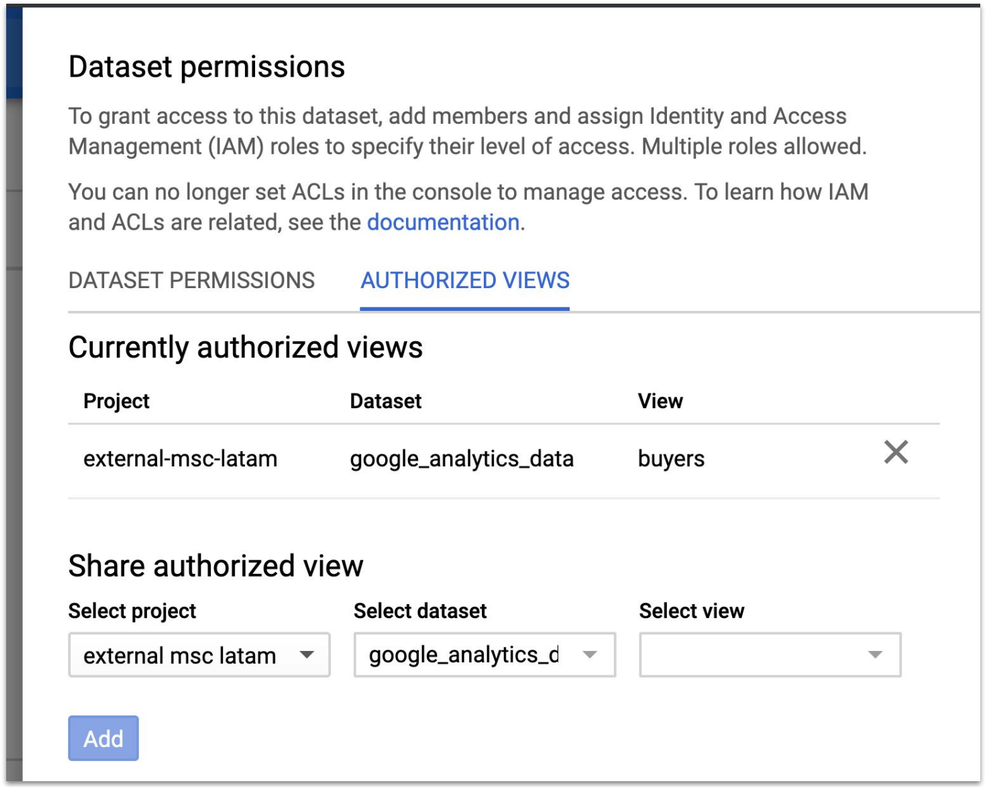
BigQuery: Controlling access to datasets

Authorized Views

1. **View:** View is a virtual table defined by a SQL query. When you create a view, you query it in the same way you query a table

2. **Query:** When a user queries the view, the query results contain data only from the tables and fields specified in the query that defines the view.

3. **Authorized Views:** An authorized view allows you to share query results with particular users and groups without giving them access to the underlying tables.

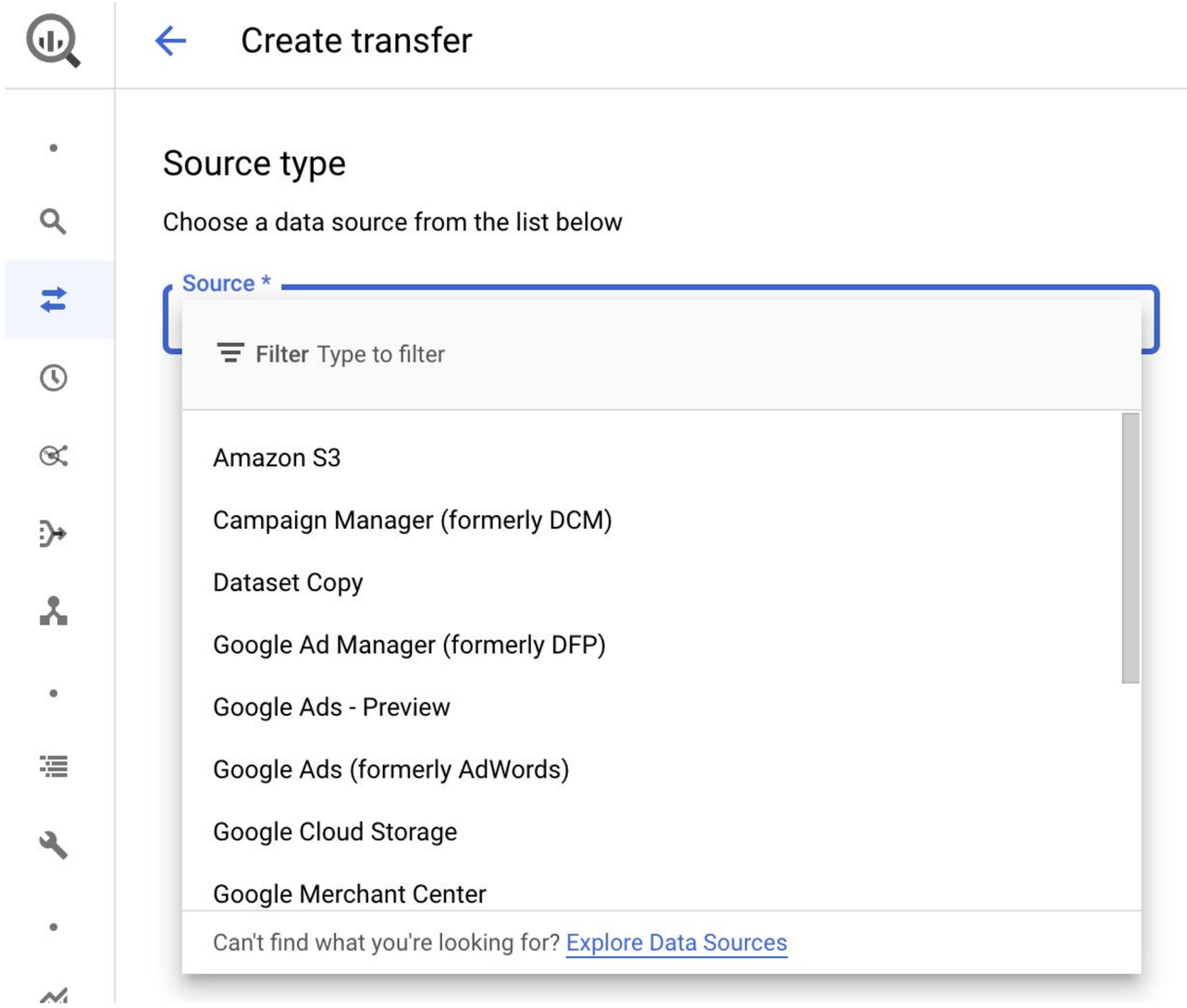


Exam Tip: *Authorized Views were especially useful when there were no table/column-level permissions. However, they're still often-used way to selectively share access to datasets (and they pop up on the exam!).*
MAKE SURE TO UNDERSTAND HOW TO CREATE AND SHARE SUCH A VIEW.

BigQuery - Data Transfer Service

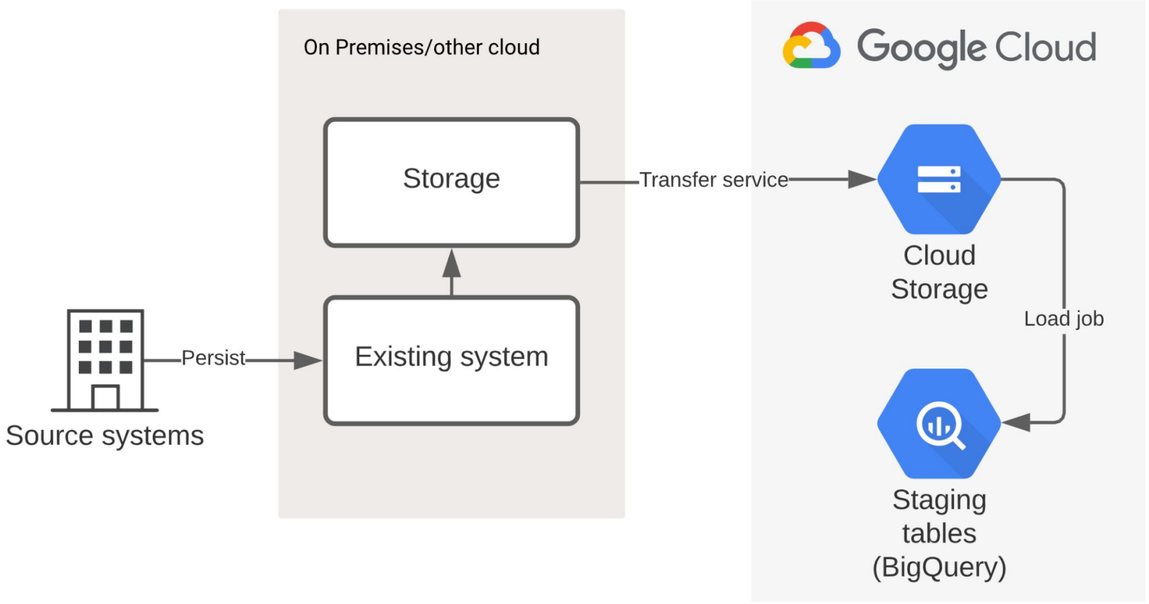
Mostly useful for regular data transfers to BigQuery

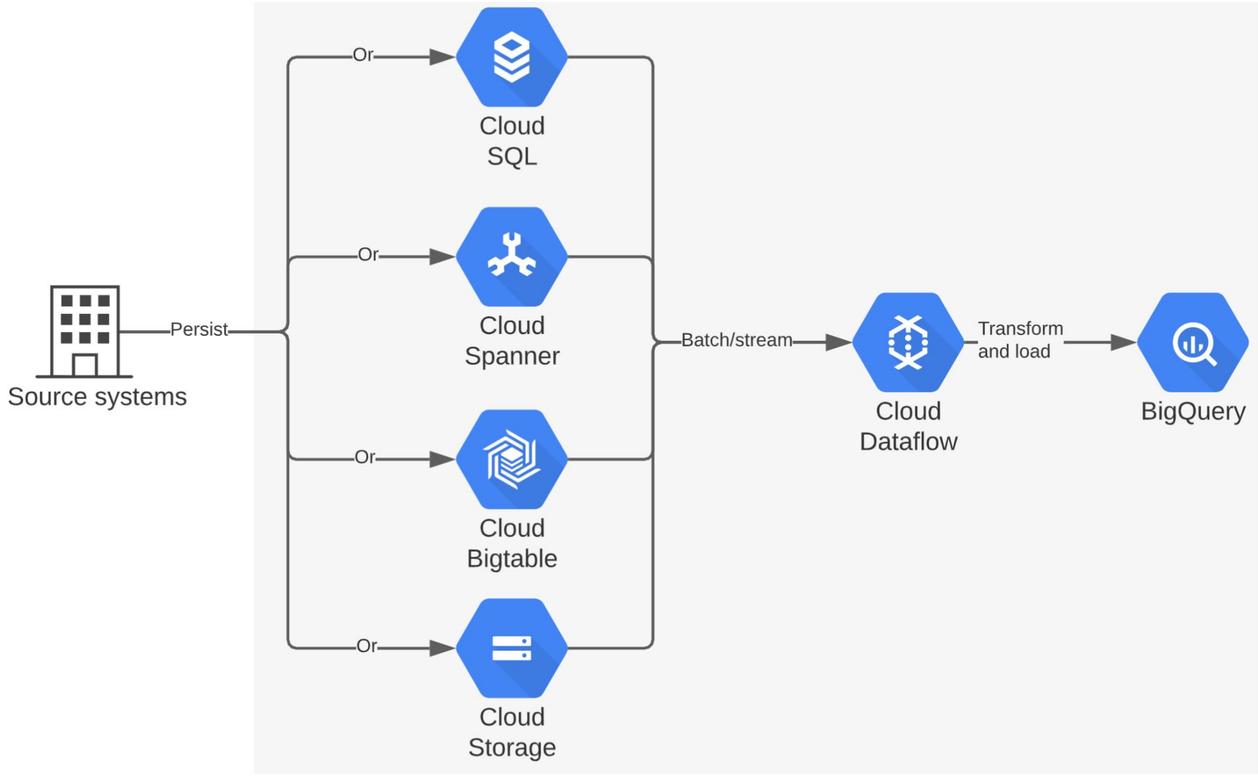
- BigQuery Data Transfer Service automates data movement from various sources into BigQuery on a scheduled, managed basis.
- You can initiate data backfills to recover from any outages or gaps.



BigQuery - Batch vs Streaming inserts

Most common architectures



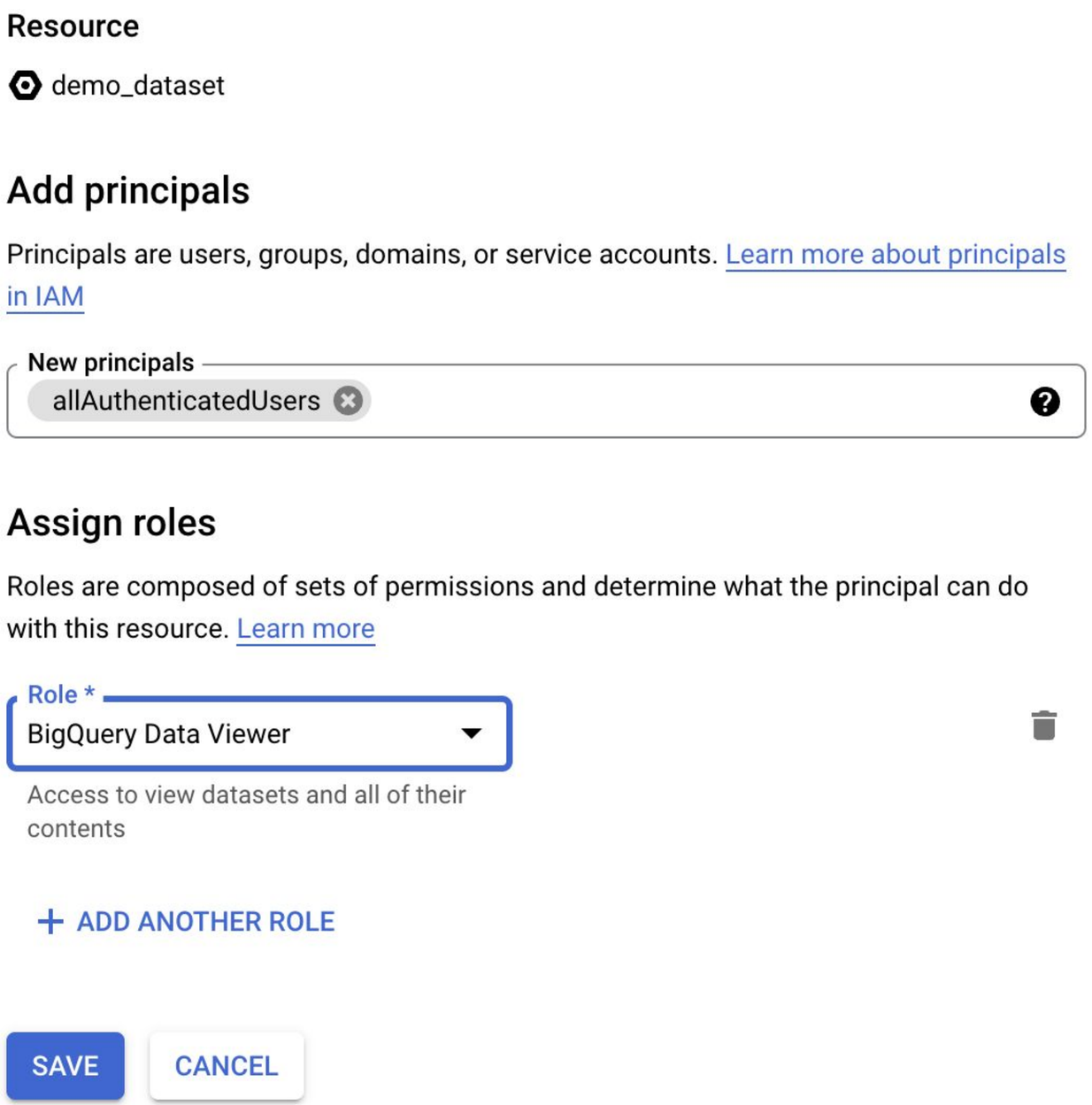


Exam Tip: *There is additional cost for streaming (both inserts and reads) in BigQuery.*

BigQuery: Sharing Datasets with others

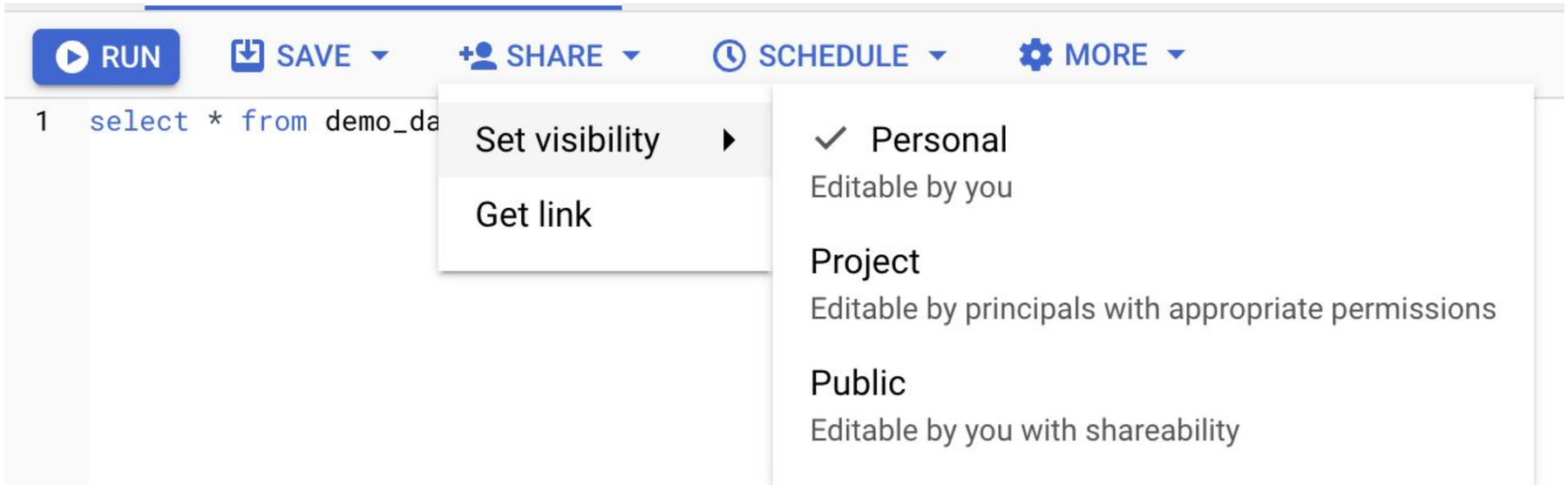
AllAuthenticatedUsers

The special setting allAuthenticatedUsers makes a dataset public. Authenticated users must use BigQuery within their own project and have access to run BigQuery jobs so that they can query the Public Dataset. The billing for the query goes to their project, even though the query is using public or shared data. In summary, the cost of a query is always assigned to the active project from where the query is executed.



BigQuery: Sharing Queries with others

Mostly for collaboration

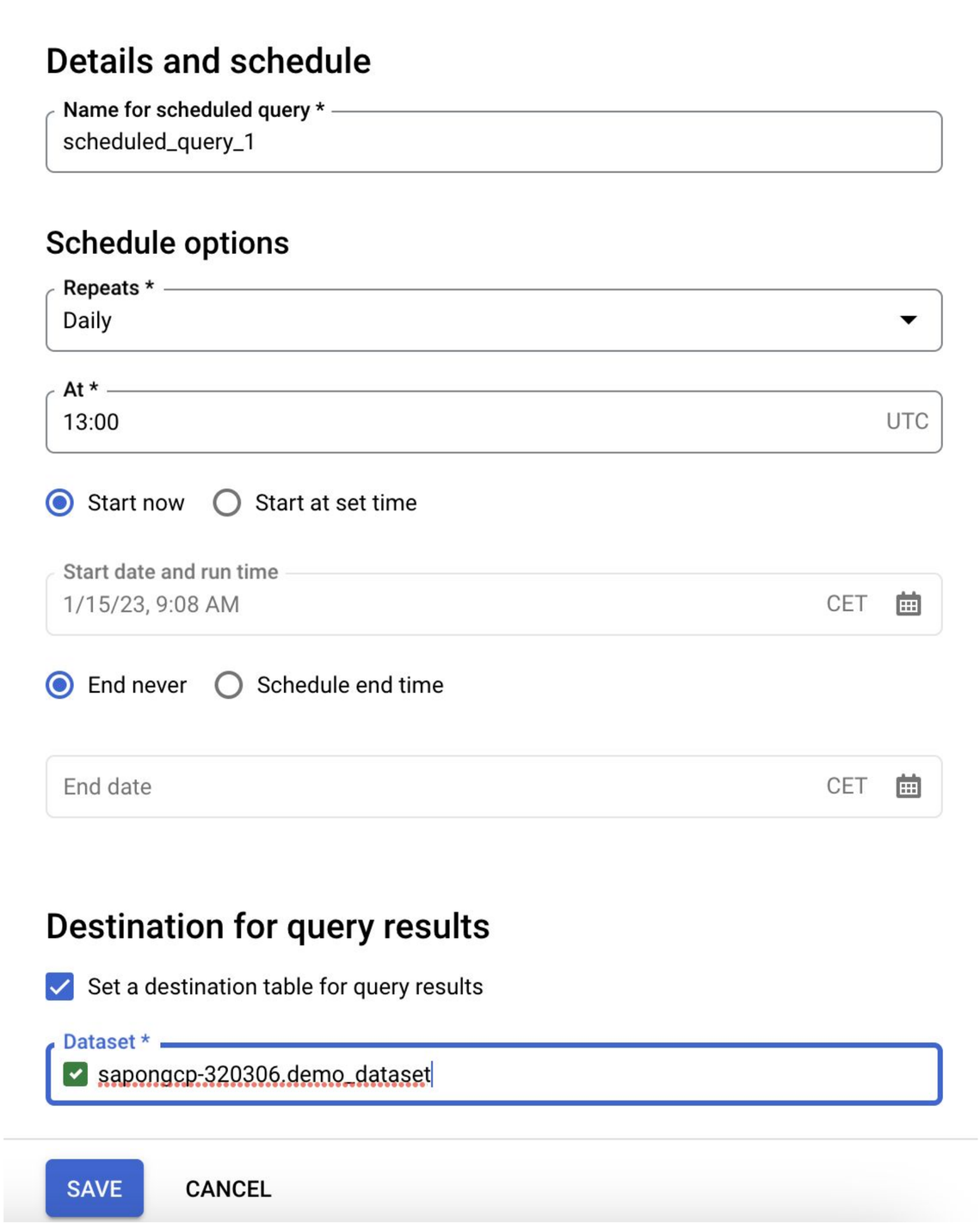


- Query needs to be saved first, before it's shared;
- Can share incomplete / invalid queries -> collaboration;
- Project-level saved queries are visible to principals with the required permissions;
- Public saved queries are visible to anyone with a link to the query;

BigQuery: Scheduling queries

Mostly useful for regular execution

- Scheduled queries use features of BigQuery Data Transfer Service.
- If the destination table for your results doesn't exist when you set up the scheduled query, BigQuery attempts to create the table for you.
- You can set up a scheduled query to authenticate as a service account.



BigQuery: Query results caching

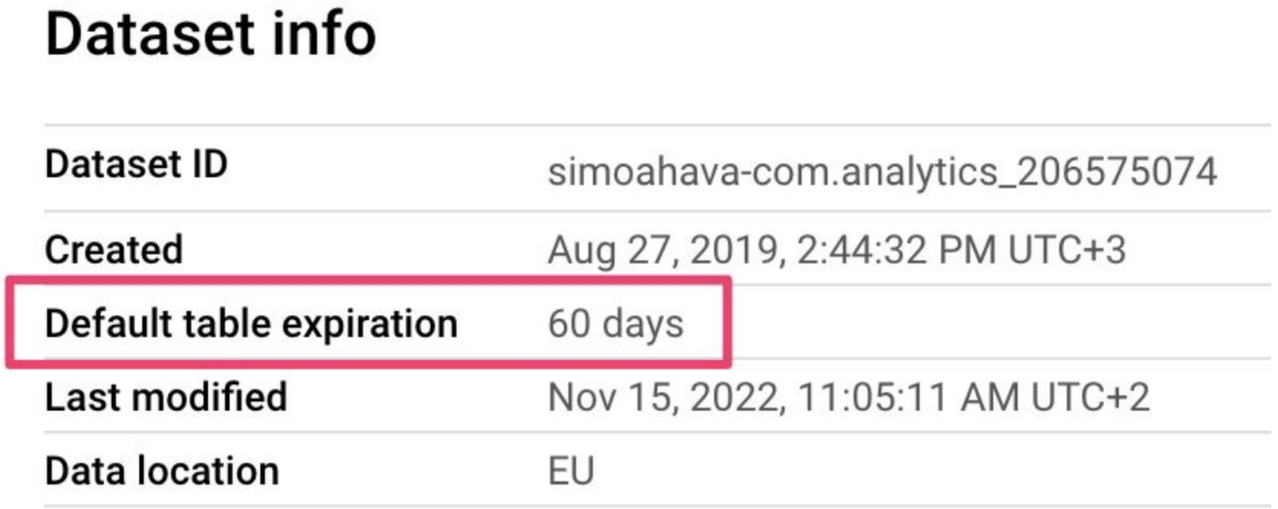
Limit Access by Data Lifecycle Stages

- Query results are cached to improve performance and reduce costs for repeated queries
- Cache is per user
- Still subject to quota policies
- Cache results have a size limit of 128 MB compressed
- No charge for queries that use cached results
- Results are cached for approximately 24 hours
- Lifetime extended when a query returns a cached result
- Use of cached results can be turned off (useful for benchmarking)

BigQuery: table/partition (automatic) data expiration

Can be set for dataset / table / partition

Best practice for data lifecycle management.
Expiration in BigQuery automatically implements retention policy.



- Dataset expiration
  - = “default table expiration time” for a dataset
- Table expiration
  - If Dataset expiration is set, each table inherits this setting by default
- Partition expiration:
  - The setting applies to all partitions in the table, but is calculated independently for each partition based on the partition time.
  - At any point after a table is created, you can update the table's partition expiration

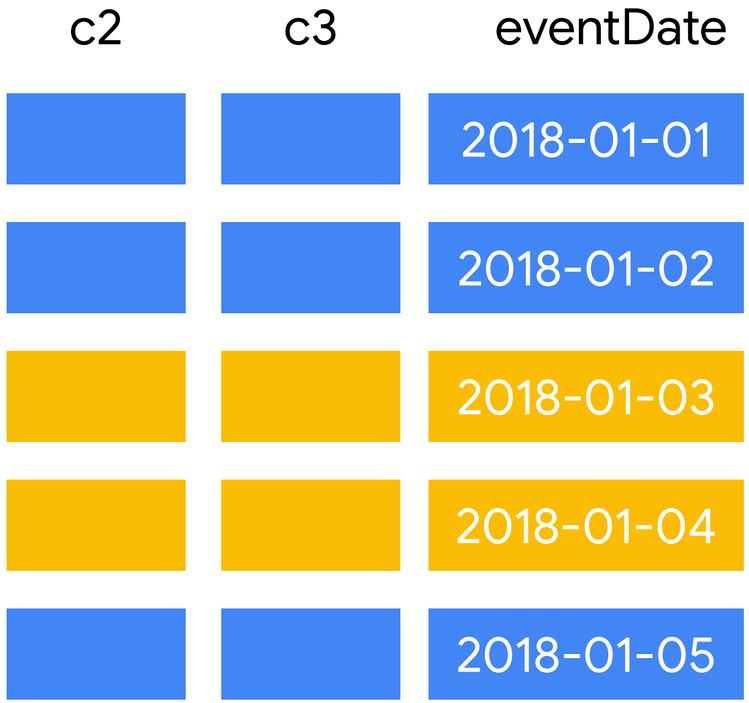
BigQuery: Table Partitioning

Partitioning versus sharding:

- Table sharding is the practice of storing data in multiple tables, using a naming prefix such as [PREFIX]_YYYYMMDD. Partitioning is recommended over table sharding, because partitioned tables perform better.

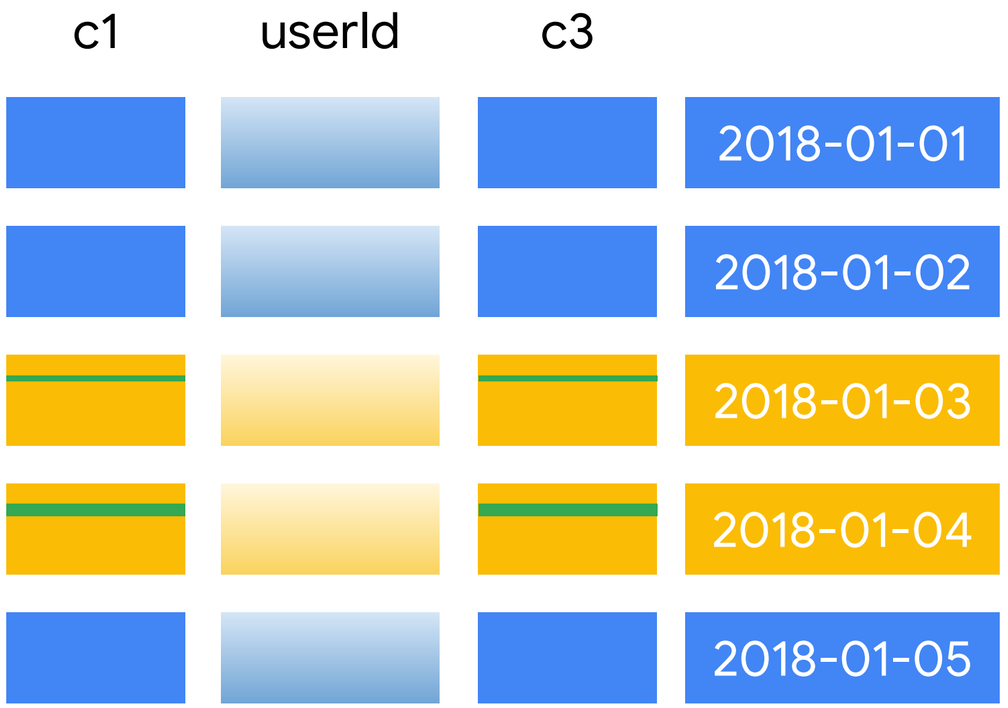
You can partition BigQuery tables by:

- Time-unit column: Tables are partitioned based on a TIMESTAMP, DATE, or DATETIME column in the table.
- Ingestion time: Tables are partitioned based on the timestamp when BigQuery ingests the data.
- Integer range: Tables are partitioned based on an integer column.




```
SELECT * FROM ...
WHERE eventDate BETWEEN
“2018-01-03” AND
“2018-01-04”
```

BigQuery: Table Clustering



SELECT c1, c3 FROM ... WHERE userId BETWEEN 52 and 63
AND eventDate BETWEEN “2018-01-03” AND “2018-01-04”

BigQuery: table partitioning AND clustering

Both partitioning and clustering can improve performance and reduce query cost

Orders table

Not Clustered; Not partitioned

Order_Date	Country	Status
2022-08-02	US	Shipped
2022-08-04	JP	Shipped
2022-08-05	UK	Canceled
2022-08-06	KE	Shipped
2022-08-02	KE	Canceled
2022-08-05	US	Processing
2022-08-04	JP	Processing
2022-08-04	KE	Shipped
2022-08-06	UK	Canceled
2022-08-02	UK	Processing
2022-08-05	JP	Canceled
2022-08-06	UK	Processing
2022-08-05	US	Shipped
2022-08-06	JP	Processing
2022-08-02	KE	Shipped
2022-08-04	US	Shipped

Orders table

Clustered by Country; Not partitioned

Order_Date	Country	Status
2022-08-04	JP	Shipped
2022-08-04	JP	Processing
2022-08-05	JP	Canceled
2022-08-06	JP	Processing
2022-08-06	KE	Shipped
2022-08-02	KE	Canceled
2022-08-04	KE	Shipped
2022-08-02	KE	Shipped
2022-08-05	UK	Processing
2022-08-06	UK	Canceled
2022-08-02	UK	Canceled
2022-08-06	UK	Processing
2022-08-02	US	Shipped
2022-08-05	US	Processing
2022-08-05	US	Shipped
2022-08-04	US	Shipped

Orders table

Clustered by Country; Partitioned by Order_Date (Daily)

	Order_Date	Country	Status
Partition: 2022-08-02	2022-08-02	KE	Shipped
	2022-08-02	KE	Canceled
Clusters: Country	2022-08-02	UK	Processing
	2022-08-02	US	Shipped

	Order_Date	Country	Status
Partition: 2022-08-04	2022-08-04	JP	Shipped
	2022-08-04	JP	Processing
Cluster: Country	2022-08-04	KE	Shipped
	2022-08-04	US	Shipped

	Order_Date	Country	Status
Partition: 2022-08-05	2022-08-05	JP	Canceled
	2022-08-05	UK	Canceled
Cluster: Country	2022-08-05	US	Shipped
	2022-08-05	US	Processing

	Order_Date	Country	Status
Partition: 2022-08-06	2022-08-06	JP	Processing
	2022-08-06	KE	Shipped
Cluster: Country	2022-08-06	UK	Canceled
	2022-08-06	UK	Processing

Exam Tip: *You can combine partitioning with clustering. Data is first partitioned and then data in each partition is clustered by the clustering columns.*

BigQuery - table partitioning vs clustering

Decision making

- Clustering gives you more granularity than partitioning alone allows
- Use clustering if your queries commonly use filters or aggregation against multiple particular columns.

Use case	Recommendation
You're using on-demand pricing and require strict cost guarantees before running queries.	Partitioned tables
Your segment size is less than 1 GB after partitioning the table.	Clustered tables
You require a large number of partitions beyond the BigQuery limits	Clustered tables
Frequent mutations in your data modify a large number of partitions.	Clustered tables
You frequently run queries to filter data on certain fixed columns.	Partitions plus clustering

Use **Analytic Window Functions** for advanced analysis

- Standard aggregations
  - `SUM`, `AVG`, `MIN`, `MAX`, `COUNT`, `etc.`
- Navigation functions
  - `LEAD()` – Returns the value of a row ***n*** rows ahead of the current row
  - `LAG()` – Returns the value of a row ***n*** rows behind the current row
  - `NTH_VALUE()` – Returns the value of the ***n***th value in the window
- Ranking and numbering functions
  - `CUME_DIST()` – Returns the cumulative distribution of a value in a group
  - `DENSE_RANK()` – Returns the integer rank of a value in a group
  - `ROW_NUMBER()` – Returns the current row number of the query result
  - `RANK()` – Returns the integer rank of a value in a group of values
  - `PERCENT_RANK()` – Returns the rank of the current row, relative to the other rows in the partition

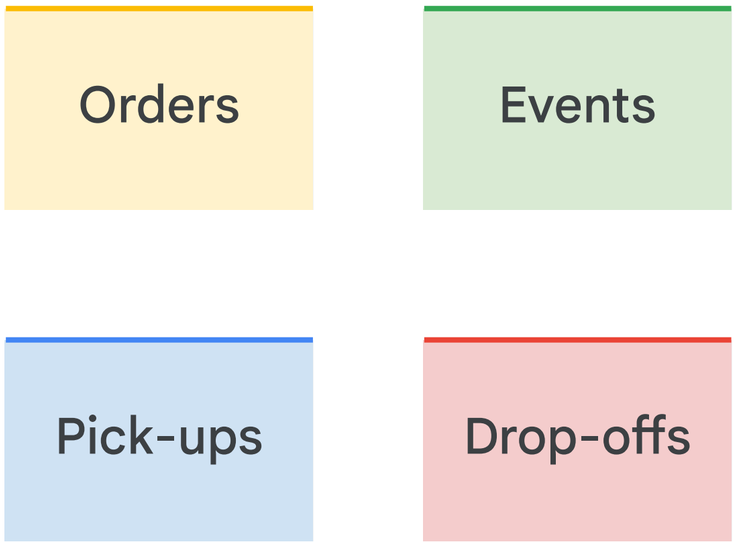
Components of a User-Defined Function (UDF)

- CREATE FUNCTION.
  Creates a new function. A function can contain zero or more named_parameters
- RETURNS [data_type].
  Specifies the data type that the function returns.
- Language [language].
  Specifies the language for the function.
- AS [external_code].
  Specifies the code that the function runs.

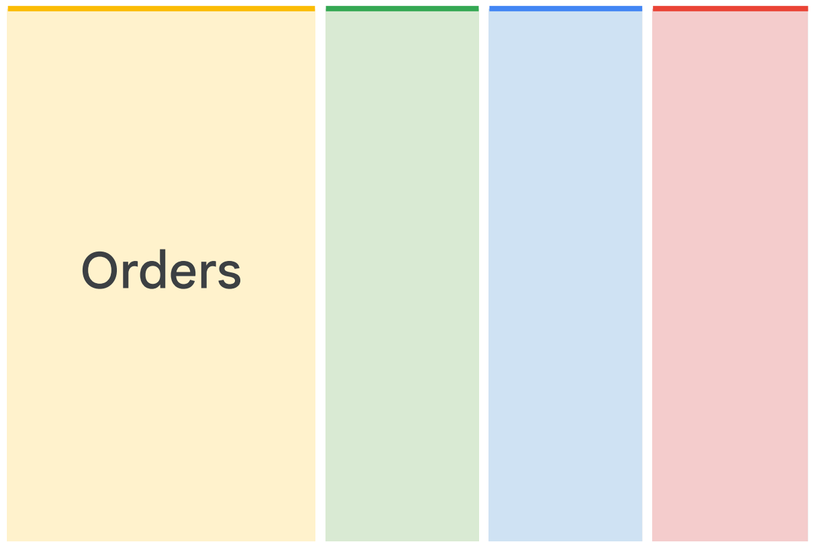
```
CREATE FUNCTION d2i_demo.nlp_compromise_people(str STRING)
RETURNS ARRAY<STRING> LANGUAGE js AS '''
  return nlp(str).people().out('topk').map(x=>x.normal)
'''
OPTIONS (
library="gs://cloud-training/datatoinsights/assets/compromise.min.11.14.0.js");

SELECT name, COUNT(*) c
FROM ( SELECT d2i_demo.nlp_compromise_people(title) names
  FROM `d2i_demo.reddit_posts`
  WHERE subreddit = 'movies'
), UNNEST(names) name
WHERE name LIKE '% %' GROUP BY 1
ORDER BY 2 DESC LIMIT 10
```

Row	name	c
1	captain marvel	198
2	will smith	118
3	ary digital	109
4	star wars	84

Should we normalize or denormalize?



VS

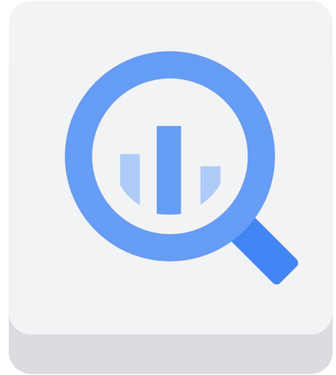


JOINs are costly

Data is repeated

BigQuery best practices (1)

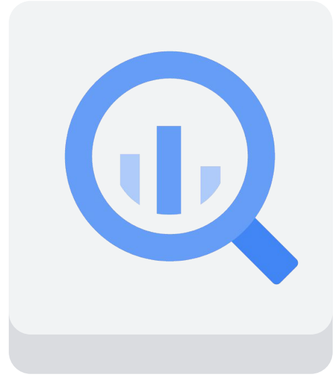
- Avoid `SELECT *`, use only the columns you need, use Preview tab (it's free)
- Filter using `WHERE` as early as possible in your queries
- Be aware of data skew in your dataset
  - Filter your dataset as early as possible (this avoids overloading workers on JOINs).
  - Hint: Use the Query Explanation map and compare the Max vs. the Avg times to highlight skew.
  - BigQuery will automatically attempt to reshuffle workers that are overloaded with data.



BigQuery

BigQuery best practices (2)

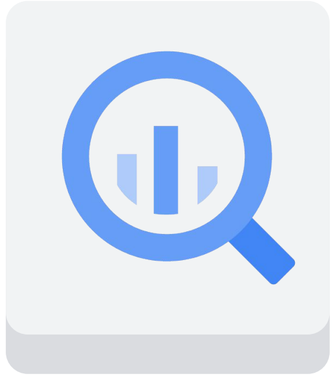
- Optimize your join patterns
- Use ORDER BY only in the outermost query or within window clauses. Push complex operations to the end of the query.
- If you need to sort data, filter first to reduce the number of values that you need to sort. If you sort your data first, you sort much more data than is necessary. It is preferable to sort on a subset of data than to sort all the data and apply a LIMIT clause*.
- Optimize aggregation functions (e.g. use an approximate aggregation function)



BigQuery

BigQuery best practices (3)

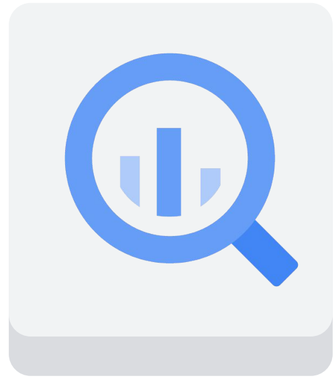
- Use cached results. Cached results are free (*)
  - Cached results will be available if there is no non-deterministic results (e.g. if you are not using functions like e.g. CURRENT_TIMESTAMP)
  - Cached results will not be used if data in underlying tables or partitions have changed.
  - Each query modification will trigger a new query (even a comment or space in the query)
  - Cached results are available only for a single user (each user have a separate cache for safety reasons)



BigQuery

BigQuery best practices (4)

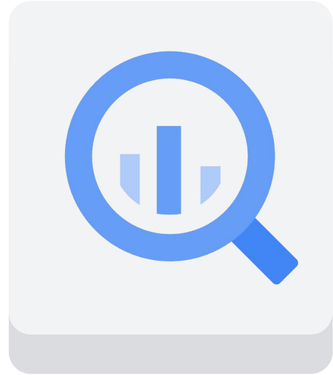
- Use the right function for the right job
  - String Manipulation Functions - FORMAT( ) ...
  - Aggregation Functions
  - Data Type Conversion Functions
  - Date Functions
  - Statistical Functions
  - Analytic Functions
  - User-defined Functions



BigQuery

BigQuery best practices (5)

- Use UDF only when it’s needed
  - Native functions will always be faster
  - There is lower concurrency limit for UDF functions
    - - for non-UDF queries: 50
    - - for UDF-queries: 6
- Be careful with JOIN and UNION - it might use a lot of data and create huge results sets.
  (e.g. it’s a great feature to use _TABLE_SUFFIX for sharded tables, but it might be expensive)



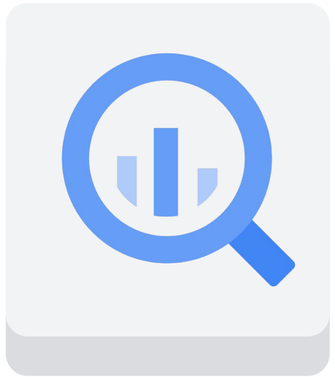
BigQuery

BigQuery best practices (6)

- Use GROUP BY wisely
  - Best when the number of distinct groups is small (fewer shuffles of data).
  - Grouping by a high-cardinality unique ID is a bad idea.

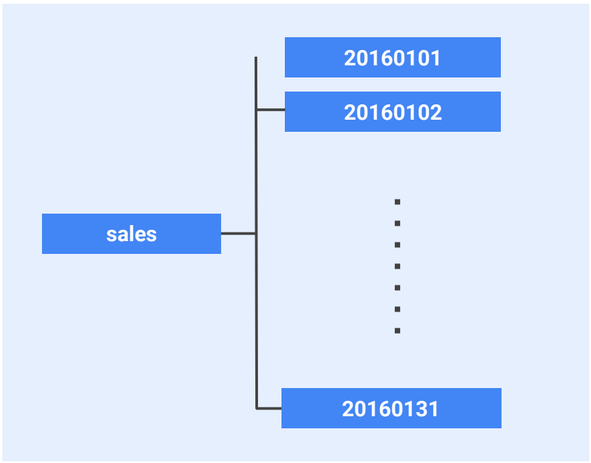
Row	contributor_id	LogEdits
1	2221364	4
2	104574	4
3	73576	4
4	311307	4
5	291919	4
6	140178	4
7	181636	4
8	3661553	4
9	3600820	4
10	4737290	4
11	938404	4
12	295955	4
13	183812	4
14	1811786	4
15	8918196	4
16	561624	4
17	5338406	4

← Do not Group on an ID

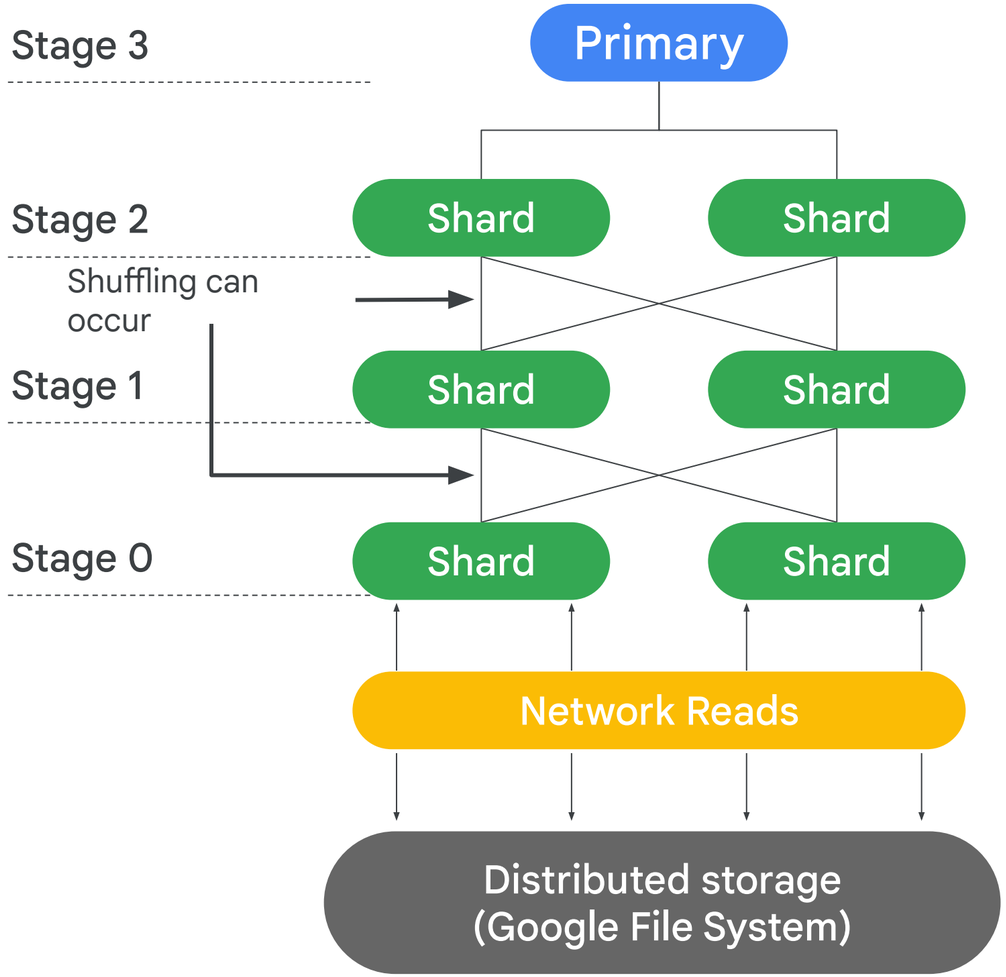


BigQuery

- Use table partitioning and clustering



Large GROUP BY means **many forced shuffles**




```
STAGE 3
READ  $10, $90 FROM __stage02_output
SORT  $90 DESC
WRITE  $100, $101 TO __stage03_output
```

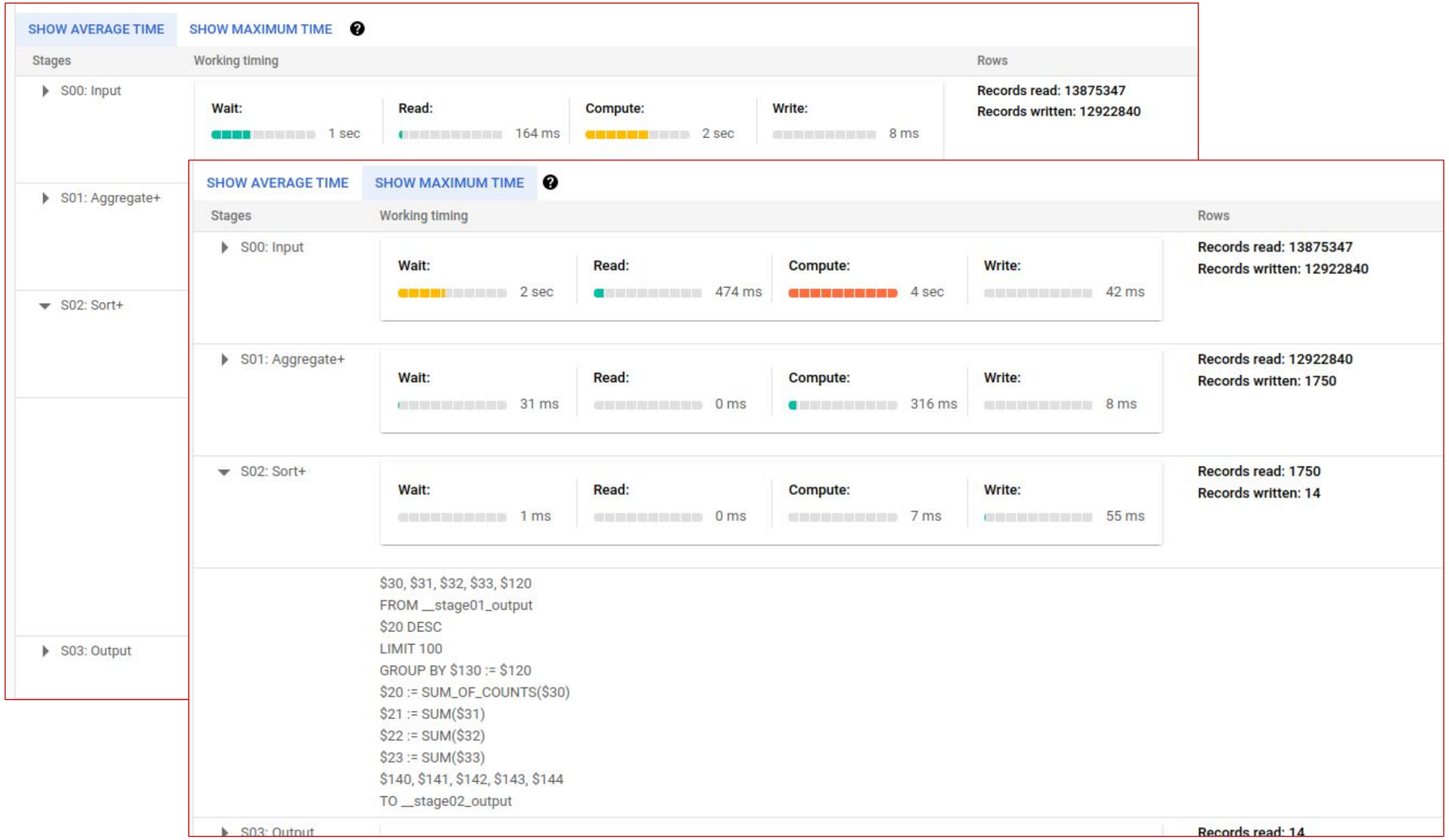
```
STAGE 2
READ  $20, $80 FROM __stage01_output
AGGREGATE  GROUP BY $90 := $80 $10 := SUM_OF_COUNTS($20)
WRITE  $10, $90 TO __stage02_output
```

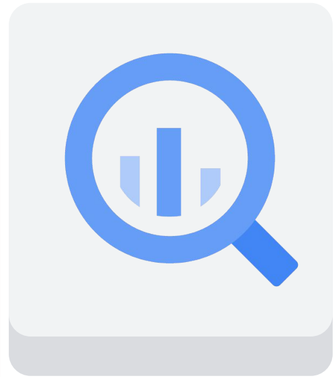
```
STAGE 1
READ  $60, $50 FROM __stage00_output
AGGREGATE  GROUP BY $80 := $30 $20 := COUNT($70)
COMPUTE   $30 := CAST(log10(CAST($40 AS DOUBLE)) AS INT64)
AGGREGATE  GROUP BY $70 := $60 $40 := SUM_OF_COUNTS($50)
WRITE  $20, $80 TO __stage01_output BY HASH($80)
```

```
STAGE 0
READ  $1:contributor_id
FROM publicdata.samples.wikipedia
AGGREGATE  GROUP BY $60 := $1 $50 := COUNT_STAR()
WRITE  $60, $50 TO __stage00_output BY HASH($60)
```

BigQuery best practices (7)

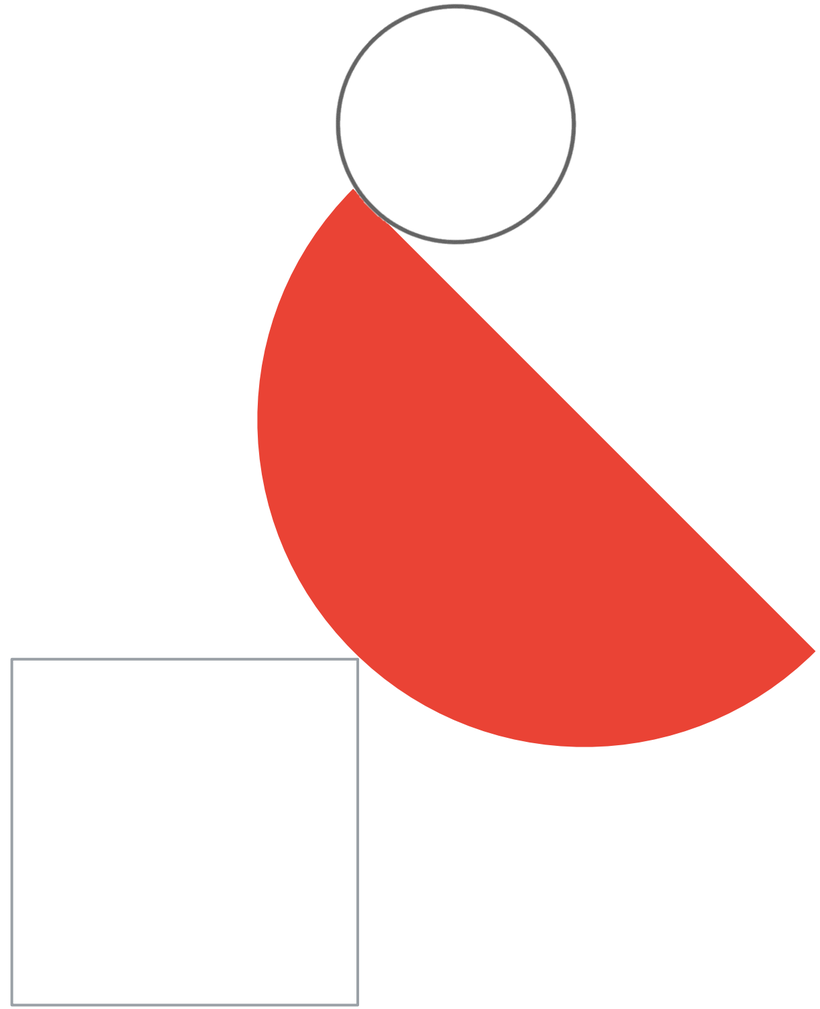
- Use query explanation map to optimize queries





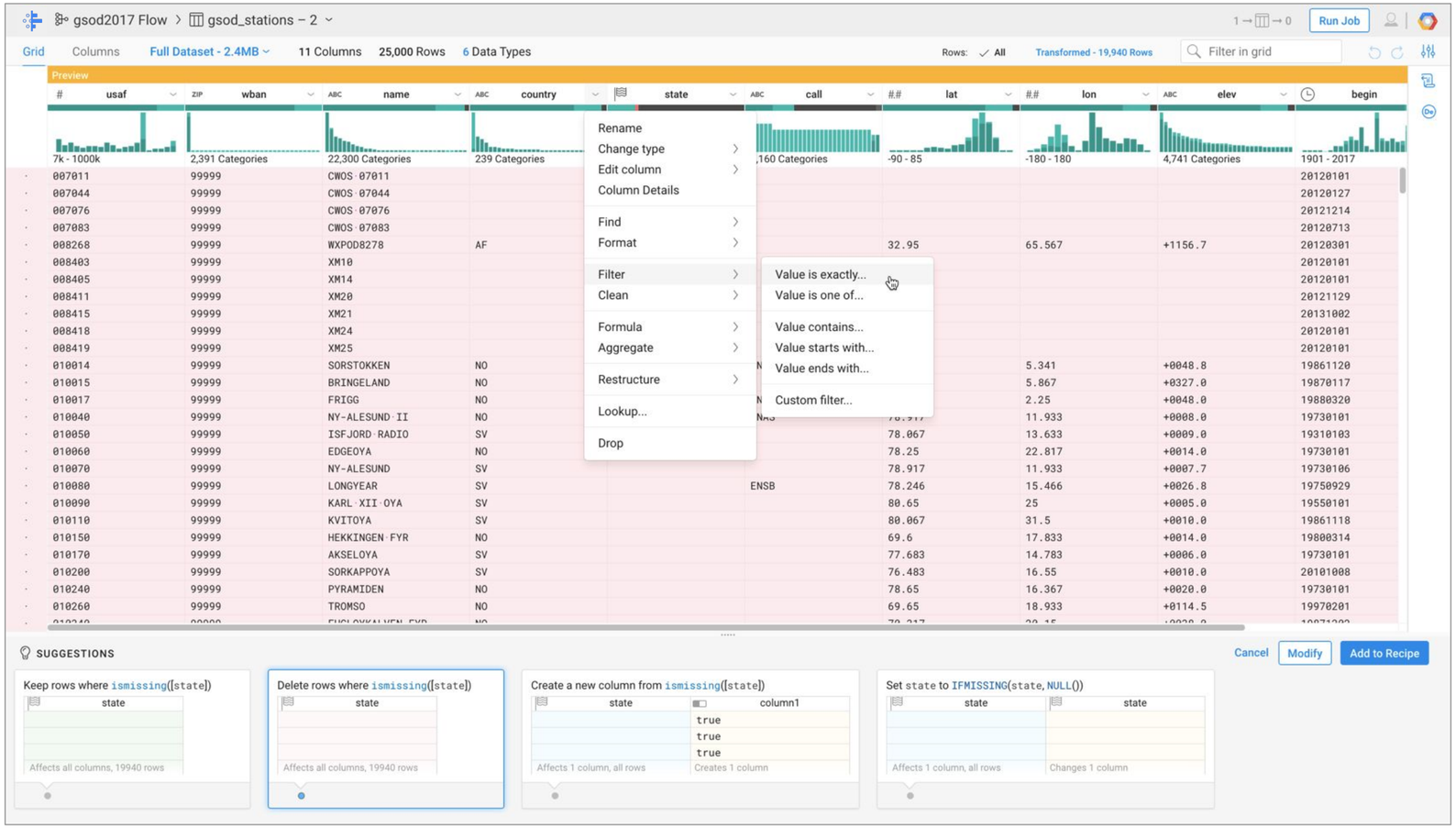
BigQuery

QUIZ time!



Dataprep

Dataprep by Trifacta for data wrangling

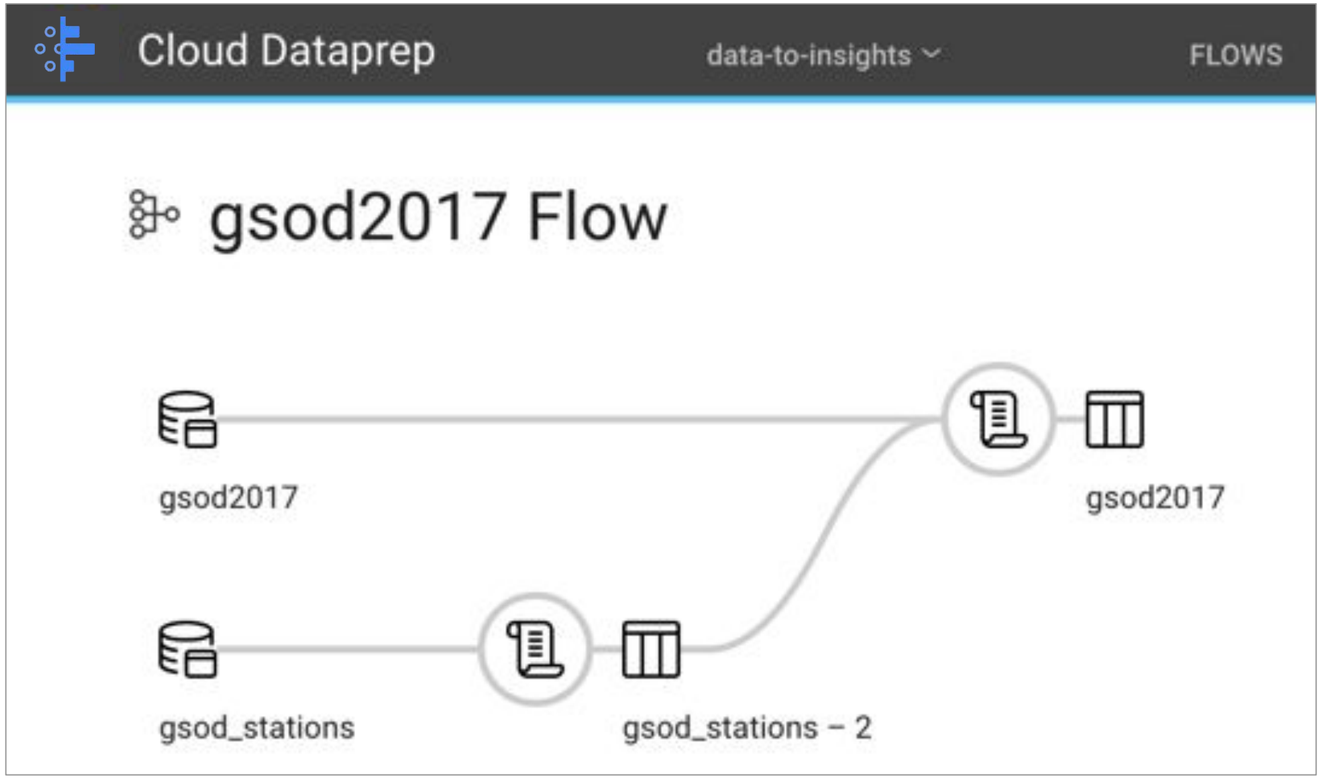


Using Dataprep for preprocessing

- Use the Dataprep GUI to create and preview data preparation steps
- Chain together multiple wranglers into a repeatable recipe
- Common tasks like record deduplication and derived fields
- Repeatable set of transformation steps build by chaining data wranglers together
- Jobs run against recipes
- Can include end-to-end steps from ingestion, transformation, aggregation, save to BigQuery
- Track completed and ongoing jobs
- See the data quality metrics for transformed datasets
- View histograms with summary statistics for each field

Dataprep

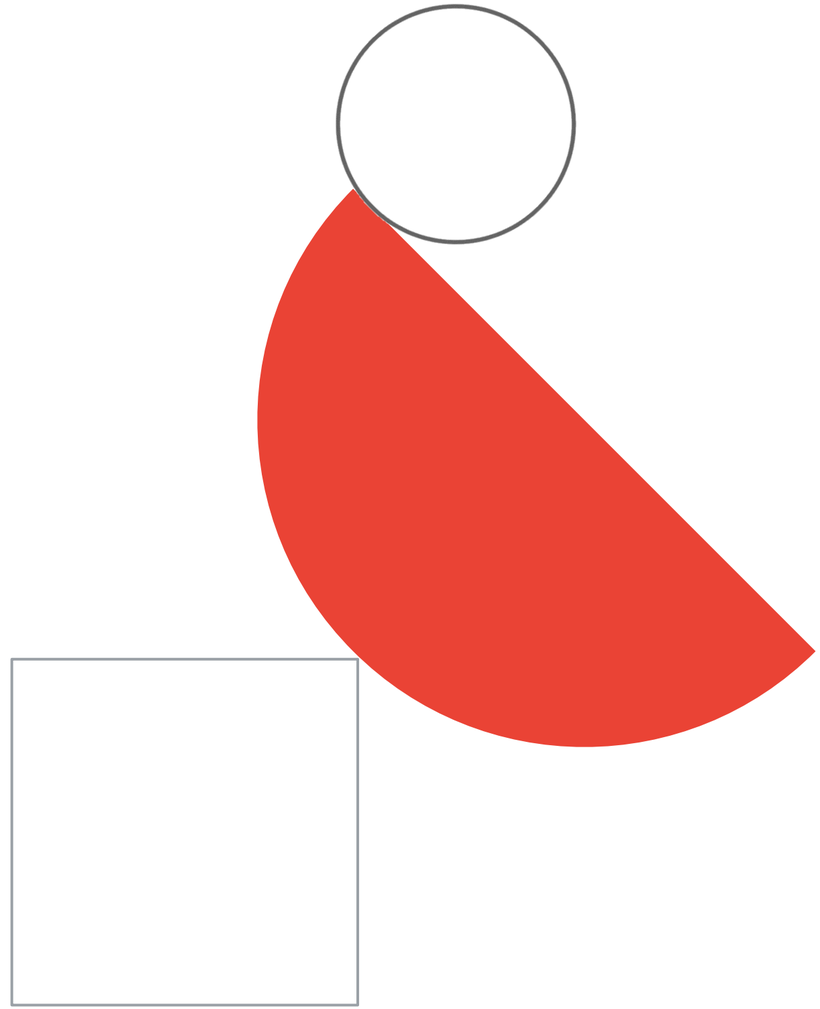
Run the Flow which includes our recipes and outputs a table



Dataplex

- Dataplex also acts as the central data governance component for modern data platforms.
- An intelligent data fabric like Google Cloud Dataplex provides the ability to discover, harvest metadata for disparate data sources and run data curation jobs using popular framework like Apache Spark.
- Dataplex makes it easy to accomplish tasks such as, data quality check, managed data security policies etc
- The data extracted from different sources are kept in designated raw zones, before subsequent transformations are applied on the raw data

[optional] Links to useful materials



OPTIONAL study materials:

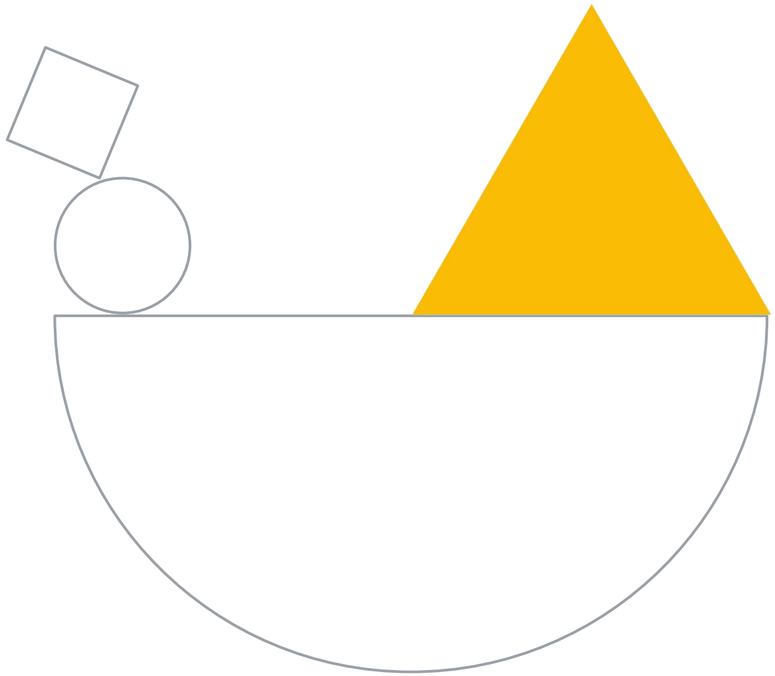
[READING]

- Drive innovation with Google Cloud smart analytics solutions
- Overview of BigQuery analytics
- Discover and protect your sensitive data
- Dataplex overview
- Cloud Key Management Service

[VIDEOS]

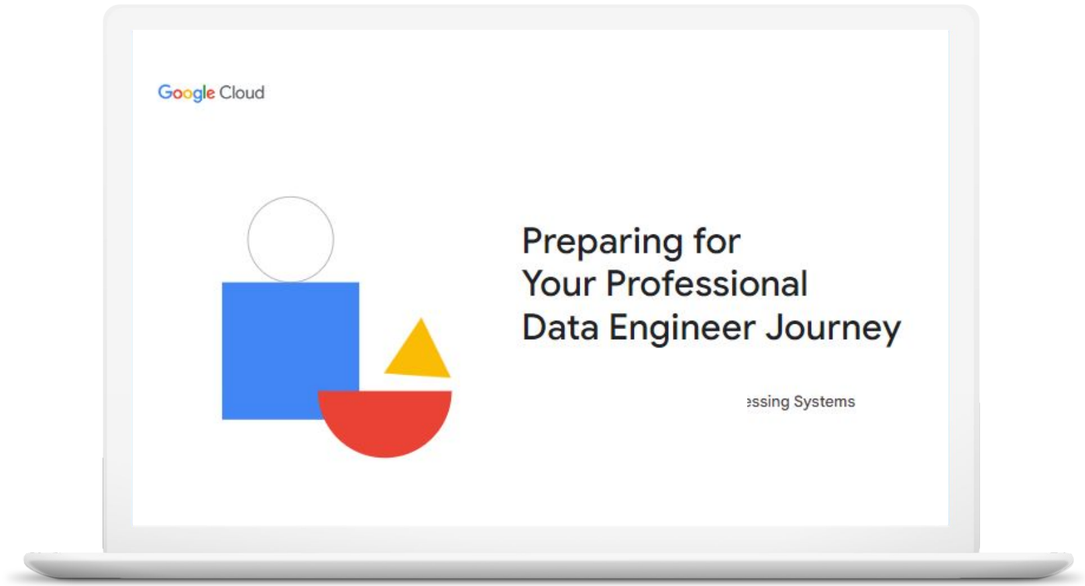
- Big Data Analytics with Google Cloud Platform
- Complete GCP END TO END Data Analysis Project for Beginners
- Data Loss Prevention (DLP)
- De-identification and inspection templates with Cloud DLP
- Build a Data Mesh on GCP with Dataplex
- How to optimize prices with DataPrep, BigQuery, and Looker
- How do I protect data in GCP using my own externally stored encryption keys?
- Cloud Key Management

Diagnostic questions



Your study plan:

Designing data processing systems



1.1 Designing for security and compliance

1.2 Designing for reliability and fidelity

1.3 Designing for flexibility and portability

1.4 Designing data migrations

1.1 Designing for security and compliance

Considerations include:

- Identity and Access Management (e.g., Cloud IAM and organization policies)
- Data security (encryption and key management)
- Privacy (e.g., personally identifiable information, and Cloud Data Loss Prevention API)
- Regional considerations (data sovereignty) for data access and storage
- Legal and regulatory compliance

1.1 Diagnostic Question 01 Discussion

Business analysts in your team need to run analysis on data that was loaded into BigQuery. You need to follow recommended practices and grant permissions.

What role should you grant the business analysts?

A. bigquery.resourceViewer and bigquery.dataViewer

B. bigquery.user and bigquery.dataViewer

C. bigquery.dataOwner

D. storage.objectViewer and bigquery.user

1.1 Diagnostic Question 01 Discussion

Business analysts in your team need to **run analysis** on data that was **loaded into BigQuery.** You need to follow recommended practices and grant permissions.

What role should you grant the business analysts?

A. bigquery.resourceViewer and bigquery.dataViewer
B. bigquery.user and bigquery.dataViewer
C. bigquery.dataOwner
D. storage.objectViewer and bigquery.user

1.1 Diagnostic Question 01 Discussion

Business analysts in your team need to r**un analysis** on data that was l**oaded into BigQuery.** You need to follow recommended practices and grant permissions.

What role should you grant the business analysts?

A. bigquery.resourceViewer and bigquery.dataViewer

B. bigquery.user and bigquery.dataViewer

C. bigquery.dataOwner

D. storage.objectViewer and bigquery.user

1.1 Diagnostic Question 02 Discussion

Cymbal Retail has acquired another company in Europe. Data access permissions and policies in this new region differ from those in Cymbal Retail's headquarters, which is in North America. You need to define a consistent set of policies for projects in each region that follow recommended practices.

What should you do?

A. Create a new organization for all projects in Europe and assign policies in each organization that comply with regional laws.

B. Implement a flat hierarchy, and assign policies to each project according to its region.

C. Create top level folders for each region, and assign policies at the folder level.

D. Implement policies at the resource level that comply with regional laws.

1.1 Diagnostic Question 02 Discussion

Cymbal Retail has acquired another company in Europe. Data access permissions and policies in this new region differ from those in Cymbal Retail's headquarters, which is in North America. You need to define a **consistent set of policies for projects in each region** that follow recommended practices.

What should you do?

A. Create a new organization for all projects in Europe and assign policies in each organization that comply with regional laws.

B. Implement a flat hierarchy, and assign policies to each project according to its region.

C. Create top level folders for each region, and assign policies at the folder level.

D. Implement policies at the resource level that comply with regional laws.

1.1 Diagnostic Question 02 Discussion

Cymbal Retail has acquired another company in Europe. Data access permissions and policies in this new region differ from those in Cymbal Retail's headquarters, which is in North America. You need to define a **consistent set of policies for projects in each region** that follow recommended practices.

What should you do?

A. Create a new organization for all projects in Europe and assign policies in each organization that comply with regional laws.

B. Implement a flat hierarchy, and assign policies to each project according to its region.

C. Create top level folders for each region, and assign policies at the folder level.

D. Implement policies at the resource level that comply with regional laws.

1.1 Diagnostic Question 03 Discussion

You are migrating on-premises data to a data warehouse on Google Cloud. This data will be made available to business analysts. Local regulations require that customer information including credit card numbers, phone numbers, and email IDs be captured, but not used in analysis. You need to use a reliable, recommended solution to redact the sensitive data.

What should you do?

A. Use the Cloud Data Loss Prevention (DLP) API to identify and redact data that matches infoTypes like credit card numbers, phone numbers, and email IDs.

B. Delete all columns with a title similar to "credit card," "phone," and "email."

C. Create a regular expression to identify and delete patterns that resemble credit card numbers, phone numbers, and email IDs.

D. Use the Cloud Data Loss Prevention (DLP) API to perform date shifting of any entries with credit card numbers, phone numbers, and email IDs.

1.1 Diagnostic Question 03 Discussion

You are migrating on-premises data to a data warehouse on Google Cloud. This data will be made available to business analysts. Local regulations require that customer information including credit card numbers, phone numbers, and email IDs be captured, but not used in analysis. You need to use a **reliable, recommended solution to redact the sensitive data**.

What should you do?

A. Use the Cloud Data Loss Prevention (DLP) API to identify and redact data that matches infoTypes like credit card numbers, phone numbers, and email IDs.

B. Delete all columns with a title similar to "credit card," "phone," and "email."

C. Create a regular expression to identify and delete patterns that resemble credit card numbers, phone numbers, and email IDs.

D. Use the Cloud Data Loss Prevention (DLP) API to perform date shifting of any entries with credit card numbers, phone numbers, and email IDs.

1.1 Diagnostic Question 03 Discussion

You are migrating on-premises data to a data warehouse on Google Cloud. This data will be made available to business analysts. Local regulations require that customer information including credit card numbers, phone numbers, and email IDs be captured, but not used in analysis. You need to use a **reliable, recommended solution to redact the sensitive data**.

What should you do?

A. Use the Cloud Data Loss Prevention (DLP) API to identify and redact data that matches infoTypes like credit card numbers, phone numbers, and email IDs.

B. Delete all columns with a title similar to "credit card," "phone," and "email."

C. Create a regular expression to identify and delete patterns that resemble credit card numbers, phone numbers, and email IDs.

D. Use the Cloud Data Loss Prevention (DLP) API to perform date shifting of any entries with credit card numbers, phone numbers, and email IDs.

1.1 Diagnostic Question 04 Discussion

Your data and applications reside in multiple geographies on Google Cloud. Some regional laws require you to hold your own keys outside of the cloud provider environment, whereas other laws are less restrictive and allow storing keys with the same provider who stores the data. The management of these keys has increased in complexity, and you need a solution that can centrally manage all your keys.

What should you do?

A. Enable confidential computing for all your virtual machines.

B. Store keys in Cloud Key Management Service (KMS), and reduce the number of days for automatic key rotation.

C. Store your keys in Cloud Hardware Security Module (HSM), and retrieve keys from it when required.

D. Store your keys on a supported external key management partner, and use Cloud

External Key Manager (EKM) to get keys when required.

1.1 Diagnostic Question 04 Discussion

Your data and applications reside in multiple geographies on Google Cloud. Some regional laws require you to hold your own **keys outside of the cloud provider environment**, whereas other laws are less restrictive and allow **storing keys with the same provider who stores the data**. The management of these keys has increased in complexity, and you need a solution that can **centrally manage all your keys**.

What should you do?

A. Enable confidential computing for all your virtual machines.

B. Store keys in Cloud Key Management Service (KMS), and reduce the number of days for automatic key rotation.

C. Store your keys in Cloud Hardware Security Module (HSM), and retrieve keys from it when required.

D. Store your keys on a supported external key management partner, and use Cloud

External Key Manager (EKM) to get keys when required.

1.1 Diagnostic Question 04 Discussion

Your data and applications reside in multiple geographies on Google Cloud. Some regional laws require you to hold your own **keys outside of the cloud provider environment**, whereas other laws are less restrictive and allow **storing keys with the same provider who stores the data**. The management of these keys has increased in complexity, and you need a solution that can **centrally manage all your keys**.

What should you do?

A. Enable confidential computing for all your virtual machines.

B. Store keys in Cloud Key Management Service (KMS), and reduce the number of days for automatic key rotation.

C. Store your keys in Cloud Hardware Security Module (HSM), and retrieve keys from it when required.

D. Store your keys on a supported external key management partner, and use Cloud

External Key Manager (EKM) to get keys when required.

Link

1.1 Designing for security and compliance

Courses

Modernizing Data Lakes and Data Warehouses with Google Cloud
- Introduction to Data Engineering
- Building a Data Lake
- Building a Data Warehouse

Smart Analytics, Machine Learning, and AI on Google Cloud
- Prebuilt ML Model APIs for Unstructured Data

Serverless Data Processing with Dataflow: Foundations
- IAM, Quotas, and Permissions
- Security

Skill Badges

Create and Manage Cloud Resources

Perform Foundational Data, ML, and AI Tasks in Google Cloud

Documentation

Secure a BigQuery data warehouse that stores confidential data
IAM basic and predefined roles reference
Creating and managing Folders
Resource hierarchy
Cloud Data Loss Prevention
InfoType detector reference
Cloud External Key Manager
Hold your own key with Google Cloud External Key Manager
Evolving Cloud External Key Manager – What's new with Cloud EKM | Google Cloud Blog

1.2 Designing for reliability and fidelity

Considerations include:

- Preparing and cleaning data (e.g., Dataprep, Dataflow, and Cloud Data Fusion)
- Monitoring and orchestration of data pipelines
- Disaster recovery and fault tolerance
- Making decisions related to ACID (atomicity, consistency, isolation, and durability) compliance and availability
- Data validation

12 Diagnostic Question 05 Discussion

Cymbal Retail has a team of business analysts who need to fix and enhance a set of large input data files. For example, duplicates need to be removed, erroneous rows should be deleted, and missing data should be added. These steps need to be performed on all the present set of files and any files received in the future in a repeatable, automated process. The business analysts are not adept at programming.

What should they do?

A. Load the data into Dataprep, explore the data, and edit the transformations as needed.

B. Create a Dataproc job to perform the data fixes you need.

C. Create a Dataflow pipeline with the data fixes you need.

D. Load the data into Google Sheets, explore the data, and fix the data as needed.

12 Diagnostic Question 05 Discussion

Cymbal Retail has a team of business analysts who need to fix and enhance a set of large input data files. For example, **duplicates need to be removed, erroneous rows should be deleted, and missing data should be added.** These steps need to be **performed on all the present set of files and any files received in the future** in a repeatable, automated process. The business analysts are **not adept at programming**.

What should they do?

A. Load the data into Dataprep, explore the data, and edit the transformations as needed.

B. Create a Dataproc job to perform the data fixes you need.

C. Create a Dataflow pipeline with the data fixes you need.

D. Load the data into Google Sheets, explore the data, and fix the data as needed.

12 Diagnostic Question 05 Discussion

Cymbal Retail has a team of business analysts who need to fix and enhance a set of large input data files. For example, **duplicates need to be removed, erroneous rows should be deleted, and missing data should be added.** These steps need to be **performed on all the present set of files and any files received in the future** in a repeatable, automated process. The business analysts are **not adept at programming**.

What should they do?

A. Load the data into Dataprep, explore the data, and edit the transformations as needed.

B. Create a Dataproc job to perform the data fixes you need.

C. Create a Dataflow pipeline with the data fixes you need.

D. Load the data into Google Sheets, explore the data, and fix the data as needed.

12 Diagnostic Question 06 Discussion

You have a Dataflow pipeline that runs data processing jobs. You need to identify the parts of the pipeline code that consume the most resources.

What should you do?

A. Use Cloud Monitoring

B. Use Cloud Logging

C. Use Cloud Profiler

D. Use Cloud Audit Logs

12 | Diagnostic Question 06 Discussion

You have a Dataflow pipeline that runs data processing jobs. You need to identify the parts of the pipeline code that **consume the most resources**.

What should you do?

A. Use Cloud Monitoring

B. Use Cloud Logging

C. Use Cloud Profiler

D. Use Cloud Audit Logs

12 | Diagnostic Question 06 Discussion

You have a Dataflow pipeline that runs data processing jobs. You need to identify the parts of the pipeline code that **consume the most resources**.

What should you do?

A. Use Cloud Monitoring

B. Use Cloud Logging

C. Use Cloud Profiler

D. Use Cloud Audit Logs

Link

1.2 Designing for reliability and fidelity

Courses

[Modernizing Data Lakes and Data Warehouses on Google Cloud]
- Building a Data Warehouse

[Building Batch Data Pipelines on Google Cloud]
- Introduction to Building Batch Data Pipelines
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

[Building Resilient Streaming Analytics Systems on Google Cloud]
- Serverless Messaging with Pub/Sub

[Serverless Data Processing with Dataflow: Develop Pipelines]
- Best Practices

[Serverless Data Processing with Dataflow: Operations]
- Monitoring
- Logging and Error Reporting
- Troubleshooting and Debug
- Testing and CI/CD
- Reliability

Skill Badges

[Perform Foundational Data, ML, and AI Tasks in Google Cloud]

[Engineer Data with Google Cloud]

Documentation

[Clean and Enhance Your Data]

[Introduction to Data Wrangling]

[Monitoring pipeline performance using Cloud Profiler | Cloud Dataflow]

1.3 Designing for flexibility and portability

Considerations include:

- Mapping current and future business requirements to the architecture
- Designing for data and application portability (e.g., multi-cloud and data residency requirements)
- Data staging, cataloging, and discovery (data governance)

1.3 Diagnostic Question 07 Discussion

You are using Dataproc to process a large number of CSV files. The storage option you choose needs to be flexible to serve many worker nodes in multiple clusters. These worker nodes will read the data and also write to it for intermediate storage between processing jobs.

What is the recommended storage option on Google Cloud?

A. Cloud SQL

B. Zonal persistent disks

C. Local SSD

D. Cloud Storage

1.3 Diagnostic Question 07 Discussion

You are using **Dataproc** to process a **large number of CSV files**. The storage option you choose needs to be **flexible to serve many worker nodes in multiple clusters**. These worker nodes will read the data and also write to it for **intermediate storage** between processing jobs.

What is the recommended storage option on Google Cloud?

A. Cloud SQL

B. Zonal persistent disks

C. Local SSD

D. Cloud Storage

1.3 Diagnostic Question 07 Discussion

You are using **Dataproc** to process a **large number of CSV files**. The storage option you choose needs to be **flexible to serve many worker nodes in multiple clusters**. These worker nodes will read the data and also write to it for **intermediate storage** between processing jobs.

What is the recommended storage option on Google Cloud?

A. Cloud SQL

B. Zonal persistent disks

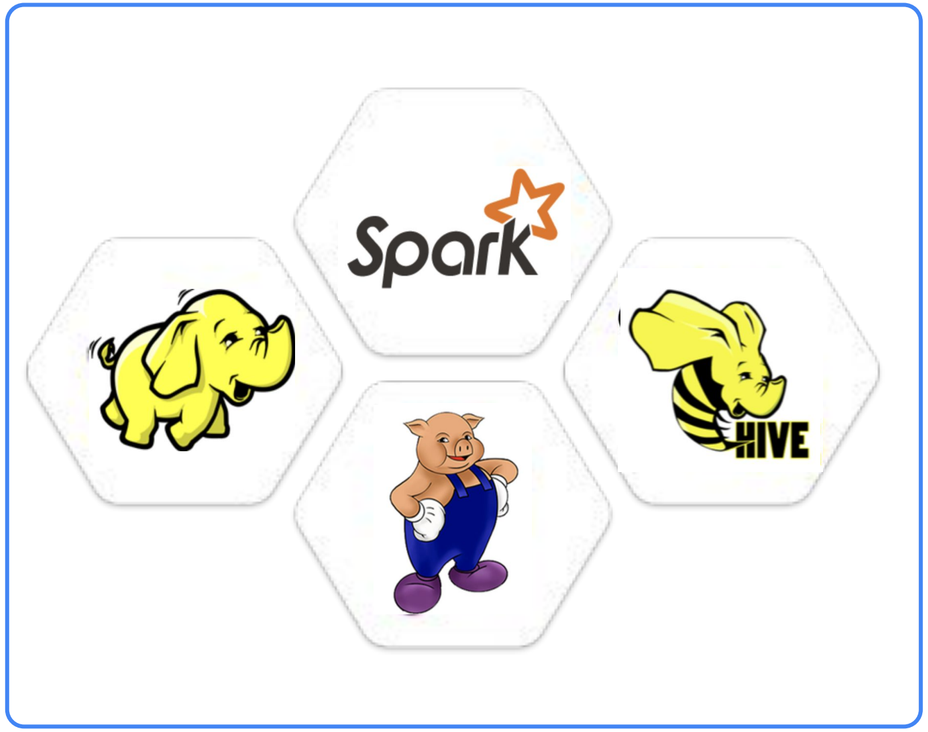
C. Local SSD

D. Cloud Storage

HDFS to GCS: a common theme for Dataproc

Use Google Cloud Storage as your primary data source and sink

- Decouple storage and compute
- Un-silo your data
- Pros:
  - HDFS compatibility, lower costs, no storage management overhead, separate storage from compute, integration with other GCP services
- Cons:
  - Increased I/O variance and request latency (esp. Important with small files), GCS is immutable (no append / truncate)



Exam Tip: *Dataproc (and GCS usage instead of HDFS) is a common choice for customers migrating from existing, on-premises Hadoop ecosystem.*

1.3 Diagnostic Question 08 Discussion

You are managing the data for Cymbal Retail, which consists of multiple teams including retail, sales, marketing, and legal. These teams are consuming data from multiple producers including point of sales systems, industry data, orders, and more. Currently, teams that consume data have to repeatedly ask the teams that produce it to verify the most up-to-date data and to clarify other questions about the data, such as source and ownership. This process is unreliable and time-consuming and often leads to repeated escalations. You need to implement a centralized solution that gains a unified view of the organization's data and improves searchability.

What should you do?

A. Implement a data mesh with Dataplex and have producers tag data when created.

B. Implement a data lake with Cloud Storage, and create buckets for each team such as retail, sales, marketing.

C. Implement a data warehouse by using BigQuery, and create datasets for each team such as retail, sales, marketing.

D. Implement Looker dashboards that provide views of the data that meet each teams' requirements.

1.3 Diagnostic Question 08 Discussion

You are managing the data for Cymbal Retail, which consists of multiple teams including retail, sales, marketing, and legal. These teams are **consuming data from multiple producers** including point of sales systems, industry data, orders, and more. Currently, teams that consume data have to repeatedly ask the teams that produce it to verify the most up-to-date data and to clarify other questions about the data, such as source and ownership. This process is unreliable and time-consuming and often leads to repeated escalations. You need to implement a **centralized solution that gains a unified view of the organization's data** and improves searchability.

What should you do?

A. Implement a data mesh with Dataplex and have producers tag data when created.

B. Implement a data lake with Cloud Storage, and create buckets for each team such as retail, sales, marketing.

C. Implement a data warehouse by using BigQuery, and create datasets for each team such as retail, sales, marketing.

D. Implement Looker dashboards that provide views of the data that meet each teams' requirements.

1.3 Diagnostic Question 08 Discussion

You are managing the data for Cymbal Retail, which consists of multiple teams including retail, sales, marketing, and legal. These teams are **consuming data from multiple producers** including point of sales systems, industry data, orders, and more. Currently, teams that consume data have to repeatedly ask the teams that produce it to verify the most up-to-date data and to clarify other questions about the data, such as source and ownership. This process is unreliable and time-consuming and often leads to repeated escalations. You need to implement a **centralized solution that gains a unified view of the organization's data** and improves searchability.

What should you do?

A. Implement a data mesh with Dataplex and have producers tag data when created.

B. Implement a data lake with Cloud Storage, and create buckets for each team such as retail, sales, marketing.

C. Implement a data warehouse by using BigQuery, and create datasets for each team such as retail, sales, marketing.

D. Implement Looker dashboards that provide views of the data that meet each teams' requirements.

1.3 Designing for flexibility and portability

Courses

[Modernizing Data Lakes and Data Warehouses on Google Cloud](#)
- Introduction to Data Engineering
- Building a Data Lake

[Building Batch Data Pipelines on Google Cloud](#)
- Introduction to Building Batch Data Pipelines

[Serverless Data Processing with Dataflow: Foundations](#)
- Beam Portability

Skill Badges

[Get Started with Dataplex](#)

Documentation

[Dataproc best practices | Google Cloud Blog](#)
[HDFS vs. Cloud Storage: Pros, cons and migration tips | Google Cloud Blog](#)
[Dataplex overview](#)

1.4 Designing data migrations

Considerations include:

- Analyzing current stakeholder needs, users, processes, and technologies and creating a plan to get to desired state
- Planning migration to Google Cloud (e.g., BigQuery Data Transfer Service, Database Migration Service, Transfer Appliance, Google Cloud networking, Datastream)
- Designing the migration validation strategy
- Designing the project, dataset, and table architecture to ensure proper data governance

14 Diagnostic Question 09 Discussion

Laws in the region where you operate require that files related to all orders made each day are stored immutably for 365 days. The solution that you recommend has to be cost-effective.

What should you do?

A. Store the data in a Cloud Storage bucket, and enable object versioning and delete any version older than 365 days.

B. Store the data in a Cloud Storage bucket, and specify a retention period.

C. Store the data in a Cloud Storage bucket, and set a lifecycle policy to delete the file after 365 days.

D. Store the data in a Cloud Storage bucket, enable object versioning, and delete any version greater than 365.

14 Diagnostic Question 09 Discussion

Laws in the region where you operate require that files related to all orders made each day are **stored immutably for 365 days**. The solution that you recommend has to be **cost-effective.**

What should you do?

A. Store the data in a Cloud Storage bucket, and enable object versioning and delete any version older than 365 days.

B. Store the data in a Cloud Storage bucket, and specify a retention period.

C. Store the data in a Cloud Storage bucket, and set a lifecycle policy to delete the file after 365 days.

D. Store the data in a Cloud Storage bucket, enable object versioning, and delete any version greater than 365.

14 Diagnostic Question 09 Discussion

Laws in the region where you operate require that files related to all orders made each day are **stored immutably for 365 days**. The solution that you recommend has to be **cost-effective.**

What should you do?

A. Store the data in a Cloud Storage bucket, and enable object versioning and delete any version older than 365 days.

B. Store the data in a Cloud Storage bucket, and specify a retention period.

C. Store the data in a Cloud Storage bucket, and set a lifecycle policy to delete the file after 365 days.

D. Store the data in a Cloud Storage bucket, enable object versioning, and delete any version greater than 365.

Know these GCS features well!

Object versioning and retention policies cannot be on at the same time. If you want to allow objects to be modified, choose object versioning. If you want to prevent deletions or changes to objects, set a retention policy.

- Controlling object lifecycle:
  - Retention policy (best for compliance)
  - Object Hold (prevent individual objects from being deleted)
  - Object Versioning (aka “automatic backups with retention policy; be aware of additional costs)
  - Object Lifecycle Management (aka “object TTL” / downgrade class to optimize costs)
- ACLs (read, write, full control on buckets or an object).
- Objects are immutable.
- Location constraints on buckets.
- Object change notifications (useful for automation, along with Pub/Sub).
- Resumable uploads (can restart from the last successful chunk).
- Strong consistency except for cached objects.
- Storage class set at object level (fine-grained performance/cost control without moving data to different buckets).
- Cloud Storage Triggers to handle events in Cloud Functions.
- Streaming uploads to GCS.
- For data encryption, GCS supports GMEK, CMEK and CSEK (most services do not support CSEK!)

14 Diagnostic Question 10 Discussion

Cymbal Retail is migrating its private data centers to Google Cloud. Over many years, hundreds of terabytes of data were accumulated. You currently have a 100 Mbps line and you need to transfer this data reliably before commencing operations on Google Cloud in 45 days.

What should you do?

A. Store the data in an HTTPS endpoint, and configure Storage Transfer Service to copy the data to Cloud Storage.

B. Upload the data to Cloud Storage by using gsutil.

C. Zip and upload the data to Cloud Storage buckets by using the Google Cloud console.

D. Order a transfer appliance, export the data to it, and ship it to Google.

14 Diagnostic Question 10 Discussion

Cymbal Retail is migrating its private data centers to Google Cloud. Over many years, hundreds of terabytes of data were accumulated. You currently have a **100 Mbps line and you need to transfer this data reliably before commencing operations on Google Cloud in 45 days**.

What should you do?

A. Store the data in an HTTPS endpoint, and configure Storage Transfer Service to copy the data to Cloud Storage.

B. Upload the data to Cloud Storage by using gsutil.

C. Zip and upload the data to Cloud Storage buckets by using the Google Cloud console.

D. Order a transfer appliance, export the data to it, and ship it to Google.

14 Diagnostic Question 10 Discussion

Cymbal Retail is migrating its private data centers to Google Cloud. Over many years, hundreds of terabytes of data were accumulated. You currently have a **100 Mbps line and you need to transfer this data reliably before commencing operations on Google Cloud in 45 days**.

What should you do?

A. Store the data in an HTTPS endpoint, and configure Storage Transfer Service to copy the data to Cloud Storage.

B. Upload the data to Cloud Storage by using gsutil.

C. Zip and upload the data to Cloud Storage buckets by using the Google Cloud console.

D. Order a transfer appliance, export the data to it, and ship it to Google.

Choose the right tool to move data to GCS...

Where you're moving data from	Scenario	Suggested products
Another cloud provider (for example, Amazon Web Services or Microsoft Azure) to Google Cloud	—	Storage Transfer Service
Cloud Storage to Cloud Storage (two different buckets)	—	Storage Transfer Service
Your private data center to Google Cloud	Enough bandwidth to meet your project deadline for less than 1 TB of data	gsutil
Your private data center to Google Cloud	Enough bandwidth to meet your project deadline for more than 1 TB of data	Storage Transfer Service for on-premises data
Your private data center to Google Cloud	Not enough bandwidth to meet your project deadline	Transfer Appliance

Exam Tips:

- *Depending on size and throughput, use gsutil / Transfer Service (low cost) / Transfer Appliance*
- ***When using Transfer Appliance, you need to execute so-called "rehydration" process*** *which will decrypt and uncompress before it's put to a destination bucket.*

1.4 Designing data migrations

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud
- Building a Data Lake
- Building a Data Warehouse

BigQuery Fundamentals for Redshift Professionals
- BigQuery and Google Cloud IAM

Documentation

Retention policies and retention policy locks | Cloud Storage

Migration to Google Cloud: Transferring your large datasets

Knowledge Check 1

Your company is very serious about data protection and hence decides to implement the Principle of Least Privilege. What should you do to comply with this policy?

A. Ensure that the access permissions are given strictly based on the person's title and job role.

B. Give just enough permissions to get the task done.

C. When a task is assigned, ensure that it gets assigned to a person with the minimum privileges.

D. Ensure that the users are verified every time they request access, even if they were authenticated earlier.

Knowledge Check 1

Your company is very serious about data protection and hence decides to implement the Principle of Least Privilege. What should you do to comply with this policy?

A. Ensure that the access permissions are given strictly based on the person's title and job role.

B. Give just enough permissions to get the task done.

C. When a task is assigned, ensure that it gets assigned to a person with the minimum privileges.

D. Ensure that the users are verified every time they request access, even if they were authenticated earlier.

Knowledge Check 2

A company collects lots of consumer data from online marketing campaigns. Company plans to use Google Cloud to store this collected data. The top management is worried about exposing personally identifiable information (PII) that may be present in this data. What should you do to reduce the risk of exposing PII data?

A. Ensure that all PII data is removed from the collected data before storing it on Google Cloud.

B. Store all data in BigQuery and turn on column level access to protect sensitive data.

C. Use Cloud Data Loss Prevention (Cloud DLP) to inspect and redact PII data.

D. Ensure that all stored data is monitored by Security Command Center.

Knowledge Check 2

A company collects lots of consumer data from online marketing campaigns. The company plans to use Google Cloud to store this collected data. The top management is worried about exposing personally identifiable information (PII) that may be present in this data. What should you do to reduce the risk of exposing PII data?

A. Ensure that all PII data is removed from the collected data before storing it on Google Cloud.

B. Store all data in BigQuery and turn on column level access to protect sensitive data.

C. Use Cloud Data Loss Prevention (Cloud DLP) to inspect and redact PII data.

D. Ensure that all stored data is monitored by Security Command Center.