Azure Fundamentals

Data Storage - blob, ADLS gen2

Amazon S3 and ADLS gen2 are object stores

data lake gen 2

Azure Data Factory

Synapse Analytics

Azure Databricks

HDInsight

===========

RDBMS

Amazon s3

Website

There are 2 requirements

1. you have to build a data platform for your DataScience team

2. you have to create a platform for your visulization or your data analytics team - AD hoc

queries

Sources -> Ingest (Azure Data Factory)

Azure Data Factory (Fully Managed and Serverless service)

==================

1. To transfer the data from Source to a Sink

blob -> adls gen2

rdbms -> adls gen2

adls gen2 -> rdbms

2. Transformations - Data Flow in Azure data factory

so do not write any code, and we just do it graphically.

Internally spark code is created and submitted on a cluster.

3. Orchestration

source -> Ingest -> Tranformations

source -> Azure Data Factory -> Data Flows/HDInsight/Databricks/Azure Synapse Analytics ->

Target

HDInsight (EMR equivalent in AWS)- Managed Hadoop cluster on Azure - Hive, spark, pig -

Hortonworks cluster (google has dataproc)

Databricks - specially for spark, optimized spark cluster

The pipeline can be triggered/scheduled using Azure Data Factory.

Data Platform for your machine learning team

source -> ADF -> DataLake Gen2 -> synapse/databricks/hdInsight/Dataflows -> blob/datalake

blob storage is slightly cheaper than your datalake storage

for adhoc reporting needs for your analytics team you would take subset of the transformed data

and keep it in

Azure Sql Database - RDBMS

Azure Synapse Analytics - Data warehouse

Tableau, PowerBI, Graffana

ADF Session - 1

=================

Its a completely managed and serverless service

1. Transfer

2. Transformations (Data Flow)

3. Orchestration

we will have a RDBMS table

I want to take this RDBMS table and put to the sink (Datalake Gen2)

Source - RDBMS table

Sink - ADLS Gen2

Azure SQL Database (SQL Server)

Azure Database for Mysql

Azure Database for postreSQL

Azure Database for MariaDB

Azure Data Studio

SQL Server Management Studio

ADF Session - 2

================

Source - Azure SQL Database

Sink - DataLake Gen2

Source -> Azure Data Factory -> Sink

Author

Monitor

Manage

Source -> Sink

Source - Azure SQL Database (done)

Sink - ADLS Gen2 ()

step 1 - is connect to both the source and sink

We do it using the Linked Service

step 2 - create datasets for both the source and sink

step 3 - create the pipeline and have a copy activity

========

```
create table courses(
course_id int NOT NULL,
course_name varchar(30) NOT NULL,
course_duration_months int NOT NULL,
course_fee int NOT NULL,
PRIMARY KEY(course_id)
);
insert into courses values(1, 'big data', 6, 50000);
insert into courses values(2, 'web development', 3, 20000);
insert into courses values(3, 'data science', 6, 40000);
insert into courses values(4, 'devops', 1, 10000);
```

ADF Session - 3

================

what is azure data factory?

Data Integration, Tranformation and Orchestration

mapping data flow - it provides us the feature of creating transformation flow and internally a

spark code is generated.

places where we should not use Azure data factory

==================

ADF does not provide storage capability

suitable to do simple/medium level tranformations but not the very complicated ones

it does not provide data streaming capabilities

its not a migration tool

==========

I will have a dataset orders.csv is hosted somewhere

https://files.cdn.thinkific.com/file_uploads/349536/attachments/c28/5fb/25b/orders.csv

1. I want to take this file from external URL and ingest it to my ADLS

A working azure account - done

subscription - done

create a resource group - done

storage account - datalake - done

azure data factory - done

linked service for source - http - done

linked service for sink - datalake gen2 - done

dataset for the source - done

dataset for the sink - done

a copy activity

a pipeline

2. we want to perform some basic transformations

remove order_date column - done

rename order_customer_id to customer_id - done

I want to know the count for each order status - done

ADF Session - 4

================

mapping data flow in ADF

source -> select transformation

ADF Session - 5

================

Assignment Time

there are 2 datasets -

1) orders.csv

https://files.cdn.thinkific.com/file_uploads/349536/attachments/c28/5fb/25b/orders.csv

you need to bring orders.csv to the datalake

2) customers.csv (I will provide you)

first you need to put this file to your datalake

then you need to ingest this to azure sql database

orders.csv (datalake)

customers.csv (azure sql database)

Mapping Dataflow

1) 2 sources

the 1st source would be azure sql database (customers)

the 2nd source would be your orders.csv in datalake

2) Join transformation

3) have a select transformation and remove 3 columns

customer_Email, customer_password, order_Customer_id

4) Filter

customer_city == "Caguas"

5) Sort transformation

```
create table customer (
customer_id varchar(50),
customer_fname varchar(50),
customer_lname varchar(50),
customer_email varchar(50),
customer_password varchar(50),
customer_street varchar(50),
customer_city varchar(50),
customer_state varchar(50),
customer_zipcode varchar(50)
)
```

customer_city == "Caguas"

ADF Continuation

==================

Ingesting (very important)

transform (rarely used)

orchestration (very important)

blob connector

http connector

azure sql database

================

Retail datasets

===================

Departments -- footwear

Categories - mens footwear, kids footwear

Products - each category have multiple products

orders

customers

order_items

products.csv

this will be coming to the blob storage

we need to bring this to our datalake

copy the products.csv from blob to datalake

I would need

2 linked Services - source blob, sink adls gen2

2 datasets - products.csv source file in blob storage, target directory in azure datalake

gen2

1 pipleline...

create a storage account

create a datalake storage also

=====

resource created

===================

resource group

normal storage account

datalake

data factory

linked service - blob

linked service - datalake

dataset - blob storage

dataset - datalake gen2

pipeline

Now we have to make this solution better

==

1. we want to pick the file as soon as it is available

2. we want to do basic validation of data before we ingest it to datalake

3. if the pipeline fails then we need to show notification (email)

validation activity

get metadata activity

if condition

copy actitity if true -> delete blob file

fail pipeline if false