cloudera®

SOLUTION BRIEF

CASK

## 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
Cask Hydrator is a code-free visual extension of CDAP for building complex data pipelines and managing them on your data hub. With Cask Hydrator, 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
- Cask Hydrator is a self-service, drag-and-drop CDAP extension that goes beyond mere data ingest, adding data science (ML), and data wrangling for advanced data pipelines on CDH Hadoop and Spark.
- Cask Hydrator runs natively on Hadoop for seamless scalability over CDH.
- Cask Hydrator supports combining Control Flow and Data Flow; this makes it easy to specify custom actions.
- Cask Hydrator integrates with Cask Tracker to provide application-level data discovery with metadata, audit, and lineage; offers seamless integration with Cloudera Navigator.
- Cask Hydrator integrates seamlessly with Cloudera Sentry for authentication and fine-grained access control of pipelines and data sets.

# Cloudera and Cask Hydrator: From Data Ingest to Machine Learning to Operational Analytics Solutions

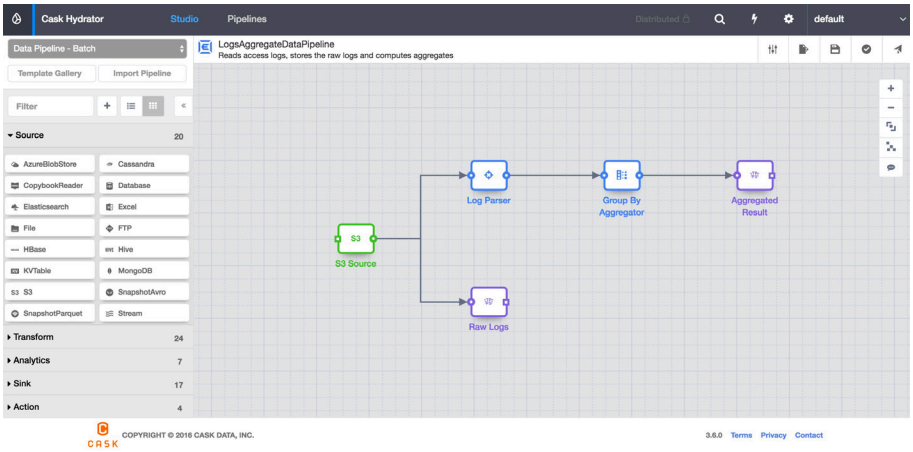
Today, enterprises are turning to an enterprise data hub (EDH), powered by Apache Hadoop, as the core platform for delivering new analytic applications. Now Cask and Cloudera are bringing together the combined power of large-scale processing, data integration, and application lifecycle management to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

## Easily Build and Run Self-Service Data Pipelines

Cask Hydrator is an interactive application for building, running, and managing data pipelines on Hadoop and Apache Spark. It is 100 percent open source and licensed under the Apache 2.0 license. Cask Hydrator 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, Hydrator 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 in 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-the-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, along with pipeline comparisons for diagnosing problems, and version management of plug-ins.

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

### Cloudera for IoT

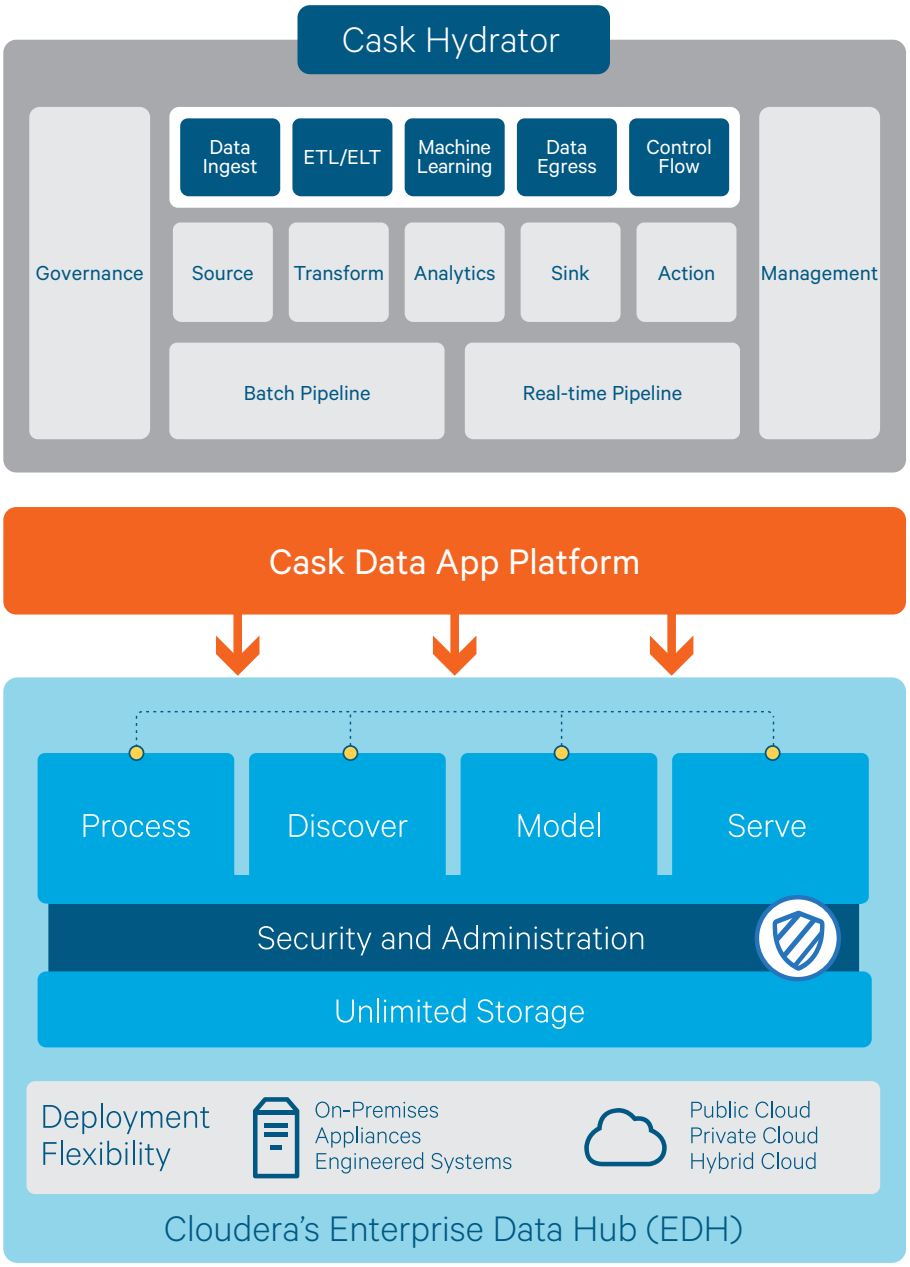
- Effectively handle both data at rest and data in motion
- Easily ingest millions of events/sec
- Industry leadership in Spark
- Real-time processing and analytics
- Hybrid Cloud deployments
- Effectively combine sensor data with other internal and external sources
- Data security beyond compromise
- Proven success across diverse IoT use cases

## Benefits of Cask Hydrator on an EDH

### Unrivaled Ease of Use

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

- Cask Hydrator Studio to simplify the creation of data pipelines; a Preview mode allows users to debug a pipeline before it is deployed to a cluster
- A rich library of pre-built plug-ins to access, transform, blend, and aggregate data from relational sources, NoSQL sources and more; native support for AVRO, Parquet and HBase




- Support for fast lookups within transforms allows 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 Cask Hydrator accelerates the design and deployment of big data analytics by up to five 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 makes it easy to specify custom actions (such as performing database bulk export or import)

### Self-Service Environment for Broad Adoption

Cask Hydrator delivers governed, best-practice, on-demand data to data scientists, data engineers, analysts, and end users 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 data sets to drive faster and more reliable analytics
- Seamless integration with visualization and data services, making data sets 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:

- Cask Hydrator runs natively on Hadoop, offering automatic recovery/fault tolerance and seamless out-of-the-box scalability
- Robust administration features include 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 through integration with Cask Tracker to track 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 percent, 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 percent 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, Salesforce, and Tableau, among others. For more information, visit the Cask website at cask.co and follow @caskdata.

cloudera®

## About Cloudera

Cloudera delivers the modern platform for data management and analytics. The world's leading organizations trust Cloudera to help solve their most challenging business problems with Cloudera Enterprise, the fastest, easiest, and most secure data platform built on Apache Hadoop. Our customers can efficiently capture, store, process, and analyze vast amounts of data, empowering them to use advanced analytics to drive business decisions quickly, flexibly, and at lower cost than has been possible before. To ensure our customers are successful, we offer comprehensive support, training, and professional services. Learn more at cloudera.com.

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304, USA



Cloudera_Cask_Hydrator_Solution-Brief_101