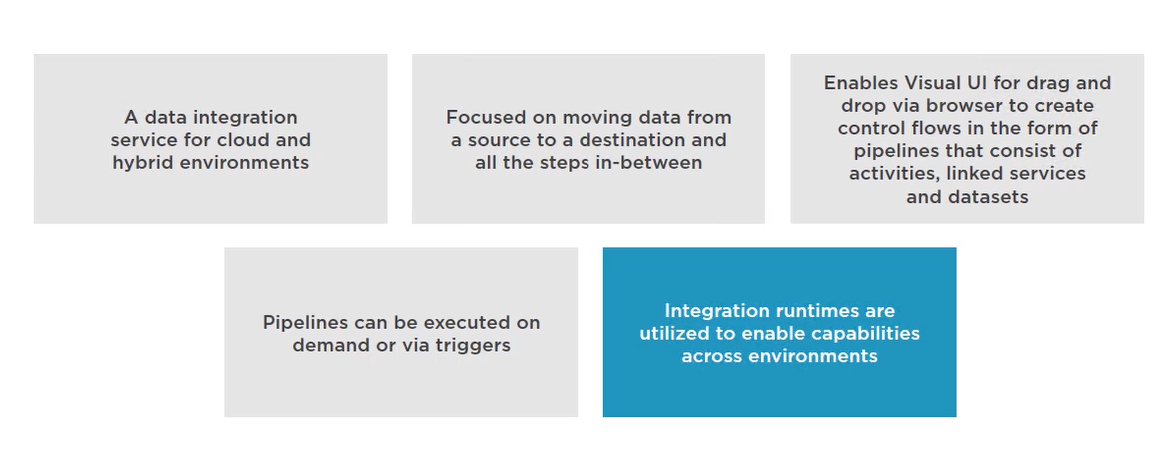
**Azure Data Factory**

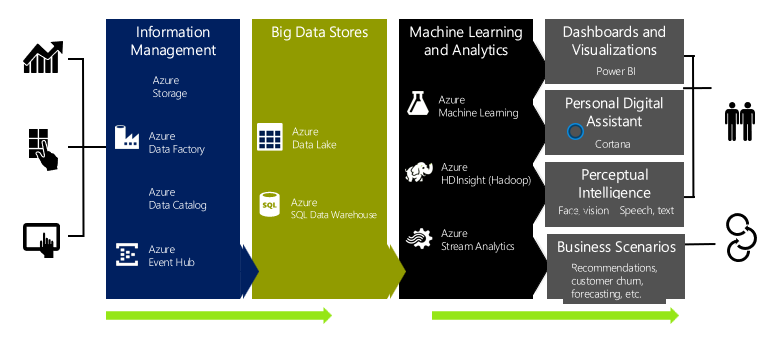
**Objectives**

* + **Understand where ADF fits in Cortana Analytics**
  + **Understand how ADF Works, and its components**
  + **Be able to deploy and manage a simple ADF implementation**
  + **Transformation of Data BLOB to BLOB**
  + **Transformation of Data SQL to BLOB**
  + **Transformation of Data SQL DB to Azure Data Lake**
  + **Transformation of Data BLOB to SQL DB**

**Azure Data Factory**

****

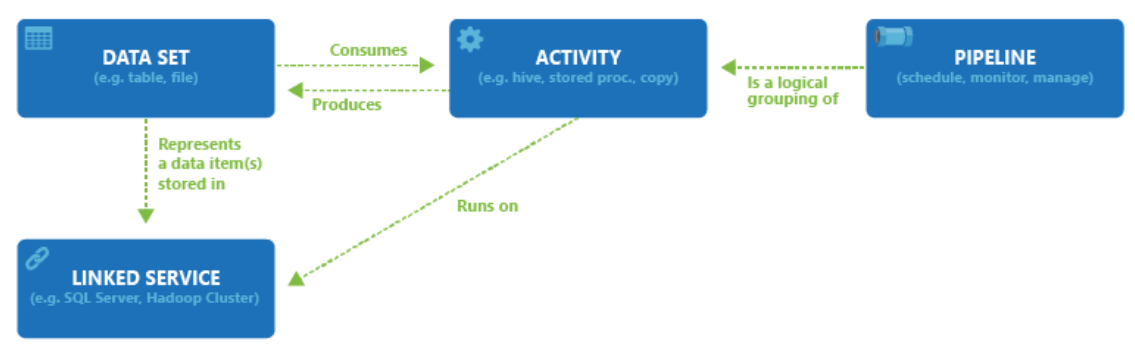
**Operationalized Analytic Solutions**

****

**Azure Data Factory- What we can do using ADF?**

**& ADF Components**

* **ADF use to manage data movement (ETL) and enrichment through the cloud and on premises.**
* **The following diagram shows the relationships among pipeline, activity, dataset, and linked service in Data Factory:**

****

**Datasets in ADF**

* **A dataset is a named view of data that simply points or references the data you want to use in your activities as inputs and outputs.**
* **Datasets identify data within different data stores, such as tables, files, folders, and documents.**
* **For example, an Azure Blob dataset specifies the blob container and folder in Blob storage from which the activity should read the data.**
* **Before you create a dataset, you must create a linked service to link your data store to the data factory.**

**Dataset type:**

**Azure Blob Storage, Azure Data Lake Store , Azure SQL DB, Azure Database for MySQL, All Azure Storage, On premises Databases and Files like Amazon S3 file system, FTP etc**

**For more details explore:**

[**https://docs.microsoft.com/en-us/azure/data-factory/concepts-datasets-linked-services**](https://docs.microsoft.com/en-us/azure/data-factory/concepts-datasets-linked-services)

**Dataset Options in ADF**

|  |  |
| --- | --- |
| **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** |

**Linked Services in ADF**

* **Before you create a dataset, you must create a linked service to link your data store to the data factory.**
* **Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.**
* **Think of it this way; the dataset represents the structure of the data within the linked data stores, and the linked service defines the connection to the data source.**
* **For example, an Azure Storage linked service links a storage account to the data factory.**
* **An Azure Blob dataset represents the blob container and the folder within that Azure storage account that contains the input blobs to be processed.**

**For more details explore:**

[**https://docs.microsoft.com/en-us/azure/data-factory/concepts-datasets-linked-services**](https://docs.microsoft.com/en-us/azure/data-factory/concepts-datasets-linked-services)

[**https://docs.microsoft.com/en-us/rest/api/datafactory/linkedservices/get**](https://docs.microsoft.com/en-us/rest/api/datafactory/linkedservices/get)

**Sample Scenario on Datasets and Linked Service of ADF**

**Sample scenario:**

1. **To copy data from Blob storage to a SQL database, you create two linked services: Azure Storage and Azure SQL Database.**
2. **Then, create two datasets: Azure Blob dataset (which refers to the Azure Storage linked service) and Azure SQL Table dataset (which refers to the Azure SQL Database linked service).**
3. **The Azure Storage and Azure SQL Database linked services contain connection strings that Data Factory uses at runtime to connect to your Azure Storage and Azure SQL Database, respectively.**
4. **The Azure Blob dataset specifies the blob container and blob folder that contains the input blobs in your Blob storage.**
5. **The Azure SQL Table dataset specifies the SQL table in your SQL database to which the data is to be copied.**

**Activity & Pipeline in ADF**

* **A data factory can have one or more pipelines. A pipeline is a logical grouping of activities that together perform a task.**
* **For example, a pipeline could contain a set of activities that ingest and clean log data, and then kick of a Spark job on an HDInsight cluster to analyse the log data.**
* **The beauty of this is that the pipeline allows you to manage the activities as a set instead of each one individually.**
* **For example, you can deploy and schedule the pipeline, instead of the activities independently.**
* **The activities in a pipeline define actions to perform on your data.**
* **For example:- you may use a copy activity to copy data from an on-premises SQL Server to an Azure Blob Storage. Then, use a Hive activity that runs a Hive script on an Azure HDInsight cluster to process/transform data from the blob storage to produce output data. Finally, use a second copy activity to copy the output data to an Azure SQL Data Warehouse on top of which business intelligence (BI) reporting solutions are built.**

**Activity & Pipeline in ADF**

* **Data Factory supports three types of activities:**
  + **Data movement activities : Copy Activity in Azure Data Factory**

[**https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-overview**](https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-overview)

* + **Data transformation activities : Transform data in Azure Data Factory**

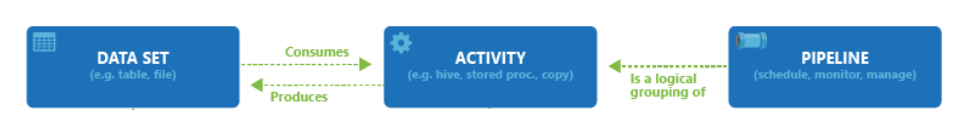
[**https://docs.microsoft.com/en-us/azure/data-factory/transform-data**](https://docs.microsoft.com/en-us/azure/data-factory/transform-data)

* + **Control activities : Web activity in Azure Data Factory**

**[https://docs.microsoft.com/en-us/azure/data-factory/control-flow-web-acActivity Options in ADF](https://docs.microsoft.com/en-us/azure/data-factory/control-flow-web-activity)**

**[tivity](https://docs.microsoft.com/en-us/azure/data-factory/control-flow-web-activity)**

* + **An activity can take zero or more input datasets and produce one or more output datasets.**
* **The following diagram shows the relationship between pipeline, activity, and dataset in Data Factory:**

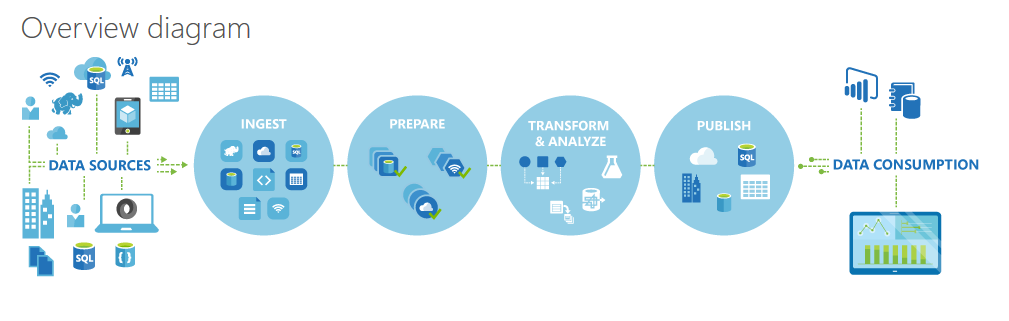
****

**Activity Options in ADF**

|  |  |
| --- | --- |
| **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** |

**ADF Logical Flow**

* **In Azure Data Factory, you can use Copy Activity to copy data among data stores located on-premises and in the cloud.**
* **After the data is copied, it can be further transformed and analyzed. You can also use Copy Activity to publish transformation and analysis results for business intelligence (BI) and application consumption.**

****

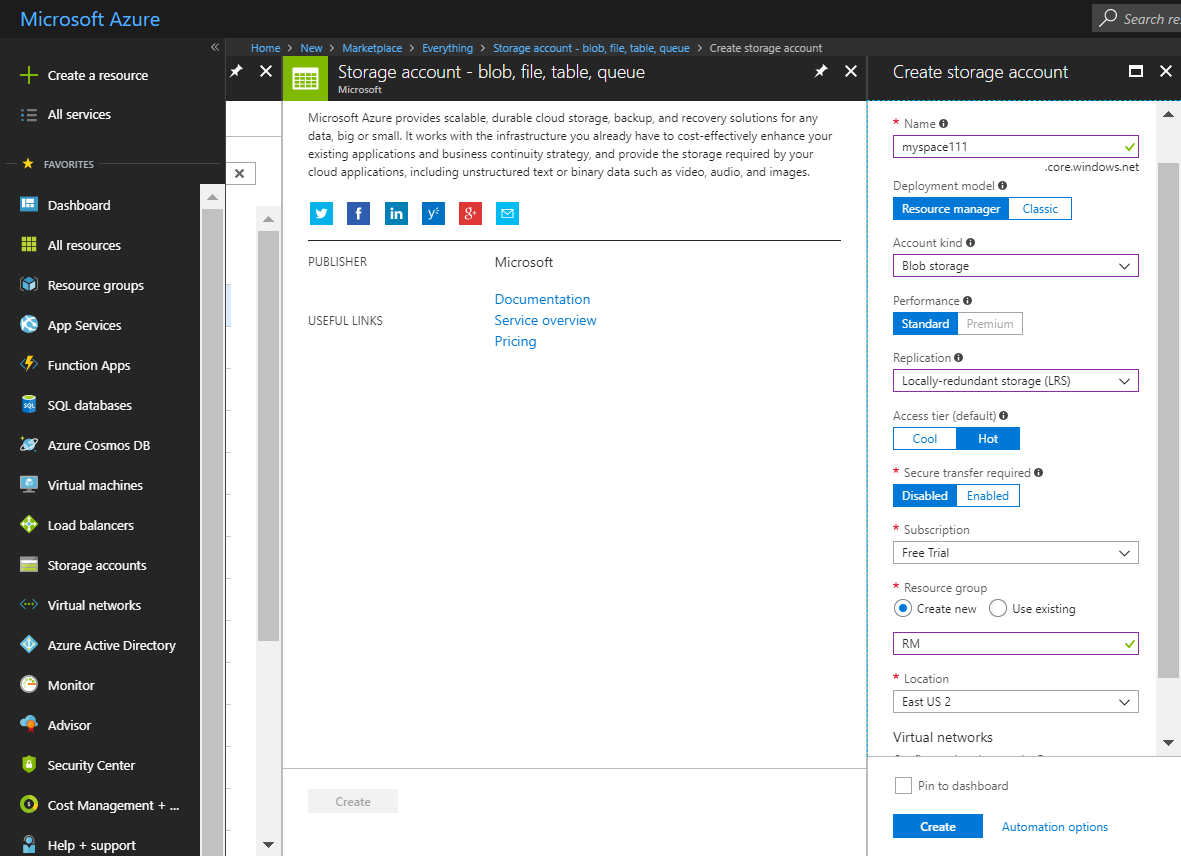
**ADF Process**

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**

**End to End Source to Target ADF Demo**

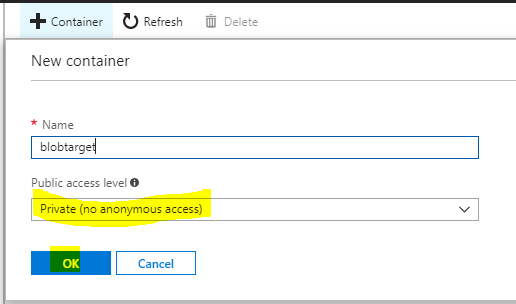
**Blob to Blob Copying data using Data Factory**

* **Deploy Blob Storage Account** 🡺 **Go To Storage Account** 🡺 **Blob Account**

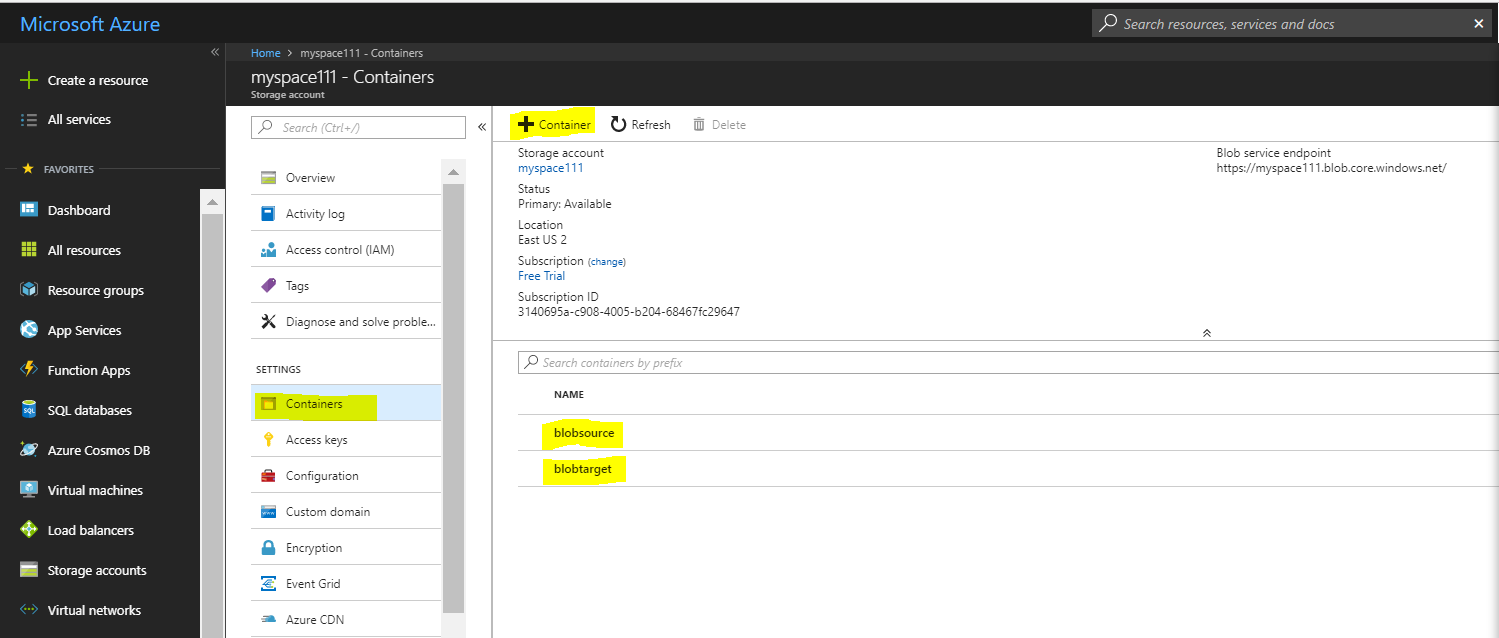
****

**Create Container for Source and Target**

* **Create container** 🡺 **Source and Target** 🡺 **load a file into source.**

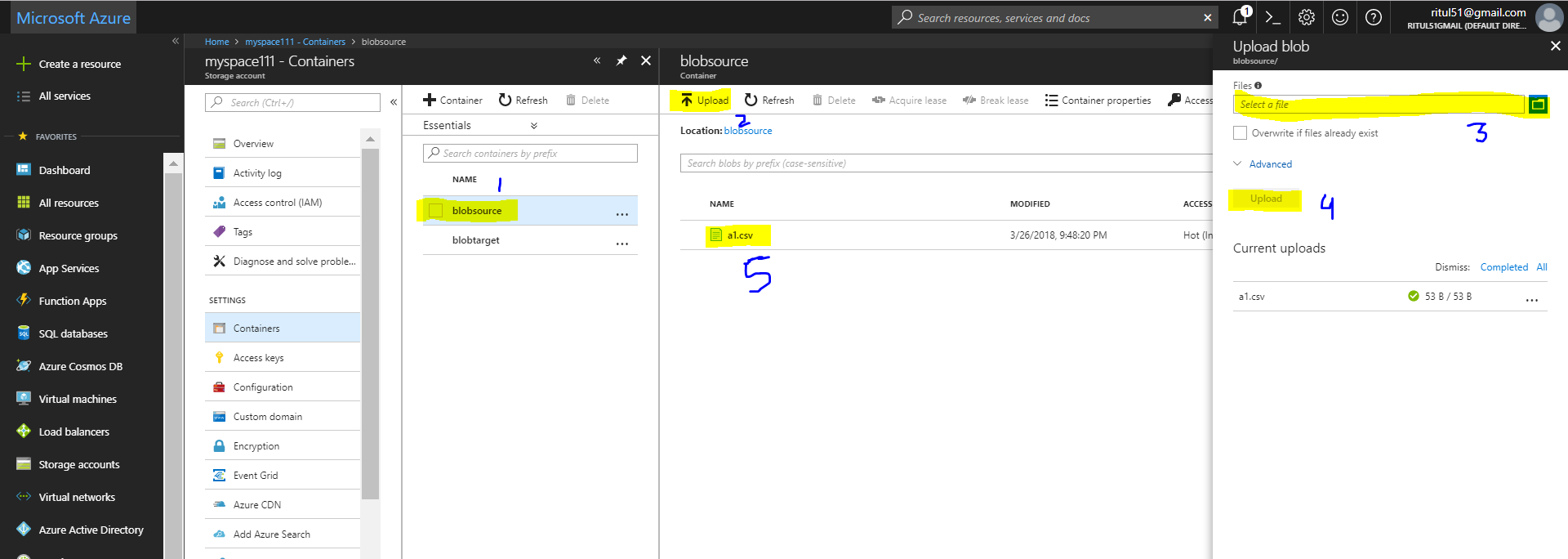
****

Create Container for Source and Target



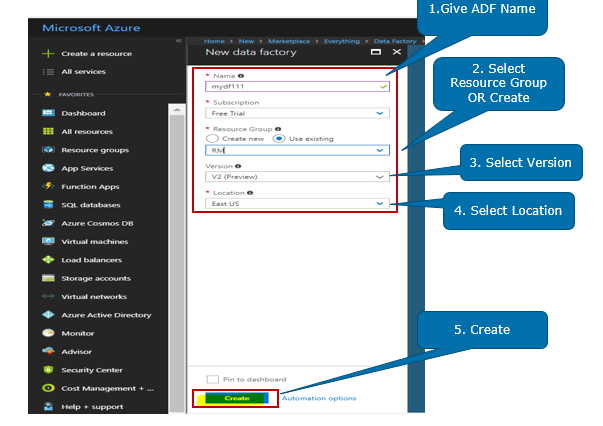
Load Data to Source Container

Select file 🡺 a1.csv 🡺 Load



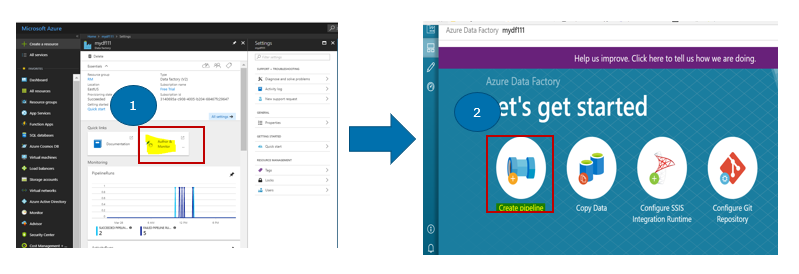
Deploy Data Factory

Go To 🡺 Create Resource 🡺Select Data factory 🡺 Fill all Details



Open Author and Monitor

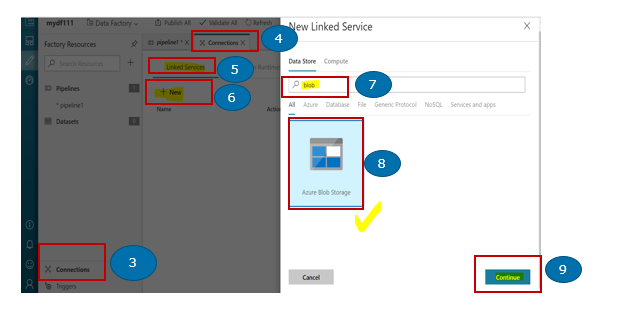
Go To 🡺 Author and Monitor



Create Linked Service

Create linked services in a data factory to link your data stores and compute services to the data factory. Here you don't use any compute service such as Azure HDInsight or Azure Data Lake Analytics. You use two data stores of type Azure Storage Blob as source & destination.

Go To Connection 🡺 Create linked Service 🡺 Click on +New 🡺 Add whatever you want to link with Data factory in this case its Blob Storage.



**Copy activity performance optimization features**