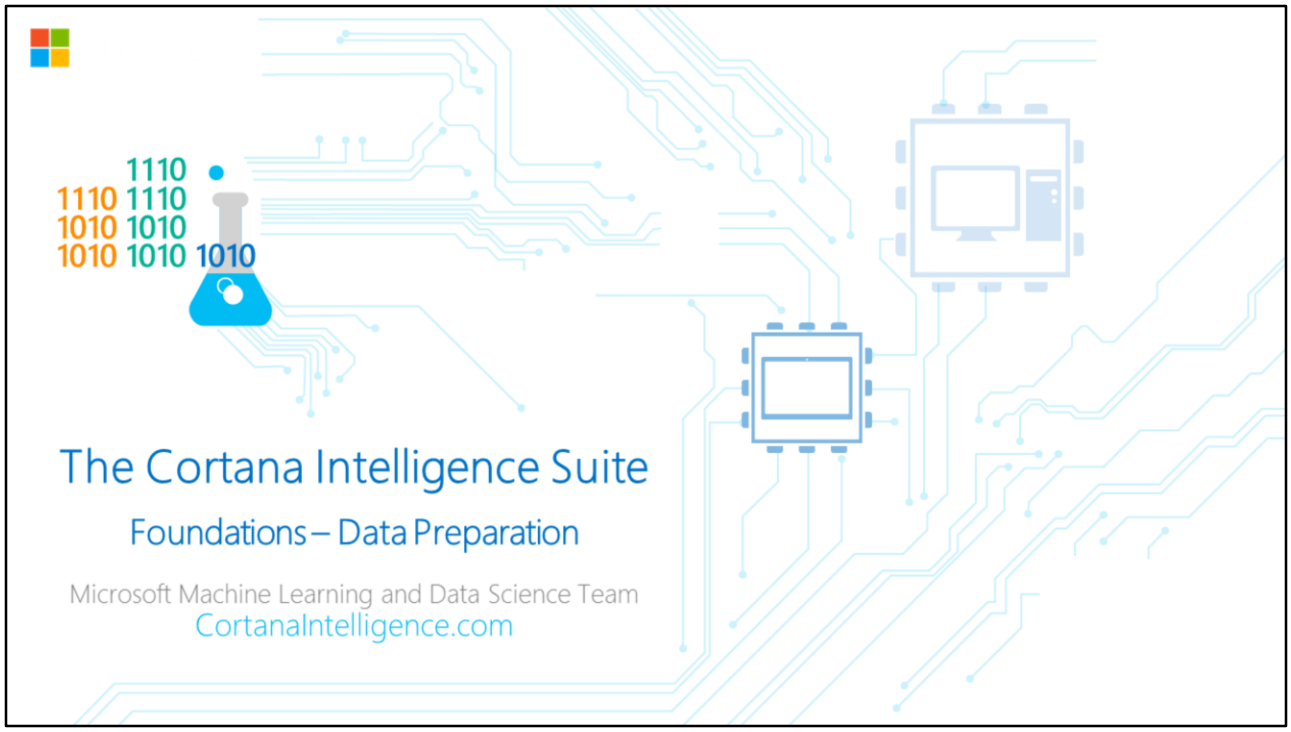


1. Main page: http://cortanaanalytics.com
2. To begin this module, you should have:
    1. Basic Math and Stats skills
    2. Business and Domain Awareness
    3. General Computing Background

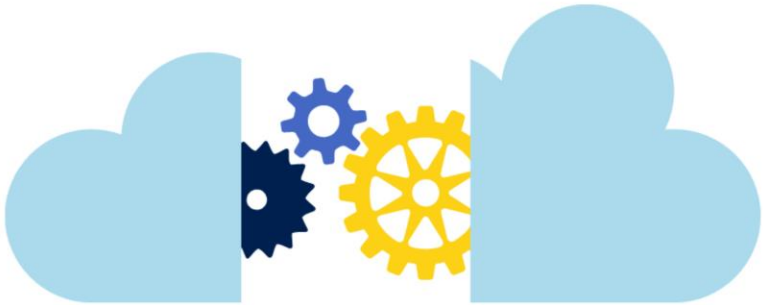
NOTE: These workbooks contain many resources to lead you through the course, and provide a rich set of references that you can use to learn much more about these topics. If the links do not resolve properly, type the link address in manually in your web browser. If the links have changed or been removed, simply enter the title of the link in a web search engine to find the new location or a corollary reference.

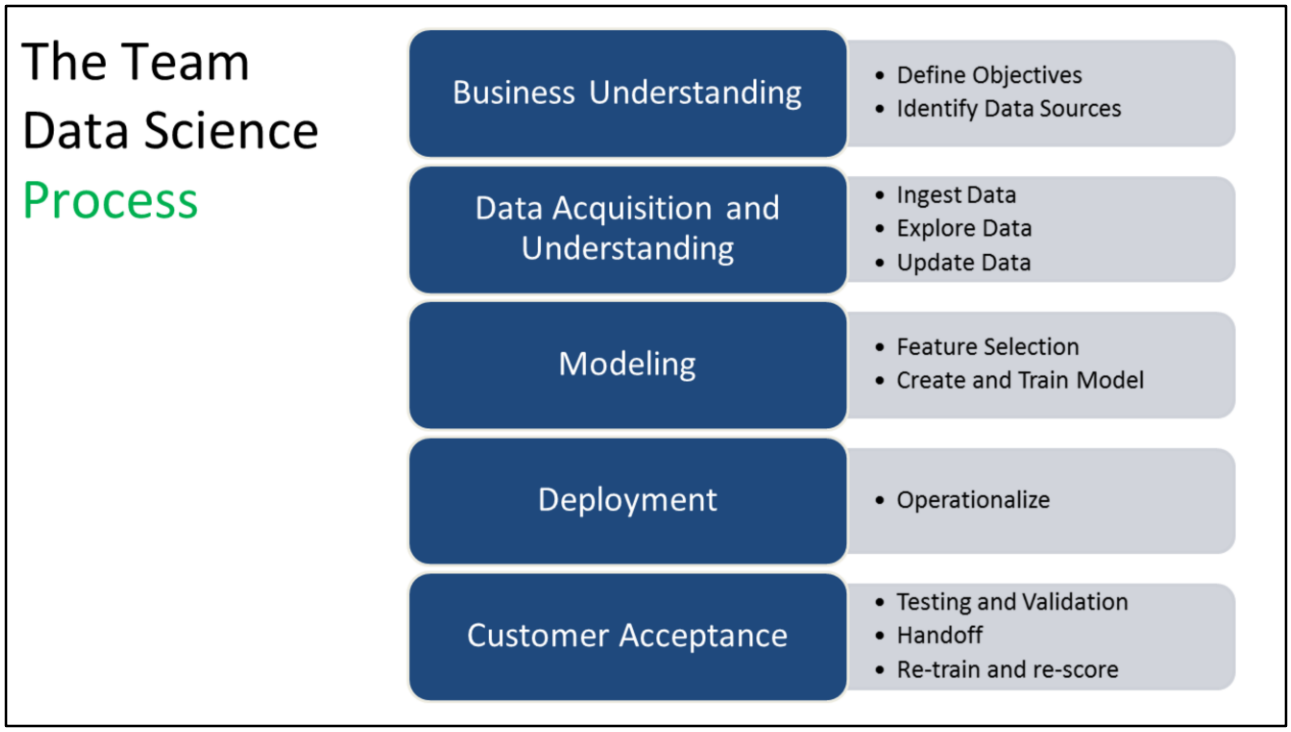
Section 3 Learning Objectives

1. Understand ADF and its constructs
2. Implement an ADF Pipeline referencing Data Sources and with various Activities
3. Understand how HDInsight can be used to process data
4. Understand the HIVE language and how it is used

1. At the end of this Module, you will:
    1. Understand ADF and its constructs
    2. Implement an ADF Pipeline referencing Data Sources and with various Activities
    3. Understand how HDInsight can be used to process data
    4. Understand the HIVE language and how it is used

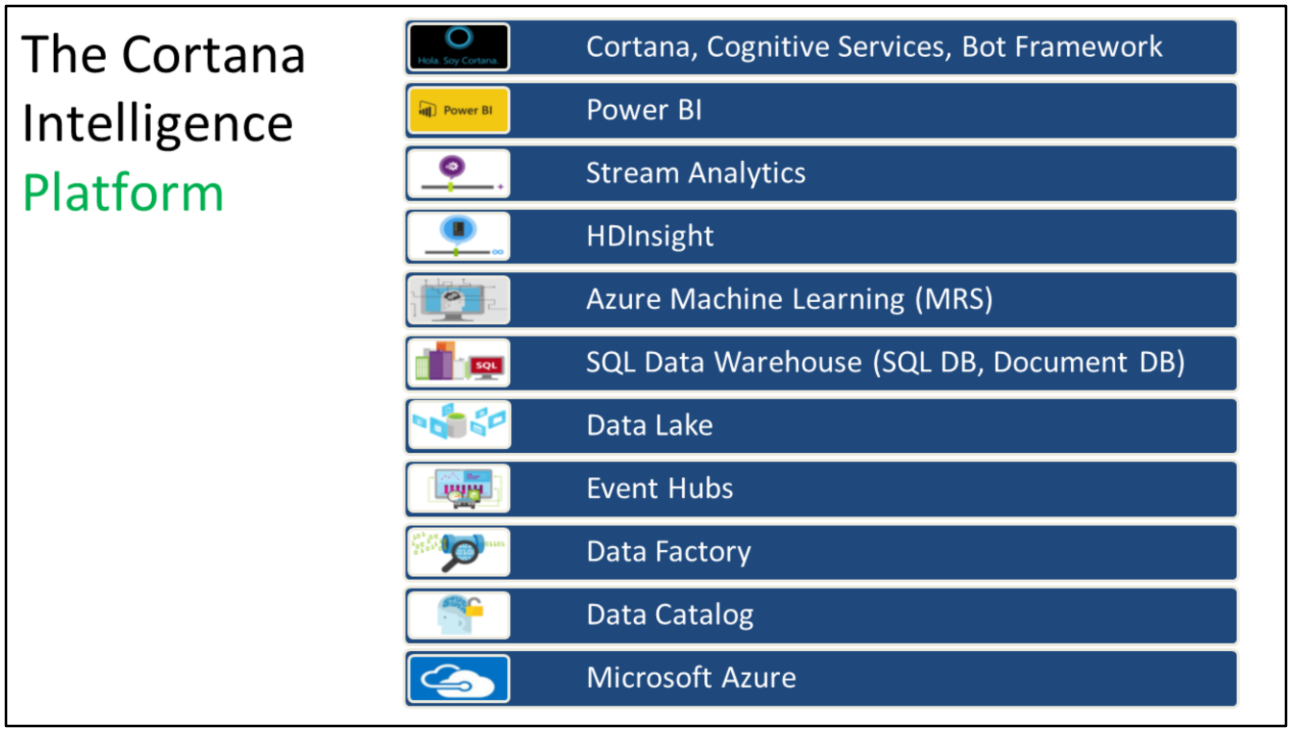
The Data Science Process and Platform





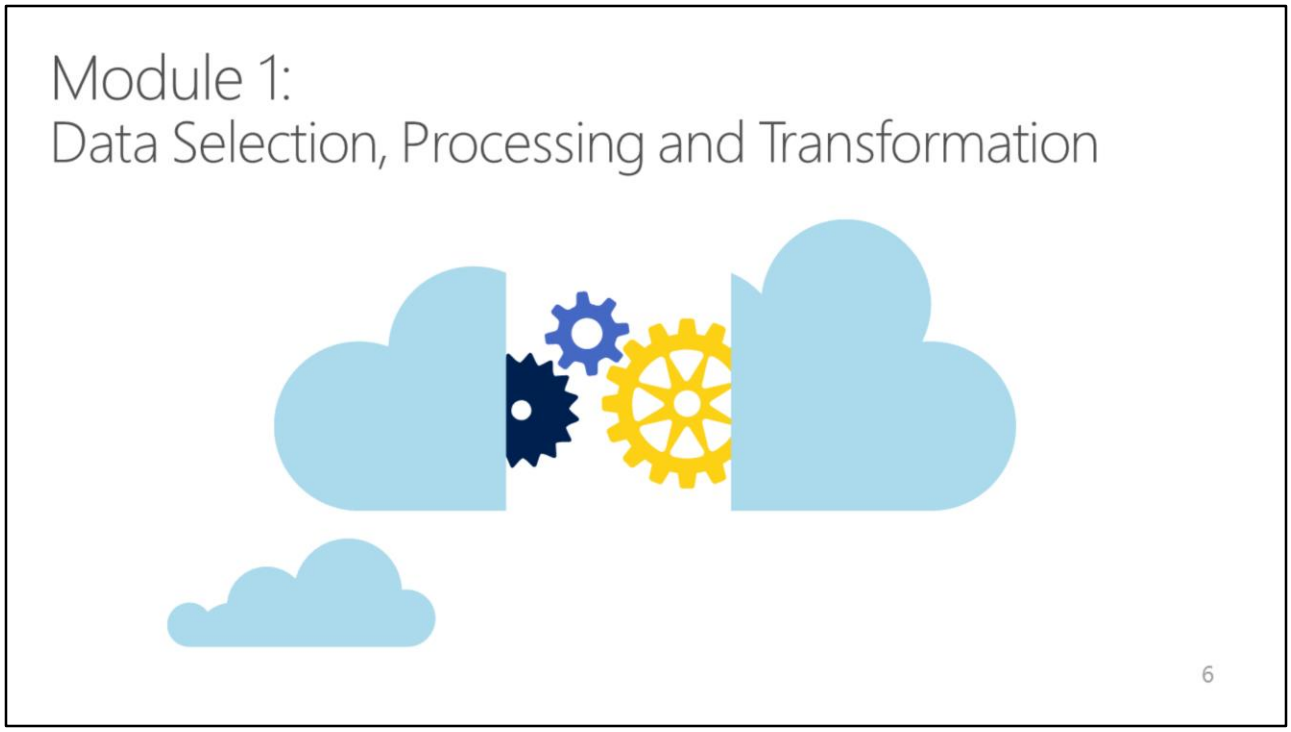
1. This process largely follows the CRISP-DM model: http://www.sv-europe.com/crisp-dm-methodology/
2. It also references the Cortana Intelligence process: https://azure.microsoft.com/en-us/documentation/articles/data-science-process-overview/
3. A complete process diagram is here: https://azure.microsoft.com/en-us/documentation/learning-paths/cortana-analytics-process/
4. Some walkthrough's of the various services: https://azure.microsoft.com/en-us/documentation/articles/data-science-process-walkthroughs/
5. An integrated process and toolset allows for a more close-to-intent deployment

6. Iterations are required to close in on the solution – but are harder tio management and monitor



1. Platform and Storage: Microsoft Azure – http://microsoftazure.com Storage: https://azure.microsoft.com/en-us/documentation/services/storage/ (Host It)
2. Azure Data Catalog: http://azure.microsoft.com/en-us/services/data-catalog (Doc It)
3. Azure Data Factory: http://azure.microsoft.com/en-us/services/data-factory/ (Move 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/ , 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/ (Scale 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/ and https://developer.microsoft.com/en-us/Cortana (Say It)
12. Cognitive Services: https://www.microsoft.com/cognitive-services
13. Bot Framework: https://dev.botframework.com/
14. All of the components within the suite: https://www.microsoft.com/en-us/server-cloud/cortana-intelligence-suite/what-is-cortana-intelligence.aspx
15. What can I do with it? https://gallery.cortanaintelligence.com/

16. Getting Started Quickly: https://caqs.azure.net/#gallery



1. Depends on the customer - https://www.microsoft.com/en-us/cloud-platform/cortana-intelligence-suite-industry-solutions

Business Case

AdventureWorks is a company that makes and sells bicycles. The sales are conducted around the world. We also support our products. But as we've made more sales in the last 10 years, we've farmed out the support function to various companies that take in maintenance and support issues in call centers around the world.

We're growing. And now we want to take our bicycles to several large retailers, but a few of them want to know a lot about our churn rate.

For over 10 years, we've collected a lot of information about our customers and of course we know a lot about our products. But since we've outsourced our call centers, we don't own the databases that hold their data – they will give us an export, though. (They support multiple customers)

We're not sure about our churn rate – we have the data of who has and has not bought again, and we think we can get the data from the call centers for the complaints and repairs, but we need a way to analyze a lot of data that has different formats to find a prediction of who will churn and who will not.

Ideally we want a list of customers we think will churn, in a structured database we could share out to our potential resellers sales staff, so they know how to target at-risk and new clients.

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

1. AdventureWorks Data Dictionary: https://technet.microsoft.com/en-us/library/ms124438(v=sql.100).aspx

Business Case

AdventureWorks is a company that makes and sells bicycles. The sales are conducted around the world. We also support our products. But as we've made more sales in the last 10 years, we've farmed out the support function to various companies that take in maintenance and support issues in call centers around the world.

We're growing. And now we want to take our bicycles to several large retailers, but a few of them want to know a lot about our churn rate.

For over 10 years, we've collected a lot of information about our customers and of course we know a lot about our products. But since we've outsourced our call centers, we don't own the databases that hold their data – they will give us an export, though. (They support multiple customers)

We're not sure about our churn rate – we have the data of who has and has not bought again, and we think we can get the data from the call centers for the complaints and repairs, but we need a way to analyze a lot of data that has different formats to find a prediction of who will churn and who will not.

Ideally we want a list of customers we think will churn, in a structured database we could share out to our potential resellers sales staff, so they know how to target at-risk and new clients.

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

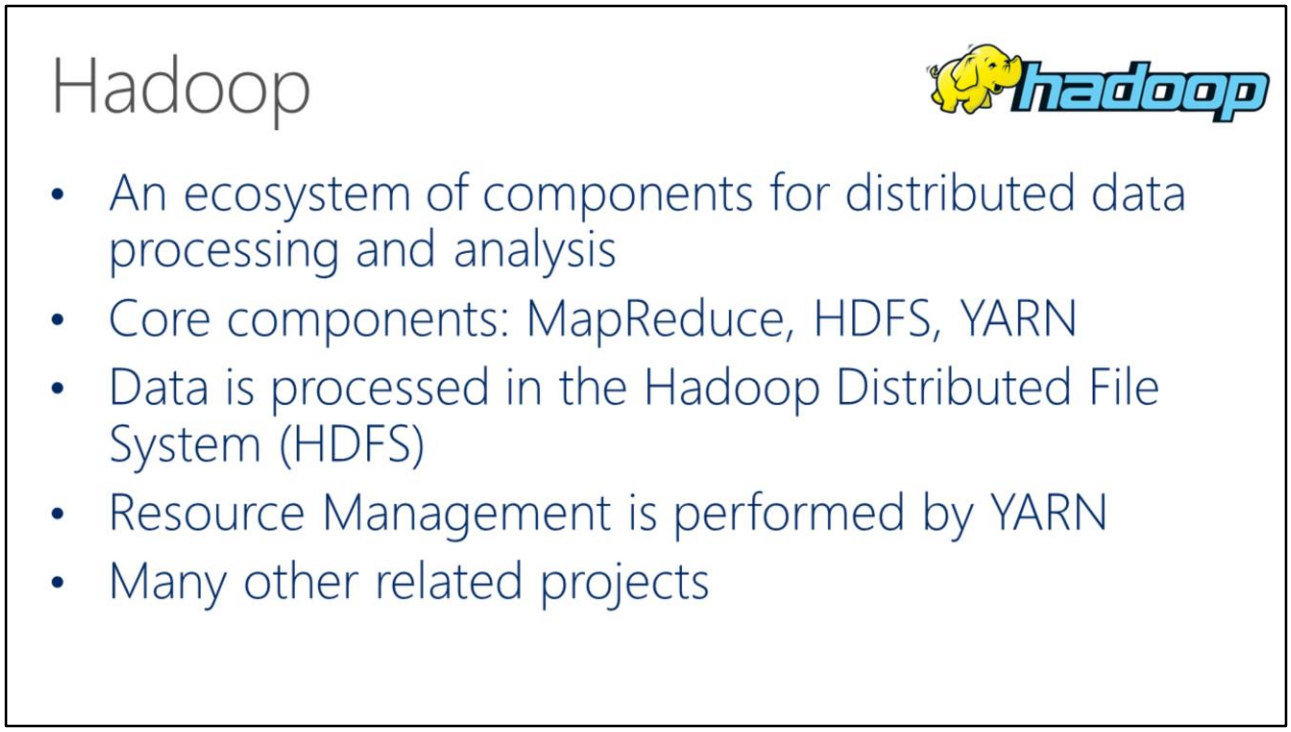
1. AdventureWorks Data Dictionary: https://technet.microsoft.com/en-us/library/ms124438(v=sql.100).aspx

Module 2:
HDInsight for Data Manipulation and Processing

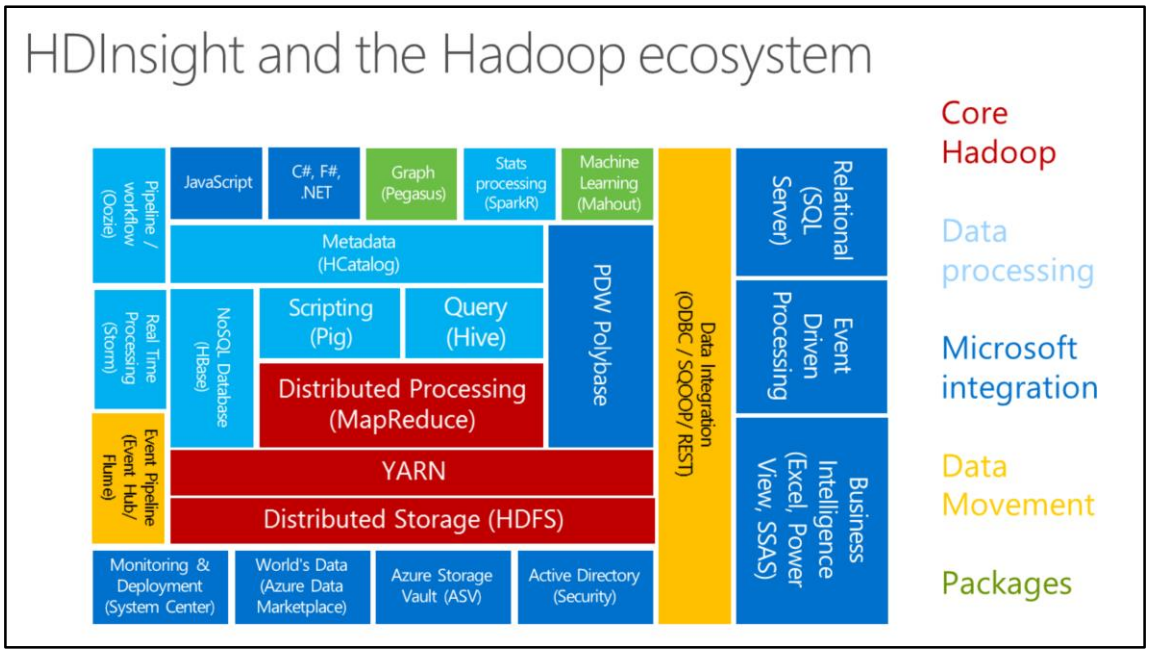
9

1. Processing, querying, and transforming data using HDInsight: https://msdn.microsoft.com/en-us/library/dn749822.aspx

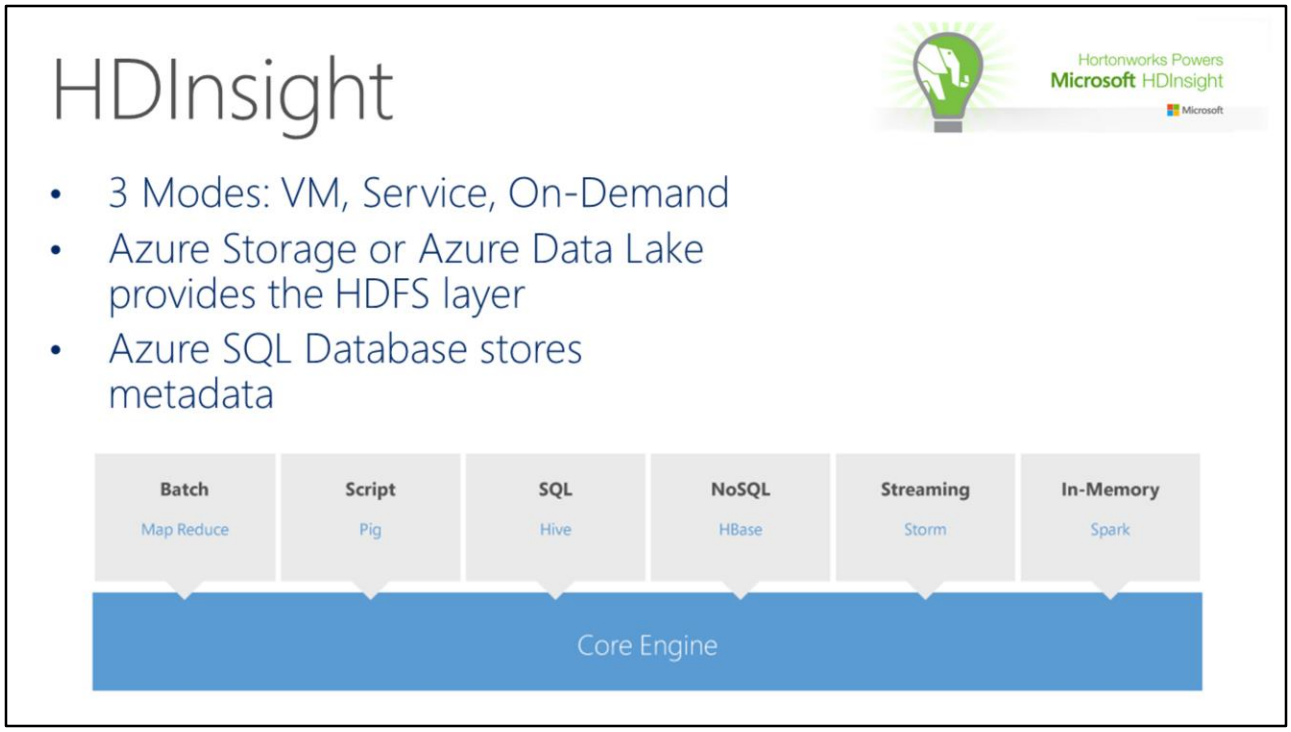
1. Primary site: https://azure.microsoft.com/en-us/services/hdinsight/
2. Quick overview: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/
3. 4-week online course through the edX platform: https://www.edx.org/course/processing-big-data-azure-hdinsight-microsoft-dat202-1x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



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



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



1. Main page: https://azure.microsoft.com/en-us/documentation/services/hdinsight/
2. Pricing for HDInsight: https://azure.microsoft.com/en-us/pricing/details/hdinsight/
3. 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

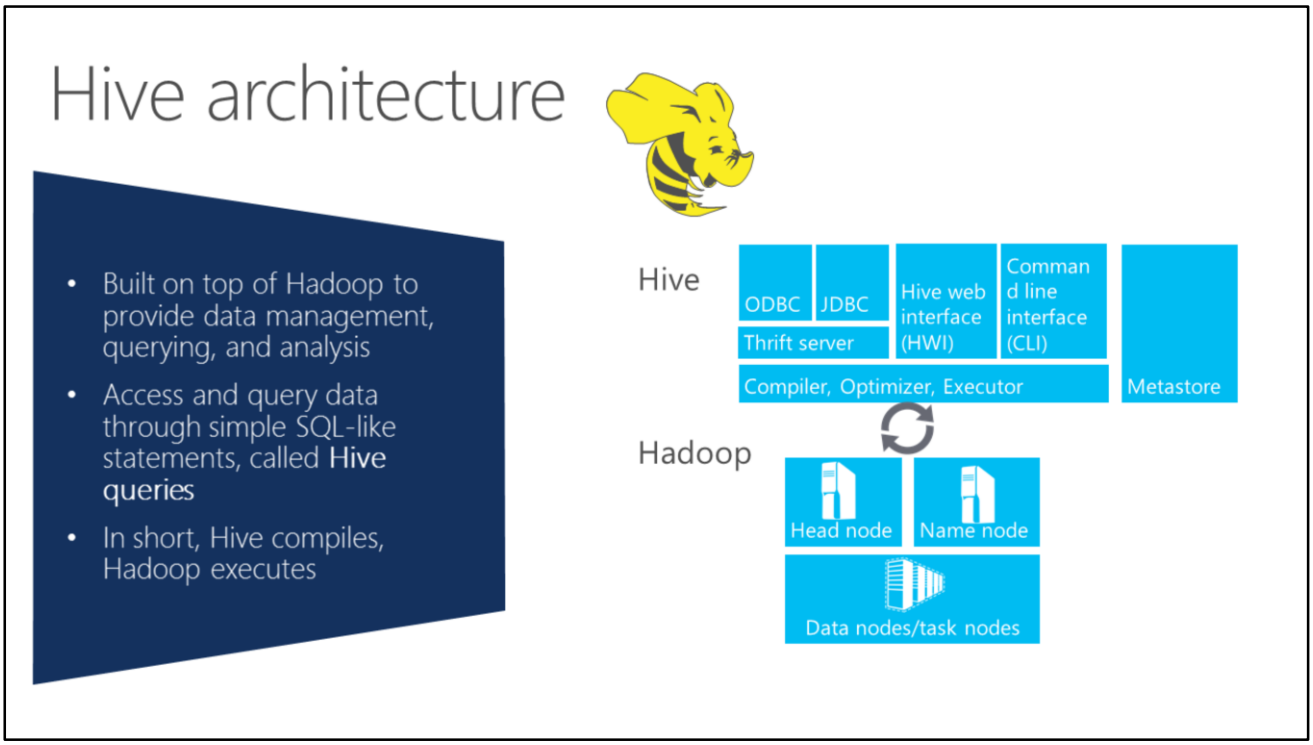
- Cluster Type: Hadoop, Spark, HBase and Storm.
  - Hadoop clusters: for query and analysis workloads
  - HBase clusters: for NoSQL workloads
  - Spark clusters: for in-memory processing, interactive queries, stream, and machine learning workloads
- Operating System: Windows or Linux
- Can be deployed 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

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

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.

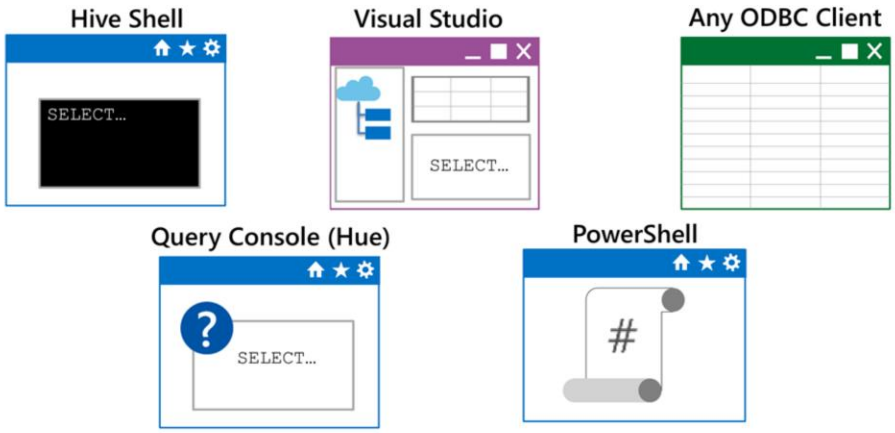
- Hive for HDInsight: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/
- Referencing user defined functions with Hive: https://msdn.microsoft.com/en-us/library/dn749875.aspx?f=255&MSPPError=-2147217396
- Using Apache Tez for improved performance: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/#usetez



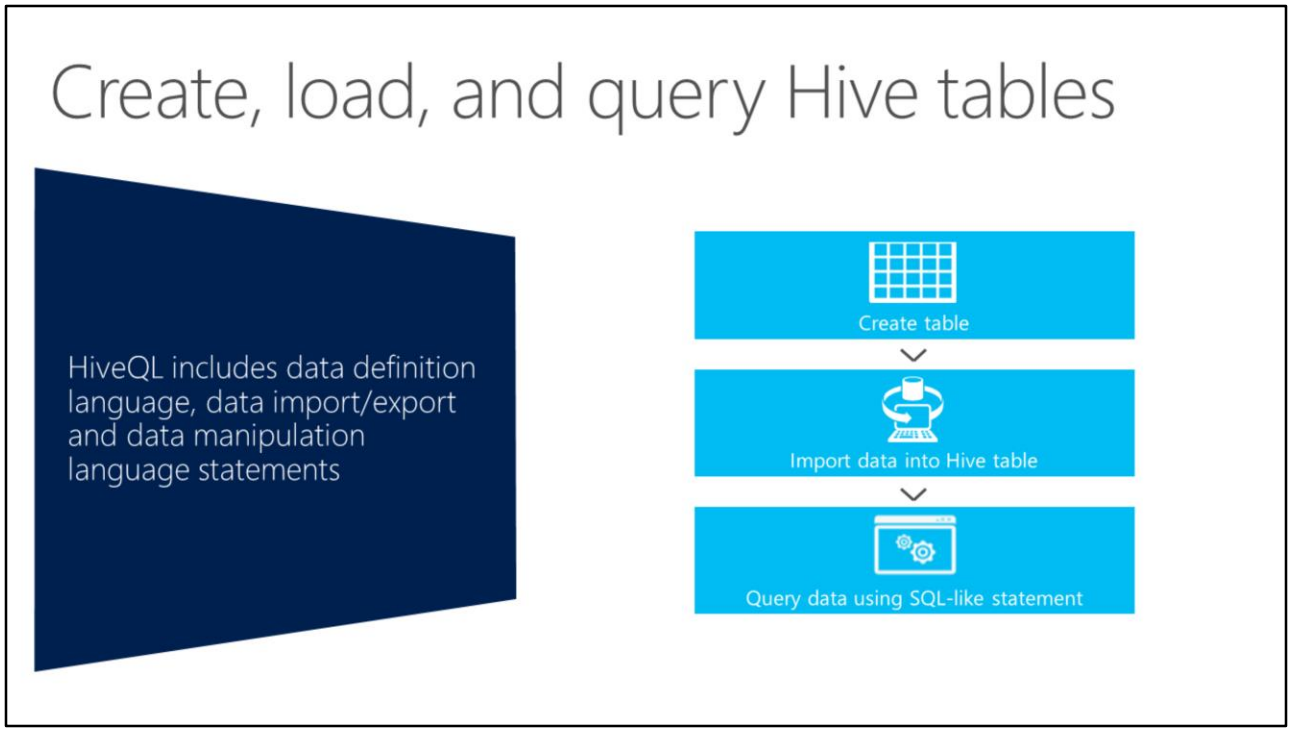
1. Hive for HDInsight: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/
2. Referencing user defined functions with Hive: https://msdn.microsoft.com/en-us/library/dn749875.aspx?f=255&MSPPError=-2147217396
3. Using Apache Tez for improved performance: https://azure.microsoft.com/en-us/documentation/articles/hdinsight-use-hive/#usetez

Hive Client Tools

- You can submit Hive Jobs using many different tools



1. https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-hive-odbc-driver/



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

Module 3:
Azure Data Factory

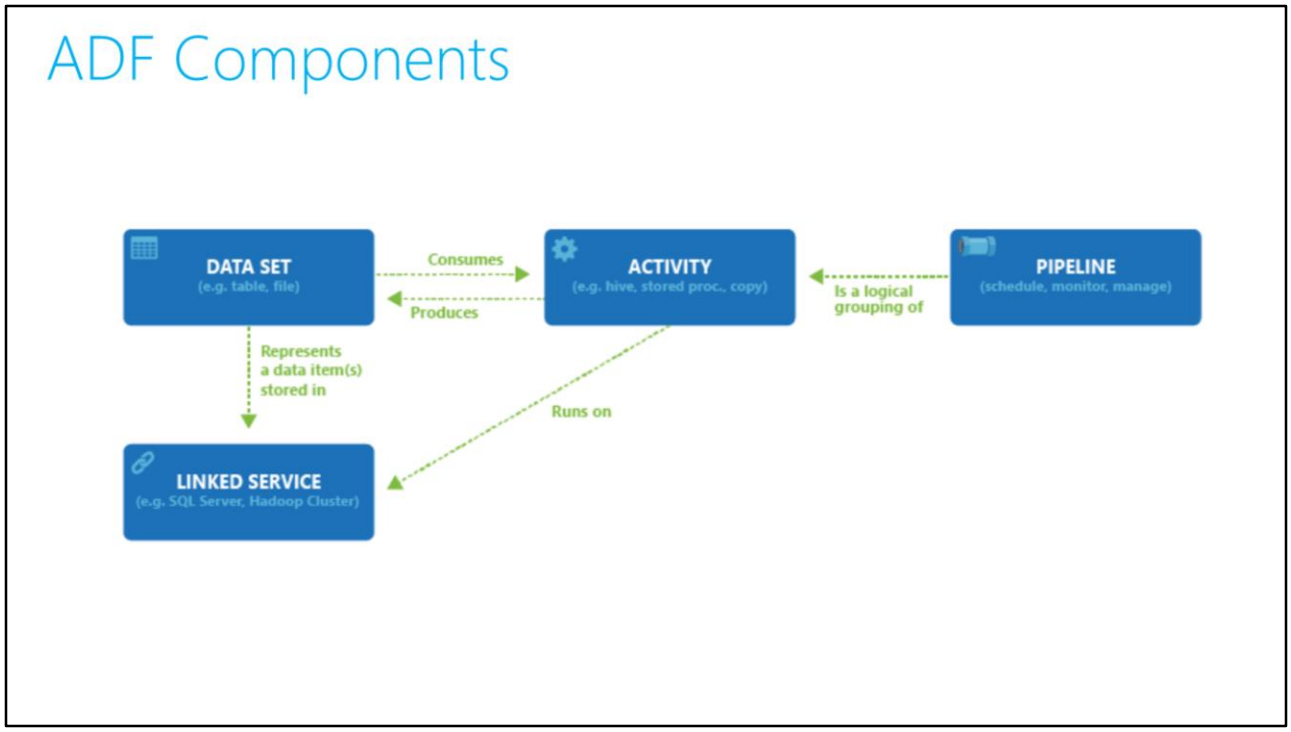
19

1. Primary Site: https://azure.microsoft.com/en-us/services/data-factory/
2. 2-minute overview video: https://channel9.msdn.com/Blogs/Windows-Azure/Introduction-to-Azure-Data-Factory/

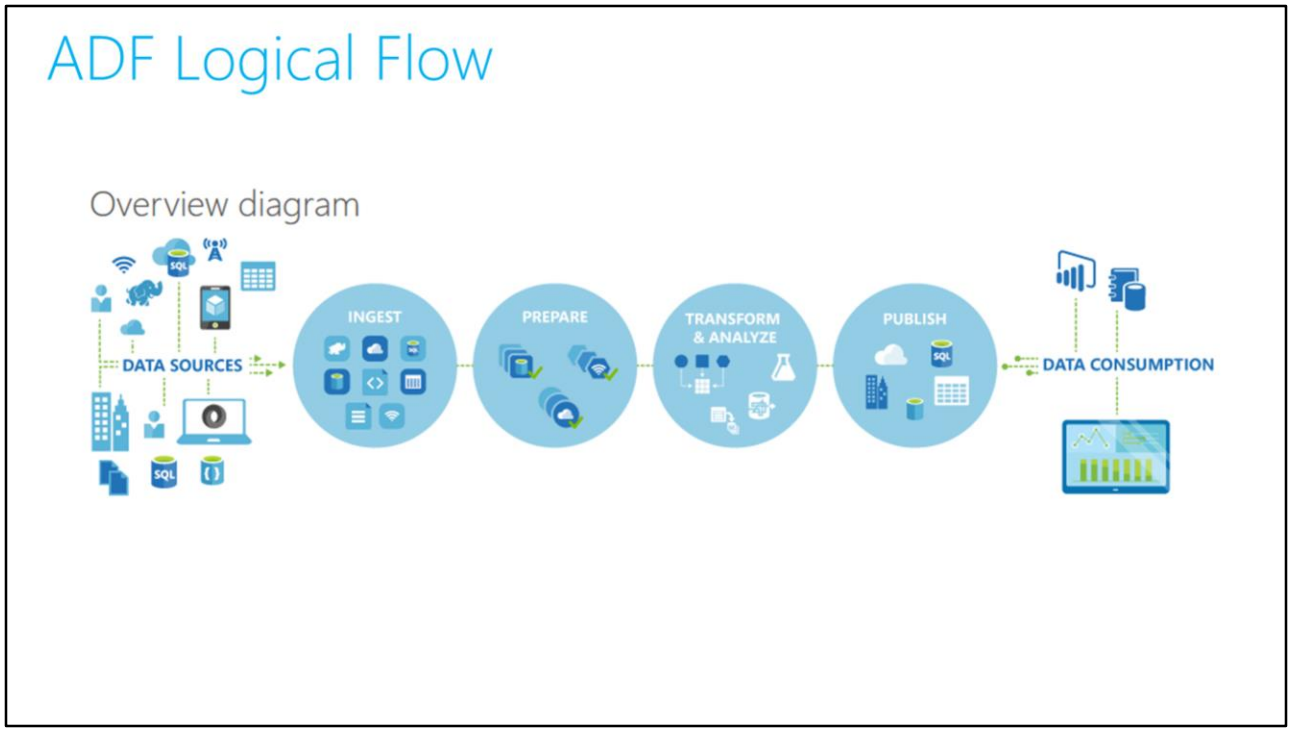
Azure Data Factory

Create, orchestrate, and manage data movement and enrichment through the cloud

1. Learning Path: https://azure.microsoft.com/en-us/documentation/articles/data-factory-introduction/
2. Developer Reference: https://msdn.microsoft.com/en-us/library/azure/dn834987.aspx



1. Pricing: https://azure.microsoft.com/en-us/pricing/details/data-factory/

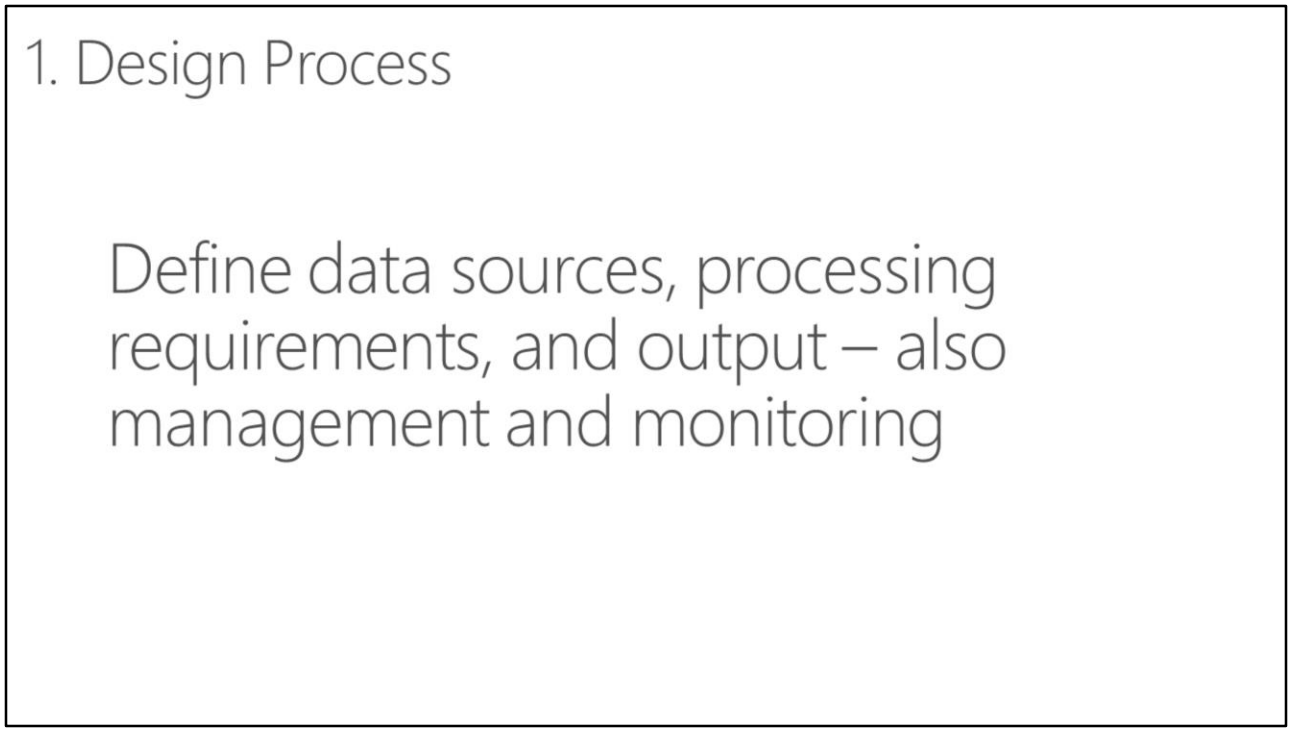


1. Learning Path: https://azure.microsoft.com/en-us/documentation/articles/data-factory-introduction/
2. Quick Example: http://azure.microsoft.com/blog/2015/04/24/azure-data-factory-update-simplified-sample-deployment/

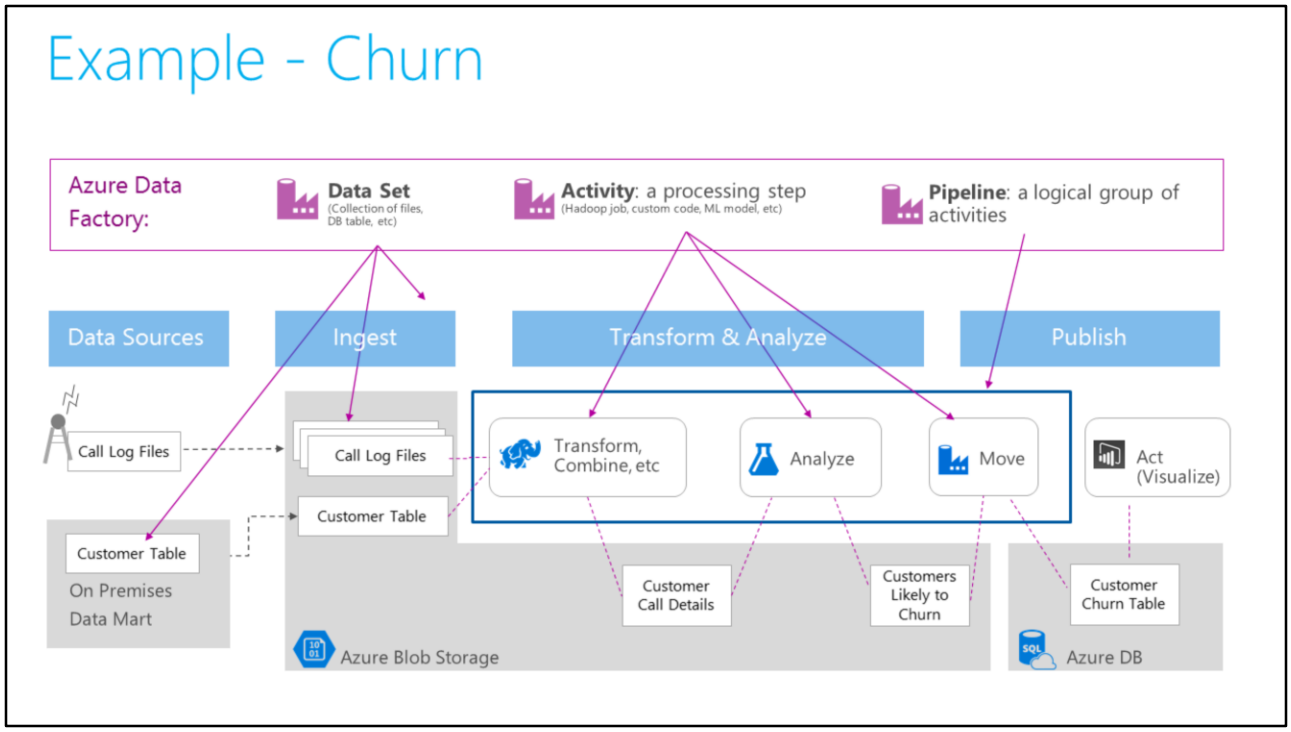
ADF Process

1. Define Architecture: Set up objectives and flow
2. Create the Data Factory: Portal, PowerShell, VS
3. Create Linked Services: Connections to Data and Services
4. Create Datasets: Input and Output
5. Create Pipeline: Define Activities
6. Monitor and Manage: Portal or PowerShell, Alerts and Metrics

1. Full Tutorial: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline/



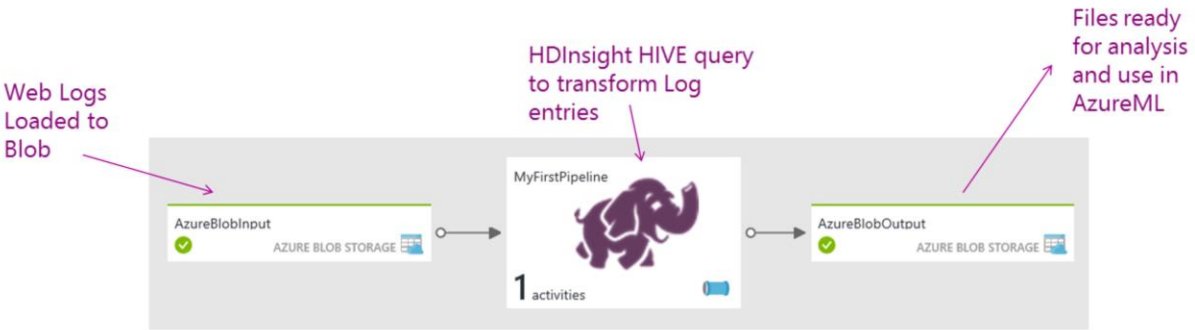
1. https://azure.microsoft.com/en-us/documentation/articles/data-factory-customer-profiling-usecase/



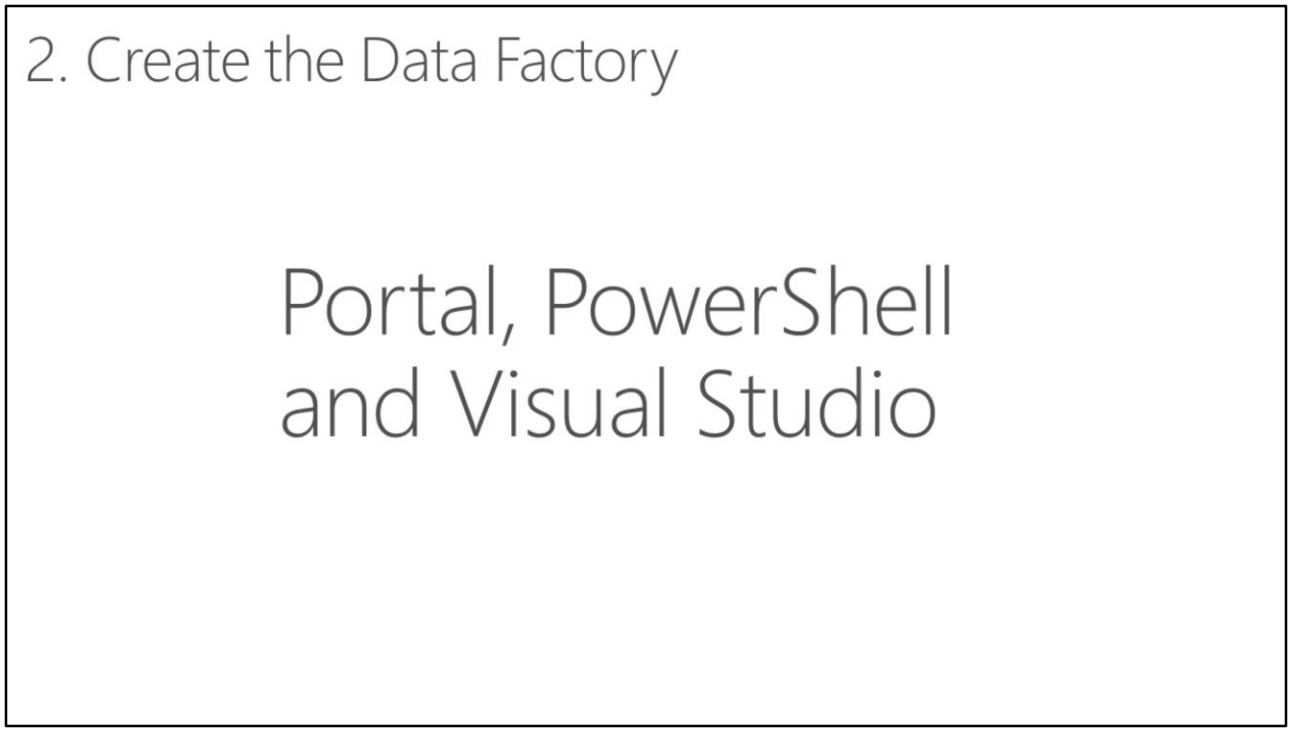
1. Video of this process: https://azure.microsoft.com/en-us/documentation/videos/azure-data-factory-102-analyzing-complex-churn-models-with-azure-data-factory/

Simple ADF:

- Business Goal: Transform and Analyze Web Logs each month
- Design Process: Transform Raw Weblogs, using a Hive Query, storing the results in Blob Storage

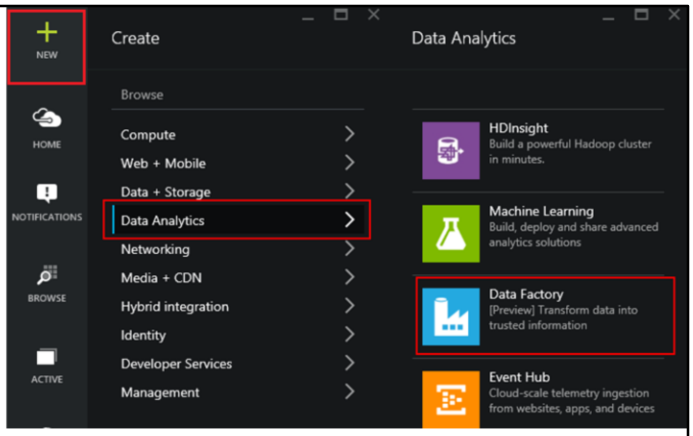


1. More options: Prepare System: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline-using-editor/ - Follow steps
2. Another Lab: https://azure.microsoft.com/en-us/documentation/articles/data-factory-samples/



1. Setting Up: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline/

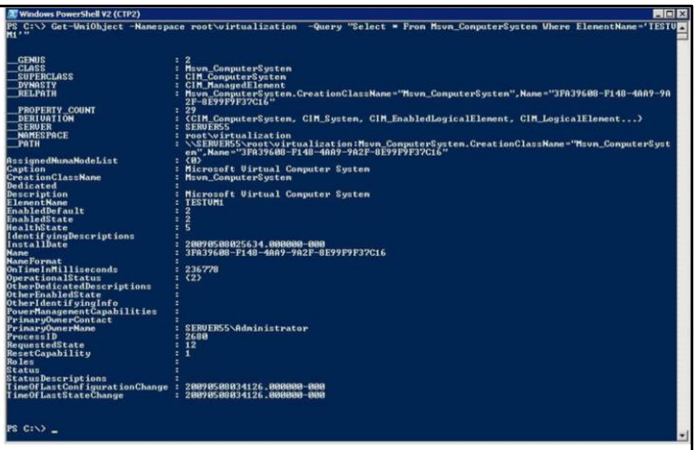
Using the Portal



- Use in Non-MS Clients
- Use for Exploration
- Use when teaching or in a Demo

1. Overview: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline/
2. Using the Portal: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline-using-editor/

Using PowerShell



- Use in MS Clients
- Use for Automation
- Use for quick set up and tear down

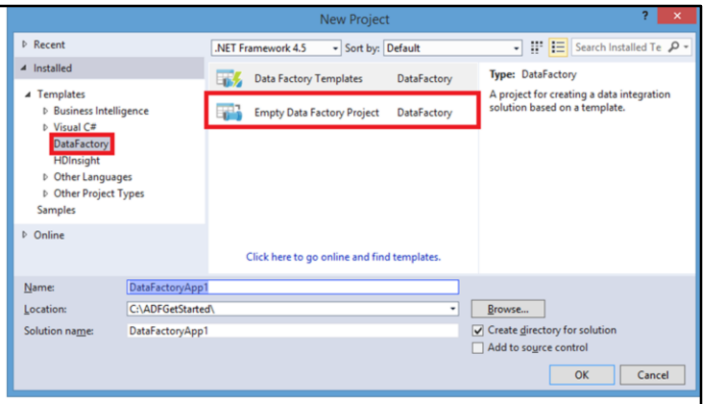
1. Learning Path: https://azure.microsoft.com/en-us/documentation/articles/data-factory-introduction/
2. Full Tutorial: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline/

PowerShell ADF Example

1. Run **`Add-AzureAccount`** and enter the user name and password
2. Run **`Get-AzureSubscription`** to view all the subscriptions for this account.
3. Run **`Select-AzureSubscription`** to select the subscription that you want to work with.
4. Run **`Switch-AzureMode AzureResourceManager`**
5. Run **`New-AzureResourceGroup -Name ADFTutorialResourceGroup -Location "West US"`**
6. Run **`New-AzureDataFactory -ResourceGroupName ADFTutorialResourceGroup -Name DataFactory(your alias)Pipeline -Location "West US"`**

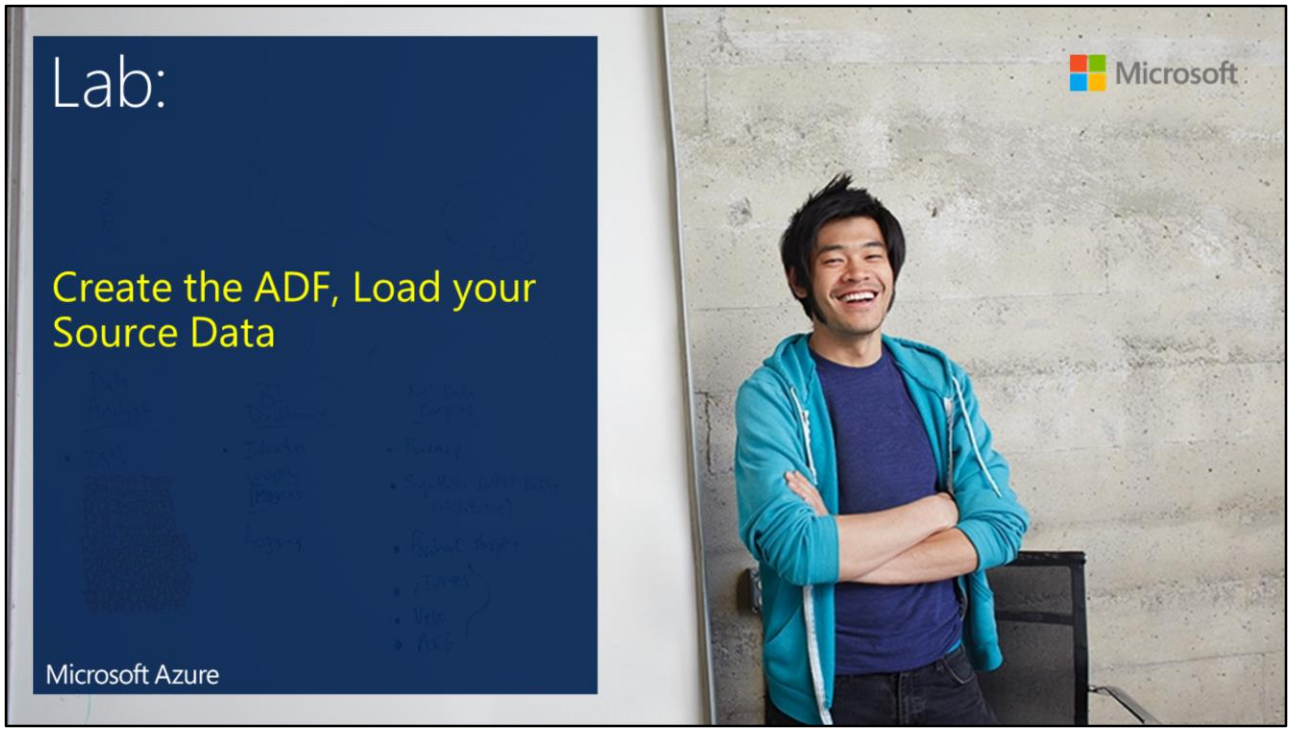
1. https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline-using-powershell/

Using Visual Studio

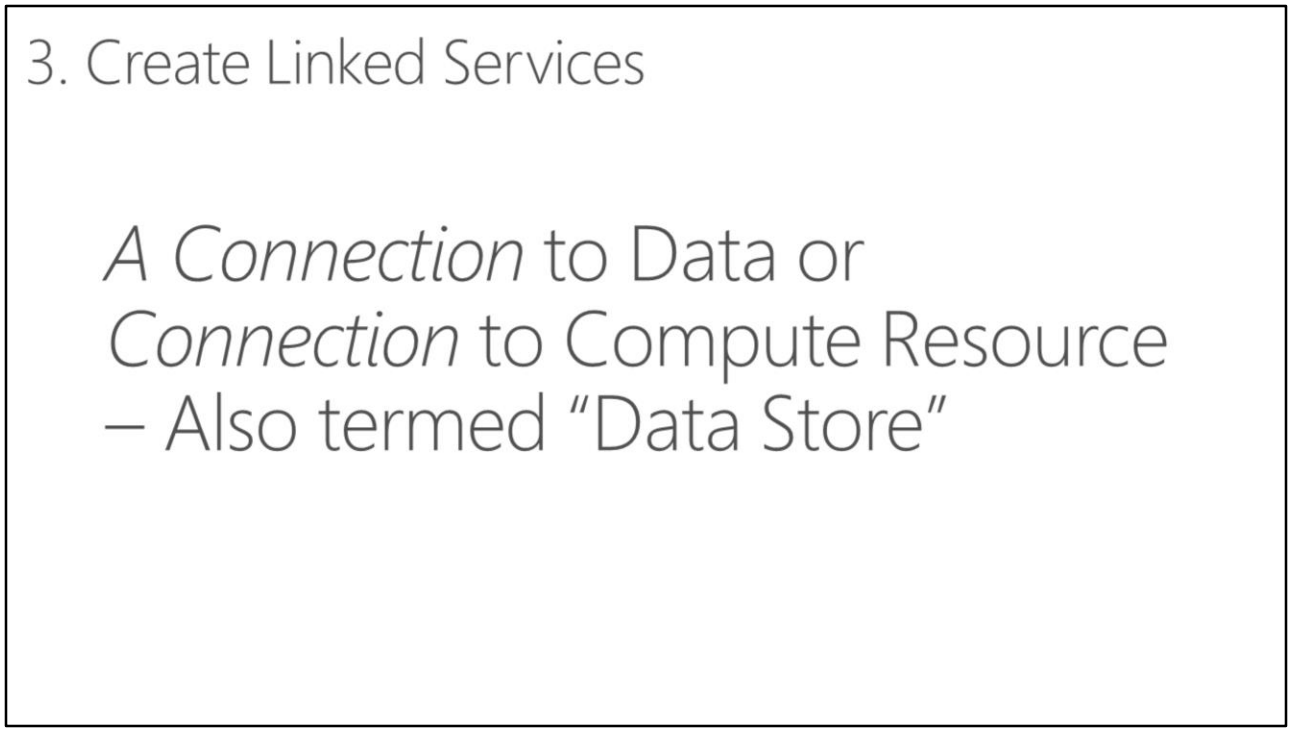


- Use in mature dev environments
- Use when integrated into larger development process

1. Overview: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline/
2. Using the Portal: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline-using-editor/



1. Open the **ADF Student Workbook** file from your \Resources folder
2. Follow the steps for **Lab 1**
3. The follow the steps for **Lab 2**
4. Note – There's a useful JSON prettifier here: http://www.jsoneditoronline.org/



1. Data Linking: https://azure.microsoft.com/en-us/documentation/articles/data-factory-data-movement-activities/
2. Compute Linking: https://azure.microsoft.com/en-us/documentation/articles/data-factory-compute-linked-services/

Data Options

Source	Sink
Blob	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, DocumentDB, OnPrem File System, Data Lake Store
Table	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, DocumentDB, Data Lake Store
SQL Database	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, DocumentDB, Data Lake Store
SQL Data Warehouse	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, DocumentDB, Data Lake Store
DocumentDB	Blob, Table, SQL Database, SQL Data Warehouse, Data Lake Store
Data Lake Store	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, DocumentDB, OnPrem File System, Data Lake Store
SQL Server on IaaS	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, Data Lake Store
OnPrem File System	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, OnPrem File System, Data Lake Store
OnPrem SQL Server	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, Data Lake Store
OnPrem Oracle Database	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, Data Lake Store
OnPrem MySQL Database	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, Data Lake Store
OnPrem DB2 Database	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, Data Lake Store
OnPrem Teradata Database	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, Data Lake Store
OnPrem Sybase Database	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, Data Lake Store
OnPrem PostgreSQL Database	Blob, Table, SQL Database, SQL Data Warehouse, OnPrem SQL Server, SQL Server on IaaS, Data Lake Store

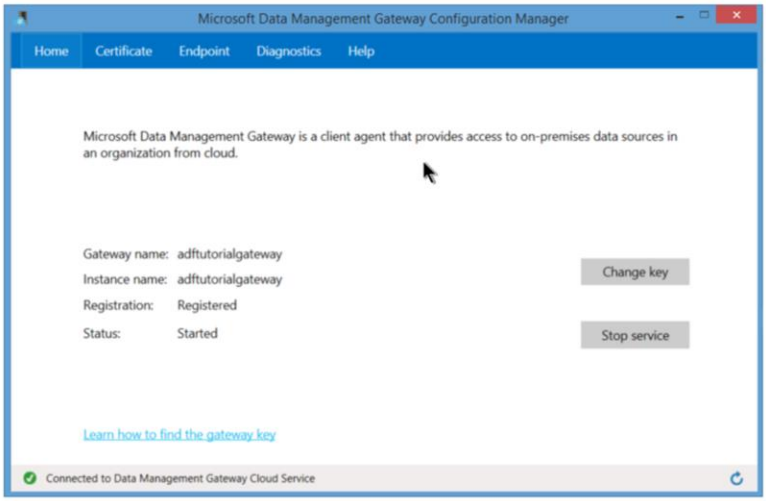
1. Data Movement requirements: https://azure.microsoft.com/en-us/documentation/articles/data-factory-data-movement-activities/
2. From on-premises, requires Data Management Gateway: https://azure.microsoft.com/en-us/documentation/articles/data-factory-move-data-between-onprem-and-cloud/

Activity Options

Transformation activity	Compute environment
Hive	HDInsight [Hadoop]
Pig	HDInsight [Hadoop]
MapReduce	HDInsight [Hadoop]
Hadoop Streaming	HDInsight [Hadoop]
Machine Learning activities: Batch Execution and Update Resource	Azure VM
Stored Procedure	Azure SQL
Data Lake Analytics U-SQL	Azure Data Lake Analytics
DotNet	HDInsight [Hadoop] or Azure Batch

1. Main Document Site: https://azure.microsoft.com/en-us/documentation/articles/data-factory-data-transformation-activities/

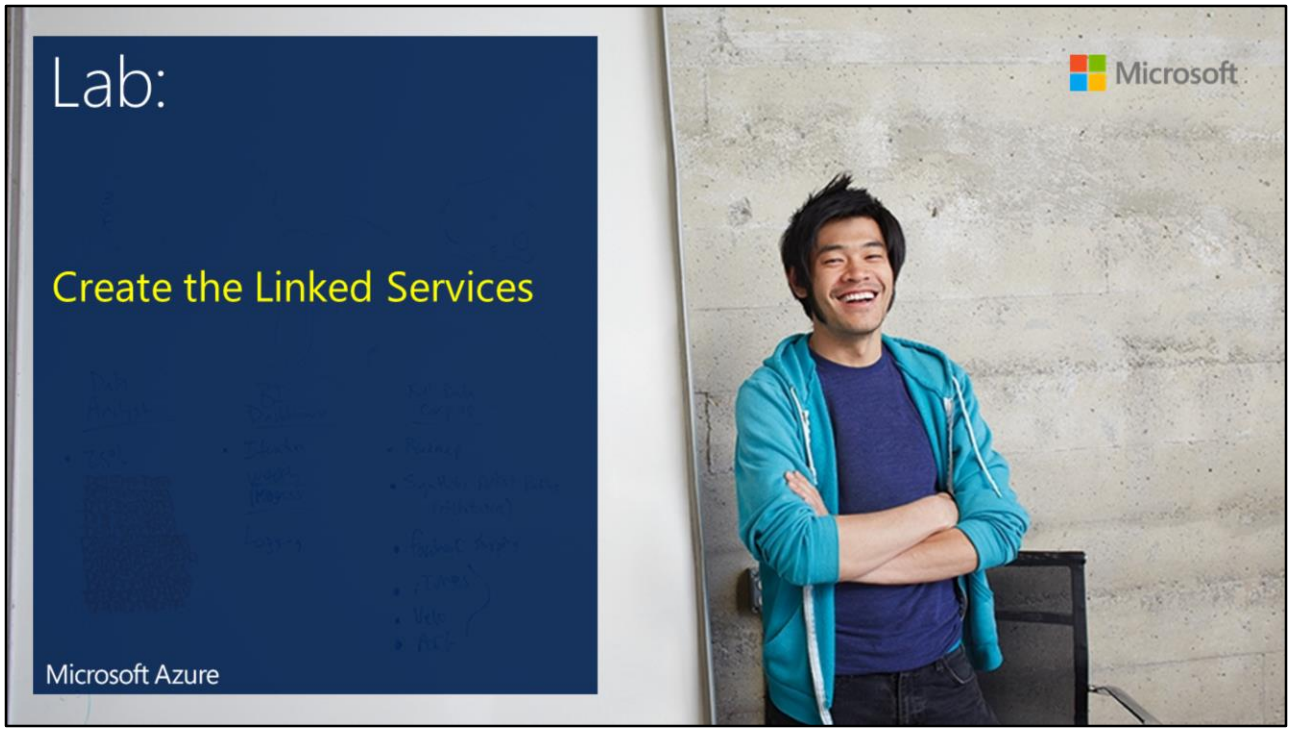
Gateway for On-Prem



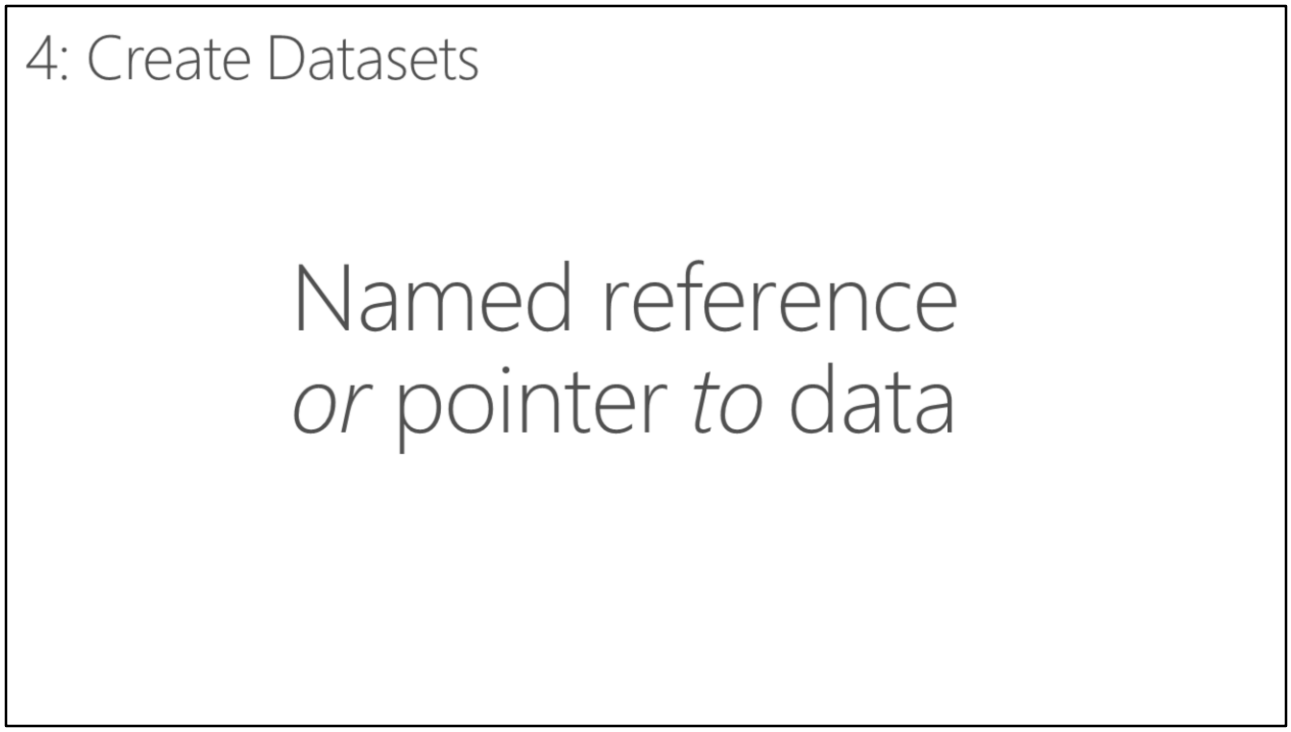
1. Activities: https://azure.microsoft.com/en-us/documentation/articles/data-factory-create-pipelines/

CreateHDICluster.txt:

```
{
  "name": "HDInsightOnDemandLinkedService",
  "properties": {
    "type": "HDInsightOnDemand",
    "typeProperties": {
      "version": "3.1",
      "clusterSize": 1,
      "timeToLive": "00:30:00",
      "linkedServiceName": "StorageLinkedService"
    }
  }
}
```



1. Open the **ADF Student Workbook** file from your \Resources folder
2. Follow the steps for Lab 3



1. Main Dataset Document Site: https://azure.microsoft.com/en-us/documentation/articles/data-factory-create-datasets/

Dataset Concepts

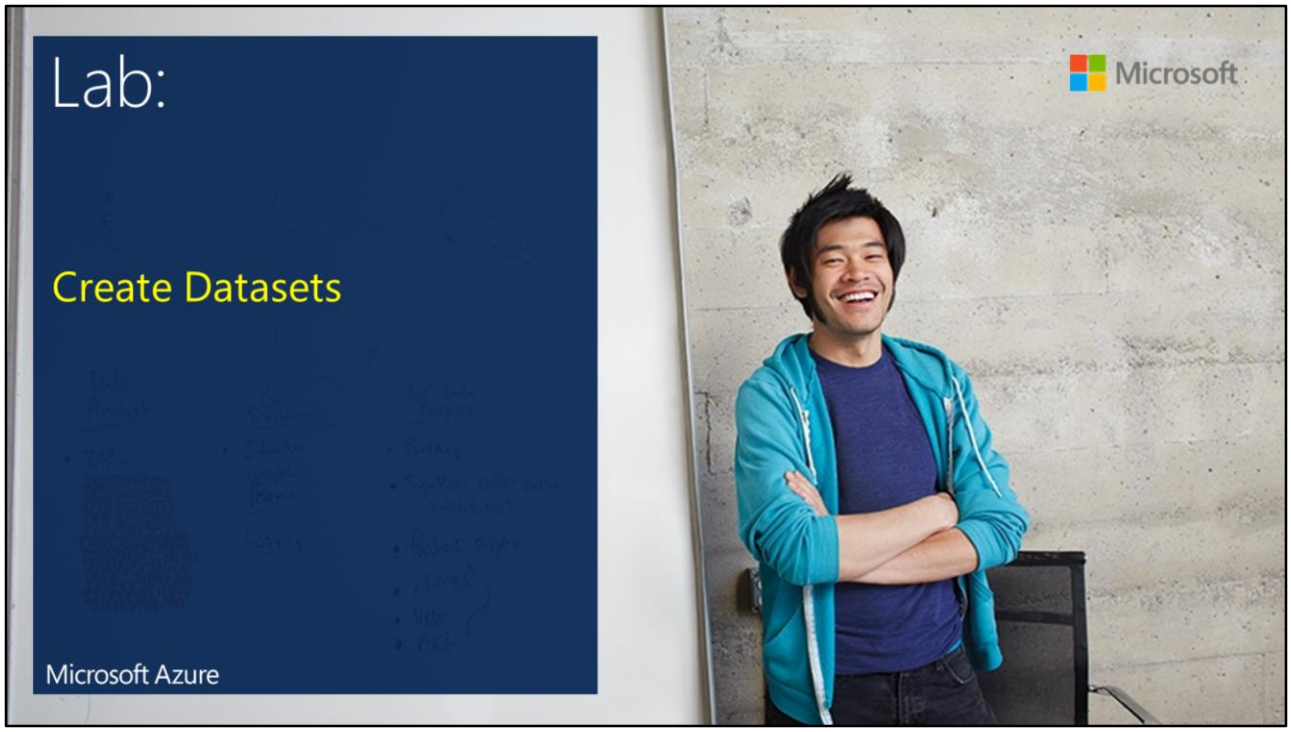
```
{
   "name": "<name of dataset>",
   "properties":
   {
      "structure": [ ],
      "type": "<type of dataset>",
       "external": <boolean flag to indicate external data>,
       "typeProperties":
       {
       },
       "availability":
       {

       },
      "policy":
       {

       }
   }.
```

1. https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline-using-editor/

AzureBlobOutput.txt:

```
{
  "name": "AzureBlobOutput",
  "properties": {
    "type": "AzureBlob",
    "linkedServiceName": "StorageLinkedService",
    "typeProperties": {
      "folderPath": "data/partitioneddata",
      "format": {
        "type": "TextFormat",
        "columnDelimiter": ","
      }
    },
    "availability": {
      "frequency": "Month",
      "interval": 1
    }
  }
```

```
}
```



1. Open the **ADF Student Workbook** file from your \Resources folder
2. Follow the steps for Lab 4

5. Create Pipelines

Logical Grouping of Activities

1. Main Pipeline Documentation: https://azure.microsoft.com/en-us/documentation/articles/data-factory-create-pipelines/

Pipeline Concepts

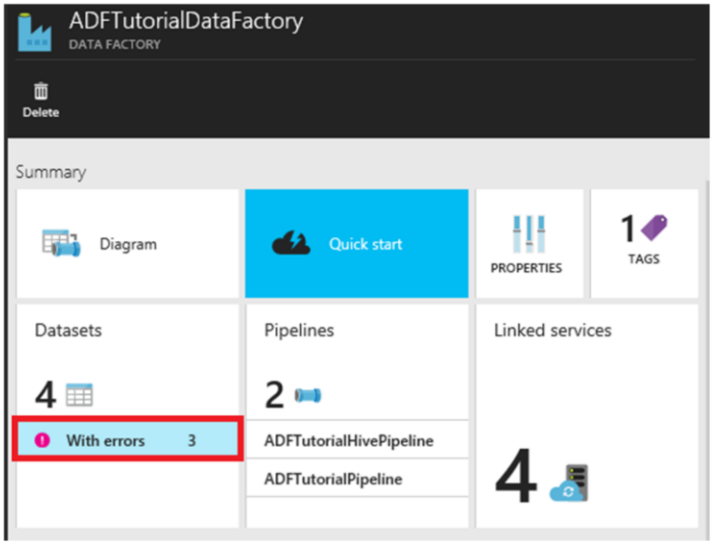
```
{
   "name": "PipelineName",
   "properties":
   {
      "description" : "pipeline description",
      "activities":
      [

      ],
      "start": "<start date-time>",
      "end": "<end date-time>"
   }
}
```

1. https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline-using-editor/
2. Activities: https://azure.microsoft.com/en-us/documentation/articles/data-factory-create-pipelines/

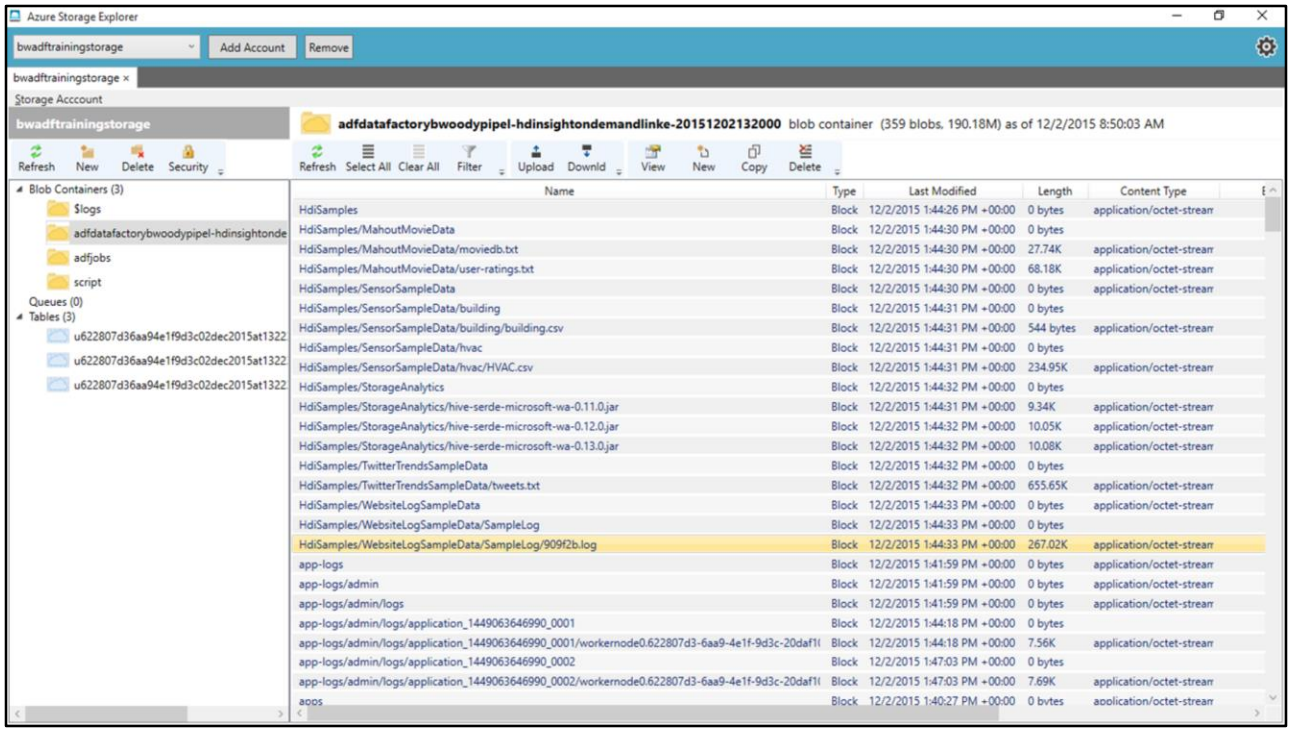
1. Open the **ADF Student Workbook** file from your \Resources folder
2. Follow the steps for Lab 5

1. Main Concepts: https://azure.microsoft.com/en-us/documentation/articles/data-factory-monitor-manage-pipelines/

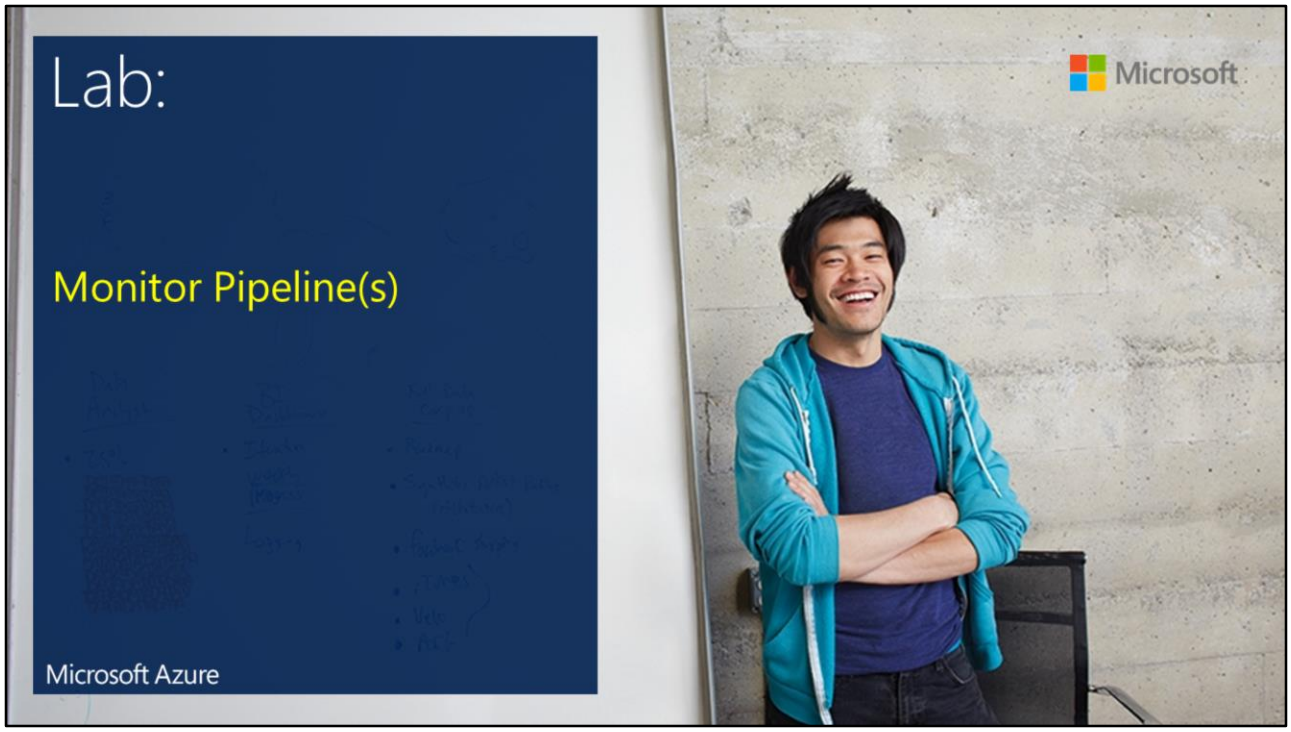
Locating Failures within a Pipeline



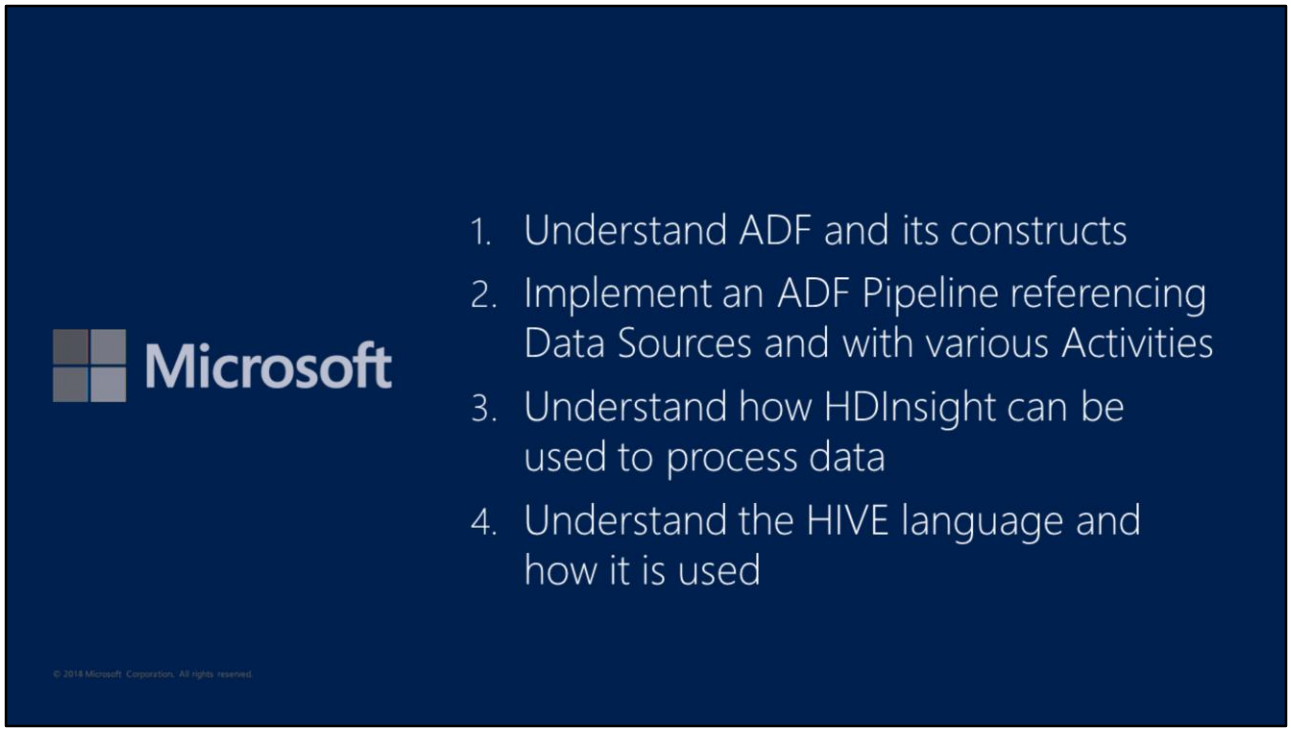
1. PowerShell script to help deal with errors in ADF: http://blogs.msdn.com/b/karang/archive/2015/11/13/azure-data-factory-detecting-and-re-running-failed-adf-slices.aspx



1. PowerShell script to help deal with errors in ADF: http://blogs.msdn.com/b/karang/archive/2015/11/13/azure-data-factory-detecting-and-re-running-failed-adf-slices.aspx



1. Open the **ADF Student Workbook** file from your \Resources folder
2. Follow the steps for Lab 6


Microsoft
1. Understand ADF and its constructs
2. Implement an ADF Pipeline referencing Data Sources and with various Activities
3. Understand how HDInsight can be used to process data
4. Understand the HIVE language and how it is used
© 2016 Microsoft Corporation. All rights reserved.

Questions?