

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.

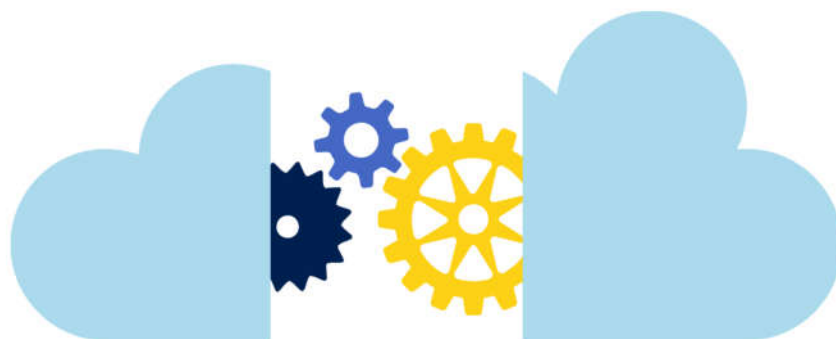
## Section 2 Learning Objectives

1. Understand how to source and document data locations
2. Understand Azure Storage Options
3. Use various methods to ingest data into Azure Storage
4. Examine data stored in Azure Storage
5. Use various tools to explore data

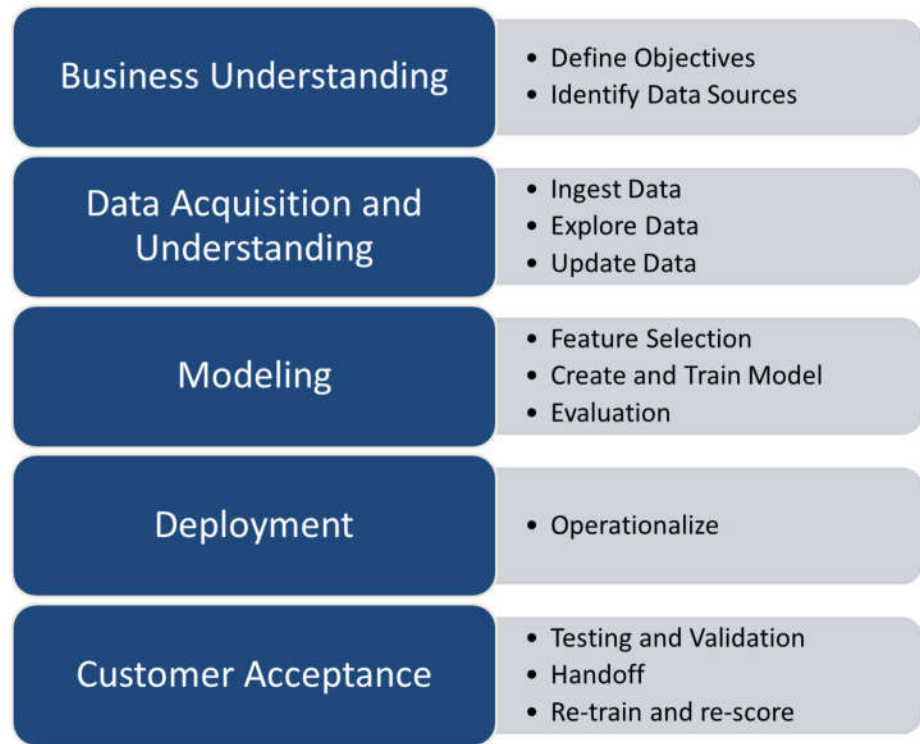


1. At the end of this Module, you will:
  1. Understand how to source and document data locations
  2. Understand feature selection
  3. Understand Azure Storage Options
  4. Use various methods to ingest data into Azure Storage
  5. Examine data stored in Azure Storage
  6. Use various tools to explore data

## The Data Science Process and Platform



## The Team Data Science Process



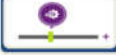










1. This process largely follows the CRISP-DM model:  
<http://www.sv-europe.com/crisp-dm-methodology/>
2. It also references the Cortana Intelligence process:  
<https://azure.microsoft.com/en-us/documentation/articles/data-science-process-overview/>
3. A complete process diagram is here:  
<https://azure.microsoft.com/en-us/documentation/learning-paths/cortana-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 to manage and monitor
7. A great resource for a team involved in data science to get started: <https://github.com/Azure/Microsoft-TDSP>

# The Cortana Intelligence Platform

	Cortana, Cognitive Services, Bot Framework
	Power BI
	Stream Analytics
	HDInsight
	Azure Machine Learning (MRS)
	SQL Data Warehouse (SQL DB, Document DB)
	Data Lake
	Event Hubs
	Data Factory
	Data Catalog
	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 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: <https://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>

## Module 1: Sourcing and Vetting Data



6

1. Data Validation: [https://msdn.microsoft.com/en-us/library/aa291820\(v=vs.71\).aspx](https://msdn.microsoft.com/en-us/library/aa291820(v=vs.71).aspx)



## Inspecting data



### Keys to quality source data

- Authority
- Amount
- Representative
- Missing values
- Types
- Ranges

1. In reference to machine learning, but applicable to all data usage: <https://azure.microsoft.com/en-us/documentation/articles/machine-learning-data-science-prepare-data/>
2. Great whitepaper on preprocessing data in R and important questions to ask: <https://info.microsoft.com/rs/157-GQE-382/images/EN-CNTNT-Whitepaper-Data-Prep-Using-R.pdf>

# Azure Data Catalog



1. Register data sources
2. Tag the data descriptions
3. Make it easy to find data in context
4. Use the data – keep it secure

1. Full example: <https://azure.microsoft.com/en-us/documentation/articles/data-catalog-get-started/>



1. Search for one on-line table involving your business scenario
2. Connect to the Azure Data Catalog
3. Add a Data Source as an HTTP site
4. Add metadata to the information
5. Save and view in Portal
6. Search for your data element based on name or tag you added
7. Add more tags
8. Add yourself as an expert
9. Search for your name as an expert

## Module 2: Azure Storage Options



10

1. Data Storage Options (Building Real-World Cloud Apps with Azure): <https://www.asp.net/aspnet/overview/developing-apps-with-windows-azure/building-real-world-cloud-apps-with-windows-azure/data-storage-options>
2. Options:  
[https://blogs.msdn.microsoft.com/uk\\_faculty\\_connection/2017/02/24/big-data-and-azure/](https://blogs.msdn.microsoft.com/uk_faculty_connection/2017/02/24/big-data-and-azure/)

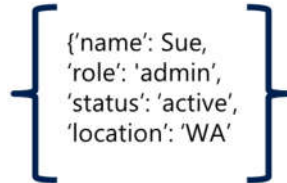
## Storage Scenarios

Unstructured data  
such as media files,  
logs, binary data,  
backups



**Blob**

Metadata (e.g. user info), in  
key-value format, fast and  
easy to query



**Table**

Messaging between  
components of your  
application



**Queue**

Shared file systems  
option – when your  
application is  
already built to use  
a SMB protocol



**File**

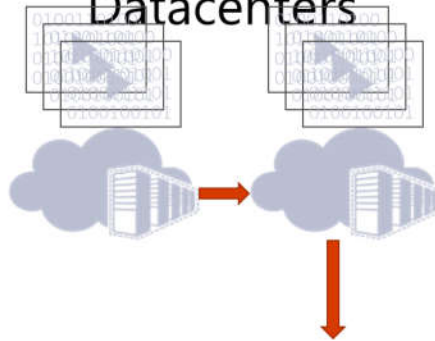
1. <https://channel9.msdn.com/Blogs/Windows-Azure/Azure-Storage-5-Minute-Overview>
2. <https://azure.microsoft.com/en-us/documentation/articles/storage-introduction/>

## Redundancy and Location

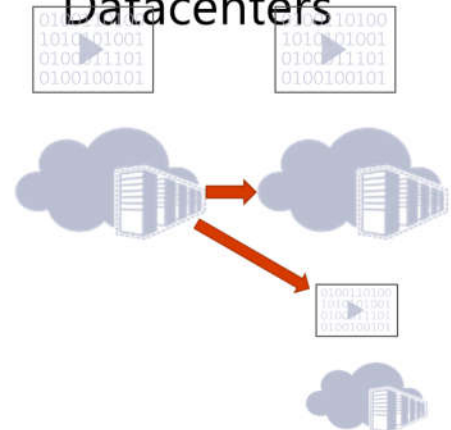
**LRS:** 3 Copies,  
1 Datacenter



**GRS:** 6  
Copies, 2  
Datacenters



**ZRS:** 3 Copies,  
2-3  
Datacenters



1. Locations and Redundancy Overview: <https://azure.microsoft.com/en-us/documentation/articles/storage-introduction/>
2. Affects on Scalability and Performance Targets: <https://azure.microsoft.com/en-us/documentation/articles/storage-scalability-targets/>
3. Pricing Details: <https://azure.microsoft.com/en-us/pricing/details/storage/>

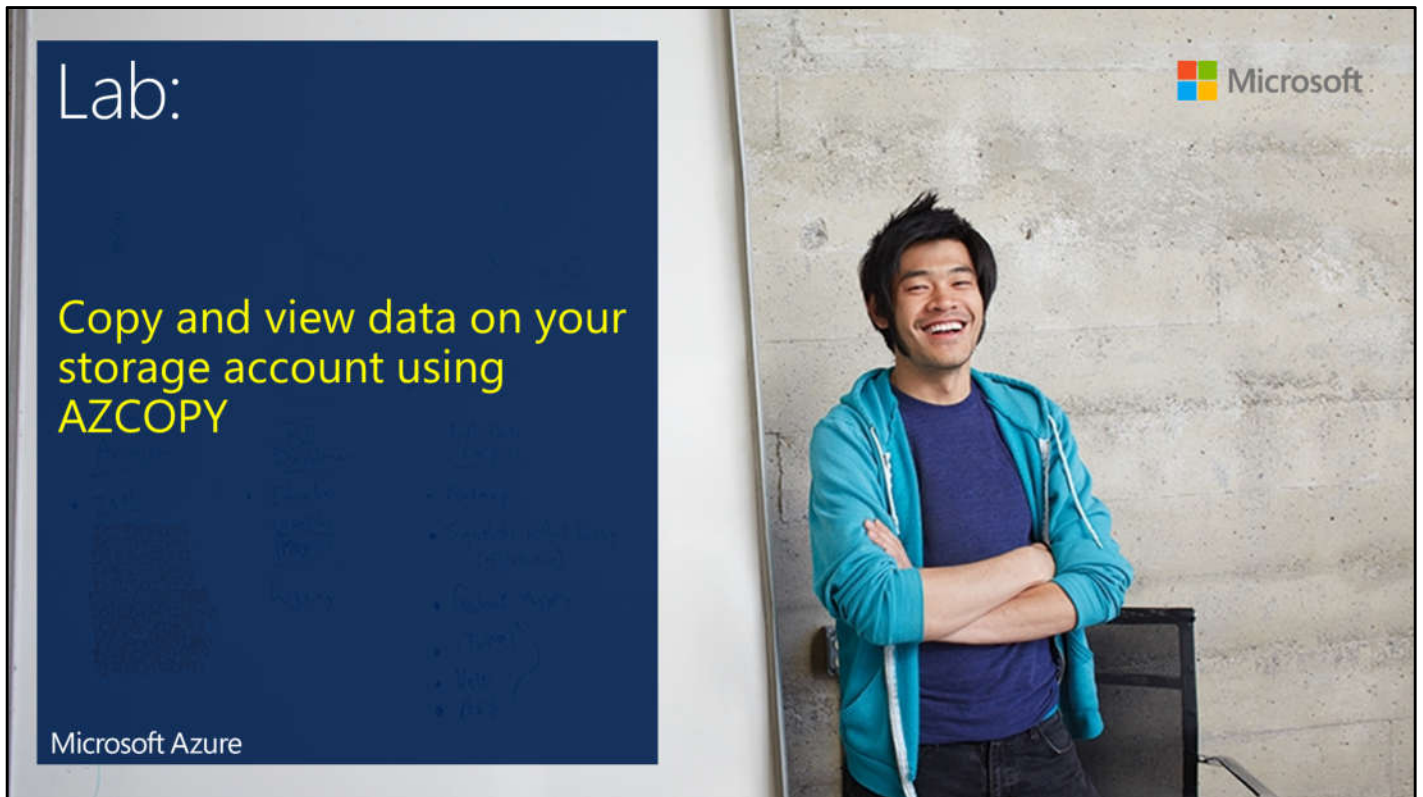
# Creating and Managing Azure Storage

- Azure Portal
- Azure PowerShell
- Azure Command Line Interface (CLI)
- Service Management REST API
- Azure Storage Resource Provider REST API



1. Azure Portal - <https://portal.azure.com/>
2. Azure PowerShell - <https://azure.microsoft.com/en-us/documentation/articles/storage-powershell-guide-full/>
3. AZCOPY - <https://azure.microsoft.com/en-us/documentation/articles/storage-use-azcopy/>
4. Azure CLI - <https://azure.microsoft.com/en-us/documentation/articles/storage-azure-cli/>
5. Service management REST API - <http://msdn.microsoft.com/library/azure/ee460799.aspx>
6. Azure Storage Resource Provider REST API - <https://msdn.microsoft.com/library/azure/mt163683.aspx>





1. Open the Azure Portal, locate your Storage Account (or create one if you have not), and a Container (or create one if you have not). Note the name of the SA and the Container, and your storage key.
2. From your DVSM, or if you installed the Azure PowerShell tools locally, open a command prompt.
  1. If you do not have the AZCOPY command, download and install it here: <http://aka.ms/downloadazcopy>
3. Navigate to this page: <https://azure.microsoft.com/en-us/documentation/articles/storage-use-azcopy/>
4. Locate the section marked "Blob: Upload - Upload single file" and follow the instructions to load one file to your storage account, using your Storage Account and storage keys.
5. Next, locate the section on the web page with instructions marked "Blob: Download - Download single blob". Follow the instructions there to copy your file to a new folder on



your local computer.

## Module 3: Data Ingestion



15

1. Example of a 3<sup>rd</sup> Party Solution: <https://www.veeam.com/fastscp-azure-vm.html>

## Options for data ingestion

- PowerShell
- Azure Data Factory
- Azure Automation
- Azure storage SDKs (.NET, Node.js, python, C++, etc.)
- Microsoft Azure Storage Explorer application (blob only right now)
- AzCopy (blob, file, and table only)
- Import/Export service



1. PowerShell in Azure Storage - <https://azure.microsoft.com/en-us/documentation/articles/storage-powershell-guide-full/>
2. Azure Data Factory data movement - <https://azure.microsoft.com/en-us/documentation/articles/data-factory-data-movement-activities/>
3. Azure Automation - <https://azure.microsoft.com/en-us/documentation/articles/automation-intro/>
4. Azure storage SDKs – for examples see <https://azure.microsoft.com/en-us/documentation/articles/storage-dotnet-how-to-use-blobs/>
5. Azure tools and SDKs in general can be downloaded here - <https://azure.microsoft.com/en-us/downloads/>
6. MS Azure Storage Explorer - <http://storageexplorer.com/>
7. AzCopy - <https://azure.microsoft.com/en-us/documentation/articles/storage-use-azcopy/>
8. Import/Export service - <https://azure.microsoft.com/en-us/documentation/articles/storage-import-export-service/>

## Connect on-prem to <anything>

### VPN Gateway

- Send network traffic from virtual networks to on-prem locations
- Send network traffic between virtual networks within Azure
- Site-to-site vs. Point-to-site
- You can connect multiple on-prem locations to a virtual network (Multi-site)
- ExpressRoute can directly connect your WAN to Azure
- Tool-Specific

1. <https://azure.microsoft.com/en-us/documentation/articles/vpn-gateway-about-vpngateways/>
2. <https://azure.microsoft.com/en-us/documentation/articles/vpn-gateway-vpn-faq/#connecting-to-virtual-networks>
3. <https://azure.microsoft.com/en-us/documentation/articles/expressroute-faqs/>

## Module 4: Data Exploration



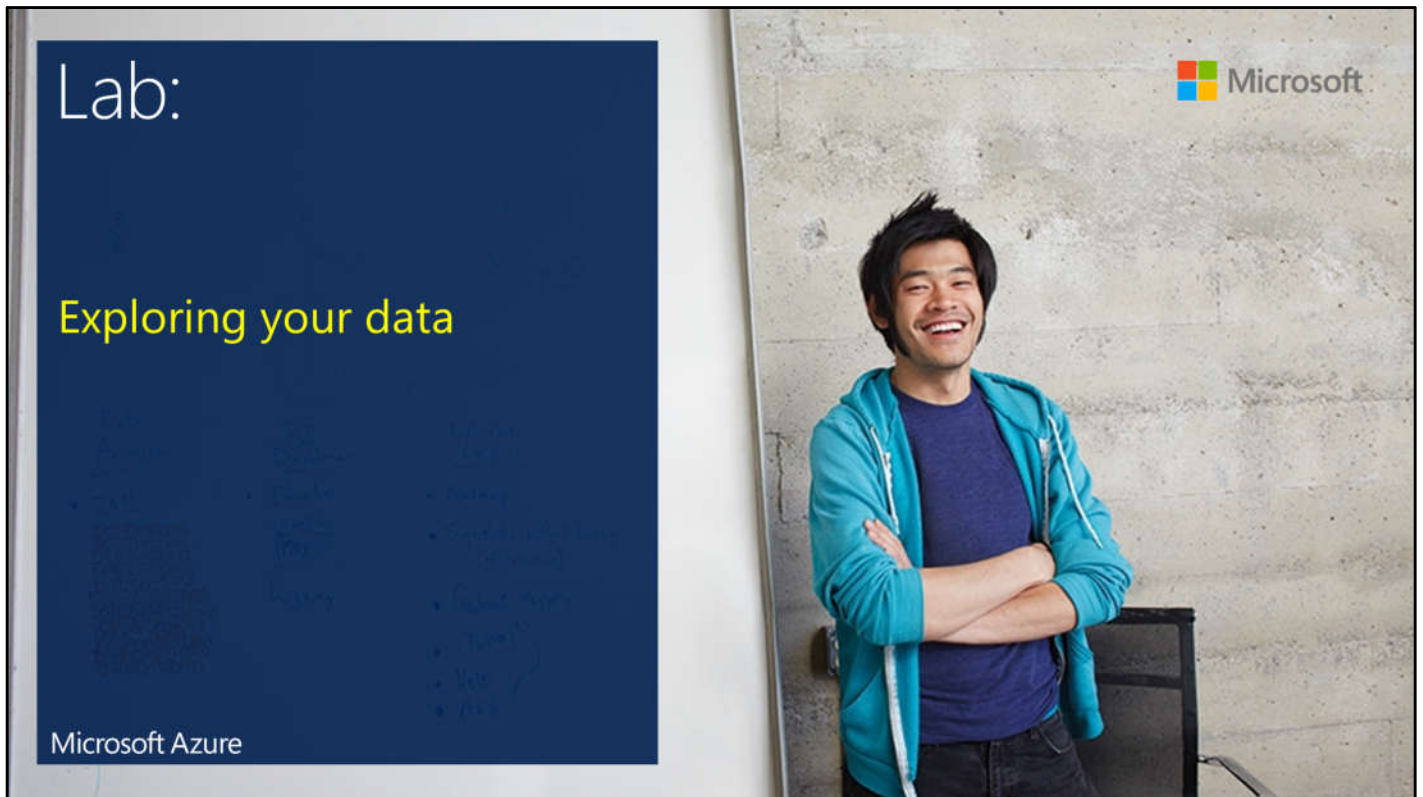
18

1. Understanding the statistics of exploring data:  
[http://danshuster.com/apstat/apstat\\_chap01.pdf](http://danshuster.com/apstat/apstat_chap01.pdf)

## Exploring Data

- MRS
- Azure ML
- HDInsight
- Other Tools

1. Data Exploration and Predictive Modeling with R - <https://msdn.microsoft.com/en-us/library/mt590947.aspx>
2. Data Exploration with Azure ML - <https://blogs.technet.microsoft.com/machinelearning/2015/09/24/data-exploration-with-azure-ml/>
3. Statistics Using Excel – <http://www.excelfunctions.net/Excel-Statistical-Functions.html>
4. Sed, awk, grep (in Windows as well) - <https://www.simple-talk.com/cloud/data-science/data-science-laboratory-system---testing-the-text-tools-and-sample-data/>



1. Using the `building.csv` and `HVAC.csv` files in your `\Resources` folder, use R, Excel, Azure ML or any other exploration tools you've seen in the class to explore the shape, size, layout, distribution and other characteristics you can find in the data.
2. Document that in any format and be ready to discuss.



1. Understand how to source and document data locations
2. Understand Azure Storage Options
3. Use various methods to ingest data into Azure Storage
4. Examine data stored in Azure Storage
5. Use various tools to explore data

© 2018 Microsoft Corporation. All rights reserved.

Questions?