Efficient Management of Big Data in Cloud Computing Environment

Abhishek Kumar

Abstract-Cloud computing offers efficacy oriented IT services to users throughout the whole world. It enables the hosting of applications from user, scientific, commercial as well as business domains. The core inspiration in Cloud Computing is that the entire system can be controlled as well as worked using simply an HTTP client. Because of the huge reduce in overall investment and greatest flexibility provided by the cloud, all the companies are migrating their applications towards cloud environment. Cloud provides the large volume of space for the storage and different set of services for all kind of applications to the cloud users without any delay and do not required any major changes at the client level.

Big data lies in the cloud. It's mainly about transforming business rather than IT by the cloud. In an enterprise cloud is the place to secure data. Therefore, to meet the changing needs, organization has to change time and budget in order to scale up infrastructure such as, software, hardware services.

This IT paper includes the major security challenges and privacy issues that are being faced while managing Big data in a cloud. It also includes the solutions to those problems.

I. INTRODUCTION

Today, Big data and cloud computing are top enterprises for IT, and when used together they provide assistances for both business and IT. IBM ranks top in the field of Big data Revenue and the company has beaten the record by it's services, hardware, software [1] which is followed by HP, Dell, SAP, Teradata, and Oracle and so on. Big data is the biggest buzzword that we hear around which is going to change the world. In terms of cloud, top companies are Amazon, IBM etc. Big data is the capability to store gigantic quantity of data generated daily in the world which could be structured or unstructured. It is an asset to the organizations and individuals [3].

Big data and Cloud Computing are top initiatives intended for IT, and when used together they provide benefits for both business and IT.

Gartner says- *"Big data" is high-volume, -velocity, -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making* [4]

The fourth one is veracity i.e. uncertainty of the data. For instance, in Twitter a lot of hash tags and colloquial languages are used (adds unwanted data) [5].

During, the last 2 years 90% of data has been created and it's mainly because of an unstructured data i.e. pictures, videos, emails, media etc. According, to the research it's going to be even bigger than the internet. Earlier, data was stored in an infrastructure using hard disks, storage media. Therefore, to handle the growing volumes of data, technologies- Big data and cloud are being implemented.

In this IT paper, the highlights of the security and privacy issues faced while Managing Big data in the Cloud has been mentioned [6], [7], [8], [9].

Big data can be managed by encryption, informing about the principles and policies that are used for governing the network.

II. BIG DATA

Business information are mainly extracted from Internet and Web data sources, for example, e-mail, Web-pages and logs, images, graphics, social networking sites (Facebook, twitter etc.). More recently, the use of Big data has grown to include other sources such as smart phones, smart utility meters, motor vehicles, aircraft engines, security equipment, telephone switches, RFID readers etc.

Apparently, there was an exploitation of data by companies which led to difficulty in handling large volumes of data by traditional approaches to data warehousing and data analytic processing. Hence, conventional methods like Hadoop distributed computing environment were used to maintain data.

Big data's value is not just the latest trend, it's the future of how we are going to guide and grow business. The vendors are focusing not only on providing products for Big data management, but also on solutions that can extract and analyze the business information embedded in the data. Hence, to speed deployment and improve time to value, these solutions are frequently offered as prepackaged hardware and software appliances and/or as a set of services for rapid deployment in a cloud-computing environment [10], [11].

III. CLOUD COMPUTING

Cloud computing services promise pay-as-you go, on demand and elastic scalability for developing and deploying many IT projects. Compared with an on-premises IT environment, cloud computing reduces upfront IT costs and enables organizations to scale their IT resources as required, while paying only for the resources they use. The cloud is therefore an ideal environment for big data projects, given the large data volumes and unpredictable nature of the analytic workloads involved. This is one of the reasons why the industry is seeing a sudden and significant jump in the use of cloud computing.

Another reason for this sudden growth is that cloud technologies are maturing and organizations are overcoming their data security issues and concerns. Barriers still remain to successful cloud adoption, however. Chief among these is complexity of integrating cloud and on-premises data and the inability of many cloud services to efficiently and rapidly move data into and out of the cloud environment – this topic is discussed in more detail below [12].

IV. DATA MANAGEMENT IN CLOUD

Most traditional data warehousing and business analytics projects to date have involved managing and analyzing data extracted from on-premises business transaction systems. In some situations, cloud services have been used for developing analytics on business transaction data stored in a cloud computing system such as salesforce.com, but these have been piecemeal and standalone projects. In fact, one of the risks of cloud computing is that it has made it easier for business groups and business users to bypass IT and purchase their own cloud-based IT services. This is why it is important for IT to partner and collaborate with the business in deploying and using cloud services to reduce risk, avoid poor technology selection, and manage data governance and data security issues [13]. For the foreseeable future, it is unlikely that many organizations will move their existing business transaction systems or sensitive transaction data for analysis purposes to a public cloud environment. However, cloud adoption for business transaction processing is increasing, especially for new projects and projects involving packaged application solutions, and so in the longer term this will lead to more traditional business transaction processing and associated analytic processing being done in the cloud. The biggest potential for cloud computing is the processing of data that already exists in the cloud. It is important to understand that big data in the cloud isn't a one-size-fits-all solution. It pays to make use of cloud services where it makes sense from the perspective of satisfying business needs, reducing costs, achieving faster time to value, and providing flexibility and scalability.

The most dismaying challenges being faced in data management- How to manage the data flood? How much of data is being generated daily? How much money to spend on collecting data? Where is the data getting dumped?

Initially, the currency of Big data is Petabyte- a quadrillion, 1015 or tens of hundreds of trillions of bytes. Digital data can be easily created and erased at the same time [14], [15].

During 1700-1950s computers were "humans" for instance- Harvard Computers or Pickering harems, were a group of women working for processing astronomical data (1880s until 1940s) at the Harvard College observatory [16].

Later, with the invention of sytems and technologies data was stored in hard disks, online databases and now in cloud. Elements of cloud computing services- IaaS, PaaS, SaaS. For example- Amazon, Google App-Engine, salesforce.com. Other examples are peer-to-peer networks on Skype, Bit Torrent.

With the rise in cloud computing, enormous data has been created which can be monetized by applications such as promotion. Cloud providers can control their network assets to enable their customers to confidently start moving more and more of their network critical workloads to the cloud. What suppose cloud providers could also directly monetize their network asset? What if networks and network services could be offered by the provider as a service i.e. network-as-a-service? Google, for example, influences (leverages) it's cloud infrastructure to gather and analyze consumer data for it's publicity network. Thus, compilation and analysis of data has been easy for companies even for those who don't have Google resources. On conflicting with the ease of data transfer and access to data, it has become easy for attackers to hack data and mine the databases. The reason for such an issue is cloud [17], [18], [19].

Importance of Big Data Management

(a) Basically for running business on data

- As the data is generated, operations run on data
- Those edintegrat data- prepared for analytics-supports visibility, monitoring, reporting
- Fresh data enable rapid responses by the business

(b) To build the business on data

  - Build partner relationships
  - Foe new customers market is formed
  - New products in the market are introduced
  - Service existing customers

V. APPROACHES/METHODS

The most straightforward method to store applications and data may be to store it in cloud. The programming model is quite similar to accessible non-cloud development models making it easy for implementation [20]. However, Data in the Cloud needs to be protected from unauthorized access:

1. Data Encryption- For protection of data in the cloud, data needs to be encrypted before storing it in cloud. Equally, data being returned from the cloud will be decrypted. For instance, Cloud safety box project is to create an Interface to a cloud storage provider that enables encryption/decryption of data available in the cloud.

2. BYOE (Buy your own Encryption) - A security model allowing cloud service customers to utilize their own encryption software plus manage their own encryption keys [21].

3. Information-Centric- This advancement of storing data in the cloud is a self-protection scheme. The data is encrypted and packaged with usage policy. Therefore, during the time of accessing data, it should be revealed itself only to trustworthy callers based upon the policy.

4, High Assurance Remote Server Attestation- This system provides a mechanism for the data owner to check how the data is being used. Hence, it will ensure whether the data is not being abused or leaked. It doesn't protect data External from the cloud but provides mechanism to ensure that security has not been breached.

5, Privacy – Will encrypt all data stored in the cloud and is similar to the encryption of Data. It's similar to encrypting data mentioned above, however, special features have been added that allow the data to be searched. Hence, this search ability allows a search query to be encoded, where the cloud can then decide if the stored data matches the encoded search query.

6, Data External from the Cloud- By storing the data outside the cloud, we can retain control of data. The major drawback of this approach lies in accessing data from the cloud-based application. For instance firewall, web services lookup running in the on-site data center, can be used to make data available to the applications. This advance will overcome troubles with the Firewall Expectations method [22]. [23], [24].

"Unless the creation, secure, storage, handling and deletion of encryption keys is carefully monitored, unauthorized parties can gain access to them and render them worthless" - Antony Ad

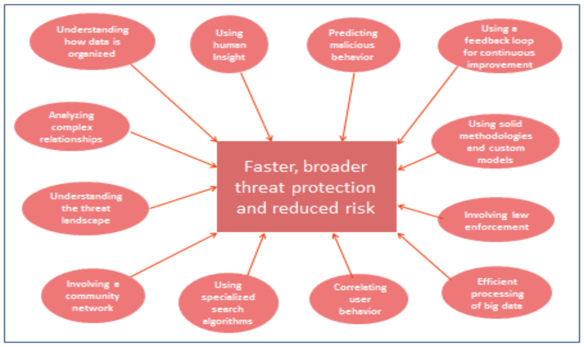
Understanding how data is organized
Using human Insight
Predicting malicious behavior
Using a feedback loop for continuous improvement
Analyzing complex relationships
Faster, broader threat protection and reduced risk
Using solid methodologies and custom models
Understanding the threat landscape
Involving law enforcement
Involving a community network
Using specialized search algorithms
Correlating user behavior
Efficient processing of big data

Figure-1 Interior components of threat protection related to big data

Conclusion drawn from the Survey includes the following issue for the Management of Big data:

- As there's a demand in IT services & rapid growth in data volume, so do the challenges with managing data. Overall, data environments are not consolidated-enterprises aren't consolidated- enterprise are still running many separate databases for applications.
- During these twelve months as well as the associated data volumes grew by more than 20% on average. In contrast, many organizations are experiencing flat or shrinking budgets.
- Close to two-fifths of enterprises either already has or are allowing to run database functions inside a private cloud.
- As the growth of data accelerates—both in terms of number of databases and in data volume—database administrators need to know how it's going to throw impact in the systems. In such environments, most respondents have managing responsibility at the database level of the technology stack in order to deal with issues. As adoption of private, hybrid and public cloud increases, the challenge of being able to control data moving into virtualized environments will accelerate as well.
- Enterprises are to find a better hold on managing database changes, leading to shorter cycle times during the database lifecycle. Conversely, the work that goes into managing the database lifecycle is still the greatest consumer of respondents' time. A majority still tremendously perform a range of tasks manually, from patching

databases to performing upgrades. As databases travel into virtualized and cloud environments, there will be a need for more comprehensive enterprise-wide testing.

- Episodes of unplanned downtime are trending upward at many enterprises. When an outage does occur, respondents admit that it's difficult to tell what the root causes may be. Merely a quarter of respondents can tell, if cross-tier components are affecting database performance.
- Close to half the respondents say they need visibility into the entire technology stack in order to do their job efficiently. Everything is included from the database applications that encompass the service being provided to the business, downward to the storage layer. The ultimate responsibility for troubleshooting application problems typically falls to database professionals, the survey finds [25], [26].

VI. SECURITY AND CHALLENGES PROBLEMS THAT ARE BEING FACED

Big data is hardly new; many organizations have collected and used large quantities of data for decades. In recent years, the idea of Big data has been taken off, in large part because organizations of all sizes and budgets now have access to infrastructure via cloud that enables Big data opportunities. While new opportunities are immense for business, it's still not clear whether numerous organizations are thinking about the security implications of Big data projects.

Deploying Cloud Computing have managed to reduce the cost of IT infrastructure costs while providing compute and storage capacity on demand, increase IT flexibility, become more competitive among emerging players and most time to time to market objectives[26],[27]. Though, it has many Business benefits, it has many challenges as well.

In June'2013, the Cloud Security Alliance (CSA) Big Data Working Group [27] released its security challenges document, which includes the following:

1. Secure computations in distributed programming frameworks

The risks identified in security of computational elements in frameworks such as MapReduce, have two specific security concerns outlined. First, the trustworthiness of the "mappers", which are the code that breaks data into pieces, analyzes it and outputs key-value pairs, needs to be evaluated. Second, data sanitization and de-identification capabilities need to be implemented to prevent the storage or leakage of sensitive data from the platform should be implemented through data sanitization and de-identification. Enterprises using complex tools such as MapReduce will need to use tools such as Mandatory Access Controls within SELinux and de-identifier routines to accomplish this, on the same note, enterprises should inquire as to how cloud providers are controlling and remediating this issue in their environments.

2. Security best practices for non-relational data stores

The use of NoSQL, and other large-scale, non-relational data stores may create security issues due to possible lack of capabilities in several issues due to a possible lack of capabilities in several vital areas, including any real authentication, encryption for data at rest or in transit, logging or data tagging, and classification. Organizations need to consider the use of separate applications or middleware layers to enforce authentication and data integrity. All passwords must be encrypted, and any connections to the systems ideally use Secure Sockets Layer/Transport Layer Security. Ensure logs are generated from all transactions around sensitive data as well.

3. Secure data storage and transactions logs

Data and transaction logs may be stored in multi-tiered storage media, but organizations need to defend against unauthorized access and ensure continuity and availability. Policy-based private key encryption can be used to ensure that only authenticated users and applications access the platform.

4. End-point input validation/filtering

In a big data implementation, numerous endpoints may submit data for processing and storage. To ensure only trusted endpoints are submitting data and that false or malicious data is not submitted, organizations need to vet each endpoint connecting to the corporate network. The working group does not have a practical set of suggestions for mitigating this concern, unfortunately, aside from the recommendation to incorporate the Trusted Platform Module chips (found in many newer endpoint devices) into the validation process where possible. Host-based and mobile device security controls could potentially alleviate the risk associated with untrusted endpoints, along with strong processes around system inventory tracking and maintenance.

5. Real-time security/compliance monitoring

Monitoring big data platforms, as well as performing security analytics, should be done in near real time. Many traditional security information can't keep pace with the large quantity (and formats) of data in use within true big data implementations. Currently, little true monitoring of Hadoop and other big data platforms exists, unless database and other front-end monitoring tools are in use.

6. Scalable and composable privacy-preserving data mining and analytics

Big data implementations can lead to privacy concerns around data leakage and exposure. There are a number of security controls that can be put in place to help organizations deal with this problem, including the use of strong encryption for data at rest, access controls to data, and a separation of duty processes and controls to minimize the success of insider attacks.

7, Cryptography enforced access control and secure communication

Historically, the popular approach to data control has been to secure the systems that manage the data, as opposed to the data itself. However, those applications and platforms have proven vulnerable time and again. The use of strong cryptography to encapsulate sensitive data in cloud provider environments, as well as new and innovative algorithms that more capably allow for key management and secure key exchange, are a more reliable method for managing access to data, especially as it exists in the cloud independent of any one platform.

8, Granular access control

Enacting fine-grained access to big data stores such as NoSQL databases and the Hadoop Distributed File System requires the implementation of Mandatory Access Control and sound authentication. New NoSQL implementations such as Apache Accumulo can facilitate very granular access control to key-value pairs; cloud service providers should also be able to articulate the types of access controls that are in place in their environments.

9, Granular audits

In conjunction with continuous monitoring, regular audits and analysis of log and event data can help to detect intrusions or attack attempts within the big data environment. The key control to focus on here is logging at all layers within and surrounding the big data environment

10, Data provenance

Provenance in this case is focused on data validation and trustworthiness. Authentication, end-to-end data protection and fine-grained access controls can help to verify and validate provenance in big data environments; cloud service providers should have these controls in place already to address other issues .

The big problem with clouds is making the storage perform and this would be the biggest reason why some people wouldn't use the cloud for big data processing.- RobertJenkins Chief Technology Officer, Cloud Sigma

VII. SOLUTIONS FOR OVERCOMING DATA SECURITY CHALLENGES

Cloud Computing has adopted in many industries sectors with adoption cum security concerns but there are ways to Ensure privacy & Prevention of Data Lose [28]

"Cloud cited as a priority for global financial services CIO. 39% of the surveyed except that more than half of all their transactions will be supported via. Cloud infrastruction & Software As a Service (SaaS) by 2015"- Gartner

Now let's see how to overcome Data Security in the cloud. Key concerns while moving to the cloud are:

1, Data Residency- The key concerns to Data Residency are :

- ACCESS: Who manages the data & have access to the data?
- LAW REGULATION: Where is the data stored & laws that apply?
- DATA BREACH: Will you know when data is breached?
- CONTRACT TERMINATION: Will data remain in the cloud even after termination of service?

Hence, solutions to Data Residency concerns are:-

- Data Encryption is mathematical solution of converting clear text data into Cipher text, which can't be read by anyone other than the customers who retains the encryption keys.
- *Tokenization* is a solution where, the actual data resigns locally in a Token Database . Randomly generated tokens are associated with data and are sent to the cloud. Hence, data can only be read by the custodian of the database.

2, Data Security- The key concerns to Data Security are:

- There has been a rise in new threats related to Data Security in Clouds. IT infrastructures are controlled by many new attackers. For instance, in an organization there's a dept. responsible for Cloud Computing Security. They have a full control over the Cloud Computing authentication framework. Therefore, in case any of the controllers are not careful, then it's easy for the attackers to access data.

Hence, solution for Data Security is:

- Porticor (VPD) Virtual Private Data System is the solution for such threats. It provides security i.e. necessary to make clouds trusted. Also, used for encryption of business- data & maintain keys while under the control [29], [30].

VIII. CONCLUSION

The goal of Managing Big Data in the Cloud is mainly for the management of data issues such as portability, integration- includes connectivity, transformation, service level integration, Business logic, management of data.

Big Data analytics is used for obtaining actionable intelligence in real time. Although, Big Data analytics have important guarantee, there are a number of challenges too. The following issues needs to be addressed:

1. Data provenance: Authenticity and integrity of data used for analytics. As Big Data expands the sources of data it can utilize, the trustworthiness of data source needs to be verified and the inclusion of ideas such as adversarial machine learning must be explored in order to identify maliciously inserted data.

2. Privacy: we need regulatory incentives and technical mechanisms to minimize the amount of inferences that Big Data users can make.

3. Securing Big Data: Securing of Big Data is very crucial for an organization. Since, it's an asset.

4. Human-computer interaction: Big Data might facilitate the analysis of diverse sources of data, but a human analyst still has to interpret result. Compared to the technical mechanisms urbanized for efficient computation and storage, the human-computer interaction with Big Data has received less attention and this is an area that needs to be raised. First step is the use of visualization tools to help analysts understand the data of their systems.

Therefore, advice for IT organizations is:

- To plan for future needs
- Assure privacy/Security
- Evaluate budget
- Be patient
- Network with others
- Consider multiple options/vendors(Only reliable vendors)
- Data virtualization
- Involvement of end users

I hope that this initial report on Managing Big Data in the Cloud outlines some of the elementary issues related to the challenges faced by organizations in managing Big Data and the importance of the cloud computing 31], [32], [33].

REFERENCES

[1] Jeff Kelly, David Floyer, Dave Vellante, Stu Miniman-Big Data Vendor Revenue and Market Forecast(2012-2017)

[2] Cloud Computing- http://www.wikibon/cloudcompuuting
http://www.linkedin.com/today/post/article/20130919142517-64875646-what-the-hell-is-the- cloud?trk=mp-author-card

[3] Gartner's glossary- http://www.gartner.com

[4] Ali Ghodsi, Vyas sekar, Matei Zaharia, and Ion Stocia. Multi-resource fair queueing for packet processing. INSIGCOMM, 2012

[5] Big data to drive surveillance society computerworld http://www.computerworld.com/8/article/92150331/

[6] Will Garside, Brain Cox- Big data Storage

[7] Bernard Marr Best-Selling Author, Keynote Speaker and Consultant in Strategy, Performance Management, Analytics, KPIs and Big Data

[8] Helen Knight, MIT News correspondent Storage System for'Big data' dramatically speeds access to information

[9] Manage the Data Lifecycle of Big data Environments http://www01.ibm.com/software/data/optim/data-lifecycle-big-data/

[10] Webinar-Workshop on the future of Big data Management http://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=246453

[11] Nessi-Big Data White Paper http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf

[12] Bob Violino- Managing Big data in the cloud

[13] http://www.ibm.com/software/data/bigdata/industry.html

[14] D.Agrawal, S Das and A.E. Abbai. Big data & Cloud Computing: New wine or just new bottles? PVLDB, 3(2):1649-1648

[15] D. Agrawal, A.E.I Abbadi, S.Antony, and S.Das, Data Management Challenges in Cloud Computing Infrastructure In DNIS, pages 1-10, 2010

[16] Sue Nelson Big Data: The Harvard Computers- http://www.nature.com/nature/journal/v455/n7209/full/455036a.html

[17] Dr. Wildmer & Partners, Attorneys-at-law-Cloud Computing & Data Protection – http://www.smartprotection.com

[18] Today's threat environment- http://www.trendmicro.com

[19] http://www.computerweekly.com

[20] James Staton, analyst Foreester- Why is BYOE important?

[21] Amazon EC2 crosses the Atlantic- http://aws.amazon,com/about-aws/whats-new/2008/12/10/amazon-ec2-crosses-the-atlantic/

[22] Blue cloud- http://www-o3.ibm.com/press.us.en/press release/26642.wss
[23] IT Cloud services user survey, pt 2: Top Benefits & Challenges- http://blogs:idc.com/ie/?p=210
[24] Why cloud security hinges the business alignment- http://www.searchcloudsecurity.techtarget.com
[25] Jessica Scarpati, site editor- Service providers anticipate SMB demand for Big data cloud Analytics
[26] Ten storage trends for 2014
[27] Top challenges of Big data in cloud- http://cloudsecurityalliance.org
[28] Overcoming Big data security Challenges- http://www.capegemini.com/customers-experience-management
[29] Data Security- http://www.porticor.com
[30] http://www.techgig.com
[31] Fern Halper (Research director,Advanced Analytics, TDWI) & Suzanne Hoffmen (Senior Director, Analyst Relation, Tableau)- Big data in the Cloud
[32] Philip Russom (TDWI Research Director, Data Management), Alic Westerfiled (Senior V.P, Enterprise, Liasion)- Preparing Data for Analytics
[33] Colin While, BI Research January'2014- Why BI in the Cloud