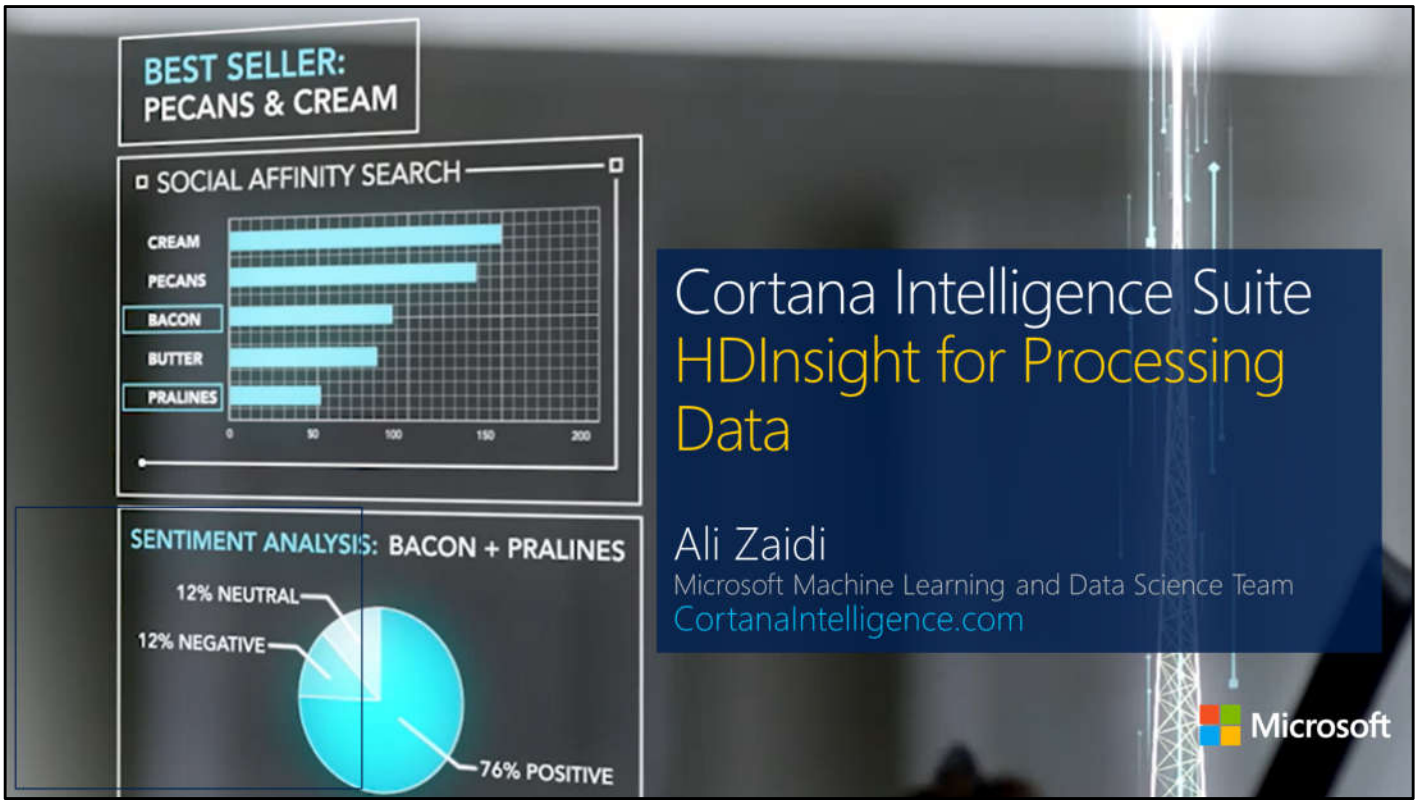
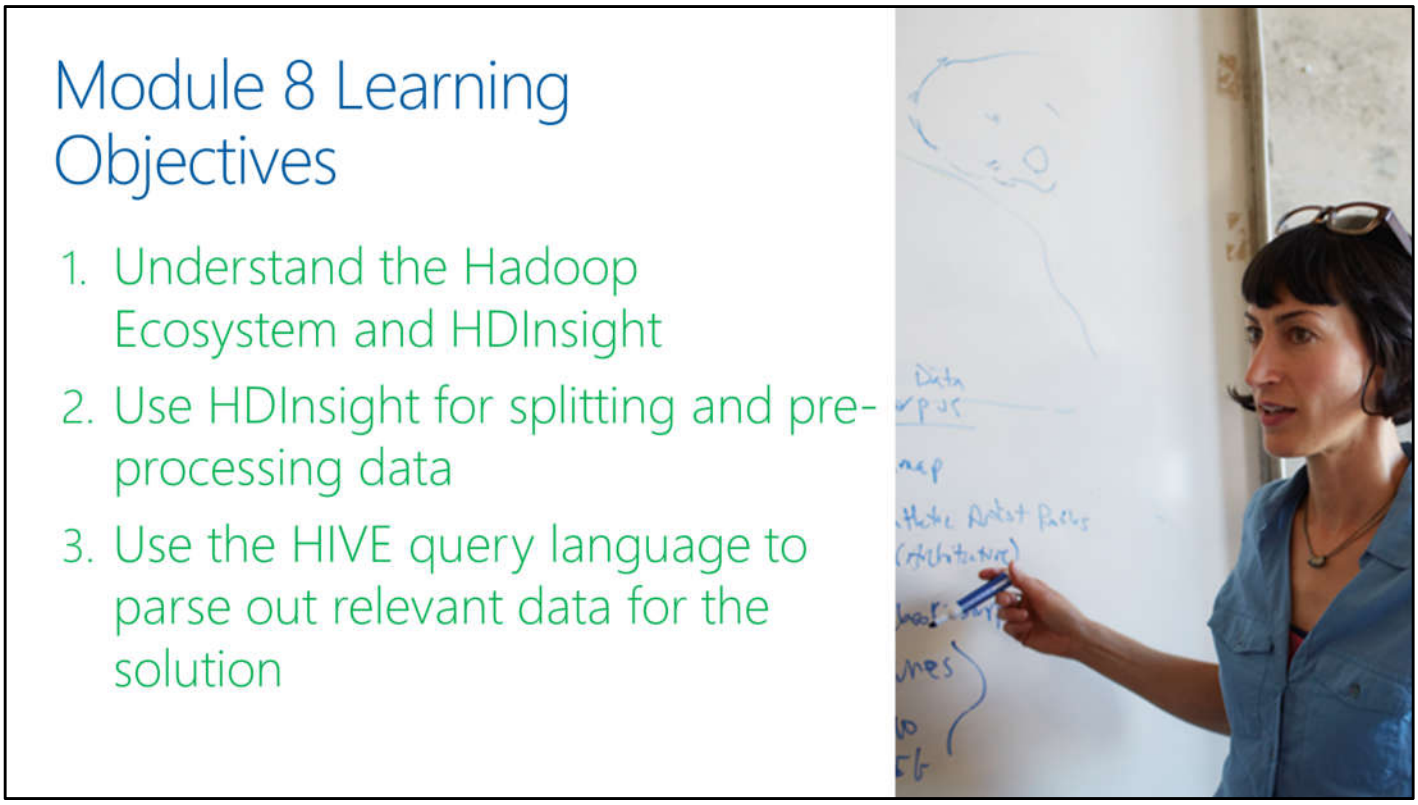
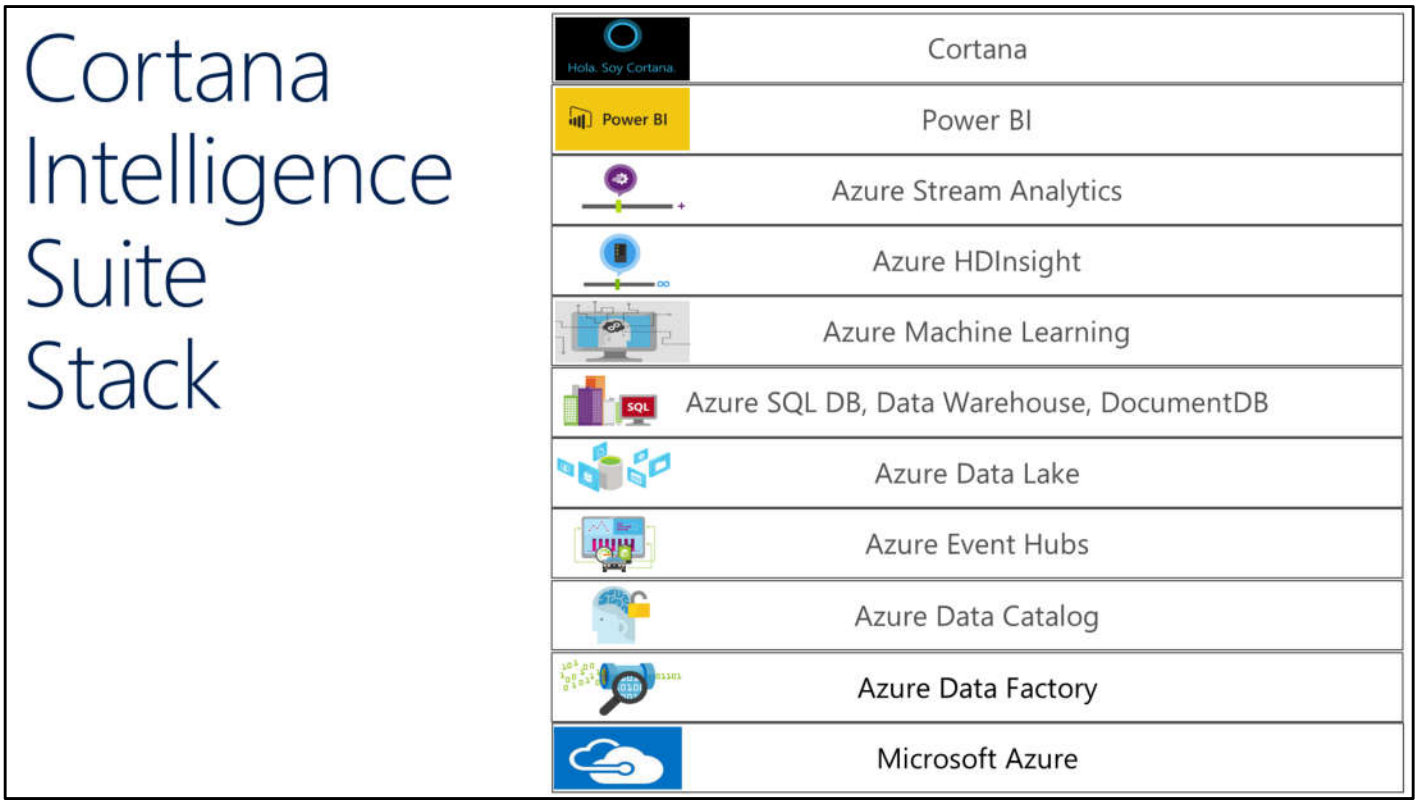


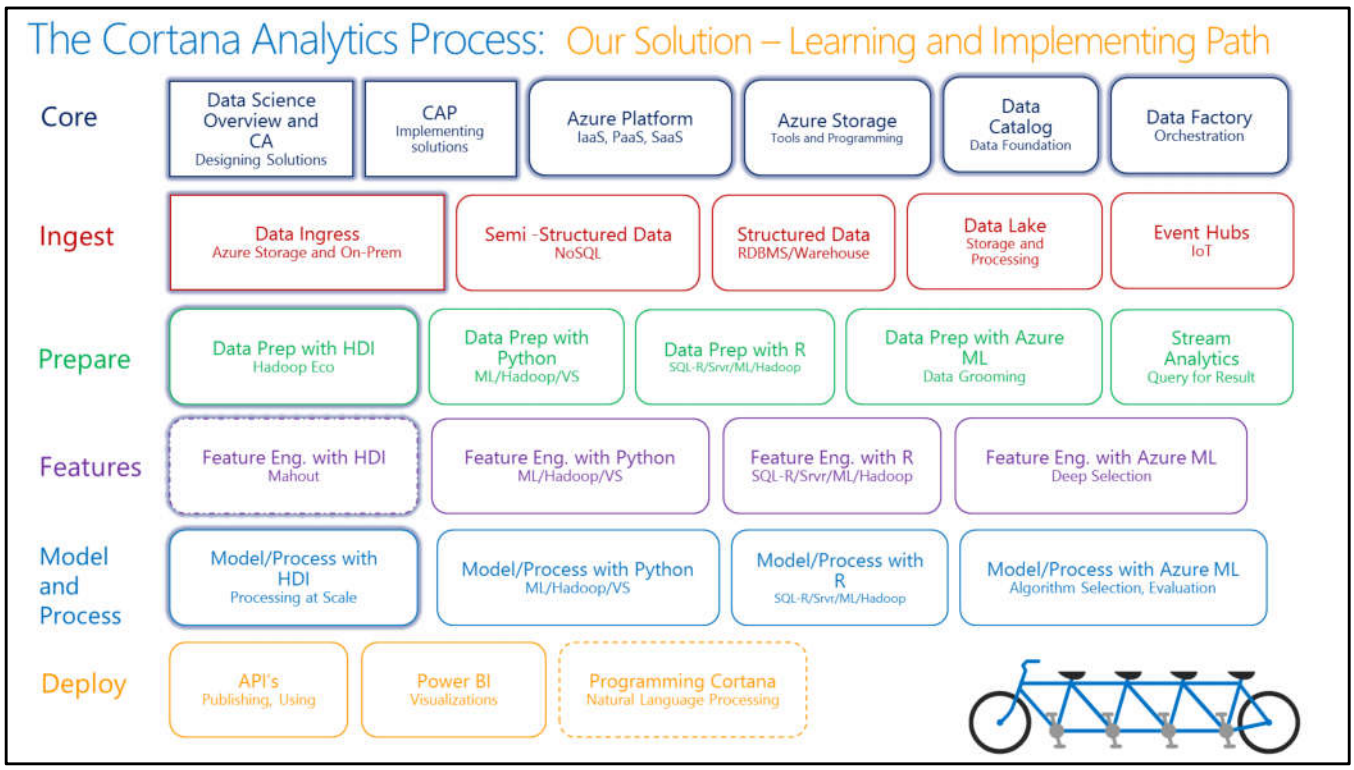
1. Main page: http://cortanaanalytics.com
2. Before you take this module, you should be able to:
    1. Understand the process for using Azure Data Factory
    2. Use Azure Data Factory to ingest data
    3. Use Azure Data Factory to leave data on prem
    4. Use Azure Data Factory to call functions 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



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



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)



1. The Cortana Analytics Process: https://azure.microsoft.com/en-us/documentation/learning-paths/cortana-analytics-process/

Business Case

AdventureWorks is a company that makes and sells bicycles. The sales are conducted around the world. We also support our products. Interestingly, the issue we're facing is in our facilities.

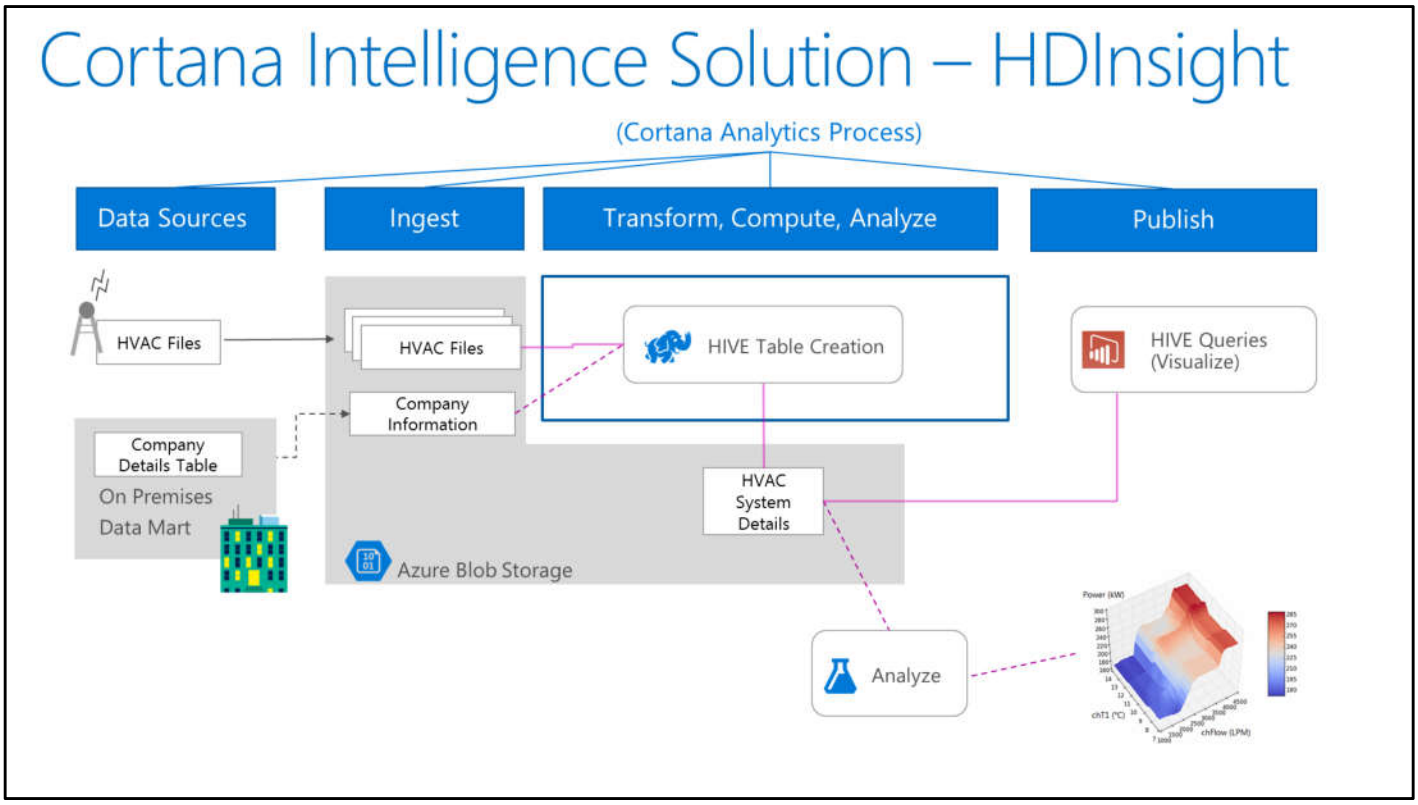
We need to know a lot more about our HVAC systems – they are critical to the machinery that creates the fine-detail parts on our products. The HVAC systems have sensors that create a lot of data – several million records a day, in fact.

We have facilities around the world, and when a facility runs "hot", we have to shift production to another part of the world. With our just-in-time manufacturing process, this has huge financial impacts. We've had situations where the systems ran hot and we shifted production to another location (at great cost), and then we found it was an anomaly in the reporting system.

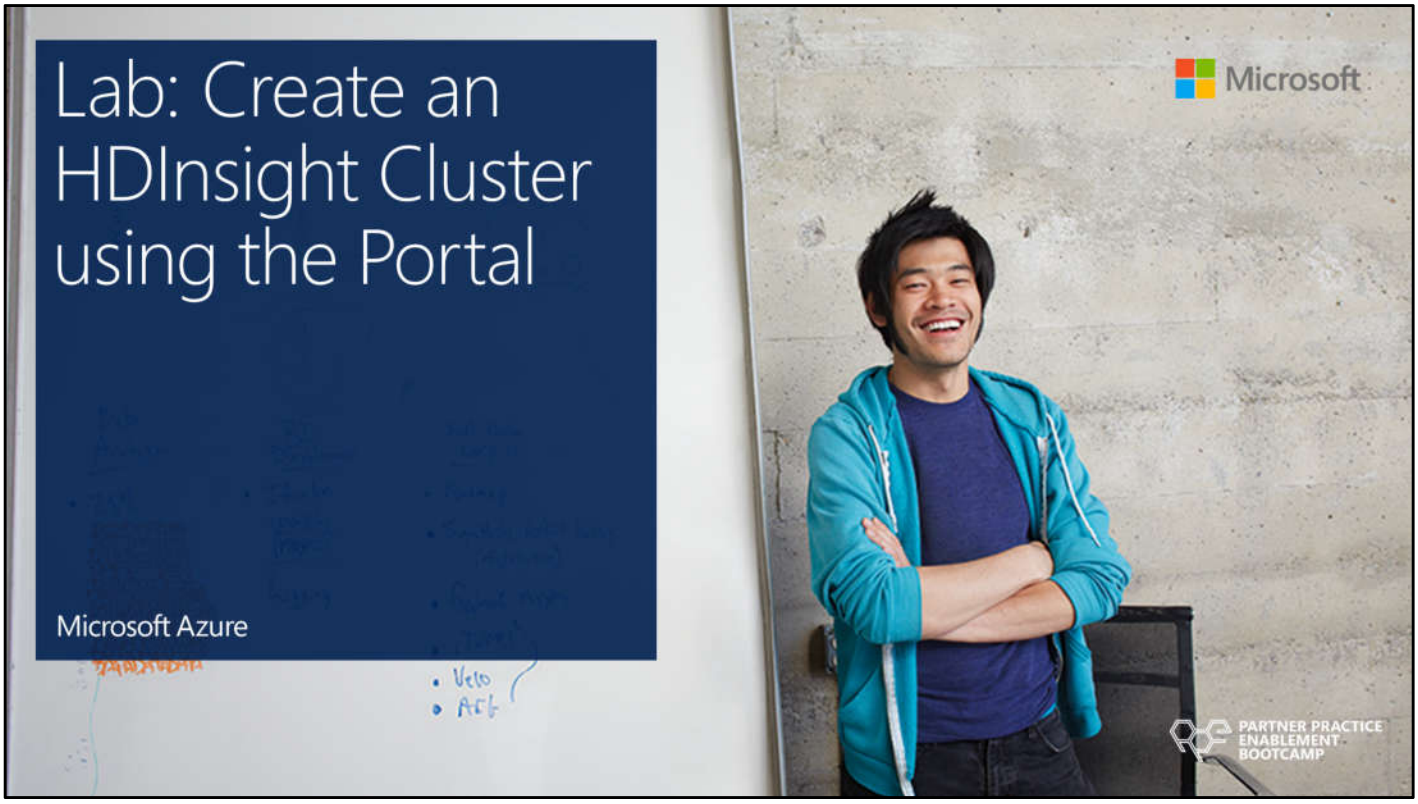
Ideally we want a graphic that shows our management team the map of where our HVAC systems are, and their daily status.

More on our in-house data: https://technet.microsoft.com/en-us/library/ms124501%28v=sql.100%29.aspx

1. The AdventureWorks Scenarios: https://technet.microsoft.com/en-us/library/ms124501%28v=sql.100%29.aspx



1. Using the Optional Machine Learning Exercise in combination with this solution: https://gallery.cortanaanalytics.com/Experiment/modeling-for-HVAC-system-2



1. Locate the following information from your earlier labs:
    1. Resource Group Name
    2. Storage Account Name
    3. Storage Account Keys
    4. Container Name
2. Log on to the Portal
3. Follow along with the instructor to start your creation process – will take up to 45 minutes.

Hadoop and HDInsight

Using the Hadoop Ecosystem to process and query data

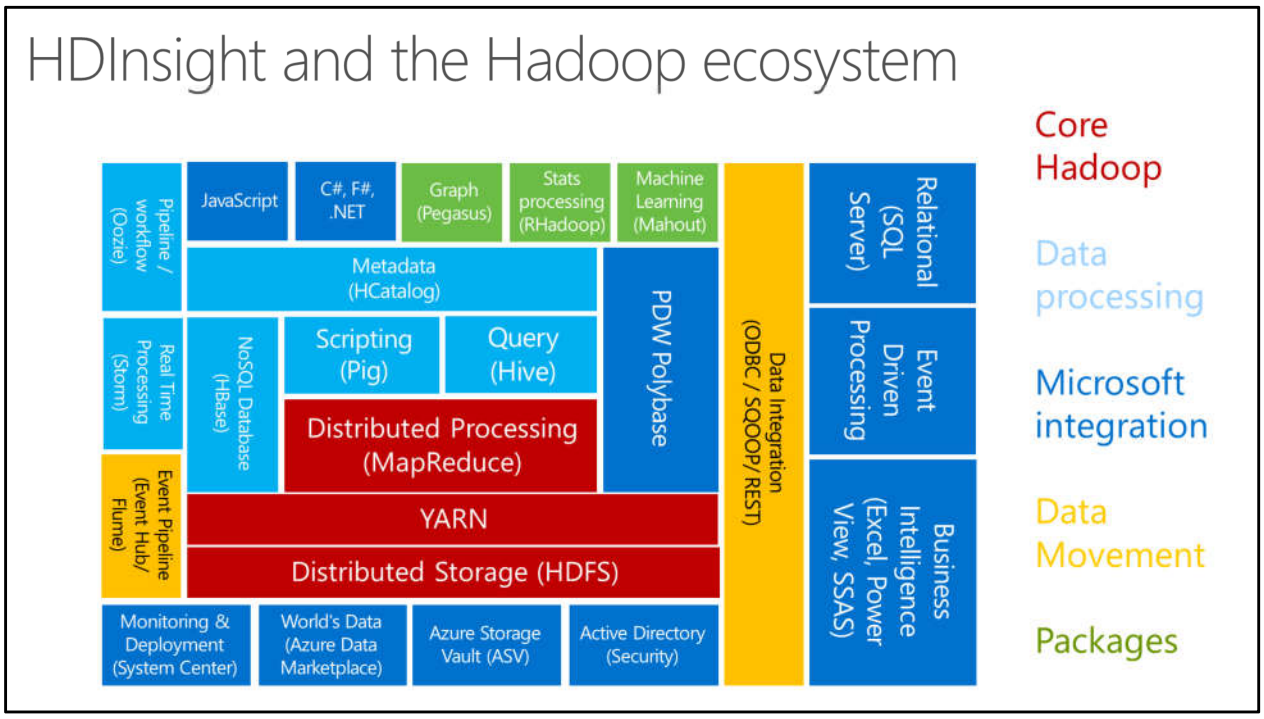
1. Primary site: https://azure.microsoft.com/en-us/services/hdinsight/
2. Quick overview: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/
3. 4-week online course through the edX platform: https://www.edx.org/course/processing-big-data-azure-hdinsight-microsoft-dat202-1x
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

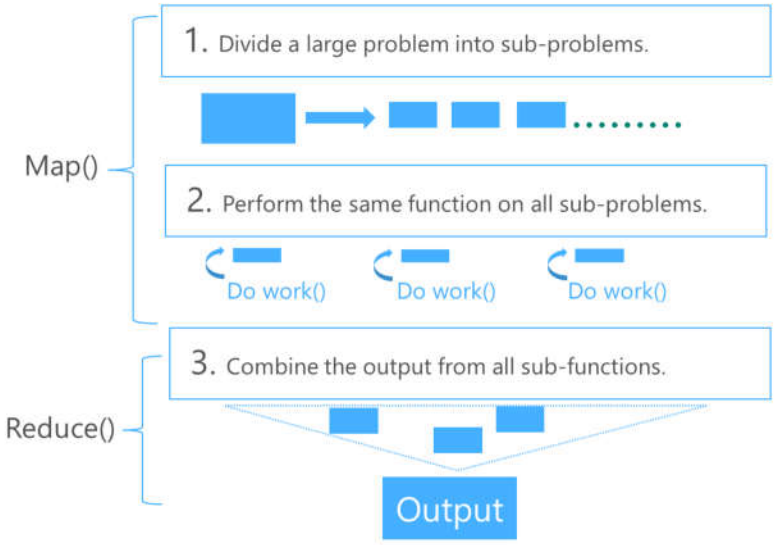
- Primary/head document: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/
- For more information about Hadoop, visit the apache foundation site: http://hadoop.apache.org/



- Full training example for the local HDP Instance: http://hortonworks.com/hadoop-tutorial/hello-world-an-introduction-to-hadoop-hcatalog-hive-and-pig/
- More detail on the Hadoop Components: http://www.datasciencecentral.com/profiles/blogs/hadoop-herd-when-to-use-what

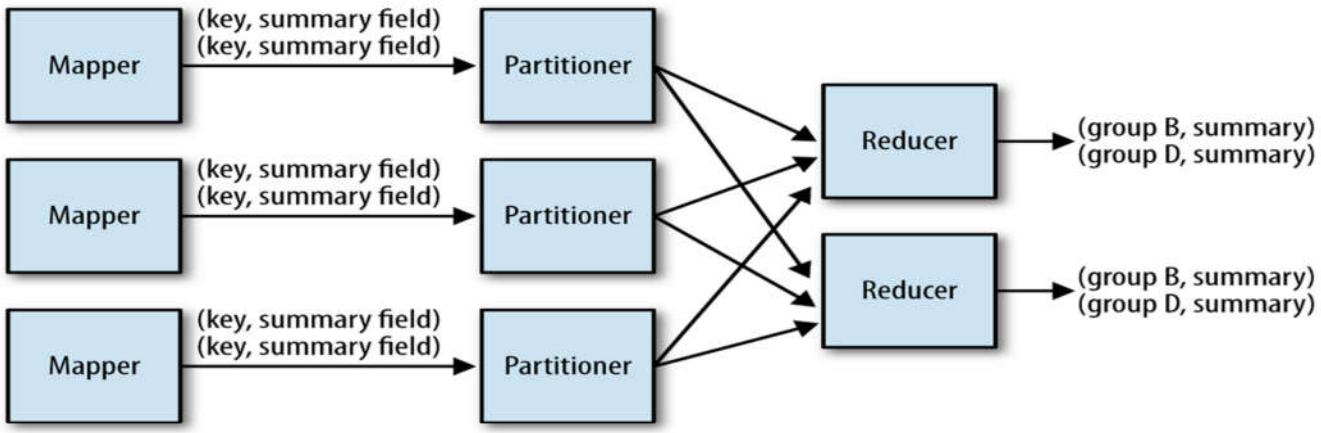
Hadoop MapReduce

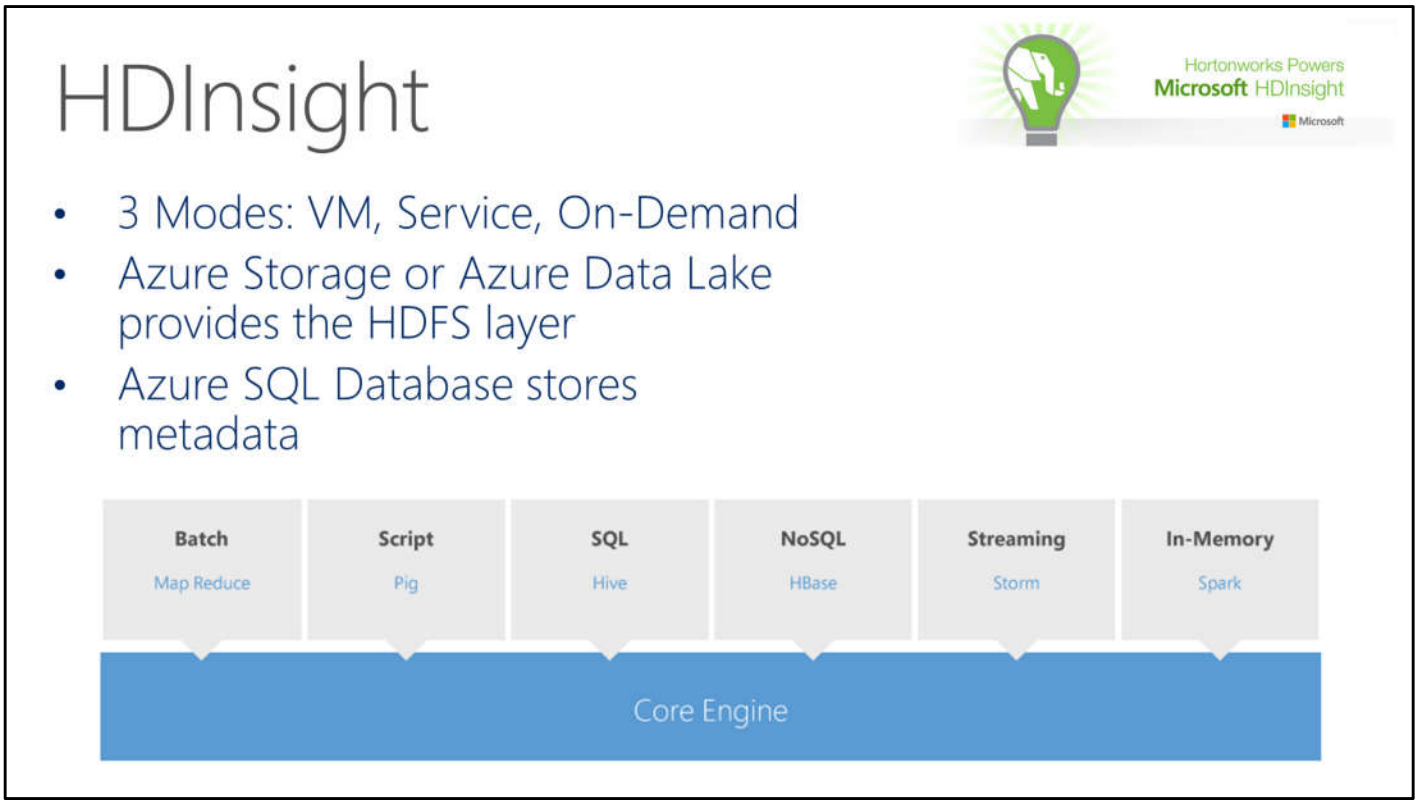
- Programming framework (library and runtime) for analyzing datasets stored in HDFS
- Composed of user-supplied Map and Reduce functions:
  - Map() - subdivide and conquer
  - Reduce() - combine and reduce cardinality



- 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





- 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 and Visual Studio
- 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

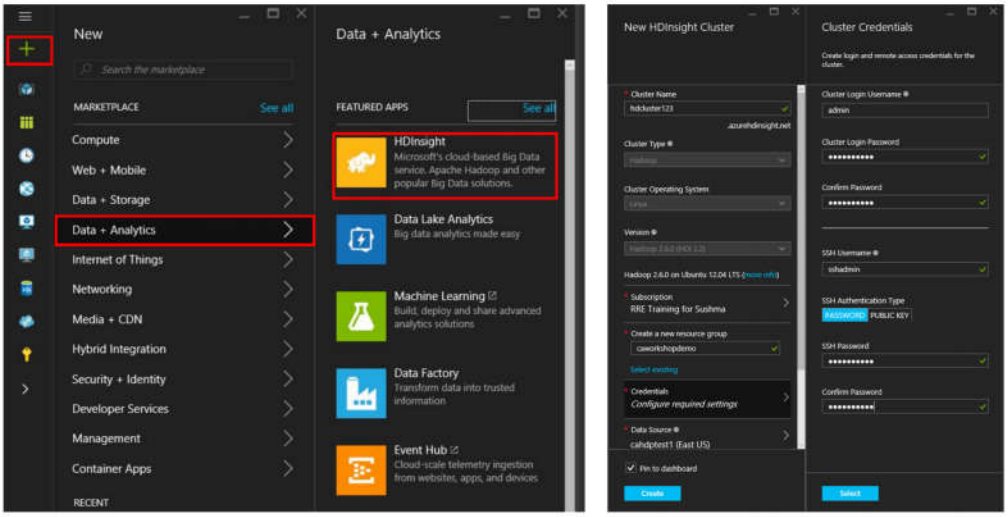
- Azure Portal: azure.portal.com
- Provisioning Clusters: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-provision-clusters/
- Different clusters have different node types, number of nodes, and node sizes.

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/

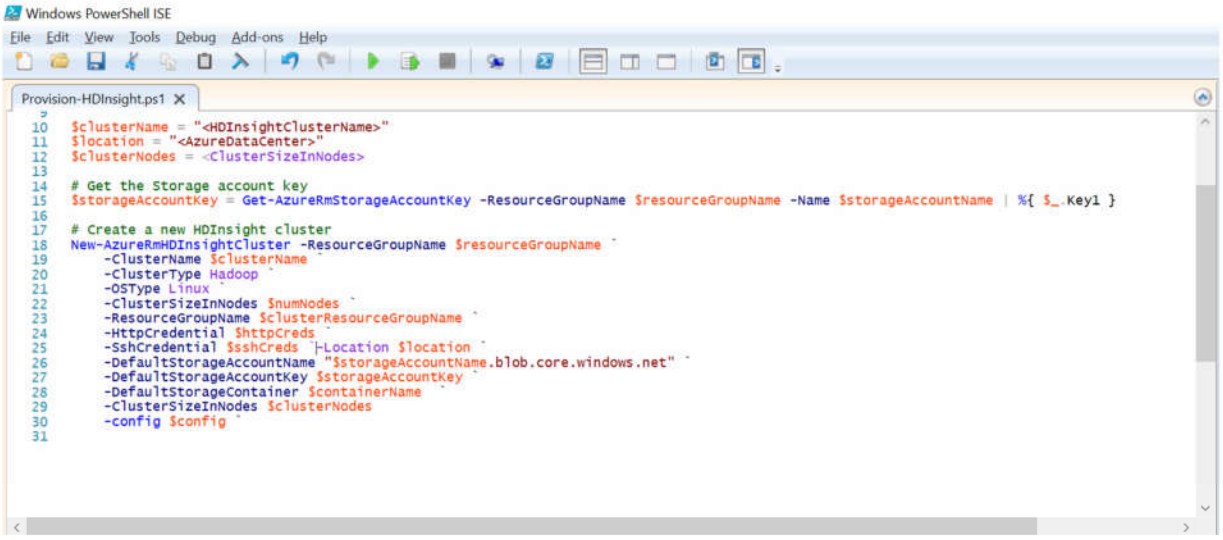
Provisioning with Azure Portal



- **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

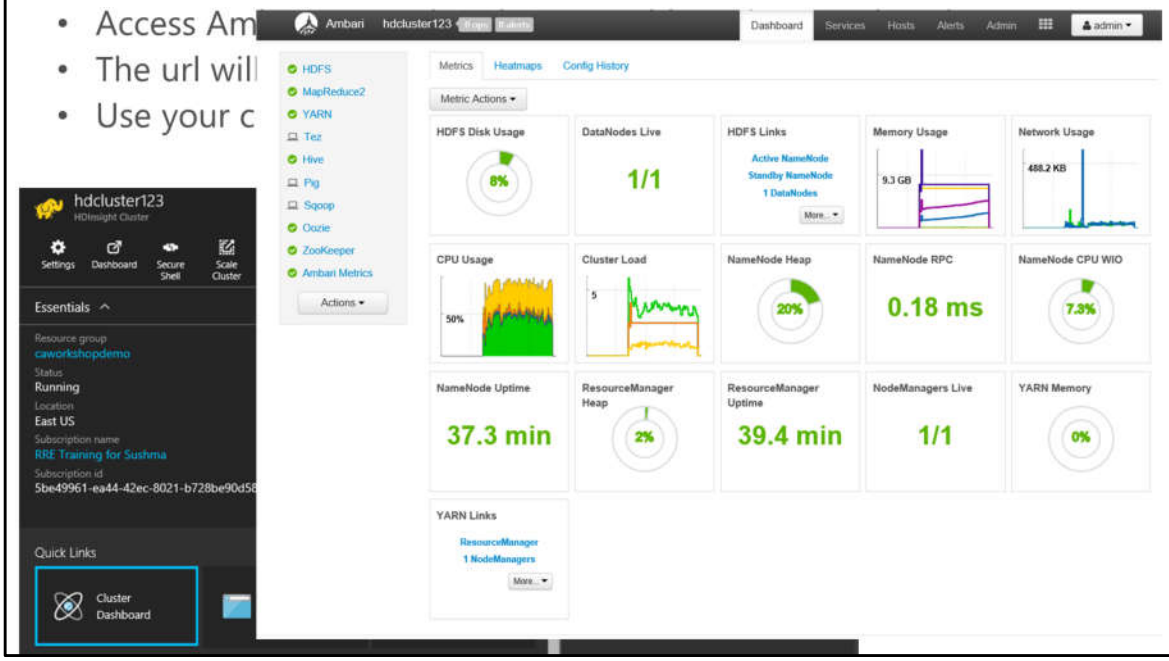
- You can provision with Azure PowerShell



- 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

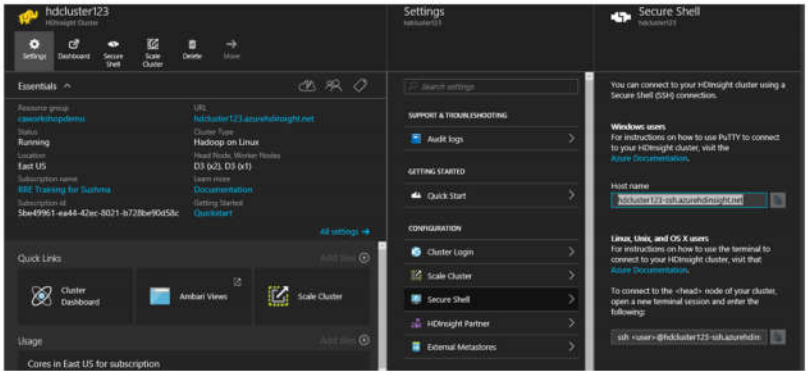
Cluster Dashboard Access - Ambari

- Access Am
- The url will
- Use your c



- 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



- In the settings tab in Azure portal, find the Secure Shell tab

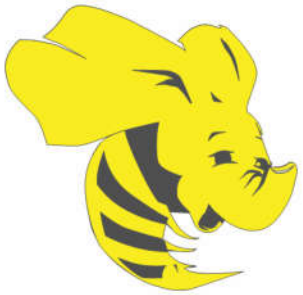
- Use the ssh credentials you created during provision to login using a terminal:
- ssh <sshusername>@<clustername>-ssh.azurehdinsight.net

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

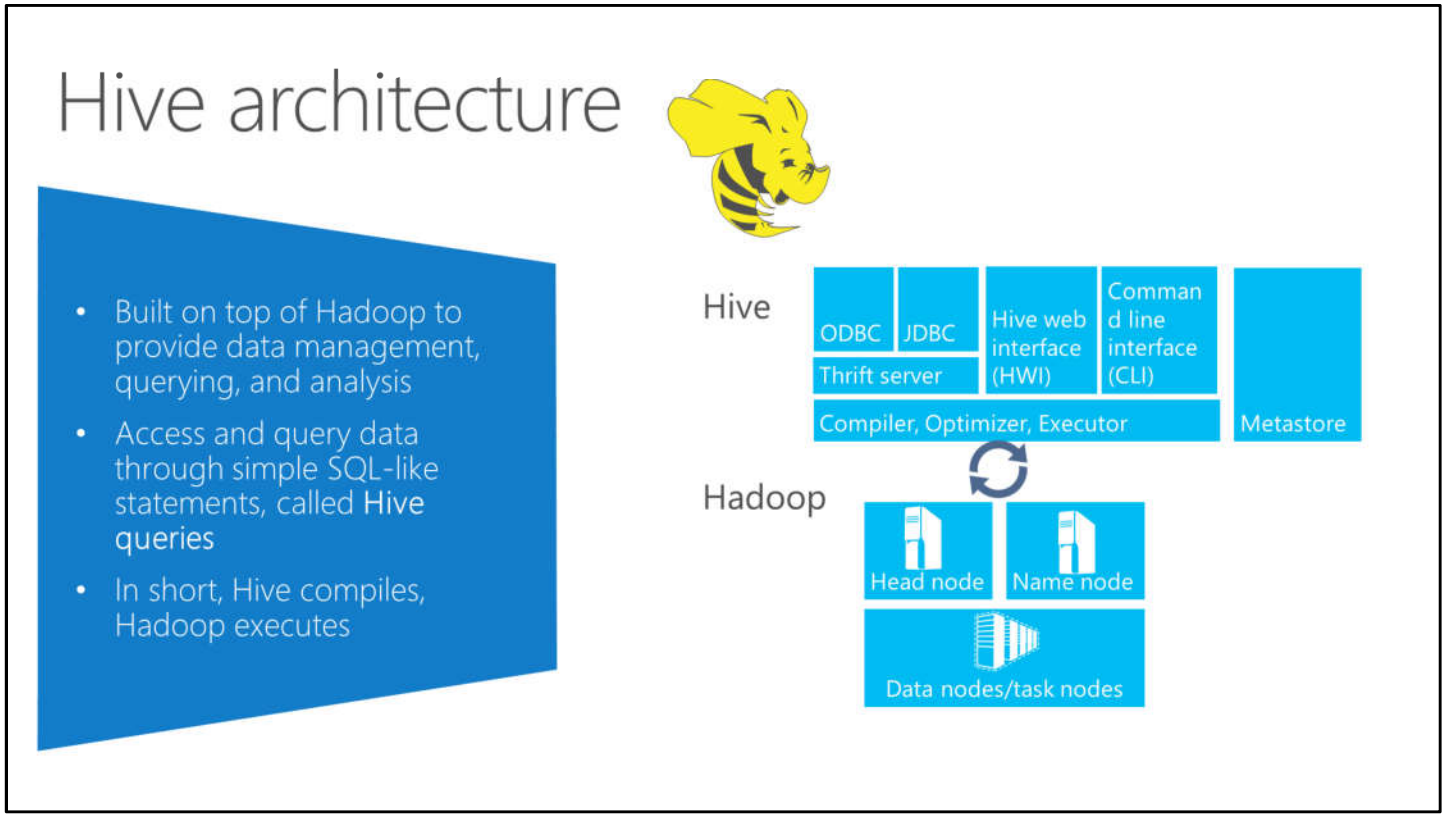
1. Open the Azure Portal
2. Note your connection string to the Ambari Cluster URL
3. If cost is a primary factor, you can practice using the emulator locally or in an Azure Virtual Machine. Open this location: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-emulator-get-started/ and follow the steps you see there.
4. 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/

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.



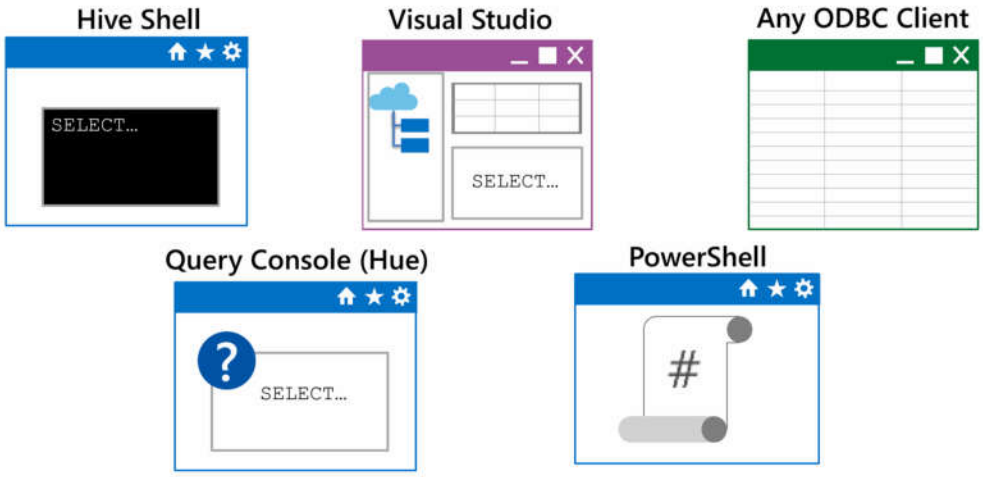
- Hive for HDInsight: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/
- Referencing user defined functions with Hive: https://msdn.microsoft.com/en-us/library/dn749875.aspx?f=255&MSPPError=-2147217396
- Using Apache Tez for improved performance: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/#usetez

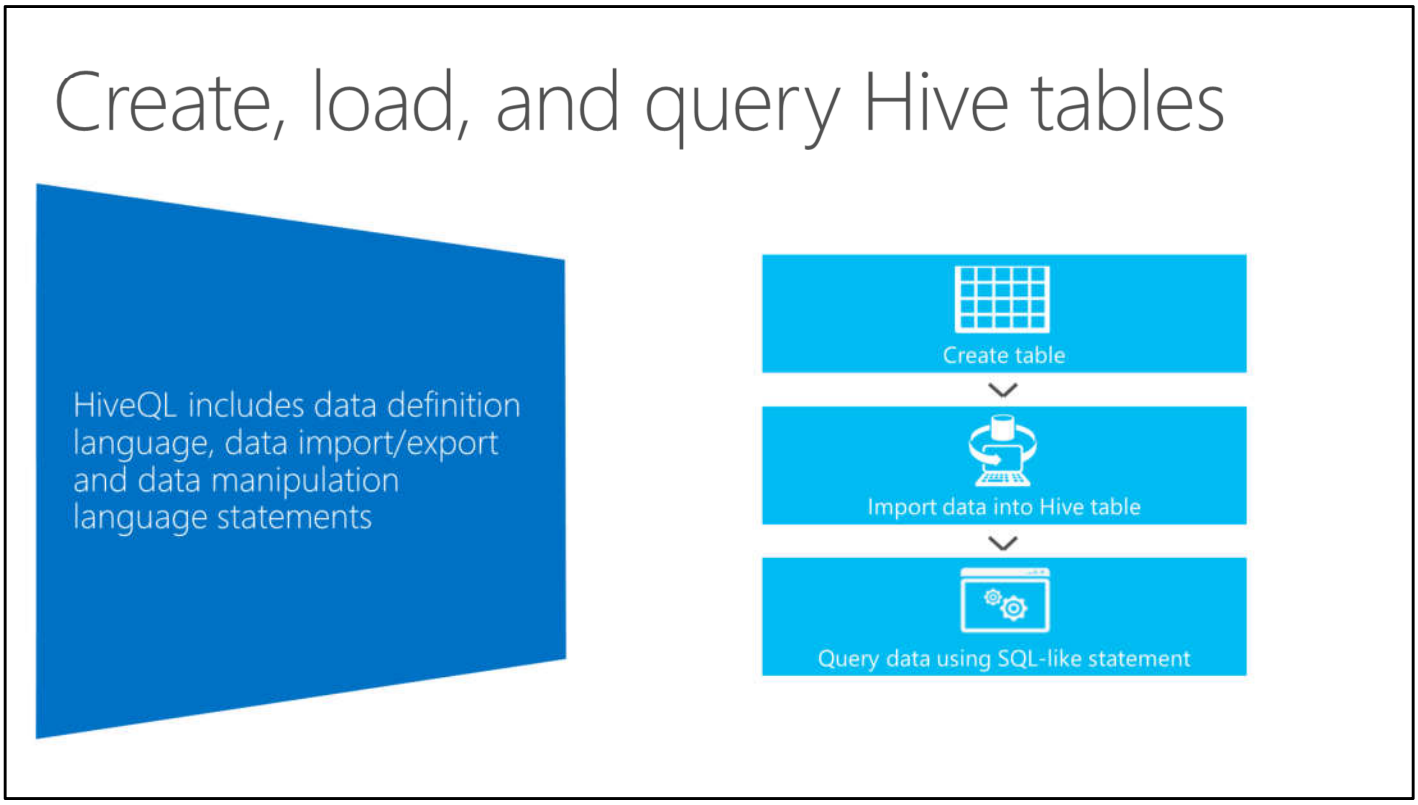


- Hive for HDInsight: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive /
- Referencing user defined functions with Hive: https://msdn.microsoft.com/en-us/library/dn749875.aspx?f=255&MSPPError=-2147217396
- Using Apache Tez for improved performance: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/#usetez

Hive Client Tools

- You can submit Hive Jobs using many different tools





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
```

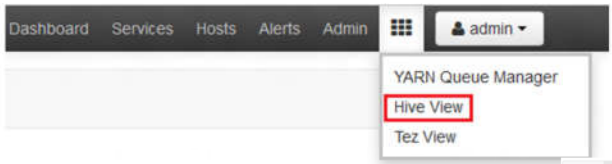
- Use the LOAD statement to load data into a table

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

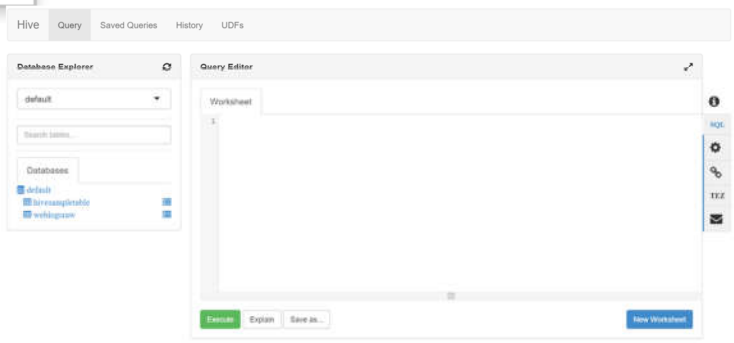
- Use the INSERT statement to insert from a separate table
- Use a CREATE TABLE AS SELECT (CTAS) statement

Hive Query in Ambari

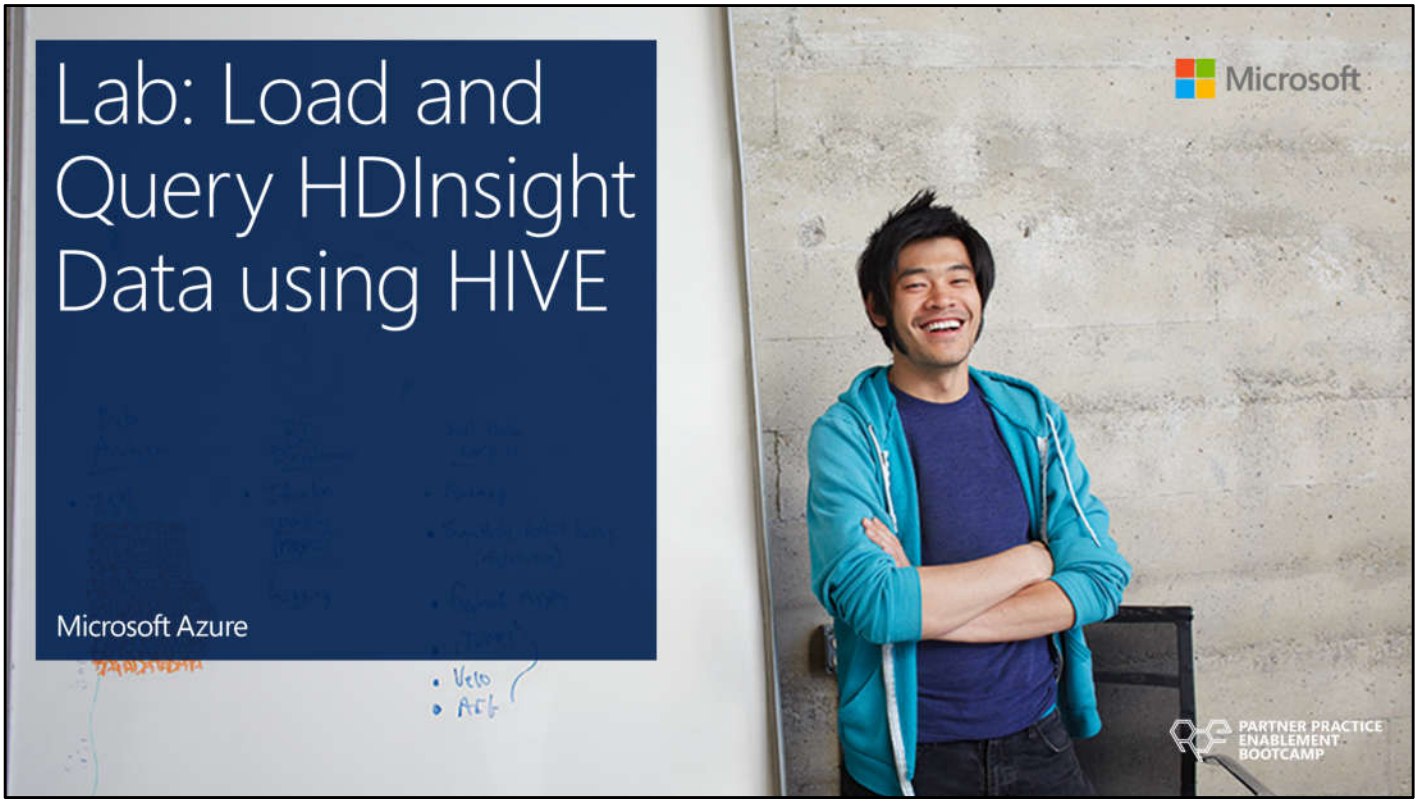
- Navigate to Ambari Views from Azure Portal.
- Select Hive view from list of views:



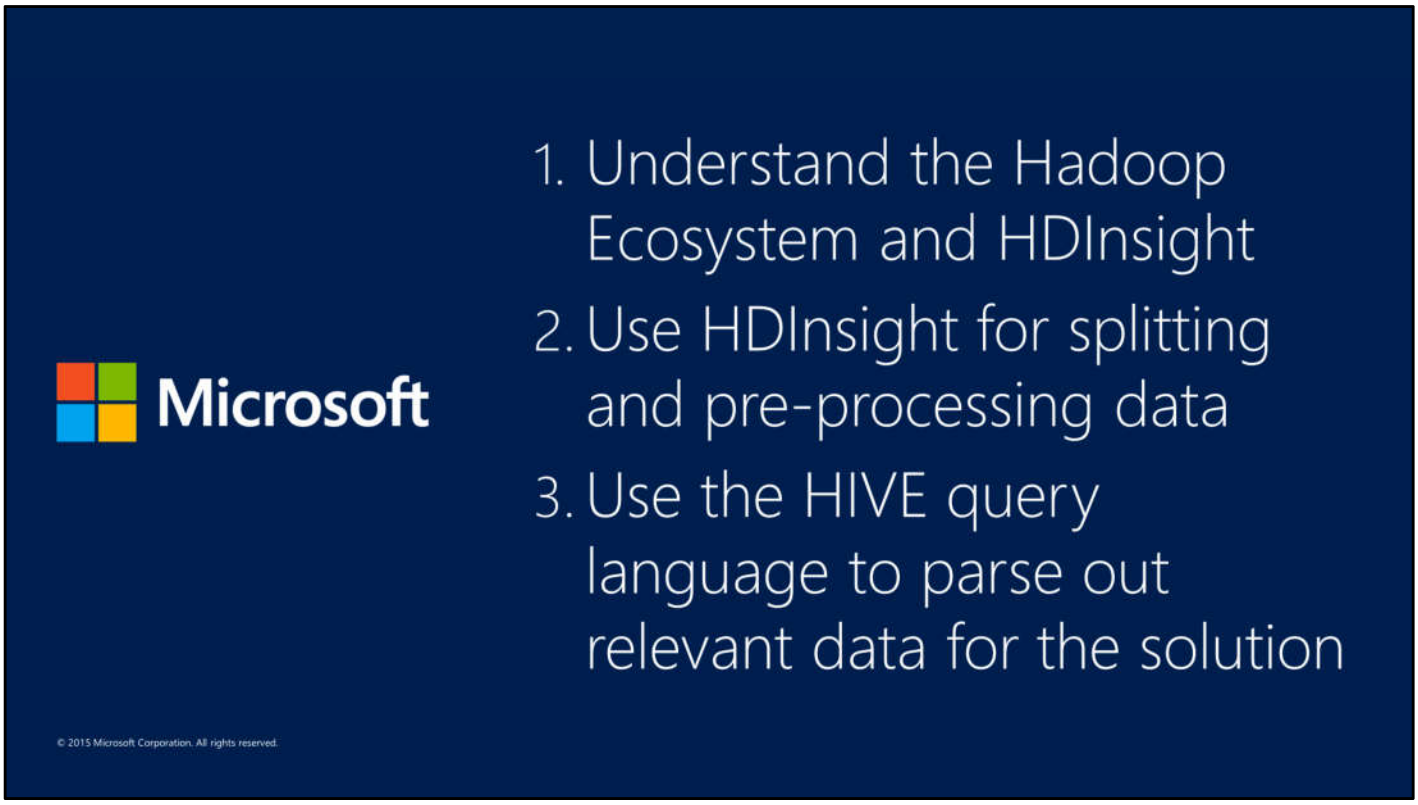
- Enter new table queries in the Query Editor
- View tables in the database explorer



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



1. Open your "Resources" Folder from your class download
2. Follow the instructions you find in the CAWHDI-SensorDataLab.docx file


Microsoft
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
© 2015 Microsoft Corporation. All rights reserved.