*International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868*
*Foundation of Computer Science FCS, New York, USA*
*Volume 5– No.7, May 2013 – www.ijais.org*

# Hadoop and Risk Analytics

Pankesh Bamotra
SCSE, VIT University
Vellore
Tamil Nadu

Saira Banu J.
SCSE, VIT University
Vellore
Tamil Nadu

## ABSTRACT

This paper brings out the specific use case of Hadoop in risk analytics which forms an important part of every organization. Risk analytics is necessary because from business perspective, business leaders in any organization run into one or other kind of risk. They need to add value proposition to any kind of risk they take on behalf of the organization. But how to assess that risk is a big challenge in today’s world. This is because of volume, veracity, velocity and variety of data that need to be analyzed is growing on an ever-increasing scale. This ultimately leads to vulnerability. Every day several petabytes of information is being stored, logged and analyzed, thus putting a bottleneck in the use of traditional RDBMS for real-time analytics. Here is when Hadoop comes as a savior. The paper talks about what Hadoop is, its programming paradigm called MapReduce, how is Hadoop different from traditional RDBMS, the technologies built on top of Hadoop, when to choose and not to choose Hadoop, its limitations and future scope. The use case of Hadoop with respect to risk analytics and that too particular to e-payments industry is also discussed.

## General Terms

Analytics, Distributed processing, Legacy systems, Payment industry

## Keywords

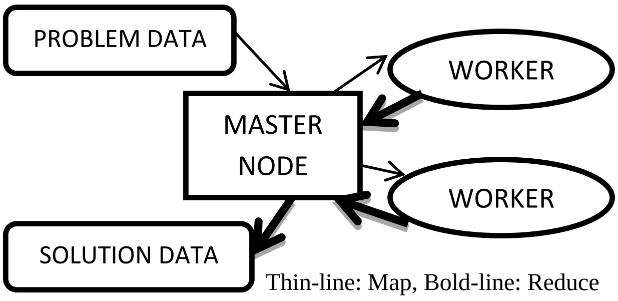
Big Data, Hadoop, Risk analysis, RDBMS

## 1. INTRODUCTION

With the explosion of information the legacy systems are facing a big time challenge. The scale of data has increased massively. Global data consumption has increased to tens of zettabytes. Undoubtedly information is aggregating from every sphere and everywhere whether its Facebook posts, tweets on twittertransactional logs of VISA customers or call records by phone companies. All this maybe just the tip of iceberg floating in the vast ocean of data. This is where Hadoop has come into action. But one thing to remember is that where Hadoop solves some of the so called Big data problems it is neither a fix to all problems nor a magic. Hadoop is still relatively new so some facts about it need to be clarified. The paper has been divided into various sections beginning with introduction to Hadoop, next section dives into its use case relative to risk analytics in e-payment companies like PayPal. Then we discuss about how Hadoop differs from the traditional databases. This is followed by the technologies built on top of Hadoop, their usefulness and limitations and finally concluding with the scope and future of Hadoop with respect to risk analytics. The objective of this paper stands at evaluating the use case of Hadoop with respect to risk analytics and that too specifically in e-payment industry and exploring the usefulness and limitations of Hadoop and its associated technologies.

## 2. HADOOP AND MAPREDUCE

Hadoop basically is an Apache Foundation’s open source project that was started by Doug Cutting in the year 2006. Since then Hadoop has completely transformed and there has been an emergence of numerous markets products calling themselves as Hadoop. Among the top vendors are Cloudera, Hortonworks and Amazon.All of them are the result of same basic and foundational platform. Hadoop is free and dramatically enhances processing speed in certain type of problems. Its primary components consists of a distributed filesystem known as HDFS[1] or Hadoop Distributed File System that provides very economical data storage making Hadoop a promising archiving store. Hadoop tends to create a balance between the economic impedance mismatch and growth in data that is supposed to create value for organization. Google introduced the MapReduce [2] paradigm that provides a distributed framework for easy processing of big data. It is one of theprimary components in Hadoop that has excellent job management framework to achieve parallel processing across various nodes or clusters. MapReduce enables utilization of the power of clusters to process bulk of information. It allows one to think in terms of what operations need to be done rather than focusing on how those operations are carried out. Thus MapReduce helps in providing a transparent view of processing the data while hiding the underlying distributed communication, networking, execution, co-ordination and fault tolerance. Each MapReduce program consists of two steps namely Map and Reduce. In the former step the job manager maps the problem into smaller chunks of problem using key-value[3] pairs and assign them to the processing nodes which in return an intermediate result in the form of key-value pair. Then in the latter step i.e. during reduce step all the intermediate results are combined to form the final result. All this computation is done on massively parallel scale and on terabytes of data.



Thin-line: Map, Bold-line: Reduce

**Fig 1: Diagrammatic view of MapReduce paradigm**

## 3. HADOOP IN RISK ANALYTICS

Risk management [4] is an important part of every business and is extremely crucial when it comes to e-payments industry. The classical example would be companies like PayPal, Google and Amazon who are in a cut-throat competition to reach the consumer's pocket through online world. Thus risk analytics plays an import role. A typical flow of payment analytics consist of the transaction itself, the rules that are triggered as a result of transaction behavior and the alerts generated consequently. Traditionally this would involves bringing the data and logs associated with the transaction to the ETL [5] platform and finally loading it to some data warehouse. This has led to emergence of several technologies like Teradata[6], Oracle [7] and Netezza [8]. However this approach is good only when dealing with not so big data at once say <1 billion rows and also changes to the data are known well in advance. But with dramatic growth in data this approach has weakened as the volume is varying and increasing daily making analytics on big sets tough and eventually infeasible. This acts a barrier when it comes to real-time analytics [9]. For explanation purpose let's take a typical example of credit card fraud detection system. Each transaction has numerous attributes and some specific sequences of transaction can be mapped to fraudulent activity. To track this various rules and modes are run that can differentiate between a normal and a fraudulent transaction. These rules are nothing but if-then conditions that generate an alert when something wrong happen. However to run these rules a lot of data needs to be analyzed like customer's present IP location, his past transactional history and many more factors. This would be a hectic and nearly impossible task when the past transactional details amounts to petabytes of information and moreover all the processing has to be fast enough to keep the customer experience good. If the fraud models consume a lot of time then a legit customer may cancel the transaction which would impact the business and is certainly not acceptable. In such a scenario Hadoop would be quite impactful. It can scale to analyze petabytes of information and due to its distributed nature it can significantly reduce the processing time as compared to the current system. Other use cases can be real time monitoring and analysis that can help prevent losses due to stolen financials and/or credentials.

## 4. HADOOP VS. RDBMS

In a data warehouse for analytical purposes two conditions are generally required – A fault tolerant and fail safe data store and faster data processing. As already explained this is achieved through HDFS in Hadoop. HDFS is a fault tolerant data store. It typically stores two to three copies of a data block on various clusters. Thus if one of the clusters is down the job can be done by fetching data on the other cluster. This adds to the other advantage of faster processing. Since the data is distributed across the cluster nodes it leads to what is called as data locality. Since some of the data may reside on the cluster where the operation is performed they is less of latency and the performance gets increased manifolds through parallel processing of data. Another significant difference between Hadoop and traditional RDBMS is that in Hadoop we don't need schemas as it can operate on the data without any transformation or normalization. One thing to be pointed here is that Hadoop not only supports structured data analytics but is indeed designed considering the semi-structured and unstructured data while RDBMS can only operate on data which fits into the Entity Relationship model [10]. One last but a noteworthy difference between Hadoop and RDBMS is that MapReduce use sort/merge to update the database which is much more efficient than the use of B-tress in the traditional databases. In nutshell these differences have been shown in Table 1.

## 5. HADOOP TECHNOLOGIES

Hadoop, as already discussed providesa framework for efficient distributed computing and as a result numerous Hadoop based technologies [11] have emerged to leverage some specific problems related to it. These are listed as below:-

1. Pig: It is a programming language based on scripting paradigm and built on top of Java to help writing the MapReduce jobs.
2. HBase: It is a distributed DBMS that is heavily influenced by Google's BigTable [12]. It provides a columnar schema for data storage on top of Hadoop.
3. Hive: It is a SQL like query language developed by Facebook that puts a querying interface in front of Hadoop.
4. Sqoop: It provides easy data movement between RDBMS and HDFS.
5. Zookeeper: It provides distributed coordination between Hadoop clusters.

**Table 1. Comparison between RDBMs & Hadoop**

| | RDBMS | HADOOP |
|---|---|---|
| DATA SIZE | Gigabytes | Petabytes |
| STRUCTURE | Static Schema | Dynamic Schema or Schema-less |
| ACCESS | Batch and interactive | Batch |
| INTEGRITY | High | Low |
| UPDATE | R/W many times | Write once, read many |
| SCALING | Non-linear | Linear |

These are just to name a few everyday a new technology for specific solution emerges. Recently Cloudera launched Impala that acts like a SQL-engine and eliminates the use of MapReduce jobs for data processing and instead provide the usual SQL like interface to intelligence tools to query HDFS and HBase. Yet another technology is Mahout that provides data mining algorithms especially designed for MapReduce framework and to work with Hadoop. HCatalog is a metadata handling tool for Hadoop that assists in easy retrieval of data from HDFS.

## 6. HADOOP AND ITS LIMITATIONS

As told in the beginning Hadoop is not a solution to every problem. There are many issues which will be evaluated one by one. Firstly Hadoop as mentioned is a framework and not a product solution. It works fine for simple queries but when it comes to complex analytics it becomes cumbersome as it results in writing Java code to support MapReduce framework that makes business analytics a complex job. Another limitation of Hadoop comes from its strength, the HDFS. HDFS was built considering efficiency. It replicated the already big data to support data locality. Theoretically 3 copies of data are available at minimum. This may pose a serious performance issue. Analytics generally need joining multiple datasets which is very slow in Hadoop and it doesn't support the concept of indices. Hadoop has very limited SQL support though products like Impala are there in the market this development is still in the age of infancy. Another point to be noted in terms of performance is that HDFS does not support query optimization i.e. there is no cost-based optimizer that can choose the cost efficient plan of executing the query. Because of this Hadoop needs bigger clusters than needed for similar database operations. Since HDFS is a filesystem there is no concept of consistency or check- points. This means one can never be sure that the result returned at the end of processing are or aren't completely true. From the development perspective Hadoop poses yet another problem of high development and maintenance cost. This is because to efficiently get the results using Hadoop one has to first translate the problem into MapReduce paradigm and optimize it to execute efficiently on the Hadoop cluster. Technologies like Hive depend on MapReduce which further reduces the querying performance.One last but often ignored limitation of Hadoop is that it is more biased towards open-source and Java community. Being open source it poses serious challenges in terms of quality and support.

## 7. HADOOP'S FUTURE

Despite the limitations listed earlier the future of Hadoop is certainly very bright. With inflow of big data steaming the traditional data management tools are becoming totally incapable of handling such huge amount of information and processing valuable information out of it. Hadoop in such a time provides an open source Java based framework for distributed and parallel processing and has emerged as a solution for analyzing massive data sets. Hadoop stands out of niche-oriented NoSQL solutions like Cassandra, CouchDB, MongoDB etc. as it provides a very generic and uniform solution to big data analytics by offering services like querying languages, database access and expanding ecosystem. Though there is a lack of tools for already SQL loving community but languages like HiveQL and DrQL are making Hadoop open to them. Companies like Greenplum have come up with product called Pivotal HD [13] that adds stability to Hadoop's gamut of being flexible, scalable, inexpensive, and fault-tolerant. It addresses the problem of analysts who are used to using BI tools and visualization softwares like Tableau. Pivotal HD increases the speed and accessibility to Hadoop approximately a hundred times than Hadoop's Hive and Cloudera's Impala. It tries to compensate the skill gap of using Hadoop in businesses by providing a SQL like interface on top of Hadoop. It offers a cost-based query optimizer which otherwise is a known limitation of the open-source Hadoop distribution.

The other limitation that critics of Hadoop have always pointed out is that there is no single and standardized distribution of it. The core functionalities of Hadoop are being developed by Apache foundation which doesn't address many of the industrial issues related to Hadoop like high availability, metadata processing etc. Thus there is an extra developmental cost in order to deal with these problems.However companies like Hortonworks and Cloudera are working on creating an industry wide consensus on Hadoop distributions.

## 8. CONCLUSION

In nutshell Hadoop certainly has a bright future in sense that as the incoming data piles up, the need for purposeful and valuable risk analytics out of big data increases. Hadoop has already proven to be low cost archival store but as it undergoes refinement it is shaping up as the future of databases and the analytics that depends on the big data. Clearly with the increase use of technology there has been an exponential increase in amount of data that is being archived for analysis purpose. E-payment companies like PayPal, VISA and MasterCard log petabytes of transactional level of data and risk analytics to identify the fraudsters and spoof has a great value for such organizations. Certainly Hadoop is not yet fully flourished and is in the stages of infancy. Companies like Cloudera, Hortonworks and others are working on it to bring about the best of Hadoop.

## 9. ACKNOWLEDGMENT

Through this paper I would like to pay my profoundest gratitude to Mr. Sanjay Singh, Manager, Risk Analytics, PayPal India Pvt. Ltd., Chennai for his immense patience in guiding me through the basics of payment industry.

## 10. REFERENCES

[1] Google's MapReduce Programming Model—Revisited, Ralf Lämmel, Science of Computer Programming, Volume 68 Issue 3, October, 2007

[2] Hadoop Operations, Eric Sammer, O'Reilly Publications

[3] Key-Value stores: a practical overview, Computer Science and Media, Ultra-Large-Sites SS09, Stuttgart, Germany

[4] Understanding Risk Management in Emerging Retail Payments, Michele Braun, James McAndrews, William Roberds, and Richard Sullivan, Economic Policy Review, Vol. 14, No. 2, September 2008

[5] A Survey of Extract–Transform–Load Technology. Panos Vassiliadis, International Journal of Data Warehousing and Mining (IJDWM), volume 5, no.3, pp. 1-27, June 2009, IGI

[6] Born To Be Parallel - Why Parallel Origins Give Teradata Database an Enduring Performance Edge Carrie Ballinger, Teradata White Papers

[7] In-Memory Big Data Analysis with Oracle Exalytics, Mark Rittman, Oracle Openworld 2012, San Francisco, October 2012

[8] IBM Netezza Analytics, IBM Data Sheet, April 2012

[9] Plenary talk: Big data and real time analytics, Rao, G. V, N Appa, Recent Trends in Information Technology (ICRTIT), 2011 International IEEE Conference

[10] Fundamentals Of Database Systems, 5/E, R Elmasri

[11] MapReduce and the Data Scientist, Colin White, BI Research, January 2012

[12] Bigtable: A Distributed Storage System for Structured Data, Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber, Proceedings of OSDI 2006

[13] O'Reilly Media Inc.'s Strata Conference 2013, Santa Clara