O'REILLY®

Google Cloud Associate Data Practitioner

Joe Holbrook

Google Cloud Associate Data Practitioner

About the Course

In this course, you will learn:

- Learn about database types, data structures, data schemes, and other important aspects of data management hosted on Google Cloud.
- Learn about data warehouses and databases on Google Cloud. (BigQuery, BigTable, CloudSQL etc)
- Understand how ETL and ELT processes work on Google Cloud Services. (DataProc, DataFlow, Data Fusion, Cloud Composer)
- Learn how to choose appropriate storage services. (Cloud Storage, etc)
- Learn about choosing appropriate data extraction Tools(Dataflow, BigQuery Data Transfer Service)

Google Cloud Associate Data Practitioner

About the Course

In this course, you will learn:

- Identify use cases for event-driven data ingestion from Pub/Sub to BigQuery
- Mapping business requirements to use cases with Google Cloud.
- How to design and deploy Eventarc triggers in event-driven pipelines.
- Describe the importance of data governance, data-stewardship, and quality controls to ensure compliance and data consistency.
- Identify the compliance, privacy, and security requirements.
- Visualize data and create dashboards in Looker given business requirements.
- Define, train, evaluate, and use ML models.
- Prepare yourself efficiently and effectively for the Associate Data Practitioner Certification

Google Cloud Associate Data Practitioner

About the course

Who is this course for?

- Beginners looking for an entry point into the data world to pass the certification.
- Business Analysts, Data Analysts, and other data professionals with some experience working with data.
- Software and other IT professionals looking to boost their knowledge and skill sets in data management and ensuring data is used to provide value to the organization.

Course Overview

What we will cover in the course

- Associate Data Practitioner Certification
- Data Preparation and Ingestion
- Data Pipeline Orchestration
- Data Analysis and Presentation
- Data Management

What is the Google Cloud Associate Data Practioner Certification?

What the certification exam is about.

What is the Associate Data Practitioner?

About the Exam

A **Google Cloud Associate Data Practitioner** is a certification that validates an individual's ability to manage, secure, analyze, and visualize data on Google Cloud.

It demonstrates proficiency in using Google Cloud data services for tasks like data ingestion, transformation, pipeline management, analysis, machine learning, and visualization, signifying a foundational understanding of data handling within the Google Cloud ecosystem.

Google Recommends 6+ months of experience working with data on Google Cloud.

Associate Data Practitioner Exam Objectives

Understanding the subject area that can be tested.

Associate Data Practitioner Exam Objectives

About the certification

The Associate Data Practitioner exam assesses your ability to:

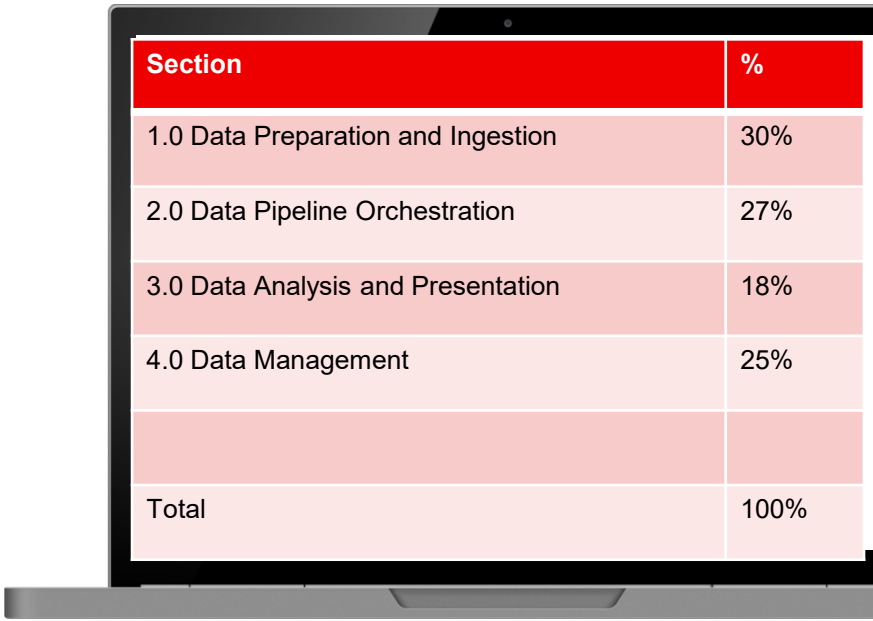
- Prepare and ingest data
- Analyze and present data
- Orchestrate data pipelines
- Manage data

Associate Data Practitioner Exam Objectives

About the certification exam

Associate Data Practitioner

- The exam is vendor-focused, meaning that Google Cloud services and practices are tested from a practical or exercise perspective.
- The exam tests identifying use cases of specific software, such as APIs, utilities, etc., that a data professional should know.



Section	%
1.0 Data Preparation and Ingestion	30%
2.0 Data Pipeline Orchestration	27%
3.0 Data Analysis and Presentation	18%
4.0 Data Management	25%
Total	100%

Associate Data Practitioner Exam Experience

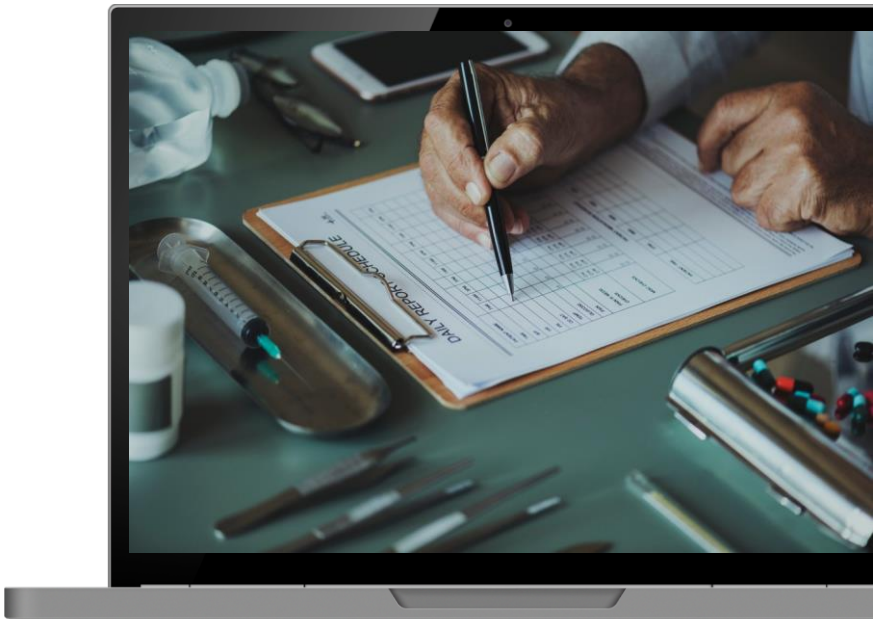
What to expect during the exam

Associate Data Practitioner Exam Experience

About the exam

Associate Data Practitioner Certification:

- Number of Questions: Between 50 - 60 questions (120 minutes)
- Passing score: Not Disclosed/Floating
- Testing Proctor is Webassessor

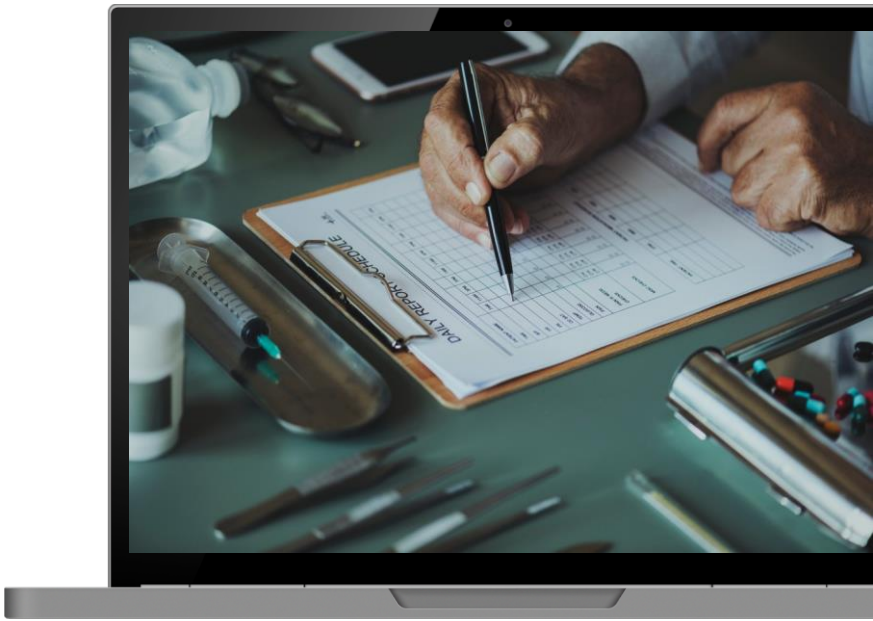


Associate Data Practitioner Exam Experience

About the exam

Identification Process

- Present ID (Federal/State)
- Primary ID: Drivers License, State ID, Military ID, Passport (Picture and Signature)
- Secondary: Credit Card/Debit Card with Signature.
- Photo is also taken of you at test center or online.

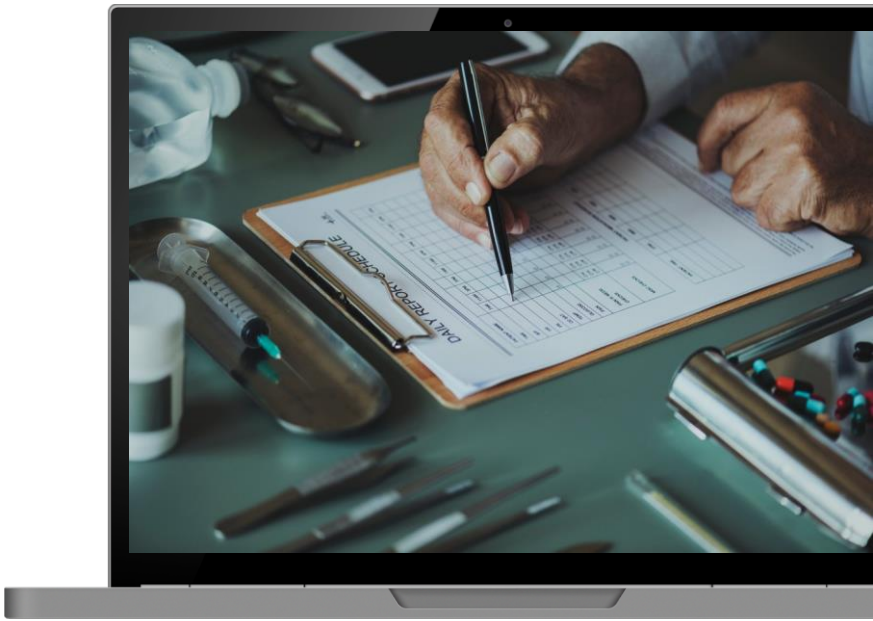


Associate Data Practitioner Exam Experience

About the exam

Exam Proctoring

- Proctor will check in and monitor your progress.
- Not an open book exam.
- The test center may provide you a dry erase board for solving problems.
- Exam will score (pass or fail)automatically once submitted or timed out.



Course Resources

Course materials provided to prepare for the certification.

Course and Exam Prep Resources

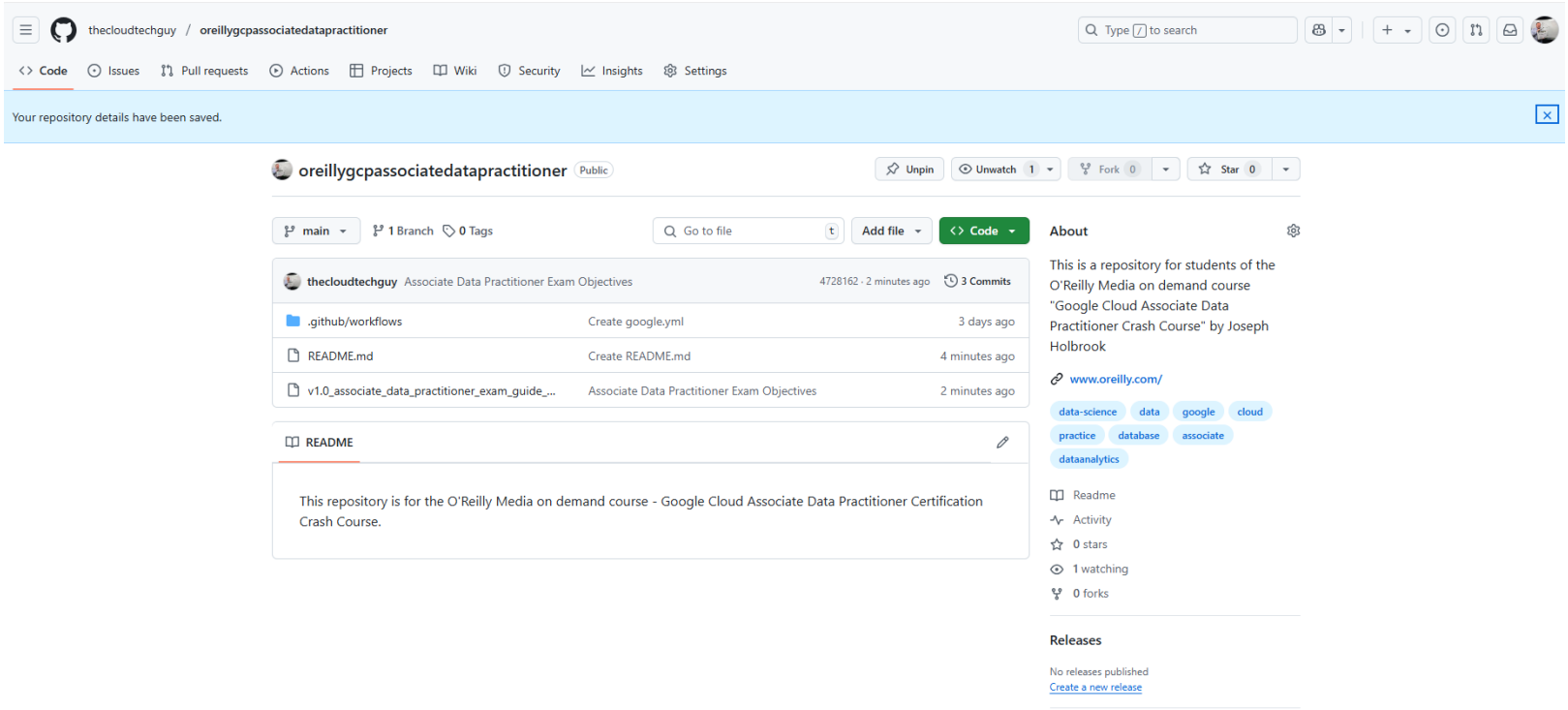
Resources available with the course

Course Resources

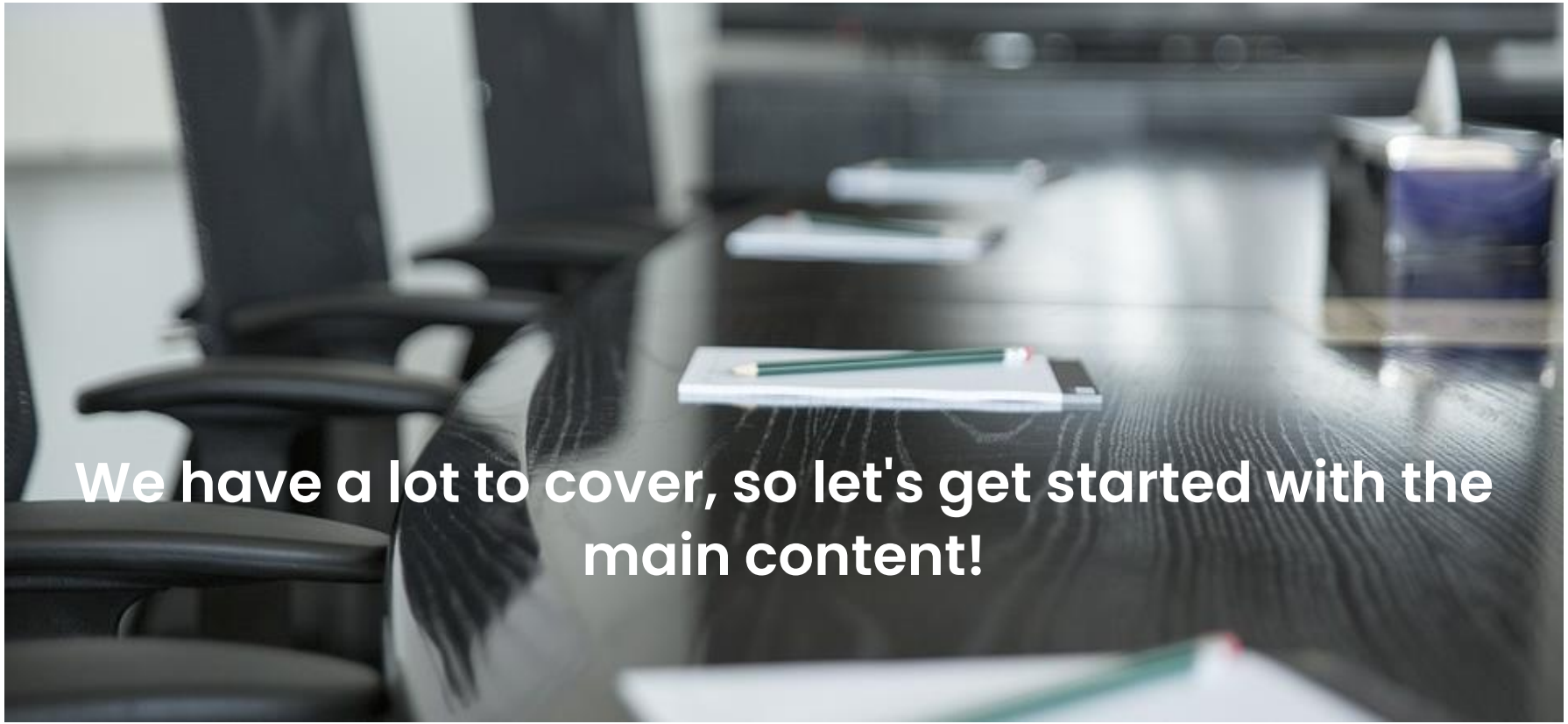
- Presentation Download
- Documentation Links
- Whiteboard Solutions
- Data Demonstrations
- Topic Discussions
- Module Quizzes
- Cumulative Exam

Course and Exam Prep Resources

Github Repository https://github.com/thecloudtechguy/oreillygcpassociatedatapractitioner



Course and Exam Prep Resources



Data Preparation and Ingestion

Diving into the data processing, preparing, mapping, formatting, and storage of data on GCP.

Module Overview

What we will cover in the Module

- Discussion – Importance of Data
- Data Quality Fundamentals
- Data Migration Transfer tools
- Data Cleaning and Profiling
- Demonstration: Data Transfer Formats
- Data Structures, Files and Types
- Demonstration - Load Data into Google BigQuery
- Understanding Storage Options
- Whiteboard - Mapping business requirements to Use Cases
- Module Review

The Importance of Data Practitioners

Discussion


Topic Discussion

Data Quality Fundamentals

Data Quality with GCP

Data Quality

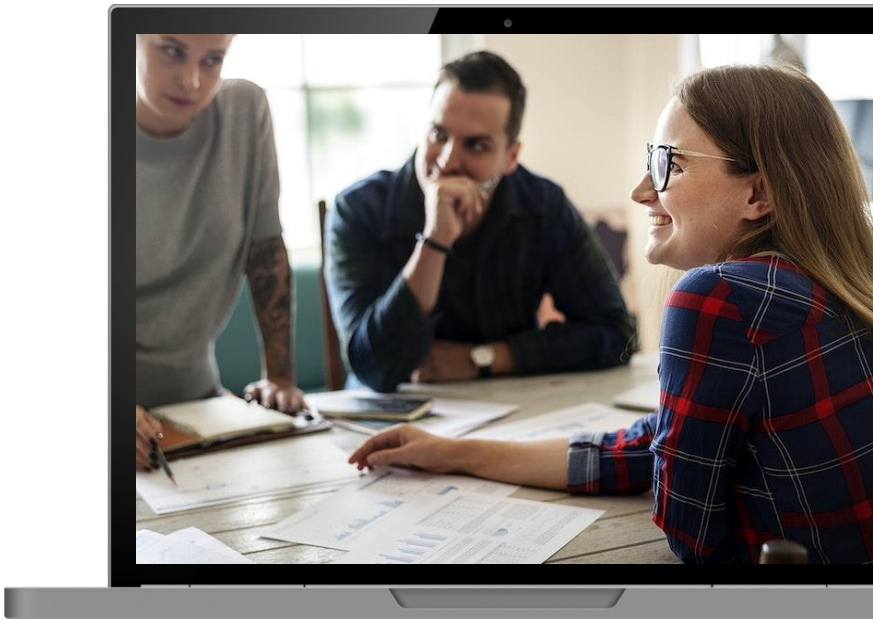
What is Data Quality?

- Data quality refers to data's overall "fitness" for its intended use.
- How reliable, accurate, and useful is your data?
- High-quality data is essential for making sound decisions, running effective operations, and gaining valuable insights.

Data Quality

Why Data Quality Matters?

- Better Decision Making
- Improved Efficiency
- Enhanced Customer Satisfaction
- Reduced Costs
- Regulatory Compliance



Data Quality

A quality dimension measures how pure the data is and a specific aspect of quality that can be evaluated and measured.
Key Dimensions of Data Quality

- **Accuracy:** This measures how correct the data is. Does it reflect reality?
- **Completeness:** This refers to whether all required data is present. Are there any missing values?
- **Consistency:** This ensures data consistency across different systems and databases.
- **Validity:** This checks if data conforms to defined rules and formats.

Data Quality

Key Dimensions of Data Quality(Continued)

- **Timeliness:** This measures how up-to-date the data is. Is it current and relevant?
- **Uniqueness:** This confirms that there are no duplicate records.
- **Fitness for purpose:** This assesses whether the data is useful for the intended task.

Data Quality Services

Comparing GCP Services

Table: Comparison of GCP Services for Data Quality

Service	Function	Data Quality Use
Dataflow	A fully managed, serverless service for stream and batch data processing. Allows you to implement data validation, cleansing, and transformation pipelines.	• Building pipelines to enforce data quality rules. • Identifying and handling data anomalies. • Performing data transformations to ensure consistency
Dataprep	A serverless, interactive data exploration and preparation service. Provides a visual interface for data cleaning and transformation.	• Profiling data to identify quality issues. • Interactive data cleansing and standardization. • Creating reusable data preparation recipes.
BigQuery	BigQuery itself has many built in functions that help with data quality. BigQuery also has data quality checks that can be set up.	• Using SQL queries to validate data and identify inconsistencies. • Implementing data quality checks as part of your data pipelines. • Data profiling and data validation

Data Quality Services

Comparing GCP Services

Table: Comparison of GCP Services for Data Quality

Service	Function	Data Quality Use
Cloud Data Fusion	A fully managed, cloud-native data integration service for building and managing ETL/ELT data pipelines with graphical interface for pipeline dev.	• Creating data pipelines that incorporate data validation and cleansing steps. • Connecting to various data sources and enforcing data quality rules during data ingestion.
Cloud Monitoring/ Logging	Provides monitoring and logging capabilities for your GCP resources	• Monitoring data pipelines and identifying data quality issues. • Logging data quality errors and anomalies. • Creating alerts for data quality issues
Data Catalog	A fully managed and scalable metadata management service. Discover, understand, and manage your data assets.	• Documenting data quality rules and standards. • Tracking data lineage and identifying potential data quality issues. • Tagging data with quality metrics
Vertex AI	Machine Learning Platform	• Machine Learning can be used to detect anomalies in data.ML can be used to find data drift, etc.

Data Migration Transfer Tools

Storage Transfer Service and Transfer Appliance

Storage Transfer Service

Service overview

Google Cloud Storage Transfer Service

Fully managed service that allows you to quickly and easily transfer large volumes of data into and between Google Cloud storage buckets.

- Scalable
- Secure
- Reliable

Storage Transfer Service

Service overview

Google Cloud Storage Transfer Service

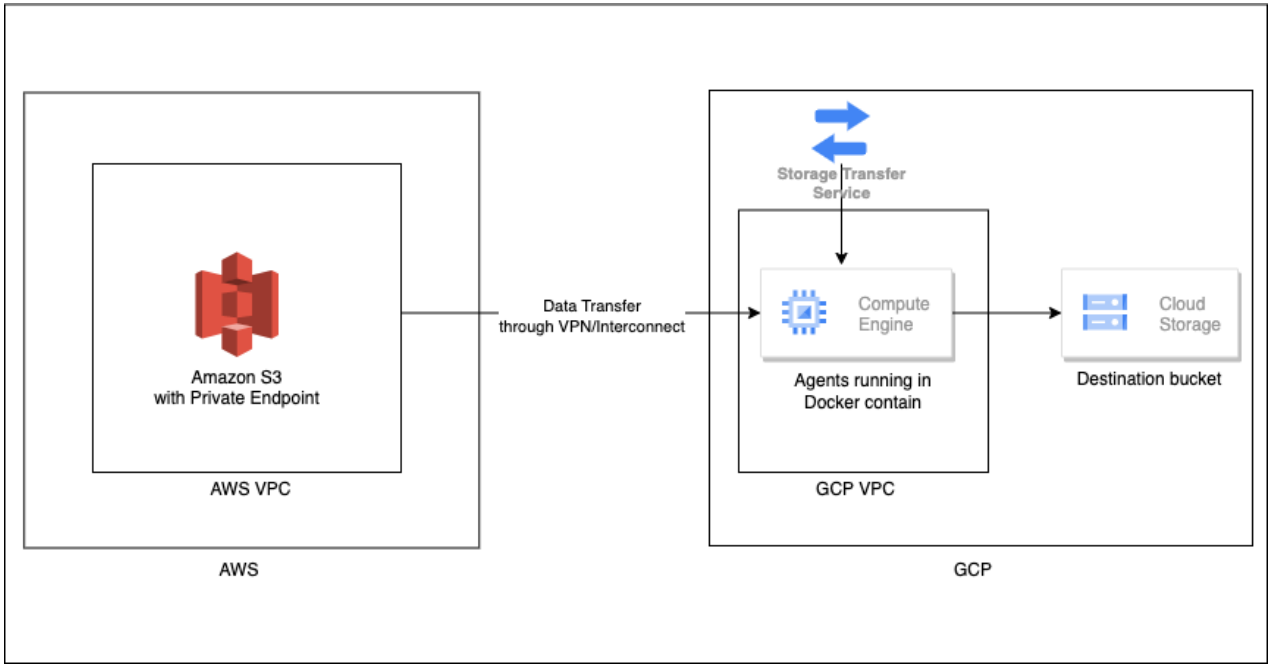
Simplifies the process of migrating data from various sources, including:

- On-premises storage
- Other cloud storage providers
- HTTP/HTTPS URLs
- Between Cloud Storage buckets

Storage Transfer Service

Service Overview

Example: Storage Transfer Service being used from AWS S3 to GCP Cloud Storage.



Transfer Appliance

Service overview

Transfer Appliance

The Google Cloud Transfer Appliance is a robust solution designed for organizations that need to move massive amounts of data to Google Cloud when network bandwidth is limited.

- High speed transfers
- Secure
- Scalable
- Low Bandwidth

Transfer Appliance

How to order a Transfer Appliance

Transfer Appliance Request Process

1. **Request the Appliance:** Request a Transfer Appliance from Google Cloud, choosing the appropriate capacity.
2. **Load Your Data:** Connect the appliance to your on-premises network and copy your data onto it.
3. **Ship It Back:** Once the data is loaded, you securely ship the appliance back to Google Cloud.
4. **Google Uploads:** Google Cloud engineers receive the appliance and upload your data to your designated Cloud Storage bucket.
5. **Data Verification and Wiping:** The data transfer is verified, and the appliance is securely wiped.

Transfer Appliance

How to order a Transfer Appliance

Formats Supported

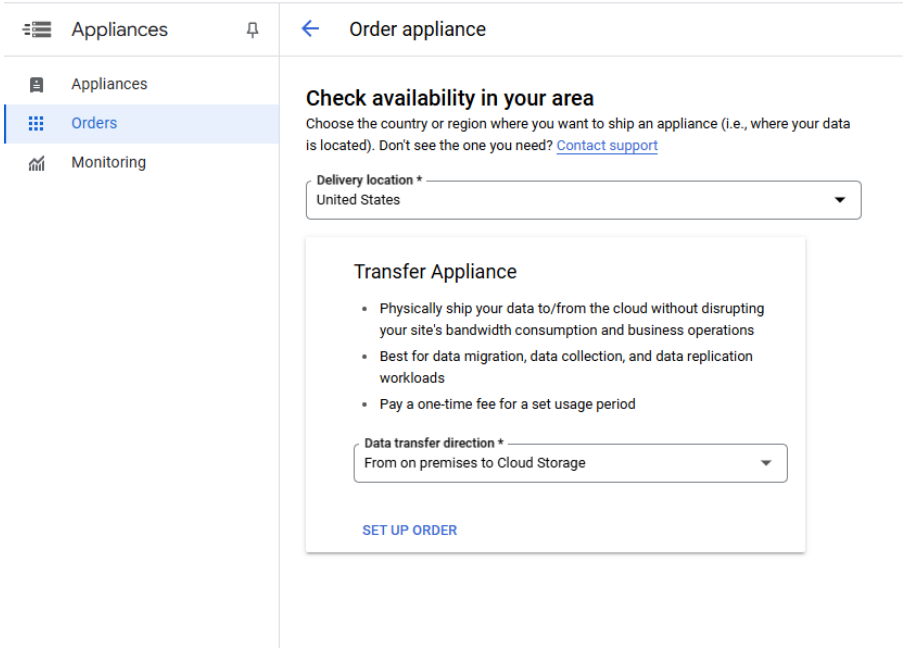
- SCP or SFTP for Microsoft Windows, Linux, and macOS.
- NFS share for Linux and macOS.
- SMB share for Microsoft Windows and Linux.
- Mount on the Appliance for NFS and CIFS.
- After you copy data to the Transfer Appliance, verify that the data transfer to the Cloud Storage bucket is complete before you delete your source data.

Transfer Appliance

How to order a Transfer Appliance

Important notes

- Online Form
- Must be approved by Google Support
- Will be shipped by Google Support
- Costs involved for set usage period
- Different versions (40TB/300TB)



Data Transfer Formats

Understanding the Factors

Data Transfer Formats

Moving data into GCP

Data Transfer Formats Factors to Consider:

The choice of format often depends on the type of data, the intended use case, and performance requirements.

- Columnar formats like Parquet and ORC are generally preferred for analytical workloads in BigQuery.
- CSV and JSON are suitable for simpler data transfers and when human readability is important.
- Data compression is also a large factor in data transfer efficiency.

Data Transfer Formats

Moving data into GCP

Table – Data Transfer Formats and Services Support

Format	Key	Services
Comma Separated Value (CSV)	• A widely used, simple format for tabular data • Easy to generate and process • Common for basic data transfers	BigQuery, Bigtable, CloudSQL, Cloud Spanner, Dataflow, Dataproc
Javascript Object Notation (JSON)	• A flexible, human-readable format for semi-structured data. • Well-suited for hierarchical data • Often used for web services and API communication	BigQuery, Dataproc
Apache Parquet	• Columnar and highly efficient for both storage and analytical queries. • Excellent compression and performance for querying large datasets.	BigQuery, Bigtable, Dataflow, Dataproc

Data Transfer Formats

Moving data into GCP

Table – Data Transfer Formats and Services Support (Continued)

Format	Key	Services
Avro	• A binary row-based format. • Designed for efficient data serialization. • Optimized for Hadoop and related ecosystems. • Known for its speed in data loading.	BigQuery, Cloud Spanner, Dataflow, Dataproc, Bigtable
Optimized Row Columnar(ORC)	• Columnar storage format similar to Parquet. • Designed for efficient data storage and retrieval (Hadoop)	BigQuery, Dataproc

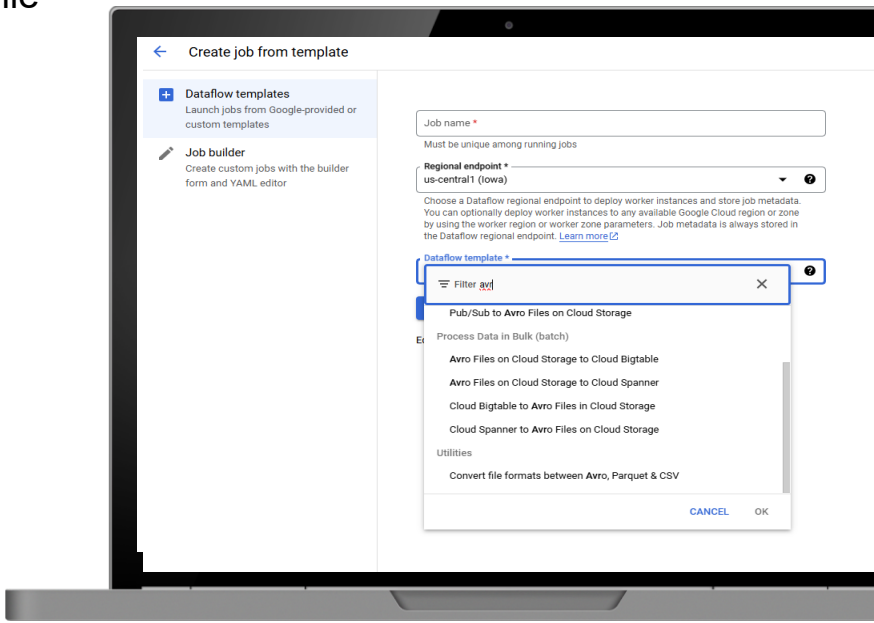
Data Transfer Formats

Templates

Dataflow Templates

Google Cloud provides **Dataflow templates** that facilitate file format conversions.
Notably, there are templates designed for conversions between:

- CSV to Avro
- CSV to Parquet
- Avro to Parquet
- Parquet to Avro



Data Profiling and Cleansing

Understanding the importance of proper Data Preparation

Data Profiling and Cleansing

Importance of data profiling and cleansing

Data Profiling and Cleansing

- **Data profiling** is the process of inspecting and analyzing data to identify its characteristics and quality. (Inventory and efficiency)

- **Data cleansing** is the process of correcting or removing errors from data. (Removing redundancy, syntax, formats, etc)

Data Profiling and Cleansing

Benefits of proper profiling and cleansing

Data Profiling and Cleansing benefits can be:

- Improved Data Consistency
- Improved Data Quality
- Enhanced Data Security

Data Profiling and Cleansing

Importance of data and its profile

Three Main Types of Data Profiling

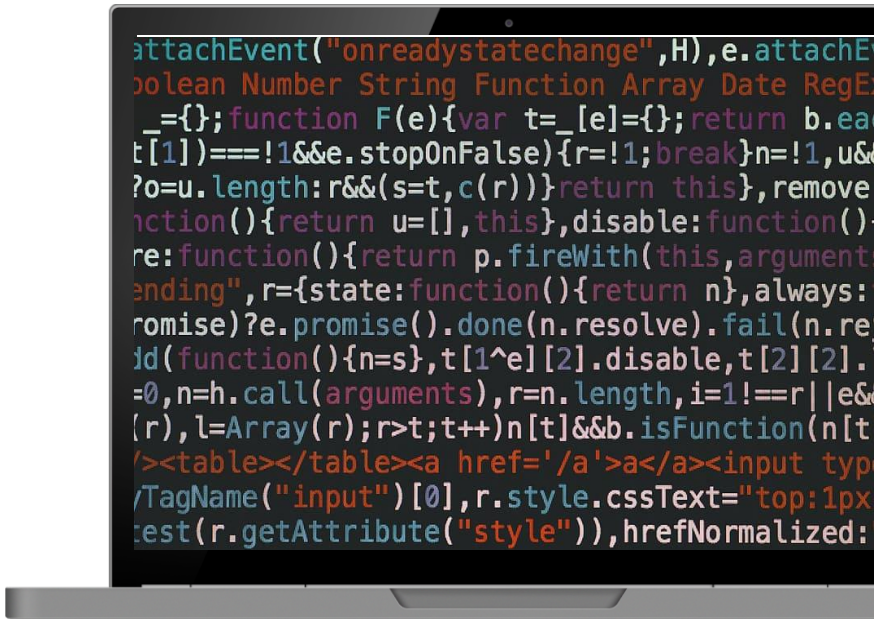
- **Structure Discovery** – Evaluating datasets to identify the number and types of fields.
- **Content Discovery** – Examining individual fields and what is contained.
- **Relationship Discovery** – Helps identify how data services can connect to.

Data Profiling and Cleansing

Data profiling and techniques

Four Main Data Profiling Techniques

- **Column Profiling** – Important step to identify datasets to match fields and datasets.
- **Cross-Column Profiling** – Next step to identify relationships between different columns or fields in the same data table.
- **Cross-Table Profiling** – Looks at the types of tables you have and the sizes.
- **Data Rule Validation** – Standardizes and cleanses the data.

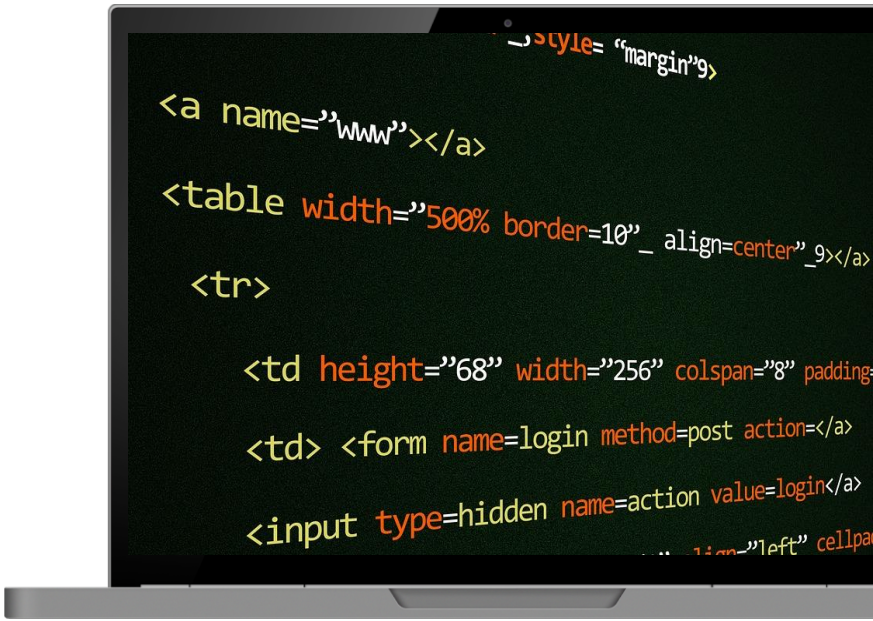


Data Profiling and Cleansing

Importance of data cleansing

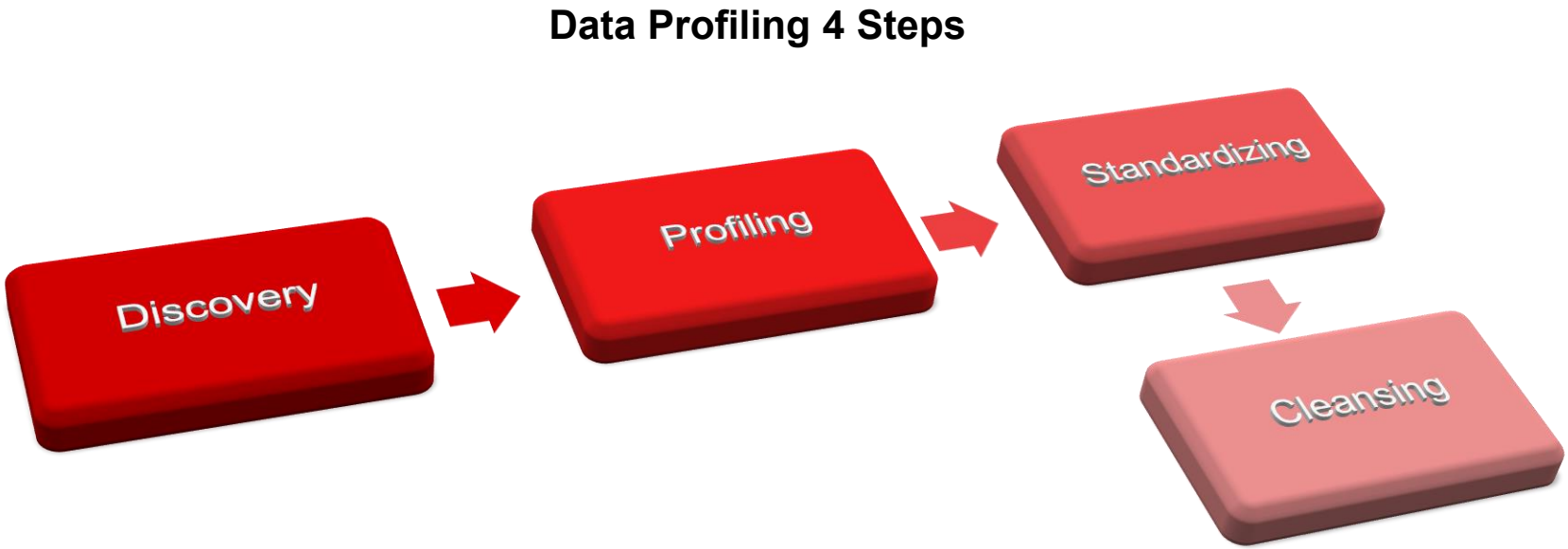
Some common examples of data cleansing:

- Removing duplicate records
- Correcting errors and typos
- Filling in missing values
- Standardizing data formats (data type validation)
- Removing outliers
- Encrypting sensitive data



Data Profiling and Cleansing

Steps for Data Profiling



Data Profiling and Cleansing

Data Profiling on Google Cloud

Dataplex

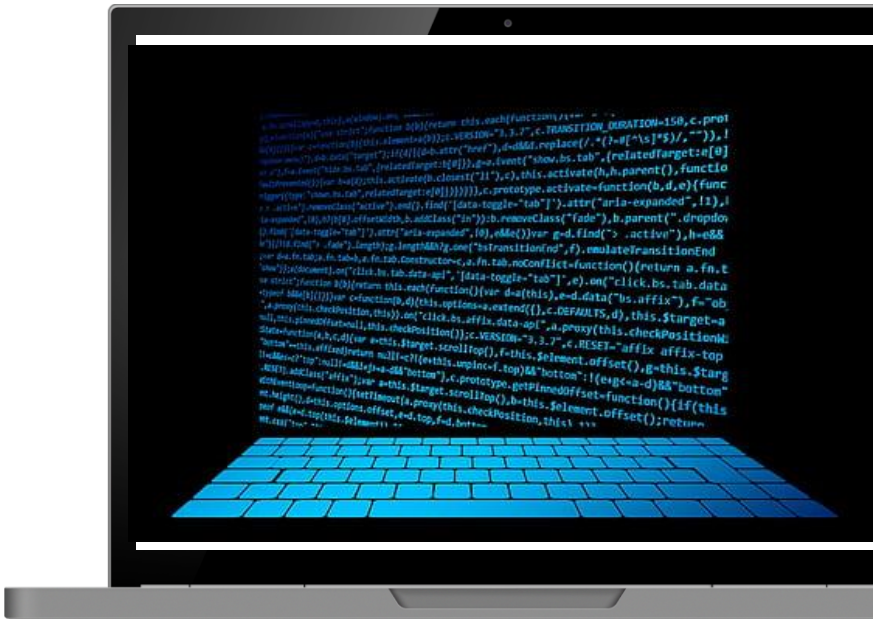
- **Dataplex** is an intelligent data fabric that enables organizations to centrally discover, manage, monitor, and govern their data across data lakes, data warehouses, and data marts with consistent controls, providing access to trusted data and powering analytics at scale.

Data Profiling and Cleansing

Data Profiling on Google Cloud

Dataplex Data Profiling

- Data profiling lets you identify common statistical characteristics of the columns in your BigQuery tables.
- It provides insights into data distribution, null counts, and other key metrics, which are essential for understanding data quality.
- It also integrates with data classification, aiding in the detection of sensitive information.



Data Profiling and Cleansing

Dataplex overview

Data Profiling with Dataplex (BigQuery Table)

- Full Table
- Partial Table
- Row or Column Filters

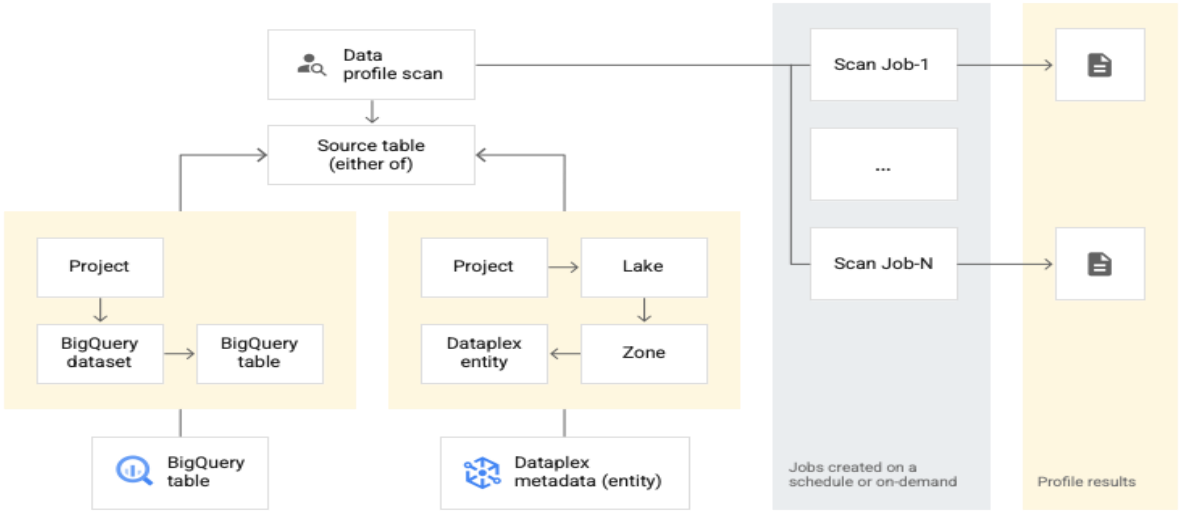


Diagram – Google Cloud

Data Structures, Files and Types

Understanding data structures and important components.

Data Structures, Files and Types

Data and its components

Overall Focus in this lesson

- Types of Data
- Data Classification
- Data Structures
- Data Formats
- Data Fields

Data Structures, Files and Types

Types of data

Types of Data

- **Qualitative** - Data that is collected in the form of words, images, or sounds. It is descriptive and cannot be easily quantified
- **Quantitative** - Data that is collected in the form of numbers. It can be easily quantified and analyzed using statistical methods.

Data Structures, Files and Types

Comparing Qualitative and Quantitative

Table: Comparing Qualitative and Quantitative

Feature	Qualitative Data	Quantitative Data
Data Types	Words, Images Sounds (Examples – Customer Feedback, Social Media, Interviews)	Numbers (Examples – Sales Figures, Customer Satisfaction, Market Share, Product Ratings)
Description	Descriptive	Can be Quantified
Analysis	Subjective	Objective
Use	Understand experiences, context	Measure, test and predict
Context	Discrete, Continuous	Nominal, Ordinal

Data Structures, Files and Types

Data Classification

Data Classification

- A **data type** is a classification of data that tells the compiler or interpreter how the programmer intends to use the data.
- Most programming languages support various types of data

Data Structures, Files and Types

Types of Data Structures

Data Structures

- Data can be **“structured”** or assigned in a specific practice.
- Data can also be unstructured with no real rules that it follows.

Data Structures, Files and Types

Types of Data Structures

The three types of Data Structures

- Structured Data
- Unstructured Data
- Semi-structured Data

Data Structures, Files and Types

Types of Data Structures

Structured data

- Structured data is data that does have a predetermined sequence, a data model, or a schema. Generally, a database (RDBMS)
- Can be human generated or machine generated
- Examples are SQL. (Key Value Pairs)

Data Structures, Files and Types

Types of Data Structures

Unstructured Data

- **Unstructured data** has no predetermined sequence, model, or schema.
- Can be human generated or machine generated
- Examples are Email, documents, social media, mobile, sensor data, satellite images, etc.

Data Structures, Files and Types

Semi-structured Data

- **Unstructured** data contains semantic tags but does not conform to the structure associated with typical relational databases.
- Can be human generated or machine generated
- Examples are XML, JSON and NoSQL

Data Structures, Files and Types

Data File Formats

Data File Format

- A **data file format** is a way of organizing data so that it can be stored and retrieved efficiently.
- There are many different data file formats, each with its own advantages and disadvantages.
- Choose the right format based on needs of application and the user requirements

Data Structures, Files and Types

Data File Formats

Table: Data File Formats

Data File Format	Description
CSV (Comma-separated values)	CSV is a simple text format that is easy to read and write. It is often used for storing tabular data, such as spreadsheets and databases.
JSON (JavaScript Object Notation)	JSON is a lightweight text format that is easy to read and write. It is often used for storing structured data, such as objects and arrays.
XML (Extensible Markup Language)	XML is a flexible text format that can be used to store a wide variety of data. It is often used for storing semi-structured data, such as documents and web pages
PDF (Portable Document Format)	File format that preserves the layout of a document, even when it is viewed on different devices.
HyperText Markup Language (HTML)	Standard markup language for creating web pages. It is a text-based language that uses tags to define the structure and content of a web page.
Text/Flat file	Two Types - Tab delimited and Comma delimited (separation of data fields)

Data Structures, Files and Types

Example

Example of a comma delimited file. (CSV)

```
Name,Age,Occupation
Joe Holbrook,40,Software Engineer
Jane Romero ,25,Teacher
```

Example of a tab delimited file. (TSV)

```
Name     | Age | Occupation
-------- | ---- | --------
John Holbrook | 40    | Software Engineer
Jane Romero | 25    | Teacher
```

Data Structures, Files and Types

Data Fields

Data Fields

- **Data fields** in a database are the individual pieces of data that are stored in a table.
- **Columns** of the table and are used to define the data that is stored in the table.
- Data types are a concern we can control when design our databases, etc.
- Each data field has a name, a data type, and a size. (unique identifier)
- Define the structure of the database.

Data Structures, Files and Types

Data Fields

Table: Comparison of Data Fields

Data Type	Description
Text	Stores any combination of letters, numbers, and symbols.
Numeric	Store numbers. The type of numeric field (integer, decimal, floating-point) determines the number of digits that can be stored and the precision of the numbers.
Date Field	Store dates and times.
Boolean	Store only two values: true or false.
Memo	A memo field can store large amount of text
Image	Stores image files

Data Structures, Files and Types

Data Fields

Table: Comparison of Data Fields (Contined)

Data Type	Description
Unique	Field that cannot contain duplicate values
AutoNumber	A number field that increments by one.
Primary Key	A primary key field is a unique identifier for each record in a table and is used to ensure each record is unique
Foreign Key	A reference to a primary key and is used to establish relationships between tables.
Other	Other types such as currency, age, DOB, etc

Data Structures, Files and Types

Data Fields

Table: Properties of Data Fields

Data Type	Description
Name	The name of the data field is a unique identifier for the data field and is used to data fields for queries, etc
Data Type	Type of Data that can be stored in the field. (Integer, float, string, and date)
Size	Specifies the size in bytes or characters
Nullability	Can a field be empty
Default Value	Value stored in data field if no value is specified

Data Structures, Files and Types

Comparing Classifications

Table: Comparison of Program Language Classification

Data Type	Description	Range
Integer	Whole Number	-2147483648 to 2147483647
Float	Decimal Point	Positive or Negative
String	Character Sequence	Any Length
Boolean	Value	True or False
Date	Specific Point in Time	1-Jan-1970 to 31-Dec-9999
Time	Specific Point in Time	00:00:00 to 23:59:59
Date Time	Time and Date	1-Jan-1970 00:00:00 to 31-Dec-9999 23:59:59

Load Data into BigQuery

Demonstration

Loading Data into BigQuery

Options and Use Cases

Key Methods to Load Data into BigQuery

Method	Description	Use Case
Loading from Google Cloud Storage (GCS)	This is a common and efficient method. You upload your data files (CSV, JSON, Avro, Parquet, ORC) to a GCS bucket and then load them into a BigQuery table.	• Large Datasets • Batch Jobs
Load from Local File	You can directly upload smaller files (CSV, JSON) from your local computer through the BigQuery web UI	• Local Testing • Development • Batch Jobs
Streaming Inserts	For real-time data ingestion, you can stream data into BigQuery tables.	• Apps with a continuous data flow • BigQuery subscription in Pub/Sub
BigQuery Data Transfer Service	This service automates data loading from various Google services (like Google Ads, Google Analytics) and other data sources	• Simply routine data transfers into BigQuery

Loading Data into BigQuery

Options and Use Cases

Key Methods to Load Data into BigQuery (Continued)

Method	Description	Use Case
Using the bq Command-Line Tool	The bq command-line tool provides a powerful and flexible way to load data from various sources.	• Scripting and Automation
BigQuery API	For programmatic data loading, you can use the BigQuery API.	• Integrate into your applications
Load from Google Sheets	Link Google Sheets to BigQuery	• Use with Google Sheets
Change Data Capture	This method enables replicating data from databases to BigQuery in near real time.	• Use DataStream • Other supported data sources

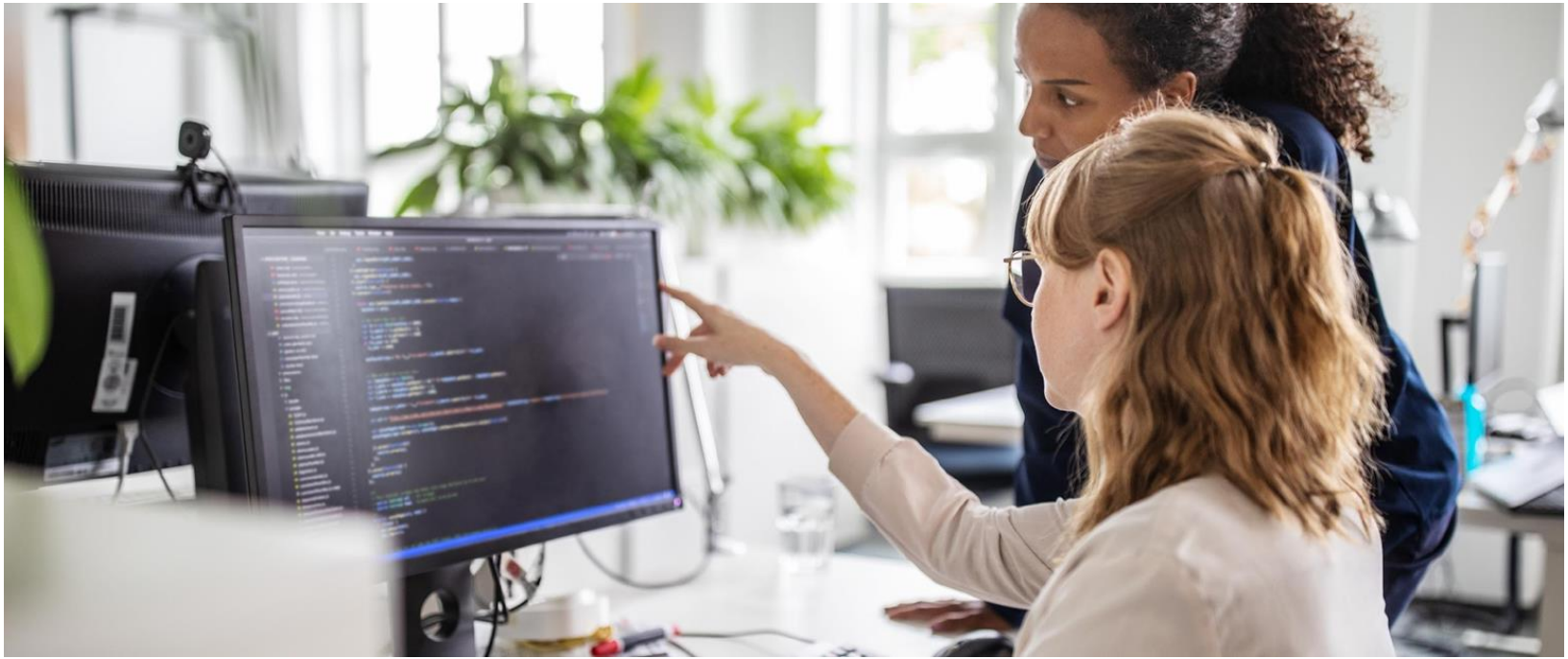
Load Data into Bigquery

Demonstration Scenario

We'll walk thru options in the BigQuery dashboard for loading data.

Loading Data into BigQuery

Demonstration



Understanding Storage Options in GCP

GCP Storage options and use cases

Understanding Storage Options in GCP

Storage Options in GCP

Storage Options

- Data is the lifeblood of modern applications.
- GCP offers a variety of storage solutions tailored to different needs.
- Choosing the right storage impacts performance, scalability, and cost.

Understanding Storage Options in GCP

Storage Options in GCP

Storage Options

- **Cloud Storage:** Object storage for unstructured data
- **BigQuery:** Data warehouse for analytics
- **Cloud SQL:** Relational database service
- **Firestore:** NoSQL document database
- **Bigtable:** NoSQL wide-column database
- **Spanner:** Globally distributed, scalable relational database

Understanding Storage Options in GCP

Storage Options in GCP

Cloud Storage

Use Cases: Images, videos, backups, archives, and unstructured data

Key Features:

- **Storage classes** (Standard, Nearline, Coldline, Archive).
- Object versioning
- Lifecycle management
- Good for storing large amounts of unstructured Data

Understanding Storage Options in GCP

Storage Options in GCP

BigQuery – Serverless Data Warehouse

Use Cases: Analytics, business intelligence, data warehousing.

Key Features:

- SQL-based querying
- Serverless Architecture
- Scalable and cost-effective
- Good for analyzing large datasets.

Understanding Storage Options in GCP

Storage Options in GCP

Firestore – NoSQL Document Database

Use Cases: Mobile, web applications, real-time data
Key Features:

- Real-time synchronization
- Scalable and flexible
- Offline support
- Good for flexible schemaless data, and real time applications

Understanding Storage Options in GCP

Storage Options in GCP

Cloud SQL – Managed Relational Database

Use Cases: Web Apps, Transactional Apps
Key Features:

- Managed MySQL, PostgreSQL, and SQL Server instances.
- Automatic backups and patching.
- High availability options.
- Good for applications that require relational database functionality.

Understanding Storage Options in GCP

Storage Options in GCP

BigTable– NoSQL Wide-Column Database

Use Cases: Large Scale Analytics, Time Series Data,IoT Apps
Key Features:

- High throughput and low latency
- Scalable to petabytes of data
- Ideal for sparse, structured data
- Good for massive amounts of structured or semi-structured data

Understanding Storage Options in GCP

Storage Options in GCP

Cloud Spanner – Global Applications, Financial, Large Scale OLTP

Use Cases: Large Scale Analytics, Time Series Data,IoT Apps

Key Features:

- Global consistency and high availability
- Horizontal scalability
- SQL with ACID transactions
- Good for global applications that require strong consistency and high availability

Understanding Storage Options in GCP

Choosing the right Storage Option

Key Considerations for Choosing Storage

- Data structure (structured, semi-structured, unstructured)
- Data volume and velocity
- Access patterns (read-heavy, write-heavy)
- Consistency requirements (strong, eventual)
- Latency requirements
- Costing model

Mapping Business Requirements to Use Cases

Whiteboard Discussion

Business Requirements

What are Business Requirements?

Business requirements are the needs of a business that a project or system is intended to meet.

- Expressed as Goals or Expectations
- Can be Functional, Non-Functional
- Business Analysis is the practice of identifying these.
- Defined in a Business Requirements Document (BRD).

Business Requirements

What are Business Requirements? (Continued)

Business requirements are determined and assessed by various methods.

- Business Stakeholders
- Distribution List
- Data Models
- Report Types

Google Cloud Associate Data Practitioner

Whiteboard Discussion

Module Review

Summary Recap

Module Review

Summary

Let’s summarize some of the important topics we covered in this module.

- Data practitioners are essential for extracting value from data.They empower organizations to make better decisions, improve efficiency, drive innovation, and mitigate risk.
- As the volume and complexity of data continue to grow, the demand for skilled data practitioners will only increase.
- Data quality refers to data’s overall “fitness” for its intended use.
- Google Cloud Transfer Appliance is a robust solution designed for organizations that need to move massive amounts of data to Google Cloud when network bandwidth is limited.
- Columnar formats like Parquet and ORC are preferred for analytical workloads in BigQuery.
- CSV and JSON are suitable for simpler data transfers and when human readability is important.
- Cloud Spanner’s use cases are large-scale analytics, time series data, and IoT apps.

Module Review

Summary

Let’s summarize some of the important topics we covered in this module.

- BigQuery’s use cases are analytics, business intelligence, and data warehousing.
- Data profiling inspects and analyzes data to identify its characteristics and quality.
- Data cleansing is the process of correcting or removing errors from data.
- A data file format is a way of organizing data so that it can be stored and retrieved efficiently.
- Various methods for loading data into BigQuery exist, catering to different data sources and use cases.
- Business requirements are specific to projects and can be expressed as goals, which should translate to a specific technology requirement.
- Cloud Storage use cases are images, videos, backups, archives, and unstructured data
- There are several important considerations when choosing storage, such as latency, regions, speed, compatibility, etc

Data Pipeline Orchestration

Understanding GCP Services for data pipelines

Module Overview

What we will cover in the Module

- Selecting a data transformation tool
- Evaluate use cases for ETL/ELT
- Choose the right GCP service to implement for a pipeline
- Demonstration: Create and manage scheduled queries
- Demonstration: Monitor Dataflow pipeline progress
- Demonstration: Review and analyze logs
- Whiteboard Discussion - data orchestration solution
- Whiteboard Discussion- Identify use cases for event-driven data ingestion
- Demonstration - Use Eventarc triggers in event-driven pipelines
- Module Review

Selecting a Data Transformation Tool in GCP

Dataproc, Dataflow, Cloud Data Fusion, Cloud Composer, and Dataform

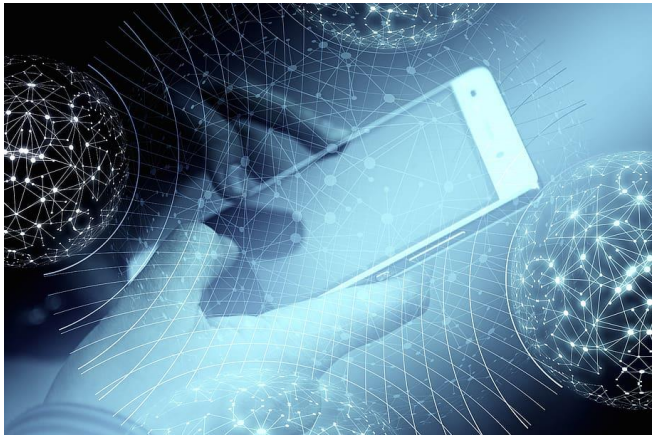
Selecting a Data Transformation Tool

Choosing a Tool

Key Considerations for Choosing a Data Transformation Tool

When choosing which tool to use, it is best to consider:

- The volume and velocity of your data.
- The complexity of your transformations.
- Whether you need batch or stream processing.



Selecting a Data Transformation Tool

Choosing a Tool

Selection Requirement: Simple, Code-Free/Low-Code Integration & Transformation:

Business Requirement: Need to quickly connect various data sources and perform basic transformations without extensive coding.

Tool Selection: Cloud Data Fusion

- Its graphical interface and pre-built connectors simplify pipeline creation.
- Ideal for users with limited coding experience
- Rapid development and deployment for straightforward integration needs.

Selecting a Data Transformation Tool

Choosing a Tool

Selection Requirement: Complex, Scalable and Real Time Data Processing

Business Requirement: Need to process high-volume, streaming data in real-time and perform complex transformations

Tool Selection: Dataflow

- Serverless and highly scalable, capable of handling large data volumes.
- Apache Beam provides flexible and powerful data processing capabilities.
- Excellent for real-time analytics and event-driven processing.

Selecting a Data Transformation Tool

Choosing a Tool

Selection Requirement: Large-Scale Batch Processing with Hadoop/Spark

Business Requirement: Need to perform complex, large-scale batch transformations using Hadoop or Spark

Tool Selection: Dataproc

- Managed Hadoop and Spark service, simplifying cluster management.
- Ideal for data engineering tasks requiring the power of Hadoop and Spark.
- Good for migrating existing Hadoop/Spark workloads.

Selecting a Data Transformation Tool

Choosing a Tool

Selection Requirement: Workflow Orchestration and Scheduling

Business Requirement: Need to orchestrate and schedule complex data pipelines, managing dependencies between tasks

Tool Selection: Cloud Composer

- Managed Apache Airflow service, providing workflow orchestration capabilities.
- Allows you to define and manage complex data workflows.
- Essential for managing dependencies and scheduling data pipelines.

Selecting a Data Transformation Tool

Choosing a Tool

Selection Requirement: **SQL-Based Data Transformation and Data Warehousing**

Business Requirement: Need to perform SQL-based transformations within a data warehouse environment, focusing on data modeling and transformation for analytics.

Tool Selection: Dataform

- Designed for SQL-based data transformation in BigQuery.
- Enables version control, testing, and collaboration for data warehouse workflows.
- Centralized transformation logic and enforces data governance.

Selecting a Data Transformation Tool

Scenarios

Tool Selection Scenarios

- **Scenario 1:** A retail company needs to analyze real-time customer behavior to personalize offers. They need to process clickstream data and perform complex aggregations in real-time.
  Tool: Dataflow

- **Scenario 2:** A financial institution needs to perform daily batch processing of large transaction datasets for fraud detection using Spark.
  Tool: Dataproc

- **Scenario 3:** A marketing team needs to integrate data from various SaaS applications into a data warehouse for reporting and analysis without requiring advanced coding skills.
  Tool: Cloud Data Fusion

Selecting a Data Transformation Tool

Scenarios

Tool Selection Scenarios

- **Scenario 4:** A data engineering team needs to build a multi-step ETL pipeline with dependencies between tasks, requiring scheduling and monitoring.
  Tool: **Cloud Composer.**

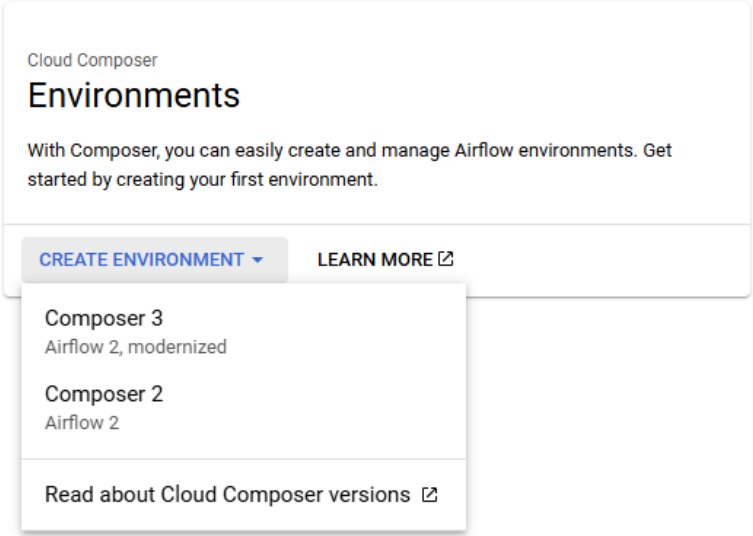
- **Scenario 5:** A data warehouse team focuses on utilizing best practice software development lifecycle practices, but for SQL based data transformations within Bigquery.
  Tool: Dataform

Selecting a Data Transformation Tool

Tools

Tool Selection Summary

- **Cloud Composer:** Orchestrates workflows; transformation logic is embedded within the workflow.
- **Cloud Data Fusion:** Fully managed ETL/ELT service with a graphical interface.
- **Dataform:** Focuses on SQL-based data modeling and transformations within BigQuery.
- **Dataproc:** Handles large-scale batch processing using Spark and Hadoop.
- **Dataflow:** Processes both batch and streaming data with Apache Beam.



Selecting a Data Transformation Tool

Considerations

Important Considerations

- **ELT vs. ETL:** BigQuery and Dataform favor ELT (Extract, Load, Transform), where data is loaded into the data warehouse first, then transformed. Dataflow and Cloud Data Fusion can handle both ETL and ELT.
- **Complexity:** Simple SQL transformations can be handled directly in BigQuery. Complex transformations may require Dataflow or Dataproc.
- **Real-time vs. Batch:** Dataflow is ideal for real-time streaming data. Dataproc and BigQuery are well-suited for batch processing.
- **Orchestration:** Cloud Composer is used to orchestrate complex data pipelines that may involve multiple transformation too

Evaluate Use Cases for ELT/ETL

Extraction, Transformation and Loading Data on GCP

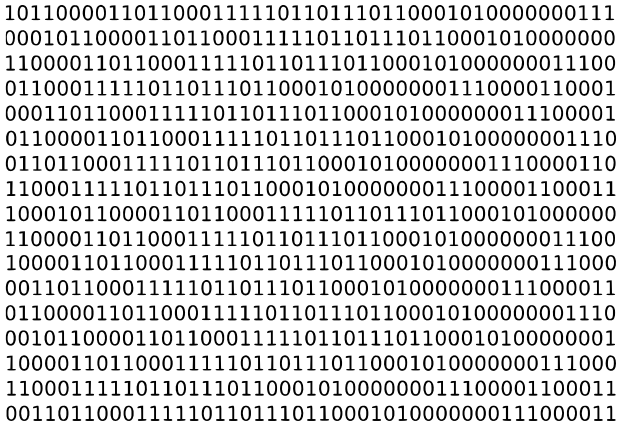
Evaluate Use Cases for ELT/ETL

Extraction, Transformation and Loading

Defining Extract, Transform, and Load (ETL)

- ETL is a process of extracting data from one or more sources, transforming it into a format suitable for loading into a data warehouse or other destination, and then loading it into the destination.

- ETL is commonly used in data warehousing, data lakes, data analytics, and machine learning.



Evaluate Use Cases for ELT/ETL

Use Cases

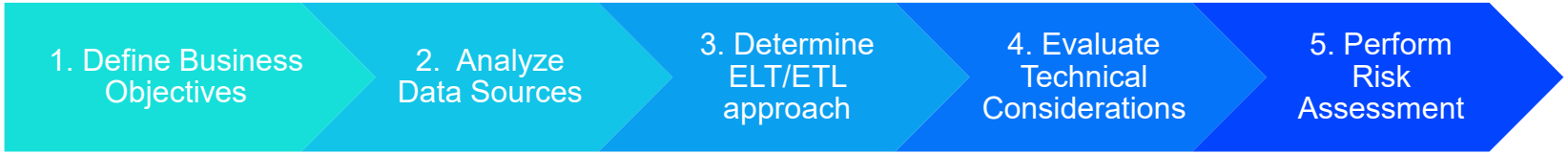
Key Questions to Ask when looking at a use case:

- What is the frequency of data updates?
- What is the expected growth of the data?
- What level of data quality is required?
- What are the performance requirements?
- What are the security and compliance requirements?

Evaluate Use Cases for ELT/ETL

Workflow

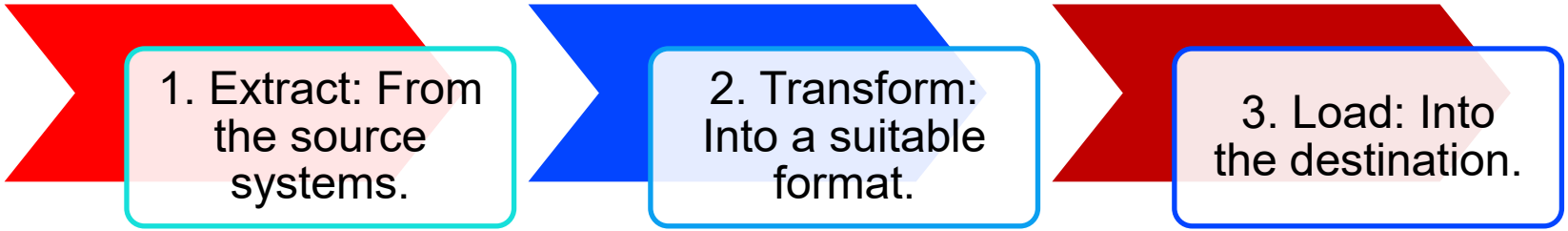
Use Case Analysis Workflow



Evaluate Use Cases for ELT/ETL

Understanding the Importance of ETL

The ETL Process



Evaluate Use Cases for ELT/ETL

Understanding the Importance of ELT

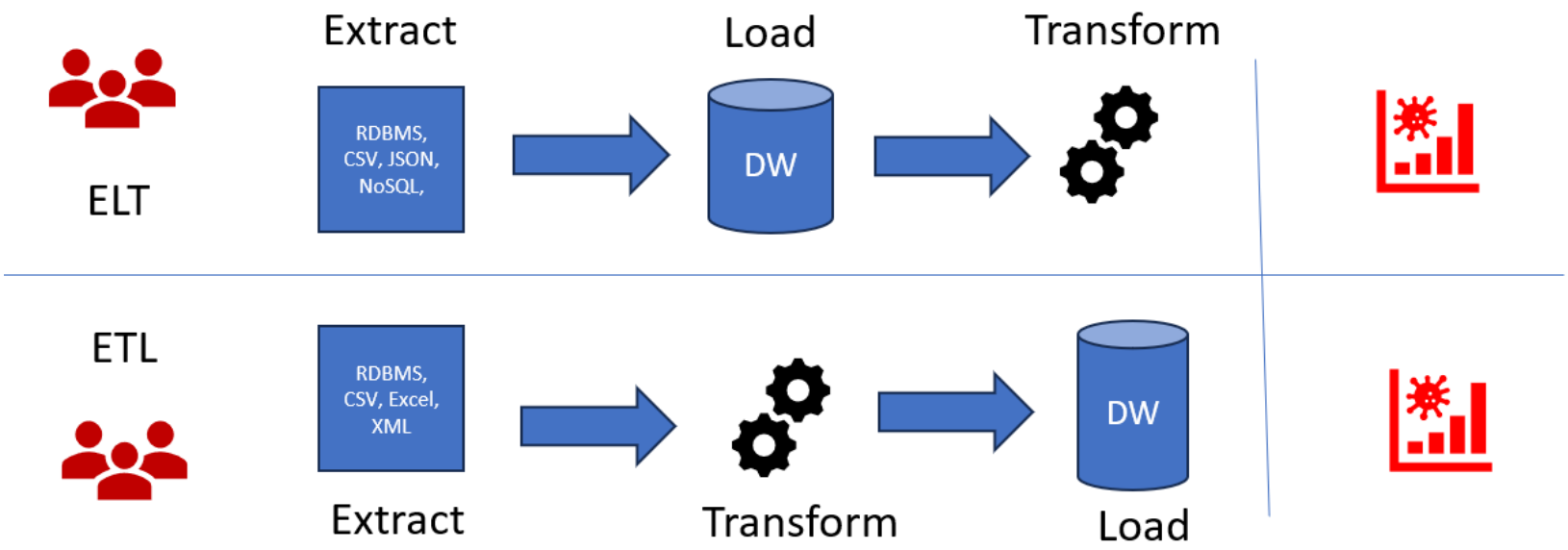
ELT Process

- Data is extracted from the source systems and loaded directly into the destination.
- Transformation is then performed in the destination.
- ELT is becoming more popular since it can be efficient than ETL.

Evaluate Use Cases for ELT/ETL

Comparing

Comparing Extract, Transform, Load (ETL) - Extract, Load, Transform (ELT)



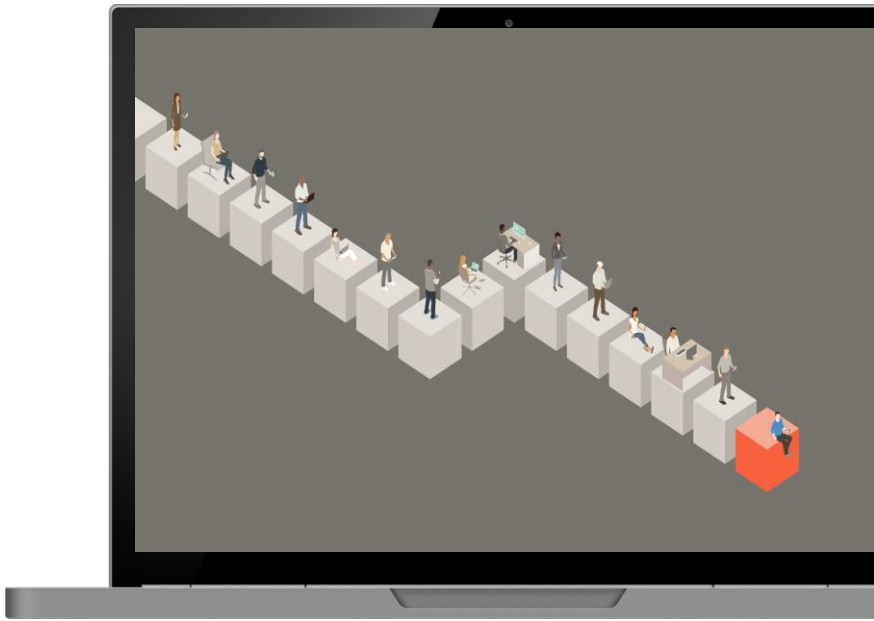
Evaluate Use Cases for ELT/ETL

Understanding the Importance of ETL

ELT Process Loading

Loading Data is the process of loading data from a source system in a data warehouse

- Full Load - Loads all the data from the data system
- Delta Load - Loads only the new or changed data. (Implemented with a watermark)



Evaluate Use Cases for ELT/ETL

Extracting Data

Extracting Data is the process of loading data from a source system in a data warehouse

- Examples of Tools - SQLDemoBench, Power BI, and Cloud Fusion are tools for extracting data from external databases
- Comma-separated Values (CSV) are commonly used for extracting data from a source.

Evaluate Use Cases for ELT/ETL

Comparing both processes

Comparing Extract, Transform, Load (ETL) and Extract, Load, Transform (ELT)

	ETL	ELT
Extraction	Raw data is extracted from disparate sources	Raw data is extracted from disparate sources
Transformation	Raw data is transformed on a secondary staging area or moved back to server (Staging Area)	Raw data is transformed within the DW (Destination)
Loading	Data loaded in the DW after transformation	Raw data is loaded into the DW
Usage	Source to Target, Intensive transformations or small data sets	Large amounts of data
Data Type	Best for structured data	Best for unstructured data
Reliability	More manageable	Less manageable

Choose GCP Services to Implement a Data Pipeline

Understanding the options

Choose GCP Services to Implement a Data Pipeline

Comparing GCP Services for ELT/ETL

Services in Google Cloud for ETL/ELT Pipelines

GCP Service	Description
Cloud Composer	• A fully managed workflow orchestration service on GCP. • Use to schedule, manage and monitor ETL/ELT pipelines • Automate workflows
Cloud Data Fusion	• A fully managed, cloud-native data integration service with a graphical interface. • Simplifies the creation and management of ETL/ELT pipelines without requiring extensive coding. • Provides pre-built connectors for various data sources and sinks.
Cloud Dataflow	• A fully managed service for executing Apache Beam pipelines. • Ideal for both batch and stream processing, making it suitable for complex data transformations. • Excellent for tasks like data cleaning, aggregation, and enrichment.
Cloud Dataproc	• A managed Apache Spark and Hadoop service. • Enables you to run large-scale data processing jobs using open-source tools. • Suitable for complex transformations and data processing tasks.

Choose GCP Services to Implement a Data Pipeline

Comparing GCP Services for ELT/ETL

Services in Google Cloud for ETL/ELT (Continued)

GCP Service	Description
Dataprep	• A serverless data preparation service. • Allows users to visually explore, clean, and prepare data for analysis. • Streamlines the data cleaning and transformation process.
Cloud Pub/Sub	• A real-time messaging service that allows you to ingest and distribute data streams. • Essential for building streaming ETL/ELT pipelines. • Enables decoupling of data producers and consumers.
Cloud Storage	• This is the foundation for storing your raw data, staging data, and output data. It's a scalable and durable object storage service. • It's essential for both ETL and ELT processes, acting as a data lake or data staging area.
BigQuery	• Serverless, highly scalable data warehouse. • Important role in ELT pipelines, where data is loaded first and then transformed using SQL. • Also used as a destination for processed data in ETL pipelines.

Monitor Dataflow Pipeline Progress

Demonstration

Monitor Dataflow Pipeline Progress

How Can you Monitor your Dataflow Pipeline?

Some Examples:

- Set up an alert in Cloud Monitoring to notify you if the error rate of your pipeline exceeds a certain threshold.
- Create a dashboard in Cloud Monitoring to visualize the throughput and latency of your pipeline.
- Use Cloud Logging to search for specific error messages in your pipeline logs.
- Use the Dataflow Monitoring UI to view the execution graph of your pipeline and identify performance bottlenecks.

Create and manage Scheduled Queries

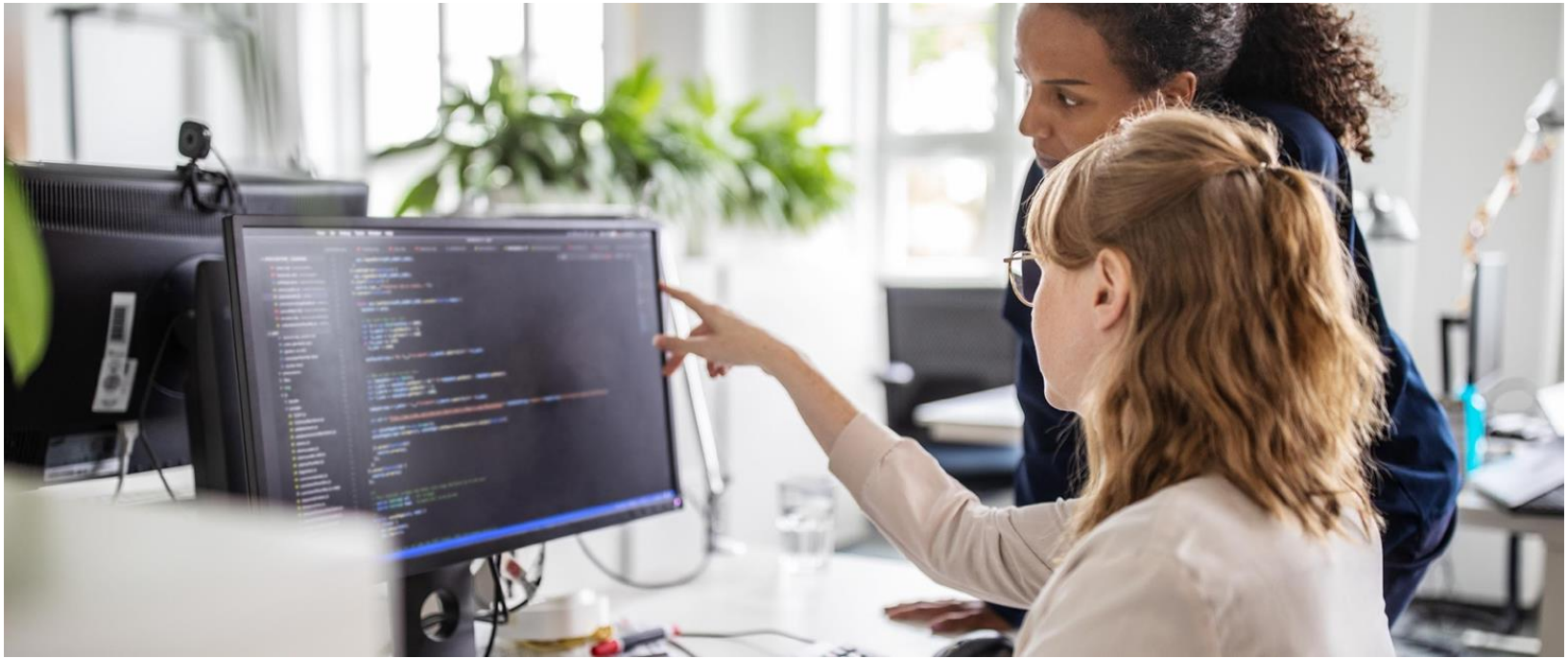
Demonstration Scenario

We'll walk thru some important concepts for monitoring a Dataflow pipeline.

- Dataflow UI

Monitor Dataflow Pipeline Progress

Demonstration



Review and Analyze Logs

Demonstration

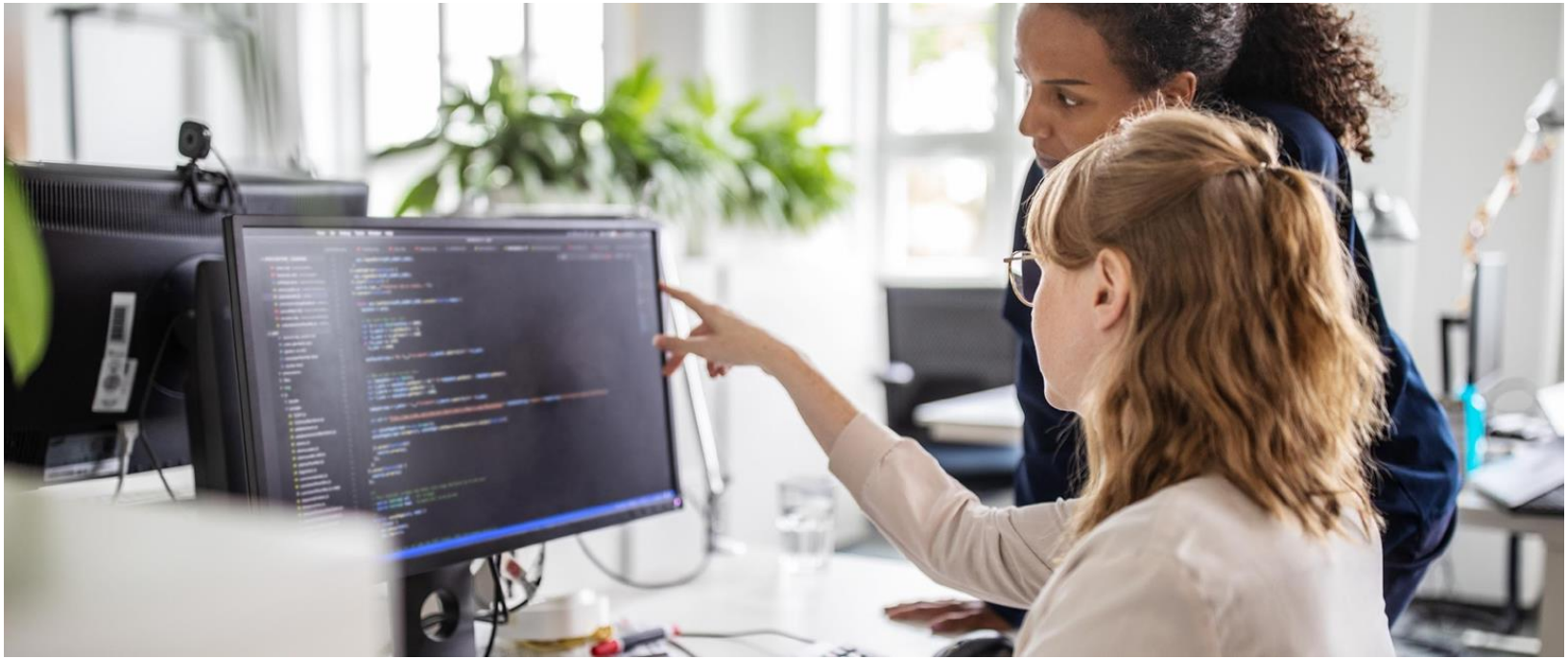
Review and Analyze Logs

Demonstration Scenario

We'll dive into Cloud Monitoring and Logging, review logs, and walk through monitoring dashboards.

Review and Analyze Logs

Demonstration



Select a Data Orchestration Solution

Whiteboard Discussion

Select a Data Orchestration Solution

Whiteboard Discussion

We'll walk through a customer challenge scenario and discuss which GCP tools would be the best or correct choices for data orchestration

Google Cloud Associate Data Practitioner

Whiteboard Discussion

Identify Use Cases for Event-driven Data Ingestion

Whiteboard Discussion

Identifying Use Cases for Event-driven Data Ingestion

Demonstration Scenario

We'll walk through a customer challenge scenario and discuss which GCP tools would be the best or correct choices for event-driven data ingestion.

Google Cloud Associate Data Practitioner

Whiteboard Discussion

Use Eventarc triggers in Event-driven Pipelines

Demonstration

Use Eventarc triggers in Event-driven Pipelines

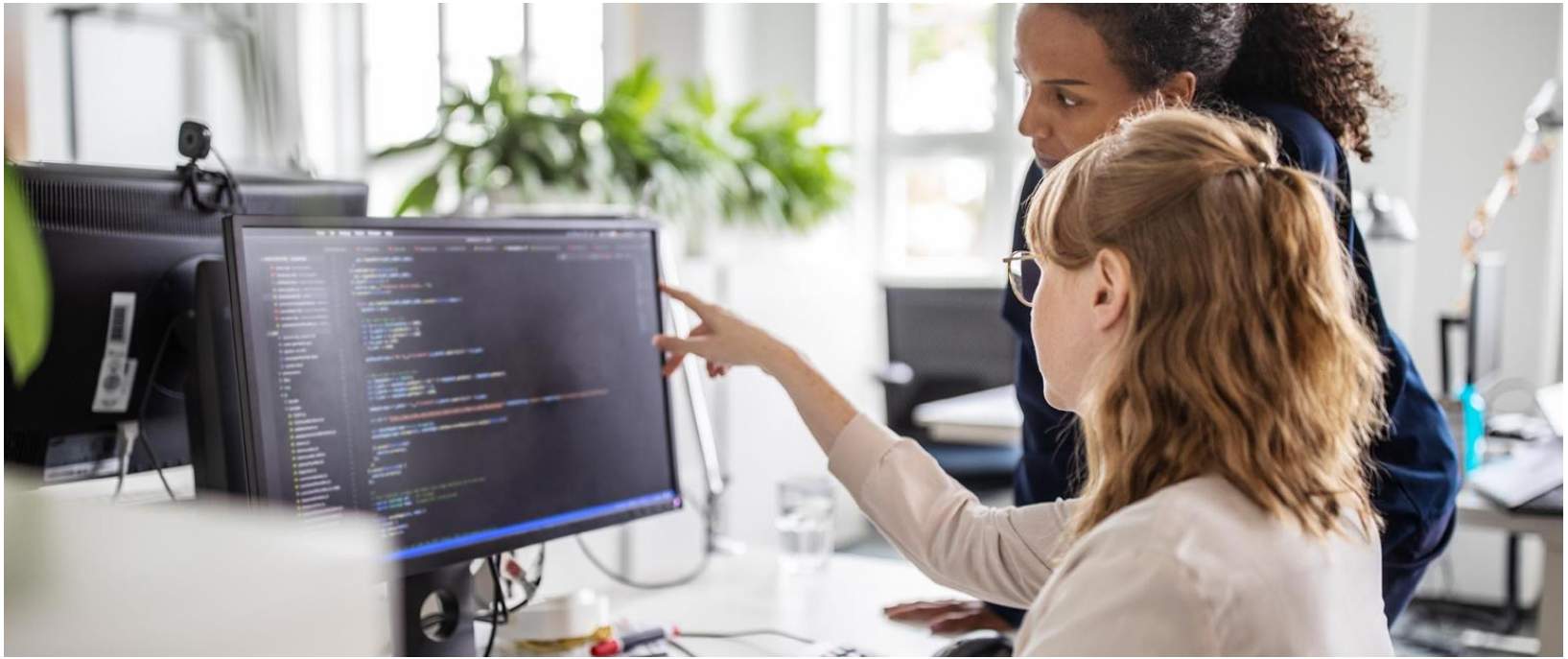
Demonstration Scenario

We'll walk thru various features and then create a trigger event in Eventarc

This upload event will trigger a Cloud Function via Eventarc, which will then process the image and store the processed image in another Cloud Storage bucket.

Google Cloud Associate Data Practitioner

Demonstration



Module Review

Summary Recap

Module Review

Summary

Let’s summarize some of the important topics we covered in this module.

- If your project requires orchestrating and scheduling complex data pipelines, as well as managing dependencies between tasks, then Cloud Composer is the best choice.
- If your project requires SQL-based transformations within a data warehouse environment, focusing on data modeling and transformation for analytics, then select Dataform.
- If your project requires processing high-volume, streaming data in real-time and performing complex transformations, then select Dataflow.
- ETL is a process of extracting data from one or more sources, transforming it into a format suitable for loading into a data warehouse or other destination, and then loading it into the destination.
- Loading Data is the process of loading data from a source system into a data warehouse And there are two types of loads.(Full and Delta)

Module Review

Summary

Let’s summarize some of the important topics we covered in this module.

- Dataflow is a fully managed batch and stream data processing service based on Apache Beam, an open-source unified programming model.
- The Dataflow job UI within the Google Cloud Console provides a wealth of information to help you track progress and diagnose potential issues with your Dataflow.
- Use the Dataflow Monitoring UI to view the execution graph of your pipeline and identify performance bottlenecks.
- Use Cloud Logging to search for specific error messages in your pipeline logs.
- Eventarc is a serverless service that allows you to build event-driven architectures. This means that applications respond to "events," which are changes in state or occurrences.

Data Analysis and Presentation

Using tools in GCP for identifying data requirements

Module Overview

What we will cover in the Module

- Identify data trends, patterns, and insights
- Demonstration: Define and execute SQL queries in BigQuery
- Demonstration: Visualize data and create dashboards
- Compare Looker and Looker Studio
- Define, train, evaluate, and use ML models
- Whiteboard Discussion: Plan an actual ML Project
- Module Review

Identify Data Trends, Patterns, and Insights

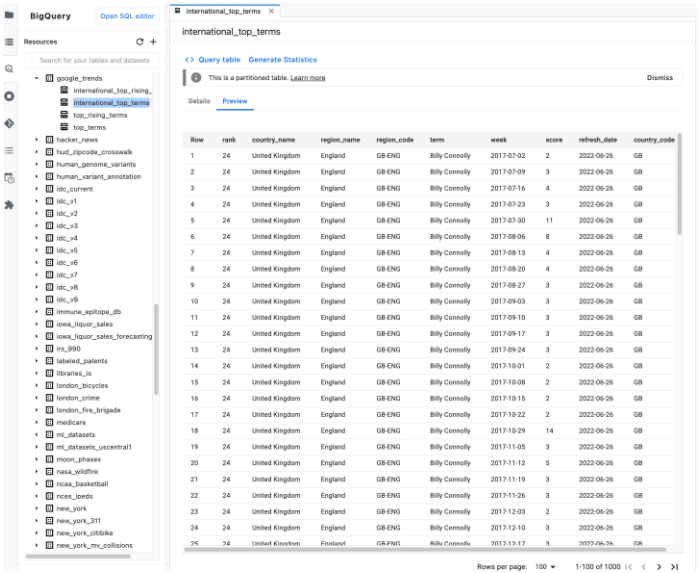
Using BigQuery and Jupiter Notebooks

Identify Data Trends, Patterns, and Insights

Using BigQuery and Jupyter Notebooks

Overall Steps

- Access your Google Cloud Platform console and navigate to BigQuery.
- Create a Jupyter Notebook instance on Google Cloud Vertex AI Workbench.
- Authenticate your Jupyter notebook to access your BigQuery dataset.
- Jupyter Notebooks can connect to BigQuery using libraries like google-cloud-bigquery and BigQuery Dataframes. This allows you to retrieve query results directly into your notebook.



Identify Data Trends, Patterns, and Insights

Using BigQuery and Jupyter Notebooks

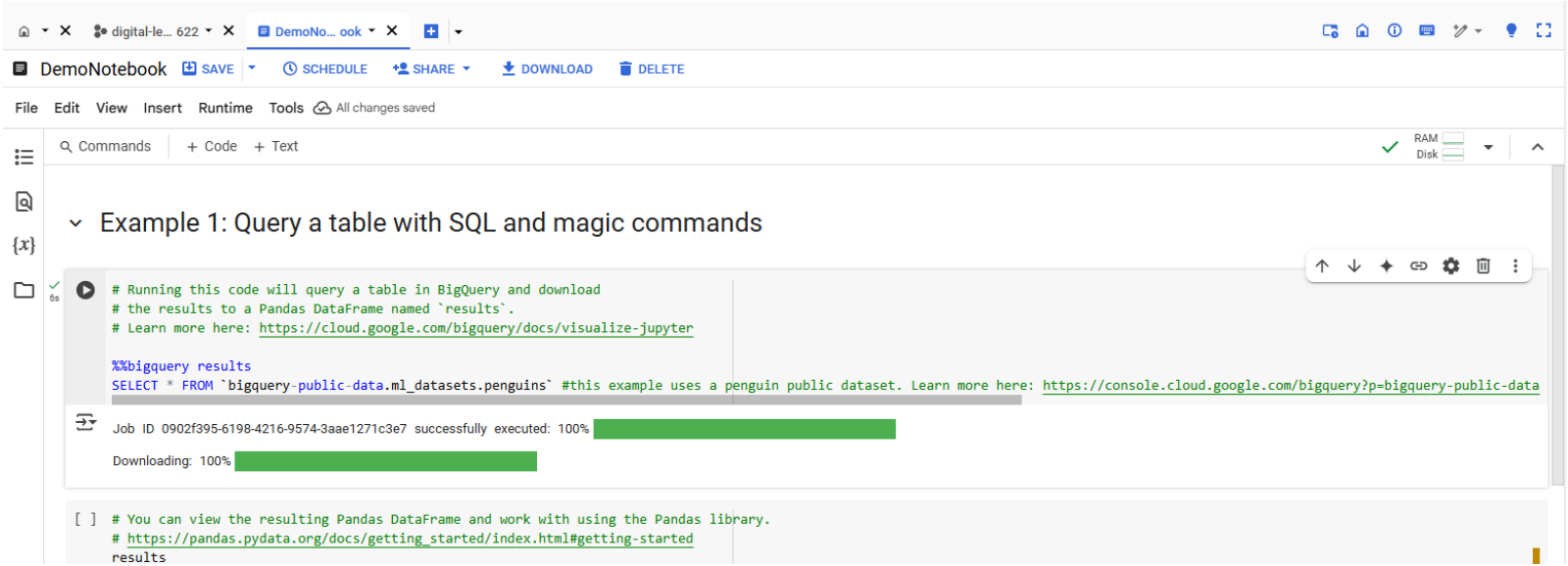
What is a Notebook

- A notebook provides an environment in which to author and execute code. A notebook is essentially a source artifact, saved as an IPYNB file.
- It can contain descriptive text content, executable code blocks, and output rendered as interactive HTML.
- Structurally, a notebook is a sequence of cells. A cell is a block of input text that is evaluated to produce results. Cells can be of three types:
- Code cells contain code to evaluate. The output or results of executed code are rendered in line with the executed code.
- Markdown cells contain Markdown text converted to HTML to produce headers, lists, and formatted text.
- Raw cells can render different code formats into HTML or LaTeX.

Identify Data Trends, Patterns, and Insights

Using BigQuery and Jupyter Notebooks

Notebook in BigQuery



Identify Data Trends, Patterns, and Insights

Using BigQuery and Jupyter Notebooks

Jupyter Notebook Environment:

You can use Google Colab, Google Cloud's Vertex AI Workbench (managed notebooks), or a local Jupyter Notebook installation.

Install necessary Python libraries:

- google-cloud-bigquery: To connect and interact with BigQuery.
- pandas: For data manipulation and analysis.
- matplotlib and seaborn: For data visualization.
- plotly (optional): For interactive visualizations.
- scikit-learn (optional): For machine learning if needed.

Identify Data Trends, Patterns, and Insights

Using BigQuery and Jupyter Notebooks

```
#Python Code for connecting to BigQuery

from google.cloud import bigquery
import pandas as pd

# Initialize BigQuery client
client = bigquery.Client()

# Example: Construct a reference to the table
table_id = "your-project.your_dataset.your_table" #replace with your table info.
table = client.get_table(table_id)

print(f"Loaded {table.num_rows} rows and {len(table.schema)} columns to {table_id}")
```

Identify Data Trends, Patterns, and Insights

Using BigQuery and Jupyter Notebooks

Jupyter Notebook Workflows

BigQuery handles large-scale data processing, while Jupyter Notebooks provides an interactive environment for in-depth analysis and visualization, leading to the discovery of valuable trends, patterns, and insights.

- Trend Analysis: Use BigQuery to aggregate data over time and then visualize the results in Jupyter Notebooks to identify trends.
- Pattern Recognition: Apply machine learning algorithms in Jupyter Notebooks to discover hidden patterns in your data.
- Insight Generation: Combine SQL queries in BigQuery with data visualization in Jupyter Notebooks to generate actionable insights.

Define and execute SQL Queries

Demonstration

Define and Execute SQL Queries

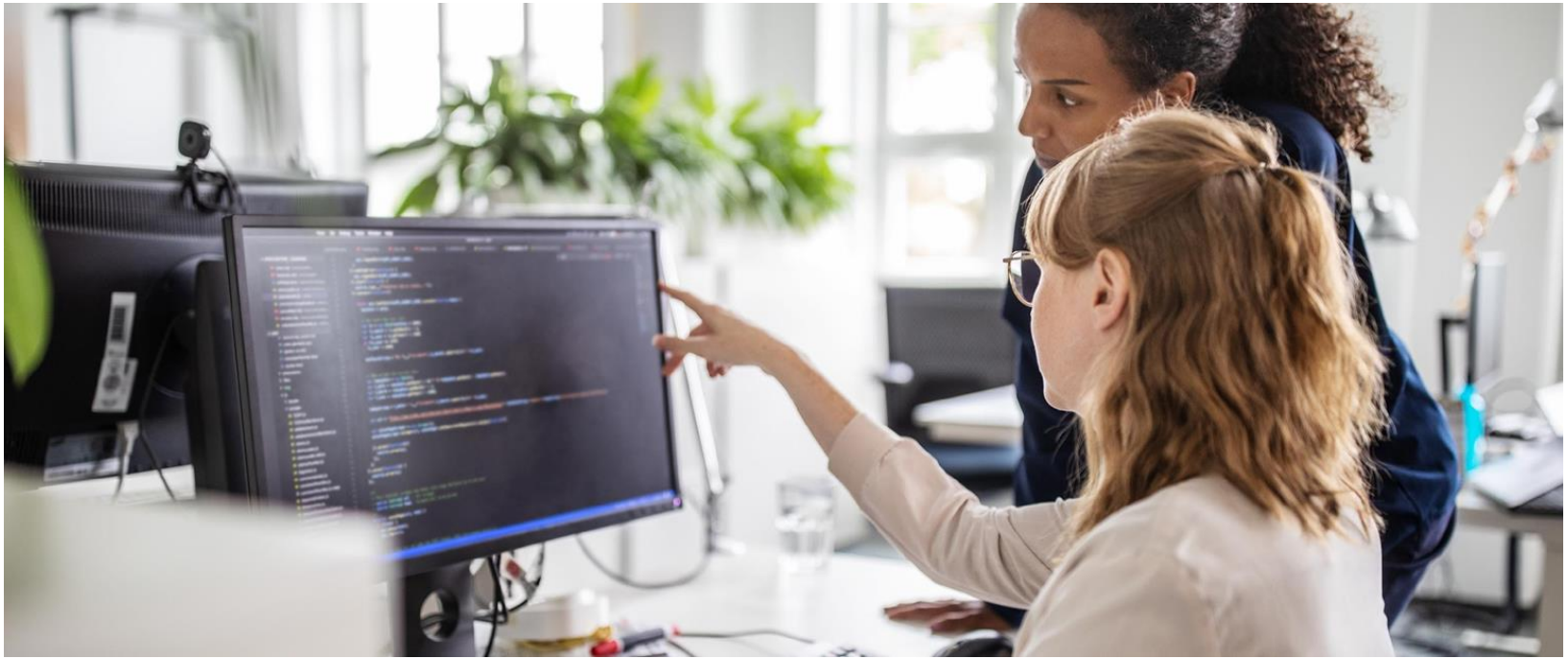
Demonstration Scenario

We'll walk through how to use BigQuery for generating queries.

We will use sample data to run queries, generate reports, and then visualize that data.

Google Cloud Associate Data Practitioner

Demonstration



Visualize data and create dashboards in Looker

Demonstration

Visualize data and create dashboards in Looker

Demonstration Scenario

Key Looker Features for Visualization and Dashboards

- **LookML:** A powerful modeling language for defining data relationships and metrics.
- **Explore:** An intuitive interface for building queries and creating visualizations.
- **Dashboards:** Interactive dashboards with filtering and customization options.
- **Visualizations:** A wide range of visualization types to suit different data and needs.
- **Scheduling and Alerts:** Automate the delivery of dashboards and reports.

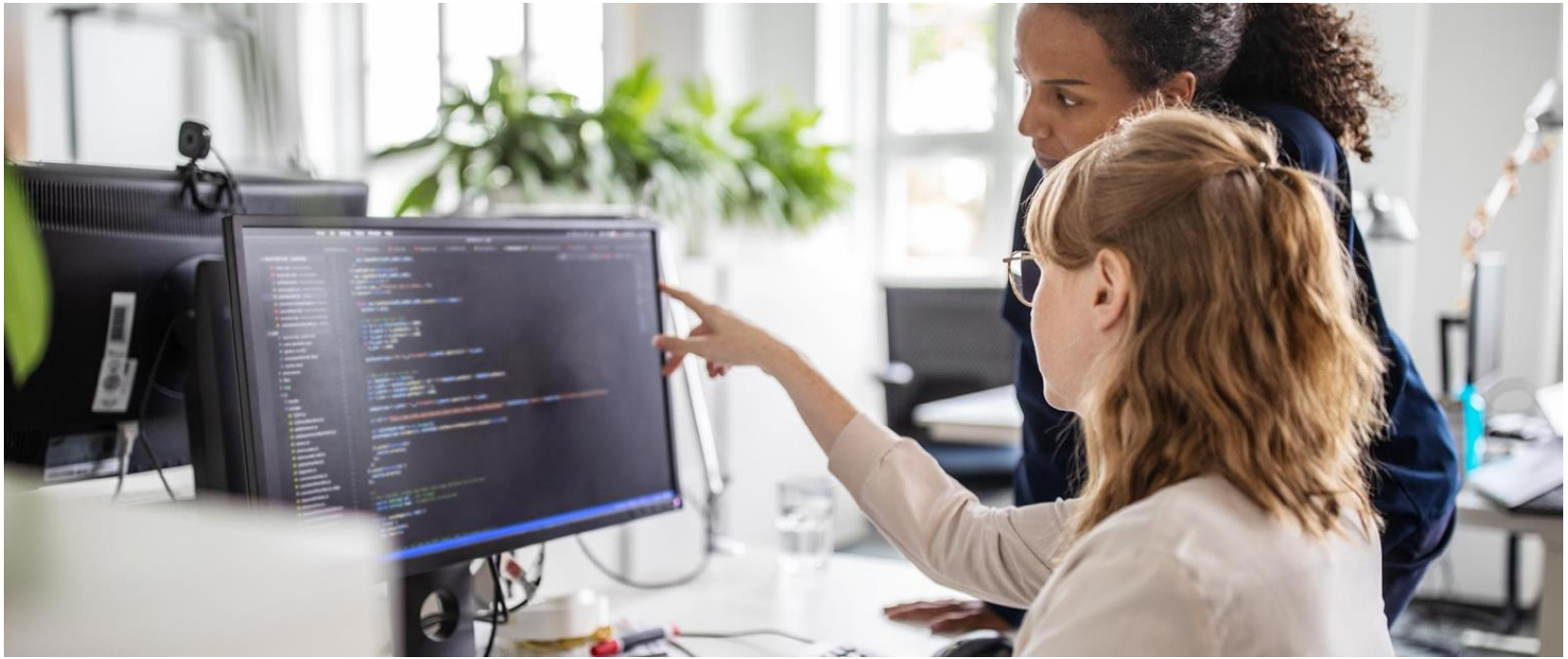
Visualize data and create dashboards in Looker

Demonstration Scenario

We'll discuss Looker Studio and walk thru some options for layouts.

Google Cloud Associate Data Practitioner

Demonstration



Compare Looker and Looker Studio

Understanding Use Cases and Features

Compare Looker and Looker Studio

Analytics Use Cases

Looker Use Cases

Looker excels in enterprise-level data modeling, governance, and deep analytics. It utilizes LookML, a powerful modeling language, to create a consistent and reliable data foundation.

- **Complex Data Modeling:** Ideal for businesses with intricate data relationships and the need for standardized metrics across the organization.
- **Embedded Analytics:** Strong for embedding analytics into applications and workflows, enabling data-driven decision-making within existing systems.
- **Enterprise-Wide Reporting:** Suitable for large organizations that require consistent, governed reporting across multiple departments.
- **Advanced Analytics:** Supports advanced analytics, including machine learning integrations and predictive modeling.
- **Data Governance:** Looker excels in data governance, providing strict control over data definitions and access.

Compare Looker and Looker Studio

Analytics Use Cases

Looker Studio Use Cases

Looker Studio is designed for user-friendly data visualization and reporting. It's accessible to a broader audience, including those without deep technical skills.

- **Marketing Reporting:** Excellent for creating marketing dashboards and reports from Google Analytics, Google Ads, and other marketing platforms.
- **Quick Data Visualization:** Ideal for quickly visualizing data from various sources without requiring complex modeling.
- **Sharing Reports:** Easy sharing of interactive reports with colleagues and clients.
- **Simple data blending:** Looker Studio enables the blending of data from multiple sources for easier reporting.

Compare Looker and Looker Studio

Analytics Use Cases

Comparing Looker to Looker Studio

Feature	Looker	Looker Studio
Data Modeling	LookML: Powerful, complex modeling language for enterprise-grade data governance.	Drag-and-drop interface: Simpler data manipulation and blending.
Complexity	High: Requires technical expertise, especially with LookML	Low: User-friendly, designed for ease of use.
Data Governance	Robust: Strong governance and security features.	Basic: Data source permissions, but less comprehensive than Looker.
Advanced Analytics	Strong: Supports machine learning integrations, predictive modeling, and complex calculations.	Limited: Primarily focused on data visualization and basic reporting.
Target Audience	Data analysts, data engineers, business intelligence professionals.	Marketing professionals, business users, anyone needing quick data insights.

Define, Train, Evaluate and use ML Models

Important Concepts

Define, Train, Evaluate and use ML Models

Important Concepts

What is Vertex AI

Vertex AI is Google Cloud's unified machine learning (ML) platform. Essentially, it's designed to streamline the entire ML workflow, from building and training models to deploying and managing them

- Unified Platform where various services are in a single environment.
- End-to-end ML workflow includes data preparation, model training, model evaluation, model deployment, and model monitoring.
- Generative AI Support, Scalability and MLOps Capabilities (Automation workflows)

Define, Train, Evaluate and use ML Models

Important Concepts

What is Vertex AI (Continued)

Tools and Services:
It provides a range of tools and services, including:

- AutoML: For automated model training.
- Custom training: For building custom models.
- Vertex AI Pipelines: For orchestrating ML workflows.
- Vertex AI Model Registry: For managing model versions.

Define, Train, Evaluate and use ML Models

Important Concepts

Defining ML Models

1. Understanding the Problem

- Defining the business objective
- Identifying the data required
- Choosing the appropriate ML task (classification, regression, etc.)

2. Feature Engineering:

- What are features?
- Importance of feature selection and transformation.

3. Model Selection:

- Overview of common ML algorithms (linear regression, decision trees, neural networks).
- Explain how to select the appropriate model for the problem.

Define, Train, Evaluate and use ML Models

Important Concepts

Training Models in Vertex AI

Data Preparation

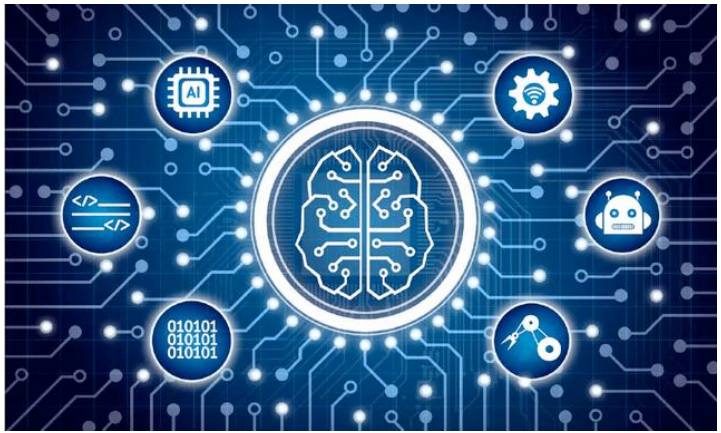
- Data storage in Cloud Storage or BigQuery.
- Data preprocessing steps.

Vertex AI Training Service:

- Custom training vs. pre-built containers.
- Distributed training options.
- Explain the training pipeline.

Hyperparameter Tuning:

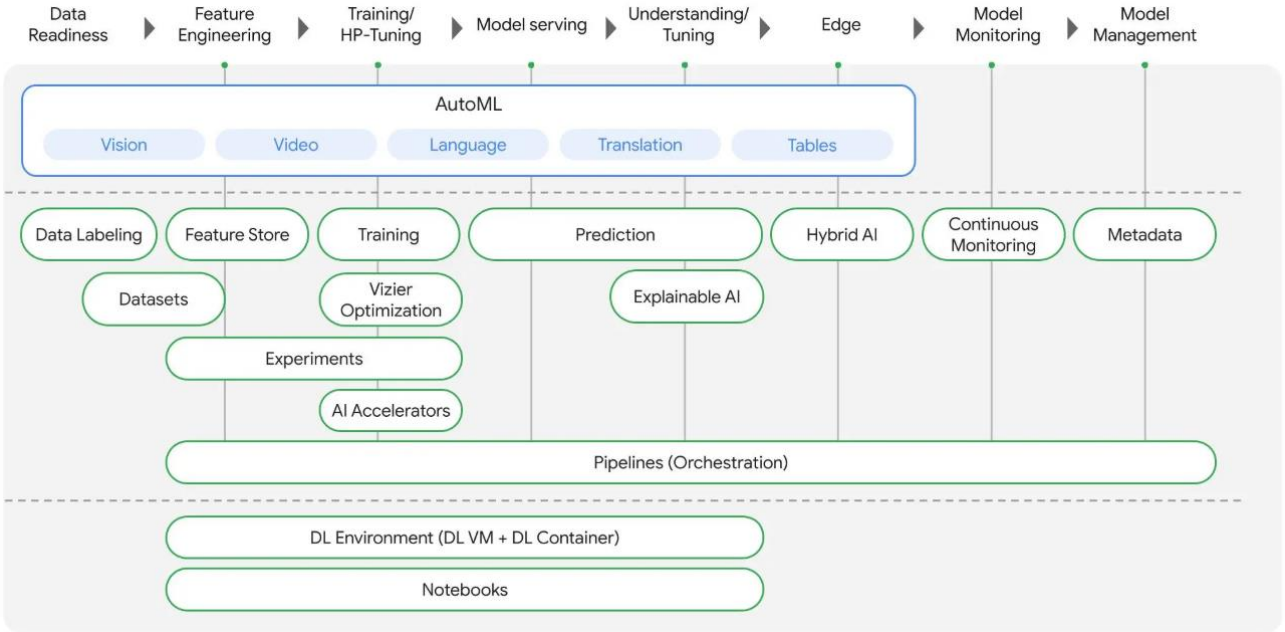
- What are hyperparameters?
- How to use Vertex AI Hyperparameter Tuning.



Define, Train, Evaluate and use ML Models

Important Concepts

Vertex AI Components and Features



Define, Train, Evaluate and use ML Models

Important Concepts

Evaluating Models (Vertex AI)

Importance of Model Evaluation:

- Preventing overfitting and underfitting.
- Measuring model performance.

Evaluation Metrics:

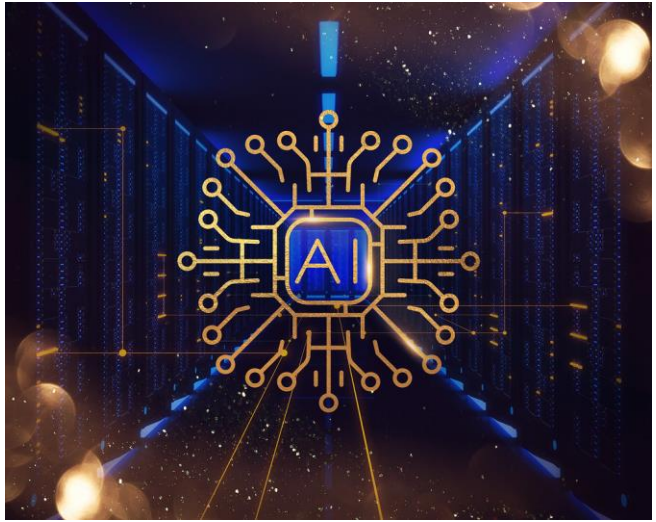
- Classification:
- Accuracy, precision, recall, F1-score, AUC.

Regression:

- Mean squared error, mean absolute error.

Vertex AI Model Evaluation:

- Explain how to use the model evaluation service.
- Confusion matrices and ROC curves.



Define, Train, Evaluate and use ML Models

Important Concepts

Using Deployed ML Models (Vertex AI)

Preventing Vertex AI Model Deployment:

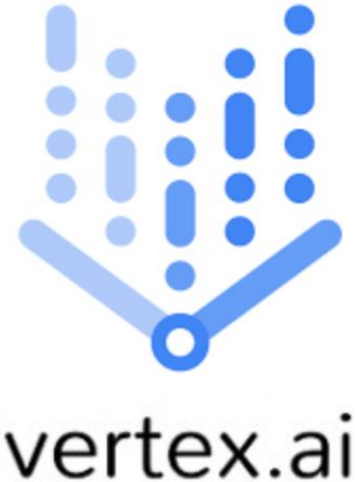
- Deploying models to endpoints.
- Online vs. batch predictions.

Making Predictions:

- Using the Vertex AI Prediction service.
- API requests for online predictions.
- Batch prediction jobs.

Model Monitoring:

- Explain the importance of model monitoring.
- Drift detection.



Plan an actual ML Project

Whiteboard

Plan an Actual ML Project

Whiteboard Scenario

We'll walk through a project scenario and then discuss what AI/ML services could be used to meet requirements.

Google Cloud Associate Data Practitioner

Whiteboard Discussion

Module Review

Summary Recap

Module Review

Summary

Let’s summarize some of the important topics we covered in this module.

- A notebook provides an environment in which to author and execute code. A notebook is essentially a source artifact, saved as an IPYNB file.
- Structurally, a notebook is a sequence of cells. A cell is a block of input text that is evaluated to produce results.
- Code cells contain code to evaluate. The output or results of executed code are rendered in line with the executed code.
- Markdown cells contain Markdown text that is converted to HTML to produce headers, lists, and formatted text.
- Raw cells can be used to render different code formats into HTML or LaTeX.
- Python Pandas are used for data manipulation and analysis in the Jupyter Notebook Environment.

Module Review

Summary

Let’s summarize some of the important topics we covered in this module.

- BigQuery primarily uses Standard SQL, which is compliant with the SQL 2011 standard and ensures greater compatibility with other SQL dialects.
- Looker is a powerful business intelligence and data visualization platform that allows you to create interactive dashboards and explore data
- Looker Dashboards has a drilldown feature that allows users to explore data more granularly.
- LookML is Looker's modeling language based on SQL. Define dimensions, measures, and relationships between tables.
- Looker is a comprehensive BI platform for deep data analysis and governance, while Looker Studio is a streamlined tool for creating and sharing visual reports.
- Vertex AI is Google Cloud's unified machine learning (ML) platform and it's designed to streamline the entire ML workflow, from building and training models to deploying and managing them

Data Management

Using Tools in GCP to meet compliance, privacy and security requirements

Module Overview

What we will cover in the Module

- Demonstration: IAM
- Demonstration: Access control for Cloud Storage
- Demonstration: Analytics Hub
- Demonstration: Lifecycle Management
- Whiteboard: Replication and High Availability
- Encryption Fundamentals
- Module Overview

IAM

Demonstration

IAM

Important Concepts

Identity and Access Management (IAM)

IAM stands for Identity and Access Management. In Google Cloud, IAM is a crucial service that enables you to control who (identity) has what level of access (role) to which resources. It's a fundamental aspect of security and access control within GCP

- Establishing Least Privileged Access with IAM in Google Cloud
- Implementing the principle of least privilege in Google Cloud using Identity and Access Management (IAM) involves granting users and service accounts only the necessary permissions to perform their tasks
- Three Key components are Member, Role and Policy
- Three types of roles in GCP IAM – Basic, Predefined and Custom

IAM

Important Concepts

Identity and Access Management (IAM) Policies

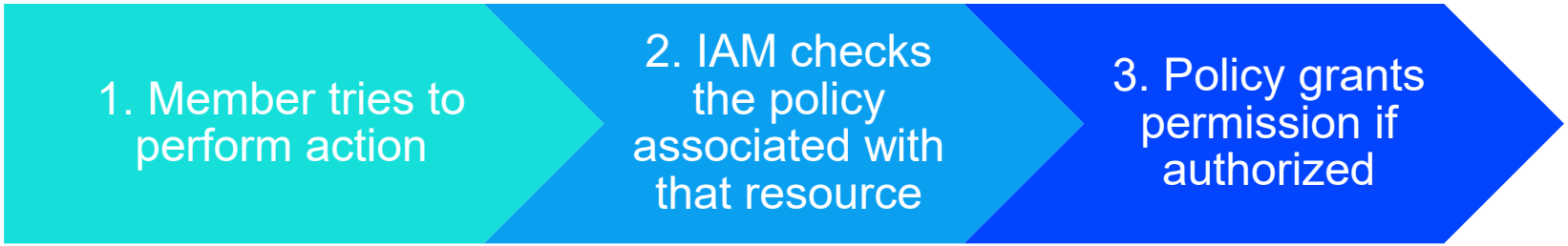
IAM policies can be applied at various levels in the Google Cloud resource hierarchy:

- **Organization**: The top level, representing your company or organization.
- **Folder:** A way to group projects and other folders.
- **Project:** The core organizational unit in GCP.
- **Resource:** Individual resources within a project

IAM

Workflow of IAM

IAM Workflow



IAM

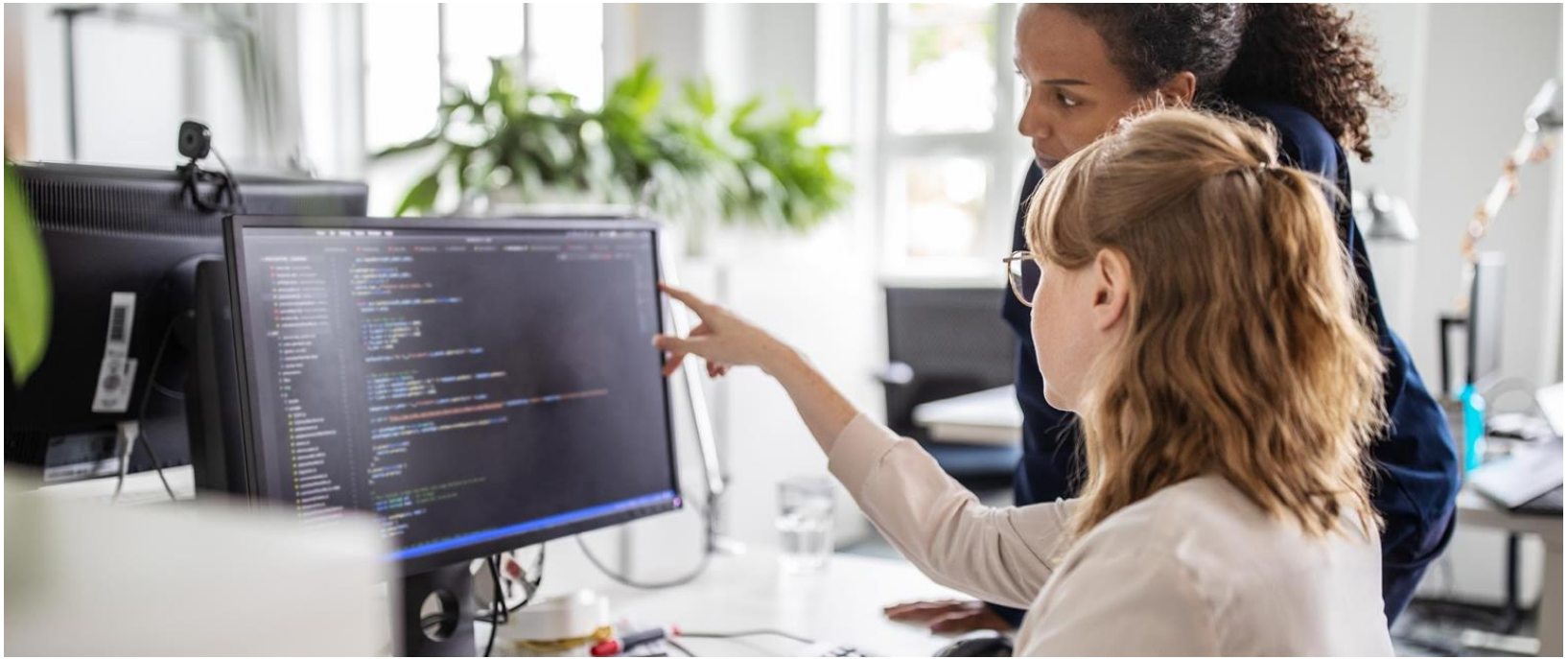
Demonstration Scenario

We'll walk thru Identity and Access Management (IAM) to secure our services.

We will walk through permissions, roles, etc, related to this certification requirements.

Google Cloud Associate Data Practitioner

Demonstration



Compare methods of access controls (Cloud Storage)

Important Concepts and Demonstration

Compare Methods of access controls

Important Concepts

Google Cloud Storage Access

Google Cloud Storage offers access control methods to secure data and manage who can access buckets and objects.

Public Access - Allows anyone online to access your Cloud Storage data.

- Achieved by granting permissions to the allUsers or allAuthenticatedUsers groups.

Use Cases

- Distributing publicly available content, such as website assets or software downloads.
- Be extremely cautious when using public access due to major security risks.

Risks can be major security risks if sensitive data is exposed, and also cost issues for excessive downloads.

Compare Methods of access controls

Important Concepts

Google Cloud Storage Access (Continued)

Public Access - Allows anyone online to access your Cloud Storage data.

- Restricts access to authorized users and service accounts only.
- Controlled through Identity and Access Management (IAM) permissions.

Use Cases:

- Storing sensitive data, such as customer information, financial records, or confidential documents.
- Most general use cases will utilise private access.

Benefits

- Enhanced security and data protection.
- Granular control over who can access your data. .

Compare Methods of access controls

Important Concepts

Google Cloud Storage Access (Continued)

IAM (Identity and Access Management)

- The primary method for controlling access to Google Cloud resources, including Cloud Storage.
- Allows you to grant roles (collections of permissions) to users and service accounts.Offers fine-grained control at the bucket and project levels.

Use Cases

- Managing access for internal users, applications, and service accounts.
- Enforcing the principle of least privilege by granting only necessary permissions.

Benefits

- Centralized access control management, Flexibility, and scalability.
- Integration with other Google Cloud services..

Compare Methods of access controls

Important Concepts

Google Cloud Storage Access (Continued)

Uniform Bucket-Level Access in Cloud Storage:

- Disables Access Control Lists (ACLs) and relies exclusively on IAM permissions.
- Simplifies access control management by applying IAM permissions consistently at the bucket level.
- This is now the recommended method of access control.

Use Cases

- Organizations that prefer a centralized and consistent access control model.
- Simplifying access management in complex environments.
- Benefits are simplified access control management, consistent application of IAM policies as well as enhanced security.

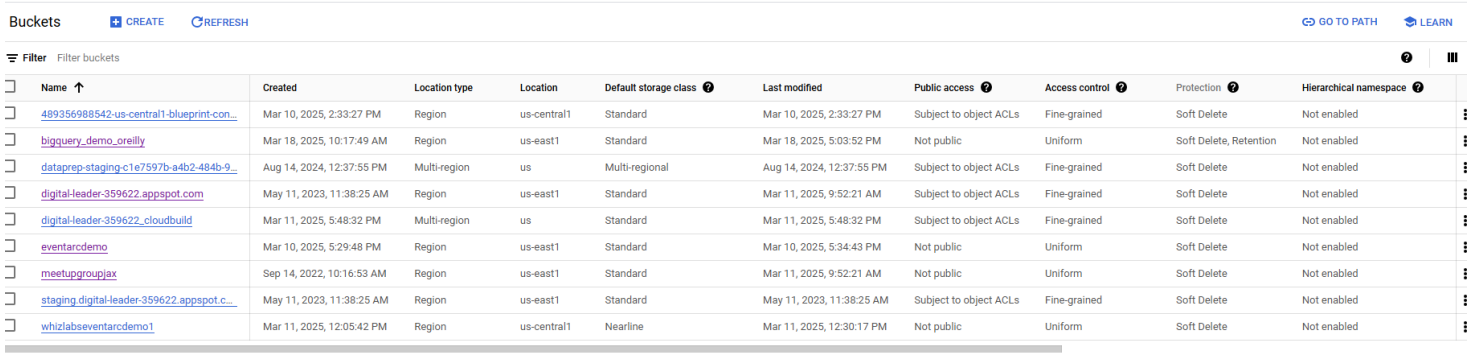
Compare Methods of access controls

Important Concepts

Google Cloud Storage Access (Continued)

IAM vs. ACLs

- IAM is the recommended method for access control due to its flexibility and scalability.
- ACLs are older and offer object-level permissions, but they can be more complex to manage.
- Uniform bucket-level access allows you to use exclusively IAM.



Name	Created	Location type	Location	Default storage class	Last modified	Public access	Access control	Protection	Hierarchical namespace
489356988542-us-central1-blueprint-con...	Mar 10, 2025, 2:33:27 PM	Region	us-central1	Standard	Mar 10, 2025, 2:33:27 PM	Subject to object ACLs	Fine-grained	Soft Delete	Not enabled
bigquery_demo_oreilly	Mar 18, 2025, 10:17:49 AM	Region	us-east1	Standard	Mar 18, 2025, 5:03:52 PM	Not public	Uniform	Soft Delete, Retention	Not enabled
dataprep-staging-c1e7597b-a4b2-484b-9...	Aug 14, 2024, 12:37:55 PM	Multi-region	us	Multi-regional	Aug 14, 2024, 12:37:55 PM	Subject to object ACLs	Fine-grained	Soft Delete	Not enabled
digital-leader-359622.appspot.com	May 11, 2023, 11:38:25 AM	Region	us-east1	Standard	Mar 11, 2025, 9:52:21 AM	Subject to object ACLs	Fine-grained	Soft Delete	Not enabled
digital-leader-359622_cloudbuild	Mar 11, 2025, 5:48:32 PM	Multi-region	us	Standard	Mar 11, 2025, 5:48:32 PM	Subject to object ACLs	Fine-grained	Soft Delete	Not enabled
eventarcdemo	Mar 10, 2025, 5:29:48 PM	Region	us-east1	Standard	Mar 10, 2025, 5:34:43 PM	Not public	Uniform	Soft Delete	Not enabled
meetupgroupjax	Sep 14, 2022, 10:16:53 AM	Region	us-east1	Standard	Mar 11, 2025, 9:52:21 AM	Not public	Uniform	Soft Delete	Not enabled
staging.digital-leader-359622.appspot.c...	May 11, 2023, 11:38:25 AM	Region	us-east1	Standard	May 11, 2023, 11:38:25 AM	Subject to object ACLs	Fine-grained	Soft Delete	Not enabled
whizlabseventarcdemo1	Mar 11, 2025, 12:05:42 PM	Region	us-central1	Nearline	Mar 11, 2025, 12:30:17 PM	Not public	Uniform	Soft Delete	Not enabled

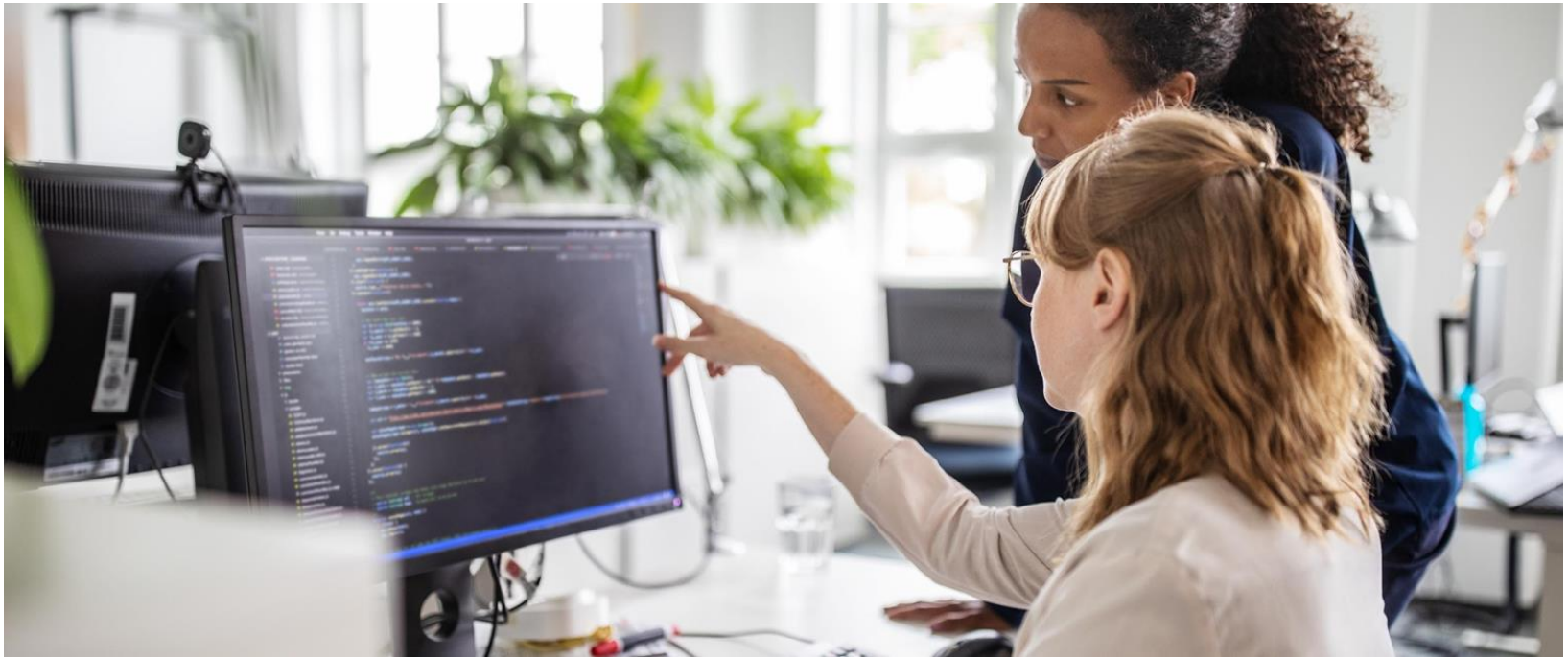
Compare methods of access controls

Demonstration Scenario

We'll walk thru Cloud Storage permissions, specially IAM and bucket permissions

Google Cloud Associate Data Practitioner

Demonstration



Analytics Hub

Demonstration

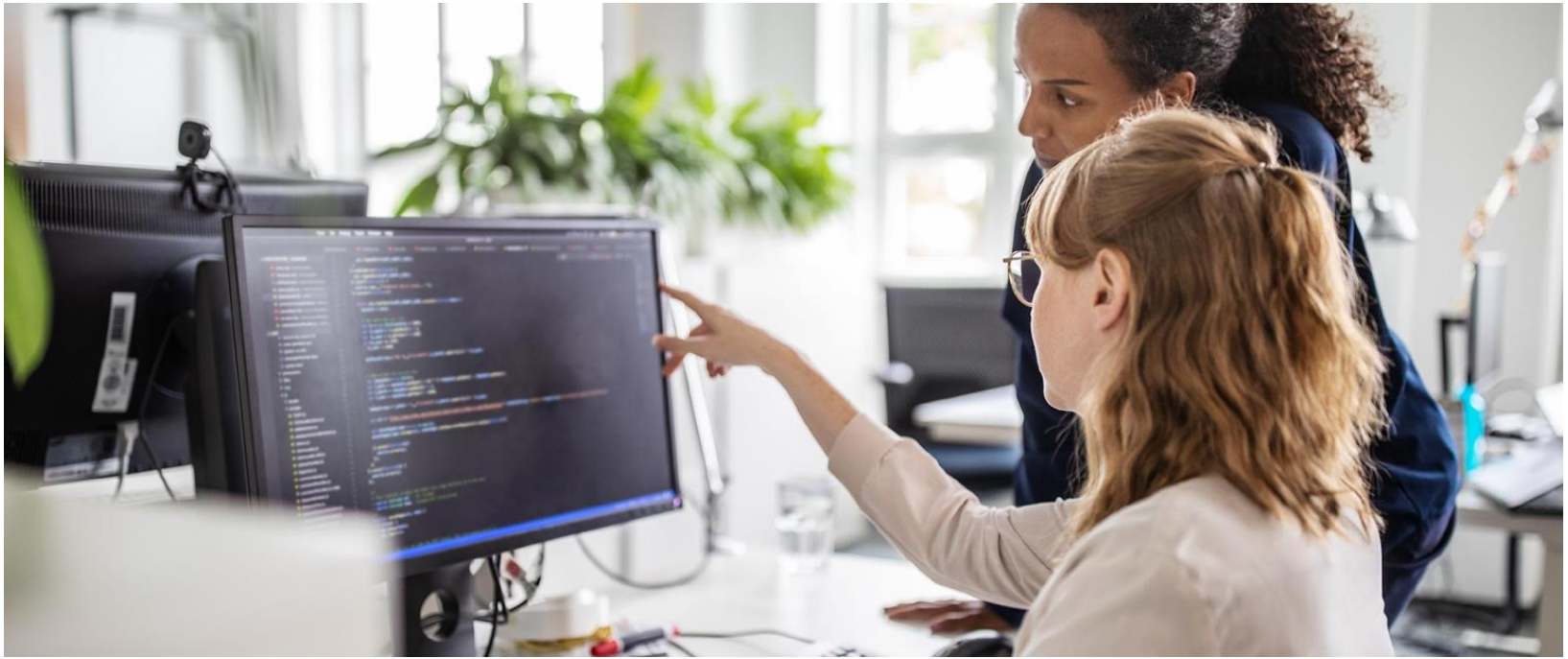
Analytics Hub

Demonstration Scenario

We'll walk thru Analytics Hub's main use case and features.

Google Cloud Associate Data Practitioner

Demonstration



Lifecycle Management – Cloud Storage

Demonstration

Lifecycle Management Cloud Storage

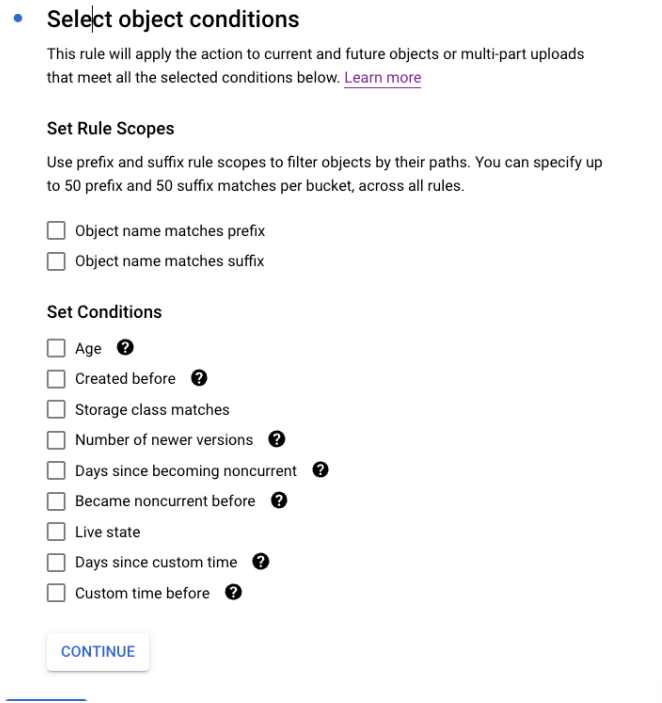
Important Concepts

Cloud Storage Lifecycles

You define lifecycle management rules in an XML or JSON document that is attached to a bucket.

Each rule consists of:

- Action: The action to take on objects that meet the conditions (e.g., Delete, SetStorageClass).
- Conditions: The criteria that objects must meet for the action to be triggered (e.g., Age, CreatedBefore, NumberOfNewerVersions, MatchesStorageClass).



Lifecycle Management Cloud Storage

Important Concepts

Cloud Storage Lifecycle Conditions

Actions specify what should happen to objects that meet the defined conditions. Google Cloud Storage provides several actions:

- **Delete:** Permanently removes objects from the bucket.
- **Set Storage Class:** Changes the storage class of objects (e.g., move objects to Nearline or Coldline).
- **Set Event-Based Hold:** Places objects under an event-based hold to prevent deletion until the hold is removed.
- **Set Temporary Hold:** Places objects under a temporary hold, which can be used for legal or compliance purposes.

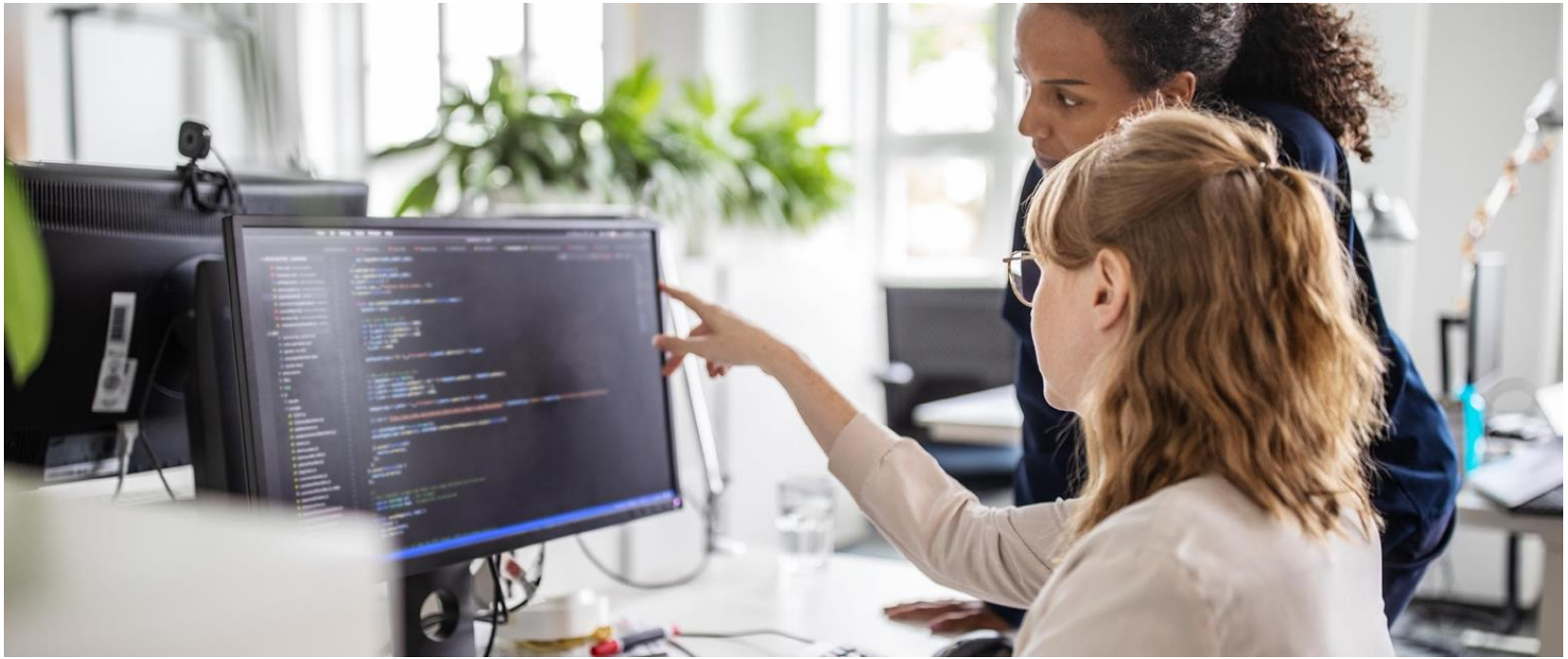
Lifecycle Management –Cloud Storage

Demonstration Scenario

We’ll walk thru Cloud Storage lifecycle management configuration steps.

Google Cloud Associate Data Practitioner

Demonstration



Replication and HA

Whiteboard

Cloud Key Management (Cloud KMS)

Whiteboard Scenario

We'll walk through how replication and high availability work with cloud storage and other services.

We will break down regions and zone concepts.

Google Cloud Associate Data Practitioner

Whiteboard Discussion

Encryption Fundamentals

Options in GCP

Encryption Fundamentals

Data Security Fundamentals

Data Encryption

Encryption is the process of encoding data using cryptographic algorithms, the product of which is encrypted data (referred to as cipher-text).

Only the intended recipient or system that is in possession of the correct key can decode (un-encrypt) the cipher-text.

Encryption Fundamentals

Data Security Fundamentals

Options in GCP

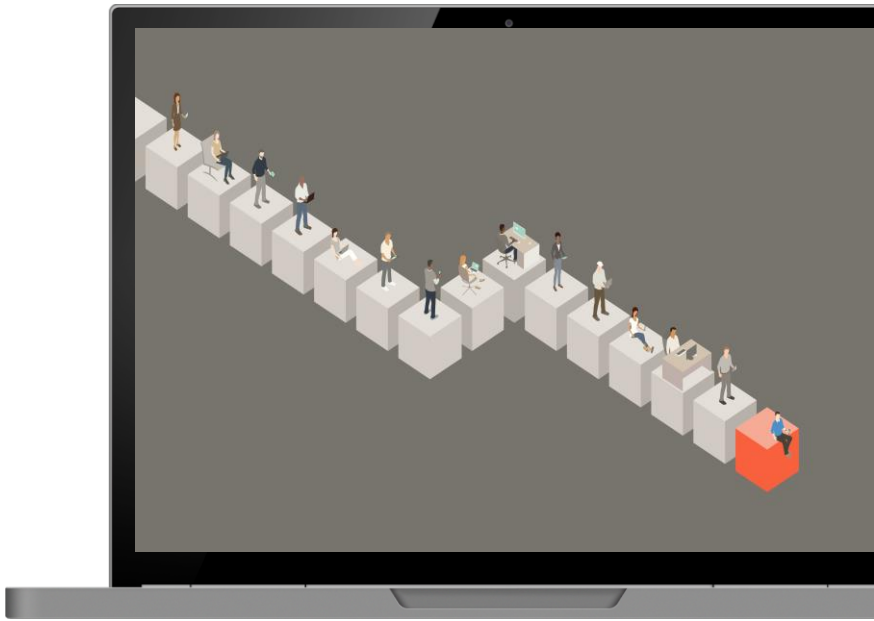
- **Data at Rest -** Data that is not actively being used or transmitted.
- **Data in Transit —** Data that is currently being transported over a median, such as a network **cable,** from one location to another.
- **Data in Use –** Data that is currently being used by an application and or an application user.
- **De-Identification -** Is the process of removing fields that could identify a user in a dataset.

Encryption Fundamentals

Important Concepts

Options in GCP

- **Data at Rest:** GCP encrypts all customer content at rest by default using one or more encryption mechanisms, including AES-256. This applies to data stored in various services like Cloud Storage, Compute Engine persistent disks, and databases.
- **Data in Transit:** GCP encrypts data in transit between its data centers and between your applications and GCP services using TLS.



Encryption Fundamentals

Important Concepts

Options in GCP

Customer-Managed Encryption Keys (CMEK)

- Cloud Key Management Service (KMS): KMS allows you to create, manage, and use your own encryption keys to protect your data.
- You can generate, rotate, and destroy keys, and grant granular access control to them.
- Integration with other services: CMEK can be used with various GCP services, such as Cloud Storage, Compute Engine, and BigQuery, to encrypt data with your own keys.

Encryption Fundamentals

Important Concepts

Options in GCP

Customer-Supplied Encryption Keys (CSEK)

- **Cloud Storage:** You can provide your own AES-256 keys to encrypt data stored in Cloud Storage.
- **Bring Your Own Key (BYOK):** For even greater control, you can use your encryption keys generated and managed outside of GCP.

Encryption Fundamentals

Important Concepts

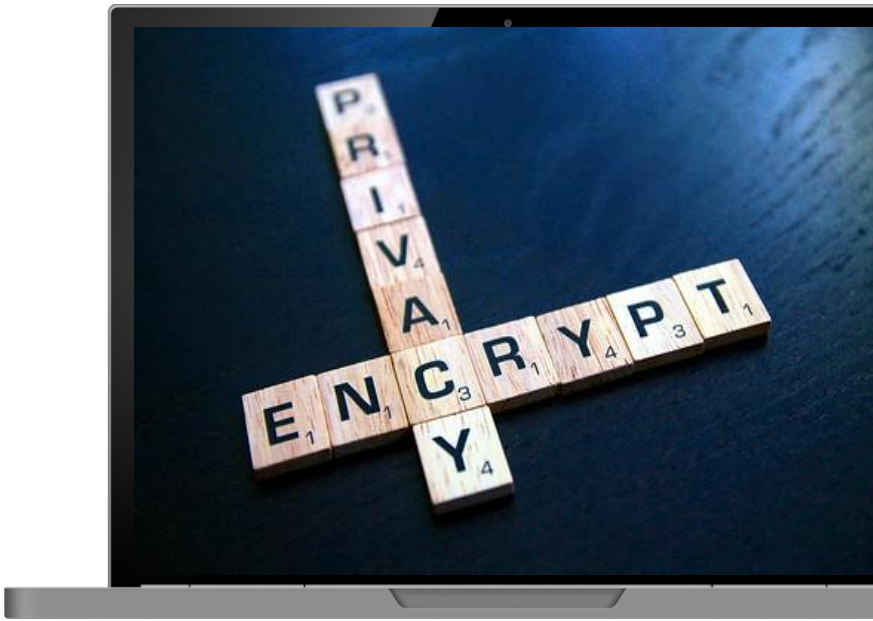
Options in GCP

Client-Side Encryption: Encrypt before uploading:

- You can encrypt your data on your own premises before uploading it to GCP.

Application-Layer Encryption: Encrypt within your application

- You can encrypt specific data fields or files within your application before storing them in GCP services. This allows for fine-grained control over encryption.



Encryption Fundamentals

Important Concepts

Options in GCP

Cloud Hardware Security Module (HSM): For highly sensitive data, you can use Cloud HSM, a dedicated HSM service that provides FIPS 140-2 Level 3 certified hardware for secure key management and cryptographic operations.

- Integrates with Cloud KMS
- Provides a vault for your keys

Module Review

Summary Recap

Module Review

Summary

Let’s summarize some of the important topics we covered in this module.

- IAM is a crucial service that enables you to control who (identity) has what level of access (role) to which resources. It's a fundamental aspect of security and access control within GCP.
- IAM policies can be applied at various levels in the Google Cloud resource hierarchy. (Organization, Folder, Project and Resource)
- Google Cloud Storage offers access control methods to secure data and manage who can access buckets and objects.
- Define Cloud Storage lifecycle management rules in an XML or JSON document attached to a bucket.
- The Cloud Storage lifecycle rule consists of two parts: the Action and the Condition that triggers it.

Module Review

Summary

Let’s summarize some of the important topics we covered in this module.

- In Google Cloud, "regions" and "zones" are fundamental concepts that define the geographical locations of your cloud resources.
- Google Cloud's Analytics Hub is essentially a data exchange platform designed to simplify and secure the sharing of data assets.
- Encryption is the process of encoding data using cryptographic algorithms, resulting in encrypted data (referred to as ciphertext).
- With Client-Side Encryption, you can encrypt your data on your premises before uploading it to GCP.
- Use Cloud HSM for highly sensitive data; you can use Cloud HSM, a dedicated HSM service that provides FIPS 140-2 Level 3

Course Closeout

Thank you for joining the course.


O'REILLY®