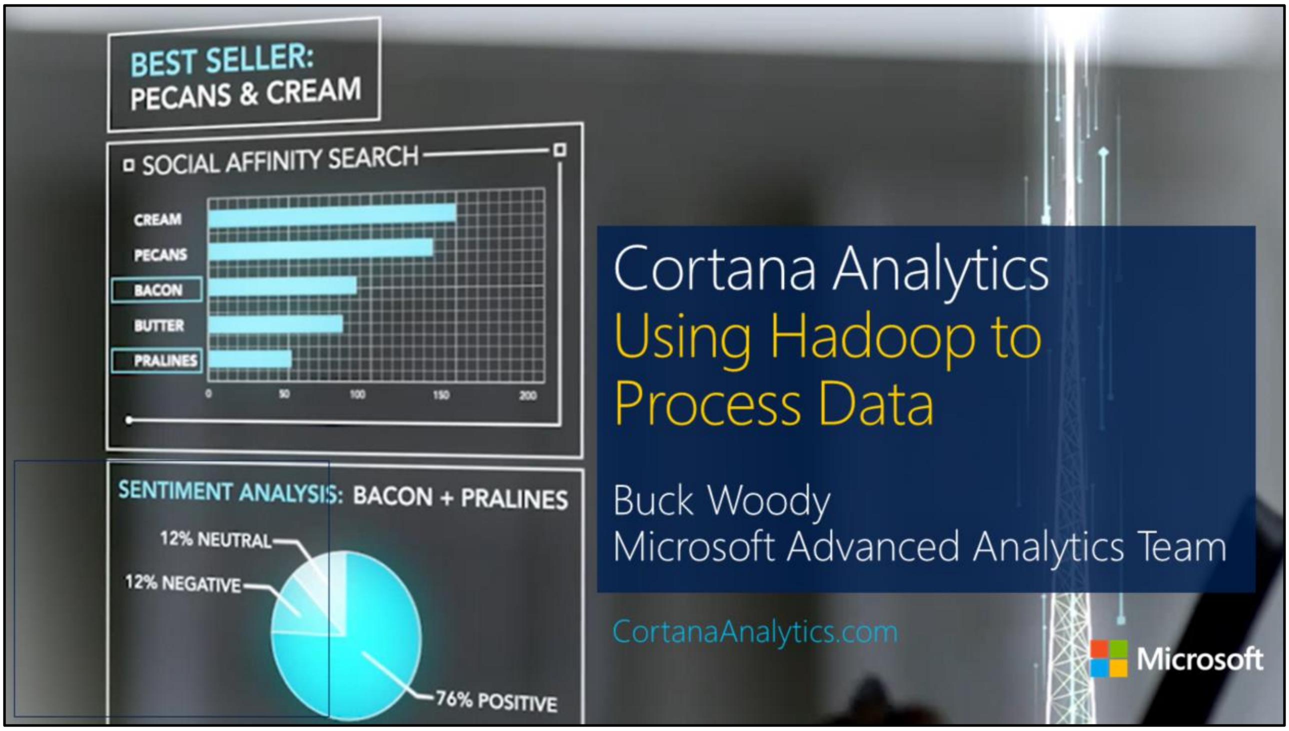


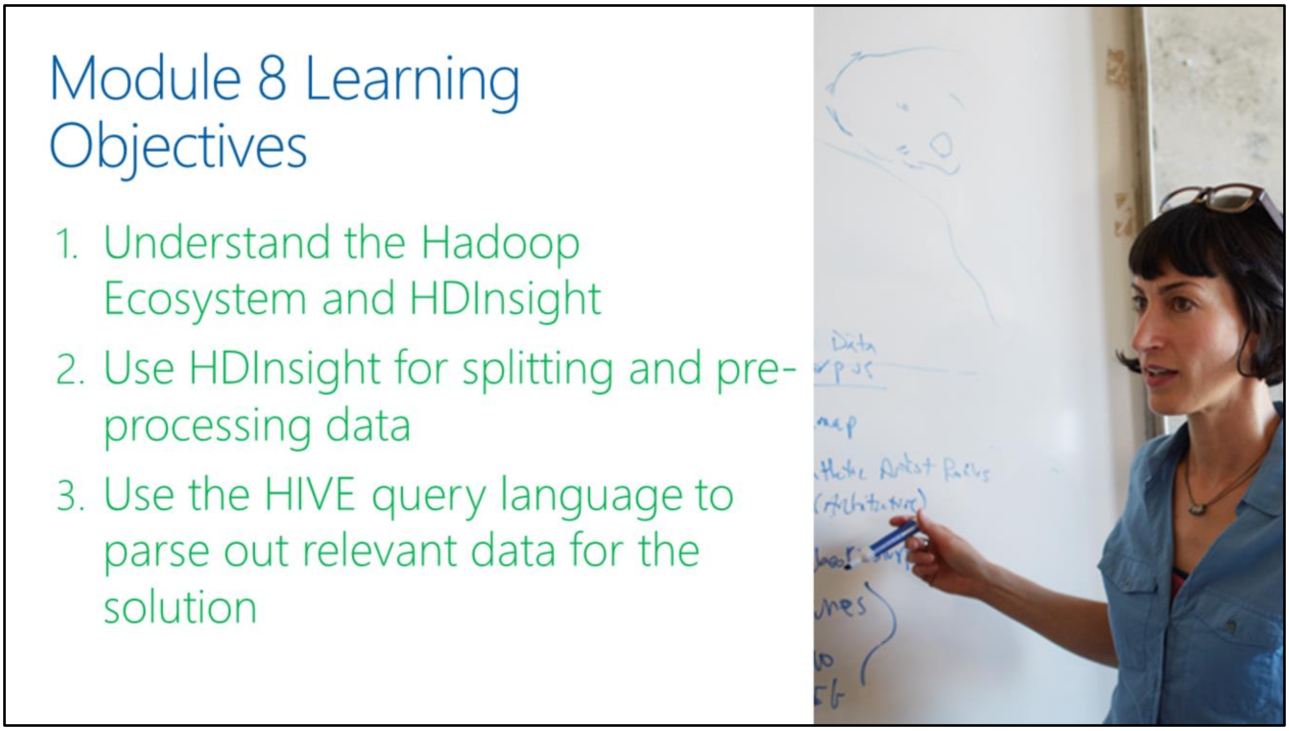
Cortana Analytics
Using Hadoop to Process Data

Buck Woody
Microsoft Advanced Analytics Team

CortanaAnalytics.com

Microsoft

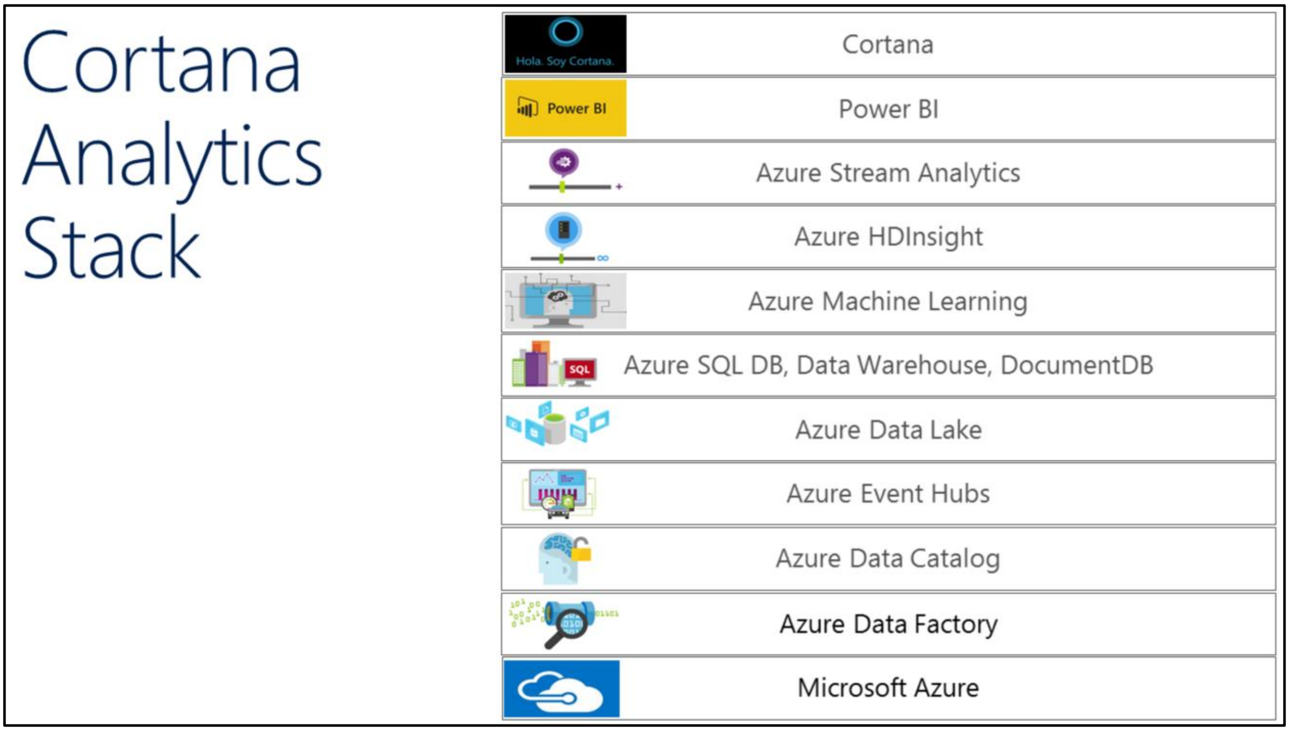
1. Main page: http://cortanaanalytics.com
2. Before you take this module, you should be able to:
   1. Understand the process for using Azure Data Factory
   2. Use Azure Data Factory to ingest data
   3. Use Azure Data Factory to leave data on prem
   4. Use Azure Data Factory to call functions to clean and shape data
   5. Use Azure Data Factory to compute analytics
   6. Use Azure Data Factory to move data to other data stores



Module 8 Learning Objectives

1. Understand the Hadoop Ecosystem and HDInsight
2. Use HDInsight for splitting and pre-processing data
3. Use the HIVE query language to parse out relevant data for the solution

1. When you finish this module, you will be able to:
   1. Understand the Hadoop Ecosystem and HDInsight
   2. Use HDInsight for splitting and pre-processing data
   3. Use the HIVE query language to parse out relevant data for the solution



Cortana Analytics Stack

Cortana

Power BI

Azure Stream Analytics

Azure HDInsight

Azure Machine Learning

Azure SQL DB, Data Warehouse, DocumentDB

Azure Data Lake

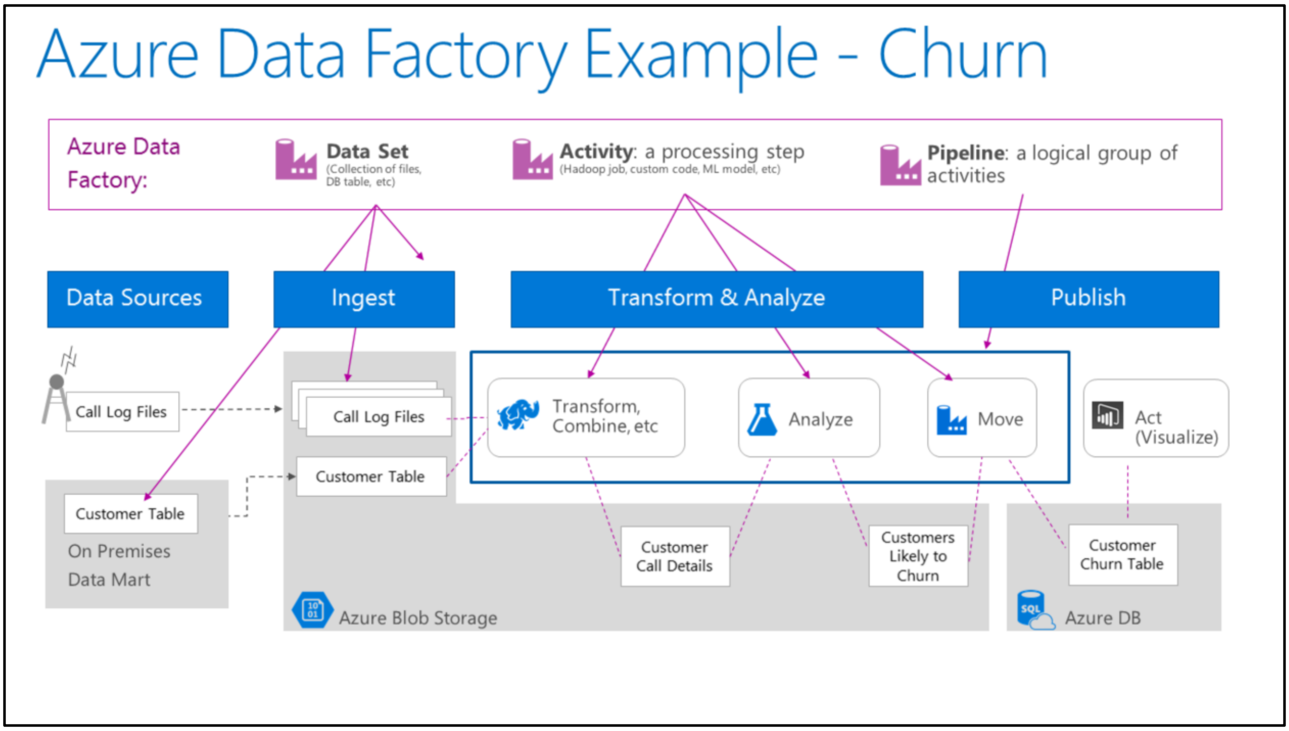
Azure Event Hubs

Azure Data Catalog

Azure Data Factory

Microsoft Azure

1. Platform and Storage: Microsoft Azure – http://microsoftazure.com Storage: https://azure.microsoft.com/en-us/documentation/services/storage/ (Host It)
2. Azure Data Factory: http://azure.microsoft.com/en-us/services/data-factory/ (Move It)
3. Azure Data Catalog: http://azure.microsoft.com/en-us/services/data-catalog (Doc It)
4. Azure Event Hubs: http://azure.microsoft.com/en-us/services/event-hubs/ (Bring It)
5. Azure Data Lake: http://azure.microsoft.com/en-us/campaigns/data-lake/ (Store It)
6. Azure DocumentDB: https://azure.microsoft.com/en-us/services/documentdb/?WT.srch=1&WT.mc_ID=SEM_JQ3fO8dU , Azure SQL Data Warehouse: http://azure.microsoft.com/en-us/services/sql-data-warehouse/ (Relate It)
7. Azure Machine Learning: http://azure.microsoft.com/en-us/services/machine-learning/ (Learn It)
8. Azure HDInsight: http://azure.microsoft.com/en-us/services/hdinsight/ (Big It)
9. Azure Stream Analytics: http://azure.microsoft.com/en-us/services/stream-analytics/ (Stream It)
10. Power BI: https://powerbi.microsoft.com/ (See It)
11. Cortana: http://blogs.windows.com/buildingapps/2014/09/23/cortana-integration-and-speech-recognition-new-code-samples/ and https://blogs.windows.com/buildingapps/2015/08/25/using-cortana-to-interact-with-your-customers-10-by-10/ (Say It)



Azure Data Factory Example - Churn

Azure Data Factory:

Data Set (Collection of files, DB table, etc)

Activity: a processing step (Hadoop job, custom code, ML model, etc)

Pipeline: a logical group of activities

Data Sources

Ingest

Transform & Analyze

Publish

Call Log Files

Customer Table

On Premises Data Mart

Call Log Files

Customer Table

Transform, Combine, etc

Analyze

Move

Act (Visualize)

Customer Call Details

Customers Likely to Churn

Customer Churn Table

Azure Blob Storage

Azure DB

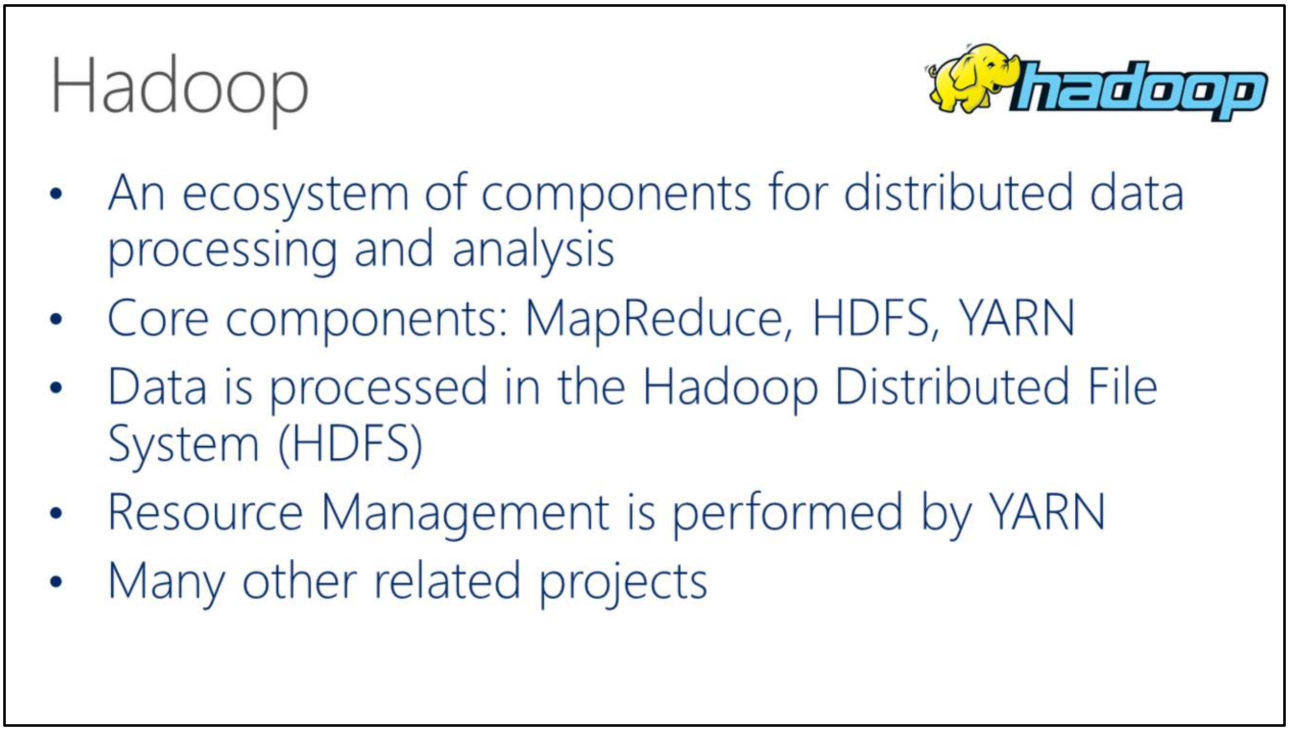
1. Full explanation of this example: https://azure.microsoft.com/en-us/blog/getting-started-with-azure-data-factory-and-azure-machine-learning-4/
2. Why use Hadoop? http://redmonk.com/sogrady/2011/01/13/apache-hadoop/

Hadoop and HDInsight

hadoop
Hortonworks Powers
Microsoft HDInsight
Microsoft

Using the Hadoop Ecosystem to process and query data

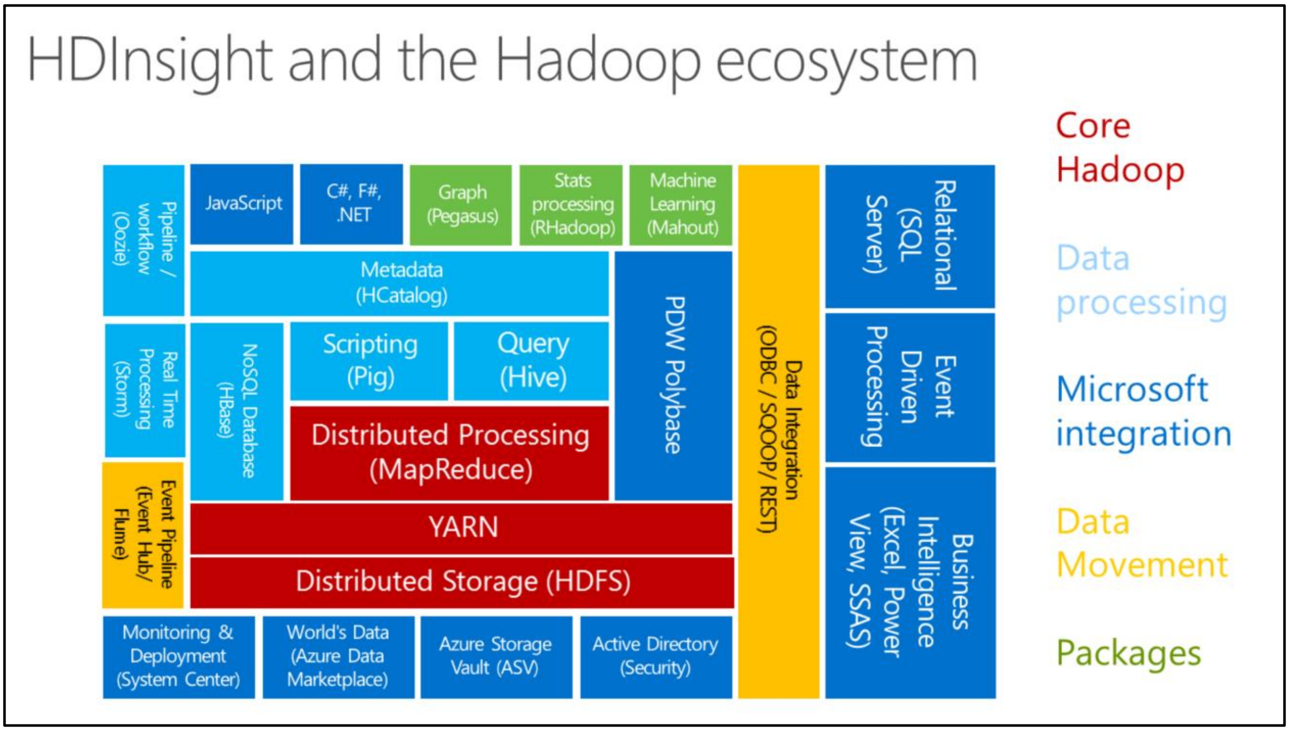
1. Primary site: https://azure.microsoft.com/en-us/services/hdinsight/
2. Quick overview: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/
3. 4-week online course through the edX platform: https://www.edx.org/course/processing-big-data-azure-hdinsight-microsoft-dat202-1x
4. 11 minute introductory video: https://channel9.msdn.com/Series/Getting-started-with-Windows-Azure-HDInsight-Service/Introduction-To-Windows-Azure-HDInsight-Service
5. Microsoft Virtual Academy Training (4 hours) - https://mva.microsoft.com/en-US/training-courses/big-data-analytics-with-hdinsight-hadoop-on-azure-10551?l=UJ7MAv97_5804984382
6. Learning path for HDInsight: https://azure.microsoft.com/en-us/documentation/learning-paths/hdinsight-self-guided-hadoop-training/
7. Azure Feature Pack for SQL Server 2016, i.e., SSIS (SQL Server Integration Services): https://msdn.microsoft.com/en-us/library/mt146770(v=sql.130).aspx



Hadoop

- An ecosystem of components for distributed data processing and analysis
- Core components: MapReduce, HDFS, YARN
- Data is processed in the Hadoop Distributed File System (HDFS)
- Resource Management is performed by YARN
- Many other related projects

- Primary/head document: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/
- For more information about Hadoop, visit the apache foundation site: http://hadoop.apache.org/



HDInsight and the Hadoop ecosystem

Pipeline / workflow (Oozie)
JavaScript
C#, F#, .NET
Graph (Pegasus)
Stats processing (RHadoop)
Machine Learning (Mahout)
Metadata (HCatalog)
Real Time Processing (Storm)
NoSQL Database (HBase)
Scripting (Pig)
Query (Hive)
Distributed Processing (MapReduce)
PDW Polybase
Data Integration (ODBC / SQOOP/ REST)
Relational (SQL Server)
Event Driven Processing
Event Pipeline (Event Hub/ Flume)
YARN
Distributed Storage (HDFS)
Business Intelligence (Excel, Power View, SSAS)
Monitoring & Deployment (System Center)
World's Data (Azure Data Marketplace)
Azure Storage Vault (ASV)
Active Directory (Security)
Core Hadoop
Data processing
Microsoft integration
Data Movement
Packages

HBase is non relational database written in Java. It uses HDFS as its file system. This database is used in scenarios where we need to store sparse data (small information found within large data; e.g., finding 50 largest items in 2 billion objects). Facebook's messaging platform uses HBase database.

Pig is a high level platform for creating map-reduce jobs for Hadoop. It uses a language called PigLatin which is very "similar" to SQL. It can be extended further to be used from other programming languages like JavaScript. Developed by Yahoo and then moved to Apache S/W foundation in 2006.

Hive is developed by Facebook and it helps in providing BI capabilities on top of Hadoop. We can analyze Hadoop's HDFS data using Hive's querying language called HiveQL which is also similar to SQL in syntax.

HCatalog is a storage management layer for Hadoop. Basically it helps presenting HDFS data in relational format.

Polybase allows SQL Server PDW users to execute queries against data stored in Hadoop's HDFS. So, as an example, users can fire queries that join data between HDFS and PDW tables!

Mahout is a set of machine learning algorithms that can use Hadoop for processing.

Oozie is a server-based Workflow Engine specialized in running workflow jobs with actions that run Hadoop MapReduce and Pig jobs.

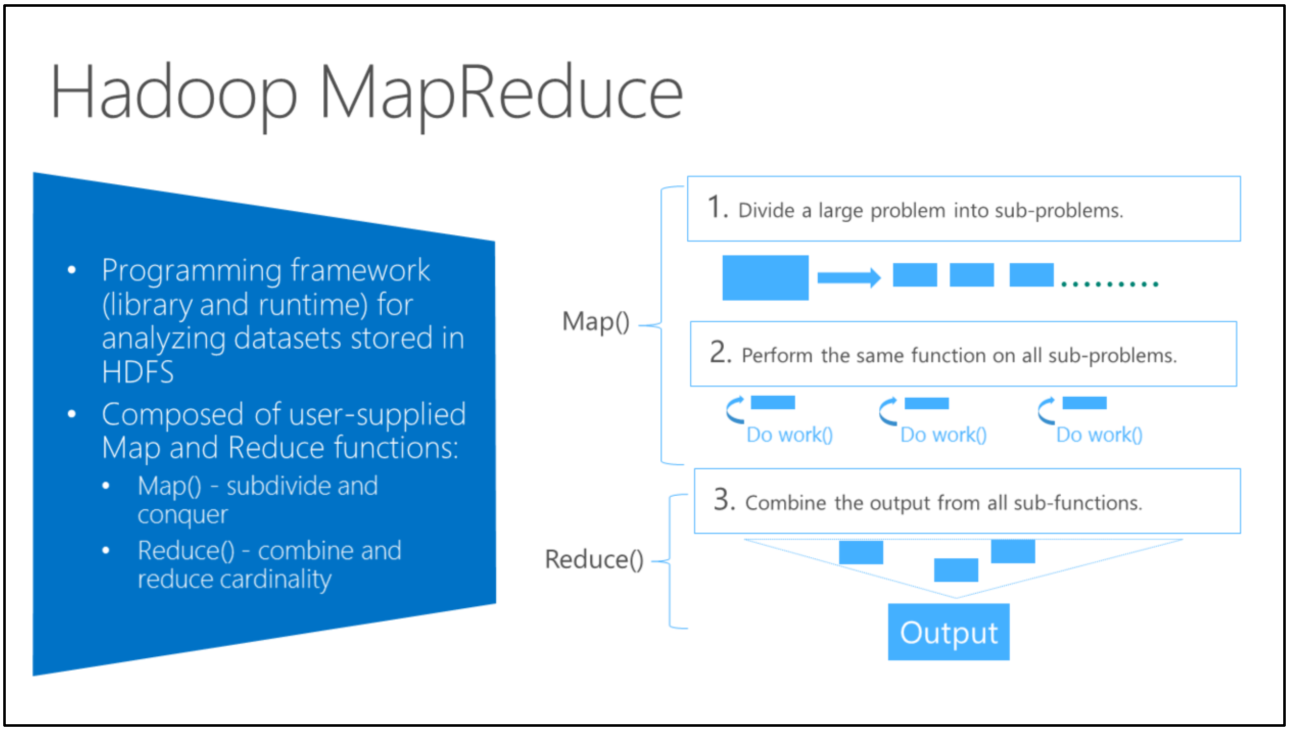
Flume is used to collect, aggregate and move large amount of log data.

Sqoop is used to transfer data between relational and Hadoop. Microsoft uses Sqoop based connector to move data between SQL Server and Hadoop.

Hive ODBC driver is used to connect different Microsoft products to connect to Hive which in turn provide connectivity with Hadoop. Products like Excel, Power Pivot, SharePoint Insights, SSRS, SSAS, etc.

REST Products can connect to Hadoop's HDFS using REST APIs as well. A common requirement for this scenario is from small devices or from apps running outside a Hadoop cluster where Hadoop's native programming is not available. This REST connectivity is possible through implementation of WebHDFS.

Microsoft's HDInsight clusters can be managed using **System Center management pack for HDInsight**



Hadoop MapReduce

- Programming framework (library and runtime) for analyzing datasets stored in HDFS
- Composed of user-supplied Map and Reduce functions:
  - Map() - subdivide and conquer
  - Reduce() - combine and reduce cardinality

Map()

1. Divide a large problem into sub-problems.

2. Perform the same function on all sub-problems.

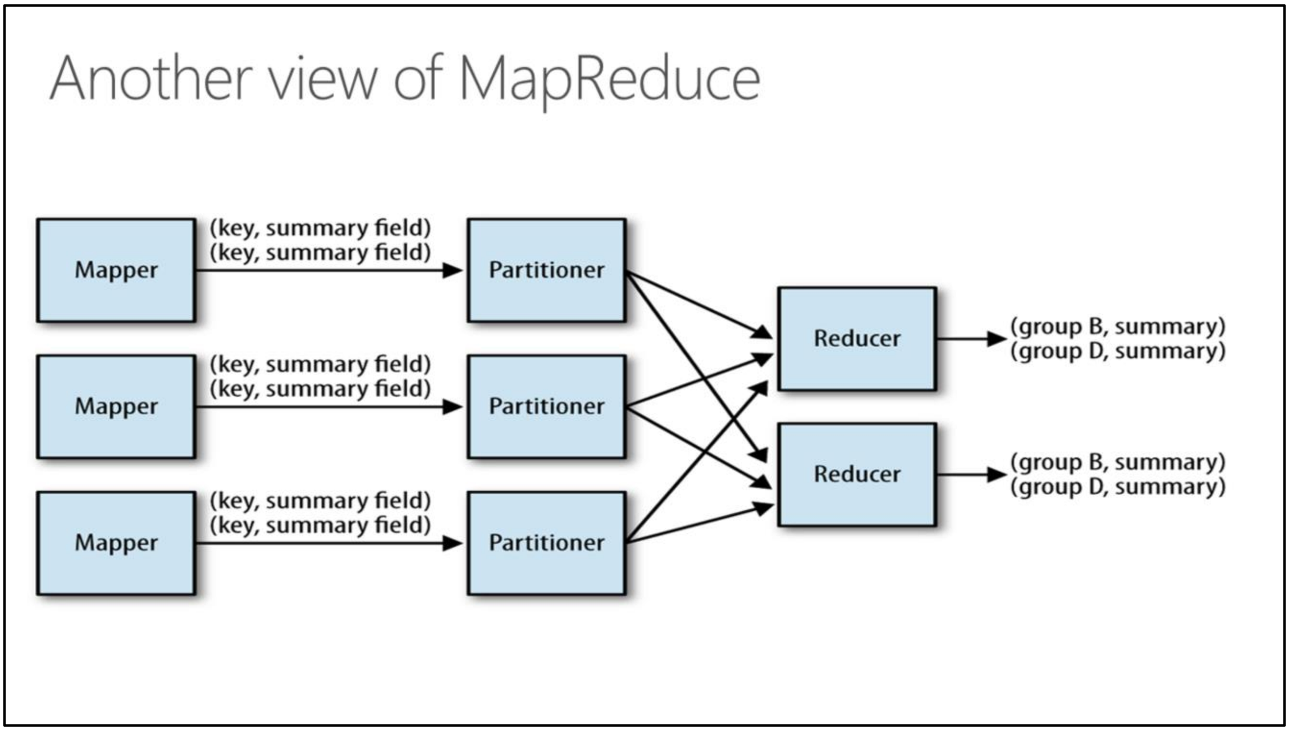
Do work() Do work() Do work()

Reduce()

3. Combine the output from all sub-functions.

Output

- Using MapReduce with HDInsight- https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-mapreduce/
- Hadoop Streaming: http://hadoop.apache.org/docs/r1.2.1/streaming.html
- Using MapReduce with HDInsight: http://www.windowsazure.com/en-us/manage/services/hdinsight/using-mapreduce-with-hdinsight/



Another view of MapReduce

Mapper

(key, summary field)
(key, summary field)

Partitioner

Mapper

(key, summary field)
(key, summary field)

Partitioner

Mapper

(key, summary field)
(key, summary field)

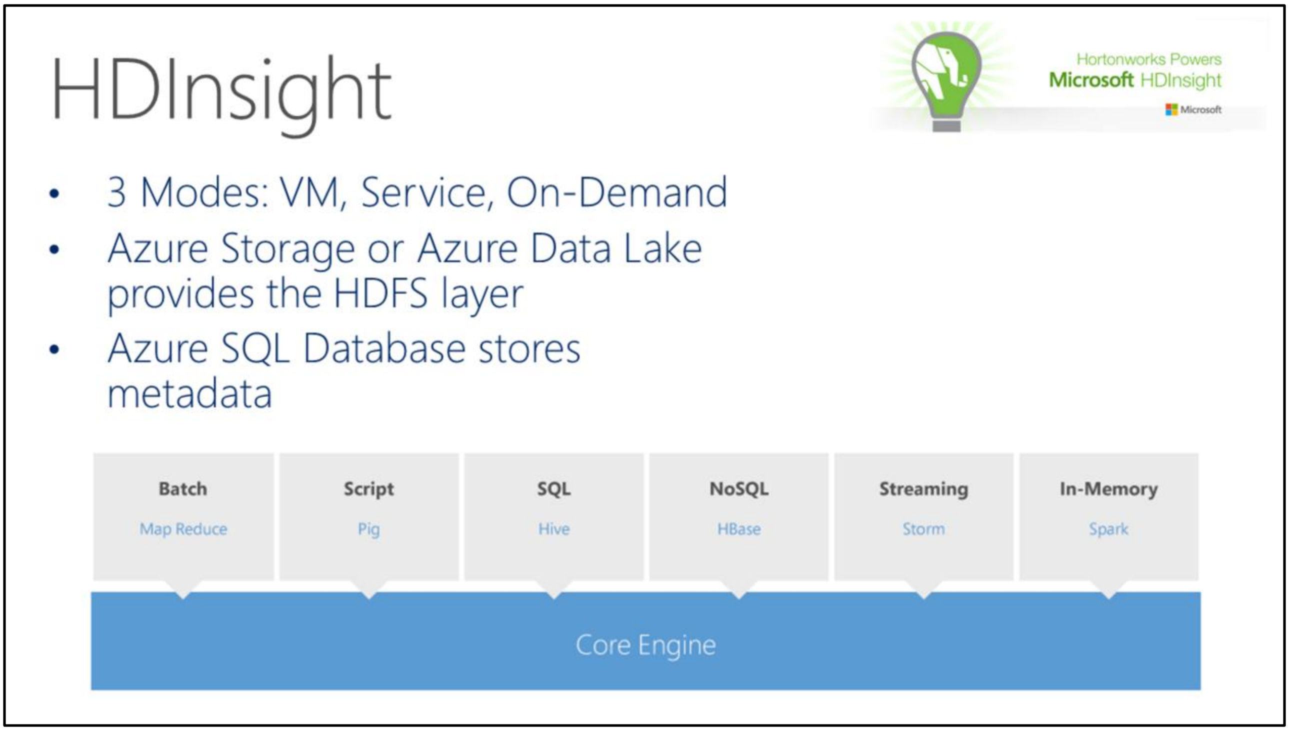
Partitioner

Reducer

(group B, summary)
(group D, summary)

Reducer

(group B, summary)
(group D, summary)



HDInsight

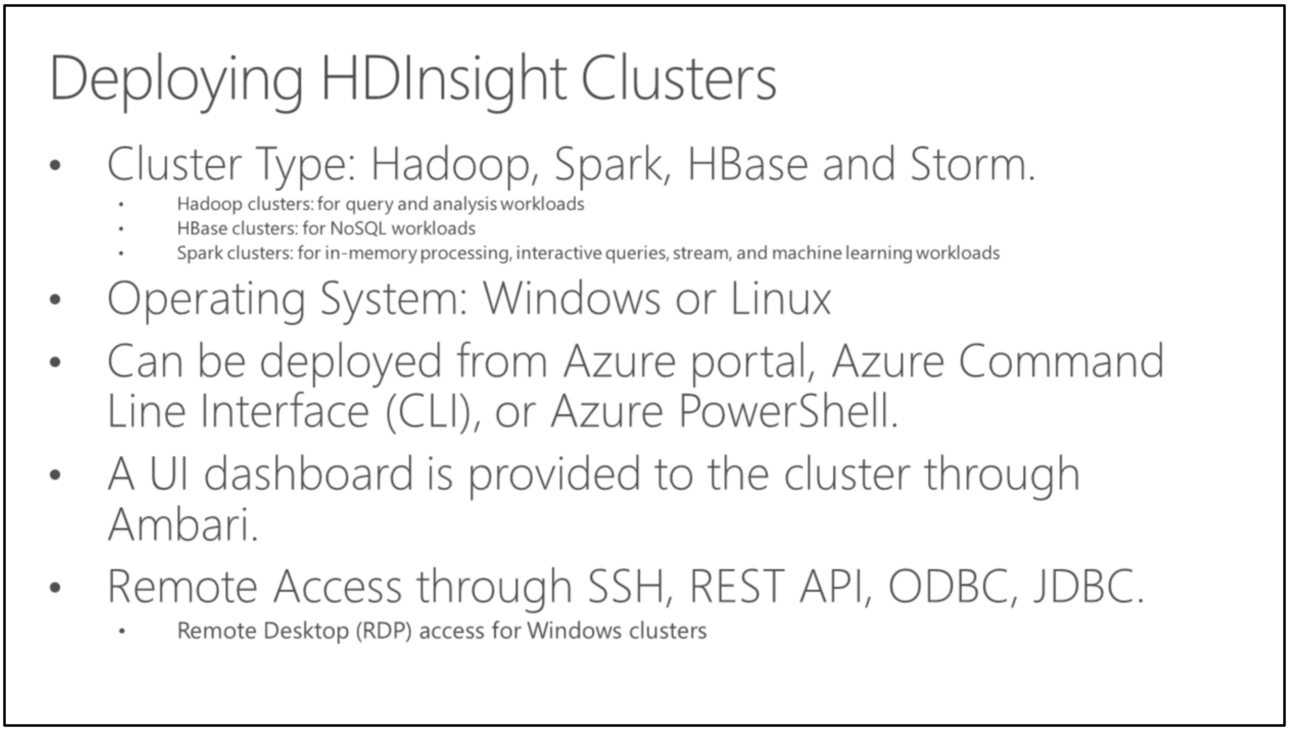
Hortonworks Powers
Microsoft HDInsight
Microsoft

- 3 Modes: VM, Service, On-Demand
- Azure Storage or Azure Data Lake provides the HDFS layer
- Azure SQL Database stores metadata

Batch	Script	SQL	NoSQL	Streaming	In-Memory
Map Reduce	Pig	Hive	HBase	Storm	Spark

Core Engine

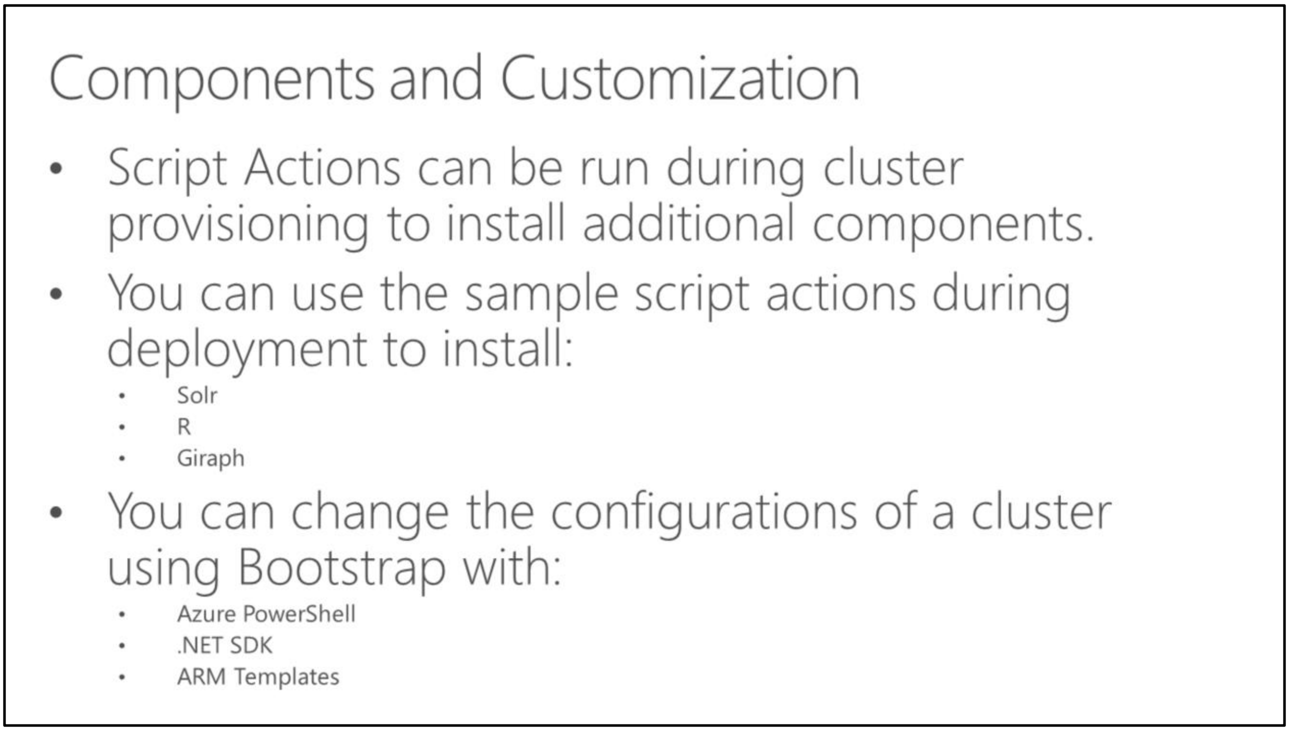
- Main page: https://azure.microsoft.com/en-us/documentation/services/hdinsight/
- Pricing for HDInsight: https://azure.microsoft.com/en-us/pricing/details/hdinsight/
- On demand HDInsight cluster: https://azure.microsoft.com/en-us/documentation/articles/data-factory-compute-linked-services/#azure-hdinsight-on-demand-linked-service



Deploying HDInsight Clusters

- Cluster Type: Hadoop, Spark, HBase and Storm.
  - Hadoop clusters: for query and analysis workloads
  - HBase clusters: for NoSQL workloads
  - Spark clusters: for in-memory processing, interactive queries, stream, and machine learning workloads
- Operating System: Windows or Linux
- Can be deployed from Azure portal, Azure Command Line Interface (CLI), or Azure PowerShell.
- A UI dashboard is provided to the cluster through Ambari.
- Remote Access through SSH, REST API, ODBC, JDBC.
  - Remote Desktop (RDP) access for Windows clusters

- Azure Portal: azure.portal.com
- Provisioning Clusters: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-provision-clusters/
- Different clusters have different node types, number of nodes, and node sizes.

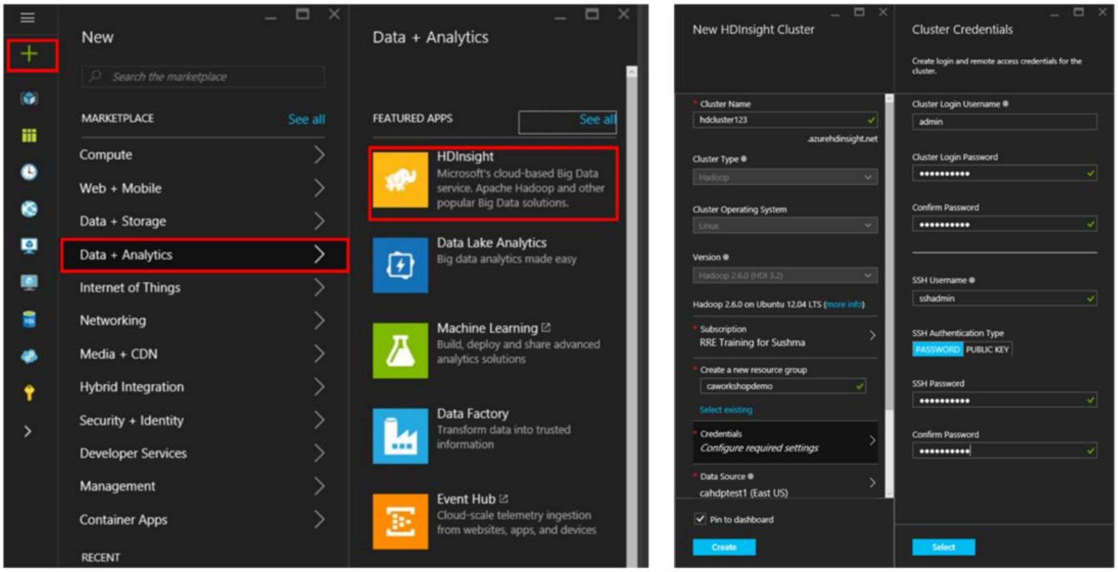


Components and Customization

- Script Actions can be run during cluster provisioning to install additional components.
- You can use the sample script actions during deployment to install:
  - Solr
  - R
  - Giraph
- You can change the configurations of a cluster using Bootstrap with:
  - Azure PowerShell
  - .NET SDK
  - ARM Templates

- Examples of script action scripts: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-customize-cluster-linux/
- Learn how to write Script Action scripts for HDInsight: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-script-actions/
- Customize HDInsight clusters using Bootstrap: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-customize-cluster-bootstrap/
- What's included in HDInsight and Supported Versions: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-component-versioning/

Provisioning with Azure Portal



New HDInsight Cluster

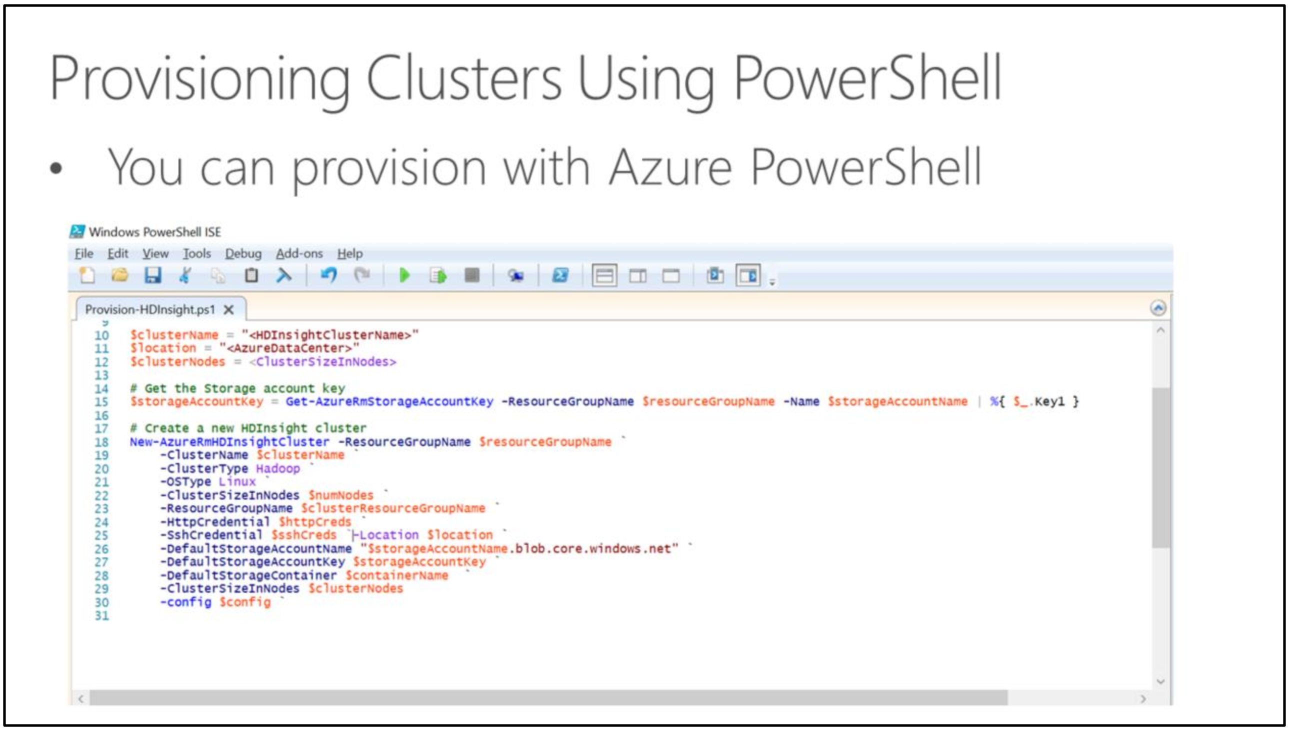
- Cluster Name: hdcluster123 .azurehdinsight.net
- Cluster Type: Hadoop
- Cluster Operating System: Linux
- Version: Hadoop 2.6.0 (HDI 3.2)
- Hadoop 2.6.0 on Ubuntu 12.04 LTS (more info)
- Subscription: RRE Training for Sushma
- Create a new resource group: caworkshopdemo
- Select existing
- Credentials: *Configure required settings*
- Data Source: cahdptest1 (East US)
- ☑ Pin to dashboard

Cluster Credentials

Create login and remote access credentials for the cluster.

- Cluster Login Username: admin
- Cluster Login Password: ••••••••••
- Confirm Password: ••••••••••
- SSH Username: sshadmin
- SSH Authentication Type: [✓ PASSWORD] PUBLIC KEY
- SSH Password: ••••••••••
- Confirm Password: ••••••••••

- **Cluster OS**: Windows (Windows Server 2012 R2 Datacenter) or Linux (Ubuntu 12.04 LTS for Linux)
- **Linux document**: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-provision-linux-clusters/
  - Great if you're familiar with linux or want easy integration with Hadoop ecosystem components built for linux
- **Windows document**: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-provision-linux-clusters/
- If you pick **windows**, you can *access the cluster via remote desktop*
- For **linux** clusters, you can use *ssh*.
  - you will need an **ssh client**, eg., *putty* if you're on windows: http://www.putty.org/
  - Can use a ssh key for credentials, or a password: https://azure.microsoft.com/en-us/documentation/articles/virtual-machines-linux-use-ssh-key/
- **Use SSH from a Linux/Unix or OS X machine**: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-linux-use-ssh-unix/
- **Use SSH from a Windows Machine**: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-linux-use-ssh-windows/
- Attach relevant resource groups or create a new one
- The cluster and default data source *must* be in the **same region**
- Deployment takes ~ 20 – 30 minutes



Provisioning Clusters Using PowerShell

- You can provision with Azure PowerShell

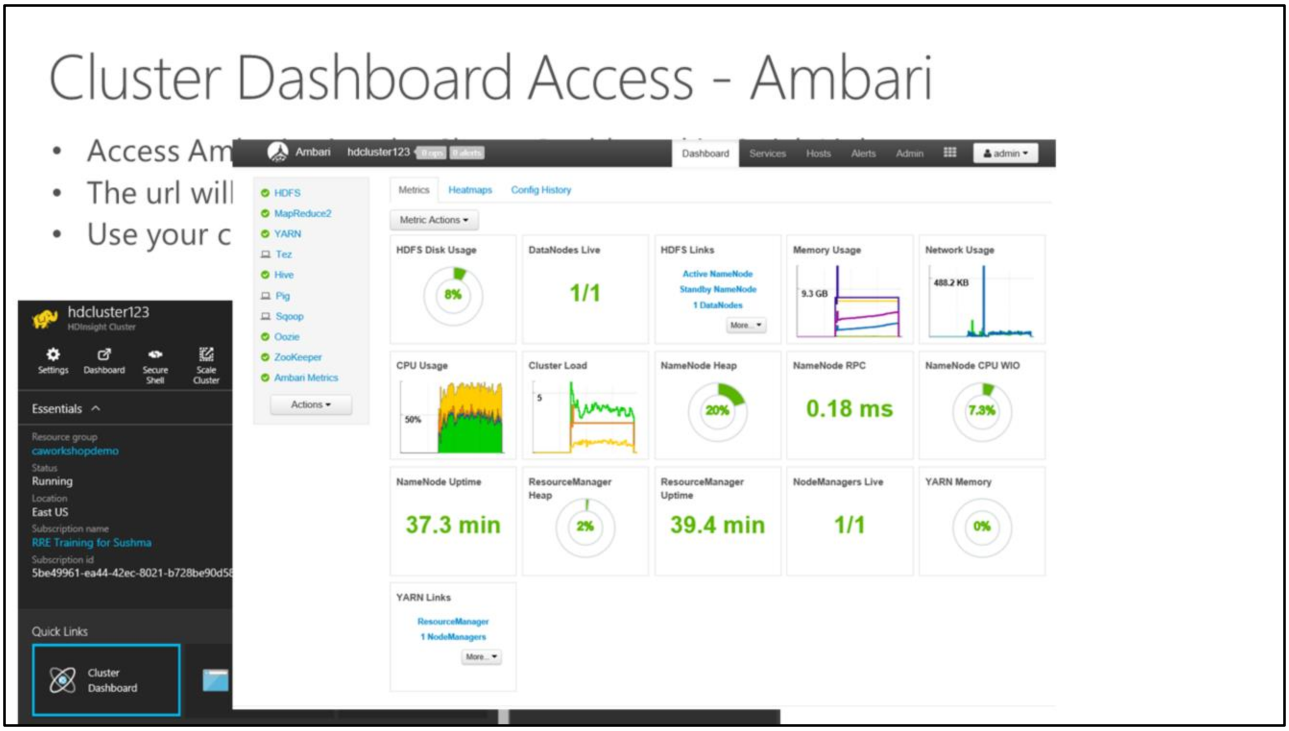
```
$clusterName = "<HDInsightClusterName>"
$location = "<AzureDataCenter>"
$clusterNodes = <ClusterSizeInNodes>

# Get the Storage account key
$storageAccountKey = Get-AzureRmStorageAccountKey -ResourceGroupName $resourceGroupName -Name $storageAccountName | %{ $_.Key1 }

# Create a new HDInsight cluster
New-AzureRmHDInsightCluster -ResourceGroupName $resourceGroupName `
    -ClusterName $clusterName `
    -ClusterType Hadoop `
    -OSType Linux `
    -ClusterSizeInNodes $numNodes `
    -ResourceGroupName $clusterResourceGroupName `
    -HttpCredential $httpCreds `
    -SshCredential $sshCreds `-Location $location `
    -DefaultStorageAccountName "$storageAccountName.blob.core.windows.net" `
    -DefaultStorageAccountKey $storageAccountKey `
    -DefaultStorageContainer $containerName `
    -ClusterSizeInNodes $clusterNodes
    -config $config `

```

- Manage HDInsight Clusters Using Azure PowerShell: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-administer-use-powershell/
- Azure Resource Manager Cmdlets: https://msdn.microsoft.com/en-us/library/mt125356.aspx
- Azure HDInsight Cmdlets: https://msdn.microsoft.com/en-us/library/mt438705.aspx



Cluster Dashboard Access - Ambari

- Access Am
- The url will
- Use your c

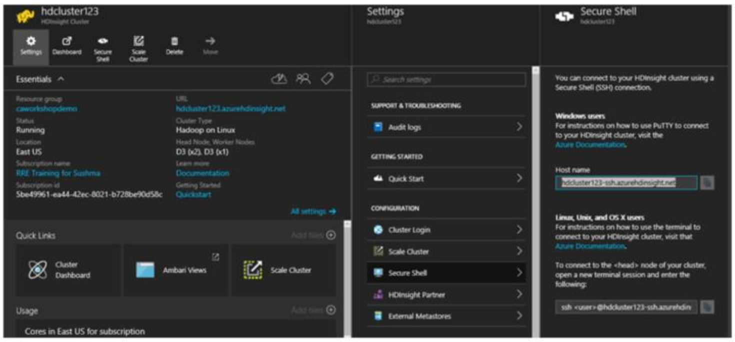
hdcluster123
HDInsight Cluster

Essentials

Resource group
caworkshopdemo
Status
Running
Location
East US
Subscription name
RRE Training for Sushma
Subscription id
5be49961-ea44-42ec-8021-b728be90d58

- Monitoring clusters with Ambari: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-monitor-use-ambari-api/
- Managing clusters with Ambari: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-manage-ambari/

Cluster SSH Access



Resource group: caworkshopdemo

URL: hdcluster123.azurehdinsight.net

Status: Running

Cluster Type: Hadoop on Linux

Location: East US

Head Node, Worker Nodes: D3 (x2), D3 (x1)

Subscription name: RRE Training for Sushma

Learn more: Documentation

Subscription id: 5be49961-ea44-42ec-8021-b728be90d58c

Getting Started: Quickstart

Cores in East US for subscription

You can connect to your HDInsight cluster using a Secure Shell (SSH) connection.

Windows users
For instructions on how to use PuTTY to connect to your HDInsight cluster, visit the Azure Documentation.

Host name: hdcluster123-ssh.azurehdinsight.net

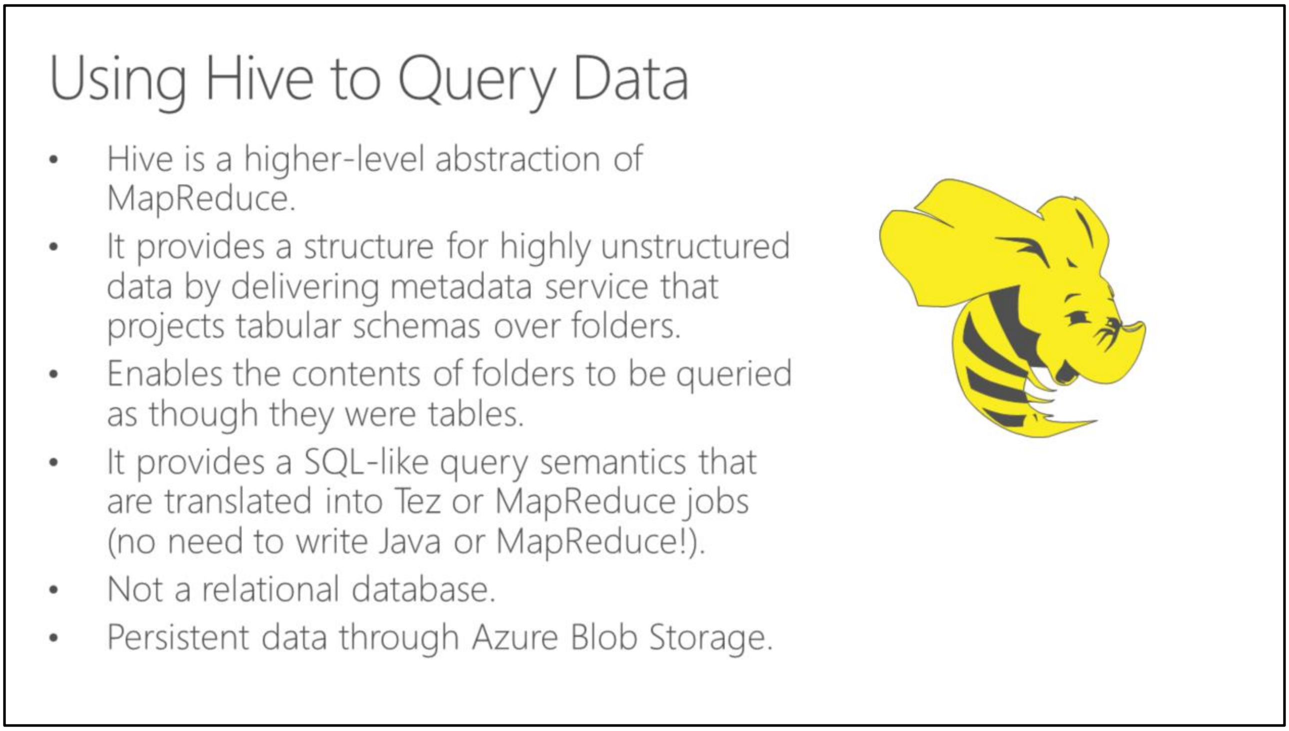
Linux, Unix, and OS X users
For instructions on how to use the terminal to connect to your HDInsight cluster, visit that Azure Documentation.

To connect to the <head> node of your cluster, open a new terminal session and enter the following:

```
ssh <user>@hdcluster123-ssh.azurehdin
```

- In the settings tab in Azure portal, find the Secure Shell tab
- Use the ssh credentials you created during provision to login using a terminal:
- ssh <sshusername>@<clustername>-ssh.azurehdinsight.net

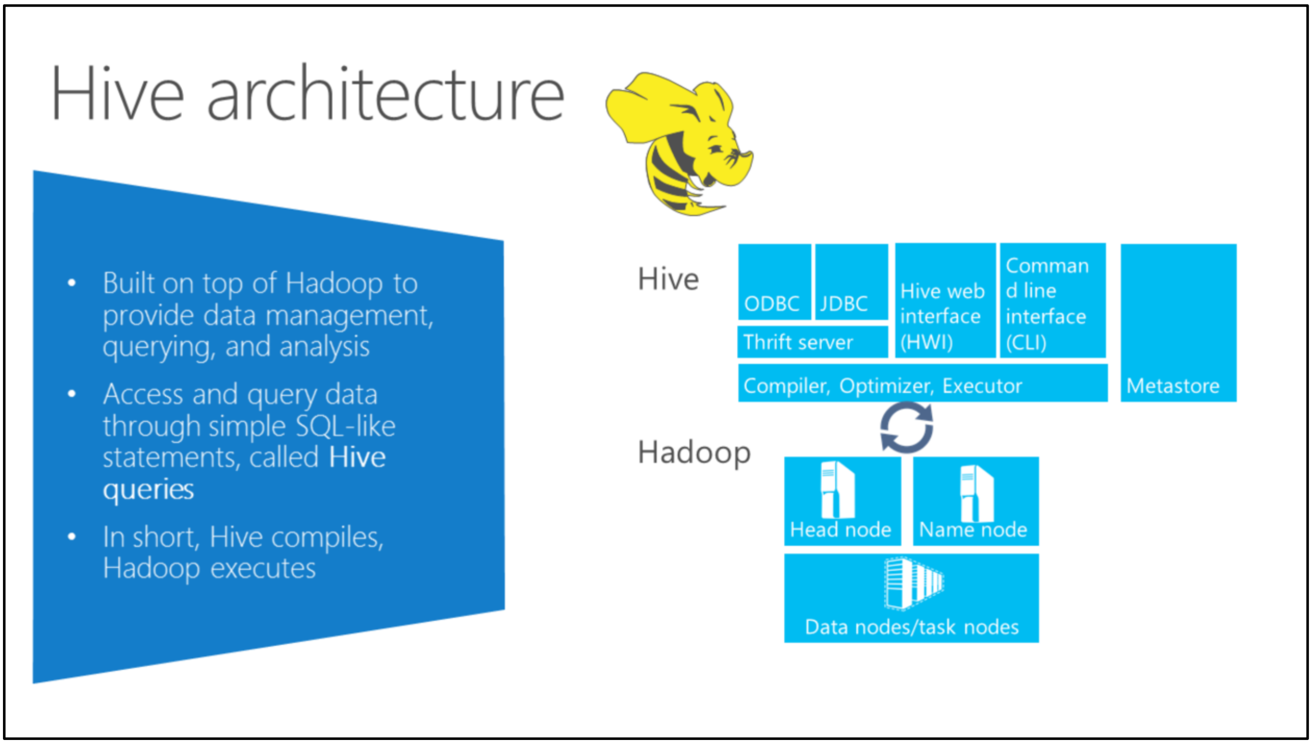
- Using ssh with linux-based Hadoop clusters: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-linux-use-ssh-unix/



Using Hive to Query Data

- Hive is a higher-level abstraction of MapReduce.
- It provides a structure for highly unstructured data by delivering metadata service that projects tabular schemas over folders.
- Enables the contents of folders to be queried as though they were tables.
- It provides a SQL-like query semantics that are translated into Tez or MapReduce jobs (no need to write Java or MapReduce!).
- Not a relational database.
- Persistent data through Azure Blob Storage.

- Hive for HDInsight: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/
- Referencing user defined functions with Hive: https://msdn.microsoft.com/en-us/library/dn749875.aspx?f=255&MSPPError=-2147217396
- Using Apache Tez for improved performance: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/#usetez



Hive architecture

- Built on top of Hadoop to provide data management, querying, and analysis
- Access and query data through simple SQL-like statements, called **Hive queries**
- In short, Hive compiles, Hadoop executes

Hive

ODBC | JDBC | Hive web interface (HWI) | Command line interface (CLI)

Thrift server

Compiler, Optimizer, Executor

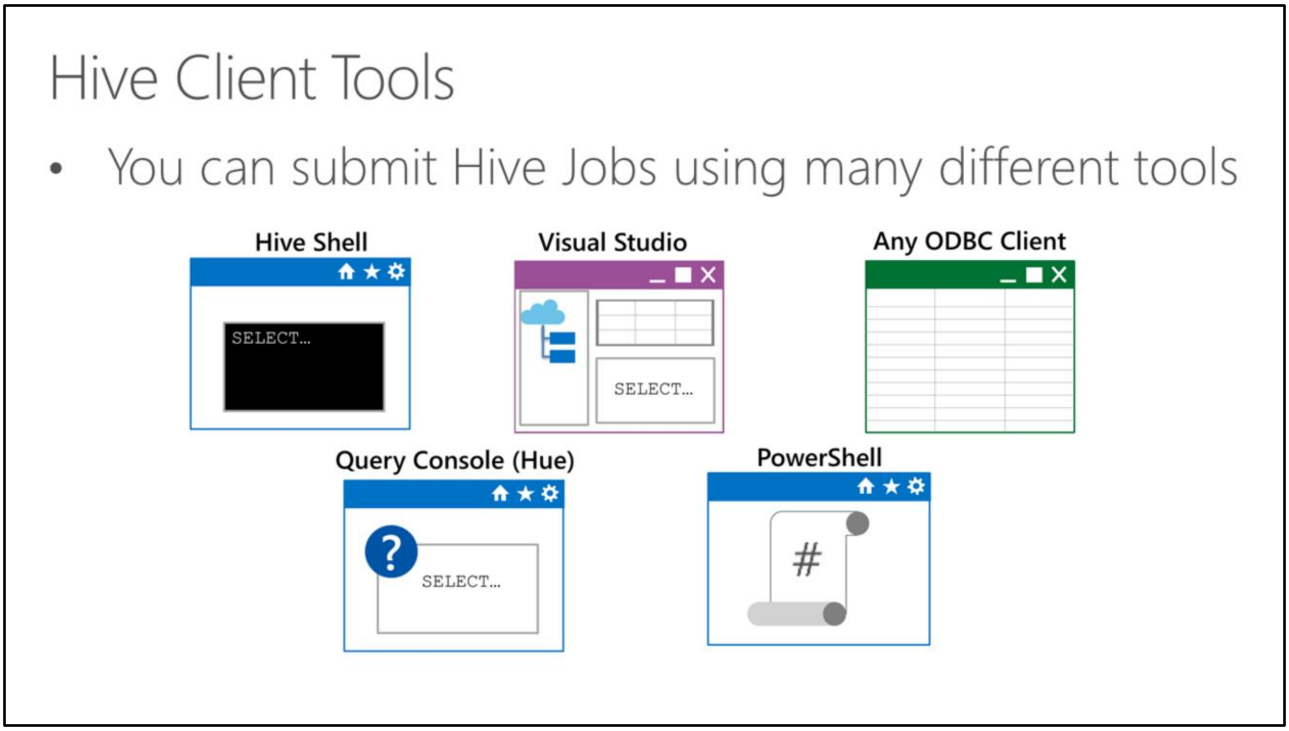
Metastore

Hadoop

Head node | Name node

Data nodes/task nodes

- Hive for HDInsight: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/
- Referencing user defined functions with Hive: https://msdn.microsoft.com/en-us/library/dn749875.aspx?f=255&MSPPError=-2147217396
- Using Apache Tez for improved performance: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/#usetez



Hive Client Tools

- You can submit Hive Jobs using many different tools

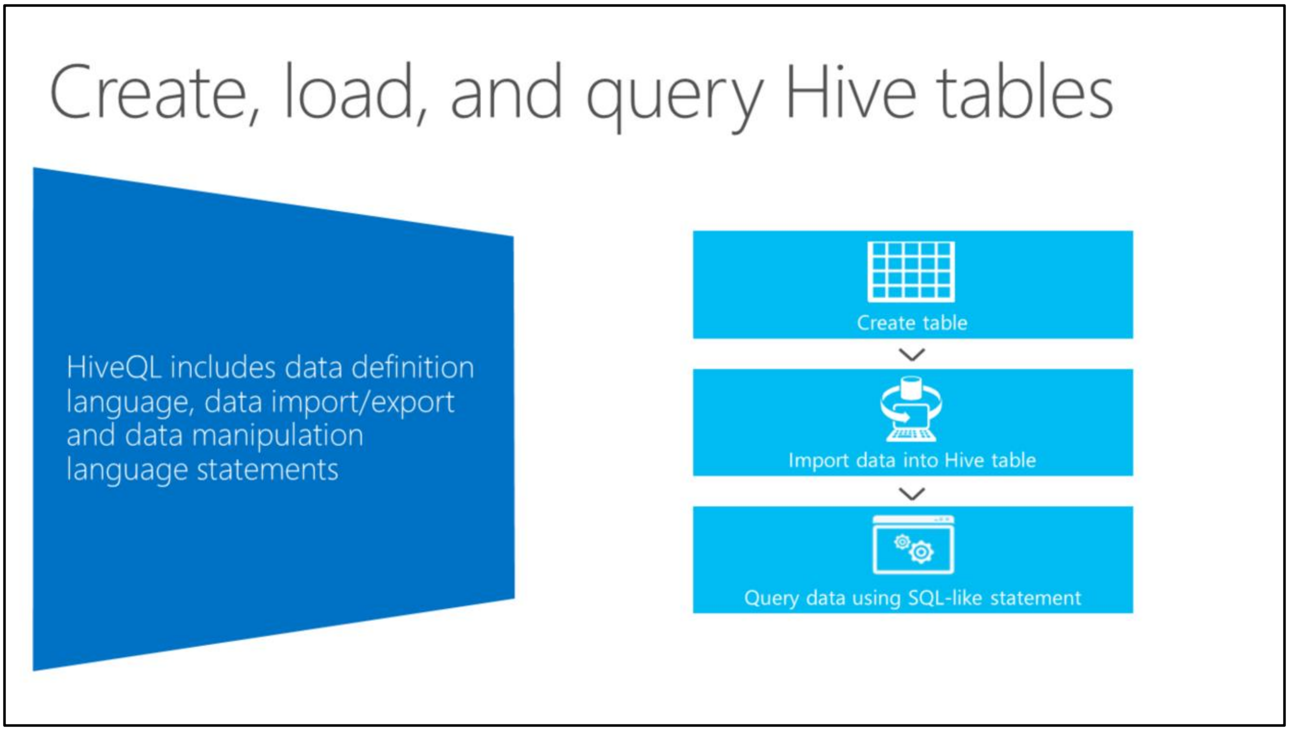
Hive Shell

Visual Studio

Any ODBC Client

Query Console (Hue)

PowerShell



Create, load, and query Hive tables

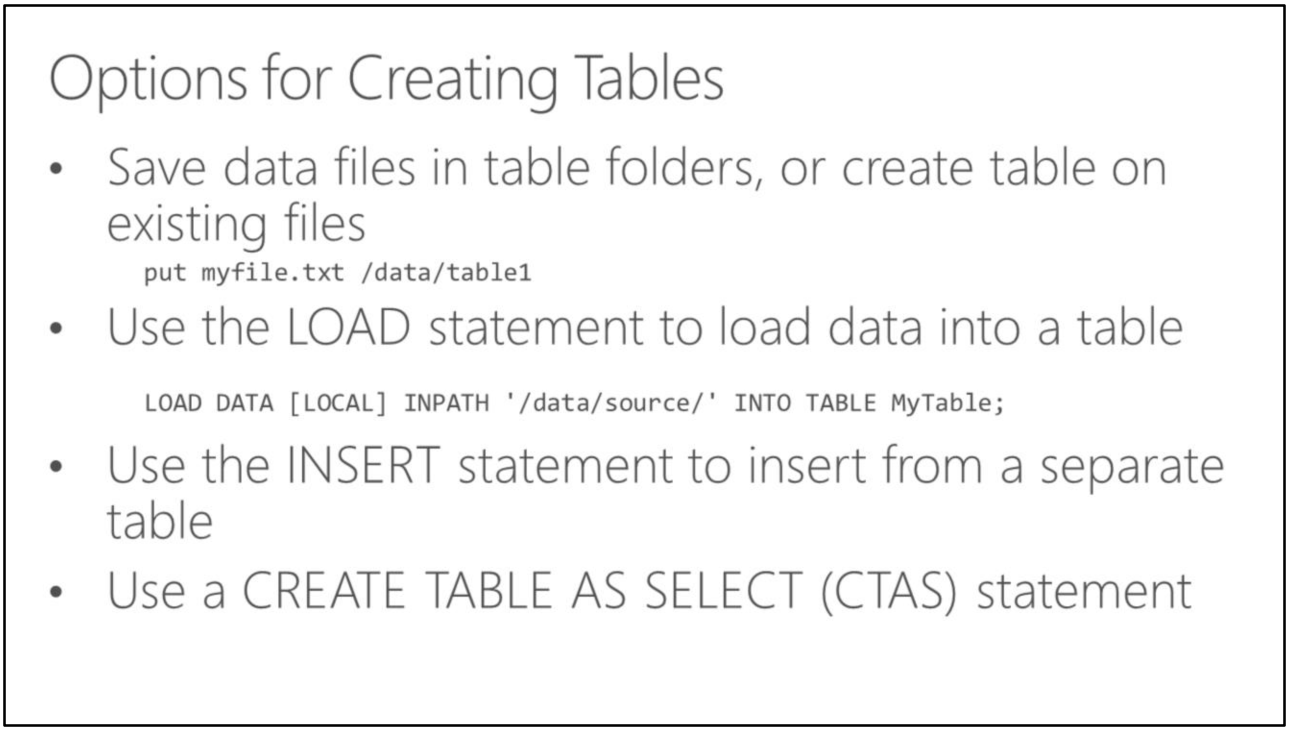
HiveQL includes data definition language, data import/export and data manipulation language statements

Create table

Import data into Hive table

Query data using SQL-like statement

1. Full tutorial on creating Hive Tables: https://www.dezyre.com/hadoop-tutorial/apache-hive-tutorial-tables



Options for Creating Tables

- Save data files in table folders, or create table on existing files

```
put myfile.txt /data/table1
```

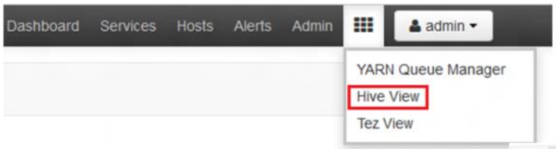
- Use the LOAD statement to load data into a table

```
LOAD DATA [LOCAL] INPATH '/data/source/' INTO TABLE MyTable;
```

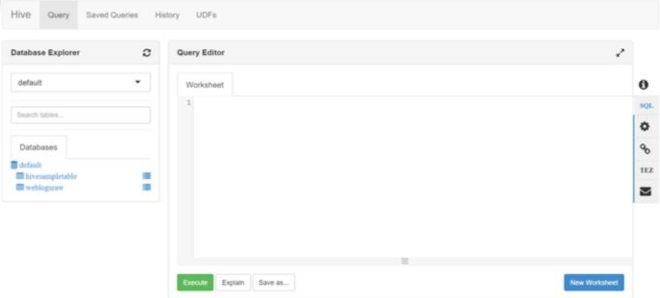
- Use the INSERT statement to insert from a separate table
- Use a CREATE TABLE AS SELECT (CTAS) statement

Hive Query in Ambari

- Navigate to Ambari Views from Azure Portal.
- Select Hive view from list of views:



- Enter new table queries in the Query Editor
- View tables in the database explorer



1. More about Ambari Views: https://azure.microsoft.com/en-us/blog/using-ambari-views-to-author-hive-and-pig-queries/

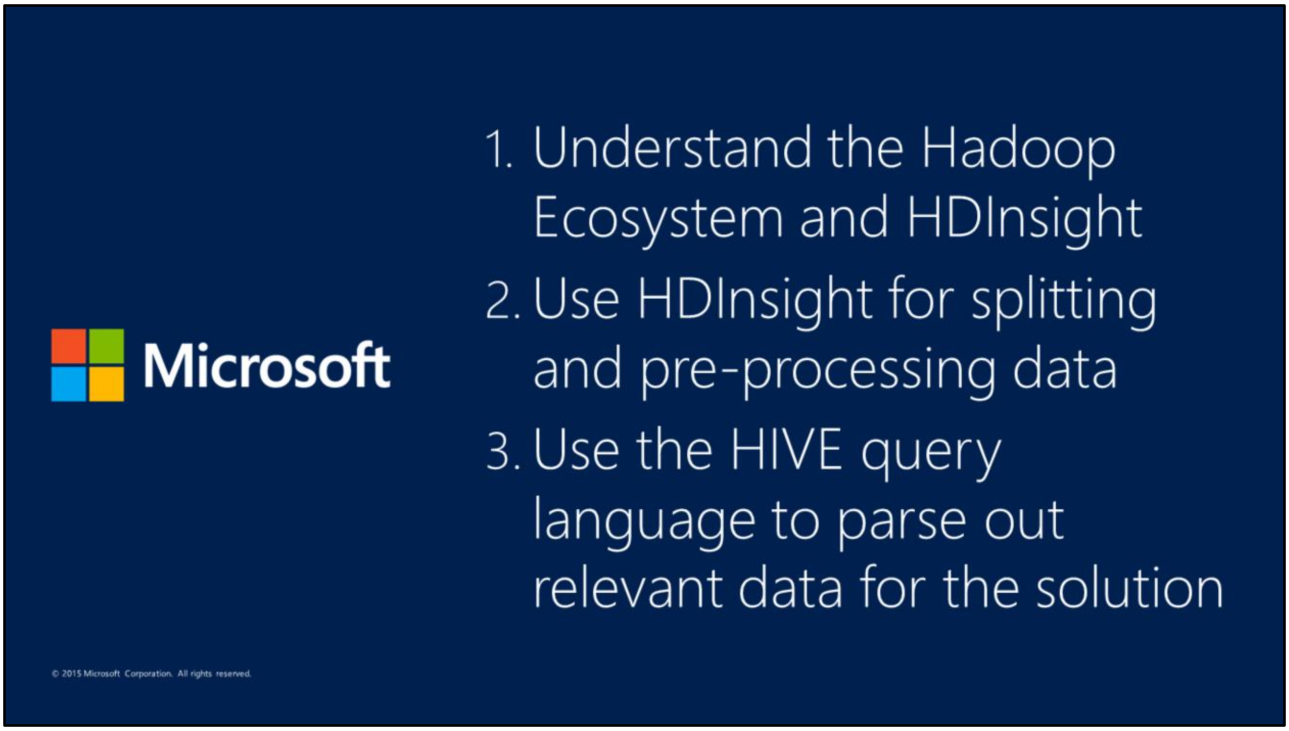
Lab: Using the HDI Emulator and tutorials for the ecostructure

Microsoft Azure

Microsoft

PARTNER PRACTICE ENABLEMENT BOOTCAMP

1. Since cost is a primary factor, you can practice using the emulator locally or in an Azure Virtual Machine. Open this location: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-emulator-get-started/ and follow the steps you see there.
2. For a more complete tutorial listing for the ecostructure, open and work through this page: https://azure.microsoft.com/en-us/documentation/learning-paths/hdinsight-self-guided-hadoop-training/



1. Understand the Hadoop Ecosystem and HDInsight
2. Use HDInsight for splitting and pre-processing data
3. Use the HIVE query language to parse out relevant data for the solution