Linux Academy | Cloud Assessments

The Data Dossier

Choose a Section

Return to Table of Contents

Choose a Lesson

What is a Data Engineer?

Exam and Course Overview

What is a Data Engineer?

Google's definition:

A Professional Data Engineer enables data-driven decision making by collecting, transforming, and visualizing data. The Data Engineer designs, builds, maintains, and troubleshoots data processing systems with a particular emphasis on the security, reliability, fault-tolerance, scalability, fidelity, and efficiency of such systems.

The Data Engineer also analyzes data to gain insight into business outcomes, builds statistical models to support decision-making, and creates machine learning models to automate and simplify key business processes.

What does this include?

- **Build data structures and databases:**
  - **Cloud SQL, Bigtable**
- **Design data processing systems:**
  - **Dataproc, Pub/Sub, Dataflow**
- **Analyze data and enable machine learning:**
  - **BigQuery, Tensorflow, Cloud ML Engine, ML API's**
- **Match business requirements with best practices**
- **Visualize data ("make it look pretty"):**
  - **Data Studio**
- **Make it secure and reliable**

Super-simple definition:

Collect, store, manage, transform, and present data to make it useful.

Return to Table of Contents

Choose a Lesson

Data Lifecycle

Batch and Streaming Data

Cloud Storage as Staging Ground

Database Types

Monitoring Unmanaged Databases

Data Lifecycle

Next

- Think of data as a tangible object to be collected, stored, and processed
- Lifecycle from initial collection to final visualization
- Needs to be familiar with the lifecycle steps, what GCP services are associated with each step, and how they connect together
- Data Lifecycle steps:
  - Ingest - Pull in the raw data:
    - Streaming/real-time data from devices
    - On-premises batch data
    - Application logs
    - Mobile-app user events and analytics
  - Store - data needs to be stored in a format and location that is both reliable and accessible
  - Process and analyze - Where the magic happens. Transform data from raw format to actionable information
  - Explore and visualize - "Make it look pretty"
    - The final stage is to convert the results of the analysis into a format that is easy to draw insights from and to share with colleagues and peers

Return to Table of Contents

Choose a Lesson

Choosing a Managed Database

Next

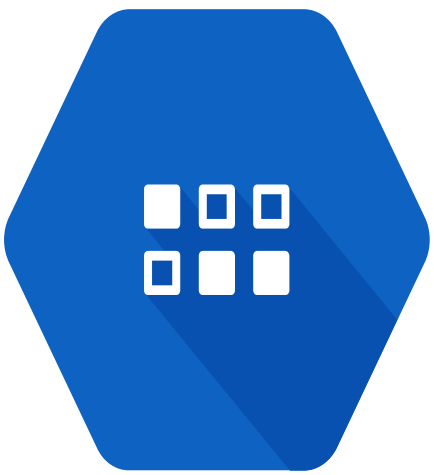
Big picture perspective:

- At minimum, know which managed database is the best solution for any given use case:
  - Relational, non-relational?
  - Transactional, analytics?
  - Scalability?
  - Lift and shift?

	Relational		Non-relational		Object - Unstructured	Data Warehouse
	Cloud SQL	Cloud Spanner	Cloud Datastore	Cloud Bigtable	Cloud Storage	BigQuery
Use Case	Structured data Web framework	RDBMS+scale High transactions	Semi-structured Key-value data	High throughput analytics	Unstructured data Holds everything	Mission critical apps Scale+consistency
e.g.	Medical records Blogs	Global supply chain Retail	Product catalog Game state	Graphs IoT Finance	Multimedia Analytics Disaster recovery	Large data analytics Processing using SQL

Return to Table of Contents

Choose a Lesson



Cloud Datastore

Cloud Datastore Overview

Next

What is Cloud Datastore?

- No Ops:
  - No provisioning of instances, compute, storage, etc.
  - Compute layer is abstracted away
- Highly scalable:
  - Multi-region access available
  - Sharding/replication handled automatically
- NoSQL/non-relational database:
  - Flexible structure/relationship between objects

Use Datastore for:

- Applications that need highly available structured data, at scale
- Product catalogs - real-time inventory
- User profiles - mobile apps
- Game save states
- ACID transactions - e.g., transferring funds between accounts

Do not use Datastore for:

- Analytics (full SQL semantics):
  - Use BigQuery/Cloud Spanner
- Extreme scale (10M+ read/writes per second):
  - Use Bigtable
- Don't need ACID transactions/data not highly structured:
  - Use Bigtable
- Lift and shift (existing MySQL):
  - Use Cloud SQL
- Near zero latency (sub-10ms):
  - Use in-memory database (Redis)

Return to Table of Contents

Choose a Lesson

Cloud Bigtable Overview

Instance Configuration

Data Organization

Schema Design

Cloud Bigtable

Cloud Bigtable Overview

Next

What is Cloud Bigtable?

- High performance, massively scalable NoSQL database
- Ideal for large analytical workloads

History of Bigtable

- Considered one of the originators for a NoSQL industry
- Developed by Google in 2004
  - Existing database solutions were too slow
  - Needed real-time access to petabytes of data
- Powers Gmail, YouTube, Google Maps, and others

What is it used for?

- High throughput analytics
- Huge datasets

Use Cases

- Financial data – stock prices
- IoT data
- Marketing data – purchase histories

Access Control

- Project wide or instance level
- Read/Write/Manage

Return to Table of Contents

Choose a Lesson

Cloud Spanner Overview

Data Organization and Schema

Cloud Spanner Overview

Next

What is Cloud Spanner?

- Fully managed, highly scalable/available, relational database
- Similar architecture to Bigtable
- "NewSQL"

What is it used for?

- Mission critical, relational databases that need strong transactional consistency (ACID compliance)
- Wide scale availability
- Higher workloads than Cloud SQL can support
- Standard SQL format (ANSI 2011)

Horizontal vs. vertical scaling

- Vertical = more compute on single instance (CPU/RAM)
- Horizontal = more instances (nodes) sharing the load

Compared to Cloud SQL

- Cloud SQL = Cloud incarnation of *on-premises* MySQL database
- Spanner = designed from the ground up for the cloud
- Spanner is not a 'drop in' replacement for MySQL
  - Not MySQL/PostreSQL compatible
  - Work required to migrate
  - However, when making transition, don't need to choose between consistency and scalability

Return to Table of Contents

Choose a Lesson

Streaming Data Challenges

Cloud Pub/Sub Overview

Pub/Sub Hands On

Connecting Kafka to GCP

Monitoring Subscriber Health

Streaming Data Challenges

Next

What is Streaming Data?

- "Unbounded" data
- Infinite, never completes, always flowing

Examples

Traffic Sensors

Credit Card Transactions

Mobile Gaming

Fast action is often necessary

- Quickly collect data, gain insights, and take action
- Sending to storage can add latency
- Use cases:
  - Credit card fraud detection
  - Predicting highway traffic

Return to Table of Contents

Choose a Lesson

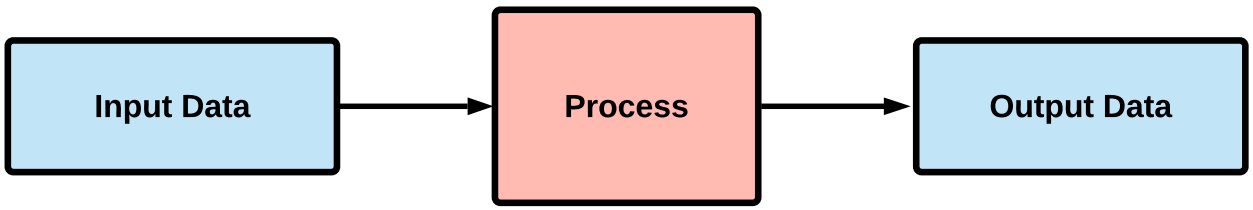
- Data Processing Challenges
- Cloud Dataflow Overview
- Key Concepts
- Template Hands On
- Streaming Ingest Pipeline Hands On
- Additional Best Practices

Data Processing Challenges

Next

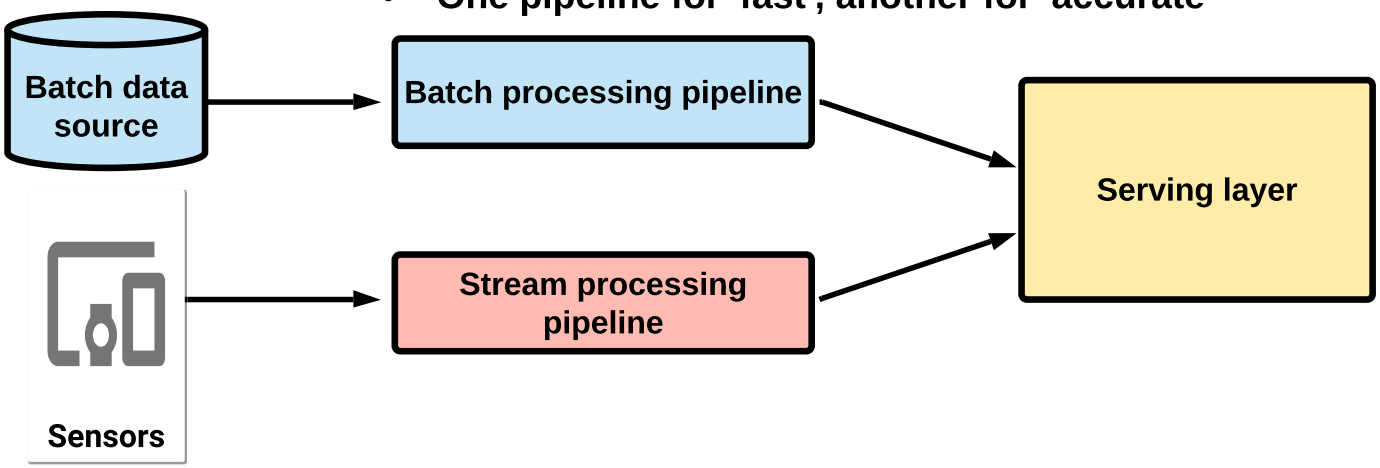
What is Data Processing?

- Read Data (Input)
- Transform it to be relevant - Extract, Transform, and Load (ETL)
- Create output



Challenge: Streaming and Batch data pipelines:

- Until recently, separate pipelines are required for each
- Difficult to compare recent and historical data
- One pipeline for 'fast', another for 'accurate'



Why both?

- Credit card monitoring
- Compare streaming transactions to historical batch data to detect fraud

Return to Table of Contents

Choose a Lesson

- Dataproc Overview
- Configure Dataproc Cluster and Submit Job
- Migrating and Optimizing for Google Cloud
- Best Practices for Cluster Performance

Dataproc Overview

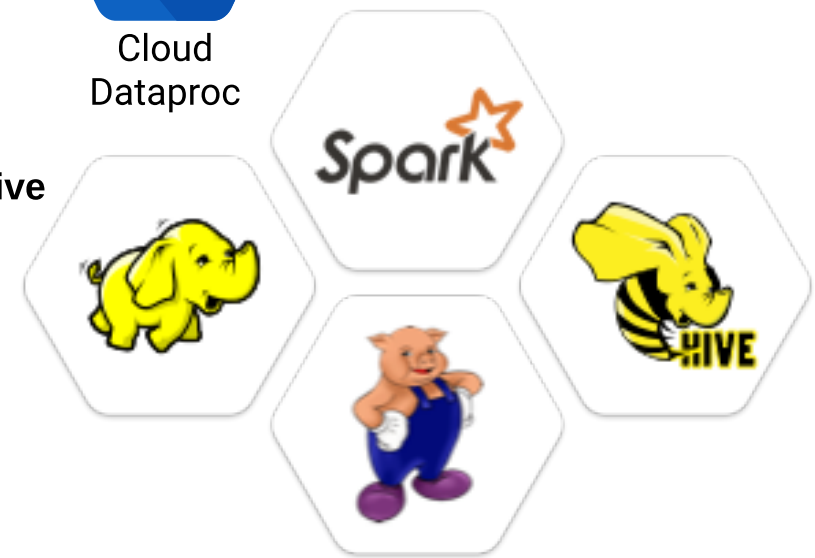
What is Cloud Dataproc?

Next

Cloud Dataproc

Hadoop ecosystem:

- Hadoop, Spark, Pig, Hive
- Lift and shift to GCP



Managed Hadoop/Spark Stack

Custom Code
Monitoring/Health
Dev Integration
Manual Scaling
Job Submission
Google Cloud Connectivity
Deployment
Creation

Dataproc facts:

- On-demand, managed Hadoop and Spark clusters
- Managed, but not no-ops:
  - Must configure cluster, not auto-scaling
  - Greatly reduces administrative overhead
- Integrates with other Google Cloud services:
  - Separate data from the cluster - save costs
- Familiar Hadoop/Spark ecosystem environment:
  - Easy to move existing projects
- Based on Apache Bigtop distribution:
  - Hadoop, Spark, Hive, Pig
- HDFS available (but maybe not optimal)
- Other ecosystem tools can be installed as well via initialization actions

Return to Table of Contents

Choose a Lesson

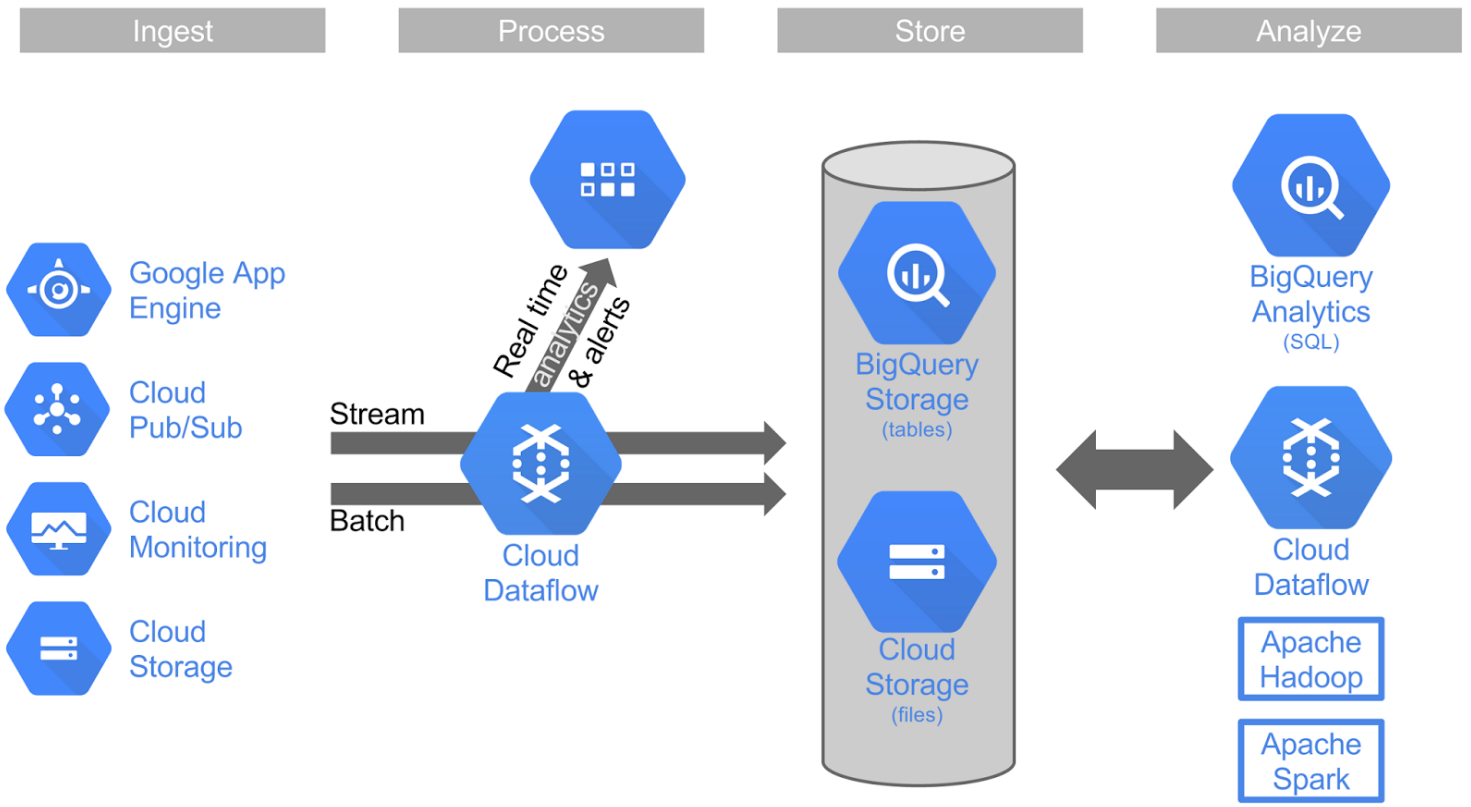
- BigQuery Overview
- Interacting with BigQuery
- Load and Export Data
- Optimize for Performance and Costs
- Streaming Insert Example
- BigQuery Logging and Monitoring
- BigQuery Best Practices

BigQuery Overview

Next

What is BigQuery?

- Fully Managed Data warehousing
  - Near-real time analysis of petabyte scale databases
- Serverless (no-ops)
- Auto-scaling to petabyte range
- Both storage and analysis
- Accepts batch and streaming loads
- Locations = multi-regional (US, EU), Regional (asia-northeast1)
- Replicated, durable
- Interact primarily with standard SQL (also Legacy SQL)
  - SQL Primer course



Return to Table of Contents

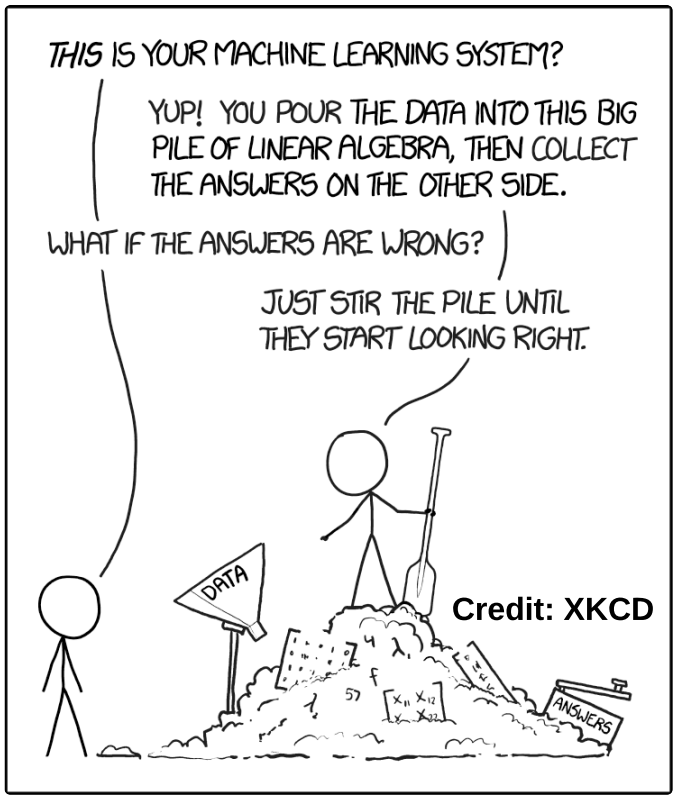
Choose a Lesson

What is Machine Learning?

Next

Popular view of machine learning...

DATA →



Credit: XKCD

→ MAGIC!

For Data Engineer:

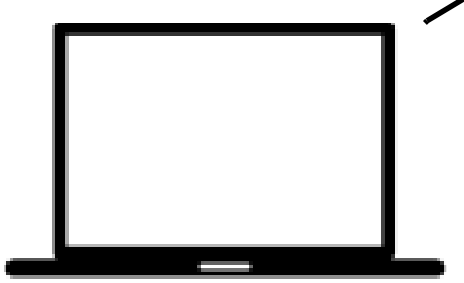
Know the training and inference stages of ML

So what is machine learning?

Process of combining inputs to produce useful predictions on never-before-seen data

Makes a machine learn from data to make predictions on future data, instead of programming every scenario

New, unlabeled image



"I have never seen this image before, but I'm pretty sure that this is a cat!"

Return to Table of Contents

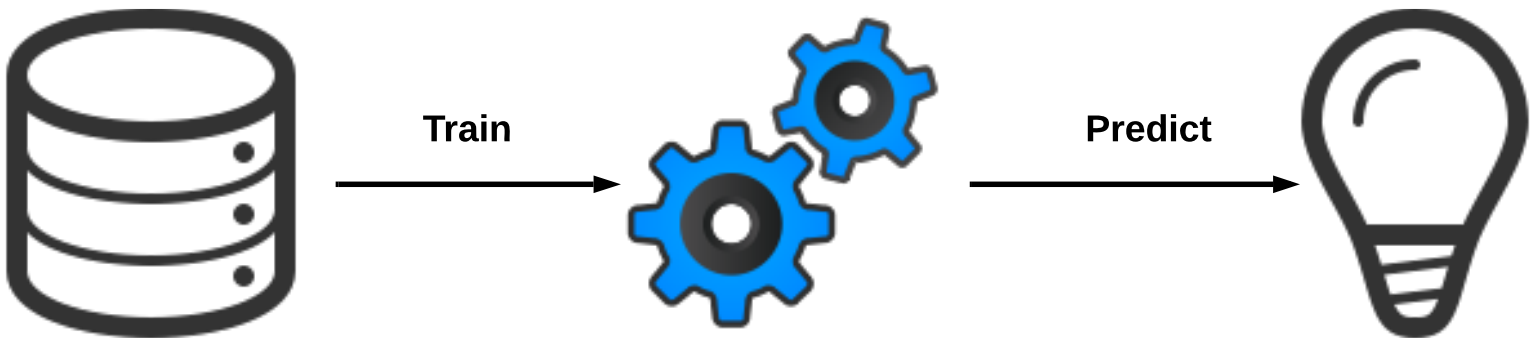
GCP Machine Learning Services

Choose a Lesson
GCP Machine Learning Services
AI Platform Overview
AI Platform Hands On

Next

Machine Learning - In a nutshell

- Algorithm that is able to learn from data



Lots of data
The more the better

ML Algorithm
Find patterns

Insight
Make intelligent decisions

Achieving this requires:
Lots of data (and data storage)
Lots of Compute
How can GCP help?

Return to Table of Contents

Choose a Lesson

Pre-trained ML API's

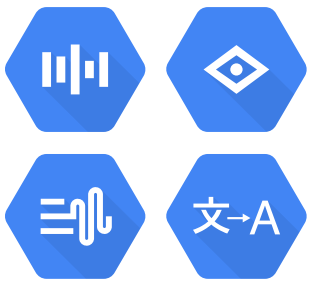
Vision API demo

Pre-trained ML API's

Next

AI Platform (Formerly Cloud ML Engine)

- Train, deploy, and manage custom ML models on managed infrastructure resources.
- You create the model, then Google provides managed infrastructure for testing it.



Pre-trained ML models

- Pre-trained models
- Common use cases (not customizable)
- Simply 'plug' into your application
- "Make Google do it"

Return to Table of Contents

Choose a Lesson

Datalab Overview

Datalab Overview

Next

What is it?

- Interactive tool for exploring and visualizing data:
  - Notebook format
  - Great for data engineering, machine learning
- Built on Jupyter (formerly iPython):
  - Open source - Jupyter ecosystem
  - Create documents with live code and visualizations
- Visual analysis of data in BigQuery, ML Engine, Compute Engine, Cloud Storage, and Stackdriver
- Supports Python, SQL, and JavaScript
- Runs on GCE instance, dedicated VPC and Cloud Source Repository
- Cost: free - only pay for GCE resources Datalab runs on and other Google Cloud services you interact with

Return to Table of Contents

Choose a Lesson

What is Dataprep?

What is Dataprep?

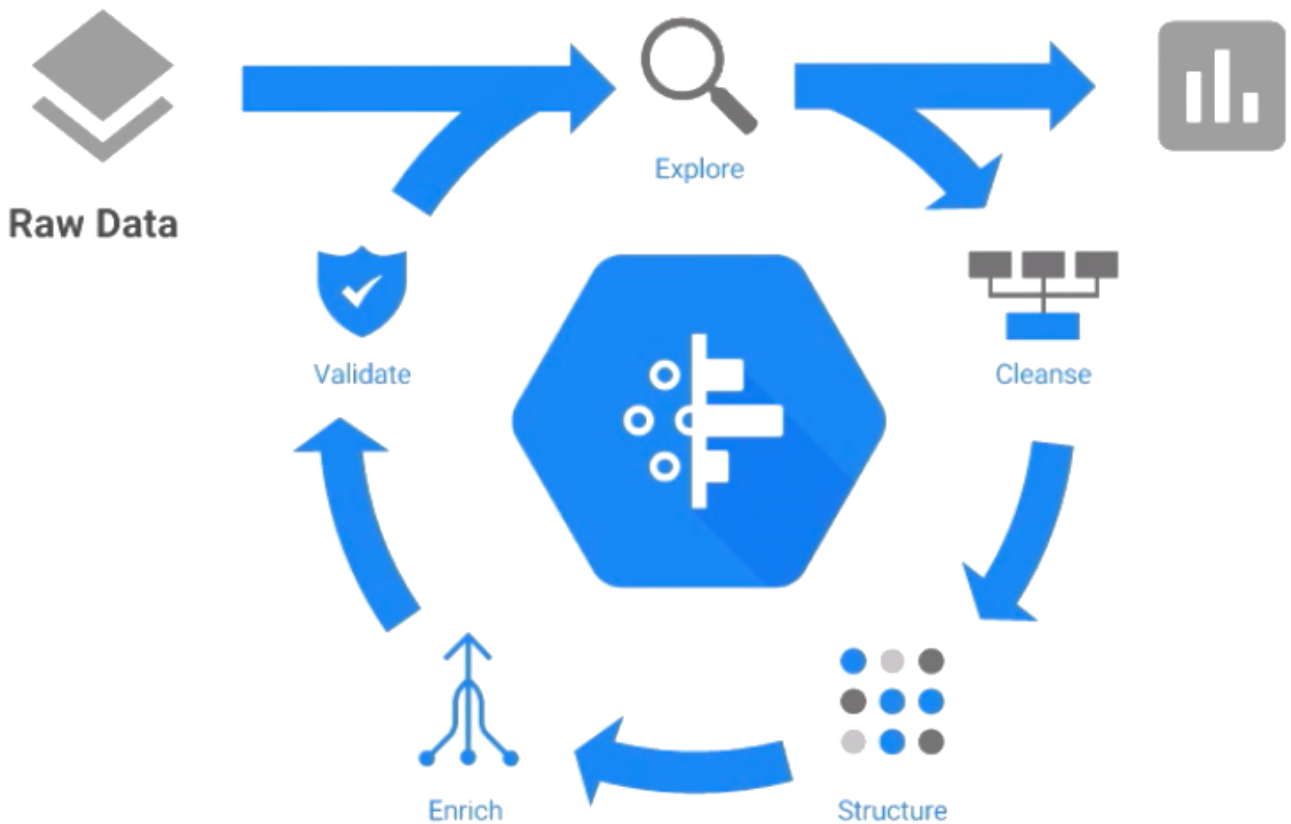
Next

What is it?

- Intelligent data preparation
- Partnered with Trifacta for data cleaning/processing service
- Fully managed, serverless, and web-based
- User-friendly interface:
  - Clean data by clicking on it
- Supported file types:
  - Input - CSV, JSON (including nested), Plain text, Excel, LOG, TSV, and Avro
  - Output - CSV, JSON, Avro, BigQuery table:
    - CSV/JSON can be compressed or uncompressed

Why is this important?

- Data Engineering requires high quality, cleaned, and prepared data
- 80% - time spent in data preparation
- 76% - view data preparation as the least enjoyable part of work
- Dataprep democratizes the data preparation process



Return to Table of Contents

Choose a Lesson

Data Studio Introduction

Data Studio Introduction

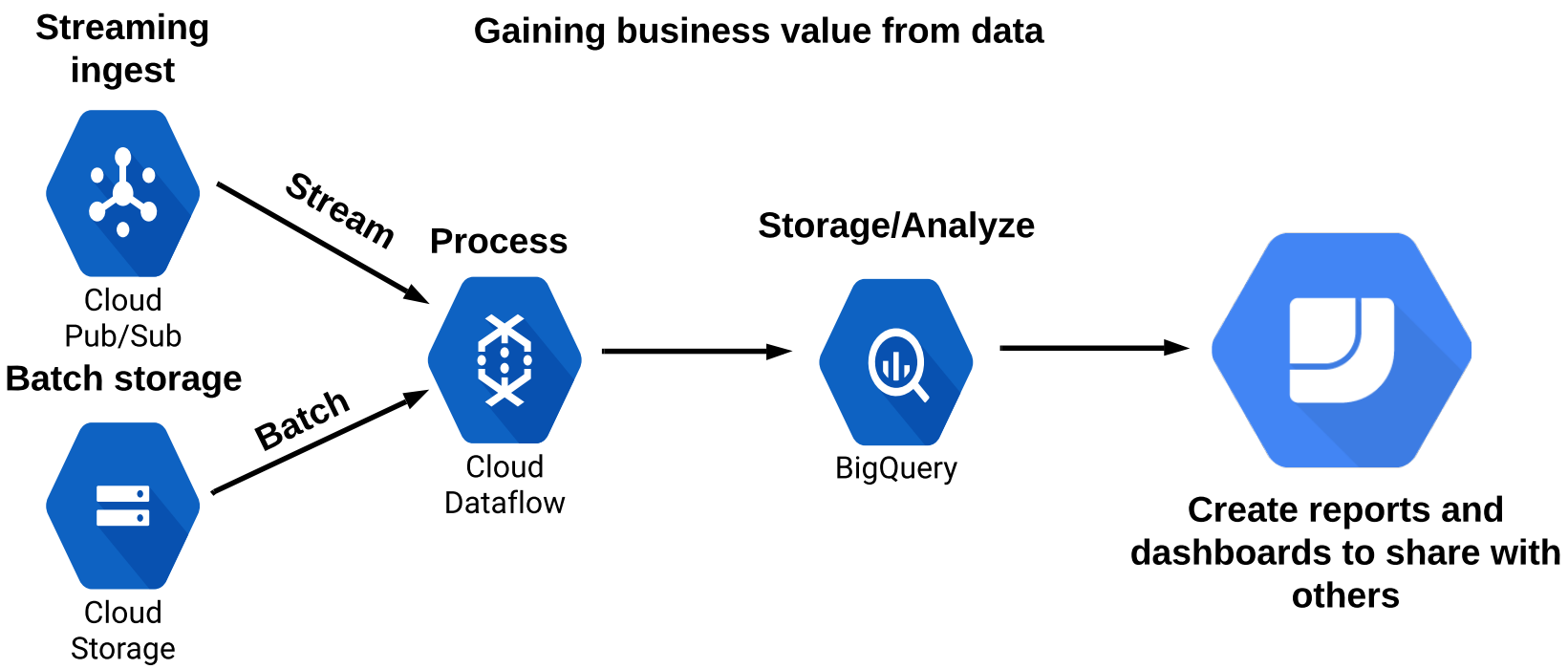
Next

What is Data Studio?

- Easy to use data visualization and dashboards:
  - Drag and drop report builder
- Part of G Suite, not Google Cloud:
  - Uses G Suite access/sharing permissions, not Google Cloud (no IAM)
  - Google account permissions in GCP will determine data source access
  - Files saved in Google Drive
- Connect to many Google, Google Cloud, and other services:
  - BigQuery, Cloud SQL, GCS, Spanner
  - YouTube Analytics, Sheets, AdWords, local upload
  - Many third party integrations
- Price - *Free*:
  - BigQuery access run normal query costs

Data Lifecycle - Visualization

Gaining business value from data



Return to Table of Contents

Choose a Lesson

Cloud Composer Overview

Hands On - Cloud Composer

Cloud Composer Overview

Next

What is Cloud Composer?

- Fully managed **Apache Airflow** implementation:
  - Infrastructure/OS handled for you

What is Apache Airflow?

- Programatically create, schedule, and monitor data workflows

Why is this important?

- Automation and monitoring
- Big data pipelines are often a multi-step, complex process:
  - Create resources in multiple services
  - Process and move data from one service to another
  - Remove resources when they complete a task
- Collaborate workflow process with other team members

How Airflow/Composer helps

- Automates the above steps, including scheduling
- Built on open source, using Python as common language
- Easy to work with, and share workflow with others
- Works with non-GCP providers (on-premises, other clouds)

Return to Table of Contents

Additional Study Resources

SQL deep dive

- Course - SQL Primer
- https://linuxacademy.com/cp/modules/view/id/52

Machine Learning

- Google Machine Learning Crash Course (free)
- https://developers.google.com/machine-learning/crash-course/

Hadoop

- Hadoop Quick Start
- https://linuxacademy.com/cp/modules/view/id/294

Apache Beam (Dataflow)

- Google's guide to designing your pipeline with Apache Beam (using Java)
- https://cloud.google.com/dataflow/docs/guides/beam-creating-a-pipeline