NavaJyoti, International Journal of Multi-Disciplinary Research
Volume 1, Issue 1, August 2016

Hadoop as a Service (HaaS) via IBM SoftLayer Cloud

Dr. Pethuru Raj
[1] *Cloud Consulting Architect, IBM Global Cloud Center of Excellence (CoE), Bangalore, India*

ABSTRACT

There are several interesting and inspiring trends and transitions succulently happening in the business as well as IT spaces. There are fresh sources pouring out a lot of usable data. The data scope, size, structure, and speed is fast-changing, the device ecosystem expands frenetically with the arrival of trendy and handy, slim and sleek, disappearing and disposable gadgets and gismos for information access and service consumption everywhere every time and all kinds of common, casually found, and cheap articles in our everyday environments are being digitized and service-enabled in order to be smart in their actions and reactions. Thus trillions of smart objects, billions of connected devices and millions of adaptive applications interact with one another over any networks purposefully and hence the amount of interaction, transaction, operation, analytical, collaboration, commercial, business, social, personal and professional data created is growing very rapidly. Now if the data getting collected, processed, and stocked is not subjected to deeper, deft and decisive investigations, then the acquired knowledge (the beneficial patterns, tips, techniques, associations, fresh opportunities and possibilities, etc.) hidden inside the data heaps go unused. For such a large amount of multi-structured data, Hadoop framework is being touted as the best way forward to squeeze out the knowledge. In this paper, you can find the ways and means of setting up and sustaining Hadoop clusters in hybrid cloud environments through the newly crafted Hadoop as a service (HaaS).

Keywords: Big Data Analytics (BDA), Hybrid Clouds, Hadoop, Knowledge Discovery and Dissemination, Decision-enablement, the Internet of Things (IoT)

INTRODUCTION

The era of big data is in full bloom with the size of the data being generated, captured, stocked, polished and processed is reaching astronomical proportions these days. This huge volume of data is becoming possible mainly due to the unprecedented levels of the adoption and adaption of proven and potential information and communication technologies (ICT). Especially the recently incorporated connectivity methods along with the fast-expanding network topologies, techniques and tools have the inherent capabilities in establishing and sustaining seamless and spontaneous connectivity among:

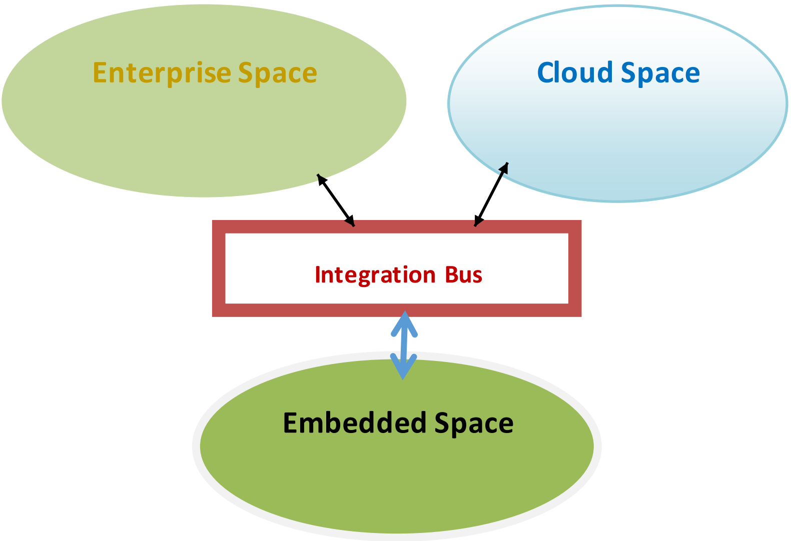
1. the growing array of digitized elements (smart objects / sentient materials) at the ground level with the embedding of edge technologies (sensors, actuators, chips, controllers, codes, tags, stickers, beacons, specks, LED lights, etc.,)
2. all kinds of personal and professional devices (physical, mechanical, electrical, electronics, etc.) in and around us and
3. scores of remotely held business, analytical, transactional, and operational applications at enterprise and cloud environments

The growing number of types of integrations and interactions among these entities as illustrated in the below picture results in a large amount of data being produced. This data explosion also challenges

Address for correspondence:

Mail Id: peterindia@gmail.com

the way decisions are being made as data is being widely recognized as the strategic asset for any individual, innovator, and institution to prepare towards the intended prosperity. It is sure that big data facilitates big insights.



Considering that every kind of enterprise across the globe is hugely betting and banking on data-driven diagnostic, predictive and prescriptive insights, the arrival of the hugely popular Hadoop framework is seen as a blessing in disguise. However, there are technical and infrastructural challenges associated with the setting up and sustaining Hadoop clusters. Cloud service providers are bringing in as much automation as possible in order to enable customers to have the right and relevant Hadoop cluster in place quickly and easily through the recently introduced Hadoop as a Service (HaaS). This paper is specially crafted to illustrate the various capabilities of the HaaS.

Describing the Big Data Paradigm

There are deeper and extreme connectivity methods flourishing these days. Integration and orchestration techniques, platforms, and products have matured significantly. The result is that Information and Communication Technology (ICT) infrastructures, platforms, applications, services, social sites, and databases at the cyber level are increasingly interconnected with devices, digitalized objects (smart and sentient materials) and people at the ground level via a variety of networks and middleware solutions. There is a strategic, seamless and spontaneous convergence between the virtual and physical worlds. All these clearly insist the point that data creation / generation, capture, transmission, storage and leverage needs have been growing ceaselessly. This positive and progressive trend is indicated and conveying a lot of key things to be seriously contemplated by worldwide business and IT executives, engineers, and experts. New techniques, tips, and tools need to be unearthed in order to simplify and streamline the knowledge discovery process out of data heaps. The scope is bound to enlarge and there will be a number of fresh possibilities and opportunities for business houses. Solution architects, researchers, and scholars need to be cognizant of the niceties, ingenuities, and nitty-gritty of the impending tasks of transitioning from data to information and then to knowledge. That is, the increasing data volume, variety, and velocity have to be smartly harnessed and handled through a host of viable and valuable mechanisms in order to extract and sustain the business value.

We now live in the age where the number of electronic devices has already overtaken the human population worldwide. There are more than 7 billion mobile devices on this planet. The data generated by billions of users over billions of devices for human to human interaction, human to machine interaction, machine to machine interaction is in the range of zettabytes. Then there are other sources of data. Apart from the mesmerizing number of mobiles, there are implantables, wearables and portables, electrical, mechanical, physical, and fixed devices in plenty in our everyday environments. Sensors, actuators, robots, machines, instruments, equipment, consumer electronics, utensils, wares, toolsets, and a wider variety of components in avionics, automotive electronics, energy grids, manufacturing plants, etc. are all contributing for the ensuing big data era. There are software applications, services and a number of diverse databases, data warehouses, data marts, data cubes, etc., emitting large quantities of data in different speeds and structures. The scope for data getting captured and crunched are also varying. The prickling and the perpetual challenges are how to make sense out of big data quickly and easily.

Pioneering data analytics platforms, information-processing and data communication infrastructures, highly synchronized processes, enabling architectures, data virtualization and visualization toolsets are the prominent requirements for knowledge discovery and dissemination out of big data. The implications of big data are vast and varied. The principal activity is to do a variety of tool-based and mathematically sound analyses on big data for instantaneously gaining big insights. It is a well-known fact that any organization having the innate ability to swiftly and succinctly leverage the accumulating data assets is bound to be successful in what they are operating, providing and aspiring. That is, besides instinctive decisions, informed decisions go a long way in shaping up and confidently steering organizations. Thus, just gathering data is no more useful but IT-enabled extraction of actionable insights in time out of those data assets serves well for the betterment of businesses. Analytics is the formal discipline in IT for methodically doing data collection, filtering, cleaning, translation, storage, representation, processing, mining, and analysis with the aim of extracting useful and usable intelligence. Big data analytics is the newly coined word for accomplishing analytical operations on big data. With this renewed focus, big data analytics is getting more market and mind shares across the world. With a string of new capabilities and competencies being accrued out of this recent and riveting innovation, worldwide corporates are jumping into the big data analytics bandwagon. This chapter is all for demystifying the hidden niceties and ingenuities of the raging Big data analytics.

Big Data Analytics: The Evolving Challenges

Big data is the general term used to represent massive amounts of data that are not stored in the relational form in traditional enterprise-scale databases. New-generation database systems are being unearthed in order to store, retrieve, aggregate, filter, mine and analyze big data efficiently. The following are the general characteristics of big data:

- Data storage is defined in the order of petabytes, exabytes, etc. in volume to the current storage limits (gigabytes and terabytes).
- There can be multiple structures (structured, semi-structured and less-structured) for big data.
- Multiple types of data sources (sensors, machines, mobiles, social sites, etc.) and resources for big data.
- Data is time-sensitive (near real-time, as well as real-time). That means big data consists of data collected with relevance to the time zones so that timely insight can be extracted.

Since big data is an emerging domain, there can be some uncertainties, potential roadblocks, and landmines that could probably unsettle the expected progress. Let us consider a few that are more pertinent:

- **Technology** – Technologies and tools are very important for creating the business value of big data. There are multiple products and platforms from different vendors. However, the technology choice is very important for firms to plan and proceed without any hitch in their pursuit. The tool and technology choices will vary depending on the types of data to be manipulated (e.g. XML documents, social media, sensor data, etc.), business drivers (e.g. sentiment analysis, customer trends, product development, etc.) and data usage (analytic or product development focused).
- **Data Governance** – Any system has to be appropriately governed in order to be strategically beneficial. Due to the sharp increase in data sources, types, channels, formats, and platforms, data governance is an important component inefficiently regulating the data-driven tasks. Other important motivations include data security while in transit and in persistence, data integrity, and confidentiality. Further on, there are governmental regulations and standards from world bodies and all these have to fully comply with in order to avoid any kind of ramifications at a later point in time.
- **Skilled Resources** – It is predicted by MGI that there will be a huge shortage of human talent for organizations providing big data-based services and solutions. There will be requirements for data modelers, scientists, and analysts in order to get all the envisaged benefits of big data. This is a definite concern to be sincerely attended by companies and governments across the world.
- **Accessibility, Consumability and Simplicity** – Big Data product vendors need to bring forth solutions that extract all the complexities of the big data framework from users to enable them to extract business value. The operating interfaces need to be intuitive and informative so that the goal of ease of use can be ensured for people using big data solutions.

Big data’s reputation has taken a bit of a battering lately thanks to the allegations that the NSA is silently and secretly collecting and storing people’s web and phone records. This has led to a wider debate about the appropriateness of such extensive data-gathering activities. But this negative publicity should not detract people from the reality of big data. That is, big data is ultimately to benefit society as a whole. There’s more to these massive data sets than simply catching terrorists or spying on law-abiding citizens.

In short, big data applications, platforms, appliances and infrastructures need to be designed in a way to facilitate their usage and leverage for everyday purposes. The awareness about the potentials need to be propagated widely and professionals need to be trained in order to extract better business value out of big data. Competing for technologies, enabling methodologies, prescribing patterns, evaluating metrics, key guidelines, and best practices need to be unearthed and made as reusable assets.

Hadoop is the leading technology for accomplishing big data analytics efficiently and cost-effectively. There are several Hadoop implementations (open source as well as commercial-grade) available today in the market. Further on, there are a variety of NoSQL databases for simplifying and streamlining big data storage and analytics. Apart from that, the traditional analytical systems such as relational databases (RDBMS), data warehouses (DW), and business intelligence (BI) tools are being accordingly enhanced to take up the impending challenges of big data There are proponents recommending hybrid architecture for big data analytics.

The Hadoop Challenges

The challenges include the following:

1. Setting up and maintaining a Hadoop Cluster
2. Writing MapReduce applications

The Hadoop ecosystem is steadily on the climb through the provision of several tools such as Hive, Pig, etc., to make big data processing simpler and smoother. Consequently, a minimal Hadoop as a Service (HaaS) offering provides a managed Hadoop cluster ready to use without the need to configure or install any of the Hadoop relevant services on any cluster nodes like Jobtracker, Tasktracker, Namenode, Datanode, and may provide secondary services like Zookeeper or HBase. Such a service also provides some of the most commonly used tools pre-installed and configured like Hive, Pig, and Sqoop. More advanced services expand this even further and include graphical interfaces and optimizations, which enable a wide user audience to utilize Hadoop transparently with common skills like SQL. Depending on the level of service, abstraction, and tools provided, Hadoop as a Service (HaaS) can be placed in the cloud stack as a Platform or Software as a Service solutions, between infrastructure services and cloud clients.

How Hadoop as a Service (HaaS) works in IBM SoftLayer Cloud?

HaaS takes the advantage of cloud infrastructures to instantiate a Hadoop cluster at the moment when it is needed and to scale it depending on demand. This happens without any work required by the user. A user provides general cluster information e.g. min, max number of nodes and what instance types, with sensible (cost free in the trial period) defaults being pre-selected for new customers. IBM SoftLayer Cloud manages the cluster that includes starting a cluster when a customer queries data with Hive or starts any type of Hadoop job. The cluster scales within the defined bounds depending on the load on it and SoftLayer shuts the cluster down when it is not used anymore. Load managed clusters ensure that customers only pay for resources they actually use.

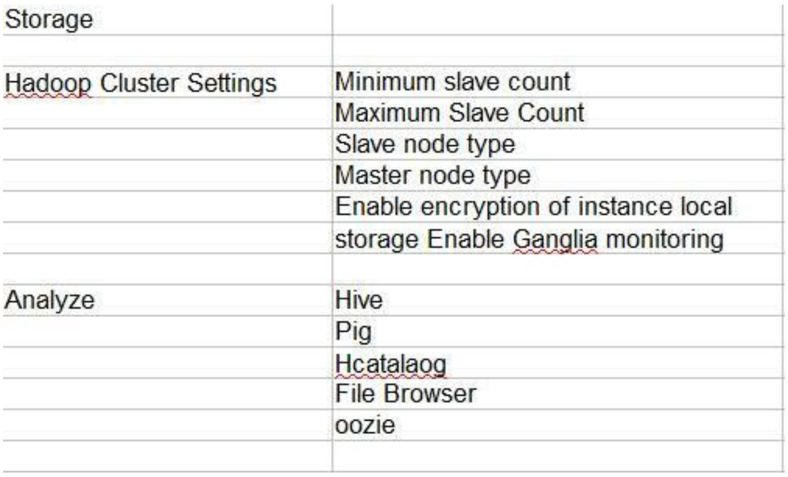
Customers, therefore, do not have to worry about Hadoop configurations, software versions, optimal settings, stability, starting or stopping a cluster. The cluster scales, appear and disappear automatically on a need basis and requires no work. This removes the operational burden and makes it effortless to use and change between small and large cluster deployment up to thousands of nodes effectively.

Hadoop and NoSQL Databases

In addition to Hadoop platforms, NoSQL databases can be provided as a service in sync up with Hadoop. Instead of using HDFS, it is possible to use NoSQL databases for data storage, querying, retrieval, and analytics.

What are the Prominent Modules for HaaS?

1. **Put up an intuitive User Portal –** The key details to be provided by users include the following:



Storage	
Hadoop Cluster Settings	Minimum slave count
	Maximum Slave Count
	Slave node type
	Master node type
	Enable encryption of instance local
	storage Enable Ganglia monitoring
Analyze	Hive
	Pig
	Hcatalaog
	File Browser
	oozie

2. **Integration with Backend Hadoop Platform –** There are many components in the consistently expanding Hadoop ecosystem. The well-known and matured modules are given below.

    1. Core Hadoop platform (Hadoop HDFS and Hadoop MapReduce)
    2. Non-relational database (Apache HBase)
    3. Metadata services (Apache HCatalog)
    4. Scripting platform (Apache Pig)
    5. Data access and query (Apache Hive)
    6. Workflow scheduler (Apache Oozie)
    7. Cluster coordination (Apache Zookeeper)
    8. Data integration services (HCatalog APIs, WebHCatalog, and WebHDFS)

The Business and Technical Cases of Hadoop as a Service (HaaS)

- Harness the robust processing power of Hadoop without needing to train or pay someone to build and maintain a system
- Build complex data flows without writing a line of code
- Quickly connect to data sources and destinations
- Provision or terminate clusters with just a few clicks
- Monitor clusters and running processes
- Adjust settings and receive error notifications right from the dashboard

Save Time and Money with Readymade Hadoop System - Data collection is growing by leaps and bounds, and it is imperative for companies to be able to use that data if they want to keep up with their competition. By using Hadoop, there is no waiting for a system to be built. Resources can be diverted to other areas directly related to the business plan instead of building and maintaining a data center.

Regular Hadoop vs IBM SoftLayer HaaS

Hadoop	Do It Yourself (DIY)	SoftLayer's HaaS
IT Infrastructures (Server, Storage, Network, etc.)	Capital investment is huge	Cloud-based, pay per use
Programming and Testing	We need to write Hadoop Applications	The Service providers could do it

Data flows	Learn and code MapReduce, Hive, Pig, Oozie. Code, debug, deploy, and manage	What to do with data and the goals are just enough.
Cluster Provisioning and Termination	Self-coding	Just a few clicks away
System Updates	Take system offline, implement updates, and miss valuable time	Cloud engineers handle it behind the scenes
Bug Fixes	Same thing as above	All the errors are handled by the Hadoop service provider.
Errors	The user needs to configure his/her own notification system	Get notified right away via the dashboard and by email
Monitor Running Jobs	The user has to sort through all of the cluster nodes to get logs and track job status	Open your browser, sign in, and they will be right there on the dashboard
Execute Data Flows/Jobs	Learn and code Oozie or other workflow management systems	Just a few clicks and the data flows are executed

CONCLUSION

Automation is at its peak everywhere. Here too, for lessening the workloads, Hadoop as a service (HaaS) is being enabled and provided to worldwide subscribers to leverage this facility and feature easily from any part of the world. All kinds of headaches can be substantially eliminated in order to have an easy and quick Hadoop cluster in place in order to proceed with the data ingestion, filtering, transformation, processing and coarse-grained analysis towards extracting viable and venerable insights that can be used by executives and other stakeholders in order to arrive at correct and timely decisions that in turn lead to strategically sound accomplishments.

REFERENCES

[1] IBM SoftLayer Cloud http://www.softlayer.com/

[2] Qubole Hadoop as a Service https://www.qubole.com/hadoop-as-a-service/

[3] Altiscale HaaS https://www.altiscale.com/

[4] High-Performance Big Data Analytics http://www.springer.com/us/book/9783319207438

AUTHOR'S BIOGRAPHY

Dr. Pethuru Raj has been working as a cloud infrastructure architect in the IBM Global Cloud Center of Excellence (CoE), IBM India Bangalore. Previously he worked as TOGAF-certified enterprise architecture (EA) consultant in Wipro Consulting Services (WCS) Division, Bangalore. He also had a fruitful stint as a lead architect in the corporate research (CR) division of Robert Bosch, India. He has gained more than 16 years of IT industry experience.