Big Data Analytics on Cloud Using Microsoft Azure HDInsight

V. Santosh Karthikeyan
Research Scholar, Department of Management Studies
St.Peter's University, Chennai, Tamil Nadu, India

Dr. N. Muthu
Professor, Department of Management Studies
Saveetha Engineering College, Chennai, Tamil Nadu, India

Abstract- Organizations are accelerating big data analytics to view the business through a more open intelligent lens and enable the percentage of growth from new types of data. The future investments are focused on information innovation in addition to the business as usual objectives. Cloud computing adds up to this innovation channel where cloud based big data solution is established to fulfill the key requirements of Hybrid architecture, enable security without any comprise and ecosystem support. Microsoft Azure is the leading vendor which provides the cloud platform to complete the big data solution. Microsoft Azure HDInsight service provides the complete Hadoop stack which enable big data developers to use it for analytics and reporting. It integrates with business intelligence tools such as Power BI, Excel, SQL Server reporting services. In this paper, Microsoft Azure HDInsight is discussed which is the big data cloud service enabling users to easily move their data to the cloud and build a strong architectural pattern to manage the data in terms of high availability, scalability, reliability and performance.

Keywords – Big Data, Analytics, Cloud Computing, Hybrid, Hadoop.

I. INTRODUCTION

Cloud based big data solution is widely established across the industries with easy access to the data both on-premise and cloud with hybrid architecture. Most of the IT organizations have their technology capabilities in the cloud and it is important for other organizations to cloud adoption. All the new solutions are cloud-based giving cloud providers an upper hand on the future business trends.

Microsoft Azure provides an open and flexible cloud platform to rapidly build, deploy and manage secure applications. This platform helps to leverage the current analytical skills and popular tools and frameworks.

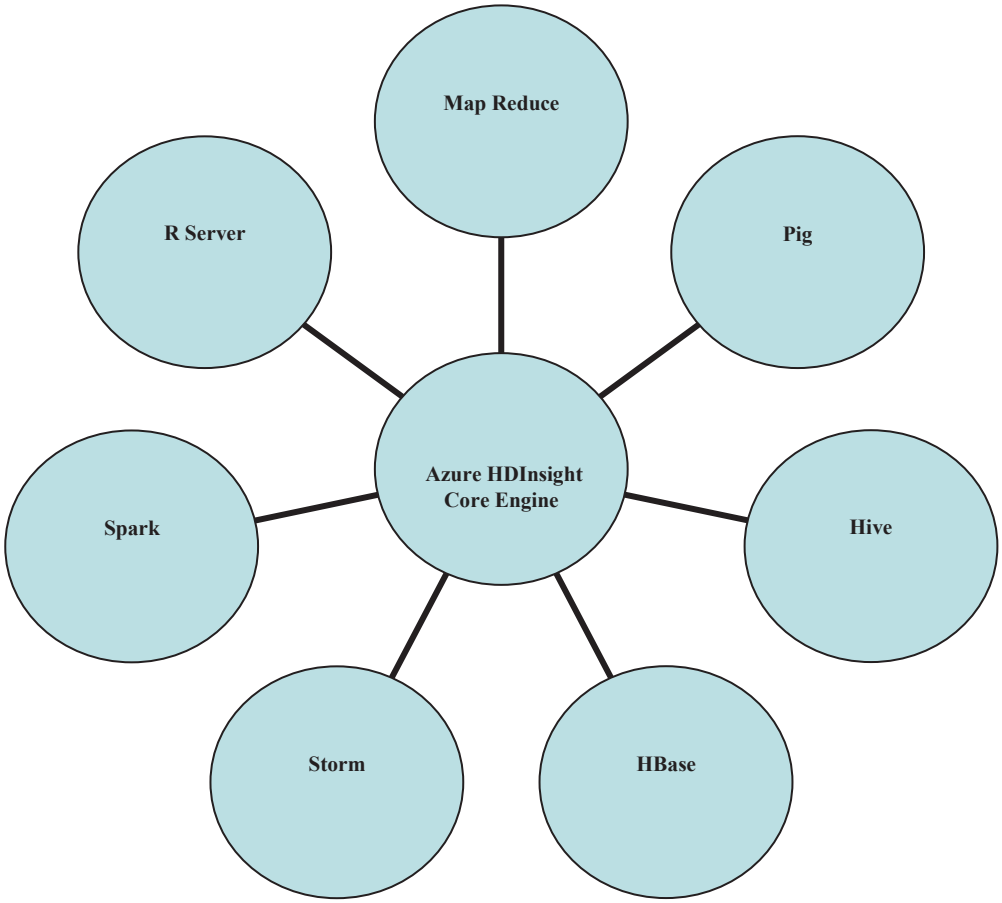
Azure HDInsight is the cloud service for using Hadoop technology ecosystem for big data solution. It enables the provision of Hadoop on cloud, Apache Spark, R server, HBase and Storm clusters. The service also includes implementation of Apache Spark, HBase, Storm, Pig, Hive, Sqoop, Oozie, Ambari and so on. Apache Spark and Storm support the real-time in-memory processing, HBase being a columnar NoSQL transactional database and Hive for SQL queries execution. There are different connectivity options which enable the solution architects to build the hybrid architecture for having the data on-premise and cloud. The storage capability in cloud is phenomenal providing the flexibility to hold the data both primary and secondary in different data centers. Customers access their data round the clock with very limited downtime and high availability cluster in the cloud. This comprehensive set of Apache big data projects within cloud promote reduced infrastructure cost, easy integration with on-premise Hadoop clusters, deployment in Windows or Linux processing unstructured and semi-structured data.

.

Big Data on Cloud:

Apache Hadoop is open source software framework which provides big data solution through multiple components i.e.; Hadoop ecosystem. Each component within Hadoop can work separately and Microsoft Azure HDInsight service provides big data solution on cloud.

Microsoft Azure is a cloud computing platform which is collection of integrated cloud services. Azure HDInsight is one of the key analytics services on cloud. Microsoft Azure can be accessed through azure portal, powershell cmdlets or through command-line interface. The Hadoop services are available through clusters on cloud which constitute the core engine as per below



Hive: Hive is a Hadoop component which performs SQL operations on HDFS. It provides a SQL-like language called HiveQL. When a hive query is executed, a map-reduce job runs in the background and the data are submitted to the cluster.
Pig: Pig is an open-source dataflow system which helps to create and execute map-reduce jobs on very large data sets. It uses a language called Pig Latin with user-defined functions (UDFs).
Spark: It is large scale data processing engine which helps for interactive data analysis.
Storm: Storm is a real-time computation system which makes it easy to process streams of data.
All the above services are readily available which can be deployed based on the requirement and it is easy to use. There is no need to manually install and configure each of these Hadoop components. The cluster can be deployed with the number of nodes needed and it just takes fifteen to twenty minutes of time to make the required instance up and running. It makes our job easy on cloud.
Another service offered by Azure is blob storage which stores the data in the cloud. Combining both HDInsight and blob storage facilitate the execution of map reduce jobs which is the data processing framework in Hadoop. It is easy to create an HDInsight cluster in the azure portal. A storage account should be created which is mapped with

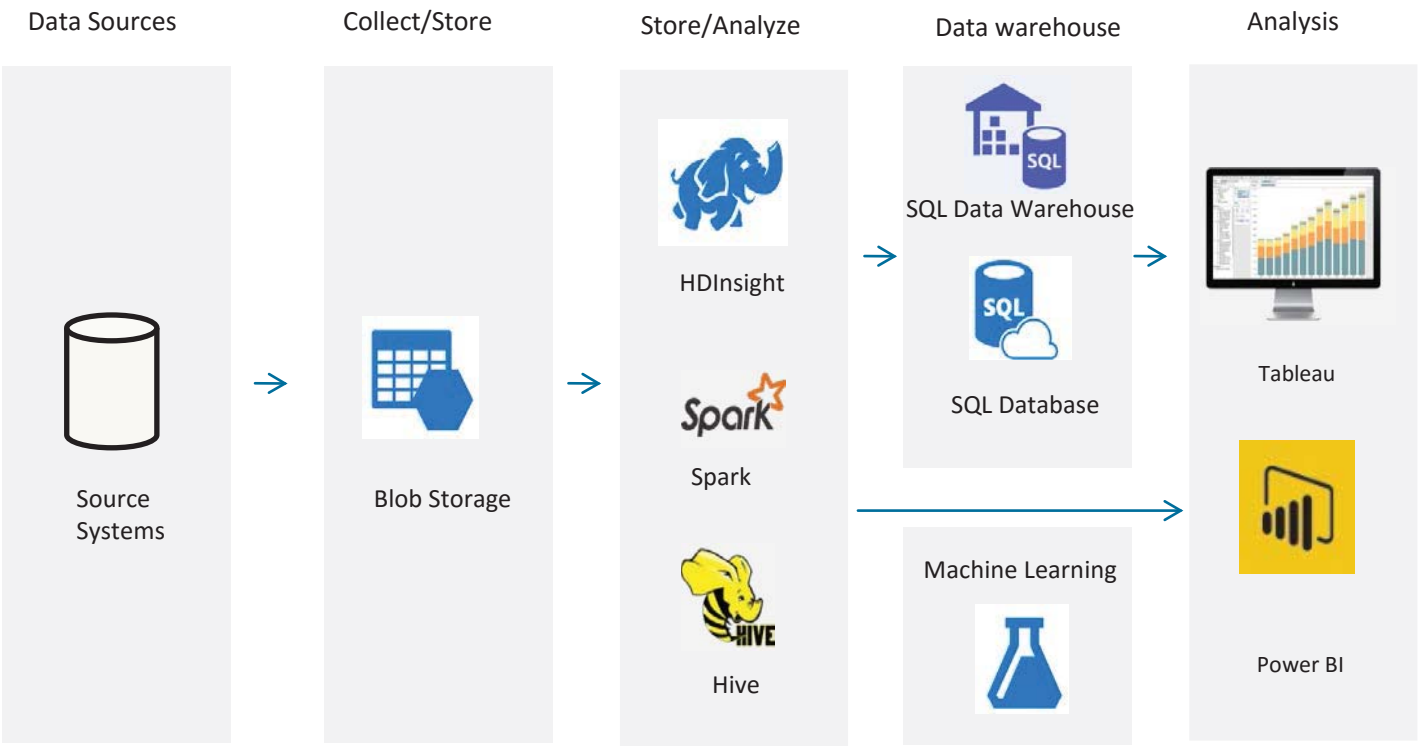
the cluster configuration. Resource Manager is the commonly used azure deployment model which is the latest one as well. Each resource created through resource manager exist within the resource group

Creating an HDInsight Cluster:

HDInsight cluster can be created in azure portal by updating the name of the cluster which becomes the part of the URL for that cluster. Then, number of nodes is provided for defining the cluster size. The next step is to provide a password. It is important to have a storage account or create one to use it in the cluster. It is recommended to have both the storage account and cluster in the same data center keeping in mind the performance and cost.

Modern Cloud Architecture:

The flexible architectural pattern can be applied based on the business need. A typical cloud architecture involving HDInsight components and azure storage services are highlighted below:

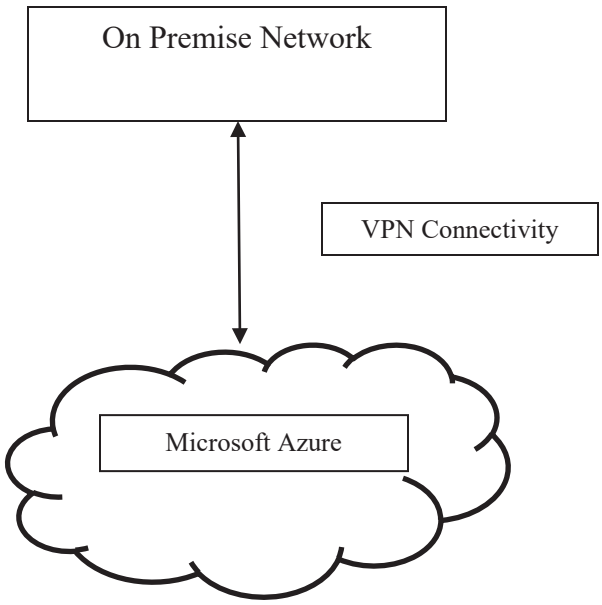


Azure HDInsight services are offered pay as you use and setting up the cluster in the cloud depends on the requirements. All the parameters like number of nodes, zone selection etc., are defined while creating the cluster and it depends on the storage account where the files are stored in the cloud. Azure HDInsight service provides the complete Hadoop stack which enable big data developers to use it for analytics and reporting. It integrates with business intelligence tools such as Power BI, Excel, SQL Server reporting services etc.

Connectivity options: Networking is an important cloud capability and there are three specific VPN connectivity options available in cloud. A point to site VPN helps to create a secure connection to the virtual network. A site to

site VPN establish a secure connection between on-premise site and virtual network. Azure expressroute lets to create a private connection between Azure datacenters and on-premise infrastructure or in a co-location environment. The third connectivity option is usually preferred for hybrid cloud architecture.

Hybrid Cloud Approach:
More than 65% of enterprises have hybrid cloud already. When the need for cloud arises, hybrid approach is the recommended one as we can leverage the existing IT investments to have the data on cloud. The movement of data between on-premise and cloud is faster and easily deployable.



Benefits of adopting Hybrid cloud in enterprise:

- Instant scalability
- Flexibility to meet specific business needs
- Opportunity to lower cost
- Enables phased approach to cloud adoption

There are some limitations as well which include networking conflict in the communication protocol and possibility of security breach through public cloud. Precautionary steps should be taken to both these factors to enable a more secured and networking set up.

II. CONCLUSION

Big Data Analytics is one of the latest technology trends in IT industry and its service on cloud platform enable business to build remarkable growth worldwide. It is said that the reasonable cost spent on applications will be consumed via the cloud. In this context, Microsoft and other leading vendors are positioned as leader in the market to provide cloud IaaS, PaaS and SaaS. The flexible architectural patterns with hybrid capability through integrated data solution enable easy movement of data to cloud with minimal cost. It is important that all the enterprises build a strategy plan to include cloud in their line of business to sustain in the competitive market. The roadmap for the cloud strategy should be clearly defined and implemented so that pricing are managed effectively aligned with the business requirements.

REFERENCES

[1] Divyakant Agrawal Sudipto Das Amr El Abbadi. Big Data and Cloud Computing: Current State and Future Opportunities. In EDBT, pages 530-533, 2011.
[2] Shang W, Jiang Z M, Hemmati H, et al, "Assisting developers of big data analytics applications when deploying on Hadoop clouds," Proc. 35th International Conference on Software Engineering, ICSE 2013, IEEE Press, May 2013, pp. 402-411.
[3] Microsoft Azure : https://azure.microsoft.com
[4] D. Agrawal, S. Das, and A.E. Abbadi. Big Data and Cloud Computing: New wine or just new bottles? PVLDB, 3(2): 1647-1648.
[5] Article on "An introduction to the Hadoop ecosystem in HDInsight" by C.J. Gronlund, 2016.
[6] Cheng, Y., Qin, C., & Rusu, F.GLADE: Big Data Analytics Made Easy. SIGMOD (pp. 697- 700). AR: ACM., 2012.