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, February 28, 2017

This paper explains CDAP - Cask Data Application Platform. CDAP provides abstraction layer on top of Apache Hadoop and other Apache Big Data Stack technologies. This paper explains CDAP technology, the kind of problems it can solve, the infrastructure and setup requirements, and its competitive landscape. The paper also provides links to learning material for CDAP.

Keywords: CDAP, Hadoop

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

1. INTRODUCTION

CDAP stands for Cask Data Application Platform. CDAP is an application development platform using which developers can build, deploy and monitor applications on Apache Hadoop. In a typical CDAP application, a developer can ingest data, store and manage datasets on Hadoop, perform batch mode data analysis, and develop web services to expose the data. They can also schedule and monitor the execution of the application. This way, CDAP enables the developers to use single platform to develop the end to end application on Apache Hadoop. This paper introduces CDAP as application development platform and explains various use cases that can be solved using CDAP. The paper also explains the CDAP deployment options and infrastructure requirements. Finally we conclude by explaining the other similar platforms and their high level comparison with CDAP. The paper also provides references to the learning material.

2. WHY CDAP

Before we understand how CDAP helps in application development, lets understand how a typical application looks like in Hadoop.

2.1. Typical Application Architecture on Hadoop

Figure 1 shows a typical application architecture on Hadoop.

There are following layers/components -

- Data Ingestion - ingest the data from data source into Hadoop. Data Ingestion tools like Apache Sqoop, Apache Flume, Apache Kafka are popularly used for Data Ingestion
- Data Storage - The data is stored in HDFS.

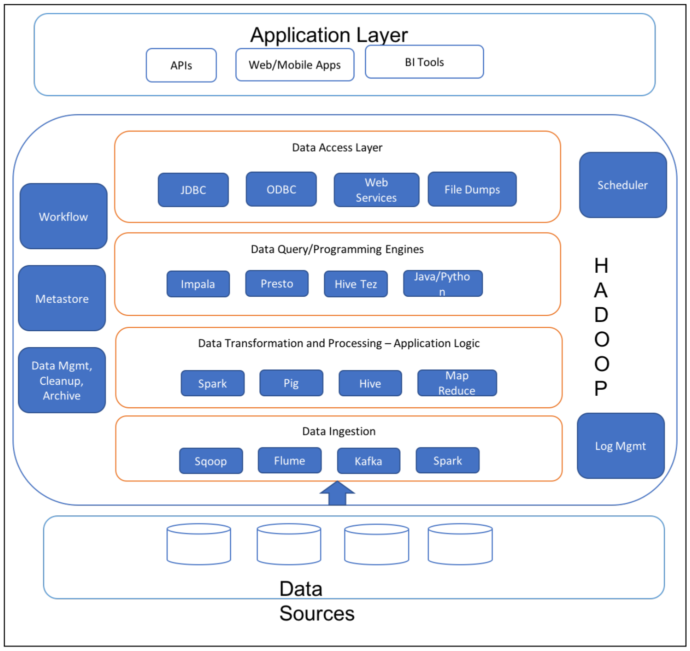

Fig. 1. Typical Application Architecture on Hadoop.

- Data Processing - The data is transformed and aggregated in Data Processing layer. The processing can involved various steps like cleansing, joining, aggregation and running machine learning algorithms. Many different tools and technologies are used to perform data processing operations - e.g. PIG, Hive, Spark are popular open source scripting technologies while Talend, Informatica are visual commercial products.
- Result Storage - The output of data processing step is again stored in HDFS
- Data Access - The end users can access the data (mainly results) using various data access mechanism like APIs, SQL interface or BI tool interface.

2.2. Why CDAP - CDAP Application Architecture

CDAP provides a common application development platform in which a developer can code all the application layers in a typical Hadoop application. 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. User need not use different tools to code different layers. He can simply code all the layers in CDAP platform. He can use same coding language (Java) to do the coding across all the layers.

Further CDAP uses native Hadoop tools for actually performing the operations. For example, the data processing operation implemented in CDAP translate to Spark jobs. Due to this, CDAP users continue to leverage the new enhancements in Apache Hadoop.

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 map reduce, 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 many 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 cluster mode deployment, CDAP provides Hadoop-distribution specific options as explained below

- Cloudera Hadoop Distribution (CDH) - Cloudera Manager is tool to deploy CDH cluster. As per CDAP documentation [1] CDAP provides CDAP-parcel 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. EMR provides 'Create Cluster Wizard' to create EMR cluster. According the CDAP documentation [2], 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 [3] provide the CDAP deployment architecture.

6. EDUCATIONAL MATERIAL

- CDPA Applications code repository in Github [4] provide sample applications which are built on top of CDAP Platform.
- CDAP Documentation [5] provides introduction to CDAP platform.

7. REPRESENTATIVE USE CASES WHICH CAN LEVERAGE CDAP

CASK [6] 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 [7] 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, trasnform 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.
- Customer 360 is another representative application which can be 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.

8. LICENSING

CDAP is licensed [8]under Apache License, Version2.0.

9. OTHER HADOOP APPLICATION DEVELOPMENT PLATFORMS

- Cascading [9] 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 [10] : 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 an application development platform over Apache Hadoop. Using CDAP developers can code multiple layers of thier data pipeline in one uniform language and tool. CDAP also can help to shield developers from different Hadoop deployment options like Cloudera, Hortonworks and EMR.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] 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
[2] ——, “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
[3] ——, “System requirements,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://docs.cask.co/cdap/current/en/admin-manual/system-requirements.html
[4] “Cdap applications,” Code Repository, May 2015, accessed: 2017-2-18. [Online]. Available: https://github.com/caskdata/cdap-apps
[5] 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
[6] ——, “Cask - the first unified integration platform for big data,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://cask.co/
[7] ——, “Cask - hydrator,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://cask.co/products/hydrator/
[8] ——, “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
[9] Cascading, “Cascading,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: http://www.cascading.org/projects/cascading/
[10] Talend, “Talend products - big data integration,” Web Page, online; accessed 18-Feb-2017. [Online]. Available: https://www.talend.com/products/big-data/