Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Business Computing on Big Data Map Reduces Action and Services

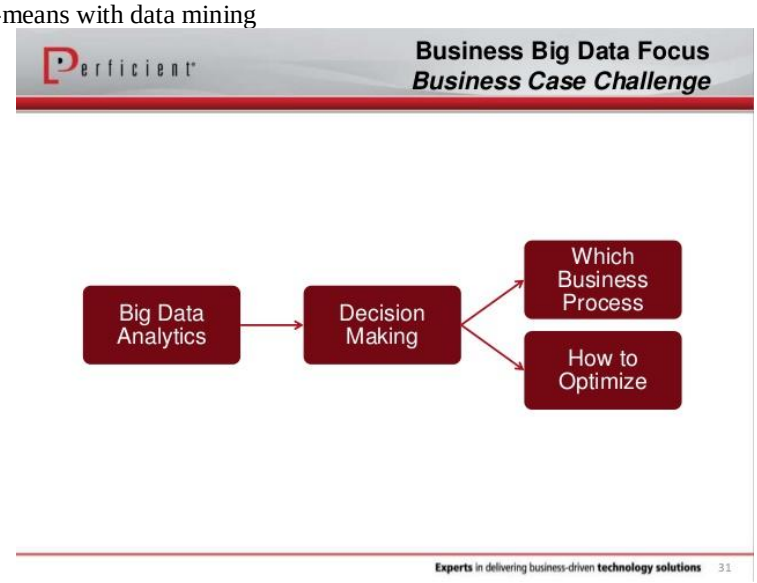
[1]N. Roja Ramani, [2]J. Sheela Jasmine
[1] Asst. Professor, Dept. of Computer Application, [2] Asst. Professor, Dept. of Computer Science
[1, 2] Annai Vailankanni Arts & Science College Thanjavur, Tamilnadu, India

Abstract: ***In this research paper, we discuss about big data process security issues for cloud computing, Big data, Map Reduce environment. The main focus is on security issues in cloud computing that are associated with big data. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries. We also discuss various possible solutions for the issues in cloud computing security and Cloud computing security is developing at a rapid pace which includes computer security, network security, information security, and data privacy. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools. Moreover, cloud computing, big data and its applications, advantages are likely to represent the most promising new frontiers in science.***

Keyword: Big Data Business tool, Varity.

I. INTRODUCTION

In order to analyze complex data and to identify patterns it is very important to securely store, manage and share large amounts of complex data. Cloud comes with an explicit security challenge, i.e. the data owner might not have any control of where the data is placed. The reason behind this control issue is that if one wants to get the benefits of cloud computing, he/she must also utilize the allocation of resources and also the scheduling given by the controls. Hence it is required to protect the data in the midst of untrustworthy processes. Since cloud involves extensive complexity, we believe that rather than providing a holistic solution to securing the cloud, it would be ideal to make noteworthy enhancements in securing the cloud that will ultimately provide us with a secure cloud. Google has introduced Map Reduce framework for processing large amounts of data on commodity hardware. Apache' distributed file system (HDFS) is evolving as a superior software component for cloud computing combined along with integrated parts such as which is an open-source implementation of Google Map Reduce, including a distributed file system, provides to the application programmer the abstraction of the map and the reduce. With Hadoop it is easier for organizations to get a grip on the large volumes of data being generated each day, but at the same time can also create problems related to security, data access, monitoring, high availability and business continuity. In this paper, we come up with some approaches in providing security. We ought a system that can scale to handle a large number of sites and also be able to process large and massive amounts of data. However, state of the art systems utilizing HDFS and Map Reduce are not quite enough/sufficient because of the fact that they do not provide required security measures to protect sensitive data. Moreover, Hadoop framework is used to solve problems and manage data conveniently by using different techniques such as combining the k-means with data mining



II. BIG DATA CHALLENGES. AMONG THESE CHALLENGES ARE THE FOLLOWING

Recognition

Identifying what's what in the data. See

Discovery--efficient ways to find the specific data that can help you

Modeling and simulation--intelligent ways to model the problems big data can solve so human inputs can result in useful outputs. See

Semantics: effective and efficient ways to contextualize the data so that it's relevant to specific individuals and groups. See

Analytics: effective ways to analyze and visualize the results of the data. See Reshaping the workforce with the new analytics

Storage, streaming and processing: efficient ways to take human inputs and act on batches or streams of big data to be able to extract insights from it. Remapping the database landscape

These disciplines are only just scratching the surface of the problem. There are sub-challenges beneath challenges. And each challenge requires its own special level of understanding. We're inefficient at targeting resources to solve specific big data challenges, because of the expanding totality of the larger problem. Each investor or willing and talented individual working the problem typically only sees a few pieces of the problem.

And then that's not to mention the issue of understanding what humans want and need to begin with, or what the natural world needs to sustain life at scale.... After all, those are the more fundamental problems we're all trying to deal with.

Storing, querying, and maintaining big data is extremely costly. There are three requirements of data for it to be worthwhile: it should be voluminous, be of a high velocity, and be of a lot of variety.

Digital product companies such as must store and recall each record of voluminous amounts of data to deliver their products. For them, Big Data is a worthwhile - necessary - enterprise. Additionally, their data meets the three criteria - there are a ton of users or file types that must be queried, and everyone is posting stuff all the time. A company that doesn't utilize Big Data to create core products may find it hard to justify the cost. What additional business value is created from storing and querying every record in a data set versus a a mere sample of those records in a data set? Additionally, the data used by these companies doesn't always meet the criteria of Big Data - there's a lot of data, and it may be coming in fast, but in the case of Utilities is there really a lot of variety To obtain valuable business insights, one needs only to analyze a sample of records rather than every record. This is especially true given a lack of variety. Non-digital products and services companies sometimes forget the laws of statistics in favor for sexy marketing.

III. BIG DATA ANALYTICAL CREATING PROBLEM

1. Storage on cloud:

When you have huge quantities of Data, the first and foremost problem would be the storage. Where would you store it? Do you buy new hardware and establish data centers or store it all in a cloud and make it somebody else's issue? What about the latency if you store it in the cloud? How fast do you want it and how often do you want it? And, what happens to the old data? Will you discard it or keep it?

All, these issues needs to be addressed before doing anything significant.

2. Security:

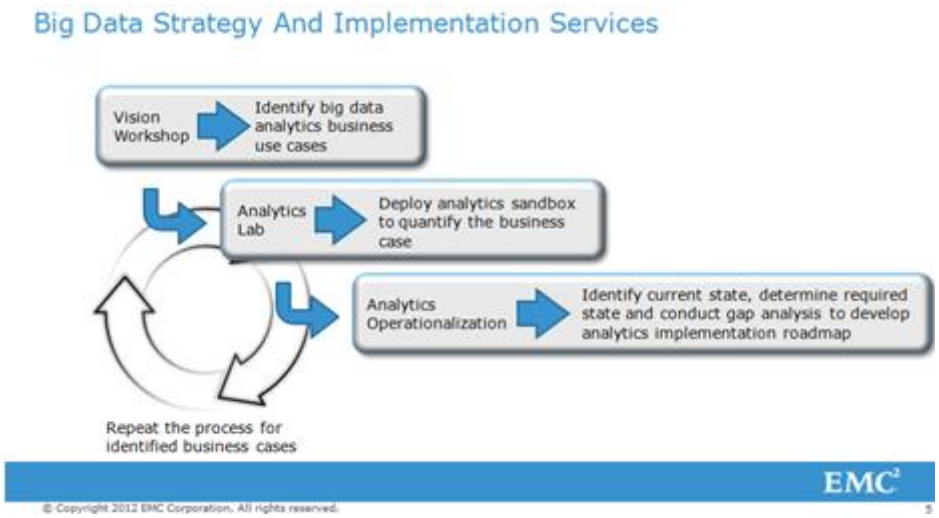
Some of the biggest data thefts have happened in the recent years and so security is one of the biggest issues to address. When you are storing the data you have to make sure you are following all the Data protection laws everywhere. Cloud storage may be an economical scenario for storage but not so much for the security. Although all the cloud storage companies employ top of the line security measures and protection to secure your data, one can never be too sure of that. And, so measures should be taken to prevent the worst case scenarios.

3. Need for Speed level:

In today's world people expect everything to be done instantaneously. Visualizations and predictions are really important to provide proper insights from the data but the challenge here is to go through the huge volumes of data and churn out beautiful graphs on top of that. For this we can always keep upgrading our systems and/or keep all data in the cache to make the access fast but still Real-time analytics is still something very impossible.

4. False Positives action:

Analytics essentially is looking at a subset of data to find a pattern and test those hypotheses on another data-set to see if you can find the same/similar pattern. So, essentially its and trial and error approach until you find the pattern that seems to be everywhere. With this approach on such huge amounts of data, it is possible to get False positives. So, the good practice is to test hypotheses rigorously and exhaustively to avoid making the wrong conclusions. Apart from these we also have to address the issues of Data quality, Data complexity.



IV. CONCLUSIONS

Data are useful only as much as the interpretations and conclusions we are able to draw from them. Putting aside concerns about the cleanliness and manageability of the data, the central problem with big data is that with enough of it, a data scientist is more likely to find support for conclusions that should not really exist. The more variables in a dataset, the greater the likelihood that one will find some random, meaningless correlation between two of them. Make the dataset bigger and this problem becomes magnified.

Despite all the excitement in machine learning and deep learning, nothing matches the ability of the human brain (yet) to draw a testable hypothesis about a given dataset using common sense about what variables might actually have a correlation. Since correlation alone does not imply causation, even if a data scientist finds something interesting in the data it must still be verified by other means.

Just because you have big data doesn't in itself imply anything useful. You might have a truckload of sand - well, great. It's better if you have an automated method to find seashell pieces in that sand, but better yet if your code allows you to do 3D image processing of whole seashells so you can tell exactly what beach in the world that sand came from. And if you are able to estimate the erosion rate of that beach from the fragmentation of the seashells so you can inform the city whether the expense of a beach wall is advisable or not, that is how you would realize the power of big data.

It really does seem that companies flocking to big data believing that it will yield something great without having the expertise to dissect, test, and derive inferences from that data are just getting a truckload of sand to haul around.

REFERENCES

[1] Abiteboul, S.; Clue, S.; Milo, T.; Mogilevsky, P.; Simeon, J.: *Tools for Data Translation and Integration*. In [26]:3-8, 1999.

[2] Batini, C.; Lenzerini, M.; Navathe, S.B.: *A Comparative Analysis of Methodologies for Database Schema Integration*. In Computing Surveys 18(4):323-364, 1986.

[3] Lakshmanan, L.; Sadri, F.; Subramanian, I.N.: *SchemaSQL – A Language for Interoperability in Relational Multi- Database Systems*.

[4] Chaudhuri, S., Dayal, U.: *An Overview of Data Warehousing and OLAP Technology*.

[5] Lee, M.L.; Lu, H.; Ling, T.W.; Ko, Y.T.: *Cleansing Data for Mining and Warehousing*. Proc. 10th Intl. Conf.

