Big Learning

Proseminar Data Mining

Sergiu Soima
Fakultat fur Informatik
Technische Universitat Munchen
Email: sergiu.soima@tum.de

***Abstract*— Big Data is a fast growing industry that can change how business will be done. I have chosen to research this field to able to "Think Big". Hadoop and MapReduce are widely used by big companies and it is worthwhile to understand how they work, and in which cases they are better than a relational Database. MapReduce is like a devide and conquer strategy. It splits the data into smaller blocks that can be processed in parallel and then puts the results of each block together to form a final result. Hadoop is a platform that implements MapReduce and eases the job of the programmers so that, for instance, they do not have to care about how jobs are shuffled and splitt. In most cases it is enough to just implement two functions map and reduce. In order to achieve this, Hadoop uses a filesystem specially designed for its needs. There are also platforms like Hive that has a SQL like query language that offers the most widely used DML and DDL functionalities in Hadoop. Using Hadoop with scripting languages is also possible by using the Apache Pig platform. Apache Pig helps create MapReduce queries more easily than with Java. The root of Apache Pig is its scripting language "Pig Latin", that helps developers to query MapReduce jobs more efficiently and comprehensively with an additional abstraction layer. The compiler can track the data of MapReduce jobs and optimize the individual jobs.**

I. Introduction

The contemporary world has drastically changed, with data growing at large scale because of virtual worlds, wikis, blogs, e-mail, online games, VoIP telephone, digital photos, instant messages (IM), tweets, traffic systems, bridges, airplanes and engine, satellites, weather sensors" [1]. The use of Big Data in enterprises and organizations has already begun. Big Data presents concepts, technologies, and methods for using the almost exponentially increasing volumes varied information even better as a sound and timely basis for decision making, increasing innovation and the competitiveness of companies. Big Data comes in where conventional approaches to information processing reach human limits. Big Data supports the economically viable production and use of decision-relevant findings from qualitative diverse and differently structured sources of information that are subject to rapid change and are on an unprecedented scale. Big Data reflects the technological progress of recent years, and it includes developed strategic approaches and technologies, IT architectures, methods, and procedures. With Big Data managers will receive a significantly improved basis for the preparation of time-critical decisions. From the business perspective Big Data illustrates how in the long term data can turn to a product. Big Data opens the prospect of the "industrial revolution of data", while cloud computing industrialized IT operations [2]. In the next pages I will describe how MapReduce can process Big Data. I will describe real life use cases and the Hadoop platform that offers an infrastructure for the developers to build MapReduce jobs. I'll also discuss Hive and Pig which build on the Hadoop platform offering languages to Query data. Additionally I'll describe the challenges of dealing with Big Data.

II. The four V's of Big Data

The 4 Vs of Data Mining are Volume, Variety, Velocity and Veracity. In the following chapter the main source of information is [3].

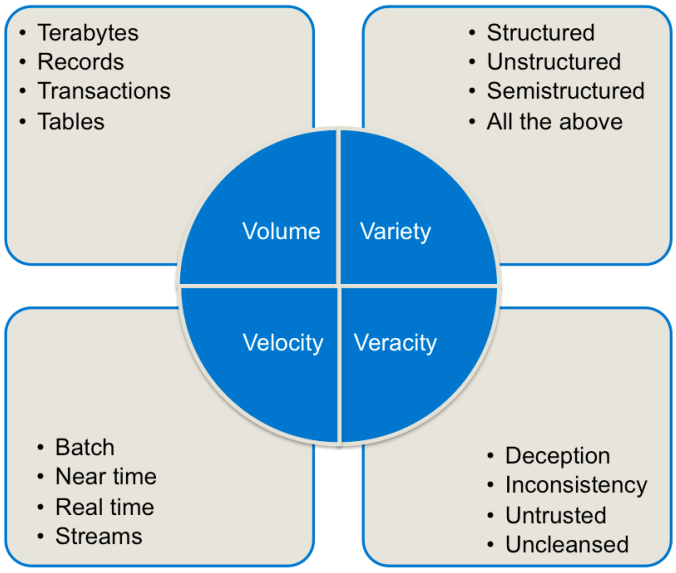

Fig. 1. The 4 Vs of Big Data

A. Volume

A lot of Organizations and Companies dispose of gigantic data volume starting from Terabytes scaling up to Petabytes. Companies have to manage huge amount of records, files and measurements. The data volume is significantly growing and this growth is perfectly described in an IDC(Gantz et al.,2011) survey: "Like our physical universe, the digital universe is something to behold 1.8 trillion giga-bytes in 500 quadrillion files and more than doubling every two years." If you open a 100 megabyte CSV in Microsoft Excel, the charging process takes already a few minutes on an average-equipped desktop

PC. On a mobile device or very simple Hardware like RaspberryPI, FritzBox or, Aruduino it is not even possible to start such processes. It is clear that bottleneck is the power of the CPU and the RAM. The storage drive itself is sufficient in the most cases for 100 megabytes easily. That's why for large datasets, in most cases "the Big" is the processing, not the storage memory.

B. Variety

Companies work with a lot of various data sources and data formats. Data come out of various sources. Some of them are unstructured like presentations, texts, videos, tweets or blogs. Semi-structured data is generated by communication between Machines. Structured data is used by companies as data holders. There are a lot of forms data could take. It could be a database, a file, or a mail. In fact any kind of observation can be transformed to data. Data is increasingly being supplemented by external data, such as social networks. Authors and veracity for external data are not always clear, which can lead to inaccurate results when analyzing data.

C. Velocity

With velocity is given by the IO speed when we access the data through a virtual machine. Huge amounts of data need to be evaluated more quickly, often in real time. The processing speed has to keep up with data growth. Thus the following challenges are connected: analysis of large data sets with answers within seconds, data processing in real-time, data generation and transmission in high speed. Having 500 megabyte hard drives was previously a sensation. Today we have gigabyte sticks in fingernail size. This example may seem a bit simple, but nevertheless it reflects the rapid development of storage hardware. With the options namely, the demands. 15 years ago 20 gigabytes for a PC was perfectly adequate. Today most of the PCs you can buy have at least 2 terabytes. With the increase o the Volume we also have a higher demand.

D. Veracity

"Not everything that counts can be counted, and not everything that can be counted counts." - Albert Einstein The Data used for analytic research have to be qualitative to get valuable information. Poor data Quality costs US consumers about $3.1 trillion per year according to an IBM survey. This uncertainty can be reduced by using more less reliable sources and combining them to get a more accurate result [4]. The amount of available data is increasingly growing. But is this data reliable? Here are some examples how generated data can be distorted:

- Advertising and spam is just focused on what people should see and not reliable.
- Automatically translated texts that are often grammatically false or have a changed meaning.
- Outdated or incorrectly categorized search results.
- Targeted misstatements or lack of information.

III. MapReduce

MapReduce is a programming framework for processing unstructured Big Data. It was developed to serve the purpose of processing enormous amount of data like clickstreams, big text files or digital versions of books.Each MapReduce program goes through the following three stages: map, shuffle and reduce. When using MapReduce the user needs to define a map function as well as a reduce function, which are able to run simultaneously on multiple machines [5]. The two functions are defined as followed:

- map (k1,v1) in list(k2,v2)
- reduce (k2,list(v2)) in list(v2)

The user needs to provide the map function with a key and a value.The map code will be applied on these parameters and it will produce an intermediate result in form of a key/value list. In the shuffle stage all the (k2,v2) records produced by the map function are collected and reorganized so that the records with identical keys will be merged into a (k2,list(v2)) record. Now we have a list of values of all the map outputs with an identical key. For each value, the reduce function will return a zero or another positive value . Usually the reduce function just sums up the list of v2 values for each key [6].

A. MapReduce in action

Let's simulate a concrete example of a MapReduce job. In this example we will count the number of occurrences of each letter in a file.

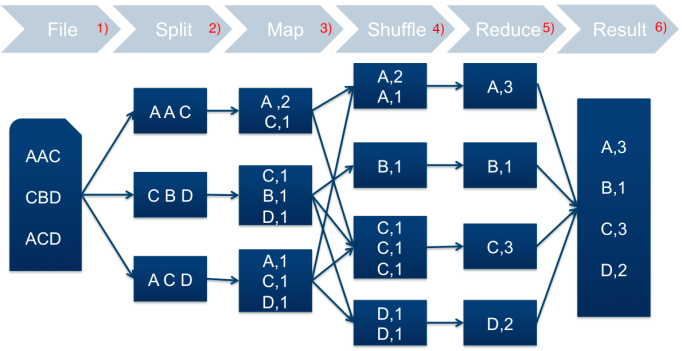

Fig. 2. Mapreduce steps

Let's say we have a file as shown in Fig. 2.1). The first step is to split the initial file to get more files that can be distributed to parallel running machines as shown in Fig. 2.2). After the files are distributed a map function, which counts the letters and produces a (k,v) pair for each letter, will process the new files. The result of the map Function is shown in Fig. 2.3). The next step is to shuffle the results by key like in Fig. 2.4). With this procedure we obtain new file for each key where we store all the (k,v) pairs which have the same key. The Reduce function is shown in Fig. 2.5). In this example we reduce by just summing up the paired values. In Fig. 2.6) we can see the final output which is the number of apparitions of each letter.

B. Advantages

The amazing part of this MapReduce job is that we can scale it for extremely big files. Customizing plays a major role when it comes to implement Map-Reduce jobs. The advantage of concrete implementations like Hadoop lies in the fact that it provides libraries with which jobs can be efficiently and easily programmed. If we use a platform like Hadoop the tasks will be automatically parallelized and distributed to different machines to get solved. Such a evaluation may sound easy but it loses its straightforwardness if they must be programmed from the Ground up so that processing can take place over a distributed cluster.

C. Use Cases

This next chapter focuses on showing real life cases where big Data is used and is inspired from the following book [6]. Map reduce has a wide range of applications. MapReduce is "the core of the distributed programming model" for a lot of Big Data applications so let's check some of the industries that it influences:

1) Social networks: Facebook, Twitter, LinkedIn are just a few examples of businesses that connect friends, groups, communities and businesses. These companies are using MapReduce to discover common friends and followers, posts that could be interesting for some users, best news feed suggestions, advertisement based on user interests.Interesting features like profile views or who read your message/post can also be processed with Hadoop.

2) Media and entertainment: Netflix and Hulu use Hadoop and MapReduce to help them find out the most popular videos for individuals according to consumers requirements and interests. By using Hadoop users can be analyzed in detail and influenced by giving them proper content to make them happy and raise their interest to use the site to use it on a daily basis. MapReduce is used to identify similar videos by using their metadata like its category, likes, owner, user groups that are watching it and clickstream logs.

3) E-commerce: E-commerce shops like Amazon, Walmart and eBay are heavy users of the MapReduce model. They use MapReduce to discover popular products according to the interests of the clients or their purchasing behaviours. In order to create this product recommendations they analyze clickstreams, website logs, purchase history and how users interact with the website. Reviews and comments reflect how customers feel about the product. This behaviours were previously analyzed by humans but by using the MapReduce model for text mining companies get a very accurate feedback of how the customers perceive their products. These big companies have a very large amount of Data that can't be discovered entirely by the users. That's why recommendations can help users to discover more products that they could need. Search logs of an e-commerce shop can be analyzed to find out which products are missing and which products are the most popular among all users.

4) Fraud detection and financial analytics: Hadoop and MapReduce is used to analyze Bank transactions and application logs to identify possible frauds, trends or business metrics. Insurance companies use their logs to find out the price and profitability of an insurance by analyzing age, gender, past and they go even further by analyzing their families and group of friends if such data is available. Following trend lines and stalking each others competitors keeps the companies up to date with the evolution of their industry and helps them to determine if they're in heading the right direction.

5) Search engines : MapReduce can help search engines to determine the most popular searches in different situations and the best search results. Mapreduce can also be used in search engines to give users an advice if they misspell a word in their search, or give them a better keyword advice.

6) Data migration: Because migrating large amount of data in the traditional way of bulk copy takes to much time, MapReduces distributed model use the map function to split the data into chunks that can be transferred in parallel. In the end the data just has to be combined back to its initial state. This can be easily done by using the reduce function. The integrity of this process is ensured through checksums.

IV. Hadoop

In 2006 Yahoo launched the open source framework "Hadoop" as a free implementation of the MapReduce frameworks forth. This development influenced today's Big Data significantly. Meanwhile, under the control of the Apache Foundation, Hadoop can save with new open source components very large unstructured data on loose clusters, use cheaper servers, and less power for data analyze. Several providers even allow others to use their cloud solution, so that the technical and financial hurdles for Big Data applications are low, and the solution to large peak loads still remains scalable. Hadoop is operated in a scale-out clusters, where one can easily add new nodes, so that an increase or decrease in accordance with the requirements can be carried out quickly. Large companies like Yahoo operating as cluster with over 4000 nodes [7]. Instead of relying on expensive hardware, Hadoop is designed to use cheap systems [8]. Hadoop believes that "Moving Computation is Cheaper than Moving Data" [9]. The concept of data locality plays a major role in Hadoop. Unlike traditional applications, in which the Data are made available to the program, the programmed code in Hadoop is distributed to the cluster to keep the need for data transport minimal. Big Data tools like Hadoop are very powerful, but it can get very difficult to program with. This is the reason why big companies strive to simplify it by developing software and libraries like Hive to add a "SQL like interface" that run on Hadoop framework and can run jobs in parallel.

A. HDFS

"The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware." [9] HDFS is a master-slave architecture. The master is the NameNode and the JobTracker and the slave component is

the DataNode and the Jobtracker. The NameNode manages all file system metadata, including directory structures, files and file accesses of clients. Data nodes manage the memory that is allocated to the respective nodes in the cluster.The NameNode is the single point of availability failure so if the NameNode goes down the DataNode will not know, how to make sense of the blocks. The NameNodes ram is backed up to a file by the secondary NameNode as a checkpoint. The nodes software is written in Java so that any machine that supports Java can run them. It is not necessary to create a separate partition for HDFS, because it relies on existing file systems like the current Ext4 (Fourth Extended File System). A significant difference is that the block size in an HDFS is on average 64 up to 128 megabytes. Traditional file systems use 1 up to 64 kilobyte blocks. If the NameNode receives a file from the client that needs to be stored in the file system it will require two further information: firstly the just mentioned block size, in which the file will be split and secondly the number of replicas that will be distributed across the cluster. The NameNode picks out so many data nodes as the client requests through the replica number of the NameNode. Their addresses are returned to the client, so that it can begin the write of the data nodes. The Client will write only one data node that passes the data to the other nodes.The NameNode sorts the data nodes for an efficient transmission before the client passes the required parameters (highest data throughput, best network connectivity, the lowest instantaneous Work-Load), so that the client knows which data node is best accessible and useful. The most common setup is three replicas that HDFS randomly distributes [9]. Now let's illustrate a concrete example on the following: Let's assume we have 10 machines and 1 Tb of data. We can split that Data into blocks and distribute 100GB on each machine. Assuming that the read speed of the hard disk is 100Mb/s we can download the whole file in just 16,6 minutes. If we would not spread the data and use a single 100Mb/s hard disk, we would need 2,7 hours.

B. Apache Hive

"The Apache Hive data warehouse software facilitates querying and managing large data sets residing in distributed storage. Hive provides a mechanism to project structure onto this data, and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL." [10] Contrary to many assumptions Hive is not a database. It uses databases to store meta-information on its tables and schemes, but it accesses to the data that has to be processed in the HDFS. This processing takes place in the form of Map-Reduce jobs that Hive individually designs and executes for each query. This leads to a high latency for individual queries so that Hive is especially suitable for large, unchanging data. It is useful for big amounts of data that normal Relational databases couldnt́ even process. For such batch data processinpg Hive is ideally suited. The advantages of MapReduce process come to this type of processing and analytic tasks to full advantage, because the processing time can be extrapolated. Hive has commands which can be subdivided into the following groups:

- DDL allows creation, modification and deletion of tables schema and views. It includes commands such as CREATE TABLE or ALTER TABLE.
- DML enables you to load and write data to and from Hive out(LOAD,INSERT,UPDATE,DELETE).
- The query language (Hive QL) allows selecting, grouping, linking and transforming data within Hive(SELECT, GROUP BY, JOINS... For more information visit [11]).

C. Apache Pig

The next chapters informations are taken from [12]. Pig is a platform for analyzing large data sets. The language was originally developed in 2006 by Yahoo, to simplify querying large amounts of data. Pig's language Pig Latin is a simple query algebra that enables one to express data transformations such as merging data sets, filtering them, and applying functions to records or groups of records. It is not a declarative language as SQL but a flow language. It's rich multivalued nested operations that allow to perform strong functions on large data sets. MapReduce requires a Java programmer. Thatś why pig opens the system for users that are familiar with scripting languages like python, ruby, or php. Pig Latin has a very easy syntax that allows allocations of data operations or transactions but doesn't have loops and control structures. There is no change in the execution of the jobs when comparing Apache Pig with MapReduce. However Pig accelerates the creation and management of MapReduce jobs. Since Pig Latin is more efficient and easier to use than Java, all companies should view the product, read the data on MapReduce queries in Hadoop. In fact, a developer does not even have to master Java to create using Pig Latin efficient MapReduce job and start a data analysis. A Pig Latin statement runs as follows: The data is read from the data system, then the data is transformed, and finally displayed or stored. Here developers have also the possibility of reading and processing files directly from the file system or from CSV files. The data is read via LOAD statements that contain the file path of the data that has to be analyzed later with Hadoop. Pig commands are always executed on data and data relationships.Pig Latin is also based on other programming languages and one can therefore find a lot of well-known functions like JOIN in SQL or FOREACH in Java. Therefore, it is possible to learn Pig relatively quickly. Relations in Pig are equivalent to tables in relational databases. The smallest unit is in Pig is the field, which is represented by a name and a data type. Based on the simple data types there are more complex data types like bags. A bag is laced by multiple tuples.

D. Apache Hive or Pig

Both of them have some similarities. If companies want to equip the MapReduce framework with a powerful query language, they should choose the Apache Hive extension of Hadoop. By using Hive developers can create Hadoop queries

by writing SQL like code. The data stored in HDFS can be directly queried with Hive. Hive is SQL oriented, and it can solve a lot of tasks in a single line.
On the other side Pig always needs more lines of code to achieve a better results optimization. This increases the overview, as with Hive queries can get complicated very quickly. However, Hive and Pig can not replace each other. They complement themselves.

V. Challenges and opportunities of Big Data

In the following I will present you challenges and Opportunities when analyzing large data sets. I will focus on data privacy because in my point of view it is the most important aspect when dealing with peoples data.

A. Privacy

A very important issue when dealing with Big data is to analyze the data according to the laws of different countries. A lot of countries protect their citizens privacy. Still it is more and more difficult to ensure data security when the data is shared and multiplied across the internet. Some organizations tried to anonymize the data through different procedures so that the data is distanced from real life identities. Scientists have although proofed that this anonymisation does not work because they managed to re-identify the specific identity that generated the data. However as Betsy Masiello and Alma Whitten have noticed if there is some anonymous data it could be re-identified but if the data is big enough there will be more uncertainty and the data will be unreliable when trying to bring it back to an individual.
It is still not clear whether reducing the amount of data is beneficial or not because then the social value of data such as "public health, national security and law enforcement, environmental protection, and economic efficiency" is also minimized. A balance should be found for the privacy by taking into account the usability of the data and its privacy risk. Some researchers have proven, that by giving users the feeling of control they are more opened to sharing and when consumers see the term "privacy policy" they believe in case of websites, that the companies will not share their personal data. In reality "privacy policies often serve more as liability disclaimers for businesses than as assurances of privacy for consumers". In a lot of situations the user can't take the right decision. It is sometimes because of misplaced beliefs or because they're not prepared to take that decision. Is it the right thing to take the decision for them? [13]

B. Scale

Managing big amount of data is more challenging than it was before. Some time ago these challenges weren't so big because of the rapid speed increase of the processors. Right now the amount of data is scaling faster than the CPU speeds that are static due to power constraints. That's why right now the only solution is to build better processing systems that also manage the power consumption of the processor [14].

C. Science

As nature.com says in its scientific Article [15], the new discovery of genomes changed the way life scientists handle data. Scientists use more often big data to prove their theories. The European Bioinformatics Institute is one of the largest biology-data repository. It stores 20 petabytes of data about genes, proteins and small molecules. The CERN particle accelerator has even more data. As described in this Article CERN's particle-collision events generate each year rough 15 petabytes of data.

D. Healthcare

Besides the challenges of Big amount of data there are a lot of other useful opportunities that need to be exploited. Information from digital records could help doctors to give a more accurate diagnosis, to reduce healthcare costs and generally increase the quality and efficiency of the service [16]. McKinsey estimates that Big Data in healthcare saves $ 300 billion every year in the US [14].

VI. Conclusion

Big Data is a very interesting industry with a great potential. Big Data is the core of some industries like social media sites, e-commerce, or search engines, but this is just the beginning. Most companies do not analyse their data enough and some of them don't even think about what they could achieve if they started to "Think Big". It is not easy to analyse data but platforms like Hadoop make it much easier and also cheaper than in the past. Hadoop is a very good solution when it comes to analysing Big Data and it also has an amazing community that can help you start using it. Hadoop uses the MapReduce algorithm that has two main functions, map and reduce. We use the map function to process the data and the reduce function to aggregate the results. Hadoop uses its own filesystem, HDFS, which is specially designed to have best performance when dealing with MapReduce jobs. This amazing platform also has great extension like Hive and Pig, which ease the job of the developers. Hive offers the SQL like language HiveQL, which is very easy to learn because most developers are already familiar with SQL. Pig offers its own scripting language Pig latin and makes the lives of developers who are familiar with python, ruby, or php easier. It can't be said that any of these two alternatives are better than the other because they rather complement each other. Big Data comes with a lot of challenges and in my point of view the biggest one is privacy. People don't have a picture of what really happens with the data they constantly generate. Some of them don't even care and others rely on the word of mouth. Data scale is natural but right now data is growing faster than CPU speed and we need new approaches to be able to manage the process costs. Healthcare is an obvious interest of human beings and data mining can increase the quality of it.

REFERENCES

[1] K. Ayankoya, A. Calitz, and J. Greyling, "Intrinsic relations between data science, big data, business analytics and datafication," in *Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology*, ser. SAICSIT '14. New York, NY, USA: ACM, 2014, pp. 192:192–192:198. [Online]. Available: http://doi.acm.org.eaccess.ub.tum.de/10.1145/2664591.2664619

[2] A. ODriscoll, J. Daugelaite, and R. D. Sleator, "big data, hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 774 – 781, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046413001007

[3] R. Birke, M. Björkqvist, L. Y. Chen, E. Smirni, and T. Engbersen, "(big)data in a virtualized world: Volume, velocity, and variety in cloud datacenters," in *Proceedings of the 12th USENIX Conference on File and Storage Technologies*, ser. FAST'14. Berkeley, CA, USA: USENIX Association, 2014, pp. 177–189. [Online]. Available: http://dl.acm.org.eaccess.ub.tum.de/citation.cfm?id=2591305.2591323

[4] T. Lukoianova and V. Rubin, "Veracity roadmap: Is big data objective, truthful and credible?"

[5] KEINE, A. data SQL, B. . MapReduce, and T. Coffing, *Aster data SQL and MapReduce, Book 12*, ser. Tera-Tom genius series, J. Nolan, Ed. [S.l.]: Coffing Pub., c2013, elektronische Ressource.

[6] K. Tannir, *Optimizing Hadoop for MapReduce*, 2014.

[7] *Hadoop Wiki*, 2013 (accessed 29 May, 2015). [Online]. Available: http://wiki.apache.org/hadoop/PoweredBy

[8] X. Pan, J. Tan, S. Kavulya, R. Gandhi, and P. Narasimhan, "Ganesha: blackbox diagnosis of mapreduce systems," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 3, pp. 8–13, 2010.

[9] *HDFS Architecture Guide*, 2013 (accessed 16 May, 2015). [Online]. Available: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[10] *Apache Hive website*, 2014 (accessed 17 May, 2015). [Online]. Available: https://hive.apache.org/

[11] Y. Huai, A. Chauhan, A. Gates, G. Hagleitner, E. N. Hanson, O. O'Malley, J. Pandey, Y. Yuan, R. Lee, and X. Zhang, "Major technical advancements in apache hive," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '14. New York, NY, USA: ACM, 2014, pp. 1235–1246. [Online]. Available: http://doi.acm.org.eaccess.ub.tum.de/10.1145/2588555.2595630

[12] *Mortardata: Pig Help and Resources*, 2015 (accessed 23 May, 2015). [Online]. Available: http://help.mortardata.com/technologies/pig/pig_help_and_resources#toc_0PigonMortar

[13] O. Tene and J. Polonetsky, "Privacy in the age of big data: a time for big decisions," *Stanford Law Review Online*, vol. 64, p. 63, 2012.

[14] N. Ammu and M. Irfanuddin, "Big data challenges."

[15] *Biology: The big challenges of big data*, 2015 (accessed 09 June, 2015). [Online]. Available: http://www.nature.com/nature/journal/v498/n7453/full/498255a.html

[16] A. A. S. Hamid Bagheri, "Big data: Challenges, opportunities and cloud based solutions," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 2, pp. 340 – 343, 2015.