A REVIEW ON INTERSECTION OF CLOUD WITH BIG DATA AND SAP HANA

[1] *MR.D.S.BARAVDE,* [2] *MISS. H.A.PATIL*

[1, 2] **Department of Computer Engineering, Ashokrao Mane Polytechnic, Vathar, Kolhapur, Maharashtra, India.**

ABSTRACT: *Cloud computing, big Data and HANA is a recent technology which allow users to access not only the files shared on internet but also the various types of services such as emails and many others applications at anytime and from anywhere in the world. Clouds are also being used to deal with the Big data to effectively store and exploit the unstructured data of the organizations. This paper presents an overview of the cloud computing big data and HANA today. It provides, how cloud is related with big data. Also some of the future trends of the cloud computing are being discussed in this paper. This paper examine the intersection of biggest treads in the storage today are big data and cloud computing with SAP HANA. In this article, we'll also look at some of the developments in big data public cloud offerings, however, let's look at how big data on the public cloud got started, Also some of the future trends of the cloud computing and Big Data and SAP HANA are being discussed in this paper.*

KEY WORDS: ***Big Data, Cloud, Hadoop; SAP HANA, Customer service provider; Customer service consumer***

1. INTRODUCTION

Cloud computing and Big Data has set a major trend in the popularity and advancement of the technology now a day. This technology has not only empowered the IT industries with a powerful tool for the advancement of the technology and also their business, but it also provides a new topic of research in academia. A simple definition of cloud may state that "Cloud Computing is a model for enabling convenient, universal and on-demand network access to shared pool of configurable computing resources (server, network, storage, application and services) that can rapidly provisioned and released with minimum management effort or service provider interaction"[1]. The idea of cloud computing is that every type of computation can be delivered to public via internet. It is changing the scenario and also affects the daily life of an individual. Any stuff can be shared across any device by users via cloud computing without any problem. Network bandwidth, software, processing power and storage are represented as the computing resources to users as the publicly accessible utility services. The different delivery models of the cloud computing are used for the different types of services to be delivered to the end user. This model provides the software, platform and the hardware as needed by the Cloud service customer (CSC).

Defining the Cloud

When defining the cloud and big data, it's helpful to consider both the consumer and producer perspectives. For consumers, the cloud is about consuming hardware or software as a service (SaaS) and the various implications of this approach. For example, pricing models and data governance may change significantly. In public clouds, the services are run by a third party, while in private clouds, they are owner-operated on premise. Consumers effectively choose the level of vertical integration for their IT; they can choose to own or outsource everything from the data center to the storage, computing, networking, and software infrastructure up to the application. For producers, on the other hand, the cloud is about the technology that goes into providing service offerings at each level [1]. The technology required to provide an application as a service in the public cloud may differ significantly from the software product that a customer installs to run an internal service. For example, virtual machines are the resource allocation units in most cloud infrastructure offerings, but they might not be used when implementing an application as a public service.

Defining Big Data

For consumers, big data is about using large datasets from new or diverse sources to provide meaningful and actionable information about how the world works. For example, Netflix can use customer data to produce shows tailored to their audiences. For producers, however, big data is about the technology necessary to handle these large, varied datasets. Producers characterize big data in terms of volume, variety, and velocity. How much data is there, of what types, and how quickly can you derive value from it? Although these are good technical descriptions of big data, they don't fully explain it. Just as adopting a service-oriented approach is the macro trend behind the cloud, there are several macro trends behind big data. The first trend is consumption; we consume data as part of the everyday activities in our personal and working lives. From booking a flight, to finding a partner, to diagnosing disease, data is driving many more decisions today than it has in the past. We

live in a relatively new social context where people increasingly want to make data-driven decisions. Related to consumption, the second trend is instrumentation. We collect data at each step in many of our activities, and much of it is now produced by machines instead of people. From supply chains to Fit bits, we collect information about all our activities with the intent to measure and analyze them. The third trend is exploration. The relatively easy access to this abundance of data means we can use it to construct, test, and consume experiments that were previously not feasible. Finally, related exploration is the concept that the data itself has value. Data is increasingly an asset, not just input to or a byproduct of a business process. This isn't a new idea of course, but in the context of consumption, instrumentation, and exploration, it's driving new business models and applications. Ultimately, big data is about the change in relationship between us and our data and, in the context of this column, the implications of this change on cloud technology.

Defining SAP HANA

SAP HANA (high-performance analytic appliance) is an application that uses in-memory database technology that allows the processing of massive amounts of real-time data in a short time. The in-memory computing engine allows HANA to process data stored in RAM as opposed to reading it from a disk. This allows the application to provide instantaneous results from customer transactions and data analyses. SAP HANA is an in-memory data platform that is deployable as an on-premise appliance, or in the cloud. It is a revolutionary platform that's best suited for performing real-time analytics, and developing and deploying real-time applications. At the core of this real-time data platform is the SAP HANA database which is fundamentally different than any other database engine in the market today. Whenever companies have to go deep within their data sets to ask complex and interactive questions, and have to go broad (which means working with enormous data sets that are of different types and from different sources) at the same time, SAP HANA is well-suited. Increasingly there is a need for this data to be recent and preferably in real-time. Add to that the need for high speed (very fast response time and true interactivity), and the need to do all this without any pre-fabrication (no data preparation, no pre-aggregates, no-tuning) and you have a unique combination of requirements that only SAP HANA can address effectively. When this set of needs or any subset thereof have to be addressed (in any combination), SAP HANA is in its elements.

2. EXAMPLES OF CLOUD COMPUTING

This section provides with some of the real life examples of cloud computing services.

a) Email: Web-based e-mails are biggest cloud computing services. Microsoft's Hotmail or Windows Live Mails are examples of cloud based email service. Using a cloud computing e-mail solution allows the mechanics of hosting an e-mail server and it is maintained by the people running the service. This means that we can access our e-mail from anywhere in the world.

b) Social Networking: The most famous example of cloud computing are the social networking websites like Twitter,Facebook, My space, LinkedIn and many others which doesn't seems to be a part of cloud computing at first glance. In social networking user finds people he already knows or like to know and shares information with them. As the user shares information with people related to him, he ultimately shares the information with peoples who are running the service. Social networking can also be used by business for its promotion among its customers.

c) Backup Services: Services like Jungle Disk, Carbonite, and Mozy allow public to automatically back up all their data to servers spread around the country or world for a surprisingly low cost.

d) Document/Spreadsheet/hosting Services: Google Docs allow users to keep and edit their documents online. Making use of Google Docs allows access and sharing of the documents from anywhere. The same document can be worked by multiple people simultaneously. Yahoo's Flickr, Google's Picasa and provides hosting for the photographs that individual wants to share with other people. Comments can be
placed on the photographs in the similar manner as on
Face book. But for the photographic enthusiast's some perks are provided by these photo hosting services.

3. BIG DATA AND CLOUD

Big Data is a term given to a collection of large number of data sets that it becomes difficult for the on-hand database management tools and traditional data processing applications to process such a large data. Search, sharing, storage, capture, transfer, visualization and analysis are the challenges of the Big data. The development of the large data set has been arising due the extraction of the extra practical information from the analysis of the single large datasets of related datasets. Also this data is unstructured. When we talk about Big Data, it doesn't only refers to data which is already gets collected from the large number of years but it also refers to the data from the social media, sensors, mobile device and many other technologies. The large sets of unstructured data such as production, financial transactions and weather data required big data analysis to bring order and get rid of light on trends, patterns and relationships, made visible only by structured and systematic analysis. It is difficult to work on big data with most of the relational database management software and other traditional tools. Instead an enormous parallel running software tools are required to run on tens, hundreds or even thousands of servers. The Big data varies depend on the capabilities of an organization to manage the

unstructured datasets and also on the capabilities of managing the different traditional database management software and tools.

There is no doubt in the fact that Big data has co-occurred with the rapid adoption of the Platform-as-a Service (PaaS) and Infrastructure-as-a Service (IaaS) technologies. IaaS allows the rapid deployment of the computation nodes while PaaS [2] lets firms scale their capacity on demand and reduce costs. The Big Data processing for enterprises of all sizes are empowered by

the cloud by relieving a number of problems, but there are still complexities in the elicitation of the business data from the big unstructured business data available.

Cloud computing [4] democratizes big data – any enterprise can now work with unstructured data at a huge scale. Cloud computing by bringing the big data analyses to the masses has provided businesses with affordable and flexible to vast amounts of computing resources on demand. Although cloud reduces the overall production cost and provides flexibility it is not suitable with all big data cases. But now it is possible to manage the big data whether structured or unstructured by a tool named Hadoop Linode is a favorite cloud for Linux users. SalesForce.com [10]- It showed the world that software can be bought as a service and also it has one of the most popular clouds for running their own home-grown applications: Heroku. Citrix is company building software for clouds. Google App [5] Engine is another spot where developers can park their apps; Google cloud storage and Google Drive are other cloud features by google for the storage and sharing of files. Microsoft Azure [11] is the cloud for the millions of developers who already write for Microsoft's platform. Rackspace [5] is a cloud provider. Rackspace's OpenStack is to cloud computing what Google's Android is to mobile device makers. Amazon- It is a leading player in the cloud computing Field.

Converging Technologies

So what is the relationship between big data and the cloud? Big data has its origins in the cloud. Apache Hadoop, one of the most widely used big data technologies today, was built on research from Google and initially deployed at Yahoo. Google invented this technology because indexing the web was infeasible with existing systems. Now companies adopting Hadoop are bringing cloud architecture into their data centers. The simultaneous rise of cloud and big data technologies isn't coincidental—they're mutually reinforcing. Big data enables the cloud services we consume. For example, SaaS lets us collect data that was infeasible or impossible in a world of packaged software. An application can record every interaction from millions of users. This service in turn drives demand for big data technologies to store, process, and analyze these interactions and inject the value of the analysis back into the application through query and visualization. The expansion of the cloud continues to drive both the creation of new big data technologies and big data adoption by making it easier and cheaper to access storage and computing resources. Companies can run their big data plat- forms on infrastructure provided as a service (IaaS) or consume the big data platform as a service (PaaS). Both models work in the public cloud and in on- premise systems. The decision for enterprises is thus a familiar one: How vertically or horizontally integrated should your infrastructure be? A spectrum of valid options exists, but cloud technology is already enabling more infrastructure outsourcing, whether it's outsourced to a cloud provider or an internal centralized IT department. Big data infrastructures also play a role in this trend. For example, recent advances in the Apache Hadoop eco- system enable more types of workloads and more tenants to share a cluster. What were once discrete systems running on their own hardware are now effectively applications running on Hadoop, sharing the same data and hard- ware resources? As this abstraction layer evolves and more projects build on it, users will be able to run more types of infrastructures on the same Hadoop cluster, which itself may be running on a cloud infrastructure. As big data infrastructures become more generic, the cloud infrastructure will add more specialized services for data storage, processing, and analysis. Future columns will examine new developments in both areas and the increasing overlap between them. Another area of exploration for this column will be technologies and trends that are leveraging both cloud computing and big data. The combination of big data, cloud computing, and new algorithms and techniques for visualizing information enables converged analytics—performing analytics on data from many different sources. These new techniques for data delivery and data management also enable cloud-based analytics as a service (AaaS). Upcoming columns will cover the development and use of converged analytics and AaaS.From security and privacy to pricing models, the combination of big data and cloud computing is having a substantial impact on the nontechnical aspects of our lives as well. There is a tension between our desire for converged analytics and cloud computing—which is about sharing more computing resources and data with increasingly diverse tenants— and our desire for better privacy controls and data protection. Usage-based pricing models are forcing us to rethink how we produce and consume technology. Future columns will look at how policies and economics are being shaped by these technological advances.

4. APACHE HADOOP Apache hadoop supports data-intensive distributed applications. Apache Hadoop is an open source software framework which is licensed under Apache v2 license. It provides the data motion and reliability to the applications. The related projects on Apache Hadoop platform are: Apache Hive, Apache HBase and others. Hadoop was derived from the

Google file system and Google's MapReduce [7] papers. A computational paradigm named MapReduce [8] is implemented by Hadoop where application is divided into number of small fragments of work. Each of the fragments thus executed on any node in a cluster. Hadoop provides the distributed file system [9] which stores data on the computation nodes. This distributed file system provides very high aggregate bandwidth across the cluster. To handle the failure of the nodes automatically, both the distributed file system and the Map Reduce are designed. This empowers the application to work on the thousands of computation-independent computers and peta bytes of data. Hadoop is written in Java programming language and is a top level project of apache which is being built and used by global community of contributors.

5. HANA and Hadoop

HANA and Hadoop are very good friends. HANA is a great place to store high-value, often used data, and Hadoop is a great place to persist information for archival and retrieval in new ways - especially information which you don't want to structure in advance, like web logs or other large information sources. Holding this stuff in an in-memory database has relatively little value.

As of HANA SP06 you can connect HANA into Hadoop and run batch jobs in Hadoop to load more information into HANA, which you can then perform super-fast aggregations on within HANA. This is a very co-operative existance.However; Hadoop is capable - in theory of handling analytic queries. If you look at documentation from Hadoop distributions like Hortonworks or Cloudera, they suggest that this isn't the primary purpose of Hadoop, but it's clear that Hadoop is headed in this direction. Paradoxically, as Hadoop heads in this direction, Hadoop has evolved to contain structured tables using Hive or Impala. And with ORC and Parquet file formats within the HDFS file system, Hadoop also uses columnar storage. So in some sense Hadoop and HANA are converging.

6. TRENDS OF CLOUD COMPUTING FUTURE

The name cloud computing sounds so new that the users unfamiliar with it think about it as a new technology in computing and data storage. But in real cloud computing has been in use for several years as it can be seen in the file sharing programs which is hosted by third party. As the computers today are connected by the computing resources hosted over the internet, it has improved the scenario of cloud computing. By this we can access the services over internet as long as the device is connected with the internet. Google docs are a very good example of this computing. It allows users with access permission to edit and manipulate the file from anywhere. Cloud computing on the entertainment side is also seen in an online game where users login, play game, save the game and log out. Whenever user returns back he can start his game on cloud from where he left it. In this way cloud computing is continuously developing and improving to meet the needs of different sectors. In the coming future many new applications for mobile devices based on cloud may be employed for making the mobile devices equipped with much better functions and power to the devices. Also cloud computing will go a step forward to deal with the big data storage of the enterprises for data processing and management. The rise in the employment of the public and private cloud together to form a hybrid cloud may also be seen in the coming future. Also stricter protocols for the security of the data on cloud may come into existence due to increase in the use of clouds both in
the public and the private sector.

7. TRENDS OF SAP HANA FUTURE

Real-time analytics – The Categories of Analytics which HANA specializes

Operational Reporting (real-time insights from transaction systems such as custom or SAP ERP). This covers Sales Reporting (improving fulfillment rates and accelerating key sales processes), Financial Reporting (immediate insights across revenue, customers, accounts payable, etc.), Shipping ,Reporting(better enabling complete stock overview analysis), Purchasing Reporting (complete real-time analysis of complete order history) and Master Data Reporting (real-time ability to impact productivity and accuracy).

Data Warehousing (SAP Net Weaver BW on HANA) – BW customers can run their entire BW application on the SAP HANA platform leading to unprecedented BW performance (queries run 10-100 times faster; data loads 5-10 times faster; calculations run 5-10 times faster), a dramatically simplified IT landscape (leads to greater operational efficiency and reduced waste), and a business community able to make faster decisions. Moreover, not only is the BW investment of these customers preserved but also super-charged. Customers can migrate with ease to the SAP HANA database without impacting the BW application layer at all.

Predictive and Text analysis on Big Data - To succeed, companies must go beyond focusing on delivering the best product or service and uncover customer/employee /vendor/partner trends and insights, anticipate behavior and take proactive action. SAP HANA provides the ability to perform predictive and text analysis on large volumes of data in real-time. It does this through the power of its in-database predictive algorithms and its R integration capability. With its text search/analysis capabilities SAP HANA also provides a robust way to leverage unstructured data.

Core process accelerators – Accelerate business reporting by leveraging ERP Accelerators, which are non-disruptive ways to take advantage of in-memory technology. These solutions involve an SAP HANA database sitting next to a customer's SAP ERP system. Transactional data is replicated in real-time from ECC into

HANA for immediate reporting, and then results can even be fed back into ECC. Solutions include CO-PA Accelerator, Finance and Controlling Accelerator, Customer Segmentation Accelerator, Sales Pipeline Analysis, and more.

Planning, Optimization Apps – SAP HANA excels at applications that require complex scheduling with fast results, and SAP is delivering solutions that no other vendor can match. These include Sales & Operational Planning, Business Objects Planning & Consolidation, Cash Forecasting, ATP calculation, Margin calculation, Manufacturing scheduling optimization (from start-up Optessa), and more.

Sense & response apps – These applications offer real-time insights on Big Data such as smart meter data, point-of-sale data, social media data, and more. They involve complexities such as personalized insight and recommendations, text search and mining, and predictive analytics. Only SAP HANA is well suited for such applications, including Smart Meter Analytics, SAP Supplier Info Net, SAP precision retailing, and Geo-spatial Visualization apps (from start-up Space-Time Insight). Typically these processes tend to be data-intensive and many could not be deployed in the past owing to cost and performance constraints.

8. CONCLUSION

Cloud computing, Big Data, SAP HANA is a recent technology which is being used at large level by the infrastructure and services Industries focusing to capture potential opportunities. This paper gives the interaction of the cloud computing technology, big data and HANA.also we discuss SAP HANA which supports many front-end tools like Lumira, BOBJ, and Dashboard etc which are used for Analytics. On the other hand, Hadoop is the core technology for big data Analytics. Many of the big industries has already implemented Hadoop and using it for various analytic purposes. This paper explores some of the future trends and application of cloud computing, big data and HANA which may lead to the enhancement and development of the technology in near future.

9. REFERENCES

[1] Wayne Jansen, Timothy Grance, “Guidelines for security and privacy in pubic cloud”, NIST Special Publication 800-144, csrc.nist.gov, 2011

[2] I. Foster, Y. Zhao, S. Lu, “Cloud computing and Grid computing 360-degree compared”, Proc. Grid Coputing Environments Workshop (GCE’08), 2008.

[3] Qi Zhang · Lu Cheng · Raouf Boutaba, “Cloud computing: state of-the-art and research challenges”, Springer, 2010.

[4] Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I, “Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility”, Future generation computer systems, Vol. 25, Issue 6, pp. 599-616, Elsevier, 2009.

[5] Strauch S, Kopp O, Leymann F, Unger T, “A Taxonomy for Cloud Data Hosting Solutions”, Published in Dependable, Autonomic and Secure Computing (DASC), Sydney, pp. 577- 584, IEEE, 2011

[6] Prodan R, Sperk M, “Scientific computing with Google App Engine”, Future generation computer systems, Elsevier, 2013.

[7] Dean J, Ghemawat S, “MapReduce: simplified data processing on large clusters”, Communications of the ACM, Volume 51 Issue 1, Pages 107-113, New York USA, ACM, 2008.

[8] Bhandarkar M, “Mapreduce programming with Apache hadoop”, Parallel & Distributed Processing (IPDPS), page 1, ISSN. 1530- 2075, Atlanta, IEEE symposium, 2010.

[9] Shvachko K, Hairong K, Radia S, Chansler R, “The Hadoop distributed file system”, Mass Storage Systems and Technologies (MSST), pp. 1-10, IEEE symposium, 2010.

[10] Bibi S, Katsaros D, Bozanis P, “Business Application Acqusition:On-premise or SaaS based solutions, Published in software IEEE, Vol. 29, Issue 3, pp. 86-93, IEEE computer society 2012.

[11] Hill Z, Li J, Mao M, Alvarez AR, Humphrey M, “Early observations on the performance of Windows Azure”, International Symposium on High Performance Distributed Computing, pp. 367-376, New York USA, ACM, 2010.

[12]Waldspurger CA, “Memory resource management in VMware ESX server”, ACM SIGOPS Operating Systems Review – OSDI '02, New York USA, Vol. 36, Issue SI, pp. 181-194, 2002.