International Journal of Engineering Research and General Science Volume 3, Issue 3, Part-2 , May-June, 2015
ISSN 2091-2730

Integration of Big Data in Cloud computing environments for enhanced data processing capabilities

Rohit Chandrashekar[1] Maya Kala [2] Dashrath Mane [3]

VES Institute of Technology,

Chembur, Mumbai

[1] rohit28chandrashekar@gmail.com **[2]** maya11kala@gmail.com

[3] dashumane@gmail.com

Abstract— Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Big data refers to not only the volume of the data but also the technology used to and processes used to analyze such huge volumes of data into usable information which cannot be performed using traditional database and software technologies.

Cloud computing is an extremely successful paradigm of service oriented computing, and has revolutionized the way computing infrastructure is abstracted and used. Cloud computing eliminates the need to maintain expensive computing hardware, dedicated space, and software.

This paper proposes on how Big Data can be integrated with the elasticity, of cloud computing environment to bring about efficient and cheaper information processing solutions.

Keywords— Big Data, Cloud Computing, Enhanced Data Processing

INTRODUCTION

Cloud Computing is a technology which depends on sharing of computing resources than having local servers or personal devices to handle the applications. In Cloud Computing, the word “Cloud” means “The Internet”, so Cloud Computing means a type of computing in which services are delivered through the Internet. The goal of Cloud Computing is to make use of increasing computing power to execute millions of instructions per second. Cloud Computing uses networks of a large group of servers with specialized connections to distribute data processing among the servers. Instead of installing a software suite for each computer, this technology requires to install a single software in each computer that allows users to log into a Web-based service and which also hosts all the programs required by the user.

There's a significant workload shift, in a cloud computing system. Local computers no longer have to take the entire burden when it comes to running applications. Cloud computing technology is being used to minimize the usage cost of computing resources [4]. The cloud network, consisting of a network of computers, handles the load instead. The cost of software and hardware on the user end decreases. The only thing that must be done at the user's end is to run the cloud interface software to connect to the cloud. Cloud Computing consists of a front end and back end. The front end includes the user's computer and software required to access the cloud network. Back end consists of various computers, servers and database systems that create the cloud. The user can access applications in the cloud network from anywhere by connecting to the cloud using the Internet. Some of the real time applications which use Cloud Computing are Gmail, Google Calendar, Google Docs and Dropbox etc.

Figure1. Benefits of Cloud Computing

Big Data focuses on achieving deep business value from deployment of advanced analytics and trustworthy data at Internet scales. Big Data is at the heart of many cloud services deployments. As private and public cloud deployments become more prevalent, it will be critical for end-user organizations to have a clear understanding of Big Data application requirements, tool capabilities, and best practices for implementation.

Big Data is said to have the following properties:

Volume: Many factors contribute towards increasing Volume streaming data and data collected from sensors etc.,

Variety: Today data comes in all types of formats

Emails, video, audio, transactions etc.,

Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.

Variability: Along with the Velocity, the data flows can be highly inconsistent with periodic peaks.

Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Big Data and cloud computing are complementary technological paradigms with a core focus on scalability, agility, and on-demand availability. Big Data is an approach for maximizing the linear scalability, deployment and execution flexibility, and cost-effectiveness of analytic data platforms. It relies on such underlying approaches as massively parallel processing, in-database execution, storage optimization, data virtualization, and mixed-workload management. Cloud computing complements Big Data by enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Big data and cloud computing are both the fastest-moving technologies emerging today. Cloud computing is associated with new paradigm for the provision of computing infrastructure and big data processing method for all kinds of resources. Moreover, some new cloud-based technologies have to be adopted because dealing with big data for concurrent processing is difficult. The current

technologies such as grid and cloud computing have all intended to access large amounts of computing power by aggregating resources and offering a single system view. Among these technologies, cloud computing is becoming a powerful architecture to perform large-scale and complex computing, and has revolutionized the way that computing infrastructure is abstracted and used. In addition, an important aim of these technologies is to deliver computing as a solution for tackling big data, such as large scale, multi-media and high dimensional data sets.

Why Big Data Analytics in Cloud computing:

Cost reduction: Cloud computing offers a cost-effective way to support big data technologies and the advanced analytics applications that can drive business value. Enterprises are looking to unlock data's hidden potential and deliver competitive advantage. Big data environments require clusters of servers to support the tools that process the large volumes, high velocity, and varied formats of big data. IT organizations should look to cloud computing as the structure to save costs with the cloud's pay-per-use model.

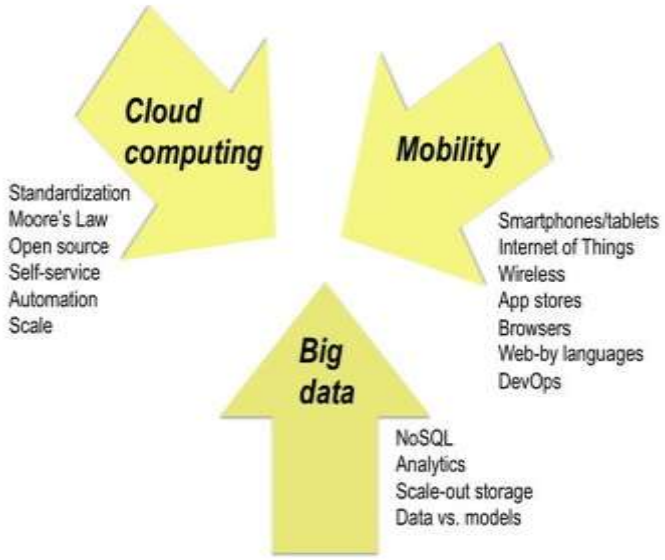

Figure2. Integration of Big Data and Cloud Computing

Reduce overhead: Various components and integration are required for any big data solution implementation. With cloud computing, these components can be automated, reducing complexity and improving the IT team's productivity.

Rapid provisioning/time to market: Provisioning servers in the cloud is as easy as buying something on the Internet. Big data environments can be scaled up or down easily based on the processing requirements. Faster provisioning is important for big data applications because the value of data reduces quickly as time goes by. .

Flexibility/scalability: Big data analysis, especially in the life sciences industry, requires huge compute power for a brief amount of time. For this type of analysis, servers need to be provisioned in minutes. This kind of scalability and flexibility can be achieved in the cloud, replacing huge investments on super computers with simply paying for the computing on an hourly basis.

SYSTEM IMPLEMENTATION

The aim of this paper is to provide a practical reference to help enterprise information technology (IT) and business decision makers of the Finance and Hospital industry as they analyze and consider the implications of big data and cloud computing on their business. The paper includes guidance and strategies, designed to help these decision makers evaluate different requirements from various actors including medical practices, hospitals, capital markets, and finance industry.

Capital Markets:

There are countless opportunities for financial services firms to leverage the benefits of cloud computing by migrating a variety of applications to the cloud. Non-core applications and such business processes as recruiting, billing and organization-wide travel management can—and should—easily move to the cloud. A number of infrastructure operations, such as data center management, data storage and disaster recovery, should also move to a cloud after a thorough evaluation of different vendors offerings and based on the flexibility of cloud vendors in documenting contracts. Although very few firms are currently using cloud computing for their core applications, different hosting architectures provided by IaaS (Infrastructure as a Service) cloud providers and new avenues in the community and hybrid cloud space, will drive more firms to move their core applications to the cloud. In fact, core solutions, such as batch processes running throughout the day, analytics and reporting applications, are perfect candidates.

A few scenarios that would be ideal for a cloud deployment include:

Risk analytics calculation:

Applications that calculate such analytics as cost of trade, current value, yields, Greeks, etc., at the level of a single security, position or portfolio are perfect candidates for a grid-based cloud. A cloud-based grid service can easily scale up or scale down depending on the data load. What's more, the applications can be seamlessly deployed on multiple grid nodes, reducing maintenance overhead. Also, since such applications only run for specific durations, dedicated hardware leads to unutilized CPU cycles, which can be optimized by a grid-based cloud. The whole solution can be implemented on a private cloud where existing computing power can be virtualized and made available as an on- demand service.

Performance attribution:

Performance attribution provides a framework for examining the relative performance of a fund versus its benchmark. It is a methodology that quantifies the success or added value of an investment strategy. Attribution allows investment managers to identify the factors of the investment process that contributed (positively or negatively) to the performance levels highlighted by performance measurement. Hence, these data- intensive processes need access to a huge amount of historical data for correctly calculating metrics. Performance attribution or benchmark rebalancing applications run at specific times of a day, like the analytics calculation processes. As such, these are ideal candidates to be deployed on a cloud, able to optimize the usage of available computing power and the scale-in and scale-out benefits of an existing grid.

Trade matching and reconciliation:

A trade matching process gets trade data from multiple brokers and counterparties and then reconciles it. This process is prone to high volumes during times of peak trading. The solution is to create a hybrid cloud where the reconciliation process can run on a public cloud for scalability and the data can reside on dedicated database servers in a private cloud. The data from multiple brokers and counterparties can be pushed to the public cloud, which can then be streamed to the private cloud. This can also help avoid creating separate connectivity to new partners and maintaining all those connections simultaneously.

Reference data virtualization:

Various types of reference data, such as security master data, positions data, holdings and book data, broker and counterparty data, etc., reside in multiple kinds of data sources. These data sources can be internal databases, file systems or external feeds. When an application needs to access data from many sources, it can be a challenge to devise strategies that connect those data sources and consolidate and aggregate the data within the application for specific needs. The recommended solution is to build a data virtualization layer that seamlessly federates these different data sources and provides different ways to access the single virtual data source. The layer should be flexible enough to mash up different streams of data according to the requirements of a particular application. Similar to the reference data virtualization layer, a transactional or operational data virtualization layer can be created to support risk management, financial analysis and compliance reporting. The goal is to make all data available through centralized data services.

Hospital Industry:

"Patient centricity" has become the key trend in healthcare provisioning and is leading to the steady growth in adoption of electronic medical records (EMR), electronic health records (EHR), personal health records (PHR), and technologies related to integrated care, patient safety, point-of-care access to demographic and clinical information, and clinical decision support. Availability of data, irrespective of the location of the patient and the clinician, has become the key to both patient satisfaction and improved clinical outcomes. Cloud technologies can significantly facilitate this trend.

Some areas where cloud computing can be implemented:

Radiologists: Using Cloud computing, radiology users can efficiently manage multimodality imaging units by using the latest software and hardware without paying huge upfront costs.

Neuroscience: Making brain mapping toolkits more accessible to non-specialist users in virtue of concealing its implementation context as well as rendering local IT infrastructure unnecessary.

Physical therapy: Using Cloud computing, effectiveness of computer-assisted learning (CAL) in physical therapy can be achieved to provide rehabilitation and recovery at a faster rate.

The key stakeholders for the Big Data Analytics Platform at a Hospital or a Medical Research Centre are Physicians / Clinicians, Researchers, Operations and Quality teams – each with their perspective and view of "tangible, measurable value" derived by implementing the platform. The platform leverages some of the finest open source Big Data technologies that include Data Mining / Machine Learning, Advanced Visualisation techniques to help address the opportunities to improve delivery of patient care at the Hospital / Medical Research Centre.

Benefits of implementing a Big Data and Advanced Cloud-based computing system:

Safety and cost effectiveness: Cloud-computing systems require very little in terms of infrastructure, and patients only need access to a Web browser to take part. The essence of cloud computing is a secure, cost-effective way to manage patient information and communication.

More accurate patient information: Cloud-based systems give patients the opportunity to complete their medical history at their convenience, which can eliminate problems that occur from information gathered over the phone.

Ease of use: When selecting a cloud-computing system, choose one that will provide patients with an intuitive interface. If patients can easily work through the system, their information will be more complete and satisfaction greater.

Prediction of 30 Day Readmission Candidates: The opportunity is to identify / predict patients at high risks by the system and thereby enable focused, directed attention to deliver care that would prevent readmissions. An accurate prediction of readmission ensures better, targeted care and interventions to the right patients at the hospital / medical research centre.

Chronic Disease Management: Real-time monitoring of patient data and its map onto advanced algorithm driven predictive models against chronic diseases and patient conditions enables generation of real-time alerts to both caregivers and patients of a deviation from the expected pathway.

Home Health Monitoring: Home health data such as weight, blood pressure, heart rate, respirations, temperature, glucose and other parameters are received, stored, analysed and then pushed to the EHR as discrete data fields. The real-time metrics generated are published as a patient scorecard and predictions around anomalies in patient condition / behaviour is immediately flagged off for relevant intervention to prevent fatalities.

Early Sepsis Detection: Severe sepsis (acute organ dysfunction secondary to infection) and septic shock are major healthcare problems, affecting millions of individuals around the world each year, killing one in four (and often more), and increasing in incidence. The ability to monitor patients in real time to determine the earliest entry point to the sepsis pathway will ensure timely treatment leading to a reduction in mortality and morbidity rates.

Clinical Research: Many pharmacology vendors are starting to tap the cloud to improve research and drug development. The 'explosion of data' from next generation sequencing as well as the growing importance of biologics in the research process is making cloud-based computing "an increasingly important aspect of R&D. Commercial cloud vendors have developed pharma-specific clinical research cloud offerings with the goal of lowering the cost and development of new drugs.

CONCLUSION

This paper described a systematic flow of survey on the big data processing in the context of cloud computing. Cloud computing provides enterprises cost-effective, flexible access to big data's enormous magnitudes of information. Big data on the cloud generates vast amounts of on-demand computing resources that comprehend best practice analytics. Both technologies will continue to evolve and congregate in the future.

REFERENCES:

[1] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li "Big Data Processing in Cloud Computing Environments" 2012 International Symposium on Pervasive Systems, Algorithms and Networks.

[2] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri "Security Issues Associated With Big Data In Cloud Computing" International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.

[3] "Deploying Big Data Analytics Applications to the Cloud: Roadmap for Success" 2014 Cloud Standards Customer Council.

[4] "Use of Big Data Technologies in Capital Markets" by Ruchi Verma, Sathyan R Mani INFOSYS.

[5] "Big Data in the Cloud: Converging Technologies" by Intel IT Center.

[6] "Impact of Cloud Computing on Healthcare" 2012 Cloud Standards Customer Council