Overview on Cloud BigLake

- *BigLake is a storage engine that provides a **unified interface** for analytics and AI engines to query multi-format, multi-cloud, and multi-modal data in a secure, governed, and performant manner.*

- *BigLake tables let you query structured data in external data stores with access delegation. Access delegation decouples access to the BigLake table from access to the underlying data store.*

- *An **external connection** associated with a service account is used to connect to the data store. Because the service account handles retrieving data from the data store, you only have to grant users access to the BigLake table.*

- *BigLake enforces fine-grained security at the table level, including **row-level** and **column-level** security.*

Supported data stores

- *You can use BigLake tables with the following data stores:*

    - *Amazon S3 by using BigQuery Omni*

   - *Blob Storage by using BigQuery Omni*

    - *Cloud Storage*

Connectors

- *You can access data in BigLake tables based on Cloud Storage from other data processing tools by using **BigQuery connectors**.*

- *For example, you could access data in BigLake tables from Apache Spark, Apache Hive, TensorFlow, Trino, or Presto.*
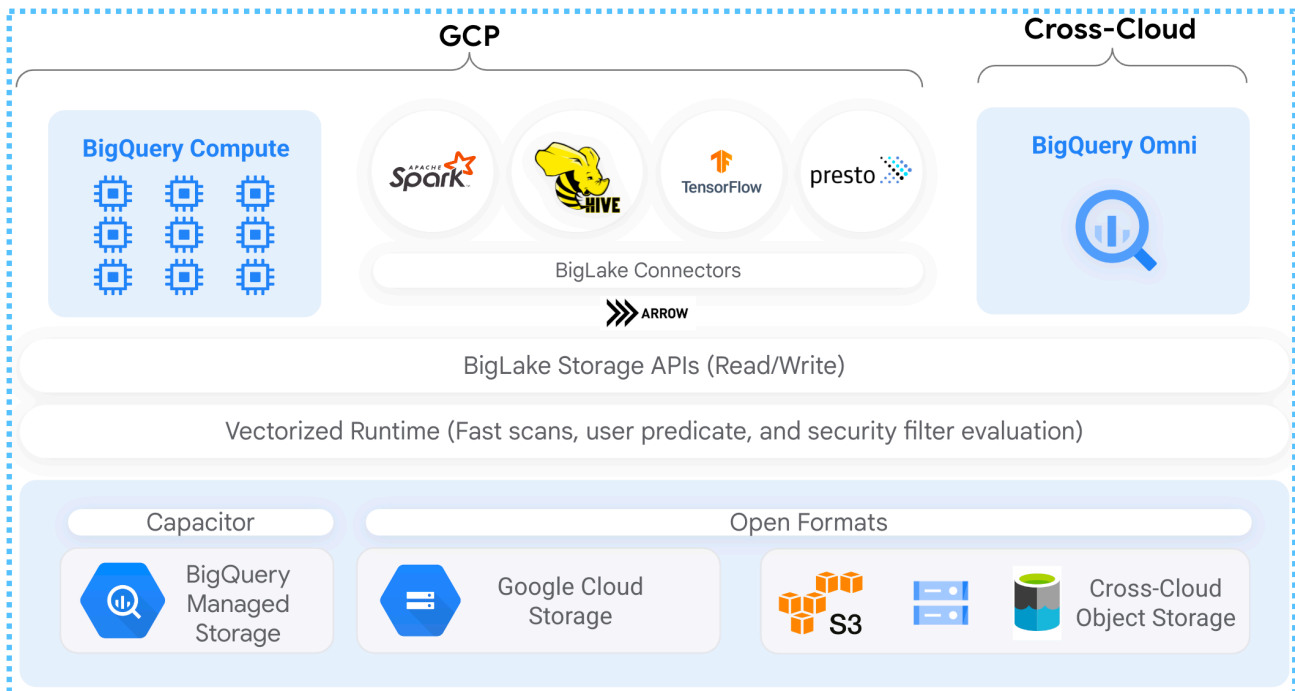
- *The BigQuery Storage API enforces row and column-level governance policies on all data access to BigLake tables, including through connectors.*
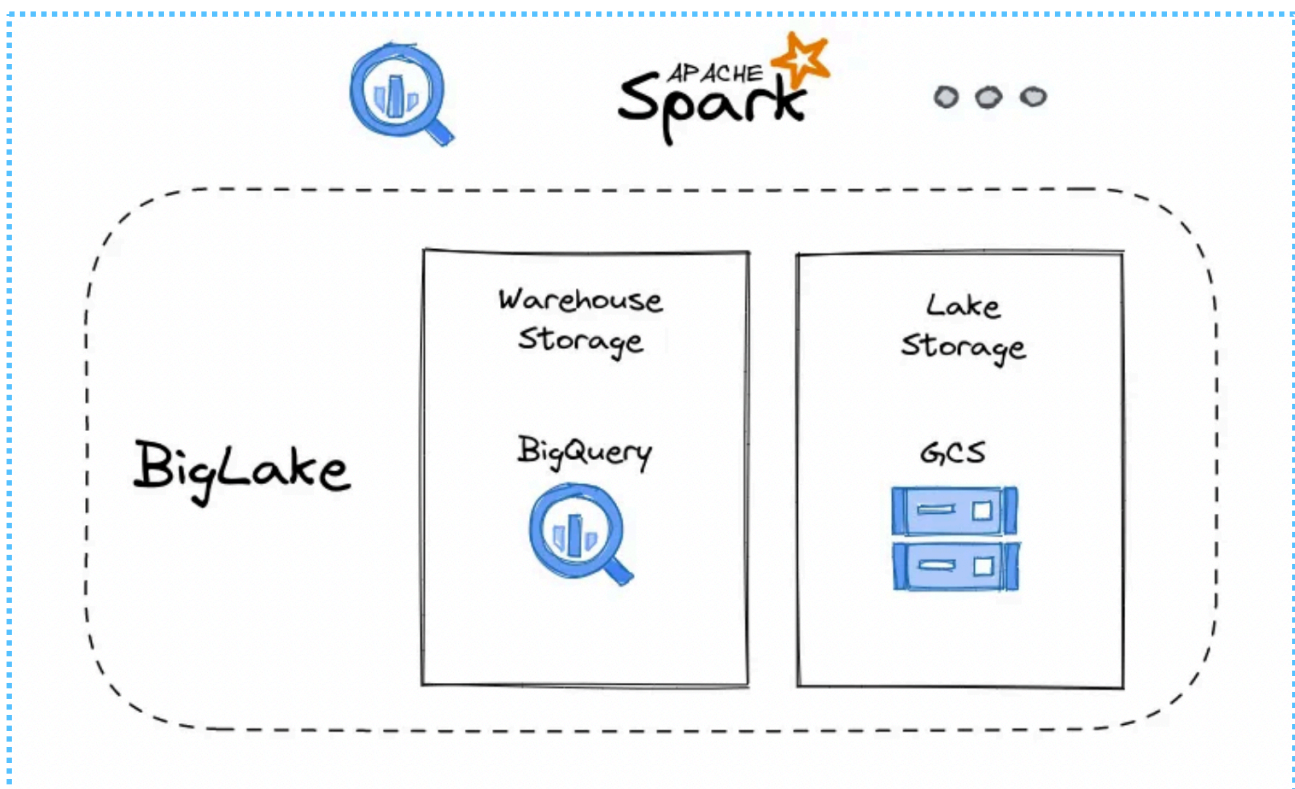
Features Of Cloud BigLake

Fine grained security controls

- *BigLake eliminates the need to grant file level access to end users.*

- *Apply table, row, column level security policies on object store tables similar to existing BigQuery tables.*

Multi-compute analytics

- *Maintain a single copy of structured and unstructured data and make it uniformly accessible across Google Cloud and open source engines, including BigQuery, Vertex AI, Dataflow, Spark, Presto, Trino, and Hive using BigLake connectors.*

- *Centrally manage security policies in one place, and have it consistently enforced across the query engines by the API interface built into the connectors.*

Multi-cloud governance

- *Discover all BigLake tables, including those defined over Amazon S3, Azure data lake Gen 2 in Data Catalog.*

- *Configure fine grained access control and have it enforced across clouds when querying with BigQuery Omni.*

Built for artificial intelligence (AI)

- *Object tables enable use of multimodal data for governed AI workloads. Easily build AI use cases using BigQuery SQL and its Vertex AI integrations.*

Built on open formats

- *Supports open table and file formats including Parquet, Avro, ORC, CSV, JSON.*

- *The API serves multiple compute engines through **Apache Arrow**. Table format natively supports Apache Iceberg, Delta, and Hudi via manifest.*

Pricing

- *BigLake pricing is based on querying BigLake tables, including:*

    - ***BigQuery pricing** applies for queries over BigLake tables defined on Google Cloud Storage.*

    - ***BigQuery Omni pricing** applies for queries over BigLake tables defined on Amazon S3 & Azure data lake Gen 2.*

    - *Queries from open-source engines using BigLake connectors:*

        - *BigLake connectors use BigQuery Storage API, and corresponding prices apply - billed on bytes read, and Egress.*

    - *Additional costs apply for query acceleration using metadata caching, object tables, and BigLake Metastore.*

    - *Ex - The first 1 TB of data processed with BigQuery each month is free.*

Limitations

- *All limitations for external tables apply to BigLake tables.*

- *BigLake tables on object stores are subject to the same limitations as BigQuery tables.*

- *BigLake does not support downscoped credentials from Dataproc Personal Cluster Authentication.*

- *BigLake tables are read-only. You cannot modify BigLake tables using DML statements or other methods.*

- *BigLake tables only support the following formats:*

  - *Avro*

  - *CSV*

  - *Iceberg*

  - *JSON*

  - *ORC*

  - *Parquet*

- *The BigQuery Storage API is not available in other cloud environments, such as AWS and Azure.*