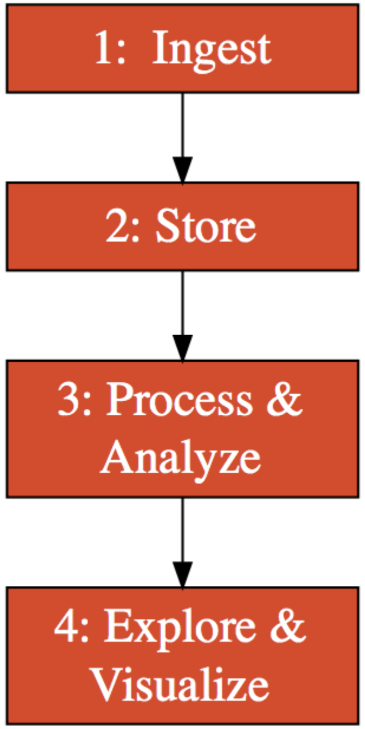
Designing Solutions
Google Cloud Platform

Data Lifecycle

- **Four Steps:**
  - **Ingest**: Stream or Batch ingest
  - **Store**: Durably and cost efficiently store data in a convenient format
  - **Process and analyze**: Convert data to information (normalizations or aggregations)
  - **Explore and visualize**: Flexibility to play with data/information. Get and share insights.



Data Lifecycle - 1 - Ingest

- **Streaming:** Pub/Sub
- **Batch:** Storage Transfer Service, BigQuery Transfer Service, Transfer Appliance, gsutil etc
- **Database migration**: Migrate data from other sources to Google Cloud
  - Database Migration Service (Simplifying migrations to Cloud SQL)
  - Batch transfer to Cloud Storage
  - Load data into database from Cloud Storage using Dataflow

Cloud Pub/Sub

Cloud Storage

Data Lifecycle - 2 - Store

Service	Solution
Cloud Storage	Object Storage (unstructured data)
Cloud SQL	Managed MySQL, PostgreSQL and MS SQL Server databases Relational, pre-defined schema, strong transactions, regional
Cloud Spanner	Horizontally scalable relational database Relational, pre-defined schema, strong transactions, high availability, and global scale
Cloud Firestore	Flexible, scalable, transactional NoSQL database
Cloud Bigtable	Managed wide-column NoSQL Petabyte scale, Real-time apps and large-scale analytical time-series workloads, single-row transactions
BigQuery	Managed data warehouse
Custom Database	Use Cloud Marketplace to deploy an open source database of your choice - MongoDB, Cassandra etc

Data Lifecycle - 3 - Process and analyze

Raw Data > Actionable Information (Clean, Transform)

Service	Solution
Dataprep	Clean and prepare data Fully managed, No-Ops Usecases: Clean data on-boarded from external sources, Prepare data for ML Visual approach for non programmers
Cloud Data Loss Prevention	Scan, discover, classify, and report on data in Cloud Storage, BigQuery, and Datastore (mask, tokenize, and transform sensitive elements)
Dataflow	More flexible ETL pipelines (Fully managed, No-Ops, Batch and Streaming)
Dataproc	Complex processing using Spark and Hadoop Needs a cluster with compute engine VMs Usecases: Machine Learning, Migrate existing Spark and Hadoop workloads

Data Lifecycle - 4 - Explore and visualize

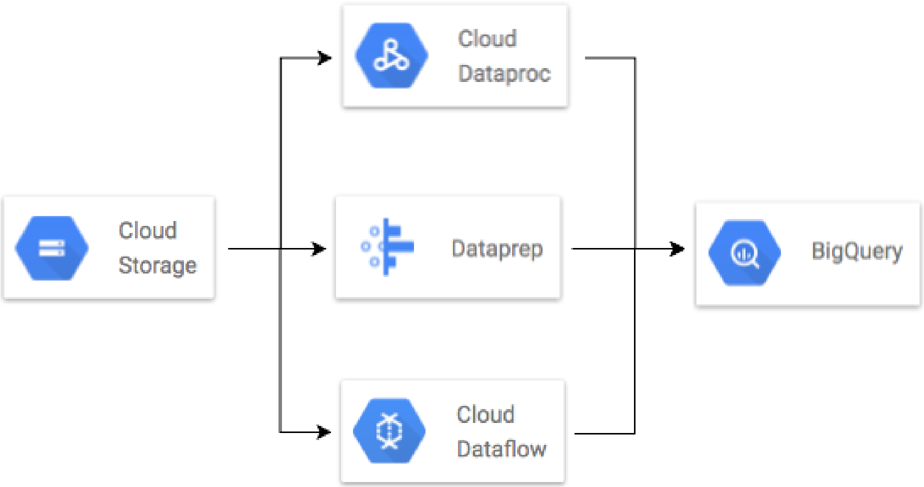
Service	Solution
Cloud BigQuery	Managed data warehouse Standard SQL, serverless, separate storage and compute
ML - Pre built models	Vision API, Speech-to-Text, Natural Language API, Video Intelligence API etc
ML - Custom models	Use AI Platform (based on TensorFlow) Use Dataflow for pre-processing
Cloud Datalab	Web based tool to explore, analyze and visualize data Based on Jupyter notebooks (Use Python, SQL queries etc) Support for popular data-science toolkits - pandas, numpy, and scikit-learn
Cloud Data Studio	Dashboarding and visualization Live charts and graphs based on data in Cloud SQL, BigQuery etc
Cloud Data Catalog	Data discovery and metadata management Unified view of all datasets Tag sensitive data using Cloud Data Loss Prevention (DLP)

Big Data & Analytics in GCP

Service	Solution
Pub/Sub	Foundation for stream analytics and event-driven systems
BigQuery	Serverless data warehouse to analyze petabytes of data Scale storage and compute separately
Google Data Studio	Managed visual analytics service
Dataflow	Data pipelines for (Stream + Batch) use cases
Dataproc	Managed Apache Spark and Apache Hadoop clusters
Dataprep	Clean and prepare data (structured and unstructured)
Datalab	Explore, analyze & visualize data on Jupyter notebooks (Use Python, SQL queries etc) Integrates well with BigQuery
Cloud Composer	Managed workflow orchestration service Create pipelines across clouds and on-premises data centers

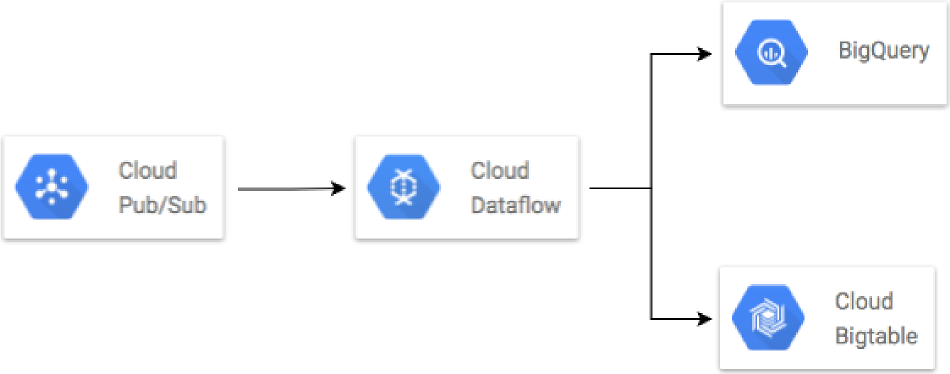
Big Data Flow - Batch Ingest into BigQuery

- Use extract, transform, and load (ETL) to load data into BigQuery
  - **Dataprep**: Clean and prepare data
  - **Dataflow**: Create data pipelines (and ETL)
  - **Dataproc**: Complex processing using Spark and Hadoop



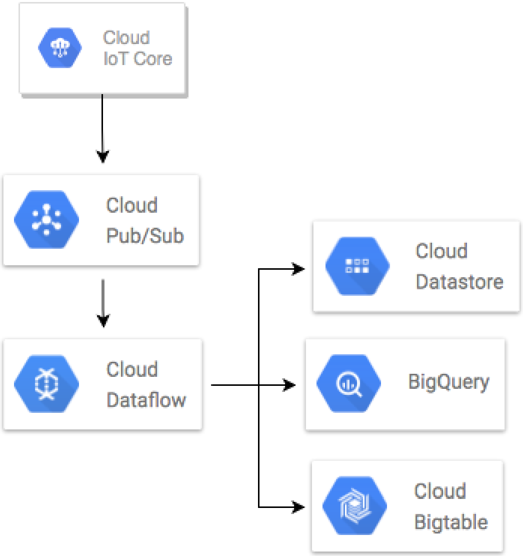
Streaming Workflow - Enable Realtime Querying

- **Query data in Realtime:**
  - **Pub/Sub** and **Dataflow:** Analyze, aggregate and filter data before storing to BigQuery
  - For **pre-defined time series** analytics, storing data in **Bigtable** gives you the ability to perform rapid analysis
  - For **ad hoc complex analysis**, prefer **BigQuery**



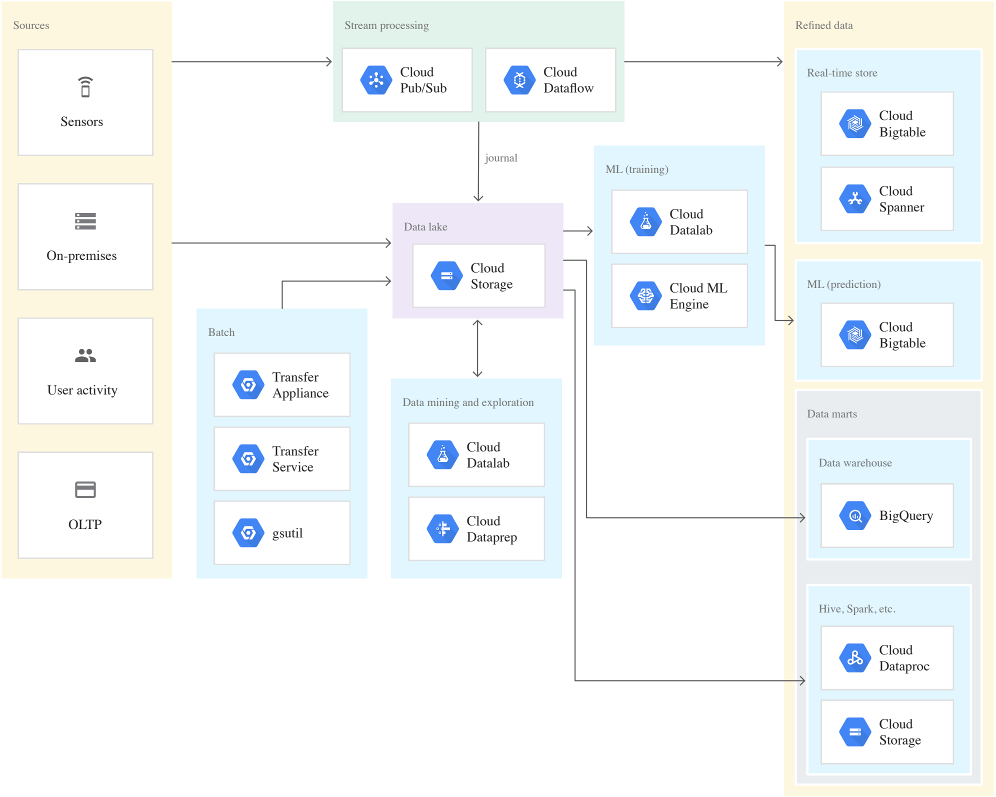
IOT

- **IoT Core**: Manage IoT (registration, authentication, and authorization) devices
  - Send/receive messages/real-time telemetry from/to IoT devices
- **Pub/Sub**: Durable message ingestion service (allows buffering)
- **Dataflow**: Processing data (ETL & more..)
  - Alternative: Use Cloud Functions to trigger alerts
- **Data Storage and Analytics:**
  - Make IOT data available to mobile or web apps => **Datastore**
  - Execute pre-defined time series queries => **Bigtable**
  - More complex or ad hoc analytics/analysis => **BigQuery**



Data Lake - Simplified Big Data Solutions

- Usual **big data solutions are complex**
- How can we make **collecting, analyzing (reporting, analytics, machine learning) and visualizing** huge data sets easy?
- How to design solutions that scale?
- How to build flexibility while saving cost?
- **Data Lake**
  - **Single platform with combination of solutions** for data storage, data management and data analytics

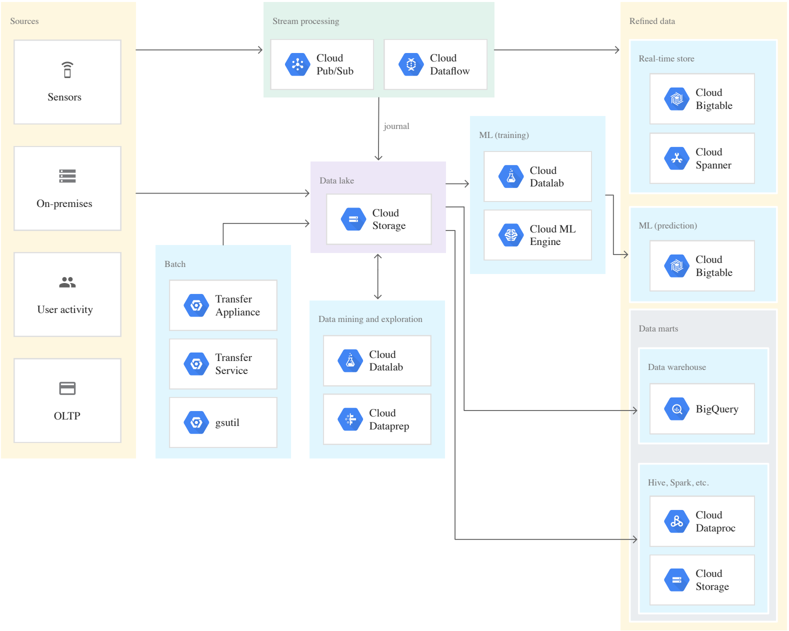


https://cloud.google.com/solutions/build-a-data-lake-on-gcp

GCP Data Lakes - Storage and Ingestion

In28 Minutes

- **Storage:** Cloud Storage (low cost + durability + performance + flexible processing)
- **Data Ingestion:**
  - Streaming data - Cloud Pub/Sub + Cloud Dataflow
  - Batch - Transfer Service + Transfer Appliance + gsutil
- **Processing and analytics:**
  - Run in-place querying using SQL queries using BigQuery or (Hive on Dataproc)
- **Data Mining and Exploration:**
  - Clean and transform raw data with Dataprep
  - Use Cloud Datalab (data science libraries such as TensorFlow and NumPy) for exploring



https://cloud.google.com/solutions/build-a-data-lake-on-gcp