Scenario Based Interview Question

Topic: Incremental Load

What is Incremental Load?

It's a data loading technique in ETL where only updated data is loaded into destination rather than the complete data.

Other ways of loading data.

Full Load: When the complete data is loaded then it's called full load.

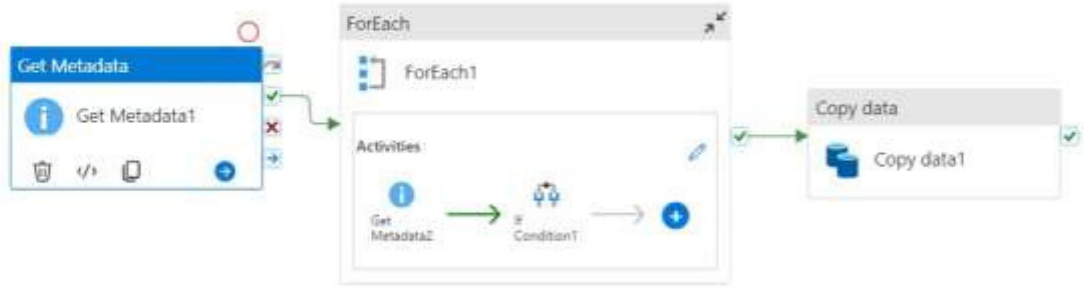
Con's of Full Load:

>> Memory issues

>> Cost is high

>> Time taken is more if data size is big.

There are various ways to implement incremental data in Azure Data Factory. I have used the below approach.

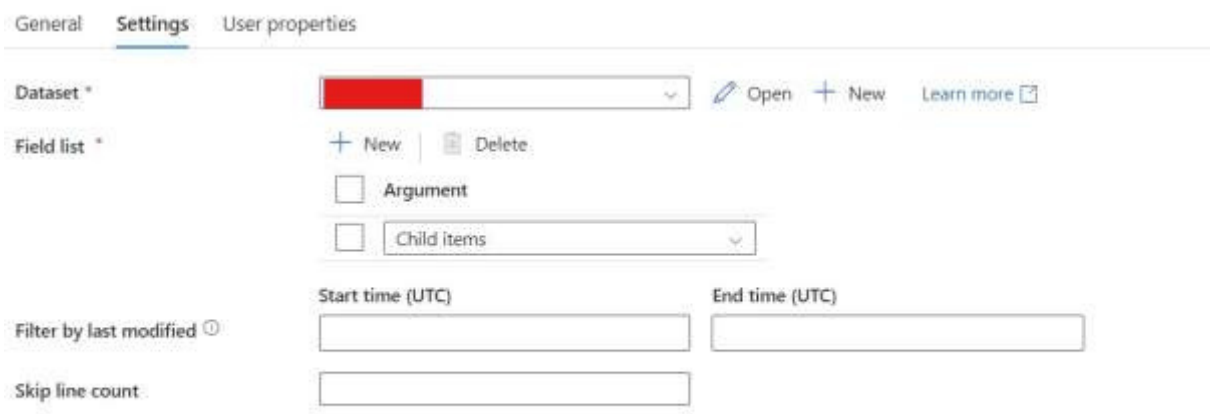


Source taken here is SFTP connection to an application(FileZilla) and destination is a Blob storage.

Internet is filled with incremental load examples for SQL server so I wanted to have a different approach and give it a twist.

Step 1: Make a successful a SFTP connection.

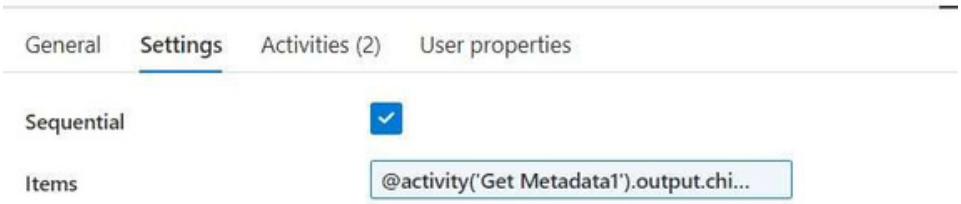
Step 2: Use a Get Metadata activity to retrieve all the files by using child items as argument.



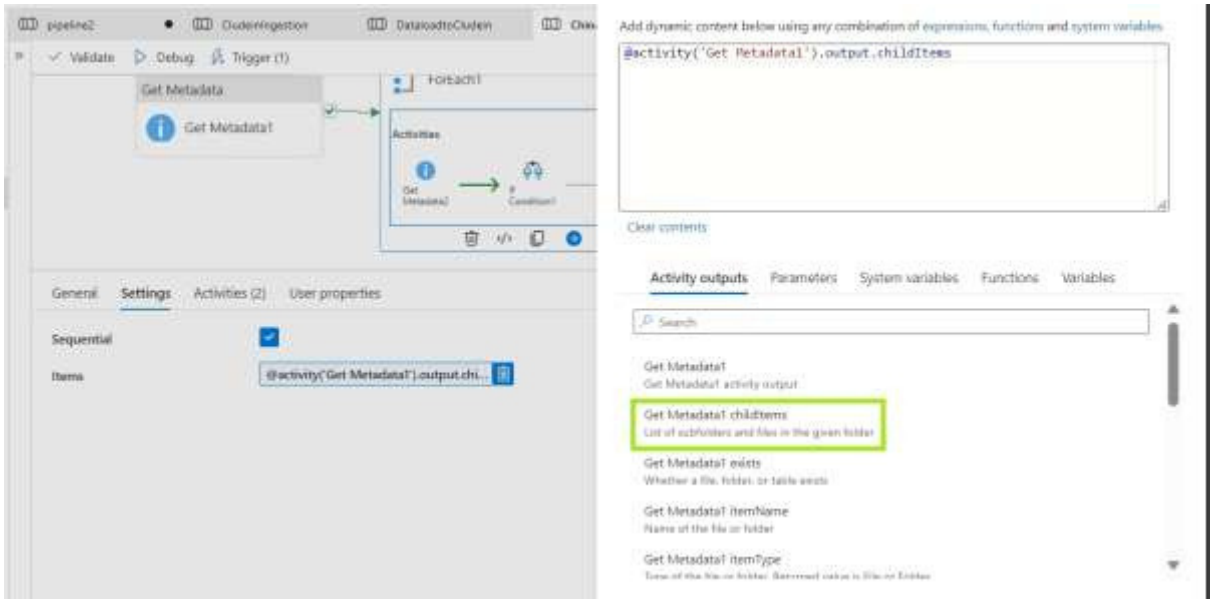
Step 3: Use a ForEach activity for iterating through all files.

I have checked Sequential option; we also have another option of using Batch.

Difference: Sequential means all the files are iterated one by one but in Batch all the files are processed at once, you need to give batch count if you use this option, the number of batch count is the number of files that would get processed all at once.

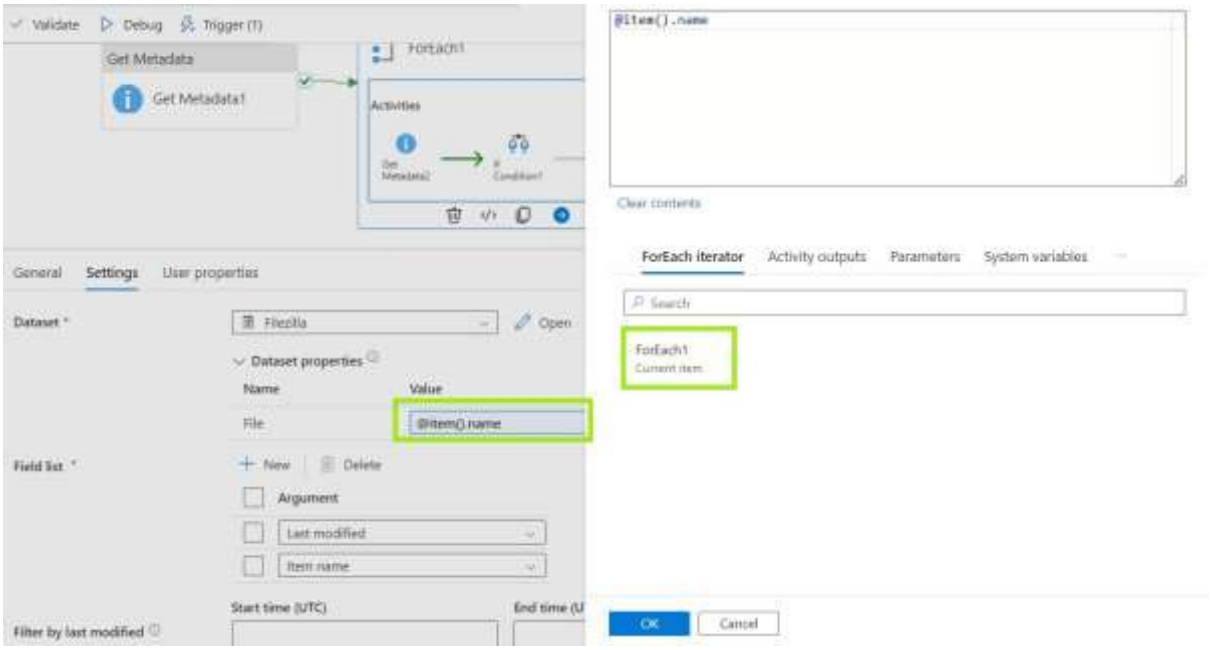


Items: Here we need to send the files that our Get Metad
ata have processed. This value must be dynamic.

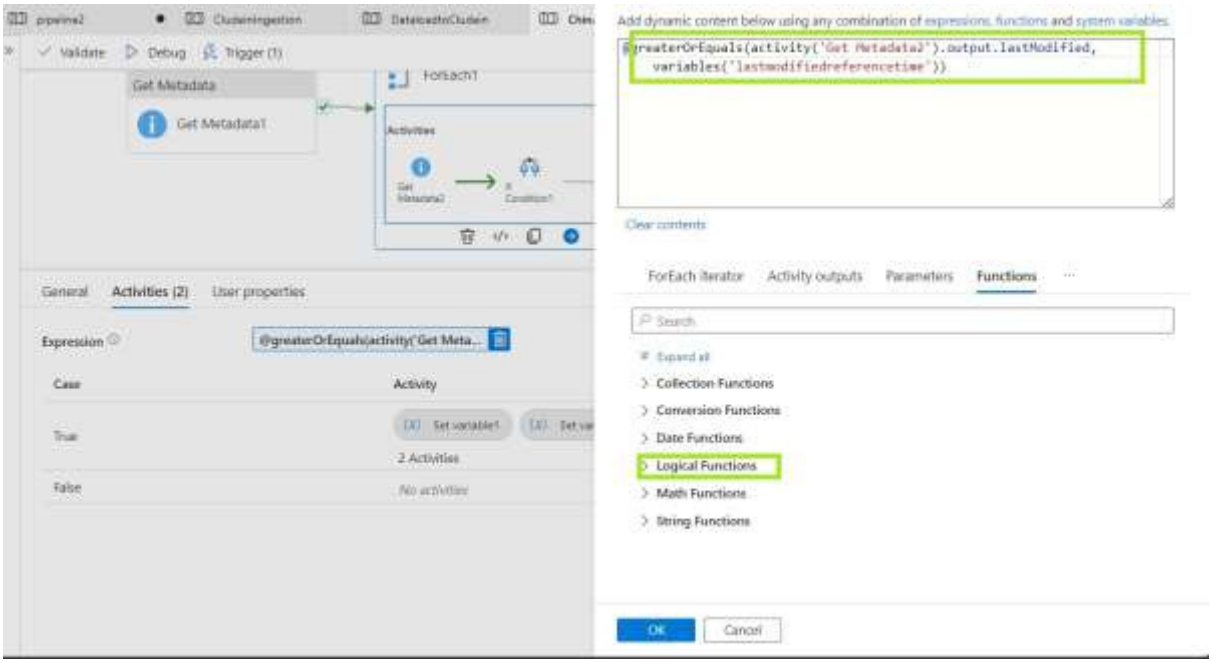


The above option must be chosen.

Step 4: We must choose activities within ForEach, here I choose another Get Metadata for this the dataset must be parameterized and pass the expression of ForEach iterator. This Get Metadata will give us the file name along with their last modified date which will be used to compare and give the latest file.

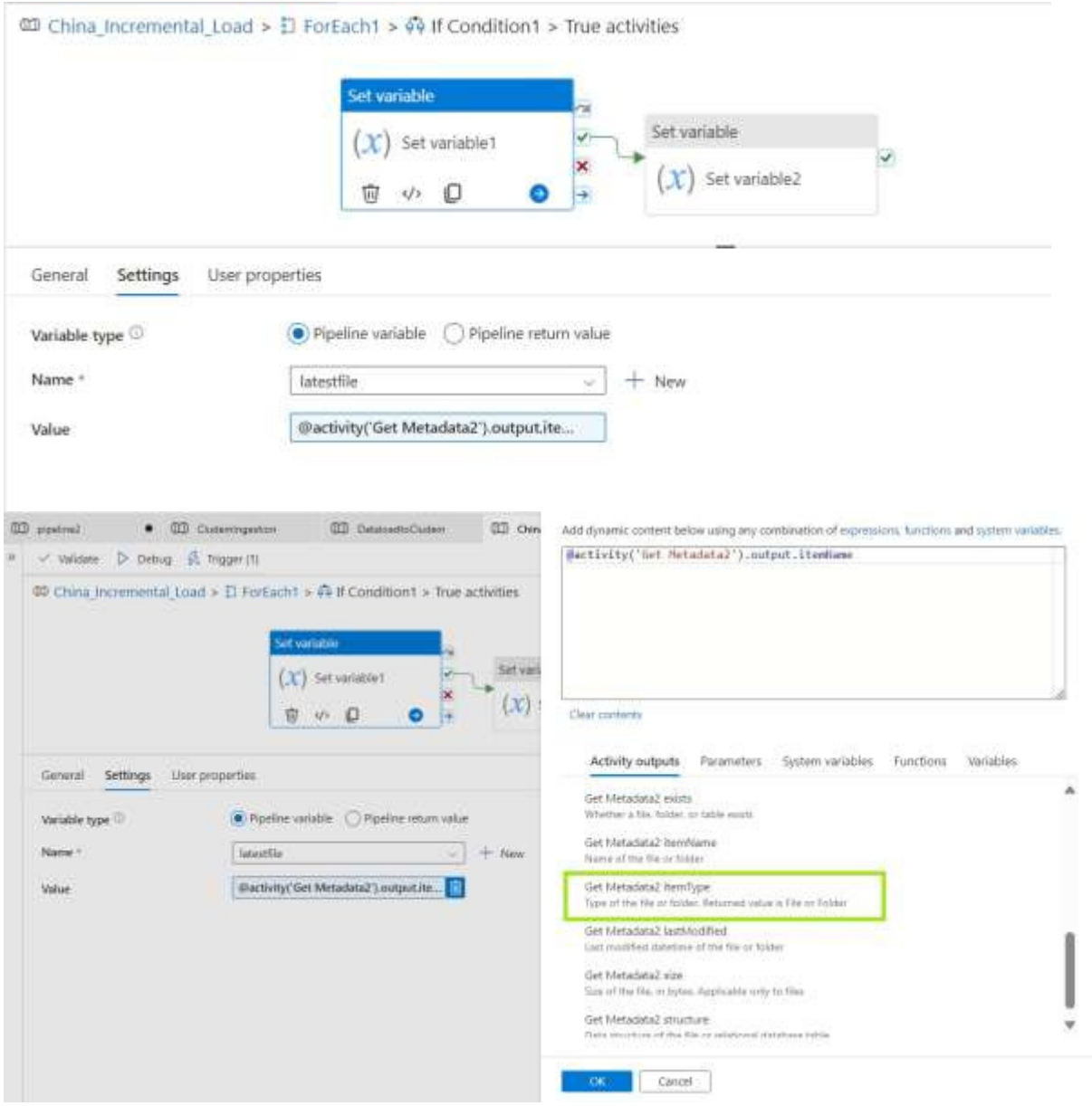


Step 5: We further use If condition activity where we would compare the timestamps of files from logical function I took greaterOrEquals function.



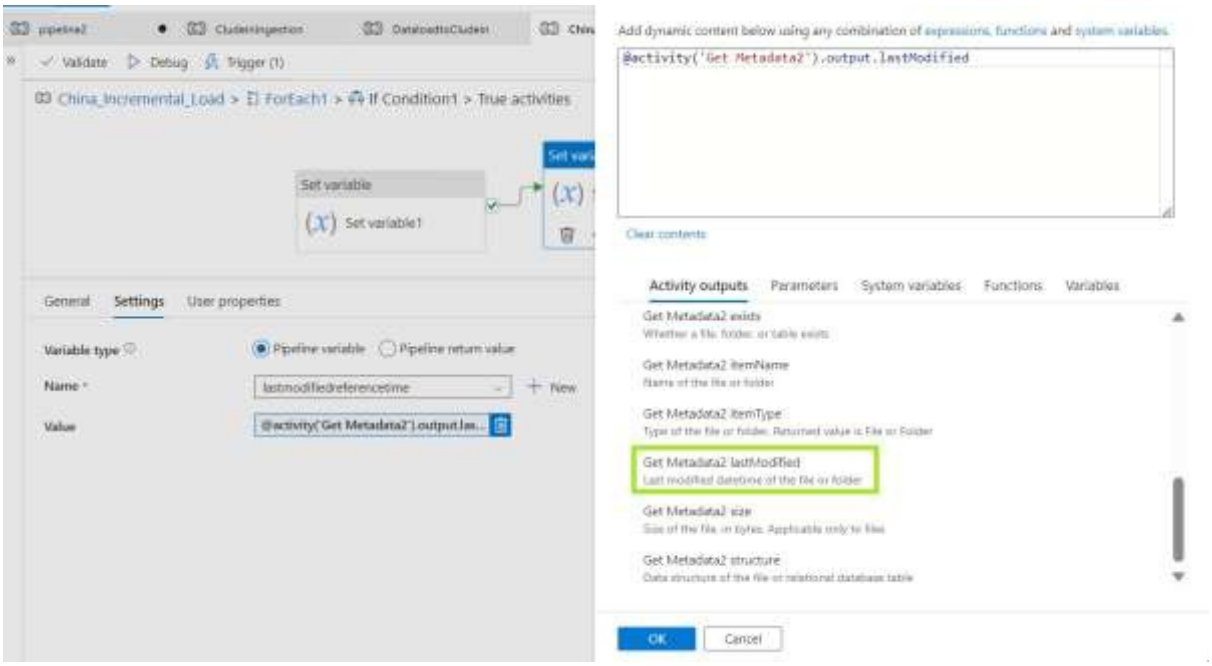
Step 6: We will use set variables to store the last modified and item name. But before that variables at pipeline level must be defined.

Set Variable1

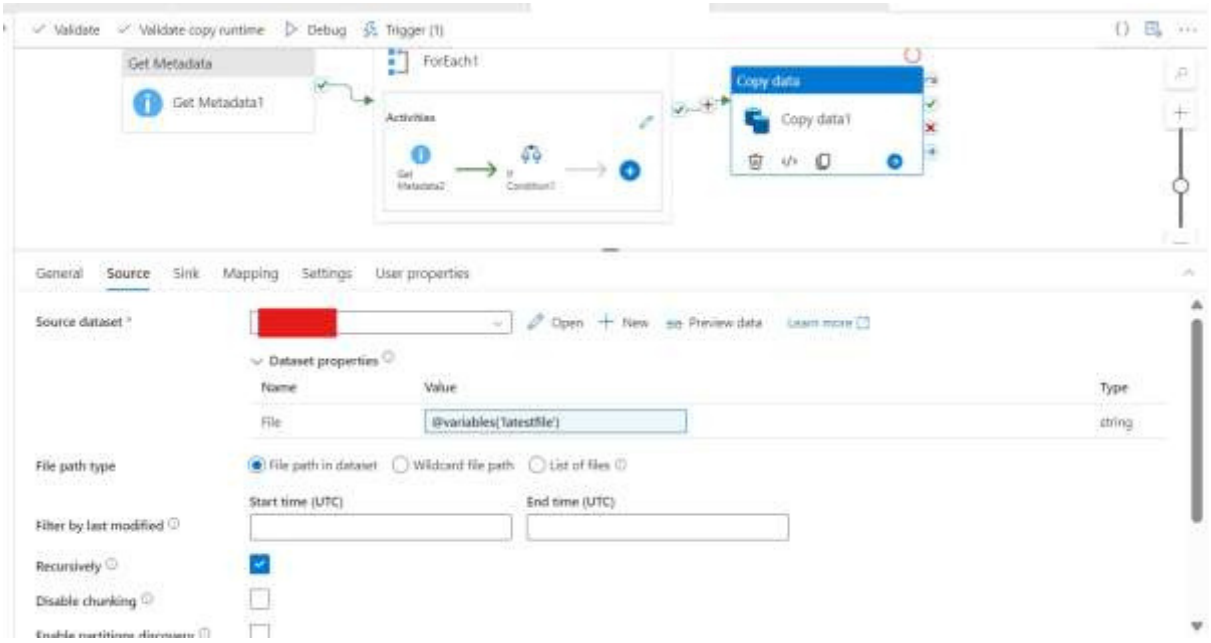


Set Variable2

Set variable2 is used for Last modified date.



Step 7: Drag a copy activity to the canvas. Set the source to the latest file variable and sink to the destination as per your requirement.



Step 8: You can Debug the pipeline before publishing.