Middle-East Journal of Scientific Research
24 (Recent Innovations in Engineering, Technology, Management & Applications): 186-193, 2016
ISSN 1990-9233; © IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.RIETMA130

Hadoop Vs Windows Azure

[1]Mrs. B. Sindu, [2]G. Sathish and [2]N. Saravanan

[1]Department of MCA, Priyadarshini Engineering College,
Vaniyambadi, Tamilnadu, India
[2]Assistant Professor, Department of MCA,
Priyadarshini Engineering College, Vaniyambadi, Tamilnadu, India

Abstract: The world is full of data's or datum's around the clock day in and day out. The data used to perform many operations, which is done as Historical manner, future analysis, decision making process and implementation of data from one way to another. The data is in unstructured manner and then it is in form of certain procedure to arranging, storing, retrieving data using Hadoop or Azure with support of MangoDB, SQLlight. etc., for effient data retrieval and analysis each one have its special features. Both has different set of tools and techniques and also has merits as well as conflicts.

Key words:

INTRODUCTION

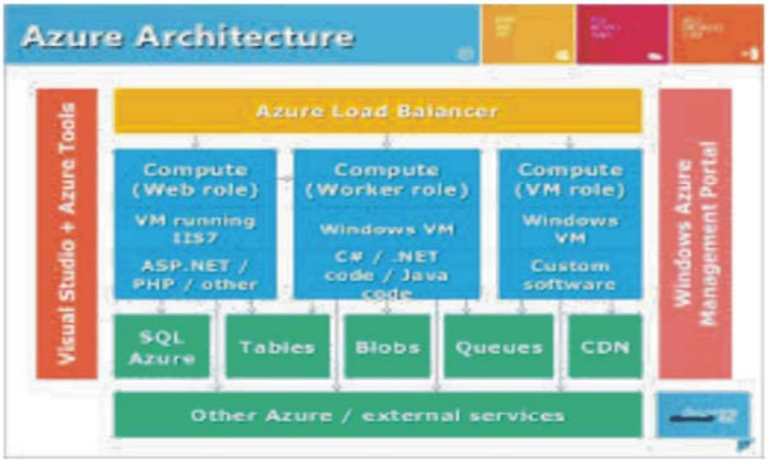
Hadoop: Hadoop is a very unusual kind of open-source data store from the Apache Foundation [1]. However, an entire ecosystem of products has evolved around the Hadoop data store, to the point where it has become its own technology category.

The central idea of Hadoop is that data is spread across many commodity, inexpensive servers, although there are several commercial distributions of Hadoop from Cloudera and Hortonworks who wrap services around the technology [2].

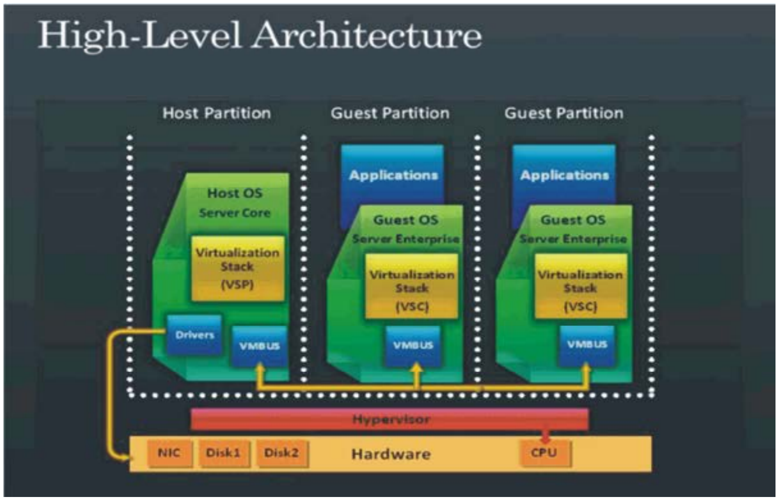
Windows Azure: Windows Azure is a cloud computing platform and infrastructure for building, deploying and managing applications and services through a global network of Microsoft-managed datacenters.

Microsoft Azure is a growing collection of integrated cloud services-analytics, computing, database, mobile, networking, storage and web-for moving faster, achieving more and saving money.

Windows Azure is very scalable and the Preview Portal enables users to have a one stop dashboard to their infrastructure, application and financial elements to their instances. However, the migration of the application from



Corresponding Author: Mrs B. Sindu, PG year Student, Department of MCA,
Priyadarshini Engineering College, Vaniyambadi, Tamilnadu, India.



one subscription to another can be a cumbersome process. Also, deployment of some products that have unsupported third party apps can be difficult as well [3].

History about Azure: Microsoft's Azure cloud started slow. It was announced in October 2008 and launched in 2009, but its appeal was limited by focus on cloud services rather than familiar Windows infrastructure, combined with awkward management tools. In 2012 Azure improved, adding a true IaaS solution based on persistent virtual machines and a user-friendly Web management portal. Azure is now growing fast and since the TechEd conference in April 2014 has added numerous new features intended to let Microsoft shops easily migrate all or part of their server infrastructure and applications to Microsoft's cloud [4].

History about Hadoop: Hadoop was created by Doug Cutting, the creator of Apache Lucene, the widely used text search library. Hadoop has its origins in Apache Nutch, an open source web search engine, itself a part of the Lucene project.

In February 2006, Cutting pulled out GDFS and MapReduce out of the Nutch code base and created a new incubating project, under Lucene umbrella, which he named Hadoop [5]. It consisted of Hadoop Common (core libraries), HDFS, finally with its proper name:) and MapReduce.

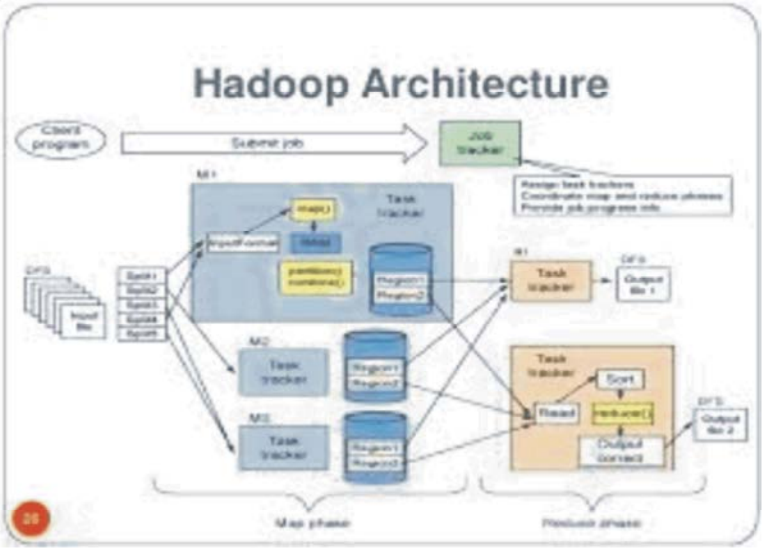


Fig. Hadoop Architecture [2]

2008 was a huge year for Hadoop. At the beginning of the year Hadoop was still a sub-project of Lucene at the Apache Software Foundation (ASF). In January, Hadoop graduated to the top level, due to its dedicated and diverse community of committers and maintainers. Soon, many new auxiliary sub-projects started to appear, like HBase, database on top of HDFS, which was previously hosted at SourceForge. Zoo Keeper, distributed system coordinator was added as Hadoop sub-project in May.

While some of the features, services and options that you'll find in Azure and AWS can't be fully compared to one another, many come pretty close. Here's our attempt at a side-by-side comparison between the two cloud platforms [6].

Features in Hadoop

Distributed Metadata: The default Hadoop architecture uses a single NameNode to store the metadata. This forces all data into a bottleneck and limits clusters to 50-200 million files. It also creates a single point-of-failure (SPOF). If the NameNode were to fail, the entire cluster would be useless [7].

Other distributions try to sidestep the problem by using a secondary NameNode. Secondary NameNodes run as a slave to the primary NameNode and only replicate data from it on a periodic basis. This means that those depending on a secondary NameNode cannot trust its data integrity.

The only real solution to the NameNode problem is to remove it. With the MapR Distribution no-NameNode solution, there are no practical limits to the number of files that can be stored on MapR. This foundational change in the Hadoop architecture distributes the metadata amongst several nodes, which is illustrated below.

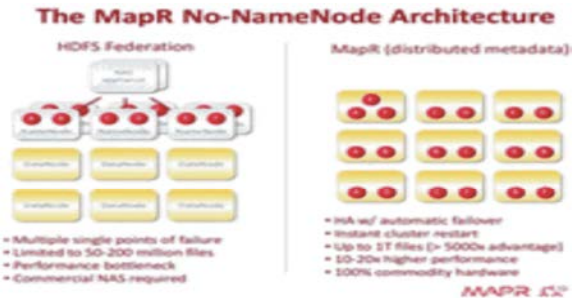
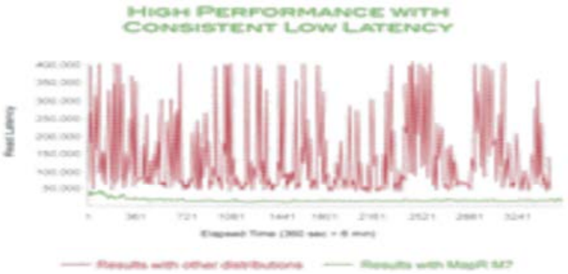

The MapR No-NameNode Architecture
HDFS Federation
MapR (distributed metadata)
• Multiple single points of failure
• Limited to 50-200 million files
• Performance bottleneck
• Commercial NAS required
• HA w/ automatic failover
• Instant cluster restart
• Up to 1T files (> 5000x advantage)
• 10-20x higher performance
• 100% commodity hardware
MAPR

Fig. MapR No-NameNode Architecture

In addition to its benefits for dependability, its database performance boost is also remarkable. With only commodity hardware, you can gain 10-20 times the performance over all other distributions that utilize the centralized metadata structure.

This feature is an architectural improvement to Hadoop that MapR initiated in its infancy. The power it adds to our offering's dependability and performance makes it untouchable by competitive offerings.

Low Latency: Your Hadoop infrastructure needs to be fast. Equally as important, it needs to stay that way. A dirty secret among many Hadoop distributions is the staggering volatility in performance and latency. The MapR M7 disk strategy obviates compactions and defragmentation that can affect performance. Because of this ability, MapR M7 achieves 5x better performance, with low 95th and 99th percentile latencies. The graph below compares the high performance and consistent low latency of the MapR M7 Edition in comparison to other Hadoop distributions.



Notice how M7's highest point of latency is much lower than the other distributions. The difference in volatility is even more shocking. With M7, you can depend on a consistent low latency experience.

High Availability: High availability (HA) refers to the capability of a Hadoop system to continue functioning, regardless of multiple system failures. For companies running mission-critical applications HA is a necessity.

The best way to ensure that your distributed system is highly available is by using an architecture that distributes the metadata. The MapR architecture increases performance and removes the SPOF.

The MapR Distribution for Hadoop provides high availability with self-healing and support for multiple failures. This means that your Hadoop infrastructure will be accessible during system failures, system upgrades and data recoveries.

Snapshots: Other distributions use the HDFS snapshot system, which has several downsides when compared to the MapR Distribution for Hadoop:

True Point-In-Time: HDFS snapshots only capture data that is closed at the time the snapshot is taken. If you are using snapshots as an automated recovery system, you will have no guarantees that the data is complete. With MapR, you can perform point-in-time recovery of all files and tables, whether they are open or not.

Supports All Applications: MapR Snapshots support all Hadoop applications by default.

No Data Duplication: MapR snapshots never duplicate your data and share the same storage with your live information. This allows clients to capture snapshots of a 1 petabyte cluster in just seconds.

As we look at these features that are exclusive to MapR, it seems obvious why our customers are continually excited about our product offering. We feel this was made apparent in our scores in the previously mentioned independent evaluation.

Features in Microsoft Azure

Microsoft Azure Grows Up: Microsoft's Azure cloud started slow. It was announced in October 2008 and launched in 2009, but its appeal was limited by focus on cloud services rather than familiar Windows infrastructure, combined with awkward management tools. In 2012 Azure improved, adding a true IaaS solution based on persistent virtual machines and a user-friendly Web management portal. Then in April 2013, IaaS features including the new VM and virtual network features moved from preview to general availability.

Azure is now growing fast and since the TechEd conference in April 2014 has added numerous new features intended to let Microsoft shops easily migrate all or part of their server infrastructure and applications to Microsoft's cloud.

Azure Files: Azure Files, now in preview, lets you create SMB 2.1 shared folders on Azure's storage service. First you create a new storage account, with options for regional or geo-redundant replication, then you create a file share using Power Shell scripts or. Net code. Access is protected using a long storage key generated by Azure. You can access the shared folder from an Azure VM using that key, just as you would on any Windows network, including the ability to map a drive letter with Net Use. The same files are also accessible over the Internet using PowerShell or REST APIs. Typical uses include migrating on-premise applications that use shared folders and storing files for a website served from multiple VMs.

Azure Site Recovery: In October 2013, Microsoft announced Hyper-V Recovery Manager, a service that enabled Azure to orchestrate site-to-site replication and recovery in event of disaster. Though it makes sense to have third-site manage recovery if your primary site fails, the customer still needed to have servers in two data centers for Hyper-V Recovery Manager to work, limiting it to the large businesses.

The service is now renamed Azure Site Recovery and lets you replicate and recover to VMs hosted on Azure itself, extending its value to businesses of almost any size. Site recovery is based on Hyper-V Replica, which keeps VMs synchronized with only a small delay. The on-premise side is configured with System Center Virtual Machine Manager.

Azure Expressroute: Azure ExpressRoute lets you connect your data center with Azure via a private link that does not travel over the Internet. The advantage is security, lower latency and higher reliability. Bandwidth is up to 1Gbps, or up to 10Gbps if you connect directly through an exchange provider (Equinix or Level 3).

At TechEd Microsoft announced general availability of the service, including an enterprise SLA (Service Level Agreement). Providers include AT and T, Equinix, Verizon, BT, Level 3, TelecityGroup, SingTel and Zadara. In order to take advantage, you need an existing VPN or Ethernet connection to your exchange provider or to have servers co-located in the exchange provider's data center.

New Azure Vm Extensions: Microsoft has developed new extensions for Azure VMs. These include support for Puppet and Chef, which lets you configure VMs with agents to manage their configurations and security extensions that let you install antimalware protection, using services from Symantec, Trend Micro, or Microsoft itself. The third-party security services are installed on a trial basis and you need to purchase a license from the vendor to continue. Microsoft's Antimalware is free while in preview. The details regarding what is protected depend on the product you choose. Microsoft Antimalware can also be enabled on other cloud services, such as Web Roles and Worker Roles.

New Azure Portal: Microsoft previewed a new Azure portal at the Build conference in early April. It is incomplete and for many operations you need to click the link to the old portal, but it adds key features. There are new tools to monitor and analyze Azure Web Sites, for example and to set up Web tests and get alerts if a site goes down. There are also devops features (bringing together development and operations), including integration with Visual Studio Online, which provides project management and source control for teams.

There is more attention paid to applications in the new portal, whereas the existing portal is focused on individual services. The new portal will also scale better as Azure adds new features.

New Networking Features: Azure has impressive networking capabilities, but with some frustrations. These are lessened following several key announcements. One is support for multiple site-to-site connections to virtual networks, essential for organizations with several sites; another is the ability to connect virtual networks to each other, such as across different Azure regions.

Another important new feature is reserving public IP numbers. Previously, you couldn't control the public IP generated for a new service. You can also now assign public IPs directly to VMs, bypassing Azure's endpoint control. If you do this, you take responsibility for firewall protection on the machine, though Microsoft says this may change in the future.

New In Azure Active Directory: Azure Active Directory (AD) is a key part of both Azure and Office 365, which uses the same directory. The directory service is free, but a premium version, now in general availability, adds multifactor authentication, security reports showing suspicious access, self-service password reset and group-based application access. Azure AD can be integrated with on-premise AD, enabling single sign-on and simplifying user management.

More significant, Microsoft says Azure AD now supports 1,200 third-party SaaS apps, including Salesforce.com, Box, Citrix GoToMeeting and even Google Apps. Azure AD Premium is also part of the Enterprise Mobility Suite announced in March, along with InTune for mobile device management and Azure Rights Management for protecting sensitive documents.

Cloud App Discovery: Another new service now in preview is Cloud App Discovery. This is not yet integrated into the Azure portal, but will be in due course. In the meantime, it's available on Azure’s site. The idea is to discover which cloud apps are in use within an organization. Microsoft's hope is that businesses will choose to integrate these apps with Azure Active Directory for easier management and control. Cloud App Discovery requires an agent running on client machines, which monitors app usage and sends the information to the service. You can see which apps are most used, categorized by type (such as Travel, CRM and Social). The risk is that employees may feel this is snooping, but the data has obvious business value.

Azure Remoteapp: Microsoft doesn't offer VDI (Virtual Desktop Infrastructure) on Azure and in fact does not allow Windows 7 or Windows 8 desktops to be hosted in any cloud. (See next slide for the lone exception.) But now in preview is Azure RemoteApp, which lets you host Windows applications in Azure and serve them, using Remote Desktop Services, to Windows, Mac, iOS and Android devices. The preview does not let you install new applications, but you will soon be able to publish custom applications. Authentication is via Azure Active Directory.

Support for iOS and Android shows that Microsoft is following through on its "any device" strategy, though Windows Phone is not yet included.

Windows 7 On Azure -- But Only For Developers: Microsoft now offers Windows 7 and Windows 8 VMs on Azure, but only to MSDN (Microsoft Developer Network) subscribers, limiting their use to test and development. Microsoft’s licensing FAQ states, "Multitenant hosting is restricted in the Product Use Rights of Windows Client, such as Windows 7 or Windows 8. Windows Client Desktops are not available on either Azure or on any other Service Provider such as Amazon or Rackspace."

Amazon Workspaces, a cloud-hosted Windows VDI offering, actually runs Windows Server 2008 configured to look like Windows 7. It's an annoying restriction, but at least developers now have a workaround, giving them a quick way to test applications running on the Windows desktop OS [8].

Benefits and Risks of Windows Azure: Ray Wang, an industry analyst, broke down the different components in a research report for companies in the vast Microsoft partner ecosystem gathering this week in Washington, D.C. for the Worldwide Partners Conference.

About 14,000 people are attending the event. Microsoft executives said the record attendance is due to a few factors, in particular the interest in how Windows Azure will play out for the future of their businesses.

Wang breaks down azure into its three categories:

- Microsoft Windows Azure
- Microsoft SQL Azure (formerly SQL Services)
- Microsoft Windows Azure Platform: App Fabric (formerly. NET Services)

He points out that companies need to really focus on what layer of the service they plan to focus their energies. Those four layers include infrastructure, orchestration, creation and consumption.

Wang spoke to 71 partners for the research report they produced, which details the benefits and risks of the new models that come with Windows Azure [9].

The Advantages: Wang provides six benefits of Windows Azure [8]:

- Faster deployment times and client adoption.
- Greater pool of development resources.
- Recurring revenue streams.
- Improved TCO and margin for differentiated IP.
- Opportunity to break out of the Microsoft client base.
- Lower application lifecycle costs.

Six Risks:

- Potential loss of account control to Microsoft.
- Increased competition for development resources.
- Shift to volume business.
- Decline in upfront profit and revenue collection.
- Accelerated globalization and market competition.
- Increased self-hosting and integration costs.

Difficult Tasks in Windows Azure [8]:

- Migration of application from one subscription to another Azure subscription
- Deployment of some products that have unsupported 3rd party apps
- De ployment of Oracle product due to version mismatch

Hadoop Advantages and Disadvantages

Advantages Of Hadoop [10]

Scalable: Hadoop is a highly scalable storage platform, because it can stores and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving many thousands of terabytes of data.

Cost Effective: Hadoop also offers a cost effective storage solution for businesses' exploding data sets. The problem with traditional relational database management systems is that it is extremely cost prohibitive to scale to such a degree in order to process such massive volumes of data. In an effort to reduce costs, many companies in the past would have had to down-sample data and classify it based on certain assumptions as to which data was the most valuable. The raw data would be deleted, as it would be too cost-prohibitive to keep. While this approach may have worked in the short term, this meant that when business priorities changed, the complete raw data set was not available, as it was too expensive to store.

Flexible: Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. This means businesses can use Hadoop to derive valuable business insights from data sources such as social media, email conversations. Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection.

Fast: Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing. If you're dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes and petabytes in hours.

Resilient To Failure: A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available or use.

Disadvantages Of Hadoop [10]: As the backbone of so many implmentations, Hadoop is almost synomous with big data.

Security Concerns: Just managing a complex applications such as Hadoop can be challenging. A simple example can be seen in the Hadoop security model, which is disabled by default due to sheer complexity. If whoever managing the platform lacks of know how to enable it, your data could be at huge risk. Hadoop is also missing encryption at the storage and network levels, which is a major selling point for government agencies and others that prefer to keep their data under wraps.

Vulnerable By Nature: Speaking of security, the very makeup of Hadoop makes running it a risky proposition. The framework is written almost entirely in Java, one of the most widely used yet controversial programming languages in existence. Java has been heavily exploited by cybercriminals and as a result, implicated in numerous security breaches.

Not Fit For Small Data: While big data is not exclusively made for big businesses, not all big data platforms are suited for small data needs. Unfortunately, Hadoop happens to be one of them. Due to its high capacity design, the Hadoop Distributed File System, lacks the ability to efficiently support the random reading of small files. As a result, it is not recommended for organizations with small quantities of data.

Potential Stability Issues: Like all open source software, Hadoop has had its fair share of stability issues. To avoid these issues, organizations are strongly recommended to make sure they are running the latest stable version, or run it under a third-party vendor equipped to handle such problems.

General Limitations: The article introducesApache Flume, MillWheel and Google's own Cloud Data flow as possible solutions. What each of these platforms have in common is the ability to improve the efficiency and reliability of data collection, aggregation and integration. The main point the article stresses is that companies could be missing out on big benefits by using Hadoop alone.

Reviews

Kamesh001: At Microsoft, with Azure, we see Platform-as-a-Service as critical piece because that's where we can attract the developers and the ecosystem around them. I came to Microsoft because I think Azure is the next great platform that developers can build interesting applications with new business models."

Alex Huai: Despite the annoying cons, the Microsoft azure is still overwhelming and more and more. net cms (http://www.codagenic.com/codagenic.../introduction.html) based websites are using it as hosting server, the developers can use the same programming languages (.net, c++) to write apps, not to mention the scalability as well as cost benefits.

Mickey Alon Insightera' Ceo And Co-Founder: The emergence of Hadoop has changed the data landscape. with Hadoop, you can gain new or improved business insights from structured, unstructured and semi-structure data sources. In addition, large volumes of data which were previously too expensive to store or siloed among departments can be gathered and analysed in one place at an affordable price.

Chief Knowledge Officer, Major Federal Agency: "MapR Hadoop is very well suited for complex environments that support users who are not tech savvy. My specific environment is providing high-end, complex computing solutions to biomedical scientists."

CONCLUSION

The datum's are shaking the world, the world is made up of data. That kind of data need to be stored, implemented, retrievied and maintained. It is so hard and difficult. Day by day security technologies are increasing as well as conflicts also coming parallel for handling huge databases. This Azure Vs Hadoop is to show the best and worst functionalities as well as merits and conflicts of Azure and Hadoop. To show the users to choose wise one to handle their large amount of data effectively.

REFERENCES

1. International Journal of Computer Science Trends and Technology (IJCST)-Volume 4 Issue 1, Jan-Feb 2016 Research Article: Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table, M. Dhavapriya and N. Yasodha, Department of Computer Science, NGM College, Pollachi, Tamil Nadu-India.
2. International Journal of Advanced Research in Computer Science and Technology (IJARCST 2014)-2(2), ver. 2(April-june2014) Big Data and Hadoop 'Dr. Rakesh Rathi',' Sadhya Lohiya' Assistant professor," PG Scholor" Deportment of computer Engineering, Govt. Engineering College, Ajmer, Rajasthan, India.

3. Abinav pothuganti/(IJCST) international Journal of Computer Science and Information Technologies, 2015. 6(1): 5222-527 “Big Data Analytics: Hadoop-Map Reduce and NoSQL Database”. ’Abinav Pothuganti’ Computer Science and Engineering, CbIT, Hyderabad, Telangana, India.
4. International Journal of Innovative Research in Information Security (IJIRIS), “Survey On Big Data Processing Using Hadoop, Map Reduce”, N. Alamelu Menaka Department of Computer Applications.
5. http://www.simonellistonball.com/technology/hadoop-hive-inputformat-azure-tables.
6. http://blogs.msdn.com/innov8showcase/archive/2010/04/14/dynamic-scaling-windows-azure.aspx.
7. http://code.msdn.microsoft.com/azurescale/Release/ProjectReleases.aspx?ReleaseId=4167.
8. http://devproconnections.com/windows-azure-development/windows-azure-development-architectural-overview.
9. http://devproconnections.com/windows-azure-development/windows-azurefor research-overview.
10. http://stackoverflow.com/questions/26608110/hdinsight-hbase-or-azure-table-storage.
11. http://www.infoworld.com/article/2606720/cloud-computing/155334-10-great-new-features-in-Microsoft-Azure.