cloudera CONNECT

SOLUTION BRIEF

CASK

INDUSTRY

Unified integration for big data

WEBSITE

www.cask.co

COMPANY OVERVIEW

Cask makes building and running big data solutions on-premises or in the cloud easy with Cask Data Application Platform (CDAP), the first unified integration platform for big data. CDAP reduces the time to production for data hubs and data applications by 80%, empowering the business to make better decisions faster.

PRODUCT OVERVIEW

CDAP

CDAP provides a code-free and visual user experience for building complex data pipelines and managing them on your data hub. With CDAP, you can ingest data from varied sources, cleanse, normalize, and transform data, build machine learning models on-the-fly, perform aggregations, run custom scripts, and more.

SOLUTION HIGHLIGHTS

- CDAP offers self-service, drag & drop data integration that goes beyond ingest of streaming and batch data from varied sources. It also adds data preparation and data science (ML) for advanced data pipelines on CDH Hadoop and Spark.
- CDAP runs natively on Hadoop for seamless out-of-box scalability over CDH.
- CDAP provides application-level data discovery with metadata, audit and lineage. It seamlessly integrates with Cloudera Navigator.
- CDAP integrates seamlessly with Cloudera Sentry for authentication and fine grained access control.

Build, Manage and Automate Data Pipelines with Cask and Cloudera

Today, enterprises are turning to an enterprise data hub (EDH), powered by Apache Hadoop, as the core platform for delivering new analytic applications. Now Cask and Cloudera are bringing together the combined power of large scale processing, data integration and application lifecycle management to make it easier for developers and organizations to productionize data pipelines more efficiently on the EDH.

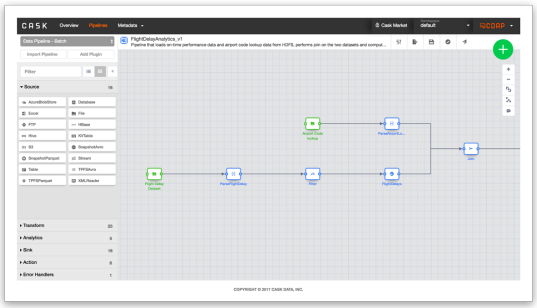
Easily Build and Run Self-Service Data Pipelines

Cask Data Application Platform (CDAP) is an interactive application for building, running and managing data pipelines on Hadoop and Apache Spark. It is 100% open source and licensed under the Apache 2.0 license. CDAP prepares, blends, aggregates and applies science to create a complete picture of your business data that drives actionable insights.

With visual tools to eliminate coding and complexity, CDAP puts big data at the fingertips of not only developers, but also of data scientists, citizen integrators and business analysts.

Integrate, Prepare and Blend

Ingest data in minutes from anywhere and any format without writing code. Prepare, cleanse, and enrich using built-in transformations. Blend data from traditional RDBMS to Data Warehouse to Hadoop.



Aggregate and Analyze

Perform step-by-step aggregation and analytics in batch or real time. Leverage state-of-art Spark ML for building models and scoring models in a unified environment without writing any code.

Automate and Operationalize

Use REST APIs or CLI tools for automating deployment and management of pipelines in different environments. Use the built-in enterprise scheduler to schedule pipelines, different notification mechanisms, aggregated pipeline logs and metrics, pipeline comparisons for diagnosing problems and version management of plugins.

Deploy, Audit and Govern

Deploy pipelines to be executed as MapReduce or Spark or Spark Streaming in case of real-time requirements. Catalog all of the datasets and metadata to support data governance. Secure your data with fine-grained access control, monitor and track user activities through audit logs.

Cloudera Enterprise Benefits

Stores and Analyzes Any Type of Data

- Leverage the full power of your data to achieve pervasive analytics, increase business visibility, and reduce costs
- Bring diverse users and application workloads to a single, unified pool of data on common infrastructure; no data movement required

Enterprise Approach

- Compliance-ready perimeter security, authentication, granular authorization, and data protection through encryption and key management
- Enterprise-grade data auditing, data lineage, and data discovery

Industry-Leading Management and Support

- Best-in-class holistic interface that provides end-to-end system management
- Open platform ensures easy integration with existing systems
- Open source to achieve stability, continuous innovation, and portability

CDAP Benefits

Accelerate Time to Value

- Rapidly deliver reliable and operational data hubs and production data apps faster and better.
- Extensible libraries and components promote reuse and further accelerate the pace of innovation.

Provide Self-Service

- Broaden the user base of your big data platform with a radically simplified developer experience and code-free extensions for non-developers.
- Reusable libraries can be assembled and run as data pipelines through drag-and-drop interfaces.

Enable Governance

- Automatic tracking of all audits and data lineage with discovery and search.
- Integrate into existing security and governance systems with authentication, authorization and audit built-in automatically.

Benefits of Cask Hydrator on an EDH

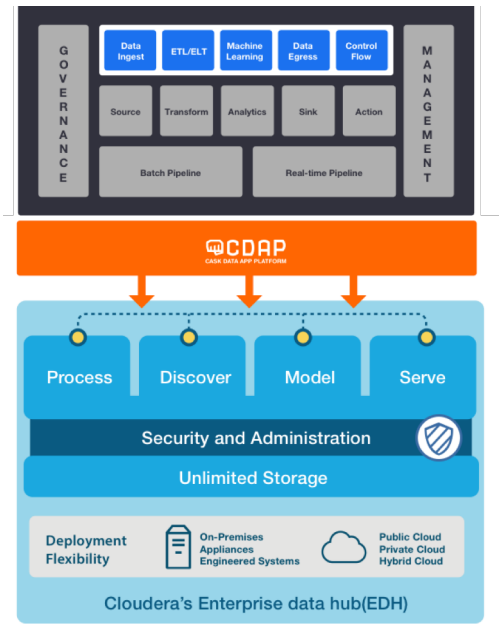
Unrivaled Ease of Use

CDAP provides intuitive drag-and-drop integrations with Hadoop and non-Hadoop storage and as well as the ability to switch between different processing technologies - MapReduce, Spark or Spark Streaming. It features:

- An interactive, drag&drop Studio environment to simplify the creation of data pipelines; a Preview mode allows users to debug pipeline before it is deployed to a cluster
- Rich library of pre-built plugins to access, transform, blend, aggregate data from relational sources, NoSQL sources and more; native support for AVRO, Parquet and HBase
- Support for fast lookups within transforms allowing users to create secondary keys during processing
- Powerful orchestration capabilities to coordinate batch and real-time pipelines, combined with notification and alerting capabilities to monitor the workflows
- Integrated enterprise scheduler for coordinating jobs within the workflows and ability to test and tune job executions

Zero-Coding Integration, Aggregation and Analytics

The intuitive interface of CDAP accelerates the design and deployment of big data analytics by up to 5 times compared to hand-coded systems:



- Complete visual integration eliminates manual programming and scripting from the process
- Streamlines analytical processes and eliminates the need for manual steps or specialized resources
- Support for Control Flow combined with Data Flow; this makes it easy to specify custom actions (such as performing database bulk export or import)

Self-Service Environment for Broad Adoption

CDAP delivers governed, best-practice, on-demand data to data scientists, data engineers, and business analysts in an agile fashion.

- Seamless self-service for transforming, aggregating and enriching large scale/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, 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
- Integration with advanced analytics such as Spark ML to operationalize predictive intelligence while reducing the build time

Native to Hadoop and Enterprise-Ready

Go beyond data ingestion to scalable and flexible management for end-to-end data pipelines with enterprise-grade capabilities:

- CDAP runs natively on Hadoop offering automatic recovery/fault tolerance and seamless out-of-box scalability
- Robust administration features including SLA monitoring, job restart, error handling and restart, and an operations central for auditing access
- Enterprise-grade security including access and version controls as well as LDAP, JSAPI, Active Directory and Apache Sentry integration
- Enhanced Data Management integration for tracking data, metadata and usage analytics

About Cask

Cask makes building and running big data solutions on-premises or in the cloud easy with Cask Data Application Platform (CDAP), the first unified integration platform for big data. CDAP reduces the time to production for data lakes and data applications by 80%, empowering the business to make better decisions faster. It lets developers, architects and data scientists focus on applications and insights rather than infrastructure and integration. CDAP accelerates time to value from Hadoop through standardized APIs, configurable templates and visual interfaces. It enables IT organizations to broaden the big data user base within the enterprise with a radically simplified developer experience and a code-free self-service environment. CDAP is 100% open source, and along with its extensions Cask Hydrator for data pipelines and Cask Tracker for data discovery and metadata, it seamlessly integrates with existing MDM, BI and security and governance solutions. Cask customers and partners include AT&T, Cloudera, Ericsson, Lotame, Microsoft, Salesforce, and Tableau, among others. For more information, visit the Cask website at cask.co and follow @caskdata.

About Cloudera

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data: The Enterprise Data Hub. Cloudera offers enterprises one place to store, process and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Founded in 2008, Cloudera was the first and is still today the leading provider and supporter of Hadoop for the enterprise. Cloudera also offers software for business critical data challenges including storage, access, management, analysis, security and search. With over 15,000 individuals trained, Cloudera is a leading educator of data professionals, offering the industry's broadest array of Hadoop training and certification programs. Cloudera works with over 700 hardware, software and services partners to meet customers' big data goals. Leading organizations in every industry run Cloudera in production, including finance, telecommunications, retail, internet, utilities, oil and gas, healthcare, biopharmaceuticals, networking and media, plus top public sector organizations globally. www.cloudera.com.