Cask Hydrator

Self-Service Ingestion and ETL for Hadoop Data Lakes

Open Source and Highly Extensible

Rich Drag-and-Drop User Interface

Build for Production on CDAP

DISCOVER

data using user and machine generated metadata

Search · Business Metadata · Machine Generated Metadata

INGEST

any data from any source in realtime and batch

FLUME · RDBMS · Kafka · EDW · N★SQL · Cloud · IOT

BUILD

drag-and-drop ETL/ELT pipelines that run on Hadoop

MapReduce · Spark · HIVE · Tigon · {JSON}

EGRESS

any data to any destination in realtime and batch

HBASE · cassandra · RDBMS · elasticsearch · EDW · Cloud

Examples of Realtime ETL Pipelines

Twitter to HBase Table

- Ingest Tweets in realtime directly from the Twitter API into an HBase Table
- Utilize natural language processing to determine social media sentiment

Kafka to HBase OLAP Cube

- Ingest events in realtime from Kafka topics into an HBase Table
- Perform OLAP aggregations over streaming events for realtime analytics and dashboards

Kafka to Elasticsearch

- Ingest events in realtime from Kafka topics into Elasticsearch indexes
- Automatically index and make all data sent through Kafka available for search and analytics through Elastic and Kibana

Amazon SQS to HBase

- Ingest events in realtime from Amazon SQS into HBase
- Store events into HBase for subsequent analytics and fast access

Examples of Batch ETL Pipelines

RDBMS/EDW to HBase Table

- Dump relational database tables into HBase Tables
- Enable data lakes, ETL offloading and new big data analytics by bringing copies of existing relational datasets into Hadoop

Amazon S3 to HDFS Avro

- Ingest files and directories from Amazon S3 into HDFS Avro files
- Bring cloud-generated data into HDFS for further processing and analysis using MR/Spark or SQL

HDFS to HDFS Parquet

- Rewrite HDFS files from raw or row-based formats into columnar-based formats like Parquet
- Simplify and automate ETL processes to enable interactive queries using engines like Impala and Presto

HDFS to/from Cassandra

- Process raw HDFS files into Cassandra tables or dump Cassandra tables into HDFS files
- Utilize Cassandra to enable interactive and online access, or load Cassandra data into Hadoop for archiving and analytics