#### **Question 1 : Assume that you are a data engineer for company ABC The company wanted to do cloud migration from their on-premises to Microsoft Azure cloud. You probably will use the Azure data factory for this purpose. You have created a pipeline that copies data of one table from on-premises to Azure cloud. What are the necessary steps you need to take to ensure this pipeline will get executed successfully?**

To move data from on-premise location to cloud, we need to have an integration runtime created because on default azure data factory will provide auto-resolve integration runtime, it cannot connect to your on-premises. Hence, we have to create self hosted integration runtime.

1. Go to azure data factory page. Navigate to manage tab and select the integration runtime.
2. Create self hosted integration runtime .

3.We have to create linked service - connection string.which creates connection to data source.

4. Create dataset for the source table

5.Create dataset for sink destination table

6.Click on create new pipeline, and add source table dataset name then sink dataset table name

7.Then we have to map both the tables.

8.Validate, if it is successful then click on publish.

#### **Question 2: Assume that you are working for a company ABC as a data engineer. You have successfully created a pipeline needed for migration. This is working fine in your development environment. how would you deploy this pipeline in production without making any or very minimal changes?**

As per question, linked service is pointing to development environment. To make it work for production environment, while creating pipeline we have to make sure that we need to send environment related information as dynamic as parameter.

**We need to create the arm template for our pipeline. ARM template needs to have a definition defined for all the constituents of the pipeline like Linked services, dataset, activities and pipeline.**

Once the ARM template is ready,  it should be checked-in into the GIT repository.  Lead or Admin will create the devops pipeline which will take up this arm template and parameter file as an input. Devopspipeline will deploy this arm template and create all the resources like linked service, [dataset](https://azurelib.com/create-dynamic-dataset-in-adf/), activities and your data pipeline into the production environment.

#### **Question 3: Assume that you have around 1 TB of data stored in Azure blob storage . This data is in multiple csv files. You are asked to do couple of transformations on this data as per business logic and needs, before moving this data into the staging container. How would you plan and architect the solution for this given scenario. Explain with the details.**

The data is huge, we should go for big data processing mechanism where we can do parallel and distributed computing.

Here we can go with Azure databricks, it is better choice because Spark is much faster than Hadoop due to in memory computation.

Next we need to create pipeline in azure data factory using azure databricks notebook as activity. We have to write all business related transformation logic into the Spark notebook. Notebook can be executed using either python, scala, java language.

When you execute the pipeline it will trigger the Azure databricks notebook and your [analytics algorithm](https://azurelib.com/azure-data-analytics/" \t "https://azurelib.com/real-time-interview-question-azure-data-factory/_blank) logic runs an do transformations as you defined into the Notebook. In the notebook itself, you can write the  logic to store the output into the blob storage Staging area.

#### **Question 4: Assume that you have an IoT device enabled on your vehicle. This device from the vehicle sends the data every hour and this is getting stored in a blob storage location in Microsoft Azure. You have to move this data from this storage location into the SQL database. How would design the solution explain with reason.**

As mentioned in the question, IOT device sends the data every hour and this is getting stored in azure blob storage which means it creates json file format for every one hour and store that file in azure blob storage. So will have many files in azure blob generated on hourly basis.

We need to create the pipeline into the azure data factory which should do the incremental load. We can use the conventional high watermark file mechanism for solving this problem.

Highwater mark design is as follows :

1. Create a file named lets say HighWaterMark.txt and stored in some place in [azure blob storage.](https://azurelib.com/create-azure-blob-storage/" \t "https://azurelib.com/real-time-interview-question-azure-data-factory/_blank)  In this file we will put the start date and time.
2. Now create the pipeline in the azure data factory.  Pipeline has the first activity defined as lookup activity. This will read the date from the HighWaterMark.txt
3. Add a one more [lookup activity](https://azurelib.com/lookup-activity-in-adf-pipeline/" \t "https://azurelib.com/real-time-interview-question-azure-data-factory/_blank) which will return the current date time.
4. Add the copy activity in the pipeline which will pull the file JSON files having created timestamp greater than High Water Mark date. In the sink push the read data into the [azure sql database](https://azurelib.com/azure-sql-database-managed-instance/" \t "https://azurelib.com/real-time-interview-question-azure-data-factory/_blank).
5. After [copy activity](https://azurelib.com/azure-data-factory-copy-activity/" \t "https://azurelib.com/real-time-interview-question-azure-data-factory/_blank) add the another copy activity which will update the current date time generated in the step 2, to the High Water Mark file.
6. Add the [trigger](https://azurelib.com/create-triggers-in-adf/) to execute this pipeline on hourly basis.

#### **Question 5: Assume that you are doing some R&D over the data about the COVID across the world. This data is available by some of the public forum which is exposed as REST api. How would you plan the solution in this scenario?**