IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Data Mining: Exploring Big Data Analytics, Hadoop and Mapreduce

Ms. Rupali Chikhale
G. H. Raisoni Institute of Information Technology, Nagpur, India
rupali.chikhale@raisoni.net

Abstract

Most internal auditors, especially those working in customer-focused industries, are aware of data mining and what it can do for an organization — reduce the cost of acquiring new customers and improve the sales rate of new products and services. However, whether you are a beginner internal auditor or a seasoned veteran looking for a refresher, gaining a clear understanding of what data mining does and the different data mining tools and techniques available for use can improve audit activities and business operations across the board.

The tremendous opportunities to gain new and exciting value from big data are compelling for most organizations, but the challenge of managing and transforming it into insights requires a new approach to analytics that has a far reaching impact on IT infrastructure. Traditional systems are unable to cope cost-effectively—if at all— with new dynamic data sources and multiple contexts for big data. Emerging technologies such as the Hadoop* framework represent completely new approaches to capturing, managing, and analyzing big data. Big data challenges plus new technologies are causing a paradigm shift that is driving organizations to reexamine their IT infrastructure and analytics capabilities.

With the fast growth of networks now-a-days organizations has filled with the collection of millions of data with large number of combinations. This big data challenges over business problems. It requires more analysis for the high-performance process. The new methods of hadoop and MapReduce methods are discussed from the data mining perspective.

Keywords: Data mining, Big data, BI, Big Data analytics, OLAP, EDA, Neural Networks, Hadoop and MapReduce technique, Advantages, Disadvantages.

Introduction

Big data analytics

Big data analytics is the process of examining large amounts of different data types, or big data, in an effort to uncover hidden patterns, unknown correlations and other useful information.

Big data analytics is the process of examining large amounts of data of a variety of types (big data) to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in business benefits, such as more effective marketing and increased revenue.

The primary goal of big data analytics is to help companies make better business decisions by enabling data scientists and other users to analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence (BI) programs. These other data sources may include Web server logs and Internet clickstream data, social media activity reports, mobile-phone call detail records and information captured by sensors. Some people exclusively associate big data and big data analytics with unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid forms of big data.

Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics and data mining. But the unstructured data sources used for big data analytics may not fit in traditional data warehouses. Furthermore, traditional data warehouses may not be able to handle the processing demands posed by big data. As a result, a new class of big data technology has emerged and is being used in many big data analytics environments. The technologies associated with big data analytics include NoSQL databases, Hadoop and MapReduce. These technologies form the core of an open source software framework that supports the processing of large data sets across clustered systems.

Potential pitfalls that can trip up organizations on big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced analytics professionals, plus challenges in integrating Hadoop systems and data warehouses, although vendors are starting to offer software connectors between those technologies.

Real-world experiences with big data analytics tools

Selecting technology is just one step in the overall big data project planning and implementation picture. Experienced users say it's crucial to evaluate the potential business value that big data tools offer and to keep long-term objectives in mind as you move forward. The articles in this section highlight practical advice on using big data analytics tools, with insights from professionals in retail, healthcare, financial services and other industries.

Opportunities and evolution in big data analytics processes

As big data analytics tools and processes mature, organizations face additional challenges but can benefit from their own experiences, helpful discoveries by other users and analysts, and technology improvements. Big data environments are becoming a friendlier place for analytics because of upgraded platforms and a better understanding of big data analytics best practices. In this section, dig deeper into the evolving world of big data analytics.

Hadoop and MapReduce

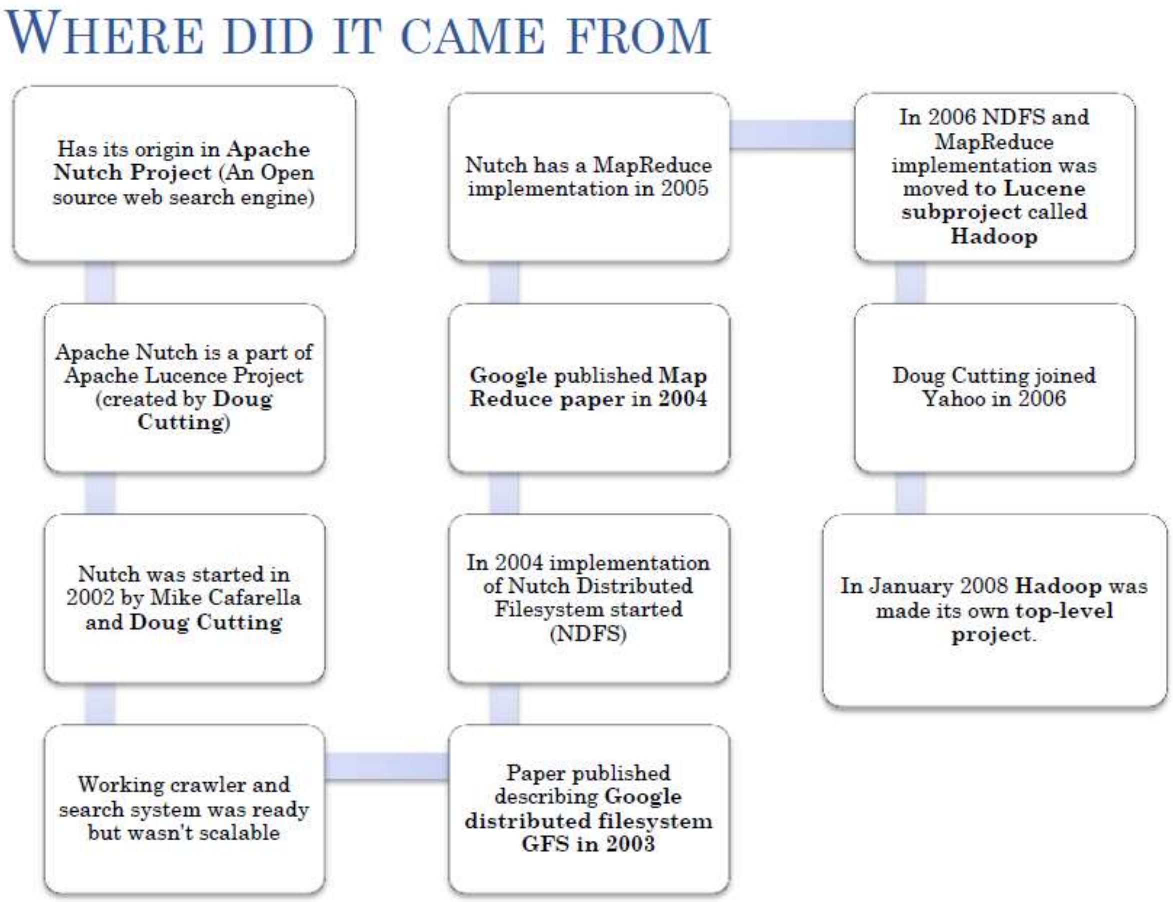
Big data [2] plays an important role in Business applications. The coverage of big data includes Data acquisition, cleaning, distribution, and best practices. Big Data contains the risk of threat analysis, predicting failures of the network data and trade control. Big data analytics found that Apache Hadoop is preferred as solution to the problems in the traditional Data Mining. It acts as extensible for recovering the failures of the data storage and processing in the distributed system.Apache Hadoop is an Open-source software framework [13] for storing and processing of large data-sets on clusters of hardware. Here the hadoop is designed with the assumptions of hardware failures that are automatically handled by the software framework. The main components of Hadoop are Hadoop distributed file system (HDFS) which is useful for large files and MapReduce which acts as heart of Hadoop. HDFS is high bandwidth clustered storage. MapReduce performs two different tasks in Hadoop programs. the First job is to map, in which it takes a collection of data where it is transformed into another set of data .After transformation the data is broken in to tuples (with key/value pairs).The job of reduce is to take the output from the map job which acts as its input and these data tuples are combined into smaller sets of tuples. In this manner the reduce job is always achieved after the map job.

HADOOP

The Apache™ Hadoop® project develops opensource software for reliable, scalable, distributed computing.

- Distributing data across machines - HDFS
- Framework to compute on top of this Distributed
- Data – Map Reduce
- Co-ordination Services – Job Tracker, Task Tracker
- New HPC infrastructures allow us to attack new problems, BUT require to solve more challenging problems.
- New programming models and environments are required
- Data is becoming a BIG player, programming data analysis applications and services is a must.
- New ways to efficiently compose different models and paradigms are needed.
- Relationships between different programming levels must be addressed.
- In a long-term vision, pervasive collections of data analysis services and applications must be accessed and used as public utilities.

We must be ready for managing with this scenario.



WHERE DID IT CAME FROM

Has its origin in Apache Nutch Project (An Open source web search engine)

Apache Nutch is a part of Apache Lucence Project (created by Doug Cutting)

Nutch was started in 2002 by Mike Cafarella and Doug Cutting

Working crawler and search system was ready but wasn't scalable

Paper published describing Google distributed filesystem GFS in 2003

In 2004 implementation of Nutch Distributed Filesystem started (NDFS)

Google published Map Reduce paper in 2004

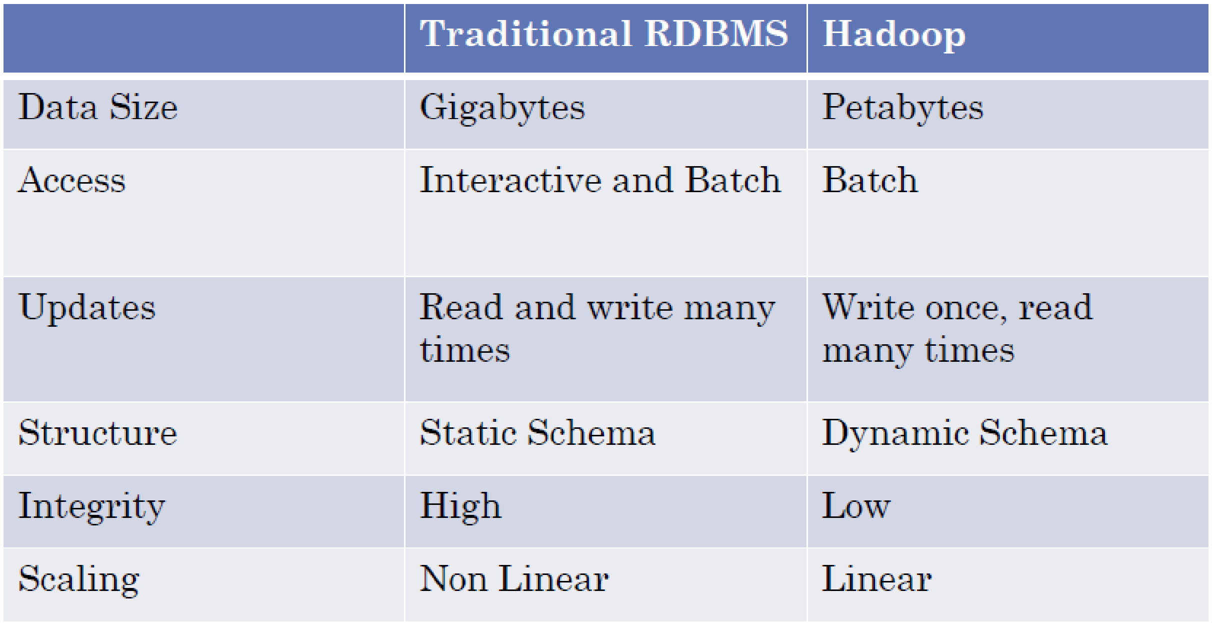
Nutch has a MapReduce implementation in 2005

In 2006 NDFS and MapReduce implementation was moved to Lucene subproject called Hadoop

Doug Cutting joined Yahoo in 2006

In January 2008 Hadoop was made its own top-level project.

RDBMS V/S HADOOP



	Traditional RDBMS	Hadoop
Data Size	Gigabytes	Petabytes
Access	Interactive and Batch	Batch
Updates	Read and write many times	Write once, read many times
Structure	Static Schema	Dynamic Schema
Integrity	High	Low
Scaling	Non Linear	Linear

Hadoop mapreduce advantages

The main advantage of Hadoop MapReduce it allows the users (even though if they are not experts) to easily handle analytical risk over Big data. It gives complete control on processing the input datasets. MapReduce can be easily used by the developers without having much knowledge of databases but with a little knowledge of java is needed. It gives satisfied performance in scaling large clusters.

- It supports distributed data and computation
- The computation is performed local to data and thus it prevents the network overload.
- The tasks are independent hence, it can easily handle partial failures such as when the nodes fail, and it can automatically restart.
- It is a Simple programming model. The end-user programmer only writes MapReduce tasks.
- HDFS stores vast amount of information.
- HDFS is simple and robust coherence model thus it stores data reliably.
- HDFS provide streaming read performance.
- Flat scalability [20]
- It has the ability to process the large amount of data in parallel.
- HDFS has capability for replicating the files which can easily handle situations like software and hardware failure.
- In HDFS the data can be written only once and it can be read for many times.
- It is more economic way as the data and processing are distributed across the clusters of personal computers.
- Vol 04, Special Issue 01, 2013 International Journal of Engineering Sciences Research-IJESR http://ijesr.in/ ACICE-2013 ISSN: 2230-8504; e-ISSN-2230-8512
- 2010-2013 - IJESR Indexing in Process - EMBASE, EmCARE, Electronics & Communication Abstracts, SCIRUS, SPARC, GOOGLE Database, EBSCO, NewJour, Worldcat, DOAJ, and other major databases etc.,
- It can be offered as on-demand service, for example as part of Amazon's EC2 cluster computing service.
- Ability to write MapReduce programs in Java, a language which even many noncomputer scientists can learn with sufficient capability to meet powerful data-processing needs.
-

Limitations of hadoop

These are major common areas where the Hadoop framework is found uncertain.

- As the both the Hadoop HDFS and MapReduce software are under active development, they are found to be uneven.
- Possibility of preventing central data leads to restrictive programming model.
- HDFS is weak in handling small files, and inadequacy of transparent compression. The design of HDFS is such that it doesn't work with random reads on small files because of its optimization for sustained throughput.
- There is a necessary of managing job flow is when there is intermediate data.
- Managing the cluster is hard in operations like debugging, distributing software, collection logs etc.
- Because of single-master model, it requires more care and may limit scaling.
- Hadoop offers high security model, but because of its complexity it is hard to implement it.
- MapReduce is a batch-based architecture which means it doesn't allow itself to use cases that needs real-time data access.

Summary

Those changes will have some positive effects on the Hadoop framework's ability to support real-time analytics and ad hoc querying. First, segregating resource management from application management and processing reduces the internal overhead of what had been the JobTracker's combined role and enables the ResourceManager to be more efficient and effective in allocating a cluster's inventory of CPU, disk, and memory resources to applications.

Conclusion

Data mining is more than running some complex queries on the data you stored in your database. To keep track of current state of business, advanced analytical technique of big data such as predictive analysis, data mining, statistics and natural language processing are to be examined. New techniques of big data such as Hadoop and MapReduce create alternatives to traditional data warehousing. Traditional Hadoop with combination of new technologies explores a new scope of study in various fields of science and technologies

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

References

1. Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees. Wadsworth, Pacific Grove, CA..
2. V. Gorodetsky, O. Karsaev, and V. Samoilov. Infrastructural Issues for Agent-Based Distributed Learning. In Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, pages 3–6. IEEE Computer SocietyWashington, DC, USA, 2006.
3. Fahrmeir L. and G. Tutz (2001) Multivariate statistical modelling based on generalized linear models 2nd edition. Springer Verlag, New York, 2001.
4. Kolyshkina, I, Petocz P. and Rylander, I. “Modelling Insurance Risk: A Comparison of Data Mining and Logistic Regression Approaches” submitted to Australian and New Zealand Journal of Statistics in October 2002
5. McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models (2nd edition). Chapman and Hall, London.
6. Salford Systems (2000). CART® for Windows User’s Guide. Salford Systems
7. SAS Institute (2002a). SAS Version 8. [On-line] http://www.sas.com/, (accessed 25/09/2002).
8. Steinberg, D. and Cardell, N. S. (1998b). The hybrid CART-Logit model in classification and data mining. Eighth Annual Advanced Research Techniques Forum, American Marketing Association, Keystone, CO.
9. Westphal, C., Blaxton, T. (1998). Data mining solutions. New York: Wiley.
10. Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning : Data mining, inference, and prediction. New York: Springer.
11. J. Dittrich, J.-A. Quian´e-Ruiz, A. Jindal, Y. Kargin, V. Setty, and J. Schad. Hadoop++: Making a Yellow Elephant Run Like a Cheetah (Without It Even Noticing). PVLDB, 3(1):519–529, 2010.
12. Edelstein, H., A. (1999). Introduction to data mining and knowledge discovery (3rd ed). Potomac, MD: Two Crows Corp.