Cask: Curbing the Complexity of Big Data

by Robin Bloor & Rebecca Jozwiak

The Painful Evolution of Hadoop

Early adopters of new technology environments are usually savvy enough to tame the rough edges of the platform. We've witnessed this time and again throughout the computer era. For example, the IBM mainframe (OS370), Unix and Windows NT were relatively immature in early releases. However, in those days, IBM, Sun Microsystems and Microsoft were quick to evolve their platforms, and third party vendors were quick to assist.

We see the rough edges with Hadoop, although sadly, the situation is more confused. Hadoop is not a fully fledged or well-integrated OS, rather it is a distributed environment made up of many open source infrastructure projects for managing big data and running big data applications, and it includes a unique file system and scheduling capability. Nevertheless, the main Hadoop distribution companies, Cloudera, Hortonworks and MapR, offer distinctly different software stacks and have distinctly different visions of how they want Hadoop to evolve.

Cloudera's management software is Cloudera Manager. They offer Impala as a SQL database and provide Cloudera Search as a search capability, all of which are proprietary. Hortonworks provides Ambari for management, Stinger for queries and Apache Solr for search. MapR offers its own proprietary file system, MapR-FS, its own database, MapR-DB, and its own streaming capability, MapR Streams. Hortonworks thinks of itself as a pure play Hadoop distro, Cloudera aspires to become an "enterprise data hub" and MapR has its heart set on being a distributed "converged data platform."

For the technology user, such platform divergence is uncomfortable enough, but it doesn't stop there. Aside from these three main distributions, the cloud vendors (Amazon, Microsoft, etc.) have their own distributions tailored to their cloud environments. If you include Spark in the equation – and you should because Spark's popularity is not far behind that of Hadoop – there's yet more diversity. While Spark can be implemented on Hadoop's HDFS, in Amazon's EMR, its data store is Amazon's S3 object store and Azure leverages the Azure Blob Store or the Azure Data Lake.

Big Data is by nature diverse and complex, and a typical big data solution will need to combine not only different datasets, but may also require using multiple storage and processing engines. Building such solutions typically requires the "stitching together" of disparate systems, which creates complexity and inefficiency. This is a contrast to the needs of enterprise IT for simplified big data management solutions, which allow them to focus on application logic and insights, rather than infrastructure and integration.

Towards the Data Lake?

Despite such divergence, many businesses have used Hadoop and Spark to build data lakes – common staging areas for data ingest, data governance and analytics applications. Companies, even those with data warehouses, now see data lakes as the natural foundation for big data applications.

Data lakes are, in practice, more complex environments than data warehouses. They cater to both structured and unstructured data, link to both external and internal data sources and feed from real-time data streams as well as batch file and table transfers. Effective data ingest requires an ability to connect to many different data sources (flat files, log files, databases of any kind, data streams, IoT devices). Effective data governance involves a

whole host of activities: data and access security, metadata discovery and management, MDM, data provenance and lineage, ETL and data lifecycle management.

A data lake is a complex processing environment that can include many capabilities aside from the analytic applications it runs. And yet, most IT Departments using Hadoop and Spark build DIY infrastructures to deploy and maintain their applications. The situation is exacerbated by the fact that Hadoop and Spark expertise is expensive and in short supply.

"CDAP's flexible architecture significantly reduces or eliminates the technical skills required to integrate Hadoop and Spark components and deploy data lake applications."

Furthermore, there is no Hadoop standard, and not all of its components are well integrated. The same can be said of Spark. There is no guarantee that any of the primary distros (or cloud vendors) will deliver the software architecture and software integration required to remedy the situation any time soon, if ever.

It is specifically to address this problem that Cask created CDAP, a software development and production environment for Hadoop and Spark.

The Cask Data Application Platform (CDAP)

CDAP is a unified integration platform for Hadoop and Spark environments and is 100% open source. It can be used to enable enterprise data lake environments or to build full-fledged, closed-loop data applications, embedding analytics or BI and using the ever expanding array of big data components. It provides a remarkably comprehensive capability, addressing both the technical and business issues that plague data lake projects.

Its most compelling capability, in our view, is that it can automatically manage (and hide) the technical complexity of a data lake environment. It does this by implementing a containerized architecture. The three key components of a CDAP application or data lake – datasets, programs and applications - are deployed as containers. This has the double virtue of technically isolating each component and providing an extremely versatile deployment capability. So, for example, CDAP and its applications can operate within a Hadoop environment on a physical cluster or a virtual cluster, in the cloud or in the data center.

In practice, CDAP customers have been able to move applications in and out of the cloud or upgrade Hadoop deployments from one release to another, even from one distro to another. CDAP's flexible architecture significantly reduces or eliminates the technical skills required to integrate Hadoop and Spark components and deploy data lake applications.

As indicated in Figure 1, CDAP currently offers support for the Cloudera, Hortonworks and MapR distributions as well as the Microsoft, Google and Amazon Hadoop cloud environments. It also offers support for IBM. With CDAP, it is possible to move applications or even a whole data lake from one to another.

Figure 1 also depicts a prominent subset of the big data components that CDAP supports: the Hive and HBase data stores, the YARN scheduling component, the Kafka messaging system and the whole Spark development and deployment environment. Cask is committed to adding other components as new ones emerge. For example, if Flink continues to increase in popularity, it will be added in. The goal is to maintain a full open source capability across all distros.

CDAP also provides a Software Development Kit (SDK) that can run on a laptop or

Cask: Curbing the Complexity of Big Data

"Perhaps the most compelling aspect of this is that the movement from application development (after the completion of testing) to full-scale deployment is remarkably swift..."

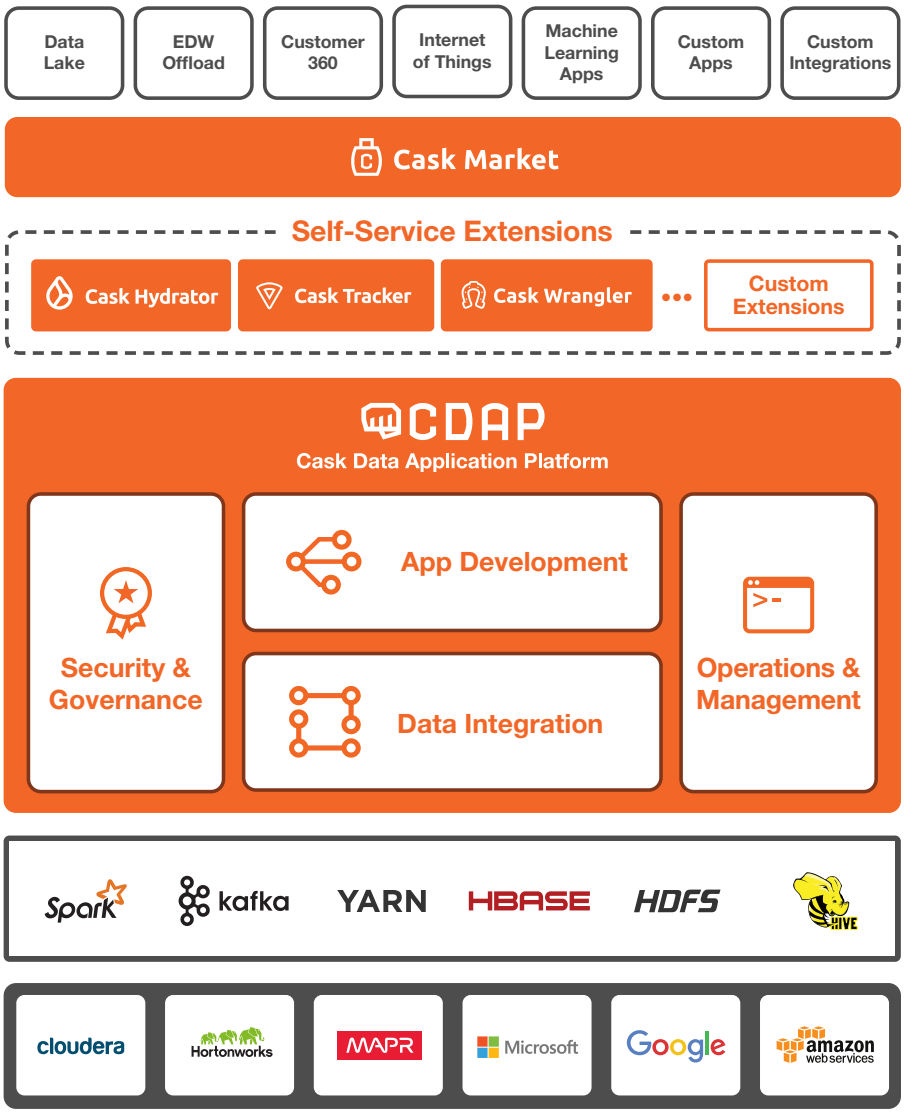

Figure 1. Cask Architecture Overview

workstation (Linux, Windows or Mac). It is a standalone CDAP instance that runs in a single JVM or Docker container and includes all of the CDAP APIs. It implements all CDAP capabilities without any need for a local Hadoop instance. It includes the CDAP web-based UI, which interacts with CDAP instances and their applications, and it also provides tools for data ingest and client authentication. The Java, Scala, Python and Ruby programming languages are supported, as are both the IntelliJ and Eclipse IDEs.

Cask provides a powerful application development and deployment environment for building and running big data solutions. Perhaps its most compelling capability is that the progress from application development (after the completion of testing) to full-scale deployment is remarkably swift, due to a set of operational tools and management capabilities designed for easy deployment of distributed applications. Typically, with Hadoop and Spark projects, deployment can be more than 50 percent of the work. With CDAP, it is almost effortless, due to its flexible and easily configurable architecture and built in functionality that is purpose-designed for data lakes and modern data applications.

Regarding security, CDAP supports Kerberos and easily integrates with encryption software and authentication software, including Apache Sentry and Apache Ranger. It comes with impressive data governance capabilities: it can identify, capture and store metadata; record lineage; and enable the analysis of data usage and usage patterns. These features require little work, as they are baked into the platform.

Another impressive aspect of CDAP is its data integration capability. Data can be ingested from any source on a scheduled basis or as a real-time feed. Metadata is captured on ingest, and thus ETL activity can easily be specified to serve data to applications or to transfer it to other destinations, including other CDAP data lakes or applications.

"Data scientists and business analysts can explore data and build applications with little or no IT department assistance."

CDAP Extensions and Cask Market

In the newest, just released version 4 of CDAP, three specific extensions, Cask Hydrator Cask Tracker, and Cask Wrangler, provide visual, self-service functionality within the CDAP environment. Hydrator is a code-free visual design and management tool that is used to create and run data pipelines, even very complex ones, using a drag and drop graphical interface. This is an intuitive tool that allows users to select data sources, including data already in the data lake, within legacy systems, cloud systems or applications, and build data pipelines. Data can be ingested into the data lake, blended with other data, indexed, tracked and delivered to its chosen destination. The data pipeline can be tested and debugged before being scheduled and executed.

As you may have assumed, Hydrator scales naturally. The user's pipeline design is a logical specification that maps to the execution environment, typically Spark. So it is easy to add more computer resource to improve performance or to switch the pipeline implementation from, say, Hadoop to Spark. Adding a new data source to a data pipeline is a relatively simple drag and drop activity. Hydrator also provides performance statistics so the performance of a pipeline can be monitored.

Hydrator delivers a powerful self-service experience that has pleased and surprised some of Cask's customers. It isn't just that building data pipelines is remarkably easy – Hydrator changes how users perceive and utilize data. Data scientists and business analysts can explore data and build applications with little or no IT department assistance.

Tracker is the second self-service extension. It provides a complete view of all data resources, making it easy to capture metadata about data sets and programs. It tracks data lineage, providing a full audit trail of what data is used where, and it captures all access patterns. It also provides the ability to search data sets based on user properties, system properties and business tags, and it can be integrated with third party MDM solutions. Tracker's usage analytics provide a view of every data set within the data lake

The third and newest extension is Cask Wrangler and makes data preparation much more interactive and intuitive within CDAP. Wrangler can be used as a standalone tool and is also directly integrated with Hydrator to allow for any preparation steps to be performed as part of production pipelines.

In addition to these three extensions, CDAP 4 also includes an app store for big data called "Cask Market." Cask Market provides an application, data and library ecosystem with pre-built Hadoop solutions, reusable templates, and third-party plugins. It is available from anywhere within the CDAP UI with a mouse click, and it allows users even with no prior experience to take advantage of the pre-built pipelines, components and solutions.

The Bottom Line

Technically, there is a great deal more to be said about CDAP than can be squeezed into a short paper. The bottom line is that it provides an extremely capable ready-made environment for building data lakes and data applications. While Hadoop may be a bewildering collection of diverse software components, CDAP is a coherent platform that allows users to leverage such components as reusable building blocks.

Additionally, it caters to security and data governance, and it provides a remarkably high level of data integration capability, which in turn makes the CDAP environment eminently suited for self-service. With the built-in Hydrator, Wrangler and Tracker extensions providing a fast and efficient way to build data pipelines, prepare data and track data lineage, the platform delivers an impressive capability that will please both the data analyst and data scientist. And Cask Market makes it extremely easy even for new users to build and deploy modern big data solutions on Hadoop and Spark.

CDAP removes a great deal of the effort and pain involved in building data lake environments and big data applications. You can think of CDAP as the operating platform and development environment that Hadoop should have been but never was. We advise companies that are considering building a data lake environment, or who have already built one and have become frustrated with the technical complexity of the task, to take a look at CDAP.

Company: Cask
Location: Palo Alto, CA

Product: CDAP
Big data application platform