

ETL Features

This is a list of features of an ETL product that can be adjusted by engineers for any specific use.

Pluggable Protocols

Ability to implement and plug an input/output connector for any data source/target.

Pluggable Creds Store

A Creds Store stores connection credential and other secrets that for security reasons should not be in the source and configuration of any specific pipeline to keep it away from hackers, users, devops and most engineers.

Pluggable Conf Store

Pipeline configurations should live separately from a specific deployment of a pipeline in order to be able to tear down, restore, redeploy etc. pipelines.

Connector Features

This assumes that data can be pulled in batches because some data sources are big.

It should be possible to configure:

- Max data size to pull/push in each batch
- Interval between batches
- Retry policy

Pipeline Features

- Ability to pull data in batches
- Each step in a pipeline should have its own executor for faster flow of data
- Scheduling
- Dependencies among pipelines
- Ability to both stream data infinitely and to reload it fully

Monitoring

- Pluggable storage and logging for all steps of a pipeline for centralized monitoring and management
- Each step in a pipeline should persist data (assets) it outputs for the next steps
- Ability to apply logic on top of assets and logging in order to handle errors and notifications

Permissions

- Ability to create users and define permissions in order to
 - allow users access to various groups of pipelines

- limit users' ability to see data that passes through and control pipelines

Programmatic Access

Ability to manage pipelines programmatically and remotely

Deployments

Provide multiple deployment targets probably through Kubernetes

Evaluated Products

dagster; pulsar; spark, flink; stitch, meltano, airbyte; prefect, airflow; aws batch + step functions; fivetran, domo, hevo; streamsets;

Winner: Dagster

- The goal of Dagster is DAG
- It is possible to build ETL on top of Dagster
- Dagster is complex

It is possible to use Dagster to build a better ETL tool but one must take into account that Dagster is a complex product.