**The following are pointers compiled for the preparation for DxC AZURE Data Engineering roles through Bootcamp:**

**Data Factory Components:**

**Pipeline**: The activities logical container

**Activity**: An execution step in the Data Factory pipeline that can be used for data ingestion and transformation

**Mapping Data Flow**: A data transformation UI logic

**Dataset**: A pointer to the data used in the pipeline activities

**Linked Service**: A descriptive connection string for the data sources used in the pipeline activities   
**Trigger**: Specify when the pipeline will be executed Control flow: Controls the execution flow of the pipeline activities

* Develop Ability to create Pipelines from different data sources, using Historical data, Incremental data.
* Develop Ability to ingest data from a Database, Event Streaming Source (real time)

**Some Use cases/scenarios: (VERY IMPORTANT)**

Use Case 1: Process, ingest Data from a variety of sources and push it to Blob storage.

Use Case 2: AZURE provides APIs - invoke these API from Python (preferably ASP .Net/Java), dynamically from ADF.

Use Case 3: On prem to AZURE Cloud – Read files from On-prem and push it to Blob Storage, AZURE SQL, ADLS Gen2.

Use Case 4: Invoke a Databrick notebook from an ADF Pipeline.

Use Case 5: Data coming on a daily basis from Blob storage, read all the data from a folder with date in name, read using a ‘for each’ loop, from a list of folders, then push into a pipeline.

Use Case 6: Logic APP flow from Data Factory and the other way round.

This explains the various types of files that are normally used within the Big Data / Data Engineering space. This is commonly asked in interviews - <https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/>

**Mock interview Feedbacks from previous batch:**

Data Flows, ETL transformations, activities.

Basics of ADF.

Basics of ADF, SQL queries, integration runtime.  
Various activities in Data Flow, setting up connections from different data sources, which integration runtimes you can use in what scenarios.

Needs to understand where to use what activiies, SQL queries to be brushed up.

Needs a bit of hands on. Scenarios of Data Flow activities.

Common themes of gaps include Data Flows not understood well, Activities, Integration Runtimes, SQL Queries, providing solutions on IR and Activities specific to a problem scenario.

**Recommended Udemy Course:**

<https://www.udemy.com/course/microsoft-azure-database-and-analytics/#instructor-1>

**Prescribed Project Course from Udemy:**

<https://www.udemy.com/course/learn-azure-data-factory-from-scratch/#reviews>

(above one is preferred)

or

<https://www.udemy.com/course/two-real-world-azure-data-engineer-projects-end-to-end/>

**Data Factory Tutorials from YouTube (Free)**

<https://www.youtube.com/playlist?list=PLMWaZteqtEaLTJffbbBzVOv9C0otal1FO>

(above one is preferred)

or

<https://www.youtube.com/watch?v=iSpVfWo4CUI>

**Databricks Tutorials**

<https://www.youtube.com/playlist?list=PLMWaZteqtEaKi4WAePWtCSQCfQpvBT2U1>

**Microsoft Documentation on AZURE Data Engineering:**

<https://docs.microsoft.com/en-us/learn/certifications/exams/dp-203>

* go to the bottom of the above page and you will find the learning paths. Entire study materials are there.

**Interview Questions:**

* What is Datalake?
* What is serving layer in datalake?
* ACL stands for? & what are the uses of ACL?
* What are different authentication methods we have in storage?
* Why should we use for azure file shares?
* Why should we use for azure Tables?
* Why should we use for azure Queues?
* What is different between blob storage & datalake?
* What is difference between Data lake & Datawarehouse?
* How many types of trigger/scheduling options available in ADF?
* What is ADF?
* What are key components of ADF?
* What is linked service?
* What is dataset?
* What is pipeline?
* What is activities & its example.
* Difference between lookup & adf?
* How to create dynamic pipeline in adf?
* What is incremental load?
* What is difference between incremental load & delta load?
* How to load data incrementally with adf?
* What is data flow?
* What is data flow debug?
* How to configure data factory with GIT?
* What is publish in ADF?
* Explain databricks architecture
* How databricks work in Nodes.
* Syntax for changing languages in adf.
* What is cluster
* How many types of cluster we have in ADB & what are the difference?
* What is notebook
* What is dataframe.
* How to mount datalake with databricks
* How to read & write data from datalake in databricks
* How to schedule databricks notebook
* How you can import libraries in databricks
* How to move pipeline from existing environment to higher environment within ADF?
* What is difference Between tumbling window & storage event trigger?
* How many types of integration runtimes?
* When should we use selfhosted IR?
* What is difference between selfhosted IR & SSIR?
* How to setup Selfhosted IR?
* Why we need Selfhosted IR?
* What are the various storage types available in Azure?
* What is data redundancy? What data redundancy options are available in Azure?
* What are some ways to ingest data from on-prem storage to Azure?
* What is the best way to migrate data from on-prem databases to Azure?
* How do you capture streaming data (e.g., website clickstream, social media feed etc.) in Azure?
* How to create external tables in synapse?
* What is data source & credentials in serverless pool
* Difference between serverless pool & dedicated pool.
* What is partitioning
* What is materialized views
* What are different types of pools available in synapse
* Openrowset used for?
* What is polybase & indexes?
* Databricks work on which runtime?
* What is row level security
* What is SSO & AAD?
* What is difference between managed identity & service principle.
* What is key vault
* What is ARM?
* How to connect ADF with key vault.
* What is RBAC
* What are different types of databases available in azure & when to use this?
* What us difference between azure sql & synapse.
* What is soft delete.
* What is purge
* What are difference between wrangling dataflow & mapping dataflow
* How to enable active directory authentication in sql
* What is availibity zone
* What is auditing.
* When to use azure monitor & log analytics
* What is resource group, subscription & management group.
* What is difference between data factory & serverless.
* What are the various windowing functions in Azure Stream Analytics?
* What data security options are available in Azure SQL DB?
* Can you explain the architecture of Azure Synapse Analytics?
* What are the data masking features available in Azure SQL Database?
* What are pipelines and activities in Azure Data Factory? What is the difference between the two?
* What are serverless SQL pools?
* What are DWUs?
* What do you understand by Data Engineering? What are the responsibilities of a Data Engineer?
* What is distribution & its types?
* What is difference between SQL & synapse
* What is difference between materialized views & views

SQL:

* What is a Database & RDBMS?
* What is a primary key?
* What is a unique key?
* What is a foreign key?
* What are the types of join and explain each?
* What is normalization & Denormalization?
* What is a View?
* What is an Index? & all the different types of indexes?
* What is subquery?
* What is a stored procedure?
* Difference between RANK & Dense RANK & ROWNUM
* Writing SQL queries as per given scenario.