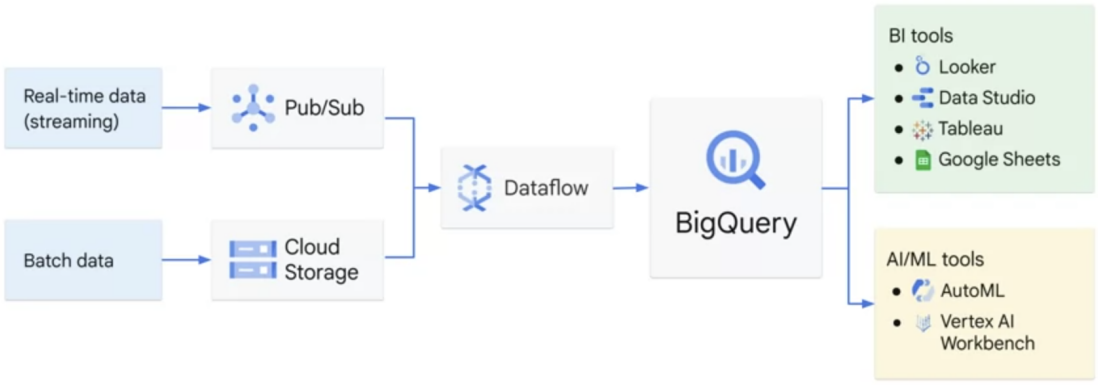
google bd and ml fundamentals

- compute - can help run big data jobs
  - compute engine - VM (IaaS i.e. raw compute, storage, network that need configuration)
  - kubernetes engine - runs containerized apps
  - app engine - full managed PaaS (resource auto provided, can focus on app logic)
  - cloud functions - run code in response to events (FaaS or serverless)
  - cloud run - fully managed PaaS for deploying containers
- storage - separated from compute for scalability
  - cloud storage - suitable for unstructured data
    - standard storage - for frequently accessed data stored for a short time
    - nearline storage
    - coldine data
    - archive storage - for infrequently accessed data stored for long time
  - cloud sql - SQL for OLTP
  - cloud spanner - SQL for OLTP (global scalability)
  - firestore - NoSQL for OLTP
  - bigquery - SQL for OLAP
  - cloud bigtable - NoSQL for OLAP
- data to AI workflow



- ingestion & processing
  - pub/sub - messaging service (distributed, encryption, at-least-once delivery)
  - dataflow
    - process (stream or batch data) by executing apache beam pipelines on google cloud after optimizing those pipeline
    - has templates for stream, batch processing and utility tasks
- storage
  - bigquery
    - managed data warehouse (storage & compute)
    - features : storage, analytics, ML
    - data encryption at rest, data replication & back up
    - auto-scaling
    - allow querying data from other google cloud storage services and cloud providers
    - data can be loaded to bigquery by batch load, streaming and generation
    - type of queries run
      - interactive query - default
      - batch query - query runs when resources are available

            - can analyze data or train ML models with SQL (importing & exporting models, auto hyper param tuning are also supported)
            - automatic train test splitting & feature preprocessing e.g. one-hot encoding
    - analytics
        - looker
            - lookML - modeling language for defining logic and permissions
            - web-based ui
            - looker API - helpful for embedding looker reports in apps
        - data studio
            - no need admin support to connect with bigquery (unlike looker)
    - machine learning
        - vertex ai workbench
            - development environment for customized data science workflow from exploring, training to deploying ML model with code
            - uses container - pre-built or custom image
        - vertex ai
            - brings together the google cloud services for ML in one place
            - can create features and share with others using feature store
            - can train and compare models using auto ml or custom code
            - models are stored in model repository to deploy, share
        - vertex ai pipelines - can help automate/monitor/govern ML systems
- data processing
    - batch processing - processing data in intervals (to wait for more data to arrive)
    - stream processing - process data on arrival (real-time analysis)
- big data challenges
    - variety - different sources or formats
    - volume - large in size

    - velocity - require real-time or fast processing
    - veracity - data quality, consistency
- dwh vs data lake
    - dwh contains structured & organized data
    - data lake contains raw & unorganized data
- phases for ML in bigquery
    - load data into bq
    - select & preprocess features
    - create model (includes training)
        - required to specify model type
        - can inspect model weight
        - can inspect model features' statistics
        - can view training progress (during model training)
    - evaluate model performance
    - make predictions
- ML options in google cloud
    - bigquery ML
        - ML using SQL and data stored in bigquery
        - only support tabular data
        - need to choose ML algo and optionally tune hyper params
    - pre-built APIs
        - using model already trained by google
            - no training data is required
            - save training time
        - can't experiment with hyper params
    - auto ml (automated ml)
        - using few clicks (no code) and own training data

        - can’t experiment with hyper params
    - custom training
        - code to train model
        - require ML expertise
        - need to choose ML algo and optionally tune hyper params
- transfer learning
    - pre-trained models (trained on similar but larger datasets) are used
    - not require too much data or computational time unlike models that are built from scratch
- neural architecture search
    - is used to find optimal model for a project
    - used in google auto ml
- forecasting model - uses multiple rows of tabular & time-dependent data from the past to predict series of numeric values in the future
- google’s ai foundation
    - infra - computer, networking, storage
    - data - bigquery, dataflow, looker
- google’s ai development platform
    - auto ml - vertex ai
    - custom training - vertex ai
    - prebuilt api
    - bigquery ml
- google’s ai solution
    - horizontal solution - can be used in any industry e.g. document ai (can extract info from document)
    - vertical/industry solution - solutions specific to industry
- ml workflow with vertex ai (iterative process)
    - data preparation

        - data uploading
        - feature engineering
    - model training
        - training
        - evaluation
            - precision - use when want to avoid false positives (not miss negative)
            - recall - use when want to avoid false negative (not miss positives)
    - model serving
        - deployment
        - monitoring
- data drifting
    - variation in production data from the testing data
    - causes drop in prediction performance of model
    - need to check data source and adjust model parameter
- artificial intelligence - everything related to computers mimicking human intelligence e.g.
    - ML model
    - robots
- ml
    - subset of ai
    - includes supervised, unsupervised, reinforcement learning
- dl
    - subset of ml
    - uses neural networks
- options for deploying ml models
    - end-point - can give immediate results
    - batch prediction - data is accumulated and predicted

- offline prediction - deploying model in env other than cloud