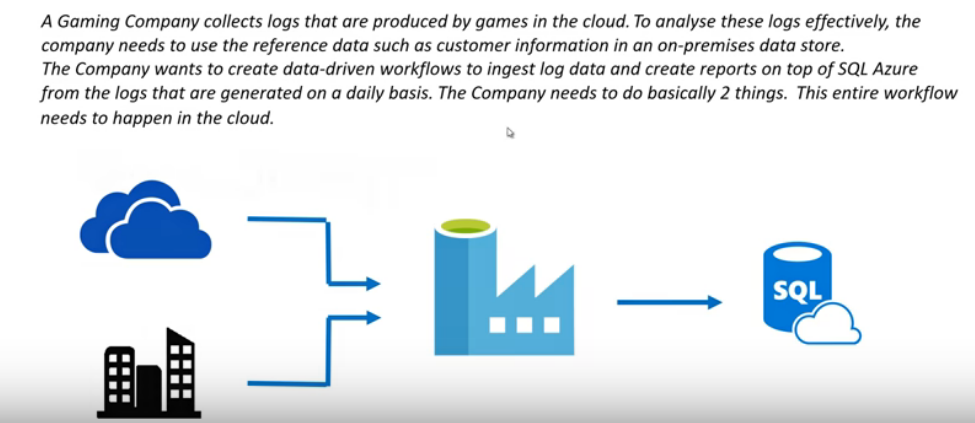
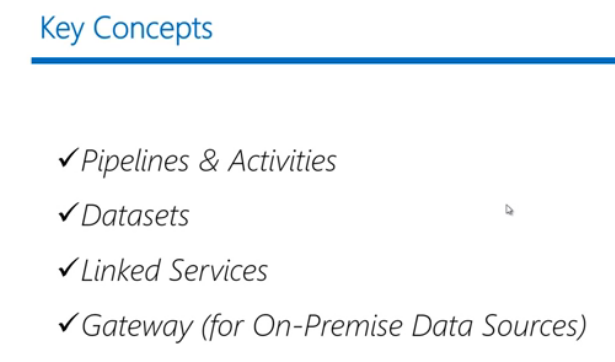
<https://blog.5nine.com/what-is-azure-data-factory-and-how-can-it-help>





Using Azure Data Factory, you can do the following tasks:

* Create and schedule data-driven workflows (called pipelines) that can absorb data from different data stores.
* Process or transform the data by using compute services such as Azure HDInsight Hadoop, Spark, Azure Data Lake Analytics, and Azure Machine Learning.
* Publish output data to data stores such as Azure SQL Data Warehouse for business intelligence (BI) applications to consume

Azure Data Factory is a cloud-based data integration service

We can load the data from different data sources from azure and on premises

It allows you to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation

A data factory can have one or more pipelines.

**Pipeline**

ADF can have one or more piplilines

A pipeline is a logical grouping of activities that together perform a task

The activities in a pipeline define actions to perform on your data

There will be 2 activities



For example, you may use a copy activity to copy data from on-premises SQL Server to an Azure Blob Storage

A pipeline allows you to manage activities as a set instead of each one individually.

For example, you can deploy, schedule, suspend, and resume a pipeline, instead of dealing with activities in the pipeline independently

**Dataset**

It is named view of data that simply references the data you want to use in your **activities** as inputs and outputs

Before you create a dataset, you must create a **linked service** to link your data store to the data factory

Dataset represents the structure of the data within the linked data stores

Dataset contains the information like

1. Schema 2. Table name 3. Location path(file based)

**Azure Blob dataset** represents the blob container and the folder within that Azure storage account that contains the input blobs to be processed

**Linked service**

Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources

Data Factory supports two types of activities:

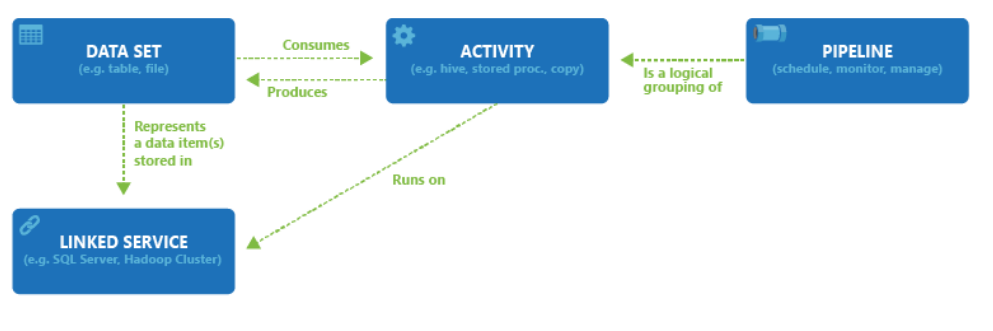
Data movement activities and data transformation activities.

Each activity can have zero or more input [datasets](https://docs.microsoft.com/en-us/azure/data-factory/data-factory-create-datasets) and produce one or more output datasets

An input dataset represents the input for an activity in the pipeline and an output dataset represents the output for the activity

**Gateway**

Incase if you want to access data from on premises you should create gateway to connect on premise to cloud. Because linked service cannot access the on premises directly



**Integration Runtime**

If you want to copy data from SQL server which is in local server (VNet) to ADF

We will download the self-hosted IF from ADF, copy the Authentication Key, and install it in local server. While installing we should give the authentication key. Once its successfully installed it will automatically connect in ADF

<https://www.mssqltips.com/sqlservertip/5812/connect-to-onpremises-data-in-azure-data-factory-with-the-selfhosted-integration-runtime--part-1/>

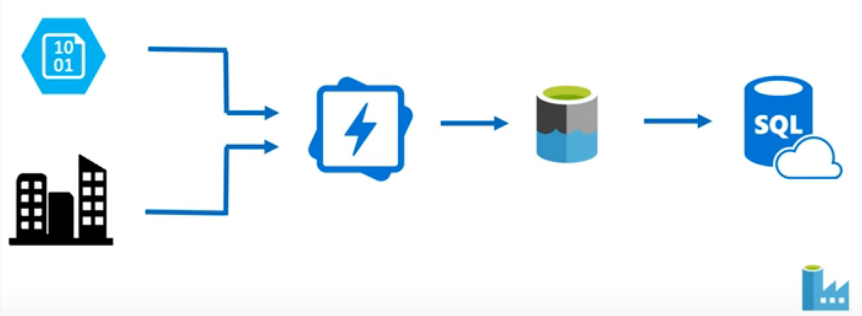
**Example**

We have blob storage that contains log, which stored everyday

Then we have customer information data in SQL, which is in on premise

We will summarize the both data’s in weekly once using **U SQL** activity and stored in Data Lake

Then summarized data copied into SQL server



Blob  Dataset  Linked service Pipeline Datalake 

Gateway  on premise 

