Azure Data Factory

**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?**

We have to move the data from the on-premise location to the cloud location we need to have an Integration Runtime created.  The reason being the auto-resolve [Integration runtime](https://azurelib.com/what-is-integration-runtime/) provided by the [Azure data factory](https://azurelib.com/azure-data-factory-tutorial/) cannot connect to your on-premises.we should create our own [self-hosted integration runtime](https://azurelib.com/install-self-hosted-integration-runtime-on-azure-vm/).

1. Go to the azure data factory portal. In the manage tab select the Integration runtime.
2. Create self hosted integration runtime by simply giving general information like name description.
3. Create Azure VM (If u already have then you can skip this step)
4. Download the integration runtime software on azure virtual machine. and install it.
5. Copy the autogenerated key from step 2  and paste it newly installed integration runtime on azure vm.

Once your Integration runtime is ready we go to linked service creation. [Create the linked service](https://azurelib.com/how-to-create-linked-service-in-azure-data-factory/)which connect to the your data source and for this you use the integration runtime created above.

After this we will create the pipeline.  Your pipeline will have copy activity where source should be the database available on the on-premises location. While sink would be the database available in the cloud.

Once all of these done we execute the pipeline and this will be the one-time load as per the problem statement. This will successfully move the data from a table on on-premises database to the cloud database.

**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?**

When you create the pipeline for migration or for any other purposes like [ETL](https://azurelib.com/azure-data-factory-copy-activity/), most of the time it will use the data source. In the above mentioned scenario, we are doing the migration hence it is definitely using a data source at the source side and similarly a data source at the destination side and we need to move the data from source to destination. It is also described in the in the question itself data engineer has developed the pipeline successfully in the development environment. Hence it is safe to assume that source side data source and destination side data source both probably will be pointing to the development environment only.  Pipeline would have [copy activity](https://azurelib.com/azure-data-factory-copy-activity/) which uses the dataset with the help of linked service for source and sink.

[Linked service](https://azurelib.com/how-to-create-linked-service-in-azure-data-factory/) provides way to connect to the data source by providing the data source details like the server address, port number, username, password, key, or other credential related information.

Hence we need to design our [Azure data factory pipeline](https://azurelib.com/lookup-activity-in-adf-pipeline/) components in such a way that we can provide the environment related information dynamic and as a part of a parameter. There should be no hard coding of these kind of information.

**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.**

First of all, we need to analyze the situation. Here if you closely look at the size of the data, you find that it is very huge in the size. Hence directly doing the transformation on such a huge size of  data could be very cumbersome and time consuming process. Hence we should think about the big data processing mechanism where we can leverage the parallel and[distributed computing advantages](https://azurelib.com/advantages-and-disadvantages-of-cloud-computing/).. Here we have two choices.

1. We can use the [Hadoop MapReduce through HDInsight](https://azurelib.com/azure-databricks-spark-tutorial-for-beginner-to-advance-level-lesson-1/" \t "_blank)capability for doing the transformation.
2. We can also think of using the spark through the Azure databricks for doing the transformation on such a huge scale of data.

Out of these two, Spark on[Azure databricks](https://azurelib.com/azure-databricks-spark-tutorial-azure-databricks-notebook-overview/) is better choice because Spark is much faster than Hadoop due to in memory computation. So let’s choose the Azure databricks as the option.

Next we need to create the pipeline in Azure data factory. A pipeline should use the databricks notebook as an activity.

We can write all the business related transformation logic into the Spark notebook. Notebook can be executed using either python, scala or java language.

When you execute the pipeline it will trigger the Azure databricks notebook and your [analytics algorithm](https://azurelib.com/azure-data-analytics/) 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.**

This looks like an a typical incremental load scenario. As described in the problem statement, IoT device write the data to the location every hour. It is  most likely that this device is sending the JSON data to the cloud storage (as most of the IoT device generate the data in JSON format). It will probably writing the new JSON file every time whenever the data from the device sent to the cloud.

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/)  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/) 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/).
5. After [copy activity](https://azurelib.com/azure-data-factory-copy-activity/) 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.