White Paper

Unified Integration:
The Key to Rapid Time to Value from Big Data

Executive Summary

The era of digital transformation and IoT is driving the necessity for modern data management, application management and IT automation as a cohesive, integrated set of capabilities. In this paper we will outline the need, challenges and solution for a unified integration platform for big data that enables a cohesive set of capabilities to deliver business agility.

INTRODUCTION

"… 100% of all large enterprises will adopt Hadoop and related technologies, such as Spark, for big data analytics within the next two years …"

-- Forrester

As organizations continue to take the data journey from structured data only databases to data lakes and connected data fabrics, a shift to more user-centric and application-centric systems has occurred. Business leaders are no longer only relying on fact-based decision-making and information analysis for their competitive advantage, but are increasingly looking to power the business with analytical applications that can drive actions in addition to generating insights. Advances in technology continue to accelerate the pace and competitive environment of business. Those organizations that are able to quickly gain insights from their data and take action through a new generation of data applications are leading their respective industries.

Apache Hadoop, Apache Spark and related technologies have been leading the big data world with their innovation in scalability, better cost efficiency in terms of storage and compute, and unprecedented ability to handle large volume, high velocity and wide variety of data. For enterprises determined to enable their technical teams to efficiently build and deploy data lakes and data applications in Apache Hadoop or Apache Spark, the next step is to adopt a big data application lifecycle approach, which merges and aligns their data integration efforts with application management as well as a secure, governed, and enterprise-ready IT operations environment.

Powered by Hadoop and Spark, a unified integration platform for big data offers a single, consistent devops framework with standardizations and pre-built integrations simplifying the combination of data storage, business logic and compute. Designed properly, it allows for powerful, self-service, flexible, secured, governed and future-proof ways of building and managing data lakes and data applications on premises, in the cloud or in hybrid environments. Yet, it is of course fair to ask the question how a unified integration platform for big data fits into an environment with pre-existing investments in Hadoop- or Spark-based data management and compute architectures.

This paper examines the role of Cask Data Application Platform (CDAP), Cask's implementation of a unified integration platform for big data as a key enabler for building and managing modern data lakes and data applications. It also covers the design principles of a unified integration platform, provides an overview of its technical components, and highlights some of the capabilities that are critical for enterprise-class deployments.

GETTING VALUE FROM BIG DATA IS HARD

"… Through 2018, 70% of Hadoop installations will fail to meet goals for cost savings and revenue generation objectives due to skills and integration challenges …"

-- Gartner

Traditional application and data management architectures have followed three-tier architectures. IT teams have made varied attempts in the past to support big data applications with similar architectures in mind, but have had challenges to create a consistent, reliable enterprise architecture. The attempts have been hindered by the very nature of big data technologies, which are often disparate and purpose-built to solve a particular technology problem, rather than a set of business problems.

Enterprises don't have a shortage of talent. But in the world of big data, the collaborative effort needed across different parts of the IT organization - developers, data scientists and IT ops – to build end-to-end solutions is significant. What is more, while often built *for*, but not *with* less technical users – such as citizen developers, citizen integrators and line of business analysts – in mind, big data solutions often fail to benefit the very people who have the most amount of domain knowledge to extract value from big data. A platform, which can actually bridge the IT/Line of Business gap by combining the required platform capabilities with necessary product usability, can remove major hurdles for big data users, helping to extract value from big data projects and unburden IT.

Complexity and Proliferation of Technology

The wide range of options of technologies and siloed vendor offerings in the big data ecosystem notably creates non-repeatable and non-reusable solutions, which is an impediment to adoption and democratization of big data for many enterprises. Often specialized skills are required to build and operate big data projects end-to-end. As organizations look to leverage more, changing and disparate data types and processing paradigms, the cost and effort to rationalize, adopt and maintain different technologies or vendor solutions becomes prohibitive.

Big data projects require significant scoping, planning, testing and deployment cycles, and a dedicated team – even more so if they are not part of a well-defined process that IT teams can follow. Heroics, not blueprints, are unfortunately still often enough common place to advance big data projects; even then, they often don't deliver any real business value or ROI. If not dealt with head-on, fragmented technology choices, the emerging disaggregation of Hadoop stacks and the lack of re-usability has the potential to lead to unacceptably long project cycles and ultimate failure of big data projects affecting business users or end customers.

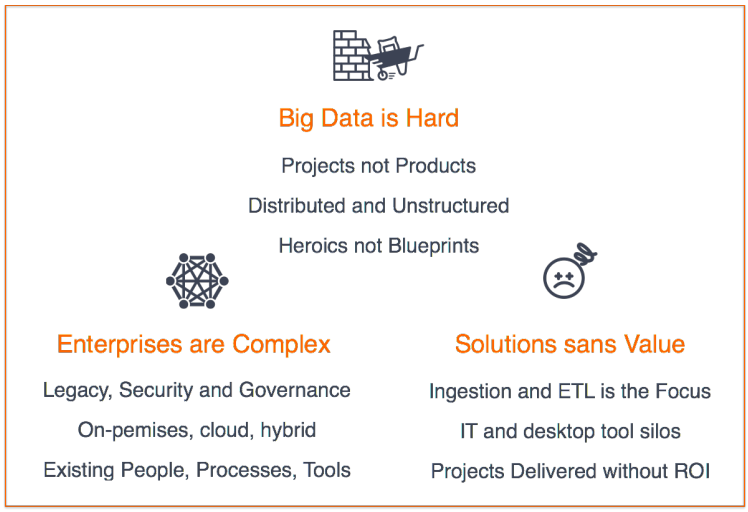

Figure 1: Big Data Challenges

Lack of Reusability

> "… Pilots and experiments are built with ad hoc technologies and infrastructure, and not created with production-level reliability in mind …"
>
> -- Gartner

While data management and processing technologies have become more fragmented in the last several years, the need for more advanced, more comprehensive analytical solutions (IoT, AI, rule-based machine learning) has grown. Continued reliance on ground-up hand-coding — whether across the enterprise or in isolated pockets of the organization — has exacerbated the problem by inhibiting reusability and draining valuable IT resources

Analytical (data) applications are taking too long to build and deploy, because of the growing complexity inherent in first-generation data architectures, which lack reusability of data integration patterns and best practices.

Complex Governance and Compliance

Most of the big data technologies start out as open-source or are open-sourced later in their life-times. This means that current enterprise big data solutions are often built using a collection of fragmented technologies that pose an undetermined, but real threat to user compliance with both internal and external rules of data access, use, distribution and alteration.

Loosely coupled Integration

Despite years of building and refining big data, IT teams still continue to struggle with high costs, delays, and suboptimal results due to the need to stitch loosely coupled big data technologies together. Complexity is a key culprit. The ceaseless introduction of new technologies — at breathtaking speed — has meant that IT professionals are perpetually taking one step forward and two back as they spend a large part of their budget on integrations of loosely coupled technologies.

Lack of Self-Service

For a long time, business users have been pressuring IT to reduce their dependency on IT to get access to the data lake and be able to see data prepared the way they need them to be. Most enterprise data lakes lack an easy-to-use environment for transforming, managing and presenting data in a business-friendly way that allows for self- service data access, exploration, and analysis.

Furthermore, lack of support for modern master data management disciplines and control, which meet IT requirements for reliable, high-quality data, affect the adoption, democratization of big data.

Pressure is mounting on IT teams to expand the scope of big data beyond its roots in just being a data lake to encompass more data insights, techniques and business areas. Growth in adoption of big data technologies, as well as growth in cloud-based data systems, poses an additional challenge in managing hybrid data environments. Dramatic changes in data volume, variety, and velocity make the traditional approach to data integration inadequate and require you to evolve to next-generation techniques in order to unlock the potential of data.

The rise of Apache Hadoop, Apache Spark and other big data technologies, and the recognition of the value and complexity they bring to the enterprise, demand a modern, unified integration platform for big data. It brings simplicity and self-service to data integration and application lifecycle management while reducing cost and inefficiencies of deploying and operating data lakes and data applications.

Anatomy of a Unified Integration Platform for Big Data

The foundation of a unified integration platform is a framework of abstraction and its associated runtime components. But the dynamics that embody big data - volume, variety, and velocity – also demand that a modern unified integration platform adapts well beyond data movement and operational data analytics, and easily accommodates ever-changing data formats in addition to the explosive growth of the amount of data. An example of the challenge presented by big data is the complexity of integrating disparate systems when building operational data applications.

Giving technical users a consistent, unified work environment to build integrated operational and analytic functions in a single data application platform helps breaking down traditional silos and achieving higher ROI.

As developers and architects seek to employ more technologies to meet demands for data driven applications, the complexity of disparate systems will only exacerbate the fundamental problem of building applications for big data. Enterprises seeking to realize the full potential of a unified integration platform for big data – such as Cask Data Application Platform (CDAP) - as the means to building data applications need a single, cohesive a system that addresses the following considerations:

Developer Productivity and Agility

Since technical resources are typically the most precious resources in an enterprise, a unified integration platform needs to help optimize developer productivity time. The platform should provide simplicity through dependency injections, standardized, pre-built patterns, and foundational support for working with Hadoop and Spark systems. The platform should be concise, but not obtuse to the developers working on building analytical applications. It should provide a collection of easy-to-use tools that makes the feedback loops short.

Containerization

Data application containers hold important components such as libraries for manipulating data, preferences, configurations, files, user code for processing data or user requests. Because resources are shared in this way, application containers can be created that place less strain on the overall resources available. Portability is also a benefit. As long as Hadoop infrastructure components are identical across systems, an application container technology can run on any system on-premises, in the cloud or in a hybrid environment, without requiring code changes, and offering maximum deployment flexibility to the enterprise.

Security

A unified integration platform for big data also should reduce the chance of malicious or accidental actions outside of the designed usage of the system, and prevent disclosure or loss of information. Technical or business users should be provided a seamless way for integrating different security components into the application they build. They should not be concerned with meeting security standards and should also have simple security best practices to follow.

Flexibility

The unified integration platform should be able to support building and managing a wide variety of application types, namely real-time and batch. It should also support a wide variety of solutions such as network analytics, sentiment analysis, customer 360, etc., combined with ability to provide abstractions to store any type of – structured and unstructured - data, while simultaneously allowing for various types of processing against the same data, regardless of structure.

Application and Dataset Lifecycle Management

The unified integration platform should provide a simpler and easier way to begin the process of creation of an application, and then proceed to testing, deployment and production. The platform should provide tools for all aspects of the lifecycle of application and dataset creation and management. As lifecycle management is a continuous process starting from creation to its retirement, it should support versioning, upgrading or downgrading, replication and resilience.

Testability

The unified platform should support easy creation and execution of test criteria. It should support basic unit tests versus functional tests for testing the end-to-end scenarios of the data application. It should support the ability to trigger backend processing through simple scripting. The platform should make it easy to integrate tests as part of the continuous integration process.

Interoperability

The unified platform should employ standard communication protocols, interfaces and data formats making it easy to interoperate with 3rd party systems. The platform should use open standards where available, and suggest standards where not available. It should make it easy for selecting the 3rd party systems to work within a complex enterprise environment.

Performance and Scalability

The unified platform should add no or very little overhead in terms of processing or accessing data from the underlying systems. It should also allow the system

to expand from terabytes (TB) to petabytes (PB) by simply scaling the underlying infrastructure without needing to tune the platform. In the same way, the unified platform should scale as the underlying system scales in terms of number of nodes it is distributed on.

Reliability and Availability

A unified integration platform for big data should be reliable in its ability as system to continue operating in the expected way over long period of times. It should not present a single point of failure. It should be highly resilient to upgrades, downgrades or failures of infrastructure components. Applications running on top of the unified platform should be able to continue to run, even when parts of the infrastructure services are unavailable. In case of hard failures, the applications should have the ability to gracefully shut down.

Ease-of-use

The graphical user interface of a modern data application is considered easy and productive when it is based on intuitive click-and-drag methods coupled with visualizations. It will enable technical and business users to explore data interdependently without requiring the technical skills of technical data analysts and data scientists.

Automation

The unified integration platform should support modern master data management techniques and controls that meet IT requirements for reliable, high-quality data. These should also be able to handle big data scale through elastic computing resources.

Microservices

Microservices break up a monolithic application into more finely grained, more manageable and economically optimized components. They can help streamline the development process, making testing easier and faster, and they also provide a high degree of composition, which is one of the keys to more flexible application development, deployment and code reuse.

Data-Driven Applications are the new frontiers of Big Data allowing business to explore and augment their existing business strategies with new and valuable data insights. The challenges presented by the technologies should be mitigated by using standard unified framework for defining and executing business requirements.

Driving Business Value with Modern, Data-Driven Applications

A unified integration platform powered by big data technologies such as Hadoop and Spark, such as Cask Data Application Platform (CDAP), is uniquely capable of meeting the current and future challenges of building and managing complex data applications. It does that by employing a fundamentally different approach that simplifies the process of building and deploying applications, compared with the conventional way of using lower level APIs and traditional application architectures. The core tenant of this approach is that big data puts data at the center of any architecture, and data applications become a way of processing, transforming or moving data. This is very different from how traditional architectures for data systems are typically perceived.

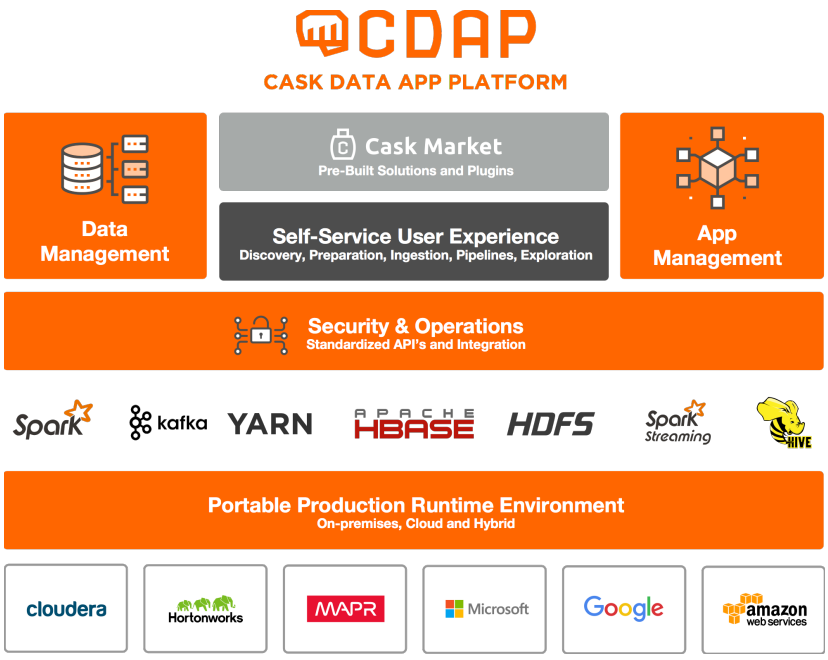

Figure 2: CDAP – The Unified Integration Platform for Big Data

Accomplishing this shift in data applications requires a comprehensive end-to-end architecture where the unified integration platform and the big data frameworks are tightly co-designed. This strategy provides a number of key capabilities that meet the demands of next generation data applications within the enterprise.

Integrated Compute and Storage

Storage (CDAP Dataset) and Compute (CDAP Program) frameworks are the core programmatic abstractions that are tightly integrated into a unified framework through a high-level notion of an "Application". "Datasets" and "Programs" are both collocated within the cluster nodes to maintain the throughput of any workload within the system, but the benefit of building applications using these higher-level frameworks make it easily to develop, manage and future-proof enterprise data applications. Thus, this design benefits all types of applications and workloads, which a business user might require for their solution - from batch-processing programs with MapReduce or Spark to real-time processing programs with Spark Streaming or Tigon, to Workflows, and on to representing more complex data patterns through Datasets like time series, OLAP Cubes and geo spatial datasets.

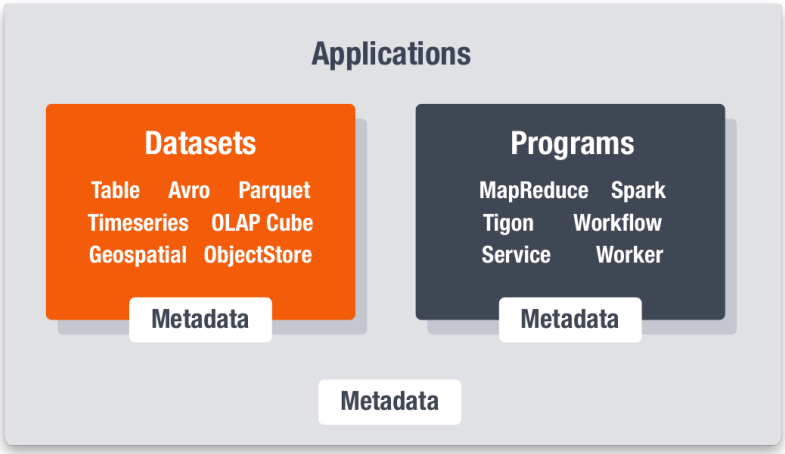

Figure 3: Cask Data Application Platform (CDAP) Architecture

Flexible Data Patterns

Enterprises have to deal with more data and complex data with greater agility and cost-saving performance. Similar to application architectural patterns, there is a strong need for having a framework that supports flexible, re-usable data patterns for big data. CDAP Dataset APIs can be used to build re-usable data patterns to achieve flexibility and re-usability

Although managing data often gets less fanfare than other IT disciplines, it is of course crucial to the well-being of enterprise big data. The architecture, design, and implementation of datasets can be very complex and present daunting tasks. The goal of having a framework for representing data patterns programmatically using CDAP Dataset APIs is to directly address this complexity, and provide solutions to common problems like administration, agility, and reusability, often using relatively simple mechanisms.

Data professionals have been working with data patterns for many years, but they have probably not explicitly recognized this. Until now, very few data patterns have been formally captured and shared with a wider big data community. Instead, they continue to be held within organizations as tacit knowledge, or expressed in the form of internal standards or guidelines. This approach makes it very hard to achieve agility and reusability within and across organizations to reduce the over-all cost of processing data.

CDAP provides pre-built datasets that are about the problems faced by those who build the data applications and services in an enterprise class analytical solution. They address the need to create the standard datasets designs and the data services that exist invisibly to the CDAP Programs that use the data; in other words, the dataset and services that exist within the data ecosystem.

Scale and Performance

Enterprises require some continuity to be maintained across big data component mutations, modeling data applications based on abstractions are needed to frame and consolidate changes at both enterprise and system levels. CDAP provides the right set of abstraction for modeling a wide range of data applications with performance and scale characteristics optimized at the system level.

From the enterprise standpoint the primary factor is the continuity and consistency of different applications provided for constantly mutating big data components. For that purpose, CDAP manages functionality, persistency, execution and performance for each abstracted component.

Definitions and performance characteristics of those abstracted system components within the CDAP provide the backbone of any enterprise data applications. But one tradeoff that is widely known but not necessarily widely understood, is the “Abstraction Optimization Tradeoff” that has been optimized over years within CDAP.

The “Abstraction Optimization Tradeoff” is between building upon layers of CDAP and achieving optimal performance. Building a data application using CDAP is typically much faster and results in easier to understand and re-use solutions, but the concern may be, it would have sub-optimal performance. Sacrificing abstraction can also lead to sub-optimal performance, but at the additional cost of a more complex, un-maintainable solution. The underlying assumption is that richer, more structured information about any optimization existing at higher layers in the CDAP stack, is not lost in the lower layers, driving optimal performance.

Security and Governance

Most enterprises no longer take for granted that their deployed applications and datasets are secure. In-spite of security being a major concern within

enterprises, most developers, architects and stakeholders who are generally aware of security necessities often make it the last item in the priorities.

Enterprises have begun to pay more attention to tools and platforms with integrated capabilities for security and governance. Developer centric tools and platforms that seek to optimize the role of the developer and increase their productivity, are now required to also have strong capabilities in terms security and governance. Authentication provides great perimeter level security, but as more critical data starts to reside on the clusters there is an eventual need for having more than just the perimeter level security - authorization, audit logs becomes critical for running any applications within enterprise ITs. Embedding standard security protocols and enforcing them across enterprise without hindering the productivity of the developers developing the applications becomes extremely critical. So, deciding on a standard framework with complete security and governance support dramatically reduces the tedious efforts of continuously chasing security needs in a tactical "after-thought" fashion. For example, organizations that have deployed CDAP can use the full suite of capabilities to address enterprise IT's security and governance needs – starting from perimeter security with integration with LDAP, AD or TLS, to the ability to authorize access to cluster entities based on ACLs, all the way to maintaining audit logs for all types of access within the big data clusters. With CDAP, organizations also get insights into how the data flows into, out of and within the cluster using CDAP's metadata capabilities.

All of the security capabilities ultimately stem from the co-engineering of the unified platform with other OSS big data security frameworks. This single design principle is what uniquely enables enterprises to bring their applications to the data, while meeting the critical security and governance requirements.

Self-Service and Microservices

Self-service is becoming more prevalent within enterprises, as LOB users of big data want to do their own big data analytics, using self-service tools such as for data ingestion, data preparation and data discovery. But challenges still exist for IT teams when it comes to providing a standardized and deployable self-service platform that is repeatable and easily consumable. They face important considerations of balancing the users' ease-of-use requirements versus common security and governance requirements.

Hadoop and Spark are complex platforms in their own right, and it is very difficult for enterprise IT to stitch together various components to deliver a fully functioning, repeatable, reliable integration platform on top of Hadoop or Spark. Combining the complexity with constant mutations of technologies makes it very difficult for IT to deliver on the promise of Platform-as-a-service for data analytics needs for different users in the organization.

As providing Hadoop as a self-service platform is becoming more complex, a microservice approach is finding wider use in the industry. Microservice architecture is gaining popularity due to the fact that a complex application is broken down into more fine-grained manageable blocks, the blocks that are closer to business requirements. Business applications can then be constructed in a self-service manner using the blocks defined by the system.

Microservices help economize the overall application development, deployment and management. They also ensure that the same functionality is not built over and over again within organizations.

With CDAP deployed over Hadoop or Spark, IT organizations are increasingly able to support their users' needs through a self-service model. They are able to provide an ease-of-user model combined with great security and governance capabilities. Combination of pipeline and metadata tracking capabilities within CDAP simplify ingestion, data discovery and data tracking using a self-service model. CDAP is also naturally suited for deploying microservices for data processing. For example, Cask CDAP pipeline plugins are a form of microservice that can be integrated in order to build an application.

For more information, please visit the Cask website at cask.co and follow @caskdata.