Private Cloud Storage in Big Data

Hasnain Nawaz
Department of Business Technology,
Dubai, United Arab Emirates

Tariq Rahim Soomro
Department of CS, SZABIST,
Dubai Campus, Dubai
United Arab Emirates

ABSTRACT

Current and upcoming era is the era of Big Data, where it comes along with challenges between Big Data processing and security that raises problems. On one side, the large amount of trust related data have to be highly available, flexible and fast processing and on the other side, the trust and security of Big Data are the challenges. This paper highlights the problems of Big Data and it explores the study of Big Data processing and security with fast emerging technology of cloud computing that is its private cloud storage.

Keywords

Big Data, Cloud Computing, Private Cloud Storage.

1. INTRODUCTION

Big Data is a methodology that describe the exponential growth and availability of data, both in the form of structured and unstructured [1]. Big Data refers to the idea that an enterprise can mine all the data it collects right across its operations to open brilliant nuggets of business intelligence whereas companies in the past had to rely on sampling Big Data. Big Data is important to business, society as the Internet has become and it may lead to more accurate analyses [2]. Big Data requires huge amount of storage space; a typical storage will be based on clustered Network-Attached Storage (NAS). Clustered NAS infrastructure comprises configuration of several NAS shells, where each NAS shell of several storage devices connected to a NAS device. The sequence of NAS devices then interconnected to permit huge sharing and searching of data [3]. In recent past business were dealing with tens of hundreds of gigabytes of storage for personal computers and today facing challenges in tens of thousands of terabytes. The current growth rate for the amount of data collected is staggering. A big task for IT researchers is that this growing rate is fast exceeding our capability in both “design appropriate systems” to handle the data effectively and “making the data as secure as possible” for decision making [4]. Cloud computing utilizes imagining of computer assets to run various existing virtual Servers on the same physical machine. This had accomplished he economies of scale, which is lower costs and charging on what service one had selected. This standardization makes it a flexible choice for computing needs [5]. In Cloud computing large combination of remote Servers are networked with each other to allow centralized data storage, services and online access to computer resources. The cloud services are categorized into three basic services as; Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). There are three type of clouds and are classified as Private Cloud, Public Cloud, and Hybrid Cloud. A public cloud is pay-as-you-go service to the general public, in public cloud the organization does not own the core technology resources and services, but those are outsourced. A private cloud is an internal data center of the organization that is not openly available to public. In private cloud configuration, assets, resources and services are owned by the system, but those are accessible internally. As the technology is owned and functioned by the organization, this type of cloud is expensive than public cloud but is more secure. The private cloud is the internal cloud residing within the company’s firewall and managed by data center experts [5]. A hybrid cloud is similar to both public and private; it uses a public cloud for some tasks and a private cloud for another task. This allows the company to preserve critical, confidential data and information within the firewall while the public cloud for non-confidential data [3].

Data storage using cloud computing is a feasible option for business seeing the use of Big Data analytic techniques. Cloud computing is on-demand network access to computing resources, which are often available in the mart that provides by an outside entity as well as inside cooperate business infrastructure that requires little management, which is likewise known as public cloud and private cloud [3].

This paper will discuss the problems of Big Data saved in a private cloud in terms of cost reductions, time reduction, high performance and availability, smarter business decision making. The paper is organized into 5 sections. Section 2 discusses the problem and issues of Big Data; Section 3 focuses on challenges of Big Data; Section 4 provides the solution of Big Data using private cloud storage; and Section 5 will discuss the paper and highlight possible future work.

2. PROBLEMS & ISSUES OF BIG DATA

The idea of big data had been widespread within Computer Science since in the earliest days of computing technology. The big data initially implied the volume of the information that couldn't be transformed (effectively) in conventional database techniques and devices. Each time another storage medium was invented, the amount of data open blasted with light that is easily accessible to others. There are three fundamental problem areas while dealing with big data: Storage issues, Management issues, and Processing issues [4].

2.1 Storage Problems of Big Data

Data is always being growing and there is a big data phenomenon occurred from last few years, which is because of popularity of new databases such as Hadoop and NoSQL databases. There is no single solution for Big Data storage as long as it depends on environment, if one asks big storage vendors no clear concerns in the industry what actual customers are looking for big data storage. Generally, one has to start looking at performance requirement, how critical the data is, is this big data that arrives every night and if one lose it, it doesn’t matter in 24hrs time or is it big data that taking a long time to collect and it's highly valuable that what one need to protect. How scalable one wants its storage system; like one buy 200 Terabyte storage and in 5 years’ time need is going more than that. If one will be storing the big data within the appliance of database definitely the performance will be good, but the storage issue occurs, which is much more expensive. Therefore, the SAN storage systems are used for big data, where the performance issues come up, but it also can be handle [6].

2.2 Management Problems of Big Data

Dealing with the data is the most difficult converse with big data. This issue initially surfaced a decade ago in the UK e-Science activities, where data was scattered geologically and "owned" and "managed" by numerous elements. Despite of the accumulation of data by manual procedures, where difficult protocols are utilized to guarantee accurateness and validity, digital data gathering is substantially looser. The efficiency of digital data configuration denies and altered technique for data gathering. Data essential regularly concentrates more on missing data or exceptions than attempting to accept each thing. Data is frequently fine-gained, for example metering of data. For the given volume it is exceptionally unreasonable to approve each data item, new approach to the data necessity and confirmation is required [4]. Source of this data are diverse, both temporary and spatially, by format, and by method for accumulation. Digital data in mediums agreeable to them, for example, reports, drawings, pictures, sound and feature recording, models, programming practices, client interface plans, etc. with or without sufficient metadata telling what, when, where, who, why and how it was gathered and its source. This data is promptly accessible for survey and examination. No doubt data and data source is a discrimination issue. There is no generally acknowledged approach to store raw data, diminished data, code and parameter decisions that created the data. Thus, there is no impeccable big data administration arrangement yet and depicts a vital gap in the research literature on big data that needs to be filled [7].

2.3 Processing Problems of Big Data

Approaches for analyzing and mining big data are basically different from traditional analysis on small samples. Big data is noisy, dynamic, heterogeneous, interrelated and untrustworthy. Noisy Big Data is more valued to tiny samples [8]. Effective processing of Exabyte's of data will require extensive parallel processing and new analytics procedures in order to provide timely and actionable data [4]. The problem with current Big Data is not synchronizing between database systems, where the data is hosted and provide SQL querying and statistical packages that perform various forms of non-SQL processing, such as data mining and statistical analyses [8].

3. ISSUES IN CHARACTERISTICS OF BIG DATA

According to [9] following are the issues in characteristics of Big Data:

Data Volume: when the data size rises, the significance of diverse data records will decrease in the measure to age, type, richness, and quality among other issues.

Data Value: The data analyzing is done by the different organizations from the data that is stored. This shows a gap between business leaders and IT professionals, the focus of business leaders is to add value to business and get more profit as compared to the IT leaders those have concerns with the technicalities of the storage and processing.

Data Velocity: The old-style schemes are not capable enough of performing the analysis on the data, which is frequently in motion and is a lot more than a bandwidth issue.

Data Variety: This data is diverse including of raw, structured, semi structured and even unstructured data, which is challenging to be handled by the existing old-style analysis systems.

Data Complexity: Trouble of big data is working by utilizing standard databases and desktop insights/ visualization bundles, requesting infinite parallel programming running on tens, hundreds, or even a great many servers.

4. CHALLENGES IN BIG DATA

In the age of big data, where visualization is becoming an important component of analytics, the organizations are looking forward to overcome several challenges to visualization and big data to obtain full advantage of visual analytics. Here are some challenges related to big data [10].

4.1 Uncertainty of the Data Management in Big Data

One disruptive aspect of big data is the use of a variety of innovative data management frameworks whose designs are proposed to support both operational and to a great extent, analytical processing. These methodologies are mostly lumped into a class referred to as NoSQL frameworks that are different from the conventional relational database management system model in terms of storage model, data access methodology, and are largely designed to meet performance demands for big data applications. The wide variety of NoSQL tools, developers and the status of the market are creating uncertainty with the data management landscape [11].

4.2 Scalability in Big Data

The scalability issue in big data has led to cloud computing, which combines different workloads with varying performance goals into very large clusters. This shows a high degree of sharing of resources, which is expensive and also brings with it various challenges [9].

4.3 The Big Data Talent Gap

It is challenging to examine the analyst and high-tech media without being blasted with content advertising the value of big data analytics and a corresponding reliance on a broad variety of disruptive technologies. These new tools from old-style relational database tools with alternative data layouts designed to increase access speed, while decreasing the storage footprint, in-memory, analytical, NoSQL data management frameworks, as well as the broad Hadoop ecosystems. The rising communities of application developers who are increasing their knowledge of tools are those covering the Hadoop ecosystem, no matter of the promotion of these big data technologies. The reality is that there is not a wealth of skills in the market for big data; the typical expert has experience through tool implementation and its use as a programming model, rather than the data management aspects. This shows that many big data experts remain inexperienced when it comes to the practical aspects of data modeling, data architecture, and data integration. There is no doubt that as more data practitioners become engaged, the talent gap will eventually close [11].

4.4 Privacy and Security in Big Data

This is the significant challenge for big data, which is sensitive and includes conceptual, technical as well as lawful essentialness. The data like (database of merchant, social networking Websites) of an individual when it is linked with external large data set, this leads to the interpretation of new facts about that individual and it is possible that these facts about that individual are private and the individual might not want that data owner to disclose his personal information to public. It is obvious that information about the people is gathered and used to add value to the business of the organization and making perceptions of their lives without their knowledge. Another concern is social stratification, where an educated individual would be taking favorable circumstances of the Big Data anticipating examination and then again disadvantaged will be effortlessly recognized and treated more badly. Big Data utilized by the law implementation will build the possibilities of certain labeled individuals to experience the forbidding effects of unfavorable results without the capability to fight back or actually having information that are differentiated [9].

4.5 Big Data Platform

It is obvious that the determined of a big data program includes processing or analyzing massive amounts of data. People had raised potentials regarding evaluating huge data sets in a big data platform, they have not been aware of the difficulty of simplifying the access, transmission, and delivery of data from the several sources and then loading those various data sets into the big data platform. The complicated aspects of data access, movement, and loading are only part of the challenge. The need to navigate extraction and transformation is not limited to structured conventional relational data sets. Analysts gradually want to import older mainframe data sets (in VSAM files or IMS structures, for example) and at the same time want to absorb meaningful illustrations of objects and ideas refined out of different types of unstructured data sources, such as emails, texts, tweets, images, graphics, audio files, and videos, all accompanied by their corresponding metadata. An extra challenge is directing the response time expectations for loading data into the program. Squeeze massive data volumes through “data pipes” of limited bandwidth will both lower the performance and may even impact data exchange [11].

4.6 Data Access, Data Source and Sharing Information in Big Data

The data in the company’s information systems are mostly used for decision making in time so it can be utilized in accurate, complete and timely manner. This complex the process of data management and government agencies to make data open for them and available in a standardized way like standard APIs, Meta data, which will help them in making decisions, business intelligence and improvements in productivity [9]. From a data currency perspective, the data source synchronization implies that the data coming from one source is not out of date with data coming from another source. With conventional data marts and data warehouses, sequences of data extractions, transformations, and migrations all provide situations in which there is a risk for information to become unsynchronized [11].

4.7 Technical Challenges in Big Data

With the Fault Tolerance capabilities in technology like cloud computing and big data is always seen that whenever the failure happen the damage should be captured in agreed on set period rather than reaching it to extreme level to avoid the whole task to start from the scratch. Fault-tolerant has been always hard, and requires complex systems. The major task is to cut the chance of failure to an acceptable level [9].

5. SOLUTION USING PRIVATE CLOUD STORAGE

With the increase of cloud computing and cloud data stores had been precursor and facilitated to the emergence of big data. Cloud computing employs visualization of assets to operate various standardized virtual servers; on the same physical machine, which make it highly available option for computing needs [5]. Big data and cloud technology goes together, big data storage requires cluster of servers for processing and secure transition, which cloud structure can readily provide with more secure in private cloud [12]. NAS (Network Attached Storage) based storage systems is the most significant to a private cloud storage infrastructure solution with complete serviceable, lower cost substitute to traditional block level SAN (Storage Area Network) storage and special purpose storage fabrics, which provides Fiber Channel Networks. NAS-based storage systems manage their physical disk units like a way that is similar to SAN manages theirs, as physical storage units are accumulated into arrays, and from those arrays are surfaced volumes, these volumes then accessible to the file system layer and made available over the network. The storage components can be managed by software (like platform of an operating system) or hardware (like RAID utility supported by SAN) across locally attached storage or through storage backed by a SAN. The key features of NAS are availability, performance, scalability, security and cost [13].

5.1 Private Cloud

Private Cloud is for single organization, which have the capability to organize and accomplish it in line with their needs to succeed in a custom-made solution [14]. Private Clouds are devoted to one connection and doesn't pass on physical assets. The asset can be utilized in-house alternately remotely. The primary requirements of private cloud deployments are security measurements and regulations that need a strict separation of an organization’s data storage and processing from malicious access through shared resources. Private cloud systems are inheritance systems with extraordinary hardware requires exceptional resource demand, like extreme memory or computing instances, which are not available in public clouds [5]. Virtualization in enterprise organizations are to cut IT costs and create a more flexible, effective and automated workload environment. The goal of private cloud model is to allow an administrator to manage technology infrastructure from a single point, where resources can be allocated as needed and its enterprise computer architecture is protected by a firewall. This is especially important for businesses that have strict compliance regulations [15]. Storage is the backbone of the private cloud. There are two types of private cloud namely, On-Premise Private Cloud, Externally-Hosted Private Cloud [16]. In On-Premise private cloud data center capabilities to support a private cloud can be assessing and on an external-hosted private cloud external providers can be used to host private cloud.

5.2 Traditional Data Center to a Private Cloud

To control and manage big data the organization should move from traditional data center to a private cloud to secure their data for high availability and performance [17]. There are five steps to migrate to private cloud: standardization, consolidation, virtualization, automation and orchestration [18].

5.2.1 Standardization

Virtualized the architecture, as much as possible, this will ease the management in reducing the components to have a single point of management.

5.2.2 Consolidation

It unifies the network where all the servers are connected through one cable.

5.2.3 Virtualized and Automation

Virtualizing the compute nodes that effectively transforms CPU into stateless processing the power and memory. Stateless hardware can be utilized for other hypervisors which result more rapid provisioning and helps in fast automation of implementing the VMs and Operating systems.

5.2.4 Orchestration

It allows a quick provision of components like storage, network and applications from a single management console and enables backup and replication services.

There are many companies in market providing infrastructure, resources, and hardware for building private cloud like HPC Storage Cloud, NetApp, VNX, EMC, Microsoft Azure, etc. This study discusses the best practice of NetApp that provides clustered Data to create a virtualized IT infrastructure for Private Cloud Computing by its best practices Services Analytics (Optimize your services), Automation (Deploy your services faster), Self-Service (Empower IT and your end users) [19].

5.3 Benefits of Private Cloud

5.3.1 Infrastructure Simplification

Single provider is to be selected to provide the storage, server, virtualization and networking layers for the private cloud. Simplification in the production data center enables just two racks to house roughly 3.6 PB of private cloud storage and 800 virtual servers. Presently more than 530 applications in 97% of the total and all desktops and virtual machines running on VMware vSphere and hosted on high-availability NetApp enterprise storage systems [20].

5.3.2 Proven Disaster Recovery

A private cloud computing solution will cut the cost and the time to recover from any outage. This is done by using virtualization [20]. Private cloud disaster recovery delivers business peace of mind by significant that data is kept safe and secure [21]. Disaster Recovery (DR) as a Cloud Service, where it has two levels of resources required Replication Mode or Failover Mode, as shown in Figure 1. During standard operation the system stays in Replication Mode and involves only a single low cost VM to act as the DR server that handles the state of synchronization. When Disaster occurs the system enters to Failover Mode, where it involves the resources to support the full application [22].

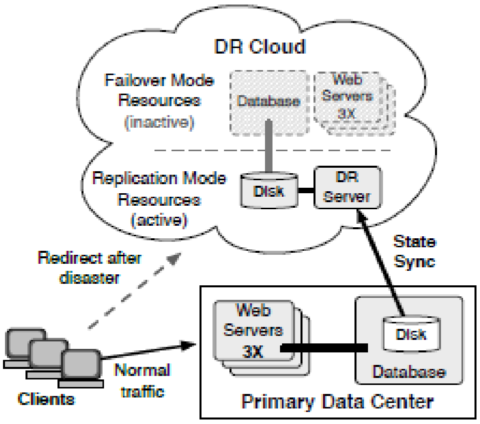

Figure 1 – Disaster Cloud Service [22]

5.3.3 Transforming Big Data into Business

IT organizations are responsible for converting, this data into actionable information. Before, requests for new business intelligence application had to wait until storage capacity could be added, after which a mirrored copy of the production database was created, which have to pass by the following process of extract, transform and load operations. With this new cloud facility model developed that pulls data from all global systems very easily, even if it is structure and unstructured data. The ease of creating zero-overhead clones no extra infrastructure cost is added to mine the data and run experiments. Clone copies the data sets are used for large scale application testing, change control, creating golden images and even for modeling the effect of different email retention policies on storage requirements [20].

5.3.4 Information Processing Capabilities

The higher information demanding of the service activity, the easier is to use that information to perform this activity at a time and location that is more efficient and results in higher quality. Cloud based system compared to the traditional computing method provide more effective data processing capabilities due to the flexible nature of the cloud computing infrastructure. It allows organizations to speedily scale up service usage and mirror information processing demands of organizations providing them with additional flexibility and scalability capabilities. This capability of cloud computing is strengthened with cloud's architecture, it is controlled interface delivered through Application Programming Interfaces (APIs) make applications more accessible by other applications and systems [23].

5.3.5 Cost and Energy

Effective implementation of a private cloud can improve the sharing of resources within an organization, where safeguarding the availability of resources to individual departments and business entities can directly respond to their demand in a flexible manner [22]. Private cloud allows organizations to have greater flexibility to shift workloads between their servers as they see spikes in usage or when they deploy new applications [24].

5.3.6 Higher Security and Privacy

Private clouds are more secure than public clouds, as private cloud using techniques, which having various pools of resources with access limited to connections made by an organization's firewall, secure leased lines and on-site internal hosting that guarantee the safe operations without any security breach [14].

6. DISCUSSION & FUTURE WORK

This study discusses the issues and challenges of big data and storage solution for big data as most of the business are directly depending on big data for future forecasting and tend analysis. This study also discusses the capabilities of the private cloud storage as it is more superior to public cloud for big data and information systems. Private cloud provides high security, availability, processing, flexibility, reliability and implementing big data is cost saving and ensures enterprises of data protection and security. There is many ongoing future research are in place to do discuss more on the security of cloud in a virtualized environment and how to handle more in coming data into cloud regardless of private or public cloud.

7. REFERENCES

[1] Thomas H. Davenport Jill Dyche, May 2013, Source: SAS the Power to Know, http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[2] Ben Rooney, Big Data’s Big Problem: Little Talent, April 29, 2012, The Wall Street Journal, http://www.wsj.com/articles/SB10001424052702304723304577365700368073674

[3] Bernice M. Purcell, Big data using cloud computing, Journal of Technology Research, http://www.aabri.com/manuscripts/131457.pdf

[4] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, Big Data: Issues and Challenges Moving Forward, 46th Hawaii International Conference on System Sciences, 2013, http://www.computer.org/csdl/proceedings/hicss/2013/4892/00/4892a995.pdf

[5] Prakash Janakiramna, Big Data Cloud Database & Computing, http://www.qubole.com/big-data-cloud-database-computing/, Retrieved December 2015

[6] Tim Stammers, Big data impact on storage infrastructure, Jan 2013, http://www.computerweekly.com/video/Big-data-impact-on-storage-infrastructure

[7] Jason, Data Analysis Challenges, 2008 http://fas.org/irp/agency/dod/jason/data.pdf

[8] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Challenges and Opportunities with Big Data, a community white paper developed by leading researchers across the United States, Executive Summary, 2011, http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf

[9] Harsh Kishore Mishra, Big Data Seminar Report, 2013 http://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report

[10] SAS Visual analytics, Five Big Data Challenges, 2013, http://www.sas.com/resources/asset/five-big-data-challenges-article.pdf

[11] David Loshin, Addressing five emerging challenges of Big Data, June 2014 https://www.progress.com/~/media/Progress/Documents/Papers/Addressing-Five-Emerging-Challenges-of-Big-Data.pdf

[12] Edd Dumbill, Big Data in the Cloud, How do the cloud offerings from Amazon, Google and Microsoft compare?, Feb 2012, http://radar.oreilly.com/2012/02/big-data-in-the-cloud-microsoft-amazon-google.html

[13] Debra and Tom Shinder, Private Cloud Storage network Storage Considerations (Part 2), 30 Jan, 2014, http://www.cloudcomputingadmin.com/articles-tutorials/private-cloud/private-cloud-storage-network-storage-considerations-part2.html

[14] Interoute, What is Private Cloud, http://www.interoute.com/cloud-article/what-private-cloud, retrieved December 2015

[15] SearchCIO, Virtualization and the private cloud: A guide for enterprise CIOs, http://searchcio.techtarget.com/Virtualization-and-the-private-cloud-A-guide-for-enterprise-CIOs, retrieved December 2015

[16] T.Swathi, K.Srikanth, S. Raghunath Reddy, Virtualization in Cloud Computing, International Jouranl of Computer Science and Mobile Computing, Vol. 3, Issue 5, May 2014, http://www.ijcsmc.com/docs/papers/May2014/V3I5201499a.pdf

[17] Insight, Private Cloud, http://www.insight.com/us/en/solutions/cloud/private-cloud.html, retrieved December 2015

[18] Ryan Huang, Six steps in migrating to private collude, May 2014, http://www.zdnet.com/article/six-steps-in-migrating-to-private-cloud/

[19] NetApp, Storage is the Backbone of the Private Cloud, http://www.netapp.com/us/solutions/cloud/private-cloud/private-cloud-technologies.aspx, retrieved December 2015

[20] NetApp, Revlonog Realizes Benefits of Private Cloud Storage, 2014 http://searchstorage.techtarget.com/NetAppSponsoredNews/Revlon-Realizes-Benefits-of-Private-Cloud-Storage

[21] CONTEGIX, Private cloud computing for disaster recovery, http://www.contegix.com/private-cloud-computing-for-disaster-recovery/, retrieved December 2015

[22] Timothy Wood, Emmanuel Cecchet, K.K. Ramakrishnan, Disaster Recovery as a Cloude Service: Economic benefits & Deployment Challenges, http://lass.cs.umass.edu/papers/pdf/HC10-dr-cloud.pdf, retrieved December 2015

[23] Anna Zhygalova, Perceived value of Cloud based Information Systems. Case: Accounting Information Systems, Masters Thesis, 2013, http://epub.lib.aalto.fi/en/ethesis/pdf/13525/hse_ethesis_13525.pdf

[24] Antony Savvas, The benefits of Private Cloud Computing, 2014, http://www.itproportal.com/2014/05/12/the-benefits-of-private-cloud-computing/