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

This review document is provided for you to achieve your best. We have listed a number of obvious opportunities for improvement. When improving it, please keep this copy untouched and instead focus on improving report.tex. The review does not include all possible improvement suggestions and if you sea comment you may want to check if this comment applies elsewhere in the document.

Abstract: remove This paper, its clear that this is a a paper so you do not have to mention it. Furthermore, its actually a technology review. Abstract has grammar errors.

1. INTRODUCTION

CDAP stands for Cask Data Application Platform

Citation

. CDAP is an application development platform using which developers can build, deploy and monitor applications on Apache Hadoop.

Grammar, make sure you check grammar we will not point out every grammer or spelling error

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

Grammar

single platform to develop the end to end

Spelling

application on Apache Hadoop. This paper introduces

not needed phrase this paper, try to avoid, Instead be concreate and say the paper is structured as follows. IN Section ... we do, use Latex refs and labels

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

Grammar

.

2. WHY CDAP

Do not use questions in section headers, and if, you need a ?

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

not a good way to introduce this section

2.1. Typical Application Architecture on Hadoop

Figure 1

not using refs and labes for images

shows a typical application architecture on Hadoop.

now what, are you explaining it?

There are following layers/components -

Grammar

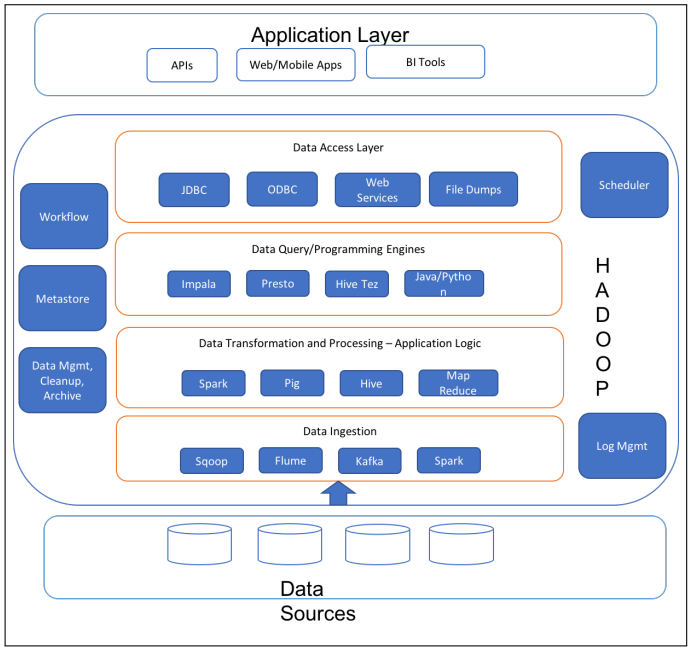

Fig. 1. Typical Application Architecture on Hadoop.

Is this copied, if so you need refernce

You have not used spell checker

. If self drawn, change image to not use white fonts on dark background but use black fonts, make contrast high e.g. boxes lighter so we can read, make sure fontsize is ok and readable

in english you use :

use latex description

inconsistent capitalization

- Data Ingestion - ingest the data from data source into Hadoop. Data Ingestion tools like Apache Sqoop

  Citation

  , Apache Flume

  Citation

  , Apache Kafka

  Citation

  are popularly used

  Term

  for Data Ingestion

- Data Storage - The data is stored in HDFS

  Citation

  .

- Data Processing - The data is transformed and aggregated in Data Processing layer. The processing can involved

  Grammar

  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.

  avoid e.g. here you may want to use such as

  PIG, Hive, Spark are popular open source scripting technologies while Talend, Informatica are visual commercial products.

- Result Storage - The output of data processing step

  Grammar

  is again stored in HDFS

  why again?

- Data Access - The end users can

  is can needed?

  access the data (mainly results) using various data access mechanism like APIs, SQL interface or BI tool interface.

  do you need a/an/the?

2.2. Why CDAP - CDAP Application Architecture

avoid question in section header

CDAP provides a common application development platform in which a developer can code all the application layers in a typical Hadoop application.

Is this repetition or something new. If it is something new, than say in addition to xyz CDAP provides ....

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.

It is unclear if this is new or already explained, see the list you had previously

User

Grammar

need not use different tools to code different layers. He can simply code all the layers in CDAP platform. He can use

Grammar

same coding language (Java) to do the coding across all the layers.

Further

Term

CDAP uses native Hadoop tools for actually performing the operations. For example, the data processing operation implemented in CDAP translate to Spark jobs

Citation

. Due to this, CDAP users continue to leverage the new enhancements in Apache Hadoop.

Which new enhancements?

Citation

3. IMPORTANT CDAP CONCEPTS

CDAP revolves around

Term

below important concepts:

- CDAP Datasets provide logical abstraction

  Grammar

  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

this seems in part replication of what was said before, would it be worth while to combine sections. Also should concepts not before architecture?

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

Grammar

Hadoop-distribution specific options as explained below

- Cloudera Hadoop Distribution (CDH)

  Citation

  - Cloudera Manager

  Citation

  is tool to deploy CDH cluster. As per CDAP documentation

  not helpful to say its in the documentation without listing the reference

  [1] CDAP provides CDAP-parcel

  Citation

  which is

  Grammar

  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

  full stop

- Amazon EMR (Elastic Map Reduce) - EMR

  Citation

  is Amazon's Hadoop distribution for the Amazon Web Services cloud. EMR provides

  Grammar

  '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,

  Grammar

  CDAP platform can be deployed on EMR when the EMR cluster is created.

inconsistent use of list as horton is outside of list

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

what is an edge node

of the Hadoop cluster. CDAP communicates with Hadoop services like Yarn

Citation

, HDFS

Citation

and HBase

Citation

. Hence CDAP needs to be installed in

Grammar

same network as that of Hadoop. However, none of the CDAP components are required to be installed on Hadoop Namenode

what is a name node

Citation

or Hadoop datanodes. CDAP documentation [3] provide

Grammar

the CDAP deployment architecture.

6. EDUCATIONAL MATERIAL

Paper writing rules do not allow a section to just have a list, it must have an introductory sentence

- 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.

You make a big deal in the abstract but you have only two links, thats fine, but people wonder why you emphasize this that much

7. REPRESENTATIVE USE CASES WHICH CAN LEVERAGE CDAP

Introductory sentence missing

inconsistent enumeration, between CASK and Customer 360, while using introduction section just for CASK and not Customer 360

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

  You have not used spell checker

  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.

mind spaces after citations

9. OTHER HADOOP APPLICATION DEVELOPMENT PLATFORMS

Introductory sentence missing

- 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

vage comparison is used, lots of ... can you be more specific

10. CONCLUSION

CDAP provides an application development platform over Apache Hadoop. Using CDAP developers can code multiple layers of thier

You have not used spell checker

data pipeline in one uniform language and tool.

I do not understand

CDAP also can help to shield developers from different Hadoop deployment options like Cloudera, Hortonworks and EMR.

Not clear enough phrased

ACKNOWLEDGEMENTS

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

many citations missing

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/

TODO LIST

- [ ] This review document is provided for you to achieve your best. We have listed a number of obvious opportunities for improvement. When improving it, please keep this copy untouched and instead focus on improving report.tex. The review does not include all possible improvement suggestions and if you sea comment you may want to check if this comment applies elsewhere in the document. 1
- [ ] Abstract: remove This paper, its clear that this is a a paper so you do not have to mention it. Furthermore, its actually a technology review. Abstract has grammar errors. 1

- [ ] Citation . . . 1
- [ ] Grammar, make sure you check grammar we will not point out every grammer or spelling error . . . 1
- [ ] Grammar . . . 1
- [ ] Spelling . . . 1
- [ ] not needed phrase this paper, try to avoid, Instead be concreate and say the paper is structured as follows. IN Section ... we do, use Latex refs and labels . . . 1
- [ ] Grammar . . . 1
- [ ] Do not use questions in section headers, and if, you need a ? . . . 1
- [ ] not a good way to introduce this section . . . 1
- [ ] not using refs and labes for images . . . 1
- [ ] now what, are you explaining it? . . . 1
- [ ] Grammar . . . 1
- [ ] in english you use : . . . 1
- [ ] Is this copied, if so you need refernce

  You have not used spell checker

  . If self drawn, change image to not use white fonts on dark background but use black fonts, make contrast high e.g. boxes lighter so we can read, make sure fontsize is ok and readable . . . 2
- [ ] You have not used spell checker . . . 2
- [ ] use latex description . . . 2
- [ ] inconsistent capitalization . . . 2
- [ ] Citation . . . 2
- [ ] Citation . . . 2
- [ ] Citation . . . 2
- [ ] Term . . . 2
- [ ] Citation . . . 2
- [ ] Grammar . . . 2
- [ ] avoid e.g. here you may want to use such as . . . 2
- [ ] Grammar . . . 2
- [ ] why again? . . . 2
- [ ] is can needed? . . . 2
- [ ] do you need a/an/the? . . . 2
- [ ] avoid question in section header . . . 2
- [ ] Is this repetition or something new. If it is something new, than say in addition to xyz CDAP provides 2
- [ ] It is unclear if this is new or already explained, see the list you had previously . . . 2
- [ ] Grammar . . . 2
- [ ] Grammar . . . 2
- [ ] Term . . . 2
- [ ] Citation . . . 2
- [ ] Which new enhancements? . . . 2
- [ ] Citation . . . 2
- [ ] Term . . . 3
- [ ] Grammar . . . 3
- [ ] this seems in part replication of what was said before, would it be worth while to combine sections. Also should concepts not before architecture? . . . 3
- [ ] Grammar . . . 3
- [ ] Citation . . . 3
- [ ] Citation . . . 3
- [ ] not helpful to say its in the documentation without listing the reference . . . 3
- [ ] Citation . . . 3
- [ ] Grammar . . . 3
- [ ] full stop . . . 3
- [ ] Citation . . . 3
- [ ] Grammar . . . 3
- [ ] Grammar . . . 3
- [ ] inconsistent use of list as horton is outside of list . . . 3
- [ ] what is an edge node . . . 3
- [ ] Citation . . . 3
- [ ] Citation . . . 3
- [ ] Citation . . . 3
- [ ] Grammar . . . 3
- [ ] what is a name node . . . 3
- [ ] Citation . . . 3
- [ ] Grammar . . . 3
- [ ] Paper writing rules do not allow a section to just have a list, it must have an introductory sentence . . . 3
- [ ] You make a big deal in the abstract but you have only two links, thats fine, but people wonder why you emphasize this that much . . . 3
- [ ] Introductory sentence missing . . . 4
- [ ] inconsistent enumeration, between CASK and Customer 360, while using introduction section just for CASK and not Customer 360 . . . 4
- [ ] You have not used spell checker . . . 4
- [ ] mind spaces after citations . . . 4
- [ ] Introductory sentence missing . . . 4
- [ ] vage comparison is used, lots of ... can you be more specific . . . 4
- [ ] You have not used spell checker . . . 4
- [ ] I do not understand . . . 4
- [ ] Not clear enough phrased . . . 4
- [ ] many citations missing . . . 4
- [ ] You have not used spell checker . . . 5