Impact of Big data in Cloud Environment

1T.Jhansirani, 2D.Satyanarayana
1,2Assistant Professor,
1G.PullaReddy Engineeering College, 2Dr.K.V.SRIT, Kurnool.

***Abstract*— With the rapid growth of emerging applications like social network analysis, semantic Web analysis and bioinformatics network analysis, a variety of data to be processed continues to witness a quick increase. Effective management and analysis of large-scale data poses an interesting but critical challenge. Currently big data is a solution to this challenge which deals with the data that have high Volume, high Velocity, high Variety and high Veracity also known as 4V's of Big Data which includes data includes emails, system logs, internal documents, business process events, and other structured, unstructured, and semi-structured data.**

***Keywords*— Volume, high Velocity, system logs, internal documents, business process events,, unstructured, and semi-structured data.**

1 INTRODUCTION

With the rapid growth of emerging applications like social network analysis, semantic Web analysis and bioinformatics network analysis, a variety of data to be processed continues to witness a quick increase. Effective management and analysis of large-scale data poses an interesting but critical challenge. Currently big data is a solution to this challenge which deals with the data that have high Volume, high Velocity, high Variety and high Veracity also known as 4V's of Big Data which includes data includes emails, system logs, internal documents, business process events, and other structured, unstructured, and semi-structured data.

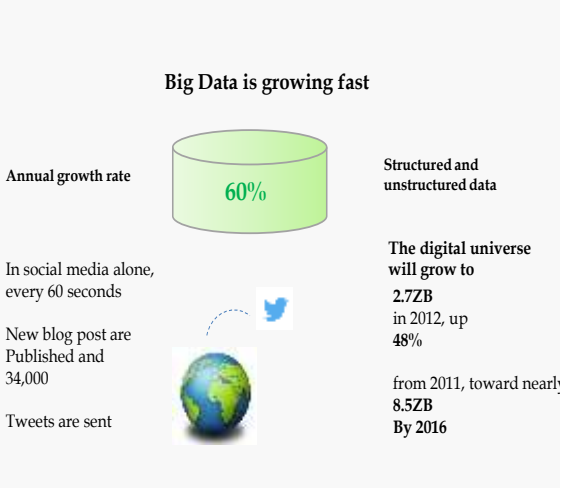

Fig: The explosion of bigdata

Examples of big data and analytics

Due to a combination of automation, consumer involvement, and market-based exchanges, big data is becoming readily available and is being deployed in a series of significant use cases that are radically disrupting traditional markets.

Here are some examples of big data: *Social media content*: A wide variety of data, including unstructured data, text, and media content, can be found at various social media websites. This data contains valuable information that can be mined by an enterprise.

Cell phone details: Currently, there are over 5 billion cell phones that can each provide useful information, such as the user's location, how the device is being used, and any functional problems that might be present in the device.

Channel-click information from set-top boxes: Users' interactions with their television set-top boxes provide valuable information about the kinds of programs and topics that interest them.

Transactional data: Today's enormous number of online transactions, each facilitated by various sources, such as mobile wallets and credit cards, are creating terabytes of data that can be mined and analysed.

Documentation: Documentation, such as financial statements, insurance forms, medical records, and customer correspondence, can be parsed to extract valuable information for further analysis.

Internet of Things: The *Internet of Things* is generating large volumes and various data from sources as varied as eBooks, vehicles, video games, television set-top boxes, and household appliances. Capturing, correlating, and analysing this data can produce valuable insights for a company.

Communications network events: Communications networks are increasingly interconnected, resulting in the need to monitor large volumes of data and respond quickly to changes. For example, IP Multimedia Subsystem networks require the

monitoring and dynamic configuration of various devices in the core and access networks to ensure real-time traffic routing while maintaining quality of service (QoS) levels.

Call Detail Records: Analysis of Call Detail Records (CDRs) enables a company to better understand the habits of its customers and, potentially, of its customers' social networks.

Radio Frequency Identification (RFID) tags: RFID tags are increasingly ubiquitous and the valuable data they contain is often ignored and not analysed due to the sheer volume and variety of data obtained from these sensors.

Traffic patterns: Today, traffic patterns can be studied based on data from on-road sensors, video cameras, and *floating-car data* (a method of determining current traffic-flow speed based on data from motorists' mobile phones). Rapid analysis of this data can be used to relieve traffic congestion.

Weather information: Weather data is now being correlated to various other large sources of data, such as sales, marketing, and product information, enabling companies to market their products more effectively and cut costs.

Big Data and Analytics platform

All big data use cases require an integrated set of technologies to fully address the business pain they aim to alleviate. Due to this complexity, enterprises need to start small, with a single project, before moving on to other issues and pursuing added value. IBM is unique in having developed an enterprise-class big data platform that allows you to address the full spectrum of related business challenges. The IBM Big Data and Analytics platform gives organizations a solution stack that is designed specifically for enterprise use. The IBM Big Data and Analytics platform provides the ability to start small with one capability and easily add others over your big data journey because the pre-integration of its components reduces your implementation time and cost. The IBM Big Data and Analytics platform addresses the following key enterprise requirements:

- **The *5 Vs***: The platform includes functionality that is designed to help with each of the 5 Vs:

Variety: The platform supports wide variety of data and enables enterprises to manage this data *as is*, in its original format, and with extensive transformation tools to convert it to other desired formats.

Velocity: The platform can handle data at any velocity, either low-latency streams, such as sensor or stock data, or large volumes of batch data.

Volume: The platform can handle huge volumes of at-rest or streaming data.

Veracity: The platform includes various tools to remove uncertainty about the target data.

Visibility: The platform provides the ability to discover, navigate, and search over a broad range of data sources and types, both inside and outside your enterprise.

- Analytics:
  - The platform provides the ability to analyse data in its native format, such as text, binary, and rich multimedia content.

– The platform can scale to analyse *all* of your data, not just a subset.

– The platform enables *dynamic analytics*, such as automatic adjustments and actions.

For example, streaming video service companies use users' past viewing behaviour to generate new recommendations, and use this recommendation data in real time to provision greater capacity for improved viewing.

- Ease of use:– The platform includes a deep set of developer user interfaces (UIs), common languages, and management consoles. This Eclipse-based development environment enables faster adoption and reduces the time spent in coding and debugging.

– The platform also provides user UIs and visualization capabilities, such as web-based analysis and visualization tools with familiar, spreadsheet-like interfaces._ Enterprise-ready

The platform has capabilities for fault tolerance across the solution stack, including enterprise-grade security and privacy features.

- Integration

The platform provides the ability to integrate with a wide variety of data sources using industry-standard protocols, such as Open Database Connectivity (ODBC), Java Database Connectivity (JDBC), and Java Message Service (JMS).

Figure 1-2 shows an overview of the IBM Big Data and Analytics platform.

Figure 1-2 IBM Big Data and Analytics platform

As depicted in Figure 1-2, the IBM Big Data and Analytics platform uses the underlying big data infrastructure, which is typically either x86 or Power servers, for running the Hadoop system and Streams components, and data warehousing appliances.

The Hadoop system provides a cost-effective way to store large volumes of structured and unstructured data in one place for deep analysis. IBM provides a non-forked, open source Hadoop version and augments it with capabilities, such as enterprise-class storage (by using an IBM General Parallel File System (GPFS™)), security (by reducing the surface area and securing access to administrative interfaces and key Hadoop services), and workload optimization (using the Adaptive Map Reduce algorithm that optimizes execution time of multiple small and large jobs).

The Stream computing ability is designed to analyse data in motion while providing massive scalability and processing of multiple concurrent input streams. The IBM Streams platform can process and analyse a wide variety of structured and unstructured data and video and audio content.

The Data Warehousing component is provided by IBM workload-optimized systems that are delivered as deep analytical appliances, with a massive parallel processing engine and the ability to handle mixed operational and analytic workloads.

The Information Integration and Governance layer gives the IBM Big Data and Analytics platform the ability to integrate with any type of data. It also provides governance and trust for big data by using capabilities, such as security sensitive data, tracking data lineage, lifecycle management to control big data growth, and master data to establish a single source of truth.

The User Interfaces in the IBMBig Data and Analytics platform are tailored for three classes of users (business users, developers, and administrators), with different types of tooling for each class. Business users can analyse a wide variety of data in an ad hoc manner using browser-based interface and spreadsheet-style interface for exploring and visualizing data.
Developers have access to various APIs and useful development environments, such as Eclipse. Developers also have access to many data-parallel algorithms, such as those algorithms for regression modelling and dimensionality modelling.

Administrative users have access to consoles to aid in monitoring and managing the systems and components of the IBM Big Data and Analytics platform.

The IBM Big Data and Analytics platform provides a number of accelerators, such as Analytics accelerators (to handle text data, mining data, and acoustic data) and Industry and Horizontal Application accelerators, such as pre-configured analytics for processing CDRs for telecom clients, and streaming options trading for financial clients.

Finally, the IBM Big Data and Analytics platform is designed for analytic application development and integration with a wide variety of third-party applications for business intelligence, predictive analytics, content analytics, and so on.

The IBM products associated with the IBM Big Data and Analytics platform and maps each product to its role in the process and the underlying big data characteristics that is being addressed
Volume:
- PureData™ System for Hadoop
- PureData System for Analytics
- PureData System for Operational Analytics
- InfoSphere Warehouse

Velocity:
- InfoSphere Streams

Variety:
InfoSphere® BigInsights
Visibility:
InfoSphere Data Explorer
InfoSphere Data Architect
Veracity:
InfoSphere Information Analyzer
InfoSphere Information Server
InfoSphere Master Data Management Family

IBM Cloud Computing Reference Architecture

The IBM Cloud Computing Reference Architecture (CCRA) is a blueprint or guide for architecting cloud computing implementations. It is based on years of experience of IBM personnel working with customers on cloud computing solutions, and is driven by broad setoff functional and non-functional requirements collected during those engagements.

The IBM CCRA (see Figure 1-4) provides guidelines and technical work products, such as service and deployment models, and defines numerous overarching adoption patterns. An *adoption pattern* embodies the architecture patterns that represent the ways that organizations typically implement cloud computing solutions.

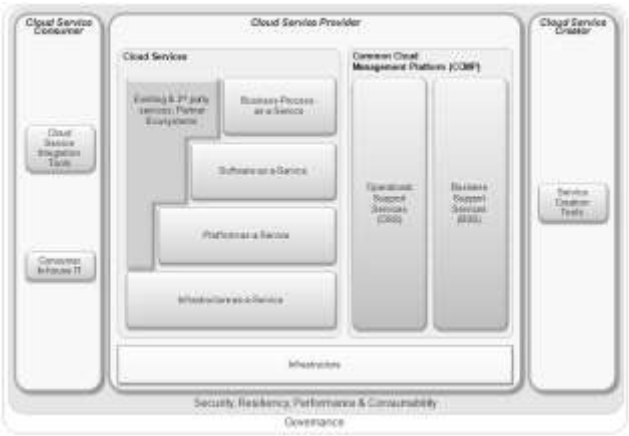

Figure 1-4 High-level view of the IBM Cloud Computing Reference Architecture (CCRA)

This paper summarizes a cloud computing environment as a data centre orchestrator that provides these functions:

_ Data storage (blocks or object storage)
_ Data processing (compute resources)
_ Data interchange (networking)

Cloud computing environments work on the principle of shared resource pools and exhibit these characteristics:

_ Ability to handle failures, such as by moving a workload off a failed VM._ Support for large workloads. Cloud computing environments are built to scale and new capacity can be provisioned quickly.
_ Programmable. Cloud computing environments typically provide APIs to connect and orchestrate all aspects of the workloads running in the cloud.
_ Utility computing, also called a *pay-as-you-go model*, with an illusion of infinite resources.

Cloud computing environments require no up-front cost and enable fine-grained billing (such as hourly billing).For this publication, the phrase *cloud computing* is used interchangeably with the phrase *Infrastructure as a service (IaaS)*.

Big data and analytics on the cloud: Complementary technologies

Big data and analytics require large amounts of data storage, processing, and interchange.

The traditional platforms for data analysis, such as data warehouses, cannot easily or inexpensively scale to meet big data demands. Furthermore, most of the data is unstructured and unsuitable for traditional relational databases and data warehouses.

Platforms to process big data require significant up-front investment. The methods for processing big data rely on parallel-processing models, such as Map Reduce, in which the processing workload is spread across many CPUs on commodity compute nodes. The data is partitioned between the compute nodes at run time, and the management framework handle sinter-machine communication and machine failures. The most famous embodiment of a Map Reduce cluster, Hadoop, was designed to run on many machines that don't share memory or disks (the *shared-nothing* model).Alternatively, cloud computing is the perfect vehicle to scale to accommodate such large volumes of data. Cloud computing can *divide* and *conquer* large volumes of data by using *partitioning* (storing data in more than one region or availability zone). Furthermore, cloud computing can provide cost efficiencies by using commodity compute nodes and network infrastructure, and requiring fewer administrators (thanks to standardizing the available offerings through the Cloud Service catalog), and programmers (through the use of well-defined APIs). However, cloud computing environments are built for general-purpose workloads and use resource pooling to provide elasticity on demand. So it seems that a cloud computing environment is well-suited for big data, provided the shared-nothing model can be honoured. But there is another large difference, the acute volatility of big data workloads compared to typical workloads in a cloud computing environment.

1.5 A large difference

Figure 1-5 shows two graphs. The graph on the left represents a typical cloud workload. The graph on the right shows a big data workload. The typical cloud uses a few resources over a long period compared to the big data workload that uses many resources over a shorter period.

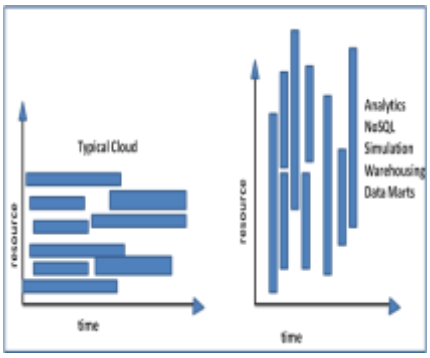

Figure 1-5 Typical cloud computing workloads versus big data workloads

Typical cloud computing workloads have volatility (certain workloads might be decommissioned after a time), but the duration of each workload is typically a few weeks to a few months. For example, communications providers often provision additional capacity to handle abnormally high workloads when new devices are launched. This new workload can require twice the provider's existing capacity, but the need might exist for just three to four weeks.

If you mentally rotate the figure counter clockwise by 90°, you get a representation of big data workloads. These workloads typically exist for less than a week (sometimes just a few hours), require massively large compute and storage capacity, and are decommissioned when the need is over. For example, an advertiser might analyze all micro-messages (such as Twitter messages) that are posted during a large public event to gauge public sentiment about its products and then change follow-up advertisements based on how the initial advertisements were received.

This differentiation has obvious implications. A big data cloud computing environment needs extreme elasticity to provision hundreds of virtual machines (VMs) in hours or minutes.

Dedicated and isolated networks are required to ensure that data replication between nodes does not affect the ingestion of incoming data.

Big data and analytics in the cloud: Bringing the two together

So, if the cloud computing environment can be modified correctly, big data and cloud can come together in beneficial ways:

_ The cloud engine can act as the orchestrator providing rapid elasticity.

_ Big data solutions can serve as storage back ends for the cloud image catalog and large-scale instance storage.

_ Big data solutions can be workloads running on the cloud.

Yet, for big data and the cloud to work together, many changes to the cloud are required:

_ CPUs for big data processing

A Graphics Processing Unit (GPU) is a highly parallel computing device originally designed for rendering graphics. GPUs have evolved to become general-purpose processors with hundreds of cores. They are considered more powerful than typical CPUs for executing arithmetic-intensive (versus memory-intensive) applications in which the same operations are carried out on many data elements in parallel fashion. Recent research has explored tightly integrating CPUs and GPUs on a single chip. So, one option is to create a resource pool with special computer chips for high performance computing for big data.

Another option of boosting computing capacity for big data in the cloud is to create a resource pool with multi-core CPUs, which can achieve greater performance (in terms of calculations per second) for each unit of electrical power that is consumed than their single-core equivalents. With quad-core and hex-core CPUs now commonplace, this is the most attractive and cost-effective way to create dedicated resource pools for big data processing in a cloud.

_ Networking for big data processing With a need to handle, potentially, petabytes of multi-structured data with unknown and complex relationships, the typical network design in a cloud infrastructure is no longer sufficient. Special considerations are required for ingesting the data into a Hadoop cluster, with a dedicated network to allow parallel processing algorithms, such as Map Reduce, to shuffle data between the compute nodes.

A minimum of the following types of network segments are required:

– Data: Dedicated to MapReduce applications with a bandwidth of 10 GB for lower latency and higher bandwidth

– Admin: A separate and dedicated network for management of all compute nodes and traffic not related to Map Reduce.

– Management: A platform for an Integrated Management Module (IMM) (can optionally share the VLAN subnet with the Admin segment)
_ Storage for big data processing
One of the biggest changes is to the storage subsystem. These changes can be addressed in two ways:
– Disk Attached Storage (DAS): The compute nodes are designed with multi-core commodity hardware with a large array of local disks. The local disks do not employ RAID and are used as just a box of disks (JBOD). In this case, built-in redundancy of big data file systems, such as Hadoop Distributed File System (HDFS), is used because they are replicating blocks across multiple nodes.
– A second option is to use a new type of storage architecture that allows storing and accessing data as objects instead of files. Rather than using traditional enterprise storage (such as storage area network (SAN) or network-attached storage (NAS), which is then pooled and provisioned dynamically), IaaS clouds are extended to provision rack-aware workloads and include support for object storage. Refer to Figure 1-6. A typical IaaS cloud is depicted on the left and an IaaS cloud that can support big data workloads is shown on the right. The object storage support of the expanded IaaS cloud enables big data workloads to scale horizontally and allows the stored objects to be accessed by any node in the server racks using a fully qualified Uniform Resource Identifier (URI). However, the primary intent of this storage is *not* to be used inside VMs for data processing.

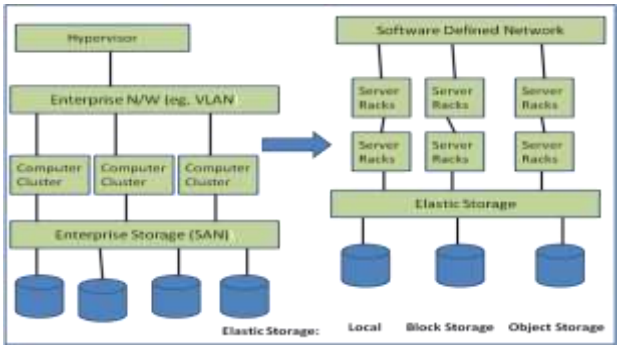

Figure 1-6 Traditional versus cloud deployments for big data

In this scenario, IaaS cloud storage can be thought of in terms of *primary* and *secondary* storage:
_ Primary storage, also called *boot storage*, is any type of storage that can be mounted on the node of a cluster. It also holds the disk images of running VMs and user data.
_ Secondary storage, or *object storage*, is a different way of storing, organizing, and accessing data on disk. An object storage platform provides a storage infrastructure to store files with significant metadata added to them (the files are then referred to as *objects*). The back-end architecture of an object storage platform is designed to present all of the storage nodes as a single pool. With object storage, there is no file system hierarchy. The cloud environments might implement object storage by adopting Open Stack Swift software. Object storage typically provides data protection by making multiple copies of the data (for example, the *three copies* of data paradigm promoted by public cloud vendors, such as Amazon and Rack space).
With these changes to the cloud architecture, we can finally bring together big data and cloud computing. The following scenarios are now possible:
_ Provision a Hadoop cluster on bare-metal hardware
_ Operate a hybrid cloud (part hypervisor for VM provisioning, part bare metal for the data store), which is the most common cloud configuration today
_ Reconfigure the entire cloud on demand

Conclusion

Big data is currently one of the most critical emerging technologies. Organizations around the world are looking to exploit the explosive growth of data to unlock previously hidden insights in the hope of creating new revenue streams, gaining operational efficiencies, and obtaining greater understanding of customer needs. Cloud computing seems to be a perfect vehicle for hosting big data workloads. However, working on big data in the cloud brings its own challenge of reconciling two contradictory design principles. Cloud computing is based on the concepts of *consolidation* and *resource pooling*, but big data systems (such as Hadoop) are built on the *shared nothing* principle, where each node is independent and self-sufficient. A solution architecture that can allow these mutually exclusive principles to coexist is required to truly exploit the elasticity and ease-of-use of cloud computing for big data environments.

References

[1]. IBM System x Reference Architecture for Hadoop: http://www.redbooks.ibm.com/abstracts/redp5009.html?Open
[2]. IBM Smart Cloud - Building a Cloud-Enabled Data Center: http://www.redbooks.ibm.com/abstracts/redp4893.html
[3]. R.L .Villars, C.W. Olofson, M. Eastwood, Big data: what it is and why you should care, White Paper, IDC, 2011,

