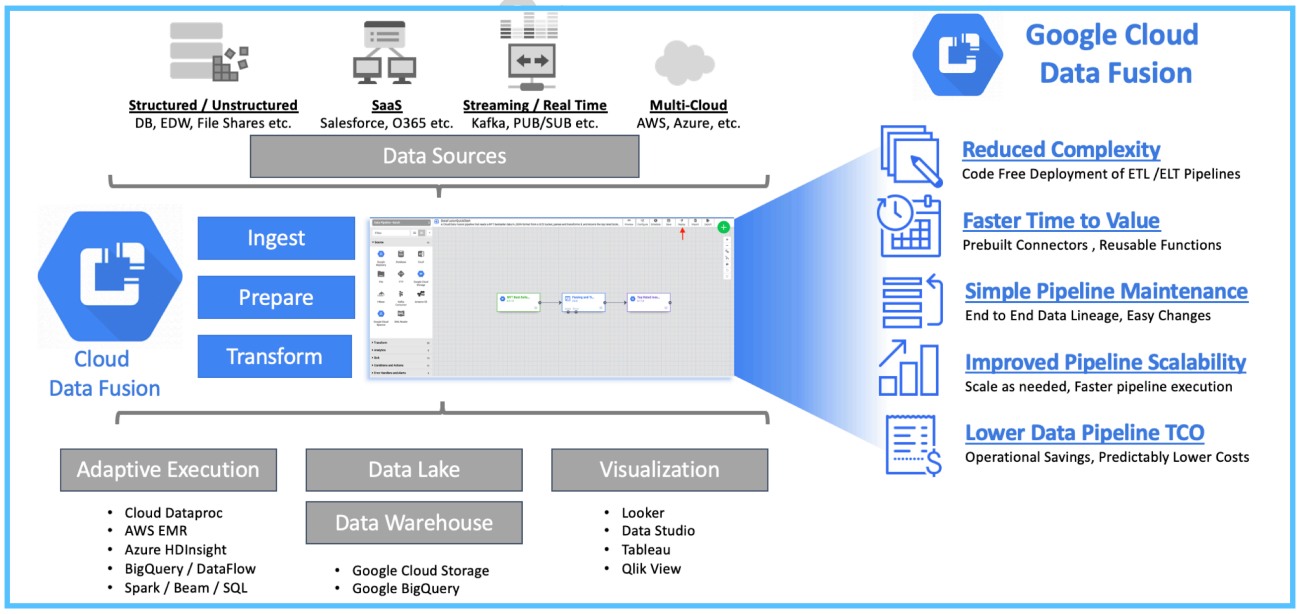
Cloud Data Fusion

- Cloud Data Fusion is a fully managed, cloud-native data integration service that helps users efficiently build and manage big data ETL/ELT data pipelines.
- Code free and Drag-n-drop tool, Data Fusion offers a graphical interface and a broad open-source library of 150+ preconfigured connectors and transformations.
- Cloud Data Fusion shifts an organization's focus away from code and integration to insights and action.
- The Cloud Data Fusion web UI allows you to build scalable data integration solutions to clean, prepare, blend, transfer, and transform data, without having to manage the infrastructure.



- Cloud Data Fusion allows non-technical users to develop pipelines using a code-free graphical interface that delivers point-and-click data integration. This removes the dependence on developers
- Its fully-managed, Google Cloud-native architecture unlocks the scalability, reliability, security, and privacy guarantees of Google Cloud .

- Integration with Cloud Identity and Access Management and Cloud Identity-Aware Proxy provides enterprise security and alleviates risks by ensuring compliance and data protection.
- Cloud Data Fusion is powered by the open-source project CDAP.
    - CDAP(Cask Data Application Platform) is a data application platform for building and managing data analytics applications in hybrid and multi-cloud environments.
    - Open source provides the flexibility and portability required to build standardised data integration solutions across hybrid and multi-cloud environments.
- Cloud Data Fusion pricing is split across two functions: **pipeline development** and **execution**.
- Editions are available
    - **Developer**
        - This edition provides a full-feature edition for product exploration and development environments with zonal availability and limitation on execution environment.
    - **Basic**
        - This edition provides comprehensive data integration capabilities.
        - Users can build batch data pipelines; connect to any data source; perform code-free transformations.
        - Limitation on simultaneous pipeline runs.
        - Recommended for non-critical environments.
    - **Enterprise**
        - This edition provides all the functionality provided in the Basic edition.

    - In addition, includes support for realtime data pipelines; interactions with data lineage; higher scalability; and high availability.
    - Recommended for critical environments.

- Interfaces
    - Using the code-free web UI
    - Using command-line tools

Features of Cloud Data Fusion

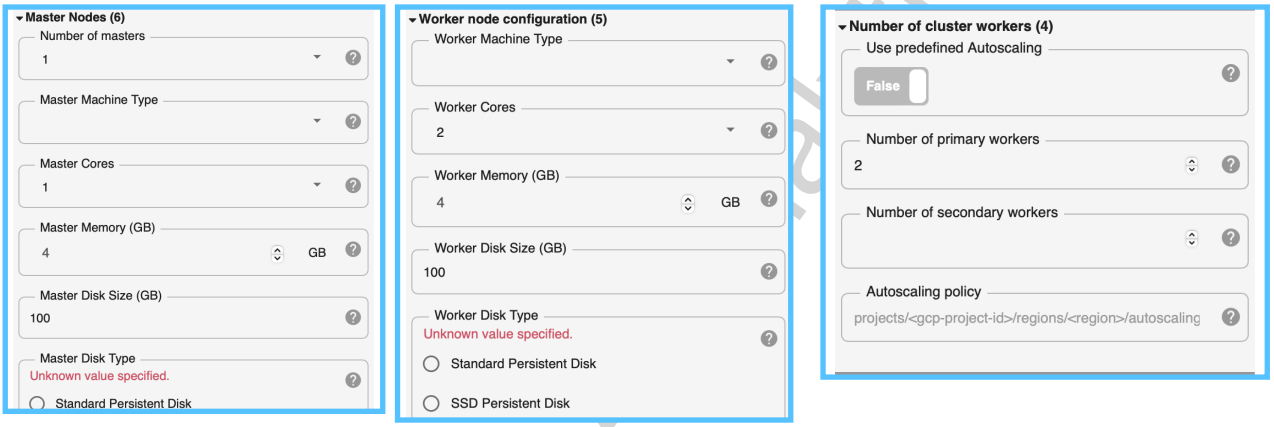
- **Code-free self-service**
    - Remove bottlenecks by enabling nontechnical users through a code-free graphical interface that delivers point-and-click data integration.
- **Collaborative data engineering**
    - Cloud Data Fusion offers the ability to create an internal library of custom connections and transformations that can be validated, shared, and reused across an organization.
- **Google Cloud-native**
    - Fully managed Google Cloud-native architecture unlocks the scalability, reliability, security, and privacy features of Google Cloud.
- **Enterprise-grade security**
    - Integration with Cloud Identity and Access Management (IAM), Private IP, VPC-SC and CMEK provides enterprise security and alleviates risks by ensuring compliance and data protection.

- **Integration metadata and lineage**
  - Search integrated datasets by technical and business metadata. Track lineage for all integrated datasets at the dataset and field level.
- **Seamless operations**
  - REST APIs, time-based schedules, pipeline state-based triggers, logs, metrics, and monitoring dashboards make it easy to operate in mission-critical environments.
- **Comprehensive integration toolkit**
  - Built-in connectors to a variety of modern and legacy systems, code-free transformations, conditionals and pre/post processing, alerting and notifications, and error processing provide a comprehensive data integration experience.
- **Hybrid enablement**
  - Open source provides the flexibility and portability required to build standardized data integration solutions across hybrid and multi-cloud environments.
- **Real time data integration**
  - Replicate transactional and operational databases such as SQL Server, Oracle and MySQL directly into BigQuery with just a few clicks using Data Fusion's replication feature.
  - Integration with Datastream allows you to deliver change streams into BigQuery for continuous analytics.
- **Batch integration**
  - Design, run and operate high-volumes of data pipelines periodically with support for popular data sources including file systems and object stores, relational and NoSQL databases, SaaS systems, and mainframes.

Demo

- Read data from BigQuery
- Wrangle – Transformation
- Sink/write it to Cloud Storage

Observation



Status

1. Provisioning - Dataproc cluster creation
2. Starting - data fusion starting
3. Running - job is in progress
4. Result - Success/failure
    1. Blue - running
    2. Red - failure
    3. Green - Success

Cloud Data Fusion pricing

- Cloud Data Fusion pricing is broken down by:
    - **Design cost:** based on the number of hours an instance is running and not the number of pipelines being developed and run.
        - The Basic edition offers the first 120 hours per month per account at no cost.

- **Processing cost:** The cost of Dataproc clusters used to run the pipelines.

EDITION	PRICE PER CLOUD DATA FUSION INSTANCE HOUR	NUMBER OF SIMULTANEOUS PIPELINES SUPPORTED	NUMBER OF USERS SUPPORTED
Developer	US$0.35	2 (Recommended)	2 (Recommended)
Basic	US$1.80	Unlimited	Unlimited
Enterprise	US$4.20	Unlimited	Unlimited

Development

For pipeline development, Cloud Data Fusion offers the following three editions:

Cloud Data Fusion Edition	Price per instance per hour
Developer	$0.35 (~$250 per month)
Basic	$1.80 (~$1100 per month)
Enterprise	$4.20 (~$3000 per month)

The Basic edition offers the first 120 hours per month per account free.

- For pricing purposes, usage is measured as the length of time, in minutes, between the time a Cloud Data Fusion instance is created to the time it is deleted.
- Cloud Data Fusion is billed by the minute.

Execution

- For pipeline execution, you are charged for the Dataproc clusters that Cloud Data Fusion creates to run your pipelines at the current Dataproc rates.