International Journal of Advanced Research in Big Data Management System
Vol.2, No.2 (2018), pp.1-8
http://dx.doi.org/10.21742/ijarbms.2018.2.2.01

A Study on Frame Work of H2O for Data Science

Se-jing Im[1], Ch. Apoorva[2] and Ch. Sekhar[3]

[1]*Chonnam National University, Korea*
[2,3]*Department of Computer Science & Engineering, Vignan's Institute of Information Technology, Visakhapatnam, AP, India*
[1]*sjlim@cnu.ac.kr*, [3]*Sekhar1203@gmail.com*

Abstract

H2O.ai is a centered around conveying AI to organizations through programming. Its leader item is H2O, the main open source stage that makes it simple for money related administrations, protection and medicinal services organizations to convey AI and profound figuring out how to tackle complex issues.H2O is an open source, in-memory, disseminated, quick, and adaptable machine learning and prescient investigation stage that enables you to construct machine learning models on enormous information and gives simple productionalization of those models in a venture situation. In this paper we are going to discuss about the components involved in H2O design, frame work requirements and Life cycle of data science.

Keywords: *Artificial intelligent, H2O, Data, Science*

1. Introduction

H2O is an open-source programming for enormous information investigation. It is delivered by the organization H2O.ai (some time ago 0xdata), which propelled in 2011 in Silicon Valley. H2O enables clients to fit a huge number of potential models as a component of finding designs in information. H2O's numerical center is created with the authority of Arno Candel, some portion of Fortune's 2014 "Major Data All Stars". The company's logical counselors are specialists on factual learning hypothesis and scientific streamlining. The H2O programming runs can be called from the factual bundle R, Python, and different situations. It is utilized for investigating and examining datasets held in distributed computing frameworks and in the Apache Hadoop Distributed File System and in addition in the customary working frameworks Linux, macOS, and Microsoft Windows. The H2O programming is composed in Java, Python, and R. Its graphical-UI is perfect with four programs: Chrome, Safari, Firefox, and Internet Explorer [1][2].

H2O is an open source, in-memory, appropriated, quick, and adaptable machine learning and prescient examination stage that enables you to manufacture machine learning models on huge information and gives simple productionalization of those models in a venture domain. H2O's center code is composed in Java. Inside H2O, a Distributed Key/Value store is utilized to access and reference information, models, objects, and so forth, over all hubs and machines. The calculations are executed over H2O's circulated Map/Reduce system and use the Java

Article history:
Received (December 17, 2017), Review Result (February 27, 2018), Accepted (June 12, 2018)

Print ISSN: 2208-1674, eISSN: 2208-1682 IJARBMS

Fork/Join structure for multi-threading. The information is perused in parallel and is disseminated over the group and put away in memory in a columnar configuration compressedly. H2O's information parser has worked in insight to figure the diagram of the approaching dataset and backings information ingest from numerous sources in different arrangements. H2O's REST API enables access to every one of the capacities of H2O from an outside program or content by means of JSON over HTTP. The Rest API is utilized by H2O's web interface (Flow UI), R authoritative (H2O-R), and Python official (H2O-Python). The speed, quality, usability, and model-sending for the different front line Supervised and Unsupervised calculations like Deep Learning, Tree Ensembles, and GLRM make H2O an exceedingly looked for after API for huge information science.

2. Architecture and working mechanism

The figure underneath indicates the greater part of the distinctive segments that cooperate to shape the H2O programming stack. The figure is part into a Top and Bottom segment, with the system cloud isolating the two segments. The best area demonstrates a portion of the diverse REST API customers that exist for H2O. The base segment demonstrates distinctive parts that keep running inside a H2O JVM process. The shading plan in the graph demonstrates each layer in steady shading however dependably indicates client included client calculation code as dark.

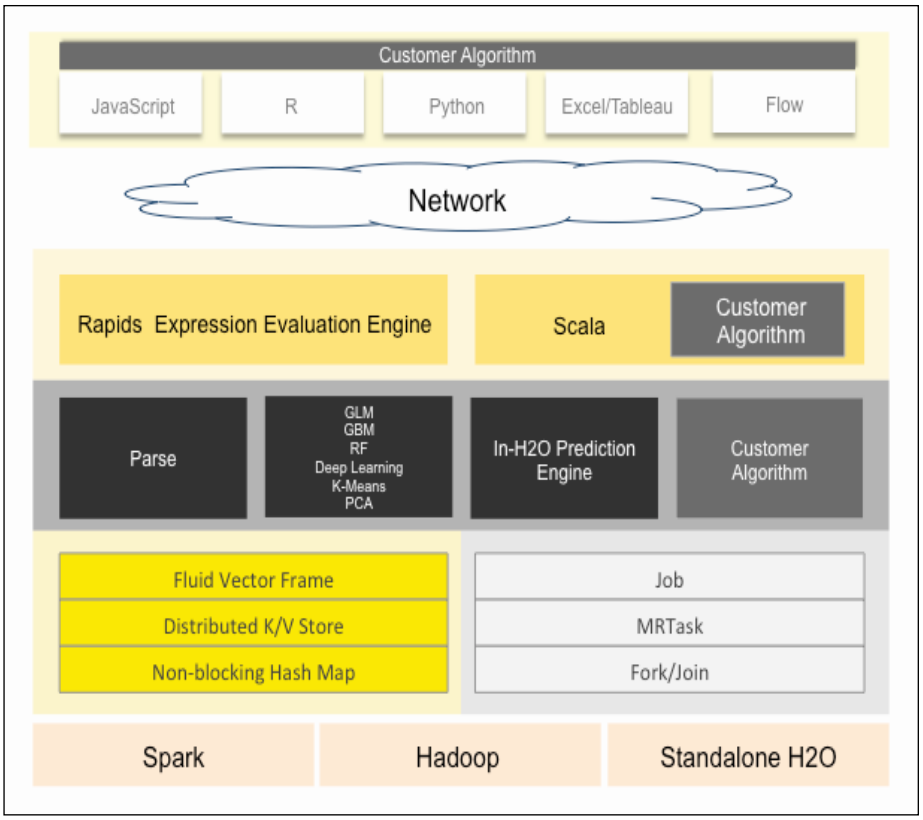

Figure 1. Frame work of H2O

H2O programming is based on Java, Python, and R with a reason to streamline machine learning for Big Data [3][4]. It is offered as an open source stage with the accompanying unmistakable highlights. Enormous Data Friendly implies that one can utilize the majority of their information progressively for better expectations with H2O's quick in-memory dispersed parallel preparing abilities. For generation sending an engineer require not stress over the variety in the improvement stage and creation condition. H2O models once made can be used and sent like any Standard Java Object. H2O models are arranged into POJO (Plain Old Java Files) or a MOJO (Model Object Optimized) organize which can without much of a stretch implant in any Java condition. The magnificence of H2O is that its calculations can be used by different classifications of end clients from business investigators and analysts (who are not comfortable with programming dialects utilizing its Flow online GUI) to engineers who know any of the broadly utilized programming dialects (e.g Java, R, Python, Spark). Utilizing as a part of memory pressure procedures, H2O can deal with billions of information pushes in-memory, even with a genuinely little bunch. H2O actualizes all regular machine learning calculations, for example, summed up straight demonstrating (direct relapse, strategic relapse, and so forth.), Naive Bayes, essential parts examination, time arrangement, k-implies grouping, Random Forest, Gradient Boosting, and Deep Learning.

To work with H2O, we need the following requirements as pre-requisites.

1) Operating Systems: H2O can work can any platform, but it needs minimum versions of Ubuntu 12.04, Windows 7, OS X 10.9 and RHEL/CentOS 6

2) Languages: It supports various languages based the requirement of work we can use the respective language. Java 7, JDK -64 bit for build H2O test. To run H2O binary we need JRF-64 bi if R or Python command line

3) Browser: It supports all varieties of browser higher versions Chrome, Firefox, Safari, or Internet Explorer.

3. Components of H2O

H2O's center code is composed in Java. Inside H2O, a Distributed Key/Value store is utilized to access and reference information, models, objects, and so forth., over all hubs and machines. The algorithms are actualized over H2O's conveyed Map/Reduce structure and use the Java Fork/Join system for multi-threading. The information is perused in parallel and is conveyed over the group and put away in memory in a columnar organization compressed. H2O's information parser has worked in insight to figure the diagram of the approaching dataset and backings information ingest from numerous sources in different arrangements.

As shown in [Table 1], the major components are involved in the architectural design of H2O are H2O Frame, Distributed K/V Store and Data in H2O. The below figure shows the components aligned structure and respective responsibilities [5][6].

Table 1. The major components

Component	Handling things
H2O Frame	• Multi-node cluster with shared memory model. • All computations in memory. • Each node sees only some rows of the data. • No limit on cluster size.

Distributed K/V Store	Objects in the H2O cluster such as data frames, Models and results are all referenced by key. • Any node in the cluster can access any object in The cluster by key.
Data in H2O	• Distributed data frames (collection of vectors). • Columns are distributed (across nodes) arrays. • Each node must be able to see the entire dataset (achieved using HDFS, S3, or multiple copies of the data if it is a CSV file).

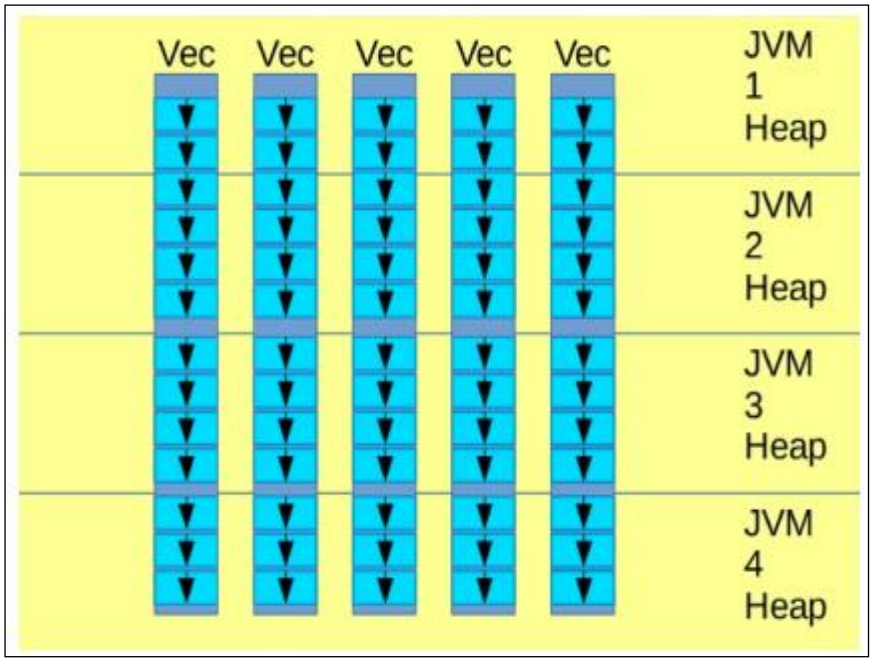

Figure 2. Distributed frame work of H2O

In-memory appropriated K/V store layer: H2O wears an in-memory (not-relentless) K/V store, with *exact* (not lethargic) consistency semantics and exchanges. The memory show is precisely the Java Memory Model, yet disseminated. The two peruses and composes are completely cache able, and average store hit latencies are around 150ns (that is 150 nanoseconds) from a NonBlockingHashMap. Give me a chance to rehash that: peruses and composes experience a non-blocking hash table – we don't experience the ill effects of a great hot-squares issue. Store misses clearly require a system jump, and the execution times are completely determined by the extent of information moved isolated by accessible data transfer capacity and obviously the outcomes are reserved. The K/V store is presently utilized hold control express, all outcomes, and the Big Data itself. You can positively manufacture a dandy chart based calculation specifically finished the K/V store, or incorporate H2O's model scoring into a low-inertness generation pipeline.

A H2O Frame speaks to a 2D cluster of information where every section is consistently composed. The information might be nearby or it might be in a H2O group. The information is stacked from a CSV document or from a local Python information structure, and is either a Python customer relative record, a bunch relative record, or a rundown of H2OVec question. H2O's parser underpins information of different organizations from various sources. The accompanying configurations are CSV, XLS, XLSX, ARIFF, etc.

4. Lifecycle of data science

Now a day all most all are based on the data driven process. Based on the data, we are able to understands or take a measure to the required needs. For data driven process we should know the flow of data processing, the below show the various stages of data processing in terms of life cycle.

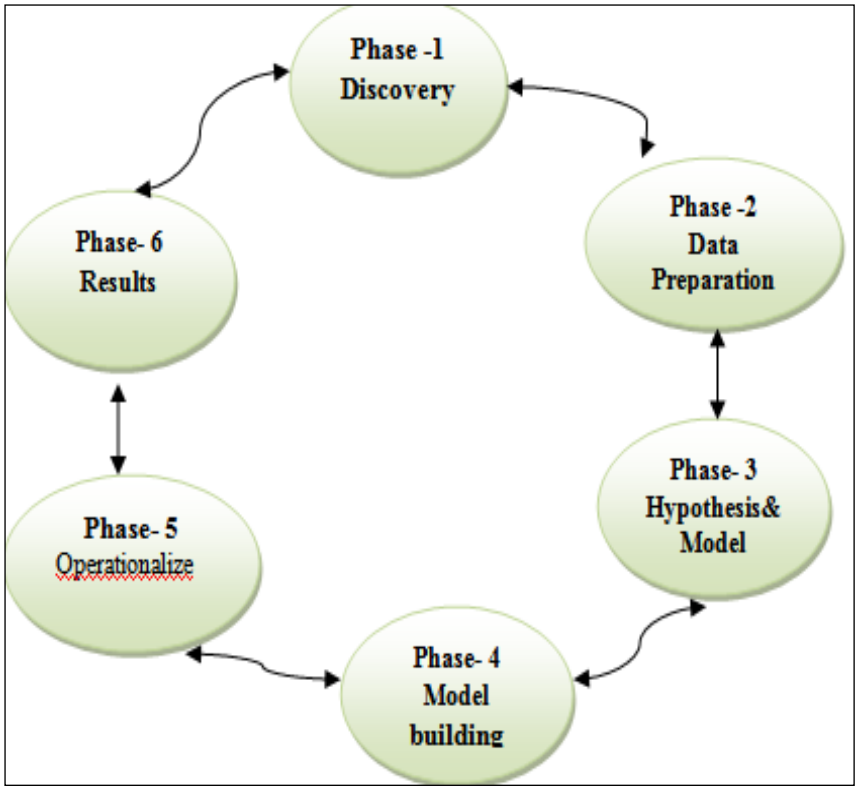

Figure 3. Life cycle of Data Science

Stage 1 - Discovery: Before you start the venture, it is critical to comprehend the different details, prerequisites, needs and required spending plan. You should have the capacity to ask the correct inquiries. Here, you survey on the off chance that you have the required assets display as far as individuals, innovation, time and information to help the venture. In this stage, you additionally need to outline the business issue and detail starting theories to test.

Stage 2 - Data arrangement: In this stage, you require explanatory sandbox in which you can perform investigation for the whole span of the undertaking. You have to investigate, preprocess and condition information before displaying. Further, you will perform ETLT

(extract, transform, load and transform) to get information into the sandbox. How about we observe the Statistical Analysis stream underneath.

Stage 3 - Model Plan: Here, you will decide the strategies and systems to draw the connections between factors. These connections will set the base for the calculations which you will execute in the following stage. You will apply Exploratory Data Analytics (EDA) utilizing different measurable recipes and perception instruments.

How about we observe different model arranging instruments.

R has a total arrangement of demonstrating capacities and gives a decent situation to building interpretive models.

SQL Analysis administrations can perform in-database examination utilizing regular information mining capacities and essential prescient models.

SAS/ACCESS can be utilized to get to information from Hadoop and is utilized for making repeatable and reusable model stream outlines.

Stage 4 - Model working: In this stage, you will create datasets for preparing and testing purposes. You will consider whether your current devices will do the trick for running the models or it will require a more hearty condition (like quick and parallel handling). You will examine different learning systems like grouping, affiliation and bunching to assemble the model.

Stage 5 - operationalize: In this stage, you convey last reports, briefings, code and specialized archives. Moreover, once in a while a pilot venture is likewise actualized in a constant creation condition. This will give you an unmistakable photo of the execution and other related limitations on a little scale before full organization.

Stage 6 - Communicate comes about: Now it is critical to assess in the event that you have possessed the capacity to accomplish your objective that you had arranged in the primary stage. Thus, in the last stage, you recognize all the key discoveries, impart to the partners and decide whether the aftereffects of the undertaking are a win or a disappointment in light of the criteria created in Phase 1.

5. Applications

The H2O venture intends to build up an expository interface for distributed computing, furnishing clients with instruments for information analysis [1]. The product is open-source and unreservedly dispersed. The organization gets charges for giving client benefit and tweaked augmentations. Enormous datasets are too huge to be in any way broke down utilizing conventional programming like R. The H2O programming gives information structures and strategies appropriate for huge information. H2O enable clients to break down and envision entire arrangements of information without utilizing the Procrustean methodology of concentrate just a little subset with an ordinary measurable package. H2O's factual calculations incorporates K-implies grouping, summed up straight models, conveyed irregular timberlands, slope boosting machines, naïve bayes, chief segment examination, and summed up low rank models [4].

Wise constant applications are a distinct advantage in any industry. Profound learning is one of the most sweltering trendy expressions here. New advances like GPUs joined with versatile cloud foundation empower the refined utilization of counterfeit neural systems to include business esteem in certifiable situations. Tech monsters utilize it, for instance, for picture acknowledgment and discourse interpretation. about some certifiable situations from various ventures to clarify when and how conventional organizations can use deep learning continuously applications.

The greater part of H2O.ai's items help to make AI more open, however Driverless AI makes things a stride facilitate by physically robotizing a significant number of the intense choices that should be made while setting up a model. Driverless AI robotizes highlight designing, the procedure by which key factors are chosen to fabricate a model [7].

Figure 4. H2O-Ais-Driverless-AI-Automates

6. Conclusion

In this paper we discussed about the fundamentals of H2O. It is Open Source Fast Scalable Machine Learning Platform for intelligent application implementations. Here we gave information about flow of H2O mechanism in terms of its architecture working nature. Also discussed about the stages of data science how the data driven application implemented. Various applications where the H2O can be help to implement.

References

[1] Candel A., Parmar V., LeDell E., and Arora A., "Deep learning with H2O," H2O. ai Inc, **(2016)**

[2] Erin LeDell et.al., "High performance machine learning in r with H2O," ISM HPC on R workshop, Tokyo, Japan, **(2015)**

[3] http://h2o-release.s3.amazonaws.com/h2o/rel-lambert/5/docs-website/developuser/h2o_stack.html

[4] Aiello, Spencer; Tom Kraljevic, and Petr Maj, with contributions from the 0xdata team, "H2O: R Interface for H2O," The Comprehensive R Archive Network (CRAN), Contributed Packages, The R Project for Statistical Computing (3.0.0.12), **(2015)**

[5] Kochura and Yuriy et al., "Performance analysis of open source machine learning frameworks for various parameters in single-threaded and multi-threaded modes," Conference on Computer Science and Information Technologies Springer, Cham, pp.243-256, **(2018)** DOI: 10.1007/978-3-319-70581-1_17

[6] John Mannes, h2O-ais-driverless-ai-automates-machine-learning-for-businesses

[7] Yuriy Kochura and Sergii Stirenko, et al., “Performance analysis of open source machine learning frameworks for various parameters in single-threaded and multi-threaded modes,” arxiv.org/ftp/arxiv/papers/1708/1708.08670.pdf

[8] “H2O Deep Learning Benchmark,” https://github.com/h2oai/h2o3/tree/master/h2oalgos/src/main/ java/hex/Deep Learning.

[9] Yuriy Kochura, Sergii Stirenko, and Yuri Gordienko, “Comparative performance analysis of neural networks architectures on H2O platform for various activation functions,” 2017 IEEE International Young Scientists Forum on Applied Physics and Engineering (YSF2017), Lviv, Ukraine, **(2017)** DOI: 10.1109/YSF.2017.8126654