CDAP EXTENSIONS

POWERED BY CDAP

CASK

Cask Data Application Platform (CDAP) Extensions

CDAP Extensions provide additional capabilities and user interfaces to CDAP. They are use-case specific applications designed to solve common and critical big data challenges.

Current extensions enable self-service ways to continuously acquire, blend, transform, apply science, and distribute data using Hadoop. They also provide solutions to centrally capture and store metadata where it can be conveniently accessed and indexed, thereby allowing users to easily search and retrieve high quality, consistent metadata.

CDAP is the de-facto platform for building data applications. It is ideal because of its simple and easy to use APIs which help maximize developer productivity, reducing TCO. Furthermore, CDAP is highly extensible and provides future proofing by integrating new technologies and supporting many different workloads.

There are currently two extensions that are packaged with and powered by CDAP:

Cask Hydrator

Cask Tracker

Cask Hydrator

A self-service, reconfigurable, extendable framework to develop, run, automate, and operate data pipelines.

Cask Hydrator

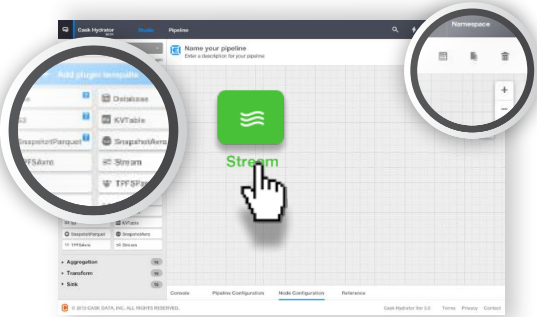
Cask Hydrator is a self-service, reconfigurable, extendable framework to develop, run, automate, and operate data pipelines on Hadoop. It is 100% open source and licensed under the Apache 2.0 license.

Hydrator prepares, blends, aggregates, and applies science to create a complete picture of your business data drives actionable insights. This solution delivers accurate, analytics-ready data and analytics to end users. With visual tools to eliminate coding and complexity, Hydrator will put big data at your fingertips.

Ease of use

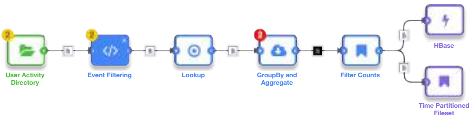
The intuitive drag-and-drop interface integrates with Hadoop and non-Hadoop storage and has the ability to switch between different processing technologies - MapReduce, Spark or Spark Streaming.

- The graphical data pipeline studio simplifies the creation of data pipelines
- A rich library of pre-built Plugins to access, transform, blend, and aggregate data from relational sources, big data sources, and more
- Powerful orchestration capabilities coordinate and combine transformations, including notifications and alerts
- Integration with enterprise schedulers allows for coordinating workflows, a testing framework, and the ability to monitor and tune jobs



Integration, Aggregation, and Analytics with Zero Coding Required

Hydrator's intuitive interface accelerates the design and deployment of big data analytics by up to five times compared to hand-coding techniques.



- Complete visual integration eliminates manual programming and scripting from the process
- Empowers users to architect big data pipelines to create complete and accurate data solutions
- Streamlines analytical processes and eliminates the need for manual steps, specialized resources, and overall complexity
- Robust support for relational data sources, No-SQL data stores, and others

Accessible for wider audience

Delivers governed, best-practice, on-demand data to data scientists, data engineers, analysts, and end users in an agile fashion.

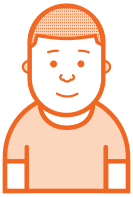
- Seamless self-service integration solutions for transforming, aggregating, and enriching large scale and variety of data
- Consistent support for batch and real-time data pipelines
- Requires minimal support from IT to support organizations and business users with reliable, repeatable, and governed data pipelines
- Automatic creation and publishing of datasets to drive faster and more reliable analytics
- Seamless integration with visualization and data services, making datasets immediately available to reports and applications
- Integrations with advanced analytics like Spark ML to operationalize predictive intelligence while reducing build time

Enterprise Ready to accelerate the Data Lake initiative

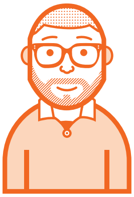
Goes beyond data ingestion to scalable and flexible management for end-to-end data pipelines with enterprise capabilities, delivering key initiatives such as Data Lakes.

- Dynamic and reusable templates that drive massive resource savings by re-using code
- Robust administration features, including SLA monitoring, job restart, error handling and restart, and an operations center for auditing access
- Enterprise-grade security including access and version controls as well as LDAP, JSAPI, and Active Directory integration
- Enhanced Data Management through integration with extensions like Cask Tracker to track data and metadata at all times
- Distribution and deployment agnostic, enables moving from on-premises to cloud and back, changing to a different distribution of Hadoop, or switching technologies for running data pipelines
- Enterprise-grade customer support with best-in class services and training

It’s built for



Developers and
ETL Developers



Data Scientists and
Data Architects

You can build

- Data Processing Pipelines - Real-time and Batch
- Data Ingestion Pipelines and Realtime Data Pipelines

Learn More About Cask Hydrator

Cask Hydrator is a CDAP extension and is 100% open source!

- Download CDAP today to try out Cask Hydrator: **cask.co/downloads**
- Visit the Cask Hydrator product page: **cask.co/products/hydrator**
- Or read more about Cask Hydrator: **cask.co/hydrator-wp**

Case Study

Information Security Analytics and Reporting

The Challenge — The customer, a Fortune 50 financial institution, created a pipeline that aggregates batched data on a secured Hadoop cluster to create daily aggregates and reports. That system performed multiple transformations, which created new datasets. The customer faced multiple issues:

1. The data pipeline was inefficient and took 6 hours to run and required manual intervention on almost a daily basis
2. Reports did not align correctly with day-to-day boundaries
3. Any points of failure required reconfiguring and restarting the pipeline, a time-consuming and frustrating task
4. Major setup and development time was needed to add new sources
5. The team could not test and validate the pipeline prior to deployment. As a result, testing was conducted directly on the cluster — a wasteful use of resources

The Cask Hydrator Solution — The customer's data development team created independent parallel pipelines that moved the data from SQL Servers and Teradata into Time Partitioned Datasets. Transformations were performed in-flight with the ability to handle error records. After completing the initial transfers, another pipeline combined the data into a single Time Partitioned Dataset and fed it into an aggregation and reporting pipeline.

Results

- In-house Java developers with limited Hadoop knowledge built and ran the complex pipelines at scale within two weeks, following four hours of training
- The data pipeline now took ~2 hours to run without any manual intervention
- The visual interface enabled the team to develop, test, debug, deploy, run, automate, and view pipelines during operations
- The new process reduced system complexity, which simplified pipeline management
- The development experience was improved by reducing wasteful cluster utilization
- Transforms were performed in-flight with the ability to handle error records
- Tracking tools made it easy to rerun the process from any point of failure

Case Study

In-Flight Brand Sentiment Analysis of the full Twitter Firehose

The Challenge — A Fortune 500 e-commerce company built a data pipeline that ingested their Twitter stream in real-time. The data was cleansed and transformed prior to conducting multi-dimensional aggregations and sentiment analysis on marketing campaigns based on Tweets. The results were updated twice daily to HBase. However, the legacy pipeline suffered on two fronts: first, latency in the existing pipeline delayed the decision making process. Second, the existing data movement process proved to be costly in time and money.

The Cask Hydrator Solution — The company's in-house team of Java developers built a real-time pipeline in two weeks using the drag-and-drop visual interface in Cask Hydrator. They developed a sentiment analysis transform using the API and then included it in the pipeline. Further, they added multi-dimensional aggregations without needing to write code using CDAP's Cube Plugin as a sink.

Results

- The analysis of Tweets in real-time allowed the business to make faster decisions on their campaigns
- The new pipeline eliminated latency between aggregation and the availability of results, producing quicker and better decisions while cutting costs
- The new pipeline cleansed, transformed, analyzed, and aggregated tweets at the rate of the full Twitter firehose in real-time
- The infrastructure was consolidated into a single Hadoop cluster
- In-house Java developers were able to build the pipeline and sentiment analysis plugin with a four hour learning curve
- Seamless transparency through custom dashboards provided easy operational insights and aggregated logs for debugging

Case Study

Encrypting and Data Masking

The Challenge — The customer, a Fortune 50 company in the Telecom sector, developed a legacy custom data pipeline that performed format-preserving encryption and data masking on a KerberOS Hadoop cluster. The pipeline extracted data from Teradata to HDFS, performed transformations, and loaded the results back into Teradata on a daily basis. This pipeline, built by a third-party service, was operationally unstable, and required constant costly intervention to keep it running.

The Cask Hydrator Solution — The self-service, code-free interface allowed the in-house team to reproduce and replace the existing pipeline. The new process performed the extraction, encryption, masking, and reload to and from Teradata in-flight. It created a copy of the data on HDFS so the team could run complex ad-hoc queries using Hive.

Results

- Using the code-free drag-and-drop visual interface, the in-house team built the pipeline in five days
- They were able to easily achieve scale with Hydrator to monitor and achieve their SLAs
- IT gained immediate insights into the performance of the data pipeline and were able to easily determine and handle failure scenarios
- Additional complex ad-hoc queries were offloaded from Teradata, further reducing overall cost

Case Study

Data Cleansing and Validating 3 Billion Records

The Challenge — The customer, a Fortune 500 company in the Financial sector, had custom-built a data pipeline to perform data validation and correction transforms. The pipeline was constructed using multiple complex technologies. Operations performed on the 3 billion records included:

- Standardization, verification, and cleansing of USPS codes
- Domain set validation, Null Checks, and Length Checks
- Regular expression validation (email, SSN, dates, etc.)

The legacy pipeline ran overnight, required multiple teams to keep it operating, and costly experts to maintain it.

The Cask Hydrator Solution — In-house Java programmers developed, tested, and ran the replacement pipeline using the drag-and-drop visual interface. The new pipeline only required limited coding in order to integrate custom regular expressions.

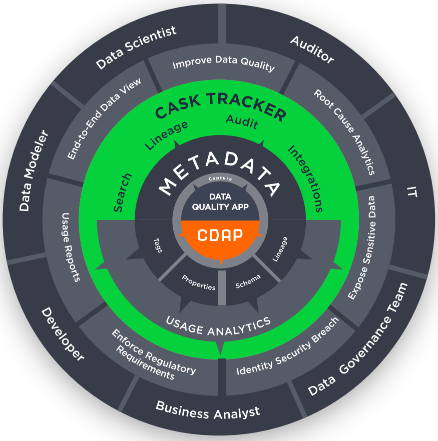
Results

- The in-house team built, tested, and deployed the pipeline in 3 days
- Processing the three billion records took less than 65% of the time compared to the custom-built pipeline
- The development team only required the standard four hour training on Hydrator before launching the project
- The new pipeline eliminated the need for costly Hadoop experts, improved performance, and reduced the number of technologies involved

Cask Tracker

A self-service framework that automatically captures rich metadata and provides users with visibility into how data is flowing into, out of, and within a Data Lake.

Cask Tracker



Cask Tracker is a self-service CDAP extension that provides users with visibility into how data is flowing through, within, into, and out of a Data Lake. It allows them to perform impact and root cause analysis and provides an audit-trail for auditability and compliance. It enables IT to oversee changes, while delivering trusted, secure data in a complex Data Lake environment. Tracker provides access to structured information that describes, explains, locates and makes it easier to retrieve, use, and manage datasets.

Tracker also provides a way to store metadata where it can be accessed and indexed. This allows it to be easily searchable and provides high quality metadata of consistent quality so users know that it can be trusted.

Linking business metadata with the underlying technical metadata provides context for collaboration across the organization and helps answer questions like:

- Who is using the dataset?
- Which datasets are used and where were they used?
- What is the dataset quality?
- What is the definition of the dataset?
- What processes use a specific dataset?
- What applications will be impacted if a dataset is modified?
- What process created the dataset?
- What other datasets did this dataset help create?

It also makes it possible to:

- Find dataset(s) based on metadata — business and technical
- Find dataset(s) based on schema fields and field types

Harvest, Index, and Track Datasets

- Immediate, timely, and seamless capture of technical, business, and operational metadata enabling faster and better traceability of all datasets
- Quickly, reliably, and accurately indexes technical, business, and operational metadata to easily locate datasets
- Understand the impact of changing datasets on other datasets or processing and queries using lineage
- Track data flow of data across enterprise systems and data lakes, no matter which process or application is moving or transforming your data
- Trusted and complete metadata on datasets provides easy traceability to resolve any data issues and improve data quality

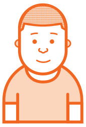
Support Standardization, Governance, and Compliance Needs

- Provide IT with traceability needed in governing datasets and easily applies compliance rules through seamless integration with other extensions
- Consistent definitions of metadata containing information about data to reconcile difference in terminologies
- Empowers business users in understanding lineage of business-critical data

Blend Metadata Analytics and Integrations

- Gain deep insights into how your datasets are being created, accessed, and processed with built-in usage analytics capabilities
- Valuable multi-dimensional usage analytics to understand complex interactions between users, application, and datasets
- Deeper and extensible integrations with enterprise-grade MDM systems like Cloudera Navigator and others for centralizing metadata repository, to deliver accurate, complete, and correct data to all

It’s built for



Developers

Data Engineers

Data Scientists

Data Architects

With Cask Tracker you can

- Search for Datasets in your Data Lake
- Debug data issues
- Improve data quality

Learn More About Cask Tracker

Cask Tracker is an extension of CDAP and is 100% open source!

- Download CDAP today to try out Cask Tracker: **cask.to/get-cdap**
- Visit the Cask Tracker product page: **cask.to/tracker-pp**

Case Study

Iterative Data Science

The Challenge — Once an organization has a process or system for ingesting data into a Data Lake, the next stage is to provide an easy way to discover, inspect, and track datasets in that data lake. In one customer example, their data scientists and business analysts found it very difficult to locate the datasets in their data lake and there was no easy way to discover datasets based on the business or technical metadata. This often led to regenerating the datasets and redoing the same work multiple times, thereby wasting time and over-utilizing cluster resources, which required the IT team to add more capacity to the clusters. Additional issues observed in this environment:

- It often took days or weeks to discover datasets needed for ideation
- Due to difficulty in locating datasets, cluster resources were over-utilized to re-create datasets that were already present on the cluster
- Multiple instances of the same datasets created confusion around their authoritative nature
- IT spending was unpredictable as more data scientists and data analysts were added to the team

The Cask Tracker Solution — Using Cask Trakcer and Cask Data Application Platform (CDAP), all datasets that were generated were tracked and indexed by their metadata — both Technical and Business metadata. This allowed data scientists and data analysts to use Cask Tracker to discover all datasets in the data lake and inspect rich metadata such as tags, schema and properties. They were then able to use the automatic lineage tracking capabilities to figure out the source of the dataset and understand the transformations that were applied on the source data. Operational metadata helped them identify the type and frequency of processing that was applied on the datasets in addition to who performed it. They also used the captured audit data to determine the freshness and activity-level of the datasets.

Results

- Time to discovering datasets on a Data Lake was reduced from days or weeks to minutes or hours
- Having an easier method to discover datasets in the data lake lowered the utilization of cluster resources in terms of compute and storage
- Lineage and audit capabilities allowed users to obtain authoritative answers to source, transform, and freshness of data, increasing transparency and their trust in the quality and nature of datasets
- Seamless integration with Cask Hydrator took them from the data discovery phase to ideation and pipeline creation in minutes, which previously required hours and sometimes days
- Ultimately IT spending became much more predictable and the collaboration between data engineers, data scientists, and data analysts improved

Cask Hydrator

Download CDAP today to try out Cask Hydrator: **cask.co/downloads**

Visit the Cask Hydrator product page: **cask.co/products/hydrator**

Or read more about Cask Hydrator: **cask.co/hydrator-wp**

Cask Tracker

Download CDAP today to try out Cask Tracker: **cask.co/downloads**

Visit the Cask Tracker product page: **cask.co/products/tracker**

650-469-DATA

Sales: sales@cask.co
Support: support@cask.co
General: info@cask.co

CASK

TAP IN @ CASK.CO

150 Grant Ave, Palo Alto, CA 94306