# CDAP Solutions

Cask Data Application Platform (CDAP)

The first unified integration platform for big data, Cask Data Application Platform (CDAP) 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.

Real-Time Analytics and IoT

Businesses are moving toward faster, enhanced integrations with real-time data streams in order to make better decisions and gain a competitive advantage. To analyze insight, retailers process sales; fleet management companies track fuel, mileage, speed; and media houses examine social media. But real-time analytics is challenging— it is dynamic analysis and reporting based on data entered into a system less than one minute before the actual time of use. IoT data can arrive from a variety of sources in an inconsistent array of formats, and video streams take an entirely different format than sensor data, yet must all be processed together. Real-time analytics is complex, costly and time-intensive.

CDAP provides a single platform for building both real-time and batch analytics applications with reliability and consistency. With its drag-and-drop interface, Cask Hydrator makes it easy to combine CDAP with the latest technologies, such as Spark Streaming for simplified data loading and transformation, to develop real-time analytics and IoT solutions. In addition, integration with Apache Sentry provides enterprise-grade security and data governance, ensuring that confidential user and sensor data is kept secure.
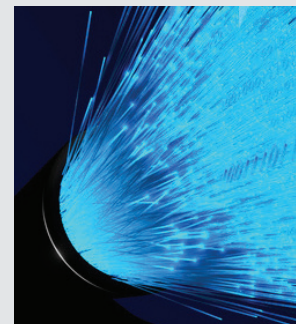
CDAP for Spark

With so much data being processed by enterprises everyday, it's essential to stream and analyze it in real-time. Apache Spark provides a framework for advanced analytics which includes a tool for accelerated queries, a machine learning library and a streaming analytics engine. However, Spark lacks many tools and reusable components as well as built-in data governance. Native Spark is also unable to support globally consistent transactions, which prevents use cases where time-ordered writes to the database are critical. Finally, Spark does not aggregate logs in a central location, which means developers must dig through each node to debug applications.

CDAP simplifies building and debugging a Spark application by offering reusable building blocks, aggregating Spark logs and making them viewable in real time. It runs each process in a container, allowing developers to run Spark and Hadoop processes in parallel on the same cluster. Packaged into CDAP, Apache Tephra is a transaction engine that supports multi-versioning and rollback, which provides globally consistent transactions on top of Spark. In addition, Cask Tracker provides a code-free governance solution for Spark applications, providing self-service access to metadata, audit trail, and data lineage analysis.

EDW Offload

The cost of maintaining a traditional Enterprise Data Warehouse (EDW) is skyrocketing as legacy systems buckle under the weight of exponentially growing data and increasingly complex processing needs. For enterprises, Hadoop significantly reduces storage and compute costs, freeing up processing times and storage in the data warehouse and enabling IT organizations to successfully meet SLAs.

But most businesses still lack the skill sets required to deploy Hadoop. Offloading EDW data into Hadoop requires developers to write complex integrations, taking resources away from building business applications. Enterprises must offload ETL processing tasks with solutions that empower developers who do not have Hadoop skills.

Configurable pipelines in Cask Hydrator save developers the complex and time-consuming process of having to write custom data pipelines from scratch, enabling quick deployment of EDW offload solutions. It also provides a graphical drag-and-drop interface for simplified pipeline building from EDW into Hadoop. In addition, Cask Tracker provides metadata audit, which makes it easy to track data flows and retrieve, use and manage datasets.

Data Lake
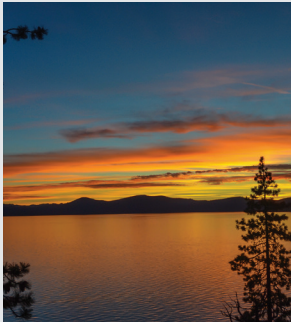
While an enterprise data lake provides a centralized data repository and makes it easy for data scientists, developers and analysts to process data to gain insights, it must support varied data types and formats - structured, unstructured and semi-structured. Other key challenges include establishing SLAs for ingestion, continually ensuring data quality and proper transport, and creating appropriate retention policies. The separation of concerns (data ingest teams, LOBs) is another important consideration, as well as the ability to guarantee a well-governed self-service environment.

CDAP speeds up the process of building and running a data lake with its unified platform of integrated application management, data integration, security and governance. It makes all data in Hadoop available for access in real-time and batch and provides a collection of pre-built building blocks to support data manipulation, data storage and key insight extraction to build smarter, more flexible end-to-end solutions. Cask Hydrator provides a rich, drag-and-drop user interface for data ingestion, exploration and transformation capabilities. Cask Tracker enables self-service data discovery with automatic metadata, audit and lineage across all databases, connectors and pipelines.

Customer 360

As enterprises introduce mobile apps, online communities, chat, social media and more into business, it's important to analyze how customers use these platforms. Each click provides behavior insight, enabling companies to analyze user buying patterns, enhance offerings, improve marketing, conduct predictive modeling and ultimately sell more. Developing this Customer 360 view relies on tools like social media listening, predictive analytics, CRM suites, marketing automation, etc.

Web analytics takes time to build, requiring multiple steps across numerous sources. For instance, aggregating user data with historical customer demographics and other recorded data is a challenge. In addition, data entry, quality, management and merging is crucial, as well as properly dealing with human errors, fuzzy matches, and incomplete records.

Cask simplifies building end-to-end data pipelines, including ingesting, blending, and aggregating data from varied source feeds. CDAP provides a simple and consistent platform to build, deploy and manage complex data analytics applications in the cloud or on-premise.

About Cask

Cask makes building and running big data solutions on-premise 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. 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.