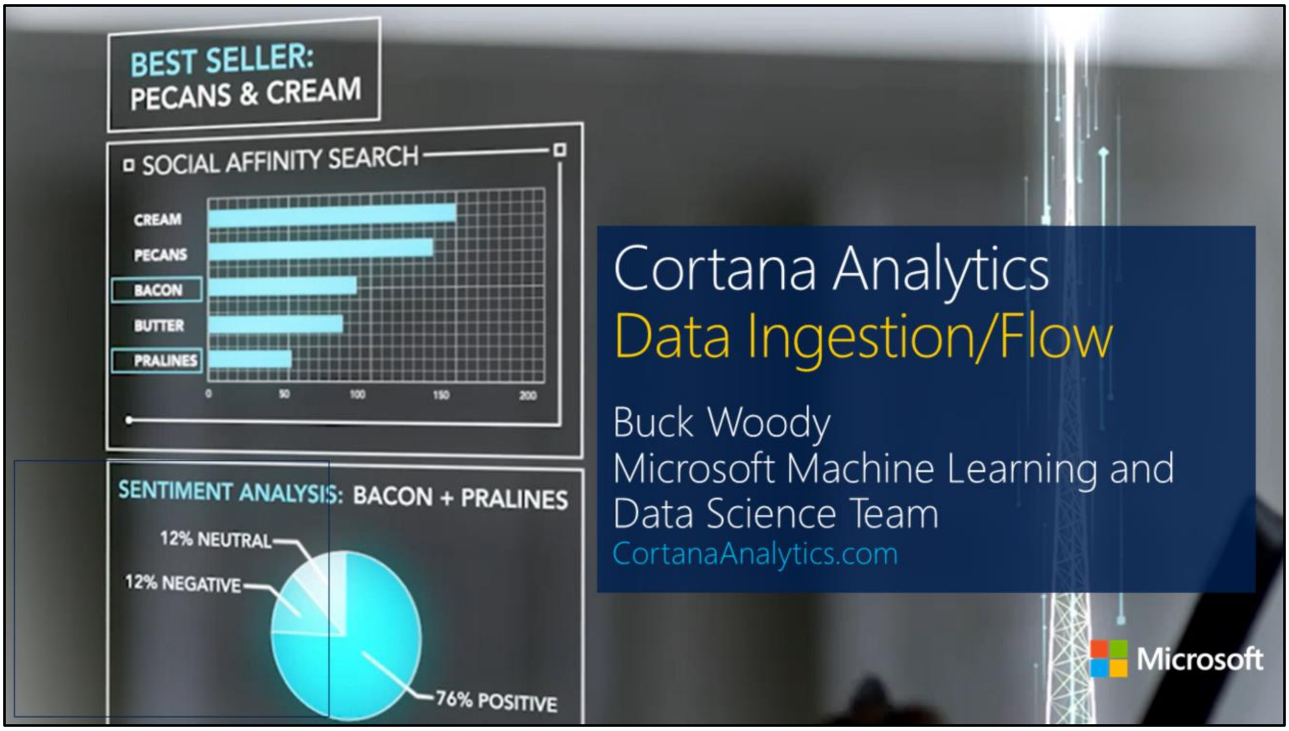
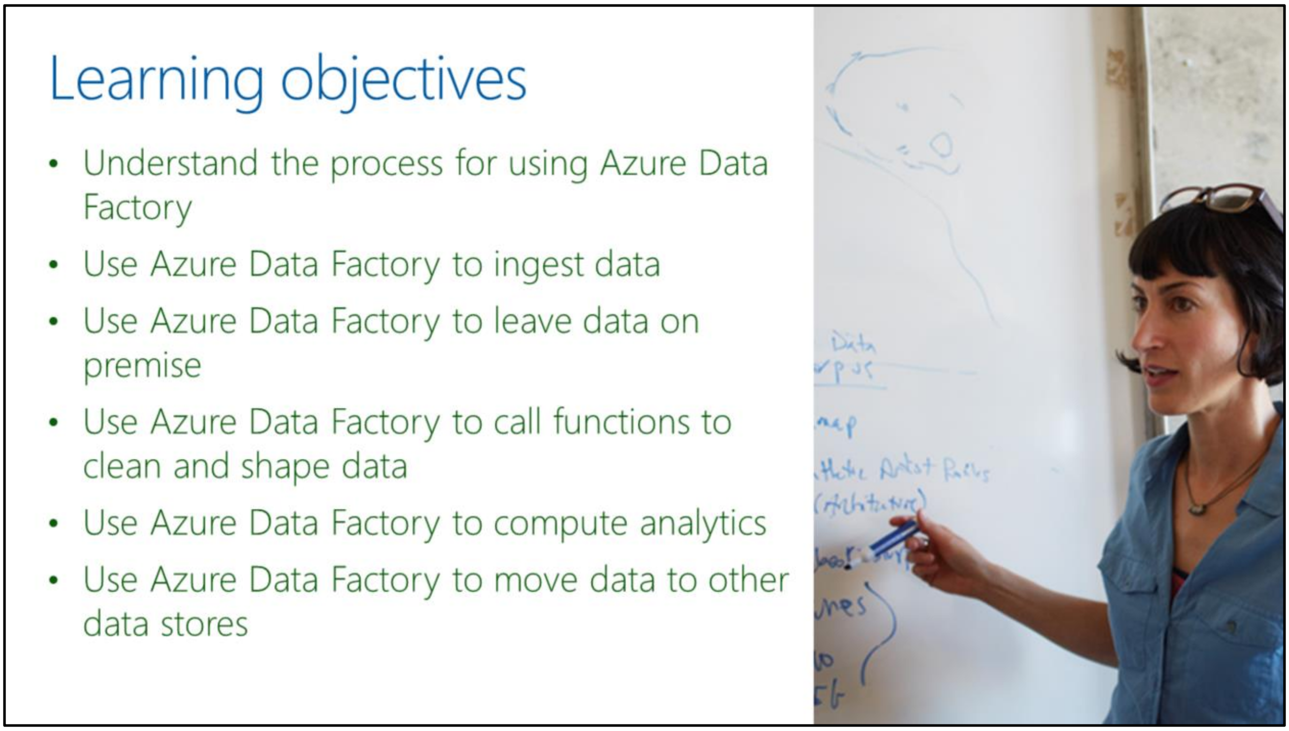


BEST SELLER:
PECANS & CREAM
SOCIAL AFFINITY SEARCH
CREAM
PECANS
BACON
BUTTER
PRALINES
0
50
100
150
200
SENTIMENT ANALYSIS: BACON + PRALINES
12% NEUTRAL
12% NEGATIVE
76% POSITIVE
Microsoft

Cortana Analytics
Data Ingestion/Flow

Buck Woody
Microsoft Machine Learning and Data Science Team
CortanaAnalytics.com

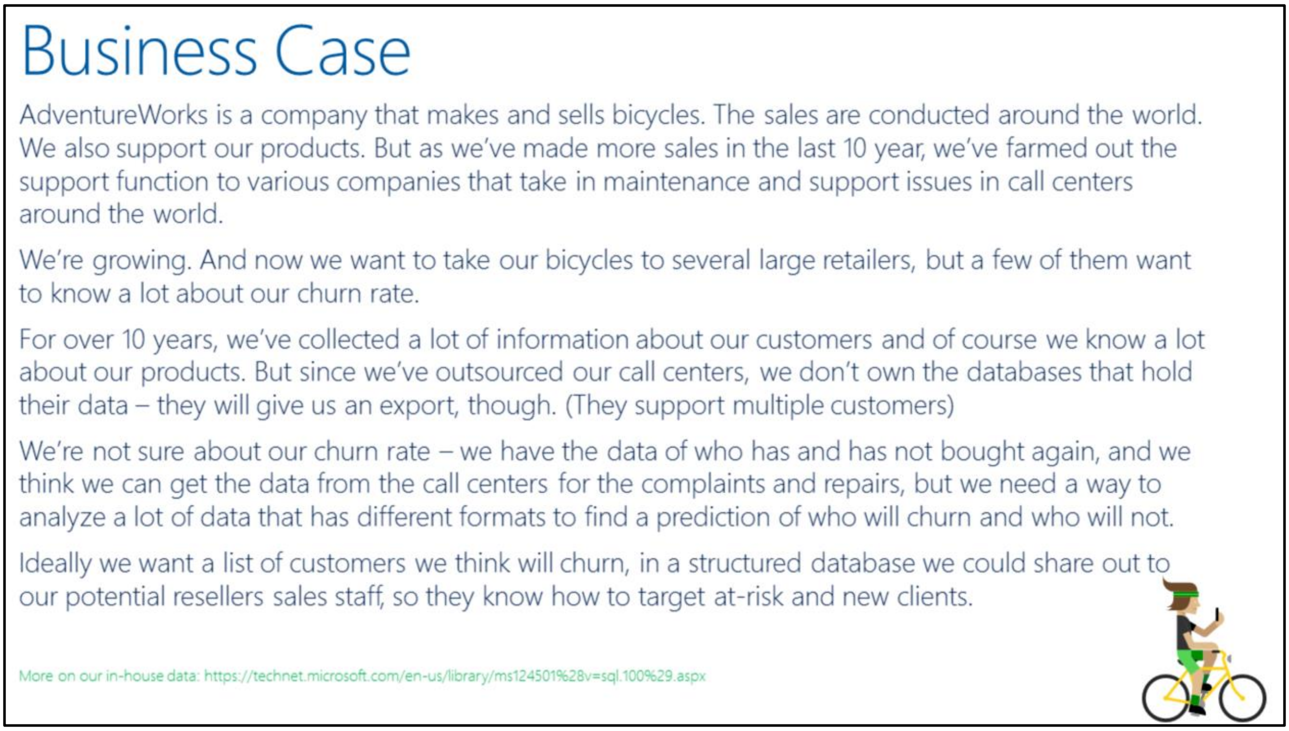
1. Main page: http://cortanaanalytics.com
2. Before starting this module, you should be able to:
   1. Use Azure Storage effectively based on region
   2. Understand parallelizing data loads
   3. Secure data access with tokens and other methods
   4. Use the appropriate data storage type for a given requirement



Learning objectives

- Understand the process for using Azure Data Factory
- Use Azure Data Factory to ingest data
- Use Azure Data Factory to leave data on premise
- Use Azure Data Factory to call functions to clean and shape data
- Use Azure Data Factory to compute analytics
- Use Azure Data Factory to move data to other data stores

1. At the end of this Module, you will 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 premise
   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



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 year, we've farmed out the support function to various companies that take in maintenance and support issues in call centers around the world.

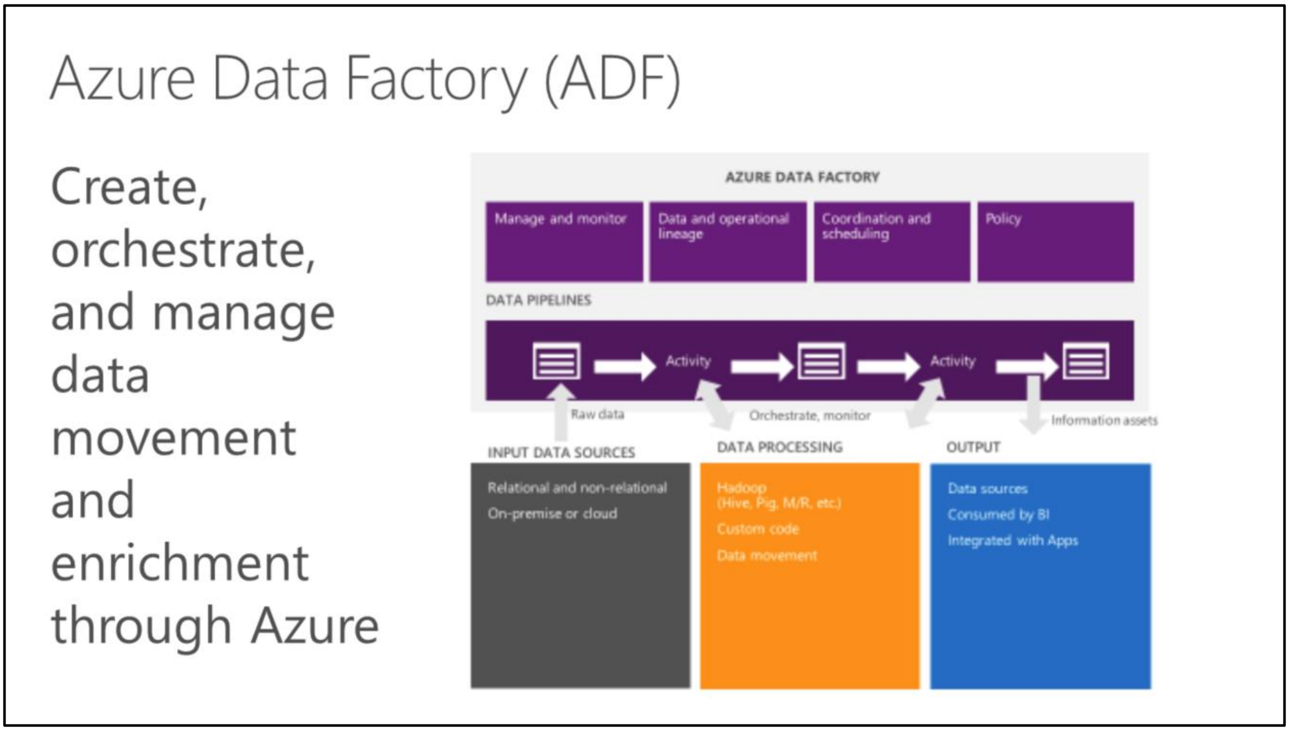
We're growing. And now we want to take our bicycles to several large retailers, but a few of them want to know a lot about our churn rate.

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.

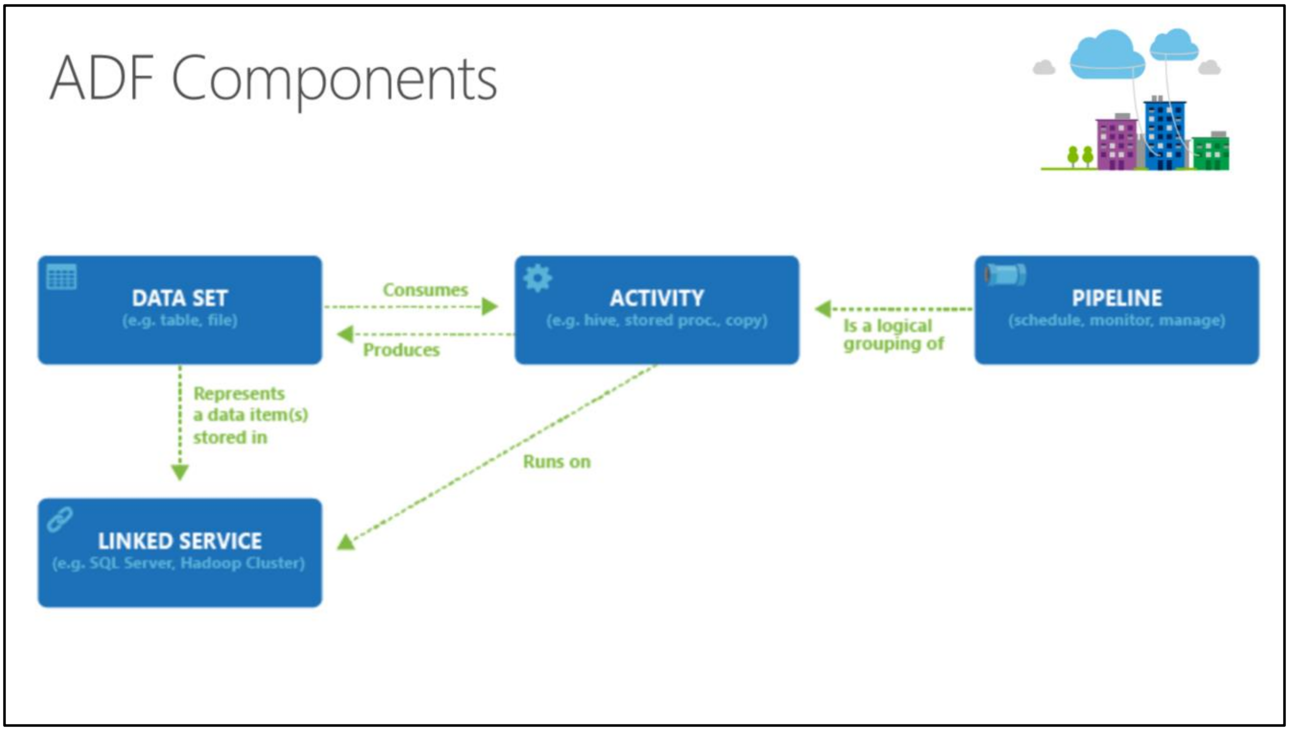
More on our in-house data: https://technet.microsoft.com/en-us/library/ms124501%28v=sql.100%29.aspx



Azure Data Factory (ADF)

Create, orchestrate, and manage data movement and enrichment through Azure

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/
3. Learning Path: https://azure.microsoft.com/en-us/documentation/articles/data-factory-introduction/
4. Developer Reference: https://msdn.microsoft.com/en-us/library/azure/dn834987.aspx;
5. Azure Data Factory Videos: https://azure.microsoft.com/en-us/documentation/videos/index/?services=data-factory
6. Collection of learning resources: https://blogs.technet.microsoft.com/dataplatforminsider/2014/10/30/the-ins-and-outs-of-azure-data-factory-orchestration-and-management-of-diverse-data/



ADF Components

DATA SET (e.g. table, file)

Consumes

Produces

ACTIVITY (e.g. hive, stored proc., copy)

Is a logical grouping of

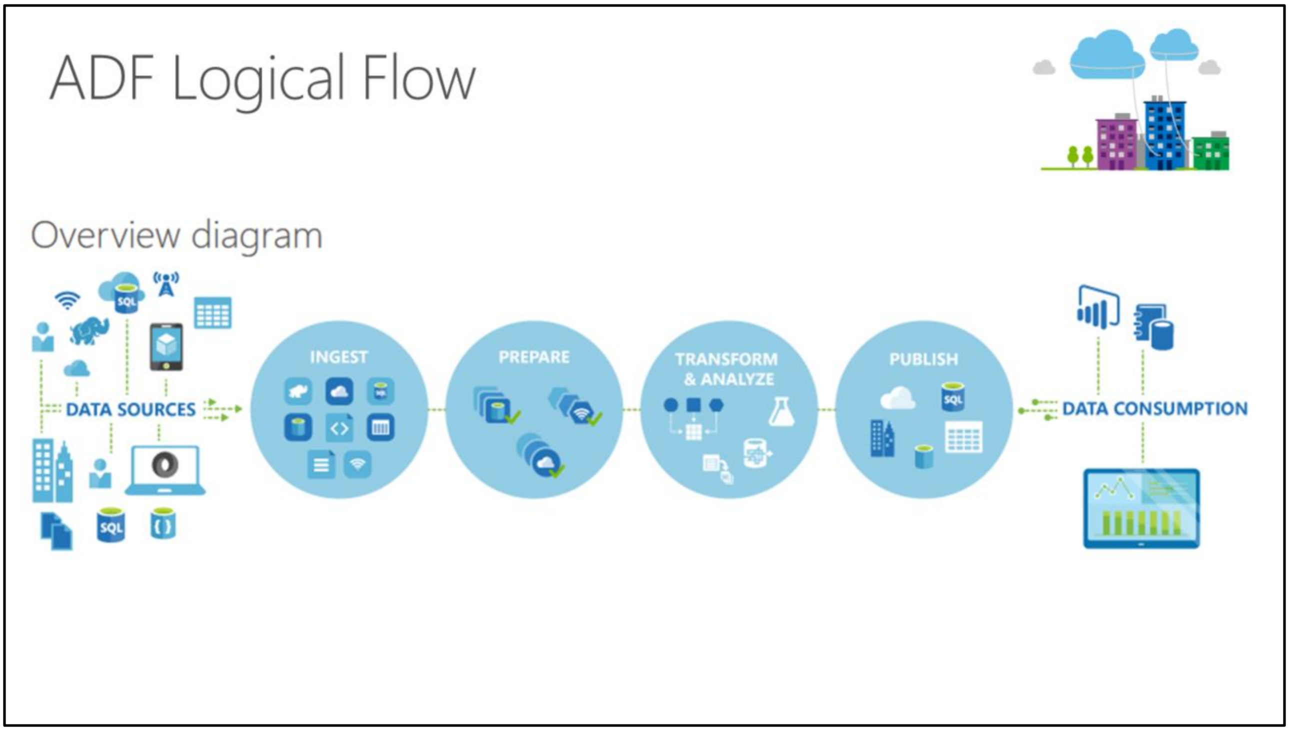
PIPELINE (schedule, monitor, manage)

Represents a data item(s) stored in

Runs on

LINKED SERVICE (e.g. SQL Server, Hadoop Cluster)

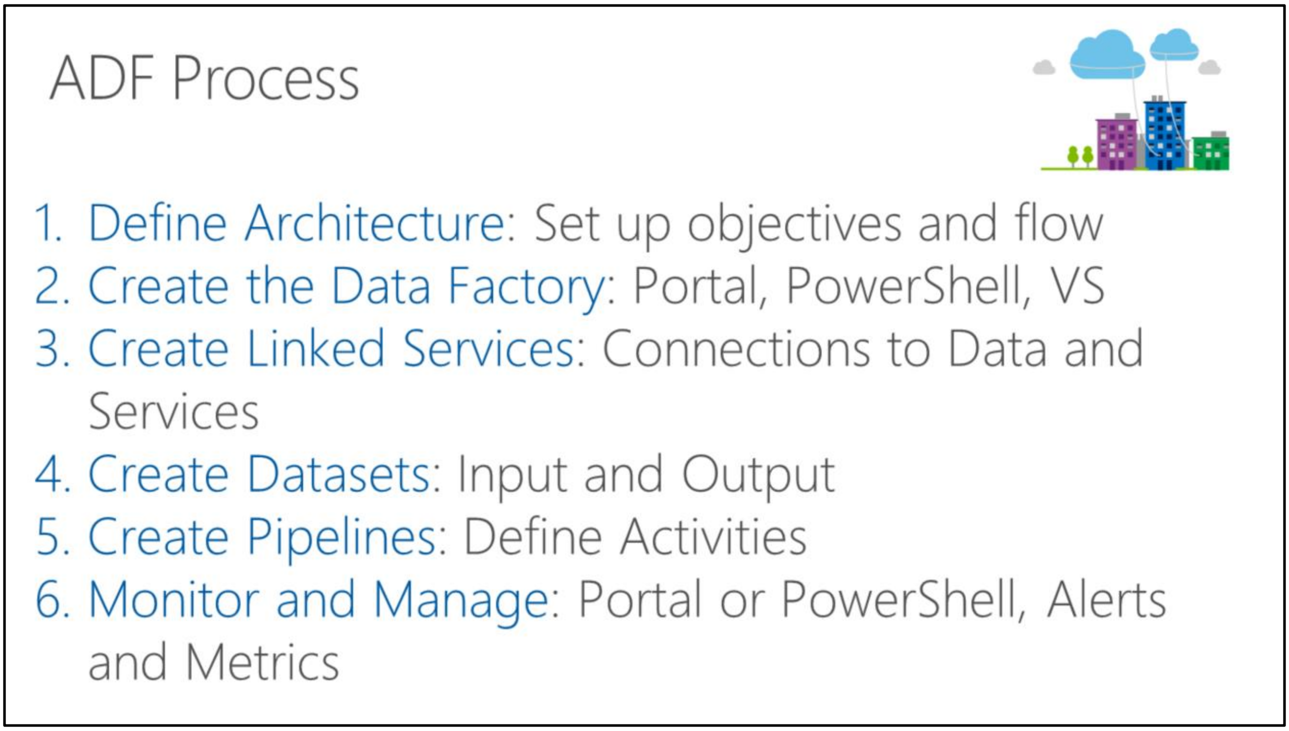
1. Pricing: https://azure.microsoft.com/en-us/pricing/details/data-factory/



ADF Logical Flow

Overview diagram

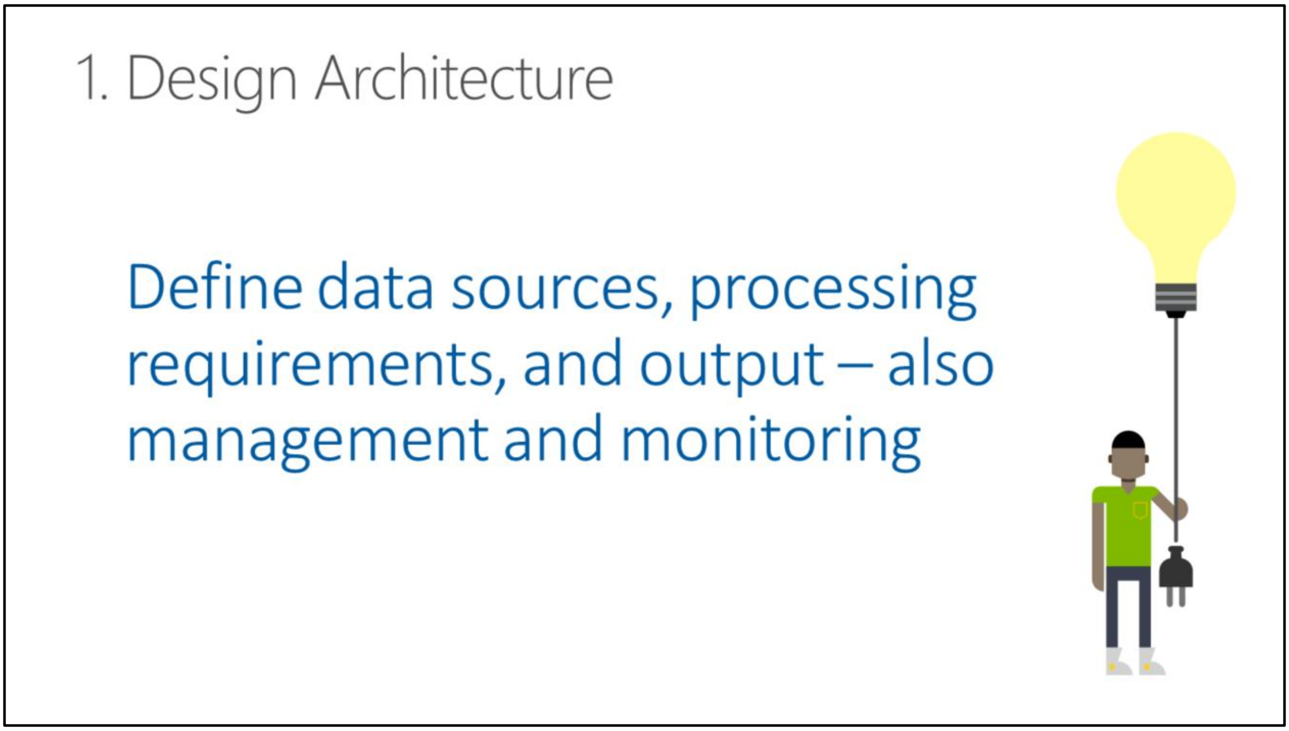
1. Learning Path: https://azure.microsoft.com/en-us/documentation/articles/data-factory-introduction/
2. Quick Example: http://azure.microsoft.com/blog/2015/04/24/azure-data-factory-update-simplified-sample-deployment/



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 Pipelines: Define Activities
6. Monitor and Manage: Portal or PowerShell, Alerts and Metrics

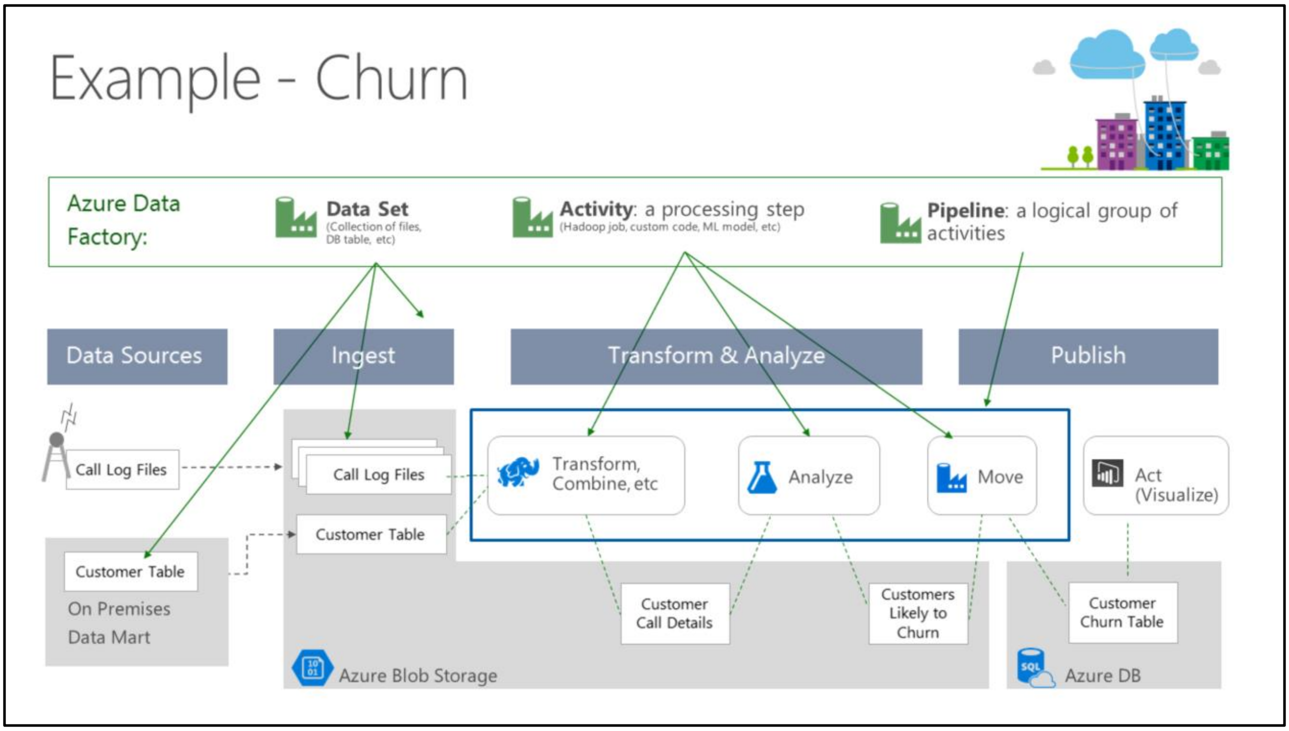
1. Full Tutorial: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline/



1. Design Architecture

Define data sources, processing requirements, and output – also management and monitoring

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



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

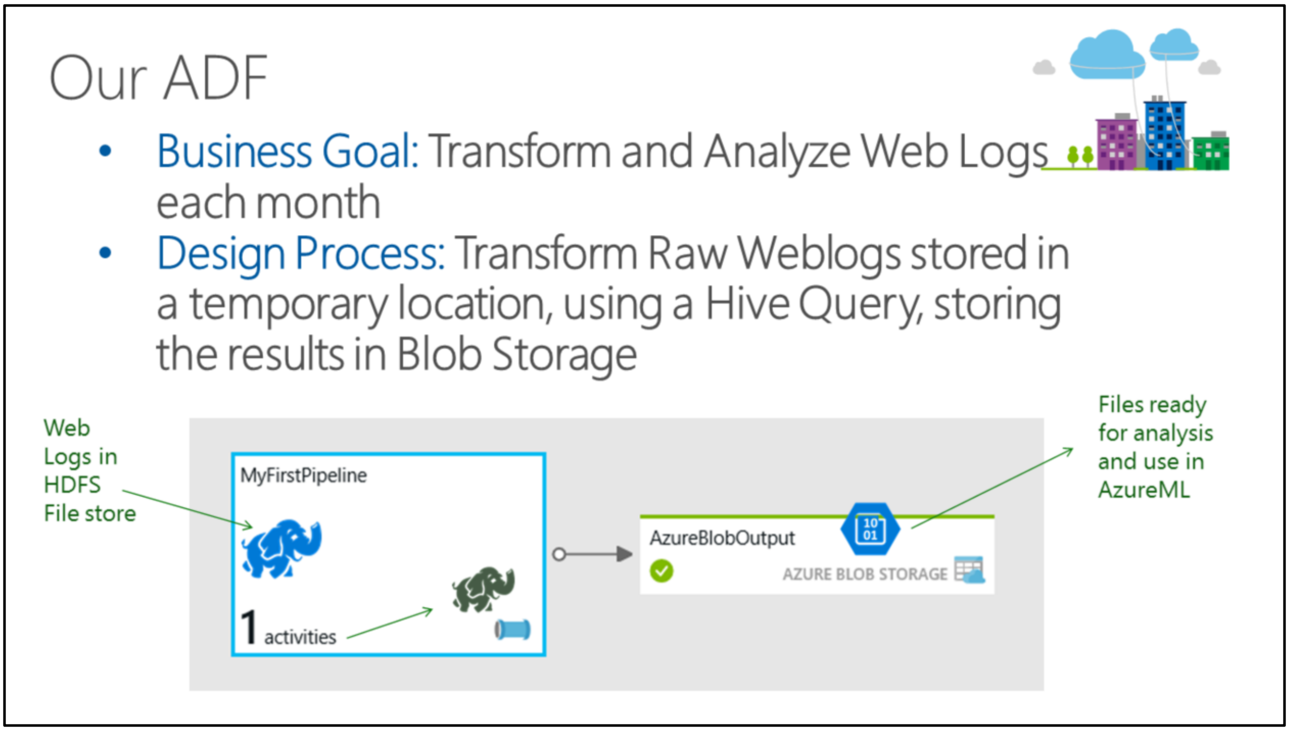
Customers Likely to Churn

Customer Churn Table

Azure Blob Storage

Azure DB

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



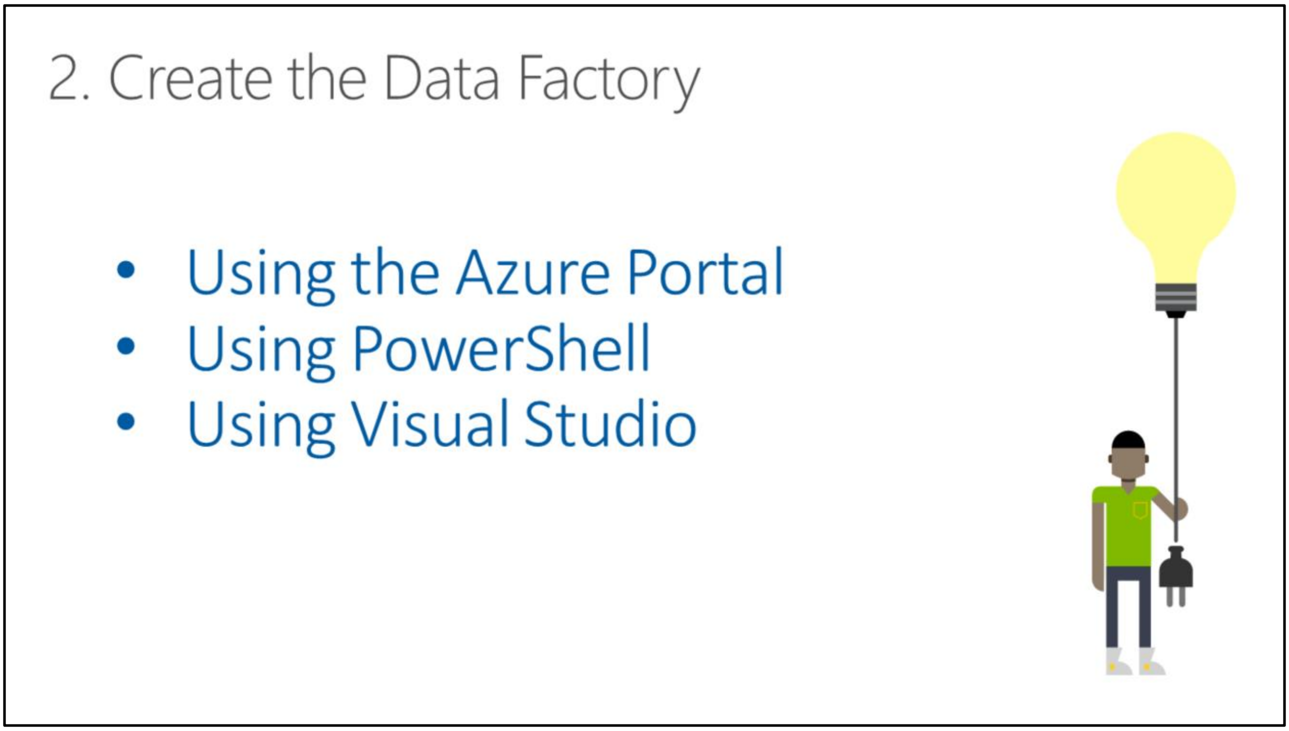
Our ADF

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

Web Logs in HDFS File store

Files ready for analysis and use in AzureML

1. Walkthrough of this example: https://azure.microsoft.com/en-us/documentation/articles/data-factory-samples/

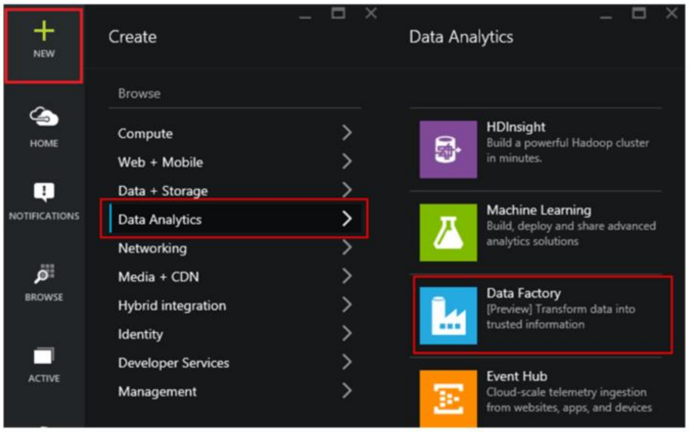


2. Create the Data Factory

- Using the Azure Portal
- Using PowerShell
- Using Visual Studio

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

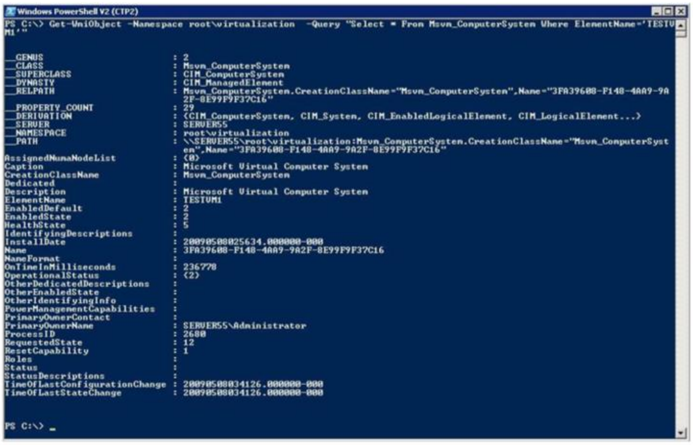
Using the Portal



- Non-MS Clients
- Exploration
- Teaching/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/

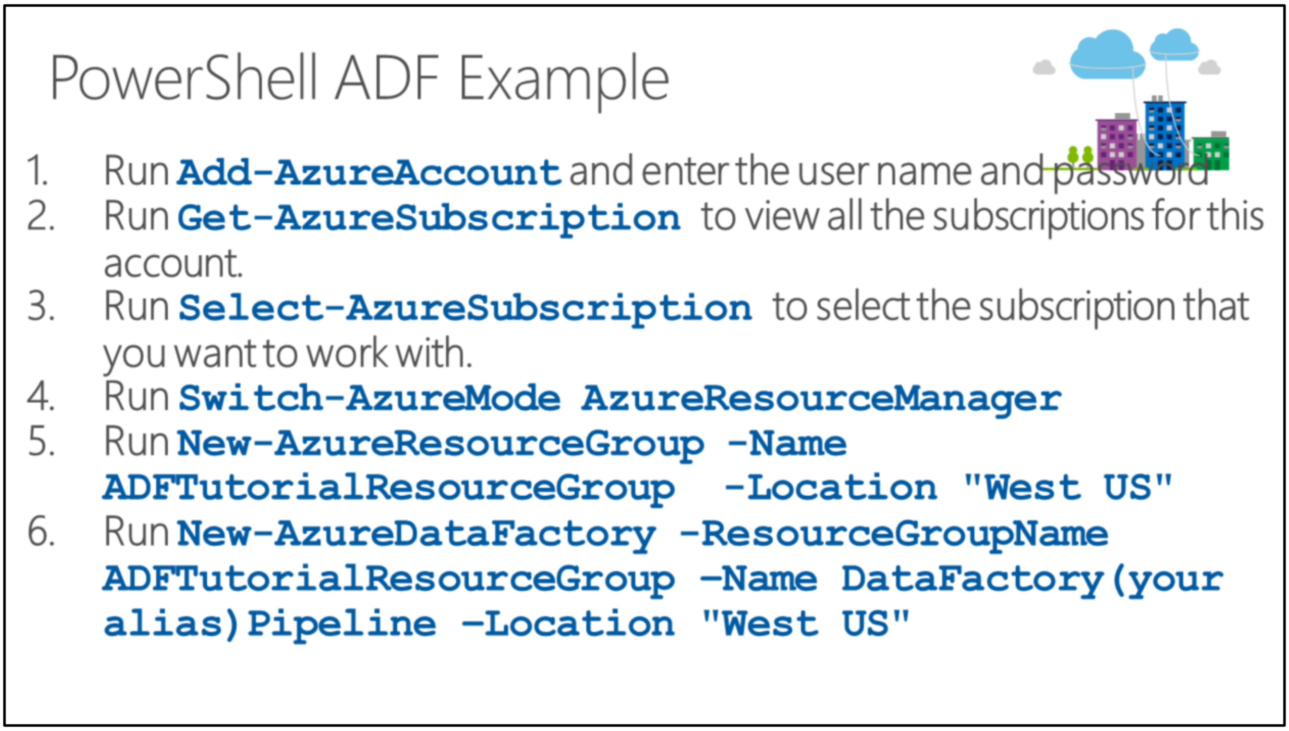
Using PowerShell



```
PS C:\> Get-WmiObject -Namespace root\virtualization  -Query "Select * From Msvm_ComputerSystem Where ElementName='TESTV
M1'"


__GENUS                       : 2
__CLASS                       : Msvm_ComputerSystem
__SUPERCLASS                  : CIM_ComputerSystem
__DYNASTY                     : CIM_ManagedElement
__RELPATH                     : Msvm_ComputerSystem.CreationClassName="Msvm_ComputerSystem",Name="3FA39608-F148-4AA9-9A
                                2F-8E99F9F37C16"
__PROPERTY_COUNT              : 29
__DERIVATION                  : {CIM_ComputerSystem, CIM_System, CIM_EnabledLogicalElement, CIM_LogicalElement...}
__SERVER                      : SERVER55
__NAMESPACE                   : root\virtualization
__PATH                        : \\SERVER55\root\virtualization:Msvm_ComputerSystem.CreationClassName="Msvm_ComputerSyst
                                em",Name="3FA39608-F148-4AA9-9A2F-8E99F9F37C16"
AssignedNumaNodeList          : {0}
Caption                       : Microsoft Virtual Computer System
CreationClassName             : Msvm_ComputerSystem
Dedicated                     :
Description                   : Microsoft Virtual Computer System
ElementName                   : TESTVM1
EnabledDefault                : 2
EnabledState                  : 2
HealthState                   : 5
IdentifyingDescriptions       :
InstallDate                   : 20090508025634.000000-000
Name                          : 3FA39608-F148-4AA9-9A2F-8E99F9F37C16
NameFormat                    :
OnTimeInMilliseconds          : 236778
OperationalStatus             : {2}
OtherDedicatedDescriptions    :
OtherEnabledState             :
OtherIdentifyingInfo          :
PowerManagementCapabilities   :
PrimaryOwnerContact           :
PrimaryOwnerName              : SERVER55\Administrator
ProcessID                     : 2680
RequestedState                : 12
ResetCapability               : 1
Roles                         :
Status                        :
StatusDescriptions            :
TimeOfLastConfigurationChange : 20090508034126.000000-000
TimeOfLastStateChange         : 20090508034126.000000-000


PS C:\> _
```

- MS Clients
- Automation
- Quick setup & 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/

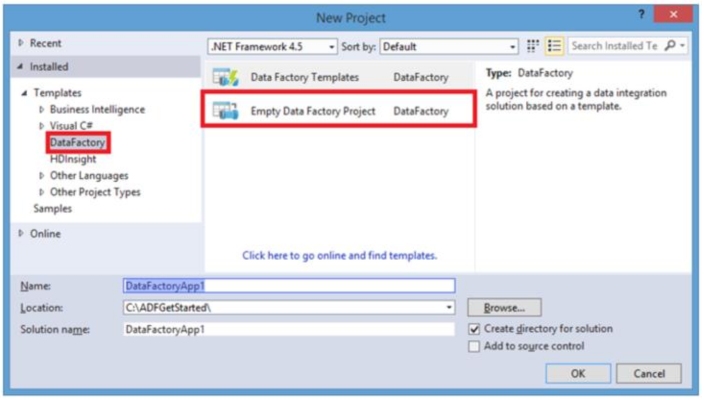


PowerShell ADF Example

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



Type: DataFactory

A project for creating a data integration solution based on a template.

Name: DataFactoryApp1

Location: C:\ADFGetStarted\

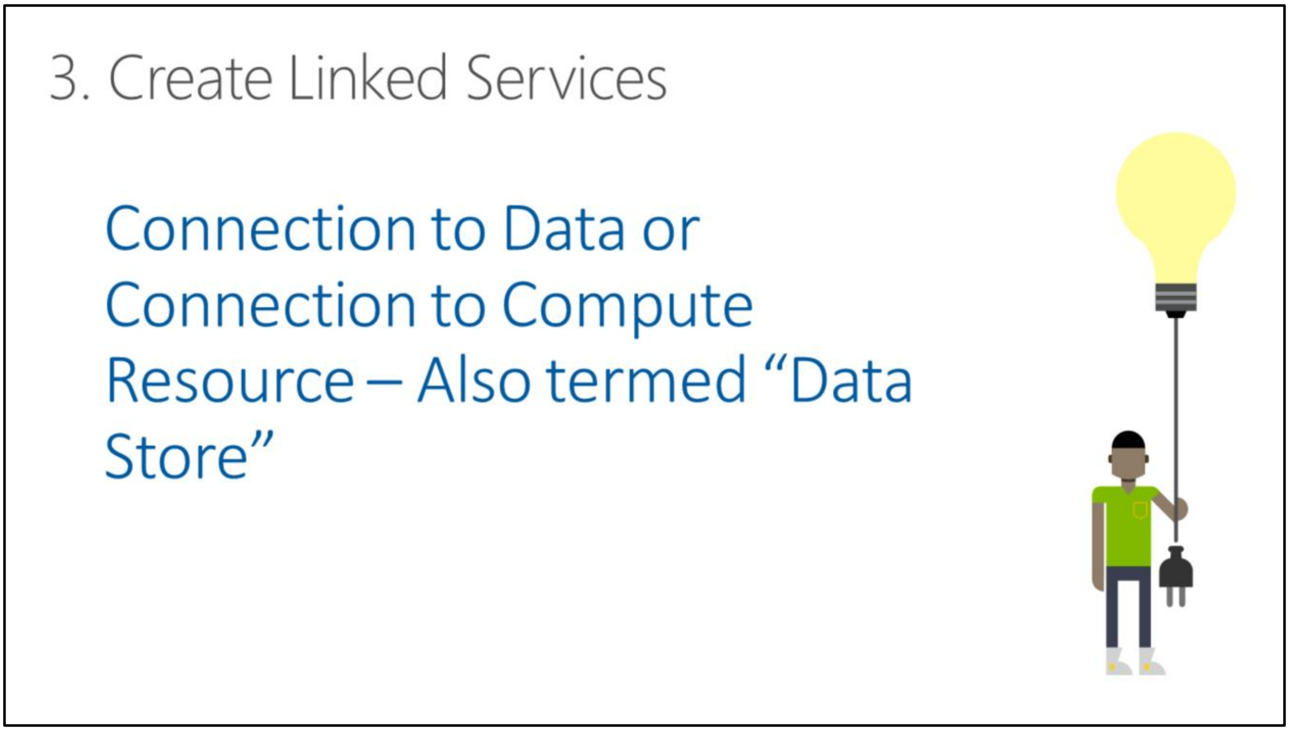
Solution name: DataFactoryApp1

☑ Create directory for solution

□ Add to source control

- Mature dev environments
- Integrated into larger dev process

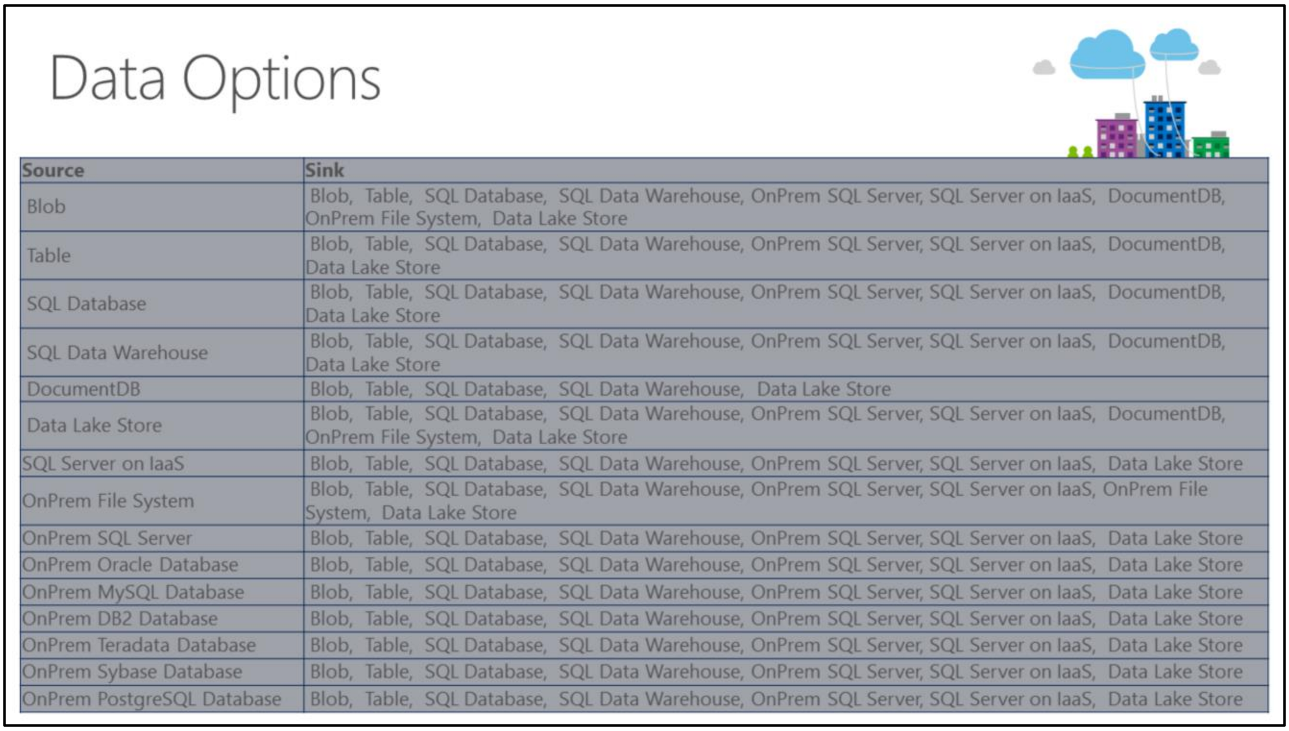
1. Overview: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline/
2. Using the Portal: https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline-using-editor/



3. Create Linked Services

Connection to Data or Connection to Compute Resource – Also termed "Data Store"

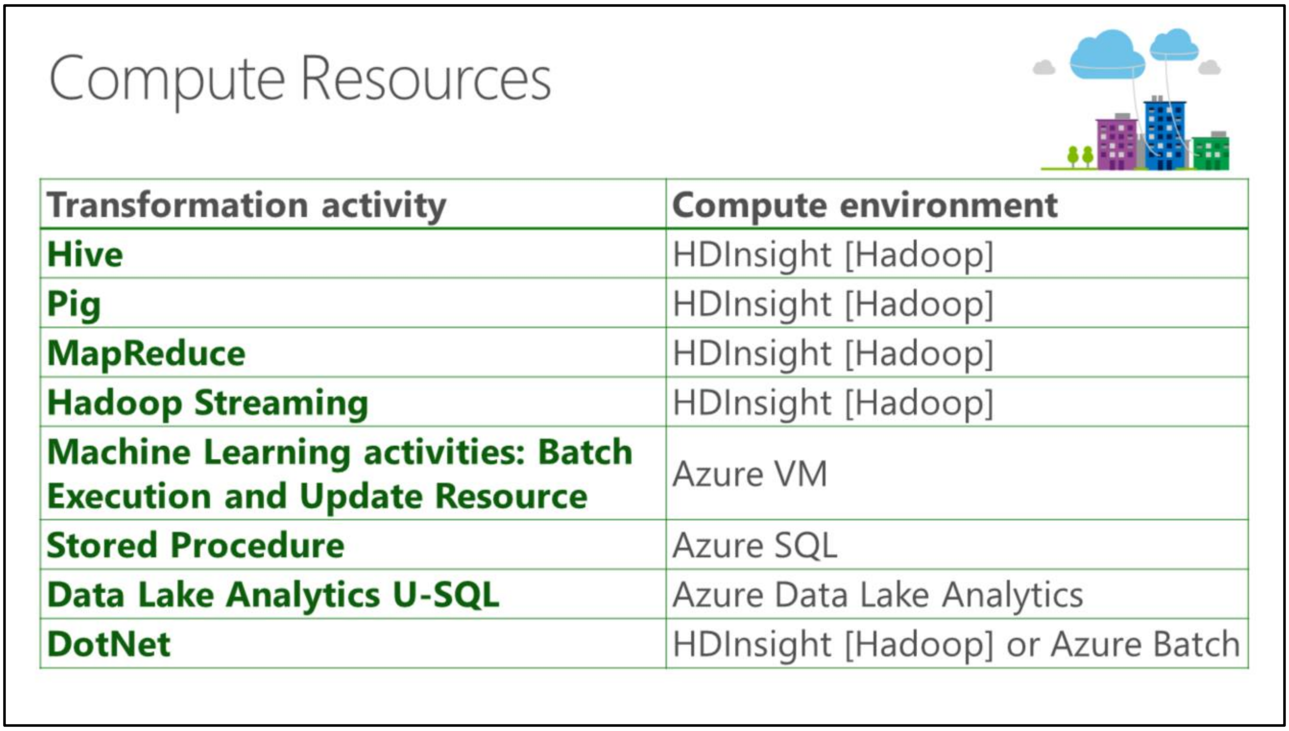
1. Data Linking: https://azure.microsoft.com/en-us/documentation/articles/data-factory-data-movement-activities/
2. Compute Linking: https://azure.microsoft.com/en-us/documentation/articles/data-factory-compute-linked-services/



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

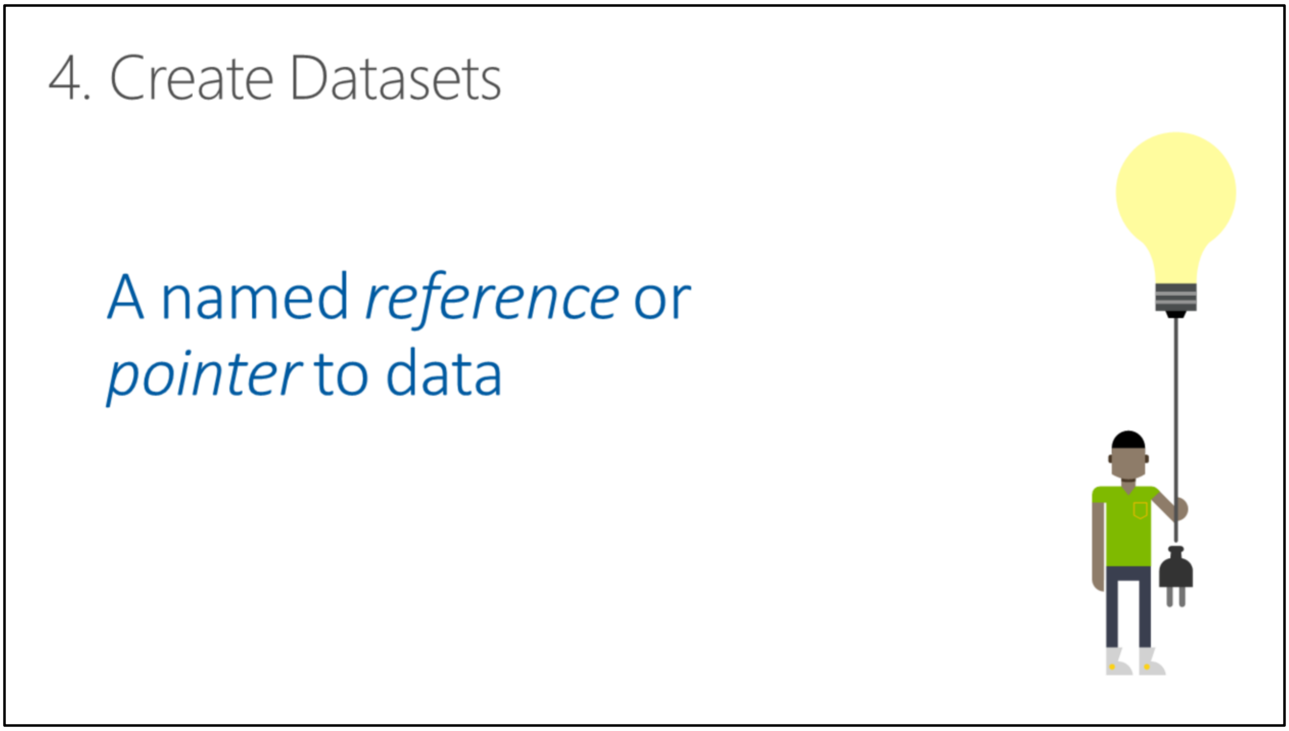
1. Data Movement requirements: https://azure.microsoft.com/en-us/documentation/articles/data-factory-data-movement-activities/
2. From on-premises, requires Data Management Gateway: https://azure.microsoft.com/en-us/documentation/articles/data-factory-move-data-between-onprem-and-cloud/



Compute Resources

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

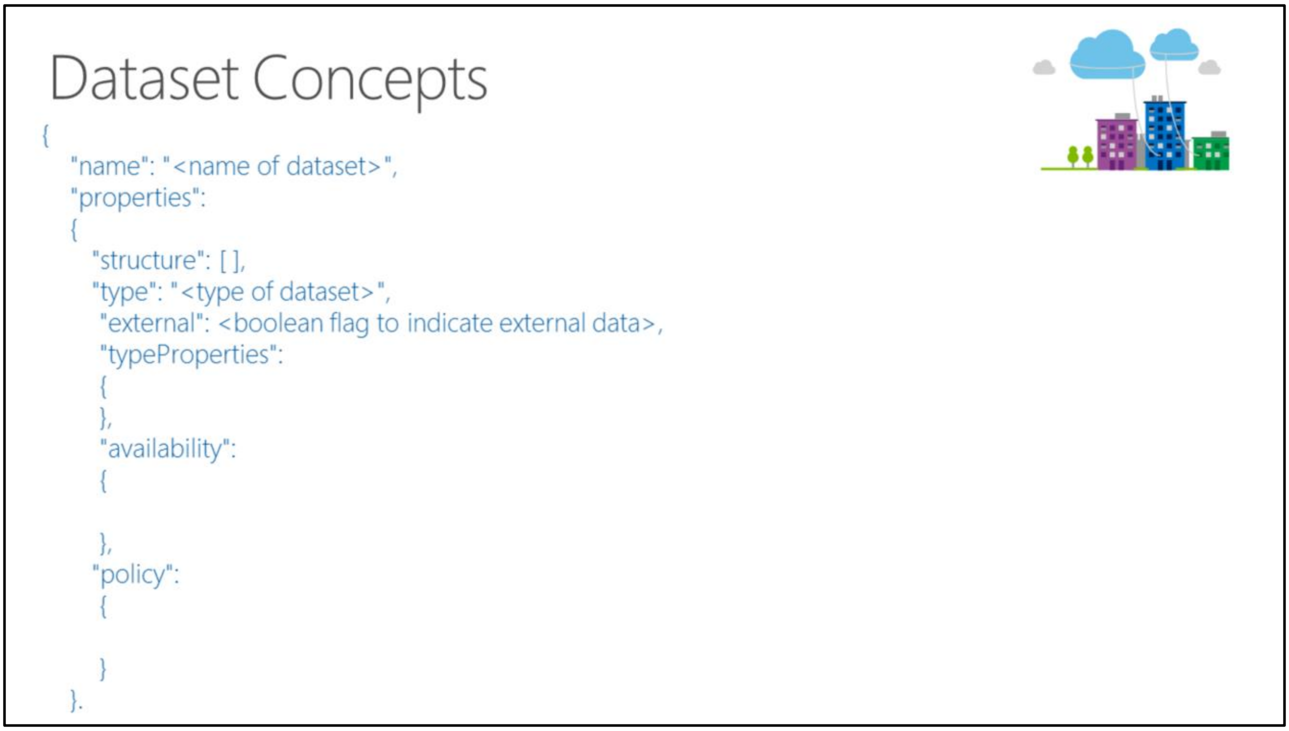
1. Main Document Sites: https://azure.microsoft.com/en-us/documentation/articles/data-factory-data-transformation-activities/



4. Create Datasets

A named *reference* or *pointer* to data

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



Dataset Concepts

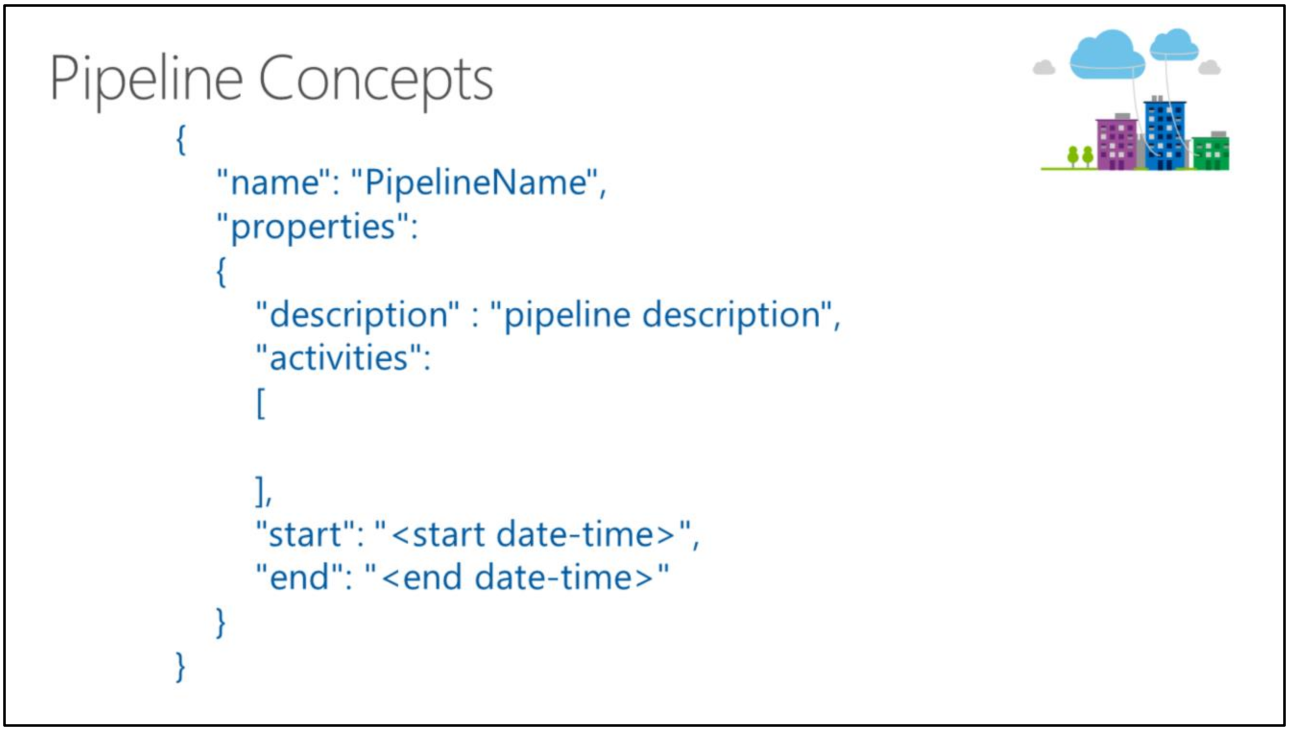
```
{
   "name": "<name of dataset>",
   "properties":
   {
      "structure": [ ],
      "type": "<type of dataset>",
       "external": <boolean flag to indicate external data>,
       "typeProperties":
       {
       },
       "availability":
       {

       },
      "policy":
       {

       }
   }.
```

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

5. Create Pipelines

Logical Grouping of Activities

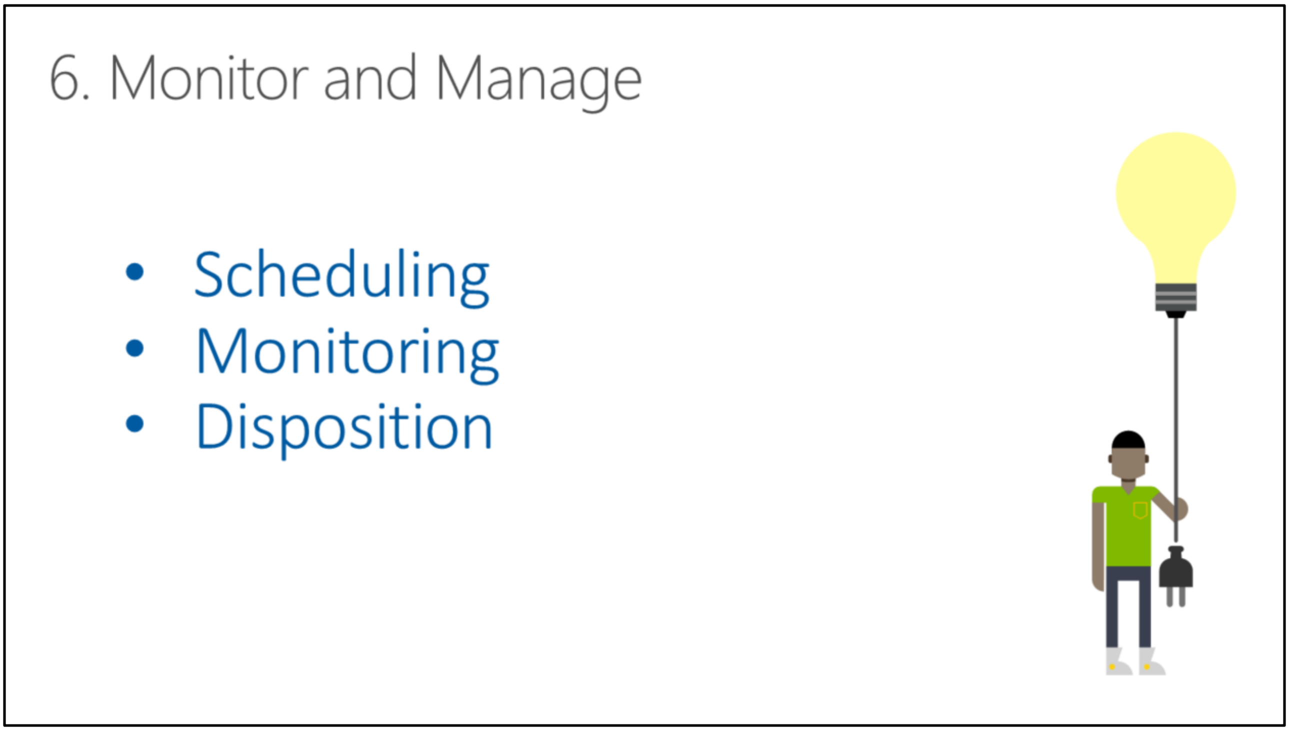
1. Main Pipeline Documentation: https://azure.microsoft.com/en-us/documentation/articles/data-factory-create-pipelines/



Pipeline Concepts

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

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

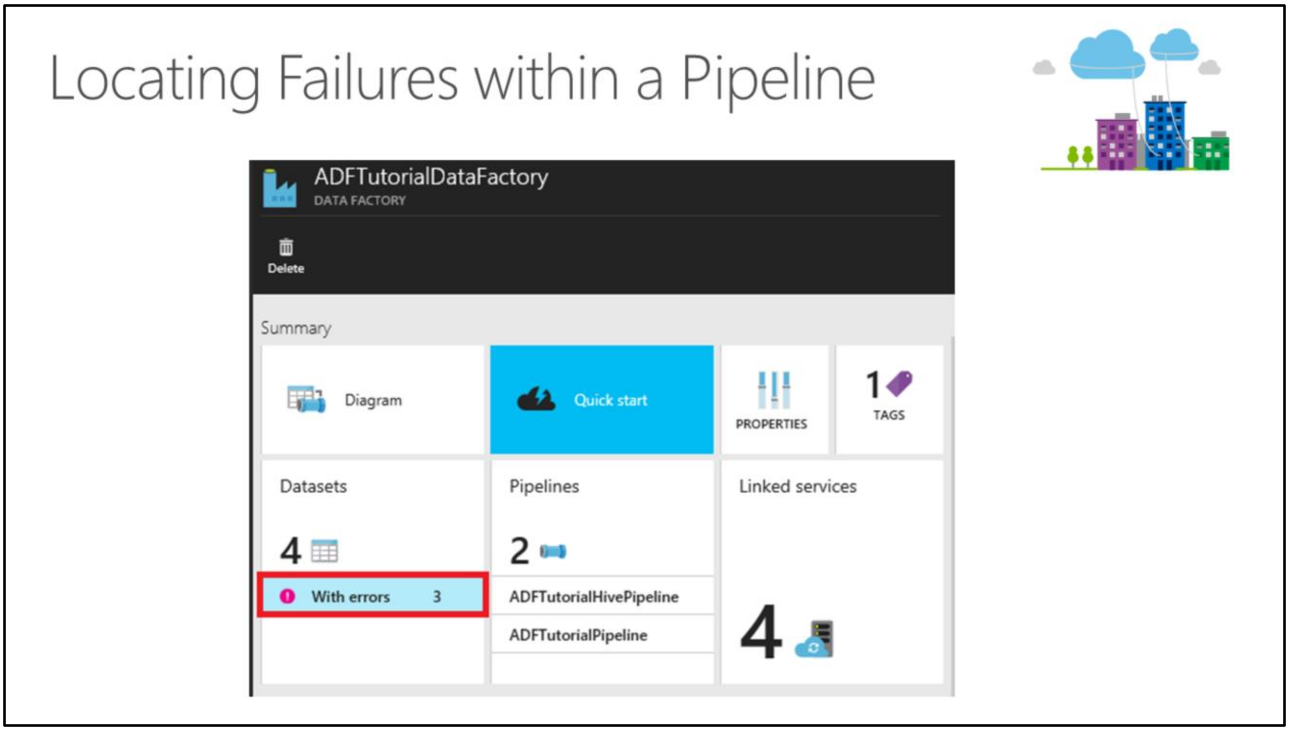
1. https://azure.microsoft.com/en-us/documentation/articles/data-factory-build-your-first-pipeline-using-editor/



6. Monitor and Manage

- Scheduling
- Monitoring
- Disposition

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



Locating Failures within a Pipeline

ADFTutorialDataFactory
DATA FACTORY

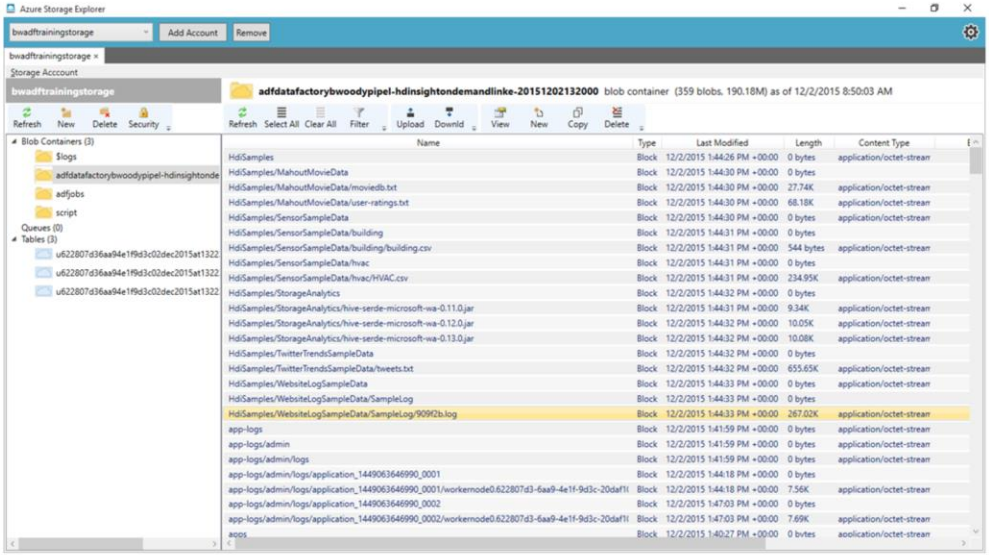
Datasets
4
With errors 3

Pipelines
2
ADFTutorialHivePipeline
ADFTutorialPipeline

Linked services
4

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

Locating Failures within a Pipeline



adfdatafactorybwoodypipel-hdinsightondemandlinke-20151202132000 blob container (359 blobs, 190.18M) as of 12/2/2015 8:50:03 AM

Name	Type	Last Modified	Length	Content Type
HdiSamples	Block	12/2/2015 1:44:26 PM +00:00	0 bytes	application/octet-stream
HdiSamples/MahoutMovieData	Block	12/2/2015 1:44:30 PM +00:00	0 bytes	
HdiSamples/MahoutMovieData/moviedb.txt	Block	12/2/2015 1:44:30 PM +00:00	27.74K	application/octet-stream
HdiSamples/MahoutMovieData/user-ratings.txt	Block	12/2/2015 1:44:30 PM +00:00	68.18K	application/octet-stream
HdiSamples/SensorSampleData	Block	12/2/2015 1:44:30 PM +00:00	0 bytes	application/octet-stream
HdiSamples/SensorSampleData/building	Block	12/2/2015 1:44:31 PM +00:00	0 bytes	
HdiSamples/SensorSampleData/building/building.csv	Block	12/2/2015 1:44:31 PM +00:00	544 bytes	application/octet-stream
HdiSamples/SensorSampleData/hvac	Block	12/2/2015 1:44:31 PM +00:00	0 bytes	
HdiSamples/SensorSampleData/hvac/HVAC.csv	Block	12/2/2015 1:44:31 PM +00:00	234.95K	application/octet-stream
HdiSamples/StorageAnalytics	Block	12/2/2015 1:44:32 PM +00:00	0 bytes	
HdiSamples/StorageAnalytics/hive-serde-microsoft-wa-0.11.0.jar	Block	12/2/2015 1:44:31 PM +00:00	9.34K	application/octet-stream
HdiSamples/StorageAnalytics/hive-serde-microsoft-wa-0.12.0.jar	Block	12/2/2015 1:44:32 PM +00:00	10.05K	application/octet-stream
HdiSamples/StorageAnalytics/hive-serde-microsoft-wa-0.13.0.jar	Block	12/2/2015 1:44:32 PM +00:00	10.08K	application/octet-stream
HdiSamples/TwitterTrendsSampleData	Block	12/2/2015 1:44:32 PM +00:00	0 bytes	
HdiSamples/TwitterTrendsSampleData/tweets.txt	Block	12/2/2015 1:44:32 PM +00:00	655.65K	application/octet-stream
HdiSamples/WebsiteLogSampleData	Block	12/2/2015 1:44:33 PM +00:00	0 bytes	application/octet-stream
HdiSamples/WebsiteLogSampleData/SampleLog	Block	12/2/2015 1:44:33 PM +00:00	0 bytes	
HdiSamples/WebsiteLogSampleData/SampleLog/909f2b.log	Block	12/2/2015 1:44:33 PM +00:00	267.02K	application/octet-stream
app-logs	Block	12/2/2015 1:41:59 PM +00:00	0 bytes	application/octet-stream
app-logs/admin	Block	12/2/2015 1:41:59 PM +00:00	0 bytes	application/octet-stream
app-logs/admin/logs	Block	12/2/2015 1:41:59 PM +00:00	0 bytes	application/octet-stream
app-logs/admin/logs/application_1449063646990_0001	Block	12/2/2015 1:44:18 PM +00:00	0 bytes	
app-logs/admin/logs/application_1449063646990_0001/workernode0.622807d3-6aa9-4e1f-9d3c-20daf1(	Block	12/2/2015 1:44:18 PM +00:00	7.56K	application/octet-stream
app-logs/admin/logs/application_1449063646990_0002	Block	12/2/2015 1:47:03 PM +00:00	0 bytes	
app-logs/admin/logs/application_1449063646990_0002/workernode0.622807d3-6aa9-4e1f-9d3c-20daf1(	Block	12/2/2015 1:47:03 PM +00:00	7.69K	application/octet-stream
apps	Block	12/2/2015 1:40:27 PM +00:00	0 bytes	application/octet-stream

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

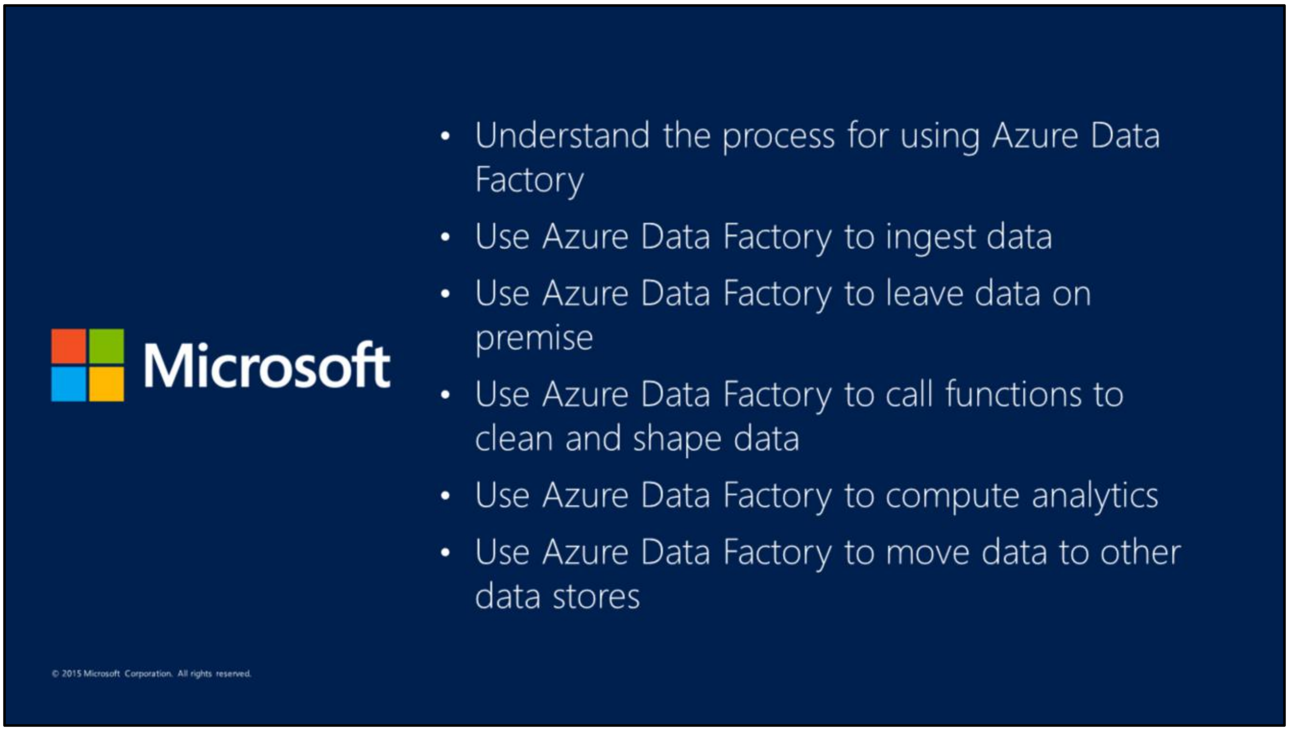
Lab – Start to Finish with Azure Data Factory

Microsoft Azure

Microsoft

PARTNER PRACTICE ENABLEMENT BOOTCAMP

1. The Lab is included in the “Resources” section of your Classroom Assets



Microsoft

- Understand the process for using Azure Data Factory
- Use Azure Data Factory to ingest data
- Use Azure Data Factory to leave data on premise
- Use Azure Data Factory to call functions to clean and shape data
- Use Azure Data Factory to compute analytics
- Use Azure Data Factory to move data to other data stores

1. Use this for Q/A time