CDAP Cask Data Application Platform

AVADHOOT AGASTI[1,*,+]

[1] *School of Informatics and Computing, Bloomington, IN 47408, U.S.A.*
[*] *Corresponding authors: aagasti@indiana.edu*
[+] *HID - SL-IO-3000*

project-000, March 27, 2017

CDAP provides application development platform on top of Apache Hadoop. CDAP services enable users to automate the task of building, executing and managing data pipelines. The CDAP studio allows users to drag-and-drop data sources, transformation tasks, and data sinks. User can chain these tasks to create data pipelines. Furthermore, CDAP provides abstraction of logical data pipeline over execution environment. Using CDAP, user can execute a data pipeline either using MapReduce or Apache Spark depending on the capability of underlying Hadoop cluster.

Keywords: CDAP, Hadoop, Name Node, Edge Node, Yarn, Apache Sqoop, Apache Flume, Apache Flink, Apache Atlas, HDFS, Apache Kafka, Apache Spark, MapReduce, HBase

https://github.com/avadhoot-agasti/sp17-i524/tree/master/paper1/S17-IO-3000/report.pdf

1. INTRODUCTION

CDAP stands for Cask Data Application Platform [1]. CDAP is an application development platform that can help developers build, deploy and monitor applications on Apache Hadoop. In a CDAP application, a developer can ingest a dataset, store and manage it on Hadoop, perform data analysis, and develop web services to expose the original and transformed dataset(s). He can also schedule and monitor the execution of the application. CDAP enables the developers to use single platform to develop the end-to-end application on Apache Hadoop.

This technology paper is structured as follows:

- Section 2 explains commonly used application architecture on top of Apache Hadoop without using CDAP. It then explains the application architecture using CDAP to ephasize the use of CDAP.
- Section 3 explains important CDAP concepts.
- Section 4 and Section 5 explain the CDAP deployment options and infrastructure requirements.
- Section 6 explains the representative use cases.
- Section 7 and 8 provide useful information about CDAP like licensing and educational material.
- Finally, in Section 9 and 10, we conclude by explaining the other similar platforms and their high level comparison with CDAP

2. CDAP: UNIFIED APPLICATION DEVELOPMENT PLATFORM

CDAP provides a unified application development platform on top of Apache Hadoop. In below sections, we explain a commonly used application architecture on top of Apache Hadoop. Then we explain the CDAP application architecture using which different components of the application architecture are unified in CDAP platform.

2.1. Application Architecture on Hadoop without CDAP

Figure 1 [2] shows a commonly used application architecture on Hadoop (without CDAP).

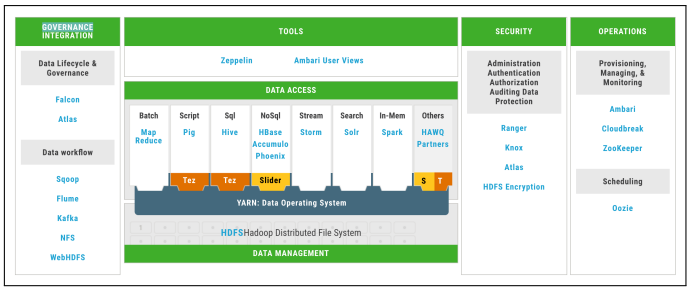

Fig. 1. Application Architecture on Hadoop.

There are following layers/components as depicted in the architecture diagram [2].

- Data Governance and Integration: This layer is responsible for ingesting the data from data source into Hadoop.

Data Ingestion tools like Apache Sqoop, Apache Flume and Apache Kafka are used for Data Ingestion while Apache Falcon and Apache Atlas are used for Data Lifecycle management.

- Data Storage: The data is stored in HDFS.
- Data Processing and Access: The data is transformed and aggregated in Data Processing layer. The processing can involve various steps like cleansing, joining, aggregation and running machine learning algorithms. Many different tools and technologies are used to perform data processing operations. PIG, Hive, Spark are open source scripting technologies which can perform Data Processing tasks.
- Tools: Tools like Apache Zeppelin provide user interface to visualize the data.
- Security and Operations: The data pipeline built using above tools can be secured using tools such as Ranger, Knox, and Atlas. The scheduling of the data pipeline is orchestrated using Oozie.

This application architecture explains that an application built on top of Apache Hadoop involve multiple tools and technologies. Further it is tightly dependent on the underlying deployment architecture and tools availability. Using this mechanism, it is difficult to segregate the application business logic from infrastructure components.

2.2. Application Architecture on Hadoop with CDAP

Figure 2 [3] explains how CDAP provides a common application development platform. CDAP provides abstractions to ingest data, store it in HDFS, process it using the application business logic, store the results in HDFS and expose web service APIs on the result data. Developers need not use different tools to implement various layers. He can implement all the layers in CDAP platform. Further, he can use Java as a common coding language to implement all the coding tasks across all the layers.

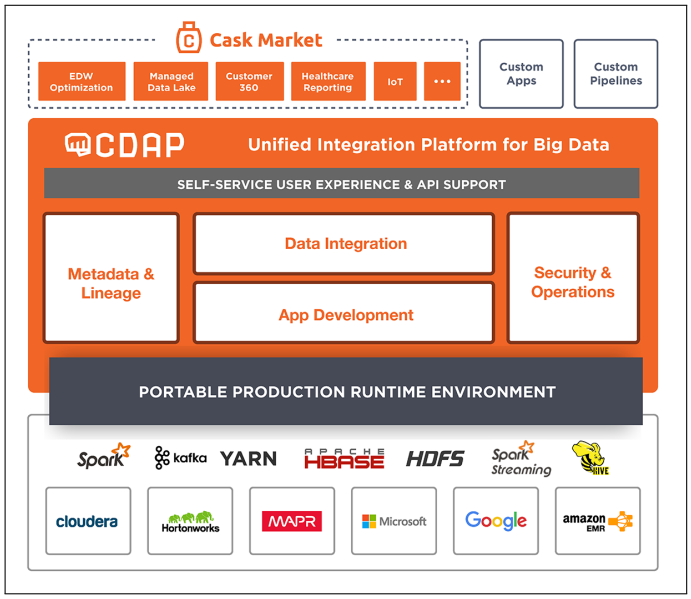

Fig. 2. CDAP Application Architecture.

3. IMPORTANT CDAP CONCEPTS

CDAP revolves around below important concepts:

- CDAP Datasets provide logical abstraction over the data stored in Hadoop. The data can be files in HDFS or tables in HBase. A dataset needs to be first declared in the CDAP. Any dataset declared in CDAP can be used in any CDAP applications or CDAP services.
- CDAP Applications provide containers to implement application business logic in open source processing frameworks like MapReduce, Spark and real time flow. CDAP applications also provide standardize way to deploy and manage the apps
- CDAP Services provide services for application management, metadata management, and streams management

4. CDAP DEPLOYMENT

CDAP provides multiple deployment options. In standalone mode, it can be downloaded as a zip file and deployed. Alternatively it is available as a standalone virtual machine. For deployment in cluster mode, CDAP provides options which are specific to underlying Hadoop distribution as explained below:

- Cloudera Hadoop Distribution (CDH) - Cloudera Manager [4] is tool to deploy CDH cluster. CDAP provides CDAP-parcel [5] which is plug in for Cloudera Manager. Once you add CDAP-parcel to your Cloudera Manager, CDAP can be deployed using Cloudera Manager and all CDAP services can be monitored using Cloudera Manager.
- Amazon EMR (Elastic Map Reduce) - EMR is Amazon's Hadoop distribution for the Amazon Web Services cloud [6]. EMR provides 'Create Cluster Wizard' to create EMR cluster. According the CDAP documentation [7], CDAP provides a bootstrap action which is executed when the EMR cluster is created. Using this mechanism, CDAP platform can be deployed on EMR when the EMR cluster is created.
- CDAP can also be deployed on HortonWorks Hadoop Distribution, MapR Hadoop Distribution and Apache Hadoop.

5. CDAP INFRASTRUCTURE REQUIREMENTS

CDAP is deployed on edge nodes of the Hadoop cluster. CDAP communicates with Hadoop services like Yarn, HDFS and HBase. Hence CDAP needs to be installed in same network as that of Hadoop. However, none of the CDAP components are required to be installed on Hadoop Namenode or Hadoop datanodes. CDAP documentation [8] provide the CDAP deployment architecture.

6. REPRESENTATIVE USE CASES

CASK [1] is the company which provides commercial distribution for CDAP. CASK has developed several applications using CDAP. Some of the applications developed using CDAP are explained below

- CASK Hydrator [9] is interactive application for building, running and managing data pipelines for enterprise data lake. CASK Hydrator is UI driven tool using which users

can ingest data from sources like traditional RDBMS, transform it, aggregate it and finally store the data into permanent storage like HDFS. CASK Hydrator provides UI drag-and-drop style abstraction to all of the above task.

- CASK Customer 360 [10] is another representative application which is built using CDAP. Customer 360 applications analyzes customer behavior data on various interaction platforms like mobile apps, online communities, customer support portals, and social media. CDAP can be used to ingest the data from these sources and perform join, unification and aggregations to derive 360 degree view of customer.

7. LICENSING

CDAP is licensed [11]under Apache License, Version2.0.

8. EDUCATIONAL MATERIAL

Below given is the list of educational material on CDAP:

- CDPA Applications code repository in Github [12] provide sample applications which are built on top of CDAP Platform.
- CDAP Documentation [13] provides introduction to CDAP platform.

9. OTHER HADOOP APPLICATION DEVELOPMENT PLATFORMS

Below given are the two other application development platforms on top of Apache Hadoop:

- Cascading [14] is another application development platform on Apache Hadoop. Cascading has many similar features like CDAP. Cascading supports Java APIs, Data Processing APIs, Data Integration APIs, Scheduler APIs, Relational Operations and scriptable interface. Cascading also support many different Hadoop distributions.
- Talend Big Data Integration [15] : Talend is integration tool using which data can be extracted from source systems, stored on Hadoop and processed in Hadoop. Although Talend is not exactly an application development platform, lot of its features overlap with CDAP. Talend provides visual interface for performing data ingestion and processing operations on Hadoop

10. CONCLUSION

CDAP provides a unified application development platform over Apache Hadoop. Using CDAP developers can implement multiple layers of their data pipeline in one uniform language and tool.

ACKNOWLEDGEMENTS

The authors thank Prof. Gregor von Laszewski for his technical guidance.

REFERENCES

[1] CASK, “Cask - the first unified integration platform for big data,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://cask.co/

[2] Hortonworks, “Hortonworks data platform,” Web Page, online; accessed 20-Mar-2017. [Online]. Available: https://hortonworks.com/products/data-center/hdp/

[3] CASK, “The first unified integration platform for bigdata that cuts down the time to production for data applications and data lakes by 80online; accessed 20-Mar-2017. [Online]. Available: https://cask.co/products/cdap/

[4] Cloudera, “Simple administration for apache hadoop,” Web Page, online; accessed 20-Mar-2017. [Online]. Available: https://www.cloudera.com/products/product-components/cloudera-manager.html

[5] CASK, “Installation using cloudera manager,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://docs.cask.co/cdap/current/en/admin-manual/installation/cloudera.html#admin-installation-cloudera

[6] Amazon, “Amazon emr,” Web Page, online; accessed 20-Mar-2017. [Online]. Available: https://aws.amazon.com/emr/

[7] CASK, “Installation on amazon emr using bootstrap actions,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://docs.cask.co/cdap/current/en/admin-manual/installation/emr.html

[8] ——, “System requirements,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://docs.cask.co/cdap/current/en/admin-manual/system-requirements.html

[9] ——, “Cask - hydrator,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://cask.co/products/hydrator/

[10] ——, “Customer 360,” Web Page, online; accessed 20-Mar-2017. [Online]. Available: http://cask.co/solutions/customer-360/

[11] ——, “Cdap product license,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://docs.cask.co/cdap/4.0.0/en/reference-manual/licenses/index.html#cdap-product-license

[12] “Cdap applications,” Code Repository, May 2015, accessed: 2017-2-18. [Online]. Available: https://github.com/caskdata/cdap-apps

[13] CASK, “Getting started developing with cdap,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://docs.cask.co/cdap/current/en/developers-manual/getting-started/index.html

[14] Cascading, “Cascading,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://www.cascading.org/projects/cascading/

[15] Talend, “Talend products - big data integration,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: https://www.talend.com/products/big-data/