Introduction to H2O

SUSHMITA SIVAPRASAD[1]

[1] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
* sushsiva@umail.iu.edu

April 10, 2017

Machine learning and data mining have been used in many data driven industries. H2O is a platform using for performing machine learning and predictive analytics for large scale data using cloud. When the data that is generated is large scale and is in terrabytes, H2O serves a very important purpose of being able to accurately predict using different algorithms and also using different programming languages through APIs. This paper gives an introduction to H2O, how this platform has impacted various industries across several domains with improved accuracy and reduced processing time. Different use cases of the H2O platform has been explained as well.

Keywords: machine learning, data mining,predictive analytics,cloud

https://github.com/SushmitaSivaprasad/sp17-i524/tree/master/paper1/S17-IR-2038/report.pdf

INTRODUCTION

H2O is an open source platform that is used to create machine learning and predictive analytics models on big datasets. It is mainly written in Javascript but have APIs for R , Python, Excel, Tableau and Flow[1]. The main algorithms implemented on the datasets are deep learning, gradient boosting , generalized linear model, distributed random forest and k-means. The algorithms implemented on the big datasets is read in a parallel manner and is then distributed and stored in memory in a compressed column format. H2O also has an inbuilt intelligence to detect and support the process of obtaining data and importing them for immediate use or storing in a database which can be obtained from various sources in different formats[1].

HOW DOES H2O WORK?

H2O is used on large dataset , usually in the range of terrabytes due to it's speed in processing the data. A company might have their dataset stored on big data systems such as Hadoop. On analyzing a data we usually choose a smaller sample dataset rather than the entire dataset to build a model due to the large processing time involved. H2O has the advantage of being able to use the entire dataset to run the algorithm on a larger dataset, the more the data we are able to analyse better the predictions would be[2]. Suppose a business is trying to understand the best product placement for optimal customer enagement, the model would be created based on the dataset formed collecting information about the interactions of the customers on the website. H2O is used to model the data with multiple algorithms using more than one machine at the same time, this way we don't have to sample the data[2]. H2O is also used to score hundreds of models in nano seconds to check which algorithm works better for that dataset.

REQUIREMENTS

Operating Systems

It works on the following operating systems
Windows 7 or later
OSX 10.9 or later
Ubuntu 12.04
RHEL/CentOS6 or later[3]

Languages

Java 7 or later
Scala 2.10 or later
R version 3 or later
Python 2.7x or 3.5x[3]

Browser

Chrome
Safari
Internet Explorer
Firefox[3]

Hadoop

Optional Cloudera CDH 5.2 or later
MapR 3.1.1 or later
Hortonworks HDP 2.1 or later[3]

Supported File Formats

H2O supports the following file types
CSV files
ORC
ARFF
XLSX
XLS
Avro
Parquet[3]

More information can be obtained from the documentation provided on the H2O website[3].

ARCHITECTURE

The H2O architecture consists of different components which combine together to form the H2O software stack.
We can divide the H2O architecture into 3 different components, top section includes all the REST API clients, middle includes the Network Cloud and the bottom section contains the different components that run within an H2O JVM process[4]. The top section contains the programming languages that can be used on the big dataset here. The REST API clients communicate with the H2O with the help of a socket connection[4]. The Network cloud consists of the different inbuilt algorithms to create the necessary model on the data, this can also contain a customized customer algorithm to analyze the required dataset and produce the desired outputs.

REST API Clients

- Javascript: The H2O Web UI is written in the javascript language.
- R: Using the H2O R package known as 'library(h2o)' the users can write their own R functions.
- Python: The Python scripts must use the REST API directly as a library for this has not been extablished yet but would be released in the future.
- Excel: H2O provides an H2O worksheet for Microsoft Excel, which allows us to import the big datasets and run the algorithms.
- Tableau: The users of H2O may pull results from H2O to create visualizations in Tableau.
- Flow: This is a notebook style WEB UI for H2O[4].

JVM Components

The H2O cloud can consists of two or more nodes which can contain a single JVM process. Each JVM process consists of language, algorithm and infrastructure (manages the resources management such as memory and CPU)[4].

USED CASES

CapitalOne

CapitalOne, a fortune 500 bank with 960 banks and 2000 ATMs accumulates terrabytes of information in real time on customer information and financial processing. H2O software was implemented in their systems and was used to reduce the process time of the algorithm implementations over the large data sets[5]. Different algorithms can be applied to find which one works best due to reduction in time consumed in processing these large datasets. A large number of hard and soft metrics were evaluated as well using machine learning frameworks[5].Using H2O they were able to process data received from credit cards the moment they were swiped, using this information new offerings were marketed to the customer based on their spending habits.

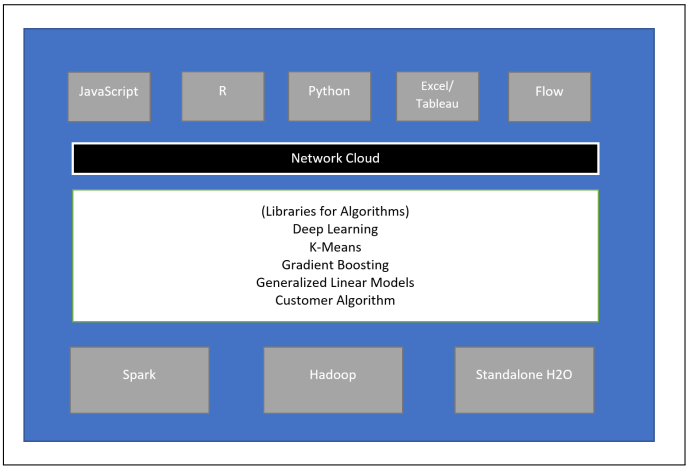

Fig. 1. H2O Architecture[4]

MarketShare

The company MarketShare have implemented H2O to optimize budgeting for marketing. Since marketing approaches are data driven features , predictive analytics under H2O was used to give a comparison on how the current state of the marketing budget is and how much is the predicted revenue[6]. Using H2O, solutions were generated as to what are the changes that can be made to improve the current projection and what an improved projection will look like. MarketShare was able to go on the cloud and use as much machines as required and get desired outputs on the large datasets. They use 25 machines for all of their clients to process the data and were able to expand the scalability of the dataset.If their datasize increases by x amount then they would add y more machines to solve the problem[6]. Scaling across lot of nodes is critical to their business as the company deals with digital log data and irrespective of the complexity of the modelling and the huge size of the data[6]. MarketShare was able to generate marketing plans and what-if cases based on the information from their customers using different predictive analytics models.

RELEVANT RESOURCES

H2O has an open source platform and hence has a community for support.

- Step by step instructions with documentation and videos have been provided for installation and to understand the work flow of H2O[1].
- Free online training videos are provided on the main webpage [7].
- H2O documentation is available on their website [1].
- h2ostream is an open source google group where H2O users can post questions and get answers.
- They have built an online community at[8] which is a discussion platform.

- They also conduct conferences around the year in United States for users to interact among one another and update new releases and happenings in the big data community[9].

CONCLUSION

Being an open source platform it gives user the flexibility to solve the problems. It is easy to set up and has a smooth installation and usage feature due to its interface with familiar programming environments using APIs[10]. Models can also be inspected during training.It can process any format of file , it can even connect to data from HDFS, S3 ,SQL and NoSQL data sources[10].It has a large scalability hence allowing large datasets to be analyzed by using multiple machines. It also conducts a real time data scoring for accurate predictions[10].

ACKNOWLEDGEMENT

A very special thanks to Professor Gregor von Laszewski and the teaching assistants Miao Zhang and Dimitar Nikolov for all the support and guidance in getting this paper done and resolving all the technical issues faced. The paper is written during the spring 2017 semester course I524: Big Data and Open Source Software Projects at Indiana University Bloomington.

AUTHOR BIOGRAPHY

Sushmita Sivaprasad is a graduate student in Data Science at Indiana University under the department of Informatics and Computing. She had completed her bachelors in Electronics and Communication from SRM University , India and her master's in International Business from Hult International Business School, UAE.

REFERENCES

[1] "H2O Documentation," Web Page, Sep. 2016. [Online]. Available: https://h2o-release.s3.amazonaws.com/h2o/rel-turing/7/docs-website/h2o-docs/welcome.html

[2] H2O.ai, "Oxdata H2O Explainer Video," Web Page, Aug. 2013. [Online]. Available: https://www.youtube.com/watch?v=UGW3cT_cZLc

[3] "Requirements of H2O," Web Page, Sep. 2016. [Online]. Available: https://h2o-release.s3.amazonaws.com/h2o/rel-turing/7/docs-website/h2o-docs/welcome.html#requirements

[4] "H2O Architecture," Web Page, Sep. 2016. [Online]. Available: https://h2o-release.s3.amazonaws.com/h2o/rel-turing/7/docs-website/h2o-docs/architecture.html

[5] Brendan Herger, "Capital One on Machine Learning using H2O ," Youtube Video, Jan. 2016. [Online]. Available: https://www.youtube.com/watch?v=L6a8oITd2L8

[6] Prateem Mandal, "MarketShare turns to H2O for Digital Marketing Analytics," Youtube Video, Jan. 2016. [Online]. Available: https://www.youtube.com/watch?v=L6a8oITd2L8

[7] "Learn H2O," Web Page, Sep. 2016. [Online]. Available: http://learn.h2o.ai

[8] "H2O Community," Web Page, p. 32. [Online]. Available: https://community.h2o.ai/index.html

[9] "H2O Meetups," Web Page. [Online]. Available: http://www.h2o.ai/events/

[10] "Why H2O," Web Page, Sep. 2016. [Online]. Available: http://www.h2o.ai/h2o/