ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 4, February 2018

Big Data Cloud Computing: Challenges and Opportunities

U.Udhayakumar
Ph.D Research Scholar
Department of Computer Science and Applications
Bharathiyar University
Coimbatore, Tamilnadu, India
udhayakumar.msc@gmail.com

Dr.G.Murugaboopathi,
Associate Professor, Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
(Kalasalingam University), Krishnankovil,
Srivilliputtur(TK). Virudhunagar Dist., Tamilnadu, India
gmurugaboopathi@gmail.com

***Abstract*— The escalation of big data cloud computing and cloud data provisions have been a forerunner and expediter to the advent of big data. Cloud computing is the commodification of computing time and data storage using reliable technologies. The emergence of cloud computing is natural to provide one of the best technology in commercial packages. The reduction of cost wise also made a wide array of applications available to smaller companies by Cloud computing. It regularly exhibits unpredictable bursting, or immense computing power and storage needs. Organizations that are faced with architecture decisions should evaluate their security concerns or legacy systems ruthlessly before accepting a potentially unnecessarily complicated private or hybrid cloud deployment. A public cloud is often achievable with the legacy process are feasible to retain a core dataset or process internally by big data due to the flexibility it provides.**

Keywords—Big data, Cloud computing, Cloud storage, Core cloud architectures.

I. Introduction

The rise of big data cloud computing and cloud data stores have been a precursor and facilitator to the emergence of big data. Cloud computing is the commodification of computing time and data storage by means of standardized technologies [1] . It has significant advantages over traditional physical deployments. However, cloud platforms come in several forms and sometimes have to be integrated with traditional architectures.

This leads to a quandary for decision makers in charge of big data in the cloud projects. Need for more and which cloud computing is the optimal choice for their computing needs, particularly for big data task. This performs regularly exhibit unpredictable, bursting, or immense computing power and storage needs. IN time being business stakeholders expect swift, inexpensive, and dependable products and project outcomes [2,3]. The implementation of cloud computing and cloud storage, the core cloud architectures, and its performance in cloud computing.

II. BIg Data Cloud providers

A decade ago IT project or start-up that needed reliable and Internet connected computing resources had to rent or place physical hardware in one or several data centers. Today, anyone can rent computing time and storage of any size. The range starts with virtual machines barely powerful enough to serve web pages to the equivalent of a small supercomputer [4]. Cloud services are mostly pay-as-you-go, which means for a few hundred dollars anyone can enjoy a few hours of supercomputer power. At the same time cloud services and resources are globally distributed. This setup ensures a high availability and durability unattainable by most but the largest organizations.

The cloud computing space has been dominated by Amazon Web Services until recently. Increasingly serious alternatives are emerging like Google Cloud Platform, Microsoft Azure, Rackspace, or Qubole to name only a few. Importantly for customers a struggle on platform standards is underway[5]. The two front-running solutions are Amazon Web Services compatible solutions, i.e. Amazon's own offering or companies with application programming interface compatible offerings, and OpenStack, an open source project with a wide industry backing [5,6]. Consequently, the choice of a cloud platform standard has implications on which tools are available and which alternative providers with the same big data processing technologies are available.

A. Cloud Storage

Professional cloud storage needs to be highly available, highly durable, and has to scale from a few bytes to petabytes. Amazon's S3 cloud storage and Microsoft Azure Blob Storage are the most prominent solutions in the space. They promise in the range of 99.9% monthly availability and 99.999999999% durability per year [7]. This is less than an hour outage per month. The durability can be illustrated with an example. [6] discussed about Submerge Detection of Sensor Nodes. Underwater networking sensor nodes provide the oceanographic collection of data and monitoring of unmanned or autonomous underwater vehicle to explore sea recourses and gathering of scientific data. The sensor network contains the statistical data about the sensor nodes. High Speed Optical communication is provided between the nodes in a point to point fashion. The design emphasis on the modulation and demodulation of the signals and thereby providing the synchronization between the nodes. The challenges include waterproofing, casing, calibration. Furthermore the research issues are outlined. If a customer stores 10,000 objects he can expect to lose one object every 10,000,000 years on average. They sometime achieve this by storing data in multiple facilities with error checking and self-healing processes to detect and repair errors and device failures [7]. This is completely transparent to the user and requires no actions or knowledge.

A company could build and achieve a similarly reliable storage solution but it would require tremendous capital expenditures and operational challenges. Global data centered companies like Google or Facebook have the expertise and scale to do this economically. Big data projects and start-ups, however, benefit from using a cloud storage service. They can trade capital expenditure for an operational one, which is excellent since it requires no capital outlay or risk [8]. It provides from the first byte reliable and scalable storage solutions of a quality otherwise unachievable.

This enables new products and projects with a viable option to start on a small scale with low costs. When a product proves successful these storage solutions scale virtually indefinitely. Cloud storage is effectively a boundless data sink. Importantly for computing performances is that many solutions also scale horizontally, i.e. when data is copied in parallel by cluster or parallel computing processes the throughput scales linear with the number of nodes reading or writing [9].

B. Cloud Computing

Cloud computing employs visualization of computing resources to run numerous standardized virtual servers on the same physical machine. Cloud providers achieve with this economies of scale, which permit low prices and billing based on small time intervals, e.g. hourly.

This standardization makes it an elastic and highly available option for computing needs. The availability is not obtained by spending resources to guarantee reliability of a single instance but by their interchangeability and a limitless pool of replacements. This impacts design decisions and requires dealing with instance failure gracefully[10].

The implications for an IT project or company using cloud computing are significant and change the traditional approach to planning and utilization of resources. Firstly, resource planning becomes less important. It is required for costing scenarios to establish the viability of a project or product. However, deploying and removing resources automatically based on demand needs to be focused on to be successful. Vertical and horizontal scaling becomes viable once a resource becomes easily deployable.

Horizontal scaling refers to the ability to replace a single small computing resource with a bigger one to account for increased demand[10 ,11]. Cloud computing supports this by making various resource types available to switch between them. This also works in the opposite direction, i.e. to switch to a smaller and cheaper instance type when demand decreases. Since cloud resources are commonly paid on a usage basis no sunk cost or capital expenditures are blocking fast decision making and adaptation [12]. Demand is difficult to anticipate despite planning efforts and naturally results in most traditional projects in over- or under-provision resources. Therefore, traditional projects tend to waste money or provide poor outcomes.

III. Cloud Big data Challenges

Vertical scaling achieves elasticity by adding additional instances with each of them serving a part of the demand. Software like Hadoop is specifically designed as distributed systems to take advantage of vertical scaling. They process small independent tasks in massive parallel scale. Distributed systems can also serve as data stores like

NoSQL databases, e.g. Cassandra or HBase, or filesystems like Hadoop's HDFS. Alternatives like Storm provide coordinated stream data processes in near real-time through a cluster of machines with complex workflows [13].

The interchangeability of the resources together with distributed software design absorbs failure and equivalently scaling of virtual computing instances unperturbed. Spiking or bursting demands can be accommodated just as well as personalities or continued growth. Renting practically unlimited resources for short periods allows one-off or periodical projects at a modest expense [14]. Data mining and web crawling are great examples. It is conceivable to crawl huge web sites with millions of pages in days or hours for a few hundred dollars or less. Inexpensive tiny virtual instances with minimal CPU resources are ideal for this purpose since the majority of crawling the web is spent waiting for IO resources [15]. Instantiating thousands of these machines to achieve millions of requests per day is easy and often costs less than a fraction of a cent per instance hour.

The mining operations should be mindful of the resources of the web sites or application interfaces they mine, respect their terms, and not impede their service. A poorly planned data mining operation is equivalent to a denial of service attack [16]. Lastly, cloud computing is naturally a good fit for storing and processing the big data accumulated form such operations.

IV. Cloud Architectures

Three main cloud architecture models have developed over time; private, public and hybrid cloud. They all share the idea of resource commodification and to that end usually virtualize computing and abstract storage layers.

A. Private cloud

Private clouds are dedicated to one organization and do not share physical resources. The resource can be provided in-house or externally. A typical underlying requirement of private cloud deployments are security requirements and regulations that need a strict separation of an organization's data storage and processing from accidental or malicious access through shared resources. Private cloud setups are challenging since the economic advantages of scale are usually not achievable within most projects and organizations despite the utilization of industry standards. The return of investment compared to public cloud offerings is rarely obtained and the operational overhead and risk of failure is significant [17].

Moreover, cloud providers have captured the trend for increased security and provide special environments, i.e. dedicated hardware to rent and encrypt virtual private networks as well as encrypted storage to address most security concerns. Cloud providers may also offer data storage, transfer, and processing restricted to specific geographic regions to ensure compliance with local privacy laws and regulations.

B. Public Cloud

Public clouds share physical resources for data transfers, storage, and processing. However, customers have private visualized computing environments and isolated storage. Security concerns, which entice a few to adopt private clouds or custom deployments, are for the vast majority of customers and projects irrelevant. Visualization makes access to other customers' data extremely difficult [18].

Real-world problems around public cloud computing are more mundane like data lock-in and fluctuating performance of individual instances. The data lock-in is a soft measure and works by making data inflow to the cloud provider free or very cheap. The copying of data out to local systems or other providers is often more expensive. This is not an insurmountable problem and in practice encourages utilizing more services from a cloud provider instead of moving data in and out for different services or processes. Usually this is not sensible anyway due to network speed and complexities around dealing with multiple platforms.

The varying performance of instances stems typically from the dependency on what kind of load other customers generate on the shared physical infrastructure. Secondly, over time the physical infrastructure providing the virtual resources changes and is updated. The available resources for each customer on a physical machine are usually throttled to ensure that each customer receives a guaranteed level of performance. Larger resources generally deliver very predictable performance since they are much closer aligned with the physical instance's performance. Horizontally scaling projects with small instance should not rely on an exact performance of each instance but be adaptive and focus on the average performance required and scale according to need.

C. Hybird Cloud

The hybrid cloud architecture merges private and public cloud deployments. This is often an attempt to achieve security and elasticity, or provide cheaper base load and burst capabilities. Some organizations experience short periods of extremely high loads, e.g. as a result of seasonality like black Friday for retail, or marketing events like sponsoring a popular TV event. These events can have huge economic impact to organizations if they are serviced poorly.

The hybrid cloud provides the opportunity to serve the base load with in-house services and rent for a short period a multiple of the resources to service the extreme demand. This requires a great deal of operational ability in the organization to seamlessly scale between the private and public cloud. Tools for hybrid or private cloud deployments exist like Eucalyptus for Amazon Web Services. On the long-term the additional expense of the hybrid approach often is not justifiable since cloud providers offer major discounts for multi-year commitments. This makes moving base load services to the public cloud attractive since it is accompanied by a simpler deployment strategy.

V. Big Data

Big Data is an umbrella term which encompasses all sorts of data which exists today. From hospital records and digital data to the overwhelming amount of government paperwork which is archived. The categorization is not possible in Big Data under one definition or description, because we are still working on it [19]. The great thing about information technology is that it has always been available for technology companies, businesses and all types of institutions.

It was the emergence of cloud computing which made it easier to provide the best of technology in the most cost-effective packages. Cloud computing not only reduced costs, but also made a wide array of applications available to the smaller companies. The cloud is growing steadily and perceiving an explosion of information across the web. Social media is completely different and common users generate loads of data every day [20]. Organizations and institutions are also creating data on a daily basis, which can eventually become difficult to manage. Take a look at these statistics on Big Data generation in the last five years

- 2.5 quintillion bytes (2.3 Trillion Gigabytes) of data are created every day.
- 40 zettabytes (43 Trillion Gigabytes) of data will be created by 2020.
- Most companies in the US have at least 100 Terabytes (100,000 Gigabytes) of stored data.

These high volumes of data present a challenge to the cloud environment and to manage and secure the essence of this data rather than stacking it. It seems like cloud computing and big data are an ideal combination for it. Together, they provide a solution which is both scalable and accommodating for big data and business analytics [15,19]. The analytics advantage is going to be a huge benefit in today's world and all the information resources will become easily accessible.

A. *Advantages of Big Data*

The advantages of big data are termed as

a) Agility

The traditional infrastructure of storing and managing data is now proving to be slower and harder to manage. It can literally take weeks to just install and run a server. Cloud computing is here now, and it can provide your company with all the resources you need. A cloud database can enable your company to have thousands of virtual servers and process the work effortlessly in only a matter of minutes.

b) Affordability

Cloud computing is a blessing in disguise for a company that wishes to have updated technology under a budget. Companies can pick what they want and pay for it as they go. The resources required to manage Big Data are easily available and they don't cost big bucks. Before the cloud, companies used to invest huge sums of money in setting up IT departments and then paid more money to keep that hardware updated. Now the companies can host their Big Data on off-site servers or pay only for storage space and power they use every hour.

c) Data Processing

The explosion of data leads to the issue of processing it. Social media alone generates a load of unstructured, chaotic data like tweets, posts, photos, videos and blogs which can't be processed under a single category. With Big Data

Analytics platforms like Apache Hadoop, structured and unstructured data can be processed. Cloud computing makes the whole process easier and accessible to small, medium and larger enterprises.

d) Feasibility

While traditional solutions would require the addition of more physical servers to the cluster in order to increase processing power and storage space, the virtual nature of the cloud allows for seemingly unlimited resources on demand. With the cloud, enterprises can scale up or down to the desired level of processing power and storage space easily and quickly.

Big Data analytics require new processing requirements for large data sets. The demand for processing this data can raise or fall at any time of the year, and cloud environment is the perfect platform to fulfill this task. There is no need for additional infrastructure, since cloud can provide most solutions in SaaS models.

B. Challenges to big data in Cloud environment

Big Data has provided organizations with terabytes of data, it has also presented an issue of managing this data under a traditional framework. Analyzing the large volumes of data often becomes a difficult task as well. In the high speed connectivity era, moving large sets of data and providing the details needed to access it, is also a problem. These large sets of data often carry sensitive information like credit/debit card numbers, addresses and other details, raising data security concerns.

Security issues in the cloud are a major concern for businesses and cloud providers today. It seems like the attackers are relentless, and they keep inventing new ways to find entry points in a system. Other issues include ransomware, which deeply affects a company's reputation and resources, Denial of Service attacks, Phishing attacks and Cloud Abuse.

Globally, 40% of businesses experienced a ransomware incident during the past year. Both clients and cloud providers have their own share of risks involved when making an agreement on cloud solutions. Insecure interfaces and weak API's can give away valuable information to hackers, and these hackers can misuse this information for the wrong reasons. Some cloud models are still in the deployment stage and basic DBMS is not only tailored for Cloud computing. Data Acts is also a serious issue which requires data centers to be closer to a user than a provider.

Data replication must be done in a way which leaves zero room for error; otherwise it can affect the analysis stage. It is crucial to make the searching, sharing, storage, transfer, analysis, and visualization of this data as smoothly as possible. The only way to deal with these challenges is to implement next-generation technology which can predict an issue before it causes more damage. Fraud detection patterns, encryptions and smart solutions are immensely important to combat attackers [17, 20]. At the same time, it is your responsibility to own your data and keep it protected at your end while looking for business intelligent solutions that can ensure a steady ROI as well.

VI. Cloud Big data Implementation

Typical cloud big data focus on scaling or adopting Hadoop for data processing. MapReduce has become a de facto standard for large scale data processing. Tools like Hive and Pig have emerged on top of Hadoop which make it feasible to process huge data sets easily. Hive for example transforms SQL like queries to MapReduce jobs. It unlocks data set of all sizes for data and business analysts for reporting and greenfield analytics projects.

Data can be either transferred to or collected in a cloud data sink like Amazon's S3, and Microsoft Blob Storage, e.g. to collect log files or export text formatted data. Alternatively database adapters can be utilized to access data from databases directly with Hadoop, Hive, and Pig. One great example is their mongoDB adapter. It gives Hive table like access to mongoDB collections.

Ideally a cloud service provider offers Hadoop clusters that scale automatically with the demand of the customer. This provides maximum performance for large jobs and optimal savings when little and no processing is going on. Amazon Web Services Elastic MapReduce and Azure HDInsight, for example, allow scaling of Hadoop clusters. However, the scaling is not automatically with the demand and requires user actions. The scaling itself is not optimal since it does not utilize HDFS well and squanders Hadoop's strong point, data locality. This means that an Elastic MapReduce cluster wastes resources when scaling and has diminishing return with more instance. Furthermore, Amazon's Elastic MapReduce and HDInsight require a customer to explicitly request a cluster every time when it is needed and remove it when it is not required anymore.

VII. CONCLUSION

There is also no user friendly interface for interaction with or exploration of the data. This results in operational burden and excludes all but the most proficient users. Qubole is a leading provider of cloud based services in this space. They provide unique database adapters that can unlock data instantly, which otherwise would be inaccessible or require significant development resource. Qubole scales Hadoop jobs to extract data as quickly as possible without overpowering the mongoDB instance.

Acknowledgment

I would like to convey my heartfelt thanks to Mr. G. Murugaboopathi, my research guidance for completion of my project. He helped me to understand and remember important details of the project and gave his guidance in completing of my project successfully.

References

[1] Agrawal, Divyakant, Sudipto Das, and Amr El Abbadi. "Big data and cloud computing: current state and future opportunities." In *Proceedings of the 14th International Conference on Extending Database Technology*, pp. 530-533. ACM, 2011.

[2] Armbrust, Michael, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee et al. "A view of cloud computing." *Communications of the ACM* 53, no. 4 (2010): 50-58.

[3] Assunção, Marcos D., Rodrigo N. Calheiros, Silvia Bianchi, Marco AS Netto, and Rajkumar Buyya. "Big Data computing and clouds: Trends and future directions." *Journal of Parallel and Distributed Computing* 79 (2015): 3-15.

[4] Fernández, Alberto, Sara del Río, Victoria López, Abdullah Bawakid, María J. del Jesus, José M. Benítez, and Francisco Herrera. "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4, no. 5 (2014): 380-409.

[5] Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. "The rise of "big data" on cloud computing: Review and open research issues." *Information Systems* 47 (2015): 98-115

[6] Christo Ananth, S.Surya, Berlin Mary, "Submerge Detection of Sensor Nodes", International Journal Of Advanced Research Trends In Engineering And Technology (IJARTET), Volume II, Special Issue XXV, April 2015.

[7] Itani, Wassim, Ayman Kayssi, and Ali Chehab. "Privacy as a service: Privacy-aware data storage and processing in cloud computing architectures." In *Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on*, pp. 711-716. IEEE, 2009.

[8] Jeyakumar, Balajee, MA Saleem Durai, and DaphneLopez. "Case Studies in Amalgamation of Deep Learning and Big Data." In HCI Challenges and Privacy Preservation in Big Data Security, pp. 159-174. IGI Global, 2018.

[9] Ji, Changqing, Yu Li, Wenming Qiu, Uchechukwu Awada, and Keqiu Li. "Big data processing in cloud computing environments." In *Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on*, pp. 17-23. IEEE, 2012.

[10] Mdarbi, Fatima Ezzahra, Nadia Afifi, and Imane Hilal. "Comparative Study: Dependability of Big Data in the Cloud." In *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*, p. 19. ACM, 2017.

[11] Priya, V., Subha, S., & Balamurugan, B. (2017).Analysis of performance measures to handle medical E-commerce shopping cart abandonment in cloud. Informatics in Medicine Unlocked.

[12] Priya. V., Cloud Service for Best Gateway in VANET, International Journal of Advanced Research in Computer Science and Software Engineering, 2014, Volume 4, Issue 4, 311-318.

[13] Priya.V., Subha S. (2014).Improving the Performance of Load Balancing in Cloud Environment Using SJF in MapReduce, International Journal of Computer & Organization Trends, 2014, Volume 7, Issue 1, 10-13.

[14] Ranjith, D., J. Balajee, and C. Kumar. "In premises of cloud computing and models." International Journal of Pharmacy and Technology 8, no. 3 (2016): 4685- 4695.

[15] Sethumadahavi R Balajee J "Big Data Deep Learning in Healthcare for Electronic Health Records," International Scientific Research Organization Journal, vol. 2, Issue 2, pp. 31–35, Jul. 2017.

[16] Ushapreethi P, Balajee Jeyakumar and BalaKrishnan P, Action Recongnition in Video Survillance Using Hipi and Map Reducing Model, International Journal of Mechanical Engineering and Technology 8(11),2017,pp. 368–375.

[17] Varghese, Blesson, and Rajkumar Buyya. "Next generation cloud computing: New trends and research directions." *Future Generation Computer Systems* 79 (2018): 849-861.

[18] Vedhanayagam P., S. S., Balusamy B., Vijayakumar P. and Chang V. (2017). Analysis of Measures to Achieve Resilience during Virtual Machine Interruptions in IaaS Cloud Service .In *Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security*ISBN 978-989-758-245-5, pages 449-460.

[19] Wang, Cong, Qian Wang, Kui Ren, and Wenjing Lou. "Privacy-preserving public auditing for data storage security in cloud computing." In *Infocom, 2010 proceedings ieee*, pp. 1-9. Ieee, 2010.

[20] Wang, Yichuan, LeeAnn Kung, William Yu Chung Wang, and Casey G. Cegielski. "An integrated big data analytics-enabled transformation model: Application to health care." *Information & Management* 55, no. 1 (2018): 64-79.