

# Technology And Methodology Analytics Of Big Data

Surekha Lanka, Sidra Eshan

MTech (CST), MSc (CS), (PhD), Himalayan University,  
 MS (cs) Faculty of Computing and Information Technology, KAA University,  
 surekha.lanka@gmail.com

**ABSTRACT:** The world of the Information Technology has changed in a dramatic way over the past few decades. With the introduction of smartphones and the use of the internet as a part of daily life, the large amount of the data is created. Generation of the data commonly known as the Big Data has created a challenge for the IT professionals. On the other hand, despite the challenges posted by the Big Data it has potential of presenting a great opportunity for the businesses and the networks to improve and optimize their services. In the research review, different technologies of the Big Data, challenges, technical details and software platforms are discussed in detail.

**Keywords :** Complexity,framework,Mapreduce,traffic management,pipeline, sensitive ,AMQP, Memcached,Unstructured, control strategies, matrix generator,data integration,extraction, cleansing, JAVA, enormous information.




## BASIC APPLICATIONS OF HADOOP

Big Data' is a term which generally refers to the sets of data whose volume, complexity and rate of evolution or growth are difficult to be managed, analyzed and processed by conventional methods. The 'Big Data' is used widely everywhere nowadays from news articles to professional magazines. The term generally means the collection, management and processing of the large amount of the data impossible to be achieved using conventional techniques. Then Scaling, error handling, self mending and defending large scale of data maintained by Hadoop. So data will be structured form or in unstructured form. When the data is larger than traditional systems are unfit to handle it. Thus, Hadoop comes into the procure. Actually Hadoop is based on a Java programming framework where larges of data set in a distributed environment. The Hadoop library contains major key components those are HDFS and Mapreduce [1]. Hadoop is very flexible and permit us to change the behavior without changing the source code.

## DATA ANALYSIS OF HADOOP TECHNOLOGIES

The existing open source data analysis of Hadoop technologies is initially to analyze the data of stock which are automated frequently. Those are MapReduce,Pig and Hive. MapReduce is a programming model for distributed parallel processing of the data. This processing takes place at each node in the networking cluster. The model presented is very effective while parallelizing batch task on a large amount of unstructured data.Basically MapReduce is used for the offline analysis of the data and is good for the general sequential data access but not for the random read/write access. Pig is a procedural data flow language which was used by programmers and researchers. The Apache pig, for instance, is built upon Hadoop and simplifies the program writing. Hadoop is very efficient for the batch processing and the ApacheHbase aims to provide real time access to the Big Data. Hive is a declarative SQLish language , Which is used by analysts for generating reports Hive provides a mechanism for structures and query data using SQL So called HiveQL and also allows programmers a map reduce to pligin their custom mappers and reducers [3].

Comparison of technologies of big data analysis

Feature			
Language	Algorithm of Map and ReduceFunctions (Can be implemented in C, Python, Java)	Pig scripting language	Like-SQL
Types/techniques	No	Yes, implicit	Yes explicit
Partitions	No	No	Yes
Sever	No	No	Optional (thrift)
Lines of, code	More lines of code	Fewer (around 10 lines of PIG = 200 lines of Java)	Fewer than MapReduce and Pig due to SQL Like nature
Complex business logic	More control for writing complex business logic	Less control for writing complex business logic	Less control for writing complex business logic
Structured vs Semi-Structured Vs Unstructured data	Can handle all these kind of data types	Works on all these kinds of data types	Deal mostly with structured and semi-structured data
Performance	Fully tuned MapReduce program would be faster than Pig/Hive	Slower than fully tuned MapReduce program, but faster than badly written MapReduce code	Slower than a fully tuned Map educe program, but faster than bad written MapReduce code

## METHODOLOGY

### BIG DATA AND NETWORK TRAFFIC MANAGEMENT

By applying the big data technologies on the network technologies, we can have an insight from the huge amount of the operational data which could not be exploited before. Telecommunication networks are becoming increasingly complex due to the increased use of the internet as a social infrastructure. Numerous kinds of data such as traffic management data, network configuration data and data from the network failures are used in the managing networks. Traffic management, data include packet count and byte count. These counts are measured per link and per-flow by the router in the management information base. There are different kinds of the network devices, including switches, routers, and servers. These network elements are developed by different manufacturers. The Big Data can be analyzed on alarms, trouble tickets, and network configurations

**TECHNOLOGY 1: MACHINE LEARNING APPLICATIONS FOR DATA CENTER OPTIMIZATION**

Both the consumers and enterprises have transferred the focus from consumer side computing to SaaS and cloud-based computing by the high speed trust on the Internet –enabled devices. The big companies like Google, Amazon and Facebook Web Services have reduced upfront capital and operating costs. This allows smaller scale companies to scale quickly and efficiently across million of users in a small amount of time. Distributed Control (DC) environments are suitable for Machine learning. The DC environments have a wide range of mechanical and electrical equipment. It is very difficult to predict any fault in these systems using conventional or traditional engineering techniques. Due to multiple software and hardware combinations it is very difficult to meet the target set points due to multiple hardware (mechanical and electrical) and software (control strategies and set points) combinations in a real time DC environment. The testing of each and feature combination will be impossible in the sensitive time constraints. To address the mentioned problems with this review, a neural network is selected as a mathematical framework for the DC environments [6]. The selected neural network can be elaborated by the following figure:

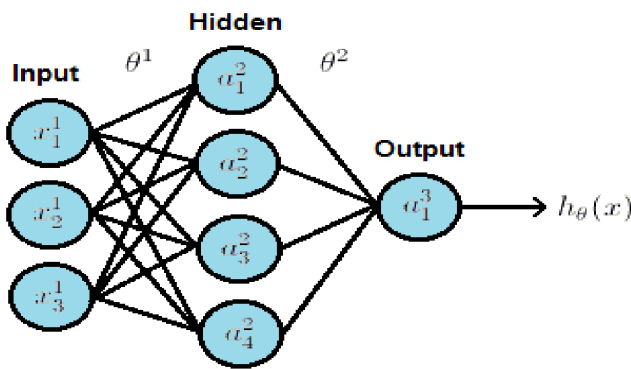


Figure: Artificial Neural Network (ANN).

The neural network selected in this study has an input matrix 'x' which an array (m x n). Where, 'm' is the number of the training examples. 'n' is the number of features. These features include IT load, weather conditions, the number of the chillers and the cooling towers. The matrix is multiplied with the model parameters known as  $\theta^1$  resulting in the hidden state matrix 'a'. In practice the hidden state matrix  $\theta^2$  interacts with the parameters to produce the output 'h<sub>θ</sub>x' [6].

**TECHNOLOGY 2: BENCHMARKING TECHNIQUE AND THE BIG DATA**

With the complexity and diversity in the data obtained by the current big data systems it can be concluded that the no single system is able to represent all applications. It is crucial to generalize the behaviors from all of the systems to a more comprehensive approach. The approach should be able to determine the conventional workload behaviors in representative application domains. There are two challenges in making this generalized approach[7]. The first challenge is the operations which are used to process the BigData in a specified application domain need to generalize and their functions should be identified. Secondly, for a given set of abstracted operations 'workload patterns' need to be identified and then specified to

them. One abstracted workload can contain one or multiple abstracted operations as well as their workflow [7]. Considering the mentioned factors the designed benchmarking layers can be elaborated by the following diagram:

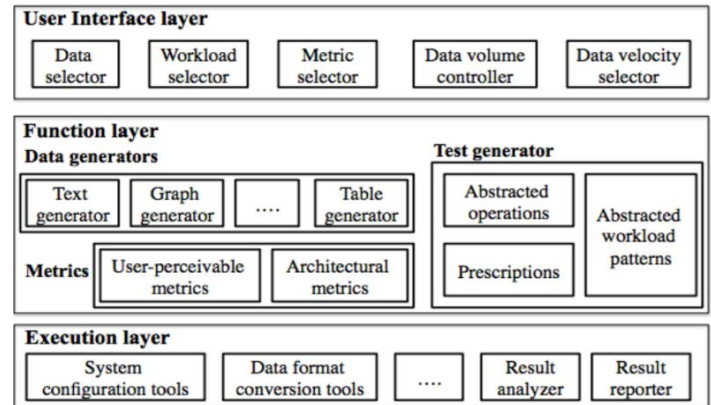


Figure: Benchmarking Layers

In the selected benchmarking system for the 'User Interface Layer' can assist the system owner to specify their benchmarking requirements. These include selected data, workloads, matrices, and the preferred data volume with velocity. The 'Function Layer' has three components known as data generators and matrices. Data generators are designed to produce data sets covering different data types and application data, keeping '4V' properties of this big data intact. The matrix generators can generate two types of matrices, one which is required or neutered by the user and the other to compare performances of the workloads from the different categories (Lu et al., 2015). The 'Execution Layer' offers numerous functions to support the execution of the benchmark over the different software stacks. The data converter transforms a generated data into a format capable of being used by this set (Lu et al., 2015).

**TECHNOLOGY 3: PIPELINE PROCESSING OF BIGDATA**

**1. DATA ACQUISITION**

The pipeline processing of the big data and the solutions using this technique are reviewed in this portion of the review. The pipeline processing of the Big Data problems can be represented by the following diagram:

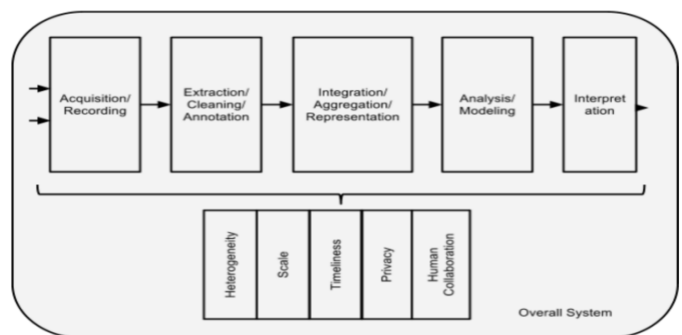


Figure: Pipeline Processing of the Big Data.

Much of the data generated by the data generating sources of the Big Data are useless. It needs to be filtered and compressed by the orders of the magnitude. One of the challenges of this filtering is not to discard the useful information presented by the sources, The data generated from these sources are also volatile or temporarily correlated. This challenge can be countered by selecting the right metadata to describe what data is recorded. Metadata capturing can reduce the human burden. The recording of the information at its birth is not useful unless this data

**2. INFORMATION EXTRACTION AND CLEANSING**

The information collected cannot be frequently available in the format ready for the analysis. The approach of thinking that the Big Data always tells the truth is not correct and an information extracting process is required that pulls out the information from the underlying resources and express it in a structured form available for analysis .

**3. DATA INTEGRATION AND PRESENTAIAION**

Data mining requires integrated, trustworthy and efficiently access data. The data mining can be used to improve the quality and reliability of the data. A problem in the current Big Data analysis is the lack of coordination available between the database systems. The hosting of the data and provision takes place in the form SQL querying. The packages that perform various forms of non-SQL processing techniques such as data mining and statistical analysis is used for the analysis of SQL based data. The privacy of the data is another huge concern. There are strict laws available to implement the privacy concerns in the Big Data analysis, but they need to be technically enforced. Declarative specification is needed not just for pipeline composition but also for the individual operations. New data, generated must be accounted for taking into the consideration the prior results or results from the existing data. Current systems for the Big Data analysis provide little or no support for such Big Data pipelines presenting a big challenge .

**PROTOCOLS USED FOR BIG DATA**

Protocols Organizations which rely on the Big Data processing have developed some protocols which are implemented within the organization. They are not public and, therefore, cannot be available for the review. In this review, we have selected the protocols which are commonly used.

**1. AMQP**

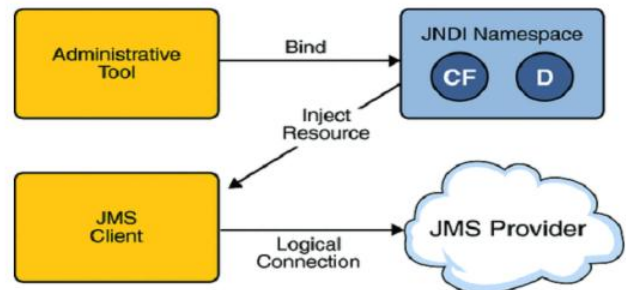
The reason for the development of the AMQP is the need for a protocol that would meet the needs of the larger companies with respect to the data acquisition. To achieve this goal 23 companies formed a group resulting in the formation of AMQP (Advanced Message Queering Protocol). It became OASIS standard in October, 2012 [10]. The AMQP system relies on the four different layers to implement its requirements. The type system is based on the primitive type of the database. AMQP allows the use of different encoding to store symbols which are the combination of several primary types. The transport layer provides the information about the transmission of the messages. The AMQP network is made up of nodes which are connected to each other via links. Messages are originated by the senders and received or consumed by the receiver. Messages are only allowed to travel if they follow the criteria set by the sender [10]. The messaging layer of the AMQP elaborates the validness of a message. According to

the protocol, a valid message should have the following properties. i.It should have a header, ii.Delivery annotation, iii. Message-annotation, iv. One property. v. The body, vi. One footer.

The transaction layer provides “coordinated outcome of otherwise independent transfers”. The final layer of the protocol is the security layer presenting the means to encrypt the content of AMQP messages.

**JMS JAVA MESSAGE SERVICE**

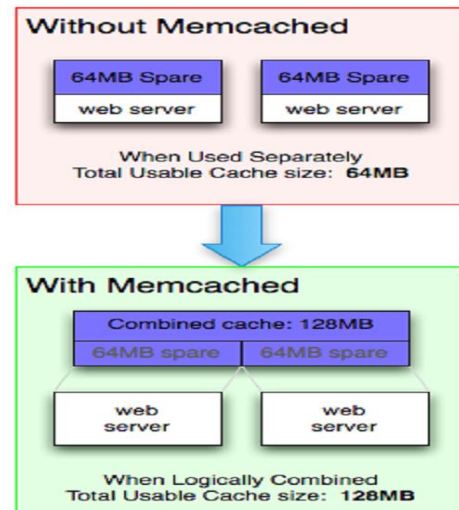
It provides, “a common way for JAVA programs to create, send, receive and read an enterprise messaging system’s message”. The administrative tools allow you to bind the destinations and connection factories into JAVA naming and Directory Interface (JNDI) [10].



The JNDI serves in this case as moderator between the hosts or clients which are supposed to exchange the messages. Currently JMS offers two types of the messaging model known as point-to-point and publisher-subscriber. The second type of the model is one to many connections.

**1. MEMCHACHEQ**

MemchacheQ is a queuing system based on the MemchacheDB, a storage system implementing the protocol. The memchache protocol is a memory caching system used by different websites to reduce the database load. It increases the speed of the websites that are operating on the top of databases. Due to its design it fairly easy to be deployed and it is available for many of the popular programming languages (Curry et al, 2014).



The memcached is used by many websites including LiveJournal, You tube, Wikipedia, Twitter And Wordpress.

## CONCLUSION

Everyone realizes that the Internet has changed how organizations work, government's capacity, and individuals live. At the same time, another, less obvious mechanical pattern is pretty much as transformative: "enormous information." Big information begins with the way that there is a considerable measure more data coasting around nowadays than any other time in recent memory, and it is being put to exceptional new employments. Enormous information is unmistakable from the Internet, in spite of the fact that the Web makes it much less demanding to gather and offer information. Huge information is about more than just correspondence: the thought is that we can gain from an expanded collection of data things that we couldn't understand when we utilized just smaller sums. Huge information helps answer what, not why, and frequently that is sufficient. The Internet has reshaped how mankind imparts. Enormous information is distinctive: it denotes a change in how society forms data. In time, huge information may change our mindset about the world. As we tap always information to comprehend occasions and decide, we are prone to find that numerous parts of life are probabilistic, as opposed to certain.

## REFERENCES

- [1] <http://www.javacodegeeks.com/2013/04/how-hadoop-works-hdfs-case-study.html>
- [2] <http://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies>
- [3] <https://hive.apache.org/>
- [4] <https://hadoopecosystemtable.github.io>
- [5] <https://www.nttreview.jp/archive/ntttechnical.php?contents=ntr201311fa1.html>
- [6] Gao, J. (2014). Machine Learning Optimization for Data Center Optimization
- [7] Hand, R., & Lu, X. (2015). On-Big Data Benchmarking. London: Imperial College London
- [8] Ghaninejad, A., Bowman, T., Tasu, A., & Ekbia, H. (2013). Big Data: Bigger Dilemmas A Critical Review. CROMI.
- [9] [http://www.slideshare.net/PhilippeJulio/hadoop-architecture/33-MANAGEMENT\\_OPS\\_WHIRR\\_AMBARI\\_Apache](http://www.slideshare.net/PhilippeJulio/hadoop-architecture/33-MANAGEMENT_OPS_WHIRR_AMBARI_Apache)
- [10] Khan, N., Yaqoob, I., Inayat, Z., Ali, M., Alam, re Articles, 11(11). M., & Gani, A. (2014). Big Data: Survey ,Technologies, Opportunities and Challenges. The Scientific World Journal, 2014(712836).
- [11] Kvernick, T., & Matti, M. (2012). Applying Big Data Technologies to network architecture. Ericsson Review.
- [12] Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and review of key research topics. AIP Publisihing.
- [13] Rusitchka, S., & Ramiez, A. (2014). Big Data Roadmap Societial Externalities. European Union.
- [14] Ryu, S. (2014). Book Review: Big Data Management, Technologies, and Applications. Healthc Inform Res, 20(1), 76. doi:10.4258/hir.2014.20.1.76
- [15] Schroeder, R., & Cowls, J. (2014). Big Data, Ethics, and the Social Implications of Knowledge Production. Oxford.
- [16] Sharma, P. (2013). Leveraging Big Data Using SAS High Performance Analytic Server. SAS Global Forum.
- [17] Shiomoto, K. (2013). Applications of Big Data Analytics Technologies for Traffic and Network Management Data - Gaining Useful Insights from Big Data of Traffic and Network Management