ANALYSIS OF ENTERPRISE DATA USING MAPREDUCE ON HADOOP

A PROJECT REPORT

Submitted in partial fulfillment for the award of the degree of

B.TECH

in

Storage Technology

by

Garlapati Saiteja 17BIT0217

Under the Guidance of

Prof. Siva Rama Krishnan S

School of Information Technology & Engineering (SITE)

DECLARATION BY THE CANDIDATE

We hereby declare that the project report entitled **"ANALYSIS OF ENTERPRISE DATA USING MAPREDUCE ON HADOOP"** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the degree of **B.Tech.(Information Technology)** is a record of bonafide project work carried out by us under the guidance of **Prof. Siva Rama Krishnan S** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Signature of the Candidate(s)

Garlapati Saiteja

ABSTRACT

Big data is presently a buzzword throughout the software industry. Many IT giants like IBM, Google and Oracle have invested billions of dollars into the research to develop frameworks that can handle the big data efficiently. In this project, we make an attempt to analyze enterprise data that cannot be analyzed locally due to size and computation limitations. We analyze these datasets and try to extract the hidden insights from the same. The project aims to design an algorithm that can group and analyze the datasets as per the requirement. Based on the results of the program, we try to identify certain patterns with the help of data visualization and predict information.

Literature Survey

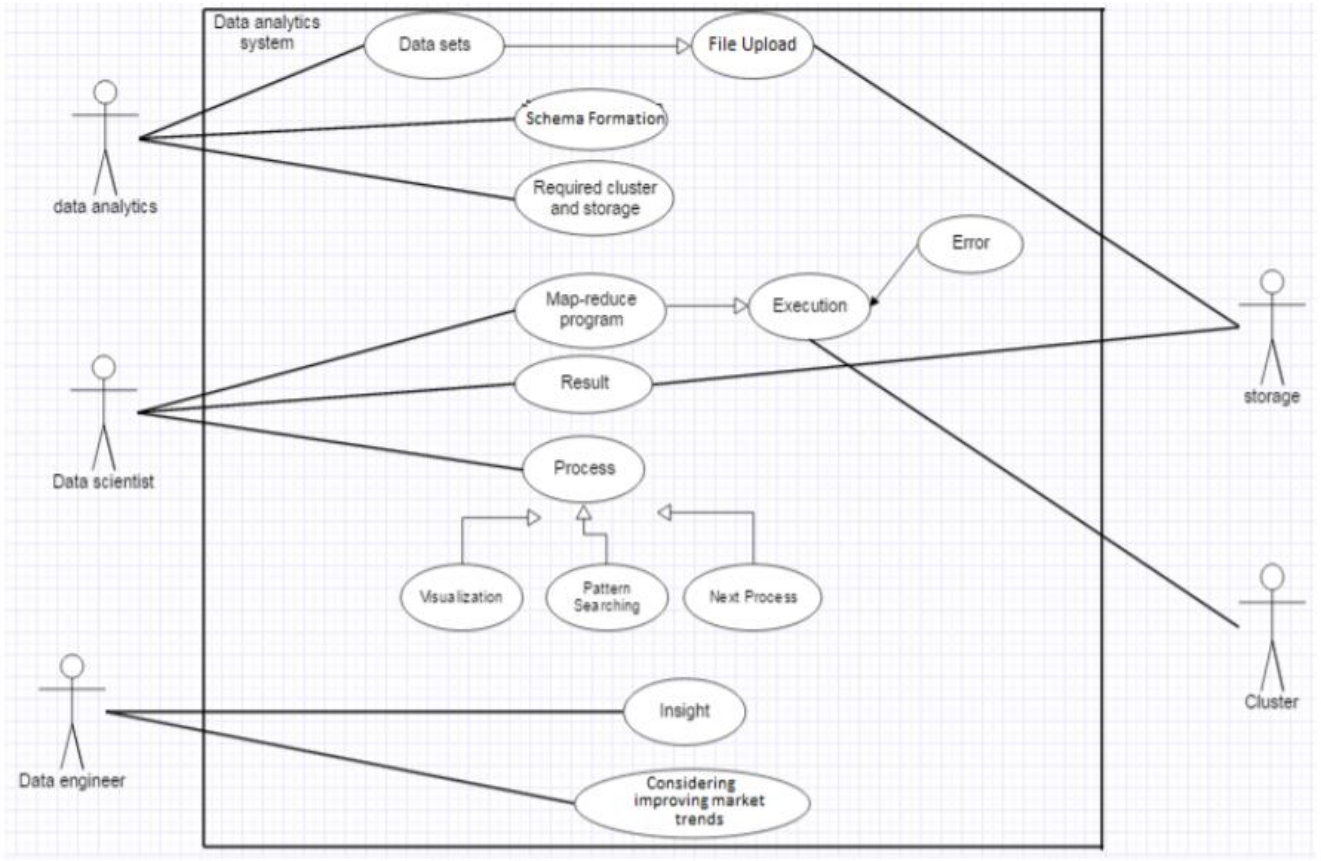
Early 2000s — Google stumbled across an obstacle while carrying out its mission — to organize information from all across the globe — which meant that it was crawling, copying, and indexing the entire Internet continuously. Back then, no software could handle the excessively large volume of data to be processed. Even Google's own custom infrastructure couldn't.

To deal with the situation, Google's engineers designed and built a new data processing infrastructure that comprised of two core components — the Google File System, or GFS, which provided fault-tolerant, reliable, and scalable storage, and MapReduce, a data processing system that allowed work to be split among large numbers of servers and carried out in parallel. Google published an academic paper [1] in 2004 describing its work.

Doug Cutting, a well-known open source software developer, thereafter decided to use the technique Google's paper described. He was working on a web crawler called Nutch [2] and was having the same problems with data volumes and indexing speed that had driven Google to develop MapReduce. He replaced the data collection and processing infrastructure behind the crawler, basing his new implementation on MapReduce. He named the new software Hadoop, after a toy stuffed elephant that belonged to his young son.

Hadoop is an open source project [3] and operates under the Apache Software Foundation today. Hadoop has become a household name and is one of the most popular technologies today to handle big data. It is a data storage and analysis system which is scalable, incredibly flexible and works under the assumption that hardware failures are common occurrences and should be automatically handled by the framework [4] – an assumption that directly leads to its fault tolerant nature. Hadoop can be deployed in a traditional on-site datacenter as well as in the cloud. Microsoft offers its cloud services via the Microsoft Azure Cloud Service platform which includes HDinsight – the service which shall be used in this project to create and deploy clusters as well as run mapreduce jobs on the data that needs to be analyzed. HDinsight offers efficient, reliable and performance centric results [5] with a pay-per-use model which is perfect for a project like the one we are aiming for.

SYSTEM DESIGN AND IMPLEMENTATION



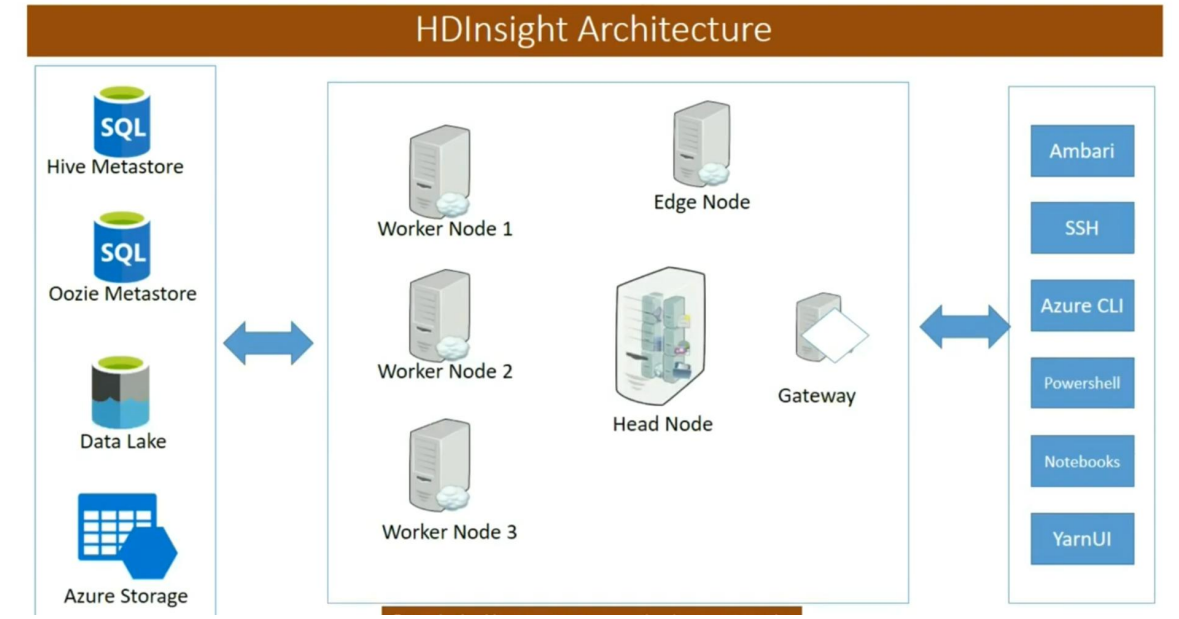
This project uses Microsoft's Azure Cloud Service which provides us with a multitude of services. HDinsight, specifically, is the service that is being used in this project.

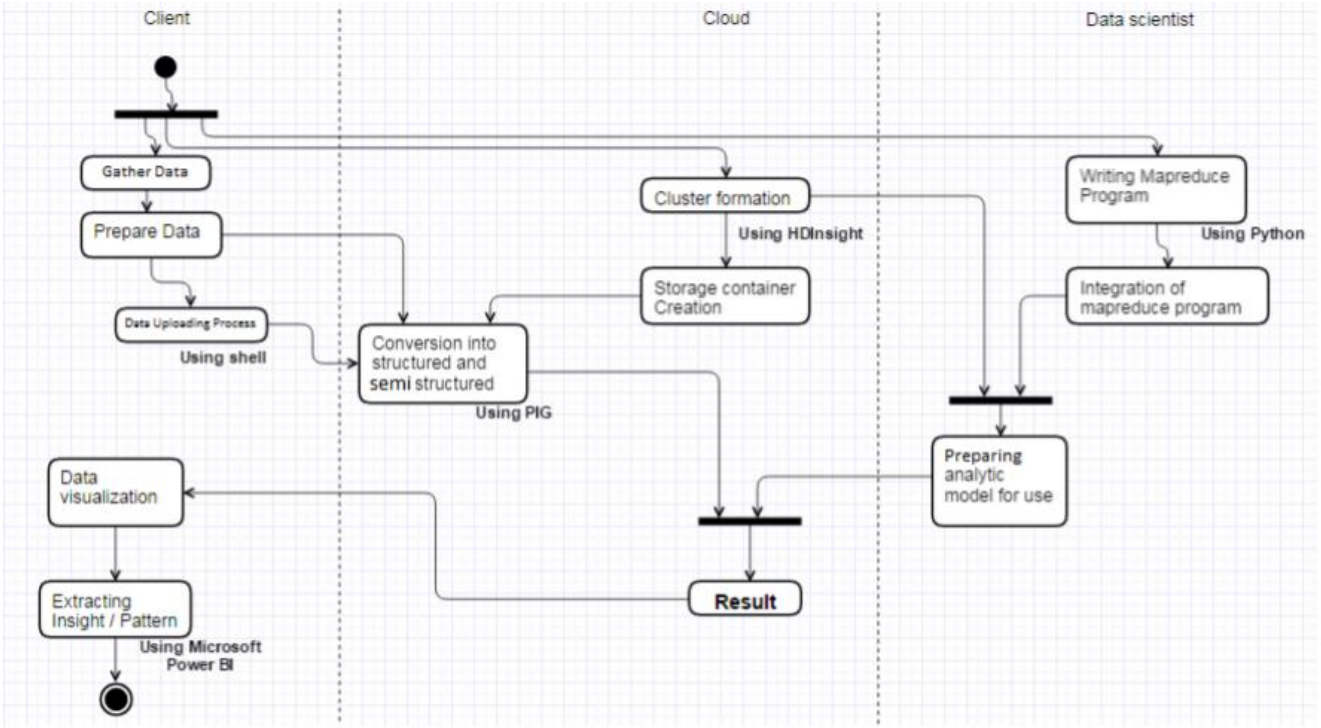
Azure HDInsight is a service that deploys Hadoop on Microsoft Azure. HDInsight uses Hortonworks HDP and was jointly developed for HDI with Hortonworks. HDInsight also supports creation of Hadoop clusters using Linux with Ubuntu. HDinsight is very flexible in its usage and we can, at any point of time, scale up the cluster by increasing the amount or type of worker/head nodes that exist in the cluster.

HDinsight is used to set up and deploy clusters on cloud. These clusters are comprised of head as well as worker nodes. Once the cluster is set up along with the storage container, the cluster can be used to upload/download data as well as to

run programs that, on a fundamental level, utilize mapreduce to perform analytical tasks.

HDinsight architecture

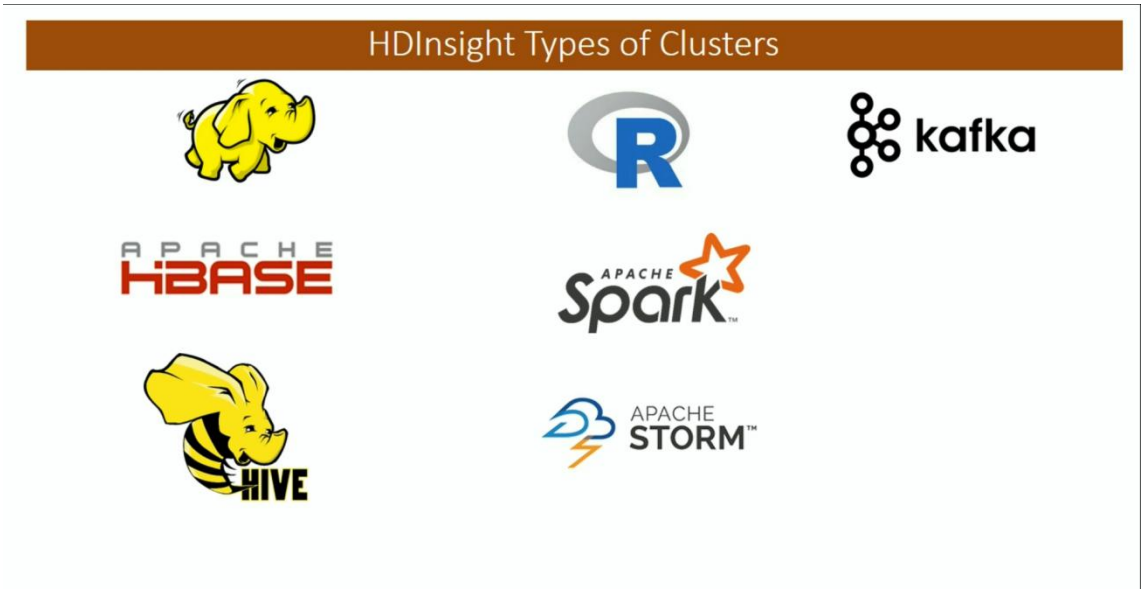




The initial step comprises of **gathering data** that needs to be analyzed. Depending on the size of the data and scope of the analytical process which determines the processing power needed, a **cluster is deployed** using HDinsight on the cloud. The Hadoop cluster, depending on the user's need, can be deployed as a cluster utilizing Linux-based machines or Windows-based machines. And depending on the choice made, HDinsight offers different services unique to the selected OS. For example, HDinsight offers SSH access to the Linux clusters while the Windows clusters have exclusive access to the Remote Desktop Access functionality which is otherwise missing from the Linux-based clusters. Since Microsoft's toolset for a Big Data Analyst comprises of a variety of tools, the entire process of cluster creation and deployment can be automated using a script by utilizing Microsoft's **Powershell** service. The underlying Hadoop framework handles distribution of data for storage using its **HDFS** and distribution of analytical work with **MapReduce** framework. Once the cluster has been deployed, data to be analyzed is **uploaded** to the cluster via Azure Command Line Interface. This data is then analyzed by running a **mapreduce program** on the various machines present on the cluster. The result is then either analyzed further for more efficiency or downloaded locally for it to be used for data visualization. Power BI is a free tool that is then used on the output of the analysis to obtain the results in a graphical format ie, pie-chart, bar graph etc.

There are three main modules in the system design. The modules comprise of the following aspects – Analysis, Data Processing & Visualization.

Types of clusters



Predictive Analysis (using Scala and Predictive Analysis Algorithms)

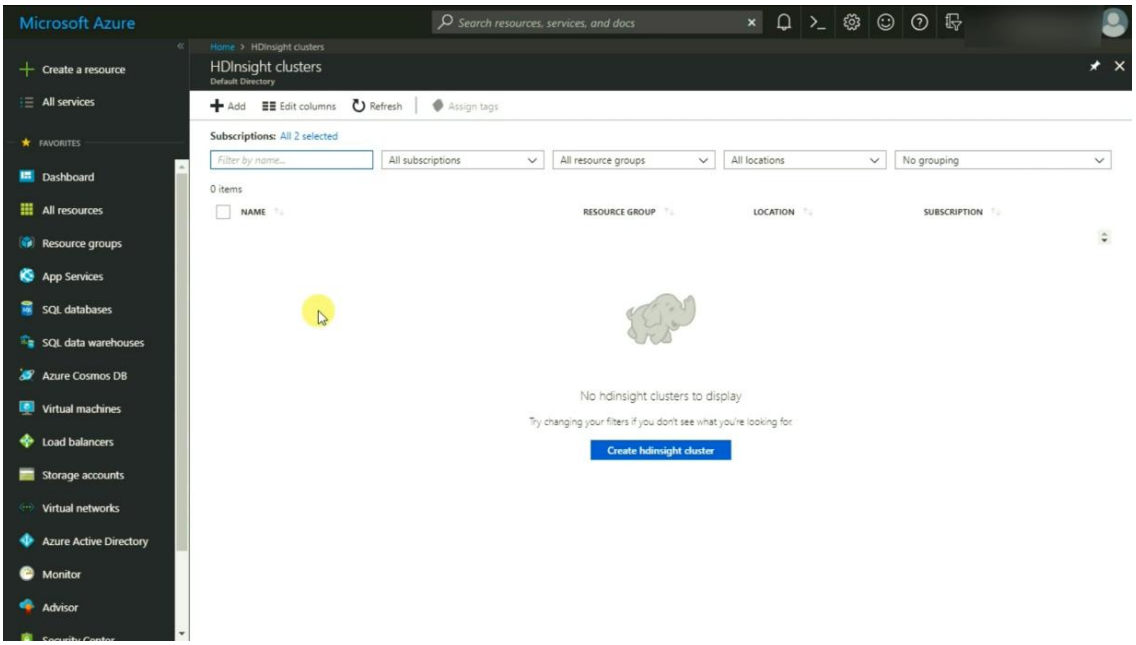
Scala is an open source language that, along with Python, can be used for the purpose of predictive analysis on the datasets. The data is first trained with a training dataset and then tested against already known data. Once a satisfactory level of training and testing has been done, the algorithm is then fed completely new data that it then predicts the needed.

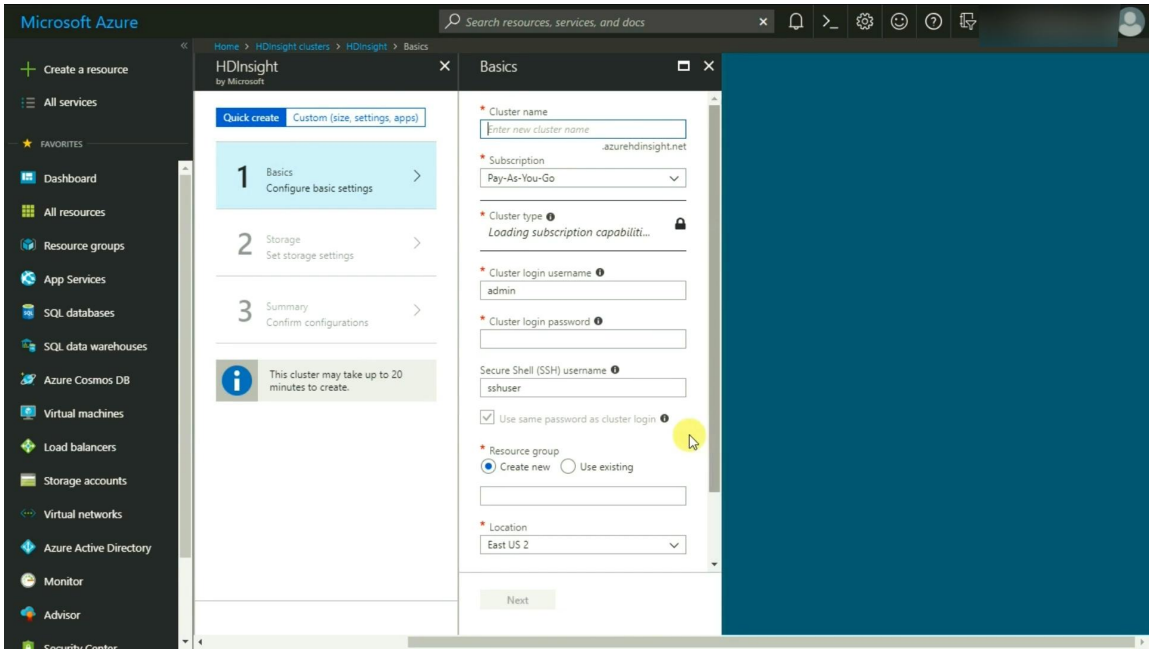
Visualization (using PowerBi)

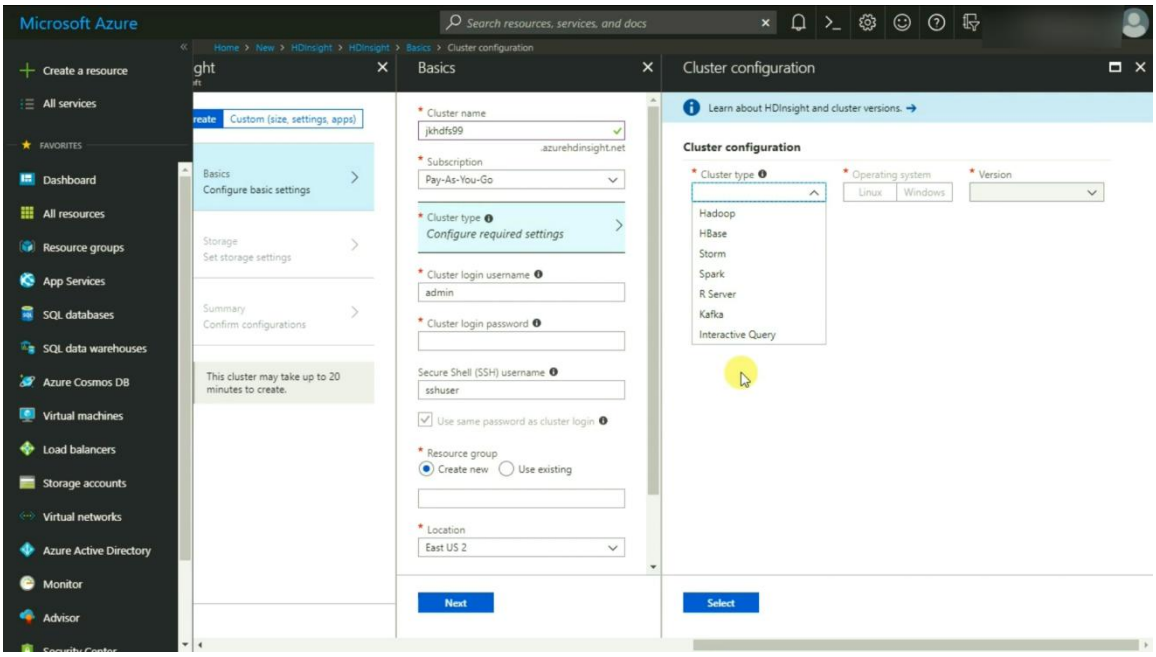
PowerBi is a data visualization tool that is provided free of cost by Microsoft. It can be used to visualize structured data such as text files with tab delimited data or data in csv formats. It is a very powerful tool that gives the user a lot of freedom when it comes to the kind of data visualization that can be used. It also enables drilling down and rolling up operations that can be used to abstract data or get into

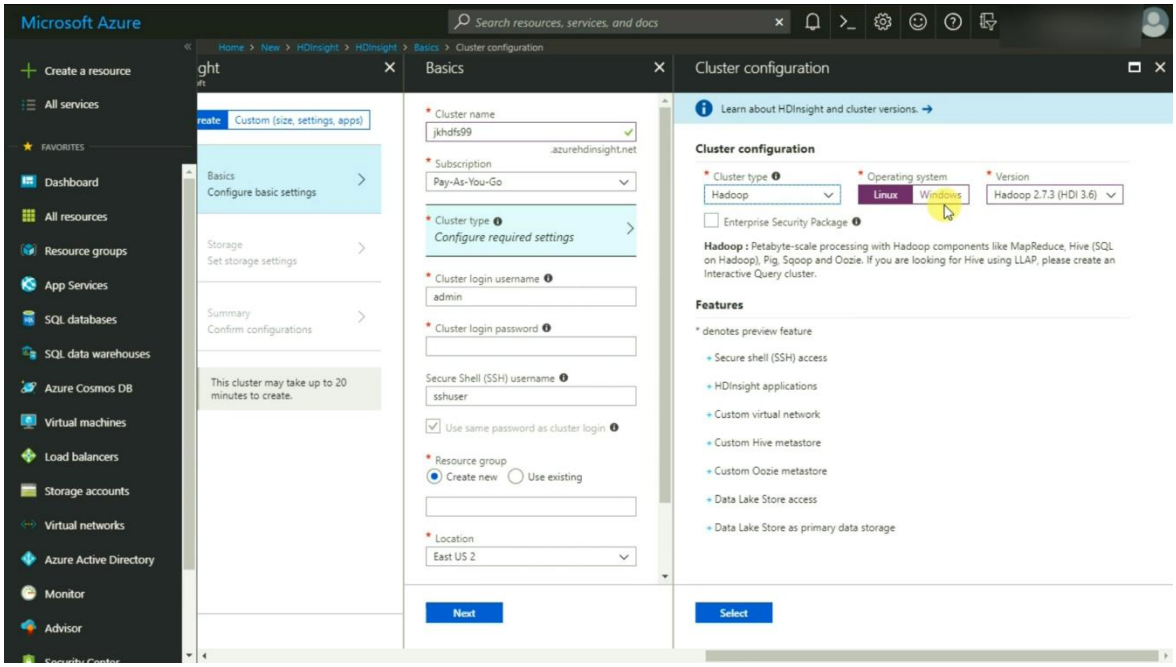
the details of the visualized data. As a result, PowerBi is the software of choice for the data visualization aspect of our project.

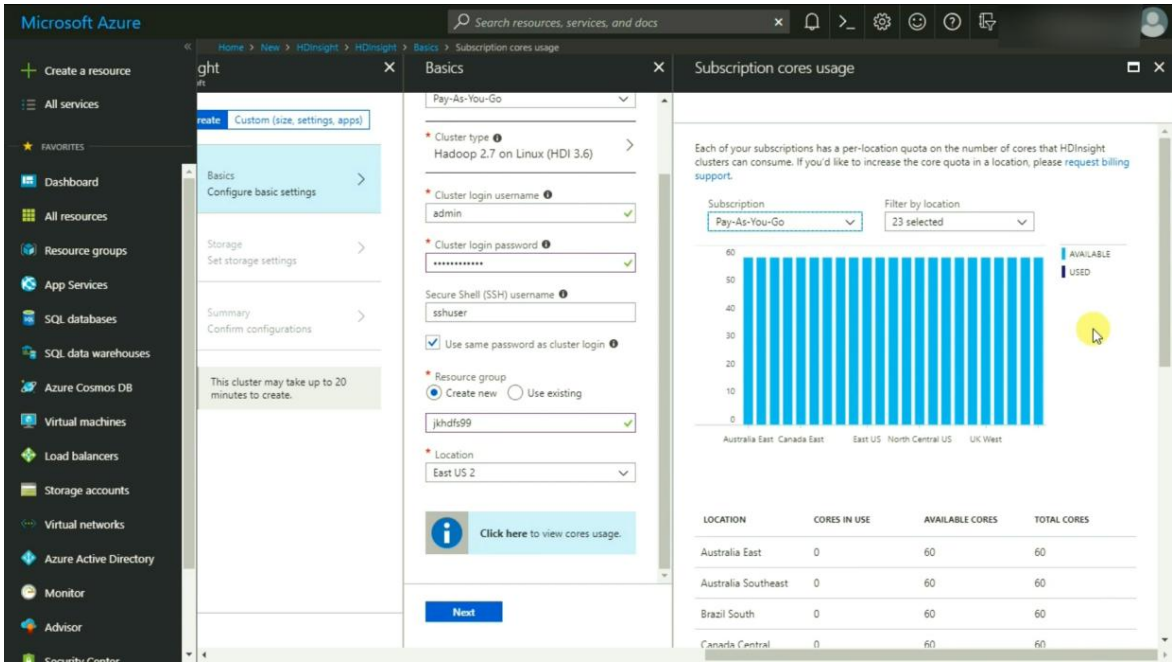
Cluster Creation

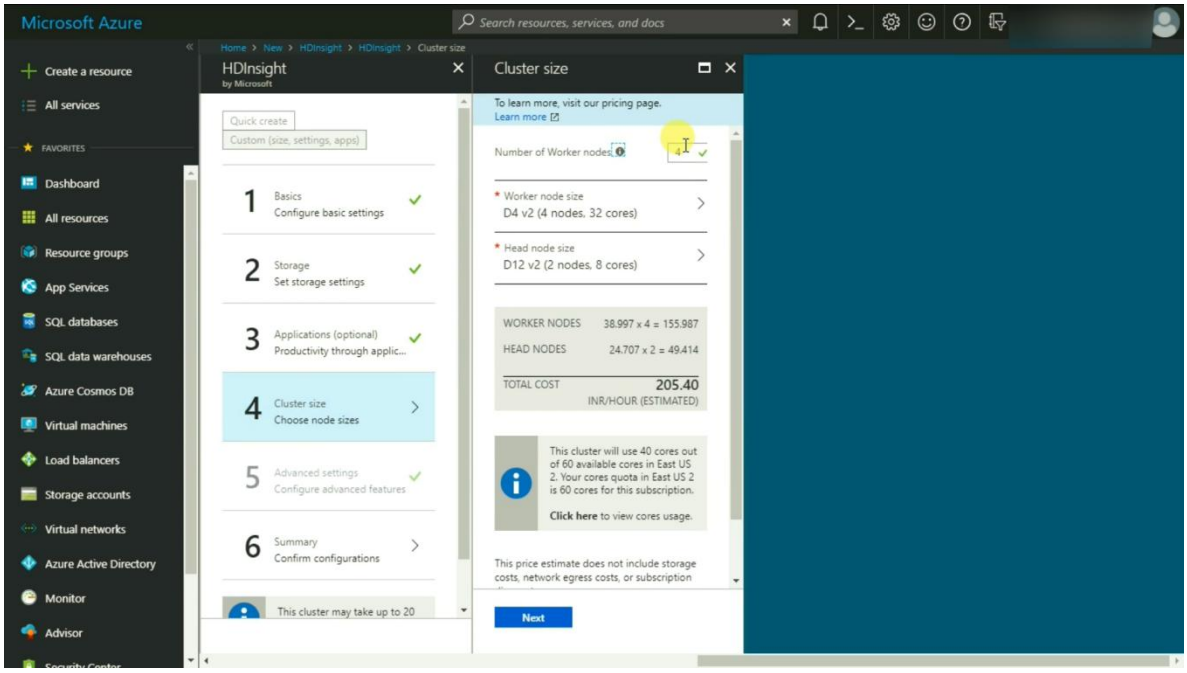


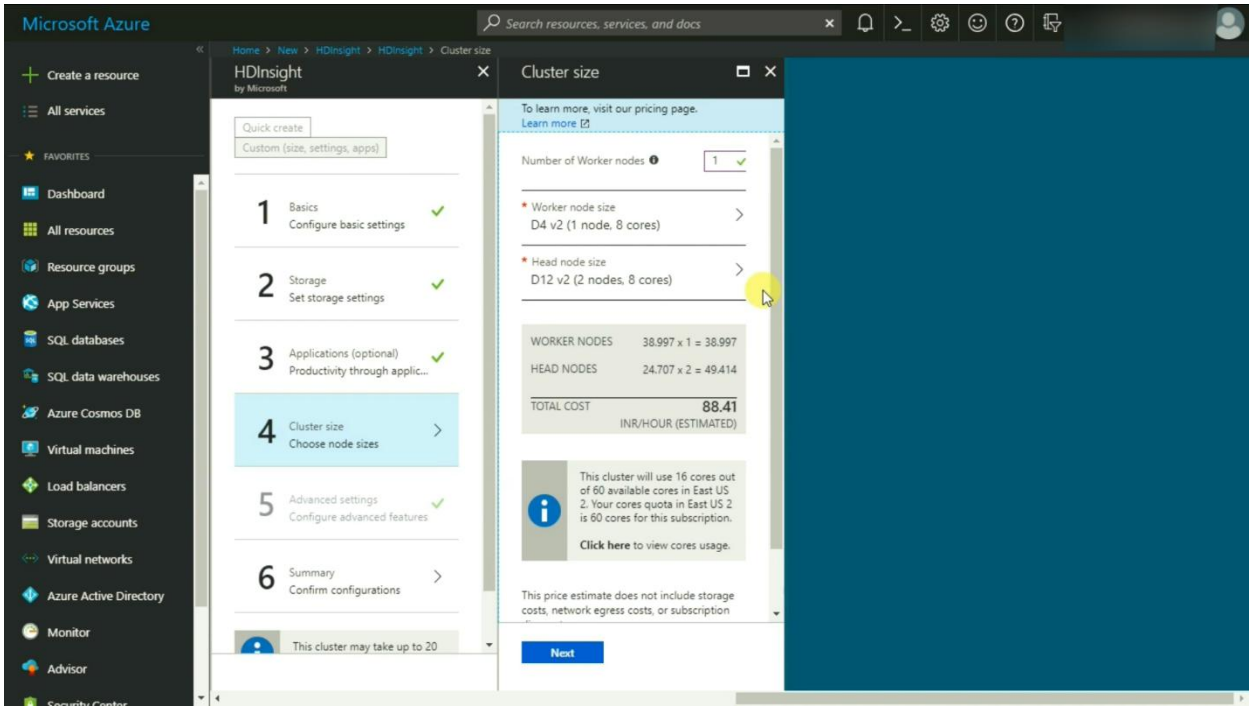


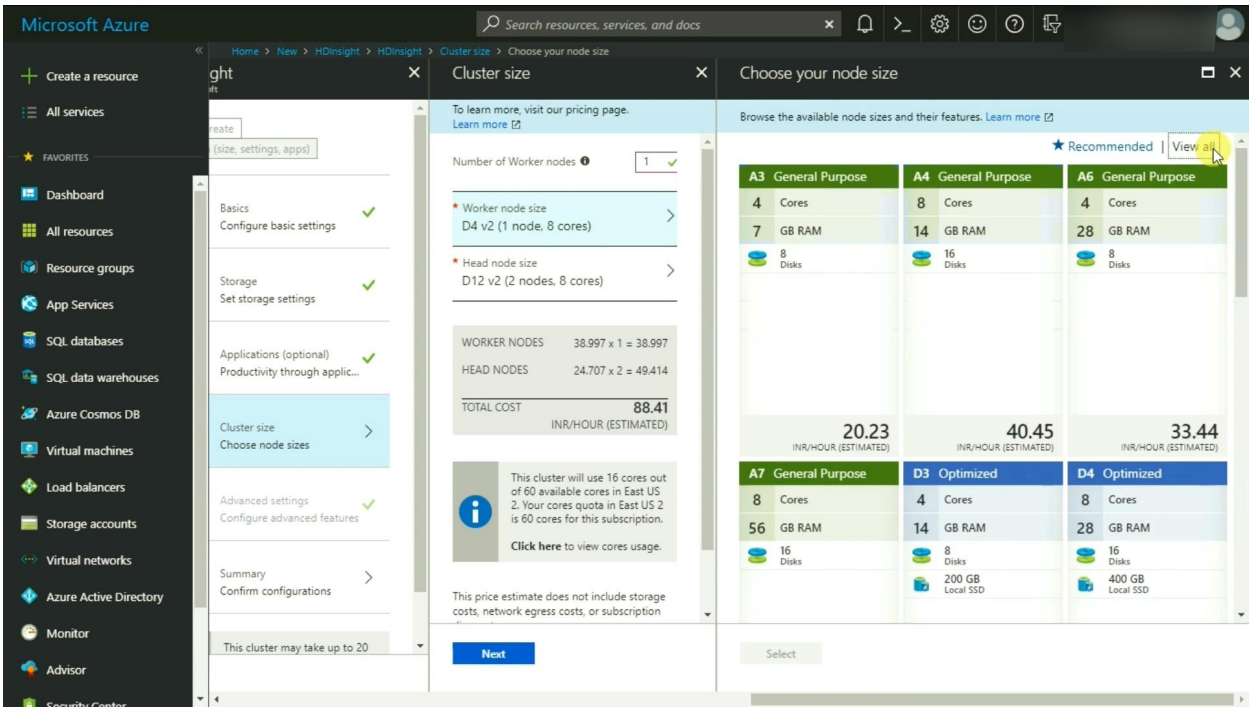

Microsoft Azure
Search resources, services, and docs
Home > New > HDInsight > HDInsight > Basics > Cluster configuration
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
SQL databases
SQL data warehouses
Azure Cosmos DB
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
Custom (size, settings, apps)
Basics
Configure basic settings
Storage
Set storage settings
Summary
Confirm configurations
This cluster may take up to 20 minutes to create.
Basics
* Cluster name
jkhdfs99
.azurehdinsight.net
* Subscription
Pay-As-You-Go
* Cluster type
Configure required settings
* Cluster login username
admin
* Cluster login password
Secure Shell (SSH) username
sshuser
Use same password as cluster login
* Resource group
Create new
Use existing
* Location
East US 2
Next
Cluster configuration
Learn about HDInsight and cluster versions.
Cluster configuration
* Cluster type
* Operating system
Linux
Windows
* Version
Hadoop
HBase
Storm
Spark
R Server
Kafka
Interactive Query
Select

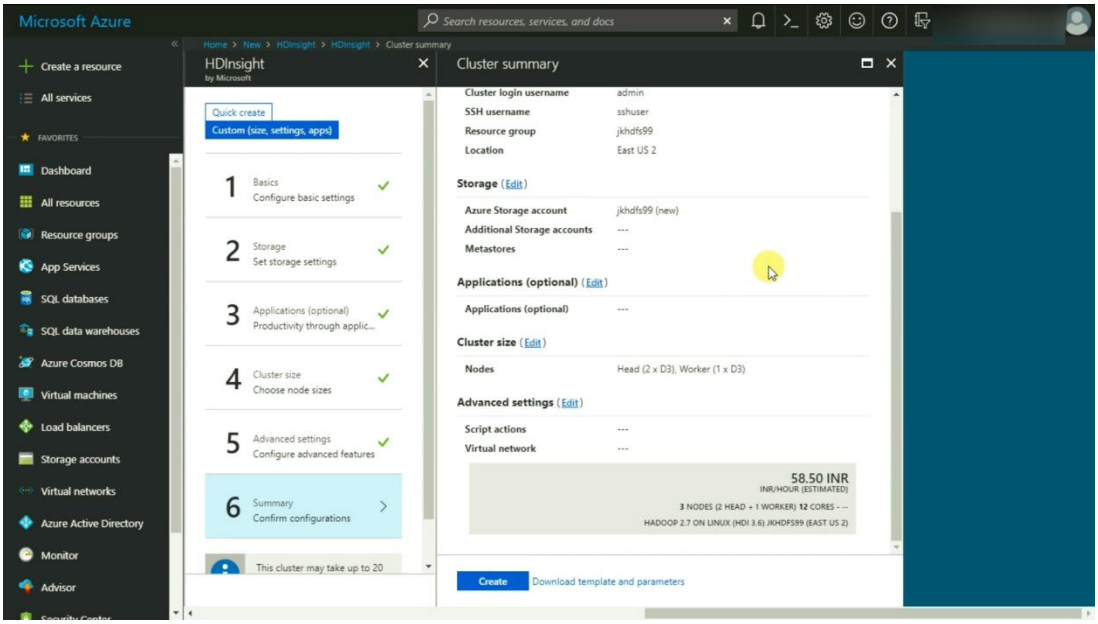

Microsoft Azure
Search resources, services, and docs
Home > New > HDInsight > HDInsight > Basics > Cluster configuration
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
SQL databases
SQL data warehouses
Azure Cosmos DB
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
Custom (size, settings, apps)
Basics
Configure basic settings
Storage
Set storage settings
Summary
Confirm configurations
This cluster may take up to 20 minutes to create.
Basics
* Cluster name
jkhdfs99
.azurehdinsight.net
* Subscription
Pay-As-You-Go
* Cluster type
Configure required settings
* Cluster login username
admin
* Cluster login password
Secure Shell (SSH) username
sshuser
Use same password as cluster login
* Resource group
Create new
Use existing
* Location
East US 2
Next
Cluster configuration
Learn about HDInsight and cluster versions.
Cluster configuration
* Cluster type
Hadoop
* Operating system
Linux
Windows
* Version
Hadoop 2.7.3 (HDI 3.6)
Enterprise Security Package
Hadoop : Petabyte-scale processing with Hadoop components like MapReduce, Hive (SQL on Hadoop), Pig, Sqoop and Oozie. If you are looking for Hive using LLAP, please create an Interactive Query cluster.
Features
* denotes preview feature
+ Secure shell (SSH) access
+ HDInsight applications
+ Custom virtual network
+ Custom Hive metastore
+ Custom Oozie metastore
+ Data Lake Store access
+ Data Lake Store as primary data storage
Select

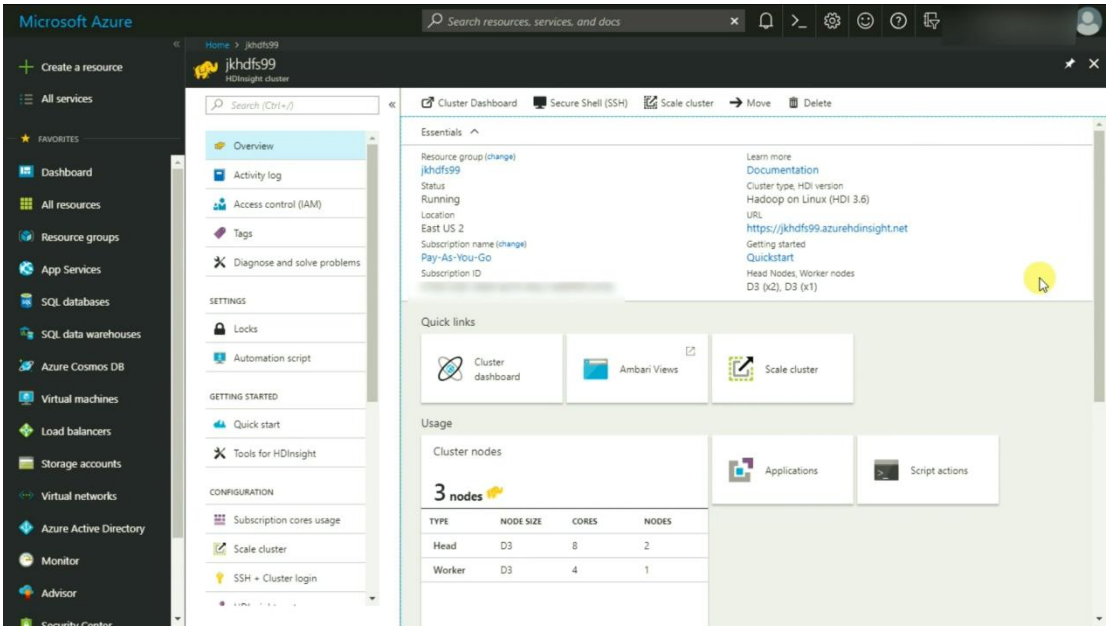

Microsoft Azure
Search resources, services, and docs
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
SQL databases
SQL data warehouses
Azure Cosmos DB
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
Home > New > HDInsight > HDInsight > Basics > Subscription cores usage
Basics
Pay-As-You-Go
Cluster type
Hadoop 2.7 on Linux (HDI 3.6)
Cluster login username
admin
Cluster login password
Secure Shell (SSH) username
sshuser
Use same password as cluster login
Resource group
Create new
Use existing
jkhdfs99
Location
East US 2
Click here to view cores usage.
Next
Basics
Configure basic settings
Storage
Set storage settings
Summary
Confirm configurations
This cluster may take up to 20 minutes to create.
Subscription cores usage
Each of your subscriptions has a per-location quota on the number of cores that HDInsight clusters can consume. If you'd like to increase the core quota in a location, please request billing support.
Subscription
Pay-As-You-Go
Filter by location
23 selected
AVAILABLE
USED
Australia East
Canada East
East US
North Central US
UK West
LOCATION
CORES IN USE
AVAILABLE CORES
TOTAL CORES
Australia East 0 60 60
Australia Southeast 0 60 60
Brazil South 0 60 60


Microsoft Azure
Search resources, services, and docs
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
SQL databases
SQL data warehouses
Azure Cosmos DB
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
Home > New > HDInsight > HDInsight > Cluster size
HDInsight
by Microsoft
Quick create
Custom (size, settings, apps)
1 Basics
Configure basic settings
2 Storage
Set storage settings
3 Applications (optional)
Productivity through applic...
4 Cluster size
Choose node sizes
5 Advanced settings
Configure advanced features
6 Summary
Confirm configurations
This cluster may take up to 20
Cluster size
To learn more, visit our pricing page.
Learn more
Number of Worker nodes
4
Worker node size
D4 v2 (4 nodes, 32 cores)
Head node size
D12 v2 (2 nodes, 8 cores)
WORKER NODES 38.997 x 4 = 155.987
HEAD NODES 24.707 x 2 = 49.414
TOTAL COST 205.40
INR/HOUR (ESTIMATED)
This cluster will use 40 cores out of 60 available cores in East US 2. Your cores quota in East US 2 is 60 cores for this subscription.
Click here to view cores usage.
This price estimate does not include storage costs, network egress costs, or subscription
Next


Microsoft Azure
Search resources, services, and docs
Home > New > HDInsight > HDInsight > Cluster size
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
SQL databases
SQL data warehouses
Azure Cosmos DB
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
HDInsight
by Microsoft
Quick create
Custom (size, settings, apps)
1 Basics
Configure basic settings
2 Storage
Set storage settings
3 Applications (optional)
Productivity through applic...
4 Cluster size
Choose node sizes
5 Advanced settings
Configure advanced features
6 Summary
Confirm configurations
This cluster may take up to 20
Cluster size
To learn more, visit our pricing page.
Learn more
Number of Worker nodes
1
* Worker node size
D4 v2 (1 node, 8 cores)
* Head node size
D12 v2 (2 nodes, 8 cores)
WORKER NODES 38.997 x 1 = 38.997
HEAD NODES 24.707 x 2 = 49.414
TOTAL COST 88.41
INR/HOUR (ESTIMATED)
This cluster will use 16 cores out of 60 available cores in East US 2. Your cores quota in East US 2 is 60 cores for this subscription.
Click here to view cores usage.
This price estimate does not include storage costs, network egress costs, or subscription
Next


Microsoft Azure
Search resources, services, and docs
Home > New > HDInsight > HDInsight > Cluster size > Choose your node size
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
SQL databases
SQL data warehouses
Azure Cosmos DB
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
Basics
Configure basic settings
Storage
Set storage settings
Applications (optional)
Productivity through applic...
Cluster size
Choose node sizes
Advanced settings
Configure advanced features
Summary
Confirm configurations
This cluster may take up to 20
Cluster size
To learn more, visit our pricing page.
Learn more
Number of Worker nodes
1
* Worker node size
D4 v2 (1 node, 8 cores)
* Head node size
D12 v2 (2 nodes, 8 cores)
WORKER NODES 38.997 x 1 = 38.997
HEAD NODES 24.707 x 2 = 49.414
TOTAL COST 88.41
INR/HOUR (ESTIMATED)
This cluster will use 16 cores out of 60 available cores in East US 2. Your cores quota in East US 2 is 60 cores for this subscription.
Click here to view cores usage.
This price estimate does not include storage costs, network egress costs, or subscription
Next
Choose your node size
Browse the available node sizes and their features. Learn more
Recommended | View all
A3 General Purpose
4 Cores
7 GB RAM
8 Disks
20.23
INR/HOUR (ESTIMATED)
A4 General Purpose
8 Cores
14 GB RAM
16 Disks
40.45
INR/HOUR (ESTIMATED)
A6 General Purpose
4 Cores
28 GB RAM
8 Disks
33.44
INR/HOUR (ESTIMATED)
A7 General Purpose
8 Cores
56 GB RAM
16 Disks
D3 Optimized
4 Cores
14 GB RAM
8 Disks
200 GB Local SSD
D4 Optimized
8 Cores
28 GB RAM
16 Disks
400 GB Local SSD
Select


Microsoft Azure
Search resources, services, and docs
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
SQL databases
SQL data warehouses
Azure Cosmos DB
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
Home > New > HDInsight > HDInsight > Cluster summary
HDInsight
by Microsoft
Quick create
Custom (size, settings, apps)
1 Basics
Configure basic settings
2 Storage
Set storage settings
3 Applications (optional)
Productivity through applic...
4 Cluster size
Choose node sizes
5 Advanced settings
Configure advanced features
6 Summary
Confirm configurations
This cluster may take up to 20
Cluster summary
Cluster login username admin
SSH username sshuser
Resource group jkhdfs99
Location East US 2
Storage (Edit)
Azure Storage account jkhdfs99 (new)
Additional Storage accounts ---
Metastores ---
Applications (optional) (Edit)
Applications (optional) ---
Cluster size (Edit)
Nodes Head (2 x D3), Worker (1 x D3)
Advanced settings (Edit)
Script actions ---
Virtual network ---
58.50 INR
INR/HOUR (ESTIMATED)
3 NODES (2 HEAD + 1 WORKER) 12 CORES - ...
HADOOP 2.7 ON LINUX (HDI 3.6) JKHDFS99 (EAST US 2)
Create
Download template and parameters


Microsoft Azure
Search resources, services, and docs
Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
App Services
SQL databases
SQL data warehouses
Azure Cosmos DB
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
Home > jkhdfs99
jkhdfs99
HDInsight cluster
Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
SETTINGS
Locks
Automation script
GETTING STARTED
Quick start
Tools for HDInsight
CONFIGURATION
Subscription cores usage
Scale cluster
SSH + Cluster login
Cluster Dashboard
Secure Shell (SSH)
Scale cluster
Move
Delete
Essentials
Resource group (change) jkhdfs99
Status Running
Location East US 2
Subscription name (change) Pay-As-You-Go
Subscription ID
Learn more
Documentation
Cluster type, HDI version
Hadoop on Linux (HDI 3.6)
URL
https://jkhdfs99.azurehdinsight.net
Getting started
Quickstart
Head Nodes, Worker nodes
D3 (x2), D3 (x1)
Quick links
Cluster dashboard
Ambari Views
Scale cluster
Usage
Cluster nodes
3 nodes
Applications
Script actions
TYPE NODE SIZE CORES NODES
Head D3 8 2
Worker D3 4 1