Hadoop a Multi Value tool for Academia: A Preliminary Study

Ibrahim Mokdad
School of Postgraduate Studies
Asia Pacific University College of Technology & Innovation
Bukit Jalil, 57000 Kuala Lumpur, Malaysia
ibrahim.mokdad@gmail.com

ABSTRACT
Hadoop is one of the recent attractive tools used to process big data; it is being adopted in various sectors. Academia however is lagging behind. Adopting it in academia would add great values. This paper presents a preliminary study to adopting Hadoop in academia. In this paper Hadoop characteristics that make it the appropriate tool to adopt in academia are presented and discussed. As this would open door for further researches in the academia context.

Categories and Subject Descriptors
B.8.0 [**Performance and Reliability**]: General; H.3.0 [**Information Storage and Retrieval**]: General; J.1 [**Computer Applications**]: Administrative Data Processing - *education*.

General Terms
Algorithms, Management, Measurement, Performance, Design, Reliability, Experimentation.

Keywords
Hadoop, Academia, MapReduce, HDFS.

1. INTRODUCTION
Hadoop is an open source tool that was made for large data set(s) processing; the tool allows for distribution of large data sets to connected nodes (computers). The nodes can scale up to thousands of nodes; each of these nodes offers local storage and processing power -or better put computation-. The nodes can be of commodity hardware and are usually are; they can also be of higher performance hardware. With this there would not be a need to rely on a single supercomputer or even pay for one.

Since Hadoop is intended for commodity machines it is for that robust and scalable [14] as it provides mechanisms to identify and handle the failures as they occur on the application layer ([11], [14]). For the data storage Hadoop uses key/value rather than relational tables; unlike the standard relational databases which by design meant for the structured data, Hadoop however is mainly to deal with the semi and unstructured data [14] which is most of what is generated nowadays [18]. To achieve that Hadoop uses MapReduce. MapReduce is a programming model and a *framework* aimed for parallelization and distribution; it combines the mapping and reducing; in which the functions are mapped over a given data set and the results are then combined (reduced). The process of parallelization is automated; hence the mapping and reducing. One can say that MapReduce splits the data to chunks (not connected to each other, so they would be executed at the same time) the outputs are classed to serve as input to the reduce operation. The framework is also responsible for scheduling task and rescheduling when failures happen

Google now has a patent for the MapReduce programming model which is the basis for Hadoop, some concerns where arose regarding Google filing law suits against MapReduce implementations especially one of Hadoop; thus would mean that Hadoop would come to sleep. However Google granted license to Hadoop and no worries are to arise ([13], [19]).

Hadoop is a tool that has attracted a lot of attention in the past few years. In a post by [3] that influenced the use of Google Trends to reflect some quick fact about Hadoop. The result on the graph clearly reflects the escalading attention on Hadoop (figure 1):

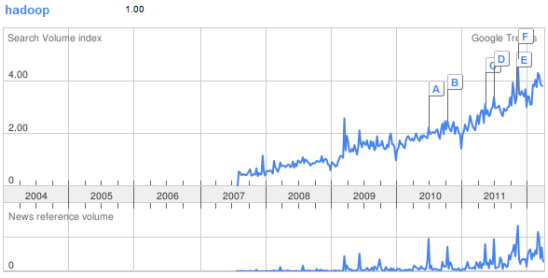

Figure 1: Hadoop on Google Trends.

Hadoop is recently widely adopted by various organizations of different types and sizes for example Adobe use it for various purposes including structured data storage. Ebay too use it; for search optimization [10]. Similarly are many other organizations. Universities should too come to harness the potential of Hadoop. However the questions would be "is it the appropriate choice for universities?" And this is what this paper will come to answer.

2. PREVIOUS WORK
To study whether Hadoop is a proper tool for universities it was necessary to visit some digital libraries to look at whether there are papers that studied the question. "Hadoop" was used as the search keyword in IEEEXplore and the result was 358 results, ranging between conference publications, articles in magazines and journals. The earliest dated back to 2008 and most of the articles were not aimed towards the question at hand. Many of the publications were aimed towards deploying Hadoop to the cloud

and towards enhancing Hadoop, which is normal; especially that it is a newly born tool. Other digital libraries like ACM were also visited but not thoroughly investigated and there was however a publication by [8] that reflected the implementation of Hadoop for the undergraduate computer science students and mentioned that it is quite useful to practice large scale data processing, however the space it occupies and the dedication it acquires might cost. [8] however proposed the cloud as it might not cost much. This finding is seen as one of the reasons for going with Hadoop.

3. HADOOP FOR UNIVERSITIES

On March 2012 Wiki.apache.org/hadoop/PoweredBy page was visited to find out about the universities currently using Hadoop. The page is updated constantly and obliges institutions to state their Hadoop cluster details to be sure of spammers. Surprisingly only 16 universities are listed. Some universities stated their type of use of Hadoop in languages other than the ones known to the author. However most were examined and their type of usage falls under three main categories 1) Teaching courses, like ESPOL University uses Hadoop for its data intensive computing course, 2) Research, like University of Nebraska Lincoln uses Hadoop for part of the physics experiments. 3) Operations, like the GIS research centre of Feng Chia University use it for the sensors' data.

That in mind, what makes Hadoop a proper tool for universities is generally its characteristics. Starting with Hadoop being open source:

3.1 Open source

Its code availability makes it accessible for everyone. This means more room for improvements and changes, rather than the closed and/or owned by one party. Community improvements are limitless; many creative minds out there can contribute if given the chance. Under the academia context, academics and students (with a bit of IT background) can utilize it to base their researches on.

3.2 Cheap

Aside from the fact that it is open source no special enterprise top notch hardware are needed. As long as a machine with a regular processor, hard drive(s) and network card are available that would be it for Hadoop. This characteristic is an attractive one especially under the university context; such that students can utilize it and be able to create their own clusters. Universities for that matter can make use of their laboratories; as to the norm universities always have laboratories in them; the universities can have the same Hardware specifications it has on its laboratories to be used for Hadoop.

3.3 Massive block size

Hadoop File System has massive block size of 64MB (and customizable ([12], [15], [22])); while the most common block size in window file system (NTFS) is 4KB [17], similarly is ext4 of Linux [16]; which makes Hadoop suitable for large data processing. This can be thought of as a bus versus a car; a bus carries a lot of passengers at a moderate speed whilst the car carries fewer passengers at a higher speed. Evidently the bus has the ability to deal with a lot of more passengers than of a car. The bus is HDFS and the other regular OS file systems are the car.

The ability to accommodate large data sets is rather an attractive advantage to universities such that universities can make use of Hadoop for various operations and researches especially Business, Engineering and Scientific related fields like to how University of Nebraska Lincoln is using Hadoop for part of the physics experiments. It can also be used for large scale computing courses like in University of ESPOL University.

Since there is lack in Analytical talent [18] to have hands on Hadoop would mean having hands on large scale data processing part of it is analytics. Let aside the fact that Hadoop is recently been used by many major organization and for students to have hands on it would enable them to be part of the major organizations.

3.4 Super Speed

Hadoop is very customizable it can be tuned in numerous ways; one obvious one is to add more computers. In the minute sort benchmark competition; Hadoop made a new record for the 500GB sort. The competition goes about measuring the time to sort 500GB of records; each record is 100 bytes. The 500GB sort is about finishing the sort in less than a minute. How Hadoop scored that well was through Yahoo's implementation of 1406 nodes each is of 2 Quadcore Xeons, 8 GB memory, and 4 SATA HDDs ([1], [24]). In that sense Hadoop would enable high performance computing for Universities. This begets another characteristic of Hadoop and is scalability.

3.5 Scalability

Hadoop allows for huge numbers of PCs to be added to a given cluster; as reflected from Yahoo's implementation thousands of machines can be used all at once. Size would not become a problem with relation to the performance; all that is needed in Hadoop when one is out of space is to add another machine or disk to the mix. Universities can connect various computer laboratories together to create one huge cluster; however scalability as to [8] can be of a hassle since it is generally accommodated with space. For that and according to [8] the cloud can be used if Hadoop's cluster grew to be really huge.

3.6 Data Integration

Hadoop makes for a proper solution where there is machine generated data; Like in the GIS research center of Feng Chia University, in which they use Hadoop for sensors' data [10]. Another example would be of Tennessee Valley Authority which is the US largest public power provider. The authority is actually using Hadoop to process the data captured from field devices ([4], [21]).

Universities generally have various machines that generate data and are commonly represented with Surveillance Systems, RFID Systems, Biometric systems and of course Computers. Universities with Hadoop can come to process that generated data.

Being able to process both the machine generated data and human generated data would mean that both data can be merged together and processed as one; thus more insights and better understanding of the possessed data. In that sense the university administration can integrate the various systems within the university to enhance operations' efficiency and quality.

The data integration could extend to merging it together with social networks data. For example Universities can trace tweets of various institutions and organizations to keep up with the latest technological advancements and/or latest discoveries; the system would filter tweets according to the type of tweet; for example

tweets concerning physics department would be sent to the physical department personnel.

3.7 Distributed Computing

Distributed is all about being connected to the network (local or public); hence reliability has to be ensured such that when crash happens it would not jeopardize the whole network; there should be a proper protocol that determines the actions that should take place when a problem occurs in the distributed system. Clearly there is a lot to consider; one has to consider network issues, communication issues, etc. Going with Hadoop most of those considerations are automated; Therefore if adopted in universities it would not require much of effort to get it running and to run applications on it.

It should be noted though that the full understanding of Hadoop requires some understanding in various fields and are; distributed computing paradigm, parallel computing paradigm, file systems, programming and Functional programming paradigm.

3.8 Fast Connection

When Hadoop is compared with other distributed systems like BOINC (Running for example World Community Grid, Rosetta@Home, SETI@Home) and folding@Home; those distributed systems evolve around sending chunks of work to an internet connected computer which would formulate a server client scenario; and when the computer is idle it uses the idle/remaining processing power to process that chunk of data (Cycle Scavenging) and when it is done it is sent to the corresponding sender ([6], [9], [23]).

From the network perspective the network in the likes of distributed systems mentioned is the Internet and is always slower than Hadoop, In a Hadoop cluster one can even use a gigabit network.

3.9 Locality and Security

In the previously stated type of distributed systems (ex: BOINC) the computers are of the general public volunteering their processing powers; there is a chance that the results would be played around with or erroneous. How this issue is overcome in those distributed systems is through sending the same chunk of data to various idle computers and the results are compared and validated [5].

The duplications that take place in the kinds of distributed systems mentioned earlier are meant for validating the results and are executed anyways. Whilst in Hadoop the duplications that take place are to have an available service and are used when failures occur. The fault tolerance techniques in Hadoop are very powerful, that when compared to other parallel SQL DBMS Hadoop would take the lead [20]. Universities can process their data without worrying much of whether failures can happen.

Another difference is that when processing the data in Hadoop it is local to the machine. Unlike the other distributed systems in which the data moves where the computation is (data moves to the volunteer). In Hadoop the data is distributed within the cluster and not through the internet; that allows for larger chunks of data and then when it resides, the code moves about (the cluster) [7]. The universities would not have to worry about their data being exposed, such that with Hadoop confidentiality of data is ensured.

Other important difference to consider is that the data transferred between the previously mentioned distributed systems is generally scientific and does not harm being read or exposed; and that would not work for confidential organizational data. Also In the previously mentioned distributed systems one cannot guarantee the specifications and the performance of the machines (or predict) whilst in Hadoop the machines are local in house/cloud; with total control over them.

3.10 Smart

Hadoop can become a smart machine simply by adding Mahout to it. Mahout is an Apache project; it is an open source machine learning library. Mahout has community support that would allow for further contributions and wider range of ideas to be put into it. It is a proper option for machine learning over larger scale of data sets. It has various implementations of various algorithms. The current ones serve for various implementations to name a few clustering, dimension reduction, vector similarity, classification, collaborative filtering and many others [2].

4. CONCLUSION

What have been presented are characteristics that would attract universities. The characteristics mentioned make Hadoop a great choice for universities' large data processing; it brings forth many values. As it is cheap, open source, reliable, proper for machine generated data, proper for academics and students to experiment on and is local to the institution.

Few universities are making use of Hadoop; universities should come to capture its values. This study opens door for further researches to take place one of which is using Hadoop as a performance assessment tool in academia.

5. ACKNOWLEDGMENTS

Many thanks to Mr. Mohamed Anwar of EXA technologies for his valuable pointers and Mr. Zailan Arabee of UCTI for his advices.

6. REFERENCES

[1] Anand, A. 2009. *Hadoop Sorts a Petabyte in 16.25 Hours and a Terabyte in 62 Seconds*. [online] Available at:< http://developer.yahoo.com/blogs/hadoop/posts/2009/05/hadoop_sorts_a_petabyte_in_162/> [Accessed 1 Mar 2012].

[2] Apache Mahout, 2011. *Algorithms* [online]. Available At :<https://cwiki.apache.org/MAHOUT/algorithms.html> [Accessed 13 Mar 2012].

[3] Aslett,M. 2010. *Google Trends: Hadoop versus Big Data versus MapReduce*. [online] Available at:< http://blogs.the451group.com/information_management/2010/11/24/google-trends-hadoop-versus-big-data-versus-mapreduce/> [Accessed 1 Feb 2012].

[4] Bisciglia, C. 2009. *The Smart Grid: Hadoop at the Tennessee Valley Authority (TVA).* [online] Available at:< http://www.cloudera.com/blog/2009/06/smart-grid-hadoop-tennessee-valley-authority-tva/> [Accessed 25 Mar 2012].

[5] BOINC, 2009. *Security issues in volunteer computing* [online]. Available at:<http://boinc.berkeley.edu/trac/wiki/SecurityIssues> [Accessed on 7 Mar 2012].

[6] BOINC, 2011. *How BOINC workds* [online]. Available at:<http://boinc.berkeley.edu/wiki/How_BOINC_works> [Accessed on 7 Mar 2012].

[7] Bortjakur, D., 2011. *HDFS Architecture Guide* [online] Available at:<http://hadoop.apache.org/common/docs/current/hdfs_design.html> [Accessed on 3 Mar 2012].

[8] Brown, R., 2009.Hadoop at home: large-scale computing at a small college. *Proceeding SIGSE '09 Proceedings of the 40th ACM technical symposium on Computer science education.* New York, USA. 41(1),pp-106-110.

[9] Folding@home, 2011. *About Folding@home* [online]. Available at:<http://folding.stanford.edu/English/About> [Accessed on 7 Mar 2012].

[10] Hadoop Wiki. 2012. *PoweredBy*. [online] Available at:< http://wiki.apache.org/hadoop/PoweredBy> [Accessed 15 Mar 2012].

[11] Hadoop, 2011. *Welcome to Apache™ Hadoop™!*. [online] Available at:< http://hadoop.apache.org/> [Accessed 10 Apr 2012].

[12] Hadoop, 2012. *Cluster Setup*. [online] Available at:< http://hadoop.apache.org/common/docs/current/cluster_setup.html> [Accessed 29 Mar 2012].

[13] Jeffrey, D. and Ghemawat, S., Google Inc., 2010. *System and method for efficient large-scale data processing*. U.S. Pat. 7,650,331.

[14] Lam, C. 2010. *Hadoop in Action.* Greenwich: Manning Publications Co.

[15] Lipcon, T. 2009. *7 Tips for Improving MapReduce Performance*. [online] Available at:< http://www.cloudera.com/blog/2009/12/7-tips-for-improving-mapreduce-performance/> [Accessed 15 Mar 2012].

[16] Mathur, A. et. al. 2007. The new ext4 filesystem: current status and future plans. *Proceedings of the Linux Symposium*. Ottawa, ON, CA: Red Hat.

[17] Microsoft Support, 2012. *Default cluster size for NTFS, FAT, and exFAT*. [online] Available at:< http://support.microsoft.com/kb/140365> [Accessed 2 April 2012].

[18] Manyika , J.et al., (2011) *Big data: The next frontier for innovation, competition, and productivity,* [online]McKinsey Global Institute. Available at < http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation> [Accessed 13 Apr 2011].

[19] O'Malley, O., 2010. License for Google's patent, *Hadoop-general mailing list archives*. [online] 23 Apr 2010, Available at:<http://mail-archives.apache.org/mod_mbox/hadoop-general/201004.mbox/%3C121803A3-CFB9-489B-96EF-027234E55D25@apache.org%3E> [Accessed on 23 Apr 2012].

[20] Pavlo, A , at al. 2009. A Comparison of Approaches to Large-Scale Data Analysis. *SIGMOD '09 Proceedings of the 35th SIGMOD international conference on Management of data*.USA.

[21] Risenberg, D. 2009. *Open-source Hadoop powers Tennessee smart grid*. [online] Available at:< http://news.cnet.com/8301-13846_3-10393259-62.html> [Accessed 25 Apr 2012]

[22] Sam, 2010. *Changing the block size of a dfs file in Hadoop*. [online] Available at:< http://stackoverflow.com/questions/2669800/changing-the-block-size-of-a-dfs-file-in-hadoop> [Accessed 1 May 2012].

[23] Sonmez, O., Grundeken, B., Mohamed, H. and Iosup, A., 2009.Scheduling strategies for cycle scavenging in multicluster grid system. *CCGRID*.,pp-12-19.

[24] sortbenchmark, 2009. *Sort Benchmark Home Page*. [online] Available at:< http://sortbenchmark.org/> [Accessed 1 Apr 2012].