Reference : <https://azure.microsoft.com/en-in/resources/videos/detailed-introduction-to-azure-data-factory/>

<https://docs.microsoft.com/en-us/azure/data-factory/introduction>

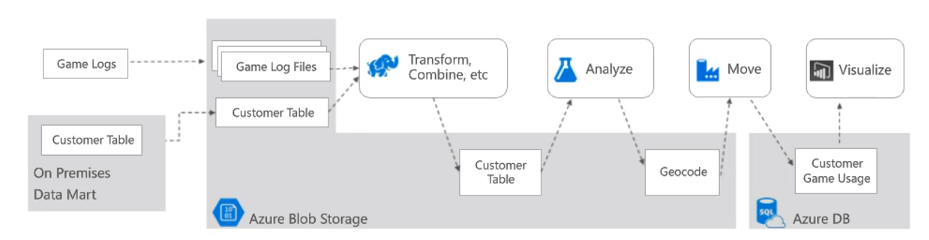
Azure Data Factory:

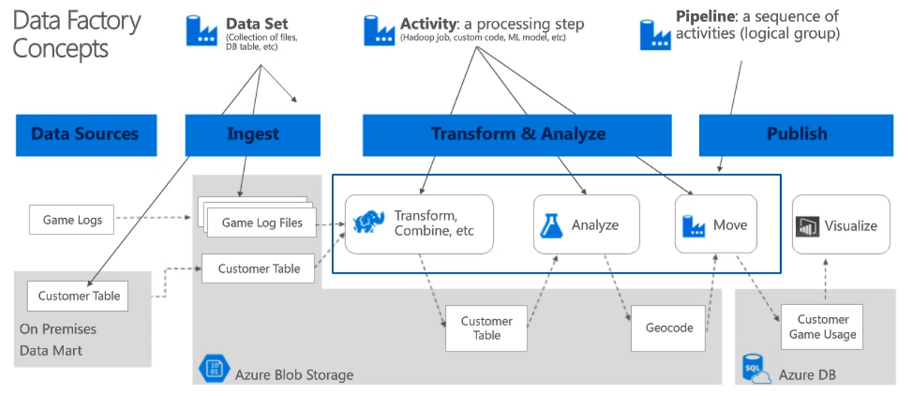


Azure blob storage (load the data from different sources and transform and preprocess in blob storage). And we can use the processed data in azure ml for analyzing and predicting eg. Churn of customer etc.



Move the processed and predicted data to Azure SQL database where it can be easily accessed for visualization using power BI.

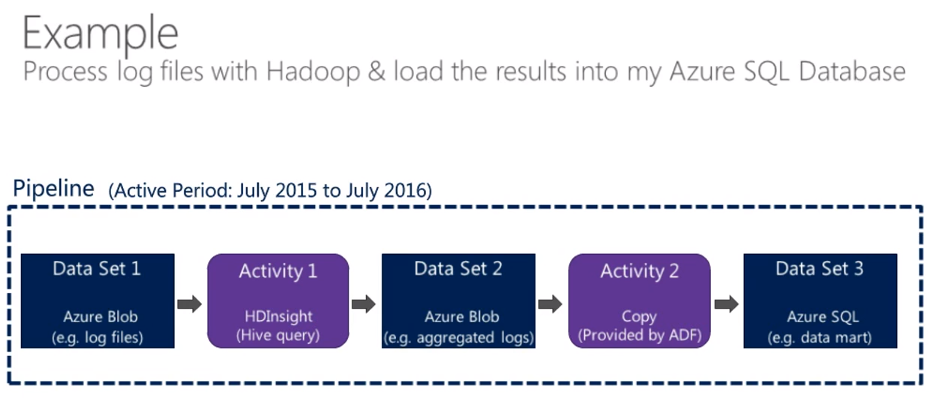




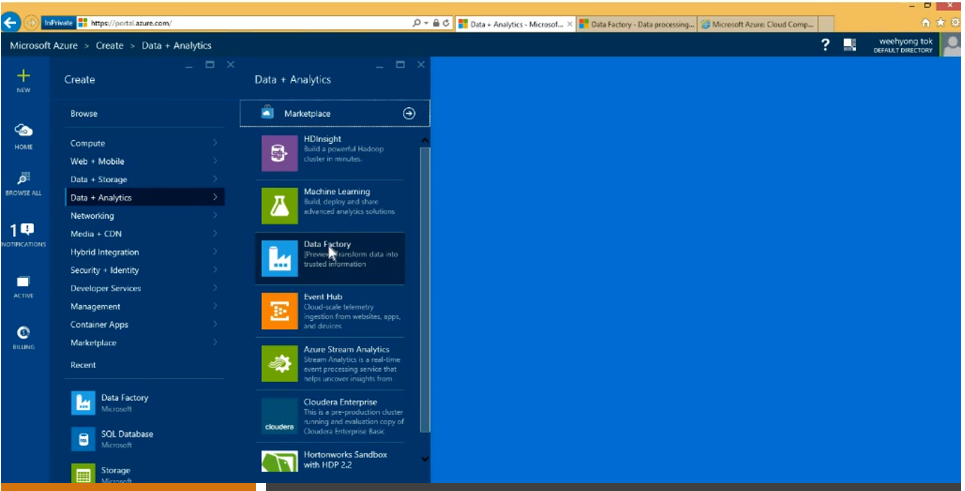
**Azure Data Factory:**

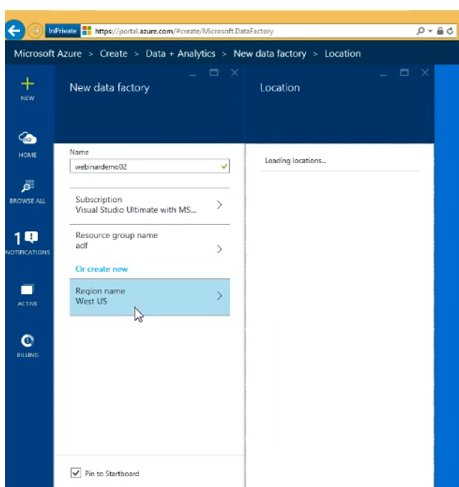
A managed cloud service for building and operating data pipelines in both on-premise and on cloud.

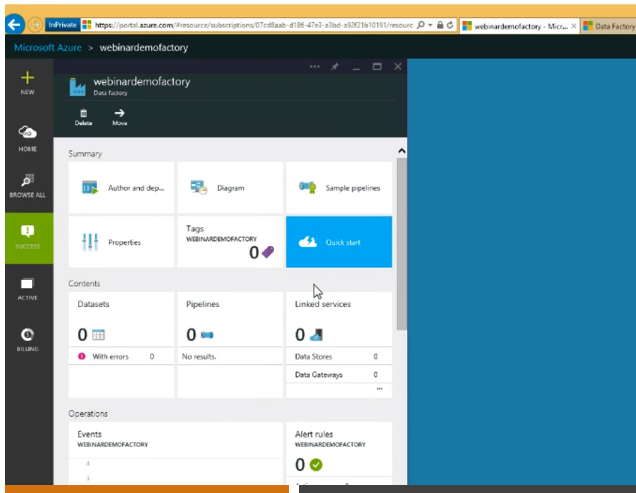




In Azure portal, select new and Data + Analytics and Data Factory



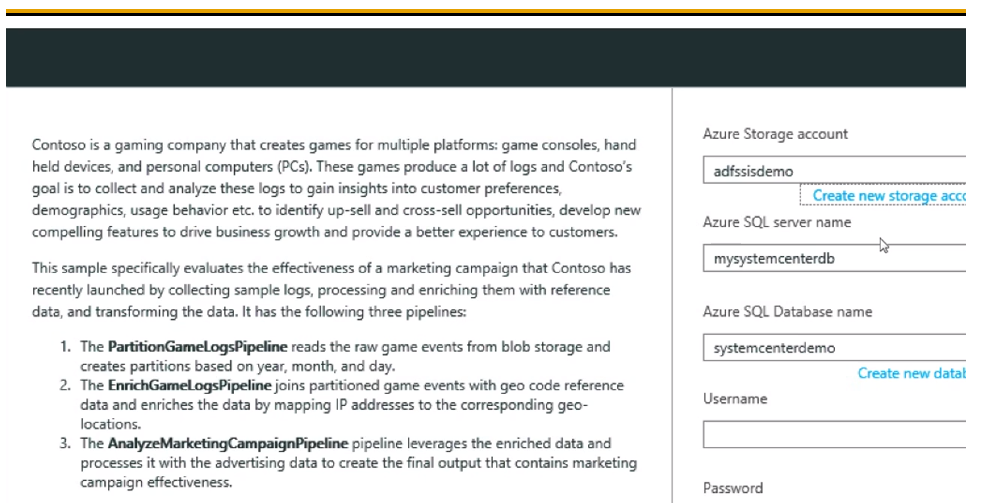


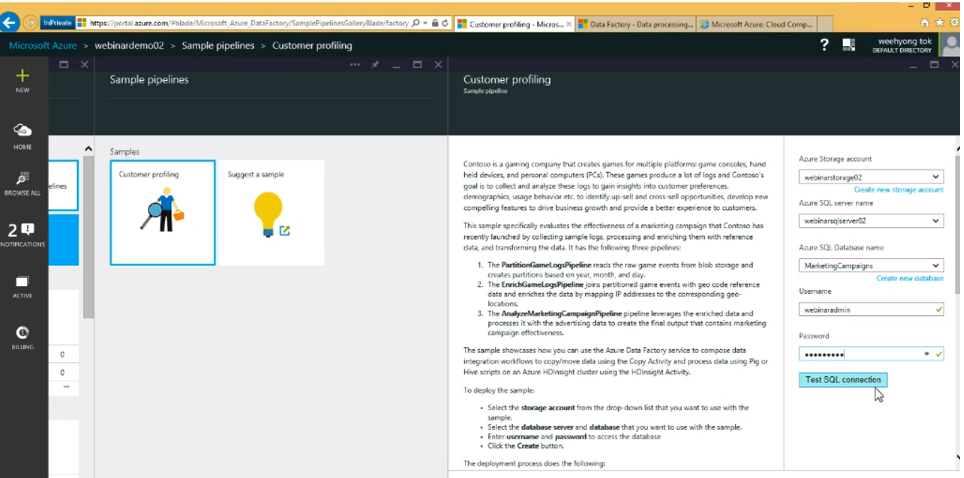


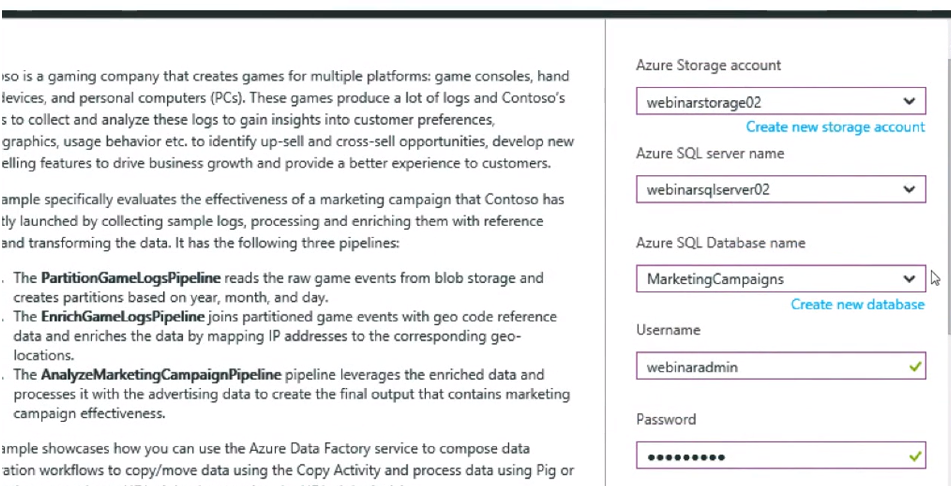
3 concepts important:

1. Data sets -> refers to Data sitting in various locations like Azure blob, Azure SQL database or on-prem file system
2. Pipelines -> grouping of activities. Data movement(copying from on-prem sql server to azure blob storage) or data processing (kick starting HDinsight cluster) this will become activity and will be grouped as pipelines
3. Linked Services -> SSIS customers uses SSIS connection manager to connect with various data sources. Linked services are like SSIS connection manager. Contains info about data sources like incase of database (username and password), in case of azure storage (account name and account key) or credentials to connect to HD cluster (username, password)

**Sample pipeline(customer profiling) in Azure:**







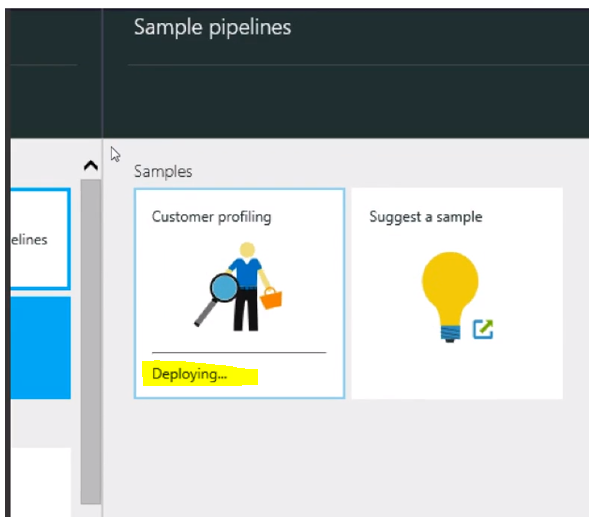
Azure storage account – deploy sample data for end to end solution

Azure SQL server name – name of azure sql server in the subscription

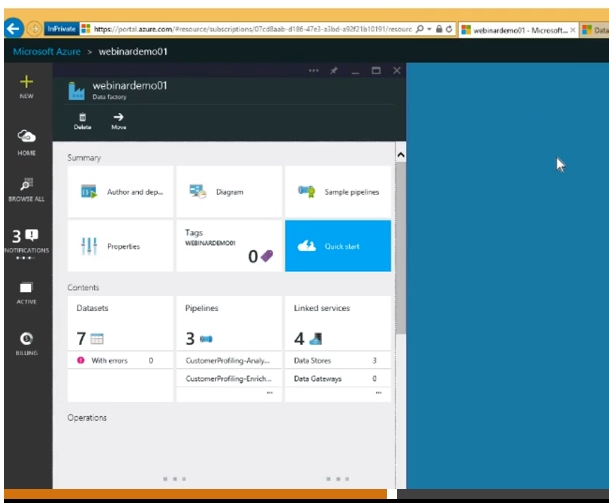
Azure SQL Database name – name of the database where the results will be loaded

Credentials (username and password) -> of Azure sql database

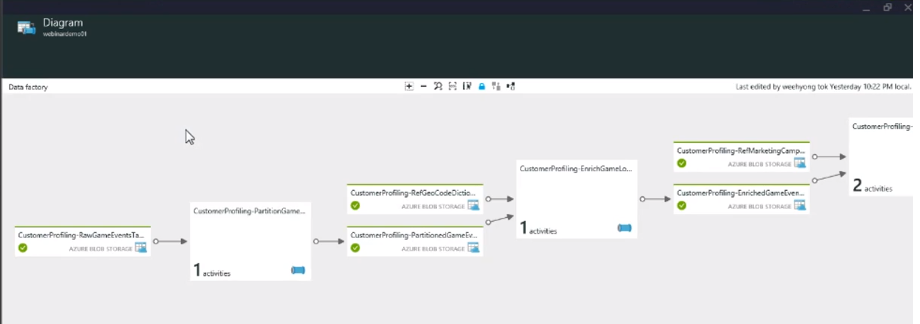
Click create.. we can see that the customer profiling pipeline gets deployed.

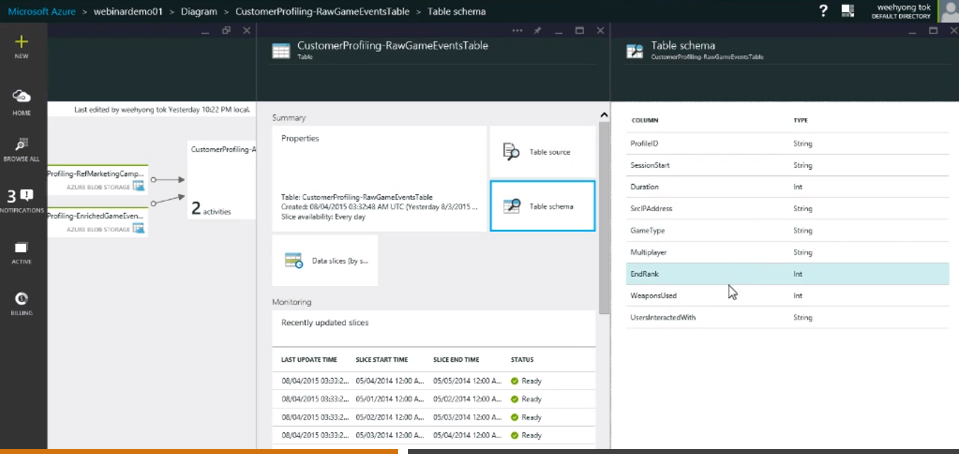


After deployment, we see as below,



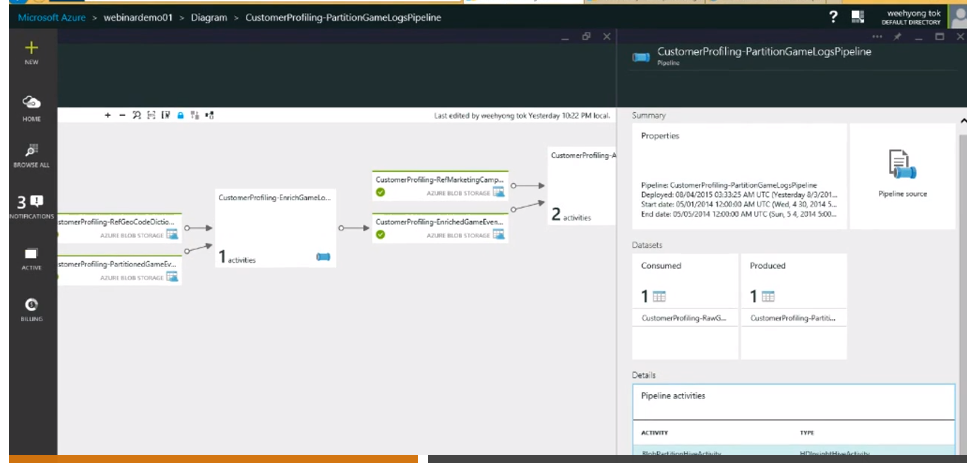
When we click on diagram, we see as below..

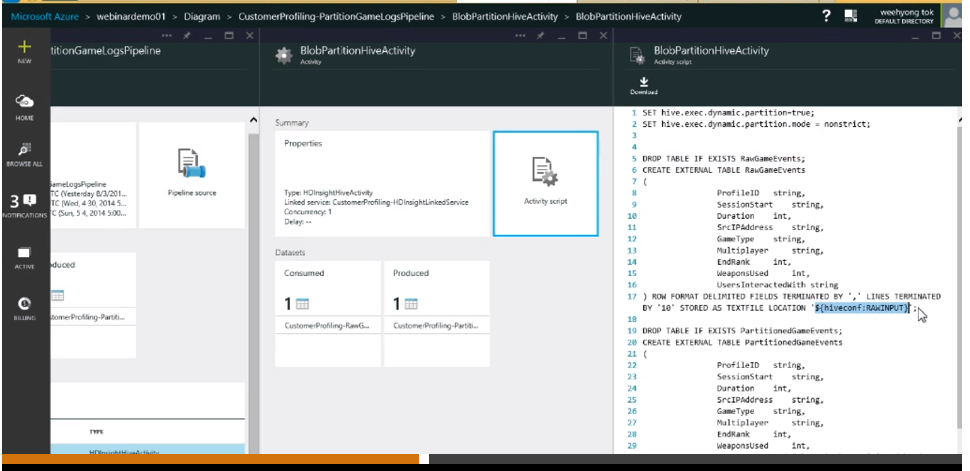


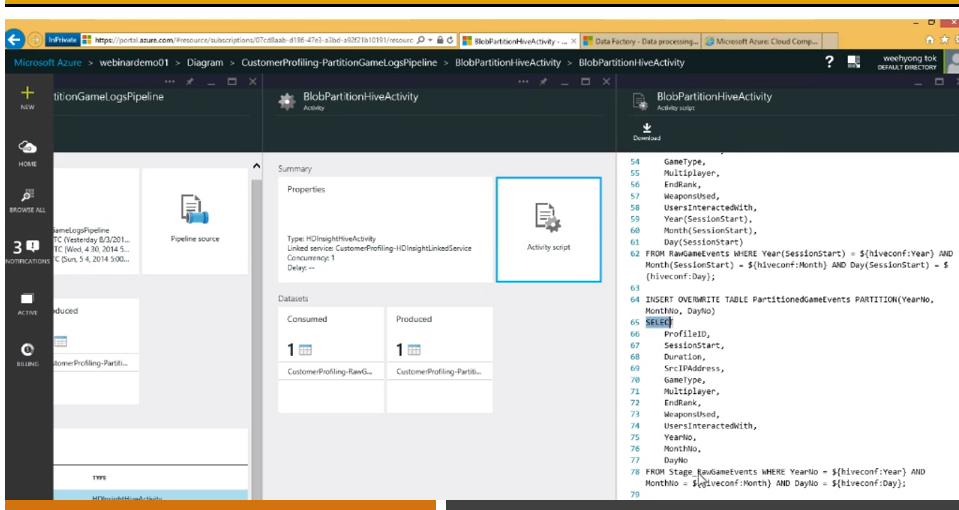


If we click on, Azure blob storage, we will see table schema details. Like duration spend by user, ip address etc.

Inside the pipeline, there is one activity which is HDinsight hive activity which is to create partition on year, month and day using hive script.



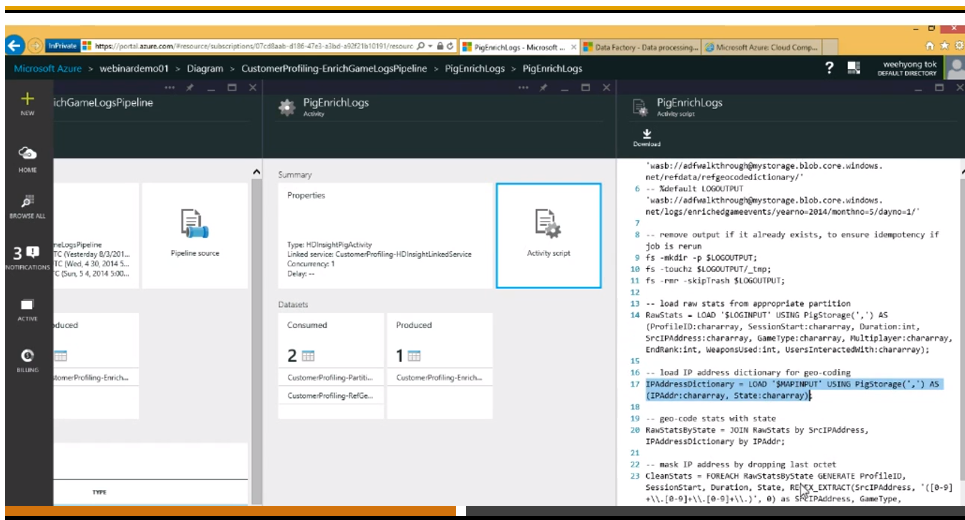


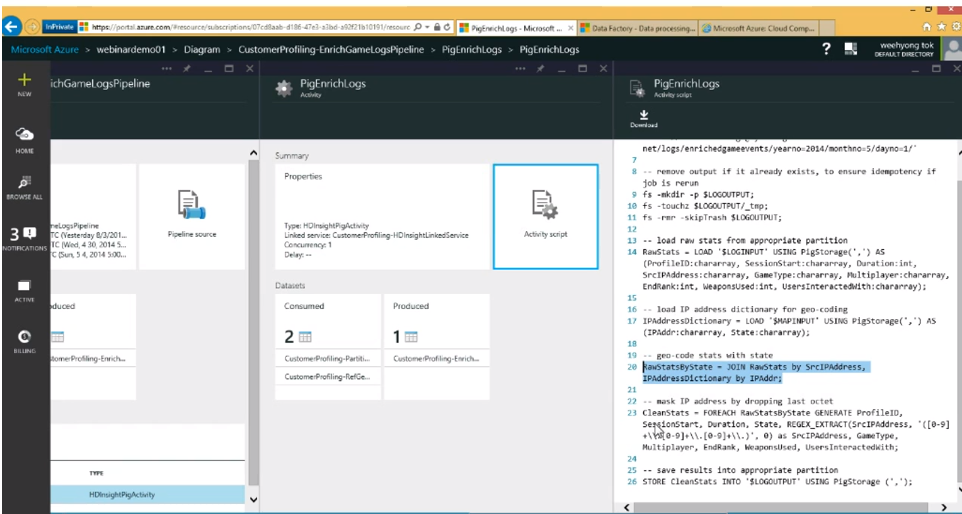


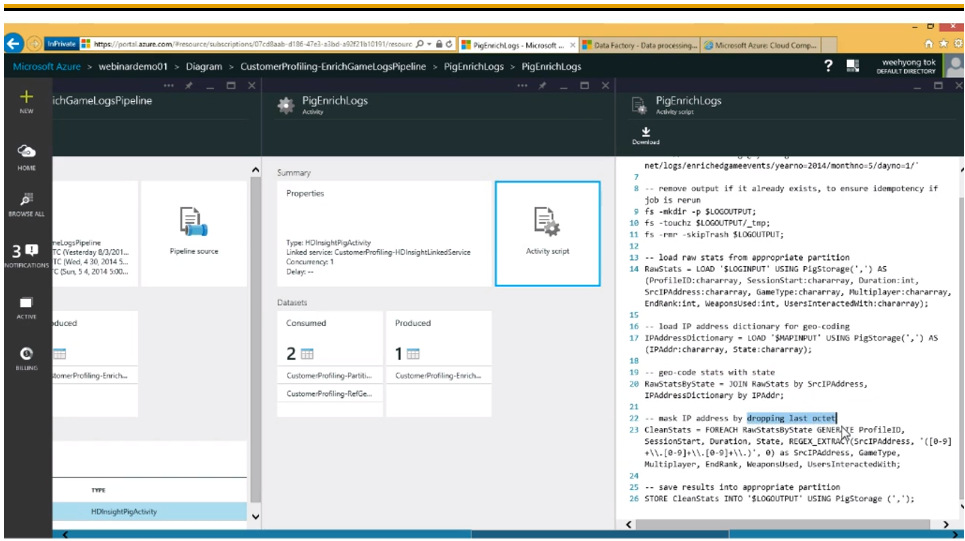
Next is a pipeline to resolve IP into geographical location.

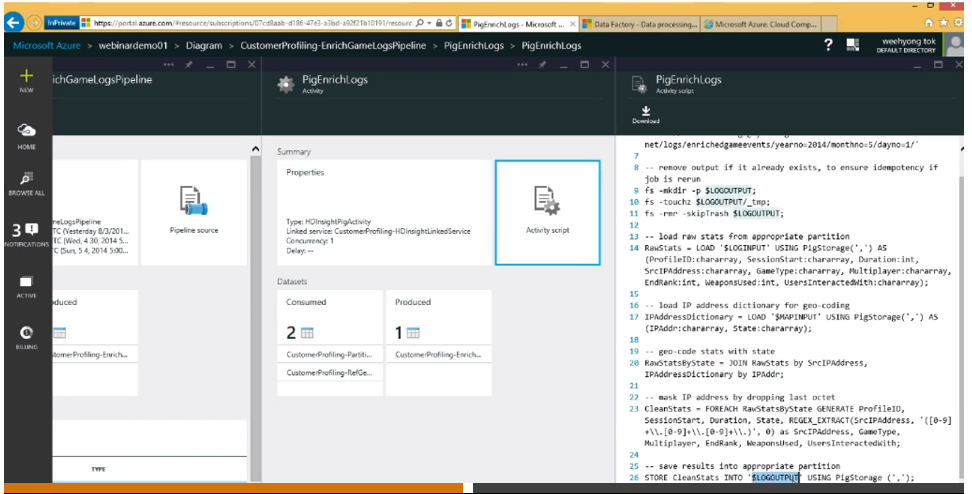
This is using HDInsightPigActivity. Reason for choosing pip is the developers choice.

In pig script, we are loading dictionary, joining with geo code with state, masking the ip address by dropping last octet and storing the data into $logout.

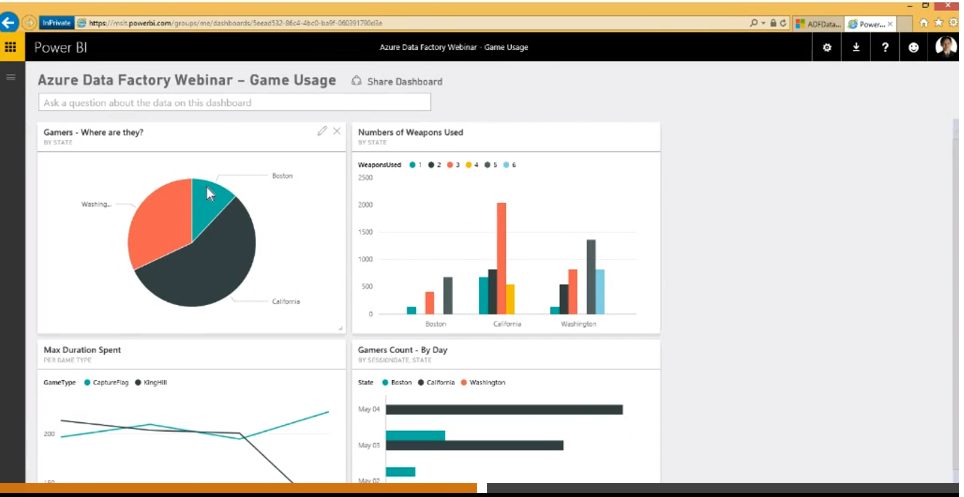




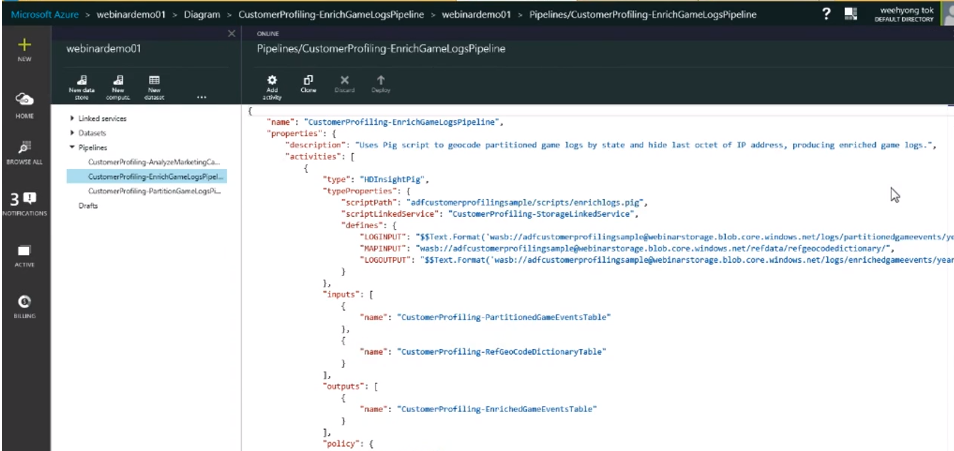


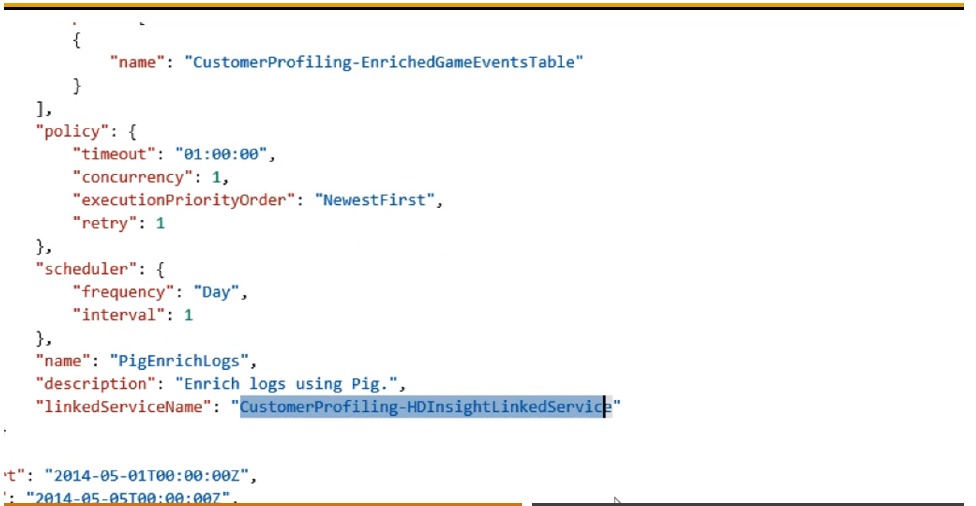


Last we load the data to azure sql database. End result is the power BI dashboard.



We can create pipeline, or dataset using json. When u select a pipeline and click on pipeline source, you will see..



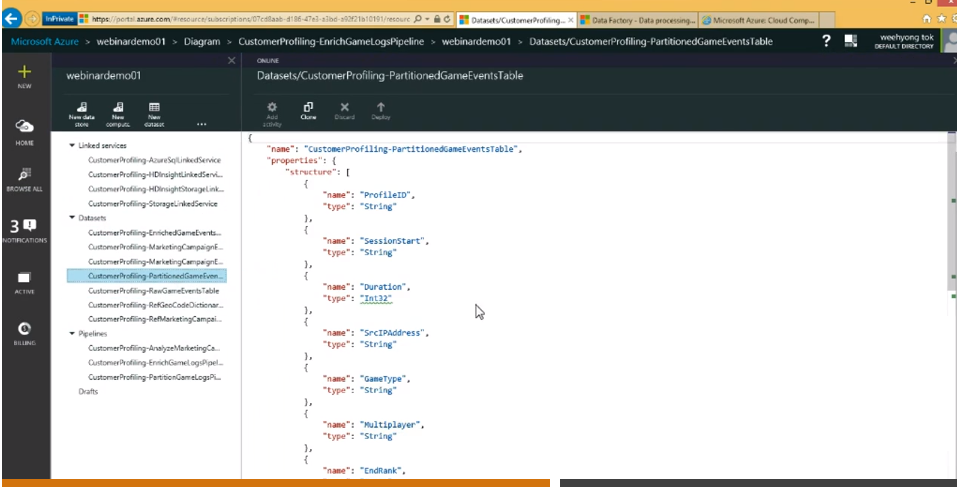


For the pig script to run, it is using the HDInsightLinkedService as mentioned above.

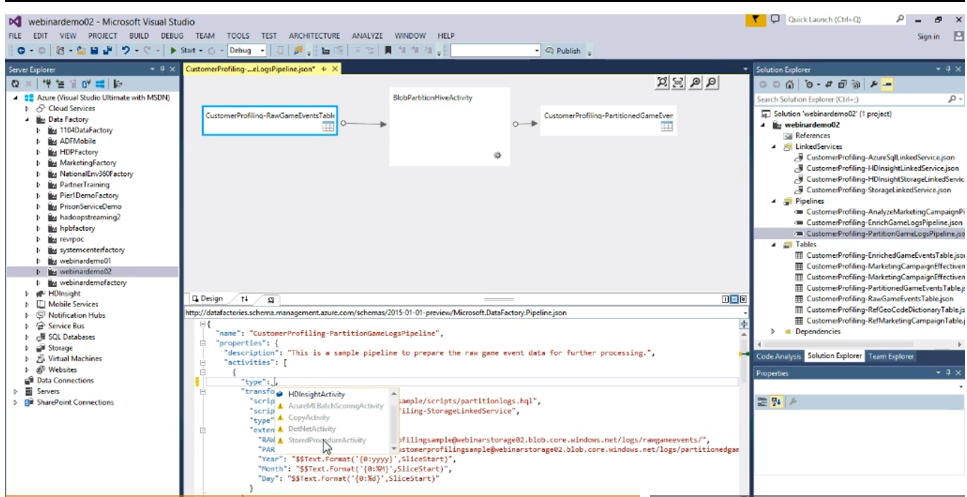


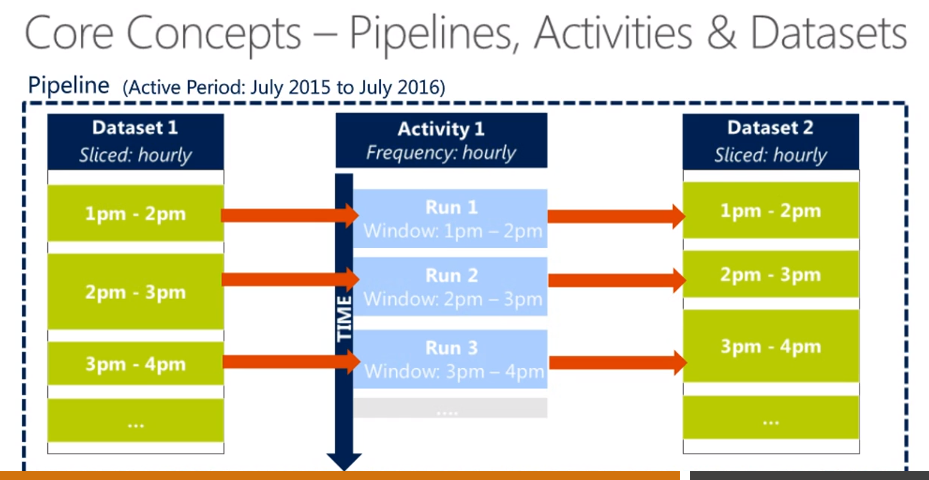
Linked service -> contains how to connect with the linked service.

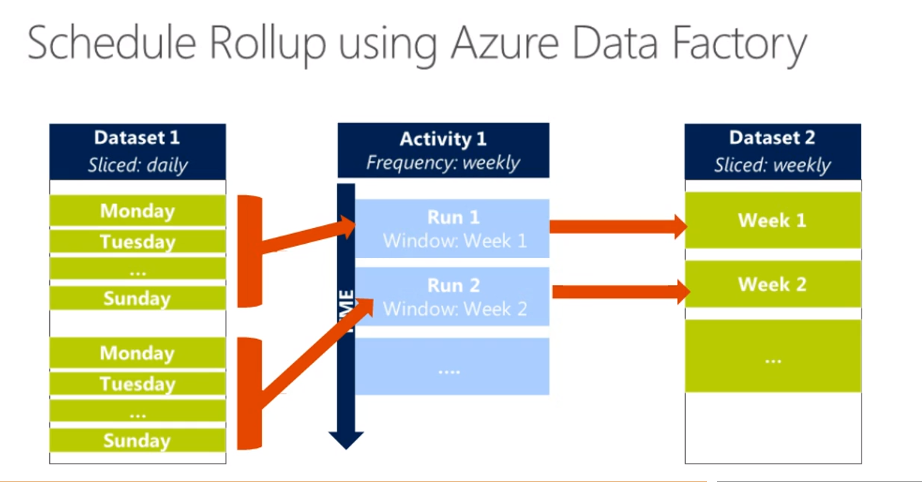
Let’s look at partition dataset.

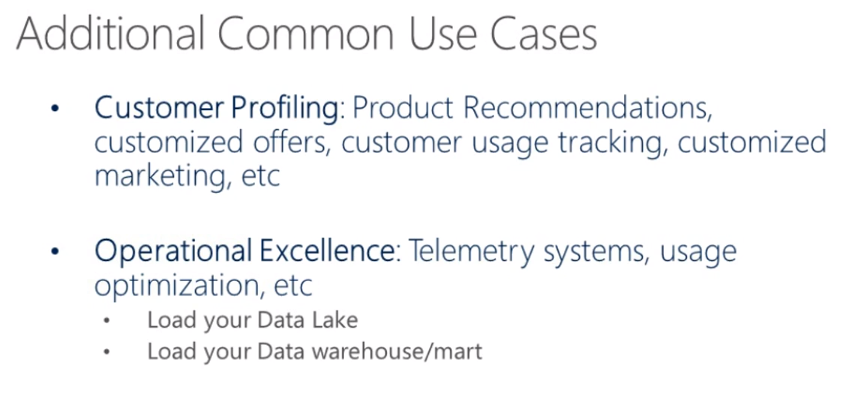


We can use visual studio by installing azure data factory component and then we can work same as SSIS.

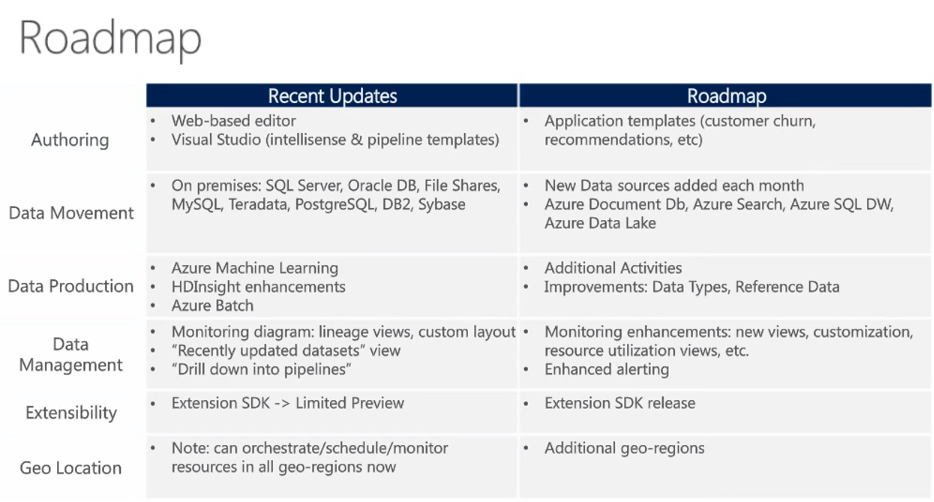


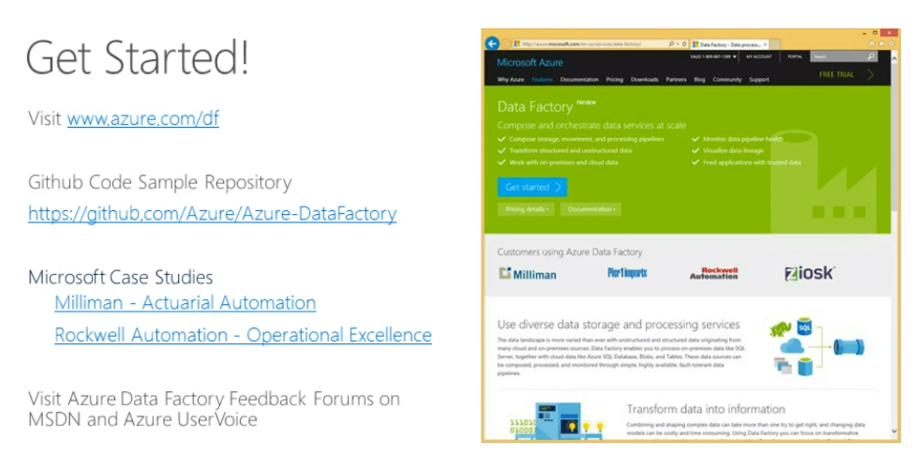






**Sample Live ADF:**





**Questions:**

Data security, ADF requires all data to be uploaded to cloud?

Ans:

Processing using bigdata like hdinsight, then u need to upload to blob..if processing is on premise, ADF orchestrates the data processing on premise.u can put on sql server database and there is no need to upload on azure blob if you are not taking advantage of other analytical services in azure.

Some finance clients, do mask the critical data on premise itself and then loads to azure blob and make use of azure machine learning and other HDinsight services and then take the data to on-premise before merging with other critical data and delivering to customers as reports.

Can we run R or power shell from azure df?

Not possible. In product backlog.

Can u invoke a web service published by azure ml and then able to do batch scoring and then push data to relevant data sources?

Yes, adf provides machine learning activity. Only thing to integrate with published web service from azure machine learning, you need URI and api key for the adf to do batch scoring and copy the data to relevant data sources supported by azure data sources.