**DATA INTEGRATION PIPELINES FOR NYC PAYROLL DATA ANALYTICS**

**Project Overview**

**Project Introduction**

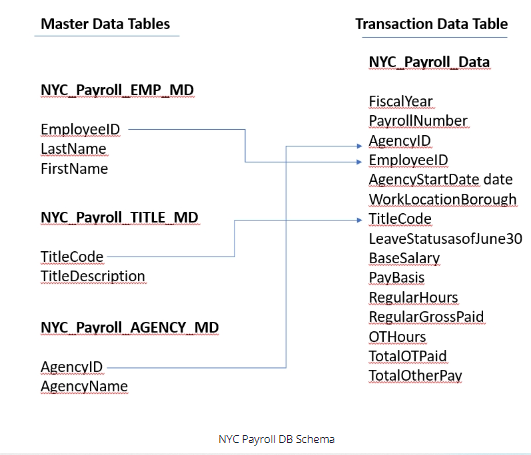
The City of New York would like to develop a Data Analytics platform on Azure Synapse Analytics to accomplish two primary objectives:

1. Analyze how the City's financial resources are allocated and how much of the City's budget is being devoted to overtime.
2. Make the data available to the interested public to show how the City’s budget is being spent on salary and overtime pay for all municipal employees.

You have been hired as a Data Engineer to create high-quality data pipelines that are dynamic, can be automated, and monitored for efficient operation. The project team also includes the city’s quality assurance experts who will test the pipelines to find any errors and improve overall data quality.

The source data resides in Azure Data Lake and needs to be processed in a NYC data warehouse in Azure Synapse Analytics. The source datasets consist of CSV files with Employee master data and monthly payroll data entered by various City agencies.

NYC Payroll DB Schema



Project Environment

For this project, you'll do your work in the Azure Portal, using several Azure resources including:

* Azure Data Lake Gen2 (Storage account with Hierarchical Namespaces checkbox checked when creating)
* Azure SQL DB
* Azure Data Factory
* Azure Synapse Analytics

Instructions for using a temporary Azure account to complete the project are on the next page.

You'll take screenshots as proof of work for this project, so remember to collect these screenshots throughout the project steps. A checklist is provided at the end of each step so you can double-check you've collected all of the deliverables.

You'll also need to create a Github repository for this project. At the end of the project, you will connect your Azure pipelines to Github and submit the URL or contents of the repository.

**Project Data**

[Download these .csv files](https://video.udacity-data.com/topher/2022/May/6283aff5_data-nyc-payroll/data-nyc-payroll.zip) that **provide the data for the project**.

**Project Environment**

**Local Machine Instructions**

**Workspace Instructions**

**Create and Configure Resources**

**Project Instructions**

For this project, you'll do your work in the Azure Portal, using several Azure resources including:

* Azure Data Lake Gen2
* Azure SQL DB
* Azure Data Factory
* Azure Synapse Analytics

Instructions for using a temporary Azure account to complete the project are on the previous page.

You'll take screenshots as proof of work for this project, so remember to collect these screenshots throughout the project steps. A checklist is provided at the end of each step so you can double-check you've collected all of the screenshot deliverables.

You'll also need to create a Github repository for this project. At the end of the project, you will connect your Azure pipelines to Github and submit the URL or contents of the repository.

When you submit your project, it will be assessed against this [project rubric](https://review.udacity.com/#!/rubrics/4857/view). Take a moment to review it now and periodically throughout the project to make sure you're meeting the requirements.

**Project Data**

[Download these .csv files](https://video.udacity-data.com/topher/2022/May/6283aff5_data-nyc-payroll/data-nyc-payroll.zip) that **provide the data for the project**.

**Step 1: Prepare the Data Infrastructure**

Setup Data and Resources in Azure

**1.Create the data lake and upload data**

Log into your temporary Azure account (instructions on the previous page) and create the following resources. Please use the provided resource group to create each resource. You will use these resources for the whole project, in all of the steps, so you'll only need to create one of each:

Create an Azure Data Lake Storage Gen2 (storage account) and associated storage container resource named **adlsnycpayroll-yourfirstname-lastintial**. Create three directories in this storage container named

* **dirpayrollfiles**
* **dirhistoryfiles**
* **dirstaging**

Upload these files from the [project data](https://video.udacity-data.com/topher/2022/May/6283aff5_data-nyc-payroll/data-nyc-payroll.zip) to the **dirpayrollfiles** folder

* EmpMaster.csv
* AgencyMaster.csv
* TitleMaster.csv
* nycpayroll\_2021.csv

Upload this file (historical data) from the [project data](https://video.udacity-data.com/topher/2022/May/6283aff5_data-nyc-payroll/data-nyc-payroll.zip) to the **dirhistoryfiles** folder

* nycpayroll\_2020.csv

**2. Create an Azure Data Factory Resource**

**3. Create a SQL Database to store the current year of the payroll data**

In the Azure portal, create a SQL Database resource named **db\_nycpayroll**

* Add client IP address to the SQL DB firewall

Create a table called **NYC\_Payroll\_Data** in **db\_nycpayroll** in the Azure Query Editor with this SQL Script**:**

*CREATE TABLE [dbo].[NYC\_Payroll\_Data](*

*[FiscalYear] [int] NULL,*

*[PayrollNumber] [int] NULL,*

*[AgencyID] [varchar](10) NULL,*

*[AgencyName] [varchar](50) NULL,*

*[EmployeeID] [varchar](10) NULL,*

*[LastName] [varchar](20) NULL,*

*[FirstName] [varchar](20) NULL,*

*[AgencyStartDate] [date] NULL,*

*[WorkLocationBorough] [varchar](50) NULL,*

*[TitleCode] [varchar](10) NULL,*

*[TitleDescription] [varchar](100) NULL,*

*[LeaveStatusasofJune30] [varchar](50) NULL,*

*[BaseSalary] [float] NULL,*

*[PayBasis] [varchar](50) NULL,*

*[RegularHours] [float] NULL,*

*[RegularGrossPaid] [float] NULL,*

*[OTHours] [float] NULL,*

*[TotalOTPaid] [float] NULL,*

*[TotalOtherPay] [float] NULL*

*)*

*GO*

**4. Create A Synapse Analytics workspace, or use one you already have created.**

* You are only allowed one Synapse Analytics workspace per Azure account, a Microsoft restriction.
* Create a new Azure Data Lake Gen2 and file system for Synapse Analytics when you are creating the Synapse Analytics workspace in the Azure portal.

Create a SQL dedicated pool in the Synapse Analytics workspace

* Select DW100c as performance level. Keep defaults for other settings.

In the SQL dedicated pool, Create master data tables and payroll transaction tables using these SQL scripts (You can execute all four sql scripts at same time) :

Create Emplyee Master Data table:

*CREATE TABLE [dbo].[NYC\_Payroll\_EMP\_MD](*

*[EmployeeID] [varchar](10) NULL,*

*[LastName] [varchar](20) NULL,*

*[FirstName] [varchar](20) NULL*

*)*

*GO*

Create Job Title Table:

*CREATE TABLE [dbo].[NYC\_Payroll\_TITLE\_MD](*

*[TitleCode] [varchar](10) NULL,*

*[TitleDescription] [varchar](100) NULL*

*)*

*GO*

Create Agency Master table:

*CREATE TABLE [dbo].[NYC\_Payroll\_AGENCY\_MD](*

*[AgencyID] [varchar](10) NULL,*

*[AgencyName] [varchar](50) NULL*

*)*

*GO*

Create Payroll transaction data table:

*CREATE TABLE [dbo].[NYC\_Payroll\_Data](*

*[FiscalYear] [int] NULL,*

*[PayrollNumber] [int] NULL,*

*[AgencyID] [varchar](10) NULL,*

*[AgencyName] [varchar](50) NULL,*

*[EmployeeID] [varchar](10) NULL,*

*[LastName] [varchar](20) NULL,*

*[FirstName] [varchar](20) NULL,*

*[AgencyStartDate] [date] NULL,*

*[WorkLocationBorough] [varchar](50) NULL,*

*[TitleCode] [varchar](10) NULL,*

*[TitleDescription] [varchar](100) NULL,*

*[LeaveStatusasofJune30] [varchar](50) NULL,*

*[BaseSalary] [float] NULL,*

*[PayBasis] [varchar](50) NULL,*

*[RegularHours] [float] NULL,*

*[RegularGrossPaid] [float] NULL,*

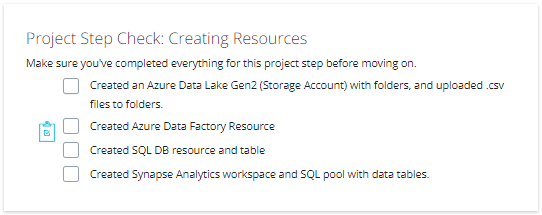
*[OTHours] [float] NULL,*

*[TotalOTPaid] [float] NULL,*

*[TotalOtherPay] [float] NULL*

*)*

*GO*



**Create Linked Services**

**1.Create a Linked Service for Azure Data Lake**

In Azure Data Factory, create a linked service to the data lake that contains the data files

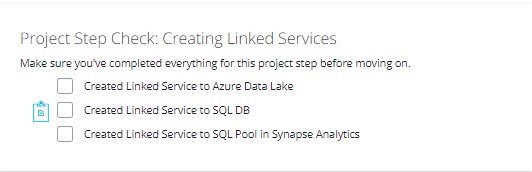
* From the data stores, select Azure Data Lake Gen 2
* Test the connection

**2.Create a Linked Service to SQL Database that has the current (2021) data**

* If you get a connection error, remember to add the IP address to the firewall settings in SQL DB in the Azure Portal

**3. Create a Linked Service for Synapse Analytics**

* Create the linked service to the SQL pool.



**Create Datasets**

**Step 3: Create Datasets in Azure Data Factory**

**1.Create the datasets for the 2021 Payroll file on Azure Data Lake Gen2**

* Select DelimitedText
* Set the path to the nycpayroll\_2021.csv in the Data Lake
* Preview the data to make sure it is correctly parsed

**2. Repeat the same process to create datasets for the rest of the data files in the Data Lake**

* EmpMaster.csv
* TitleMaster.csv
* AgencyMaster.csv
* Remember to publish all the datasets

**3. Create the dataset for transaction data table that should contain current (2021) data in SQL DB**

**4. Create the datasets for destination (target) tables in Synapse Analytics**

* dataset for NYC\_Payroll\_EMP\_MD
* for NYC\_Payroll\_TITLE\_MD
* for NYC\_Payroll\_AGENCY\_MD
* for NYC\_Payroll\_Data

Graphical user interface, text

Description automatically generated

**Create Dataflows and Pipelines**

**Step 4: Create Data Flows**

**1.In Azure Data Factory, create the data flow to load 2021 Payroll Data to SQL DB transaction table (in the future NYC will load all the transaction data into this table).**

* Create a new data flow
* Select the dataset for the 2021 payroll file as the source
* Select the sink dataset as the payroll table on SQL DB
* Make sure to reassign any missing source to target mappings

**2.Create Pipeline to load 2021 Payroll data into transaction table in the SQL DB**

* Create a new pipeline
* Select the data flow to load the 2021 file into SQLDB
* Trigger the pipeline
* Monitor the pipeline
* Take a screenshot of the Azure Data Factory screen pipeline run after it has finished.
* Make sure the data is successfully loaded into the SQL DB table

**3. Create data flows to load the data from the data lake files into the Synapse Analytics data tables**

* Create the data flows for loading Employee, Title, and Agency files into corresponding SQL pool tables on Synapse Analytics
* For each Employee, Title, and Agency file data flow, sink the data into each target Synaspe table

**4. Create a data flow to load 2021 data from SQL DB to Synapse Analytics**

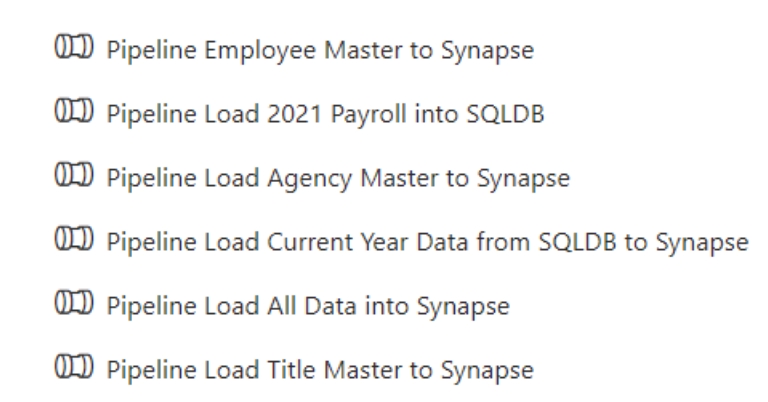
**5. Create pipelines for Employee, Title, Agency, and year 2021 Payroll transaction data to Synapse Analytics containing the data flows.**

* Select the dirstaging folder in the data lake storage for staging
* Optionally you can also create one master pipeline to invoke all the Data Flows
* Validate and publish the pipelines

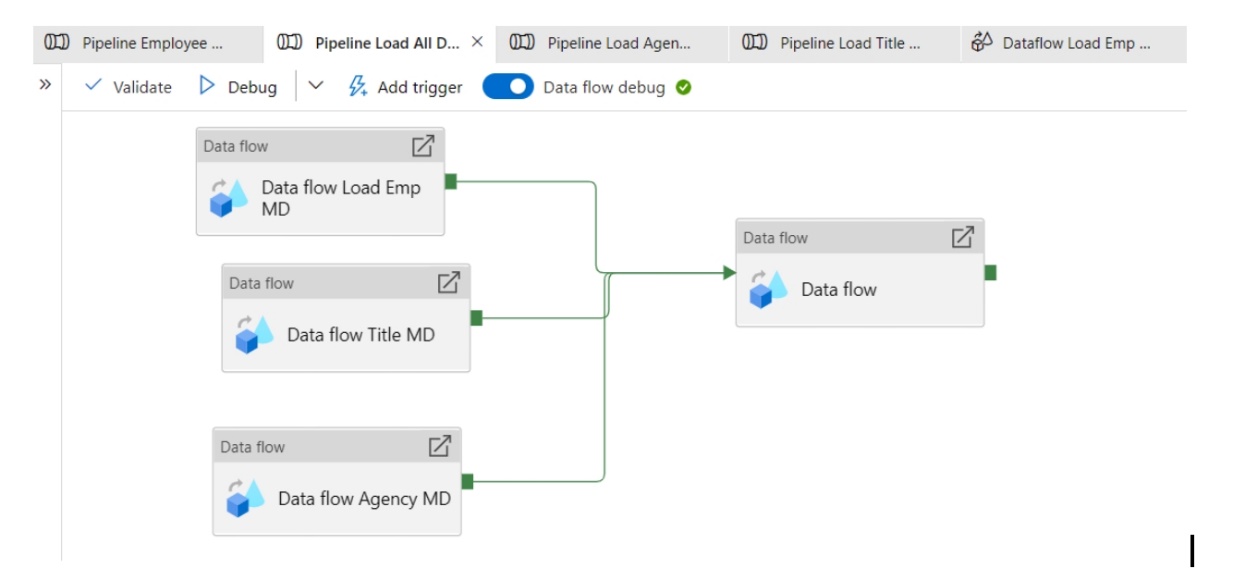
**6. Trigger and monitor the Pipelines**

* Take a screenshot of each pipeline run after it has finished, or one after your master pipeline run has finished.

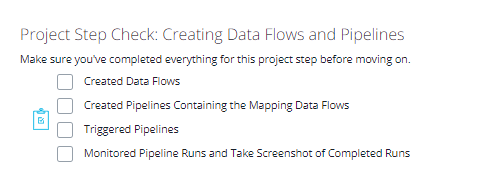
In total, you should have 6 pipelines or one master pipeline



Pipelines created in this step



Example master pipeline



**Aggregate Data Flow**

**Step 5: Data Aggregation and Parameterization**

In this step, you'll extract the 2021 year data and historical data, merge, aggregate and store it in Synapse Analytics. The aggregation will be on Agency Name, Fiscal Year and TotalPaid.

**1.Create a Summary table in Synapse with the following SQL script and create a dataset named table\_synapse\_nycpayroll\_summary**

*CREATE TABLE [dbo].[NYC\_Payroll\_Summary](*

*[FiscalYear] [int] NULL,*

*[AgencyName] [varchar](50) NULL,*

*[TotalPaid] [float] NULL*

*)*

**2.Create a new dataset for the Azure Data Lake Gen2 folder that contains the historical files.**

* Select dirhistoryfiles in the data lake as the source

**3.Create new data flow and name it Dataflow Aggregate Data**

* Create a data flow level parameter for Fiscal Year
* Add first Source for table\_sqldb\_nyc\_payroll\_data table
* Add second Source for the Azure Data Lake history folder

**4.Create a new Union activity in the data flow and Union with history files**

**5.Add a Filter activity after Union**

* In Expression Builder, enter toInteger(FiscalYear) >= $dataflow\_param\_fiscalyear

**6.Derive a new TotalPaid column**

* In Expression Builder, enter RegularGrossPaid + TotalOTPaid+TotalOtherPay

**7.Add an Aggregate activity to the data flow next to the TotalPaid activity**

* Under Group By, Select AgencyName and Fiscal Year

**8.Add a Sink activity to the Data Flow**

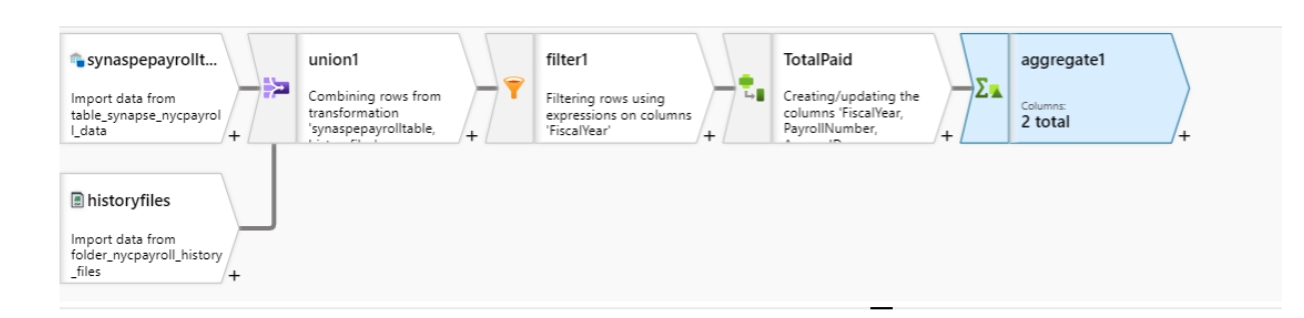
* Select the dataset to target (sink) the data into the Synapse Analytics Payroll Summary table.
* In Settings, select Truncate Table

**9.Create a new Pipeline and add the Aggregate data flow**

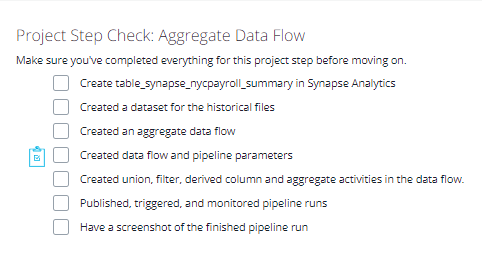
* Create a new Global Parameter (This will be the Parameter at the global pipeline level that will be passed on to the data flow
* In Parameters, select Pipeline Expression
* Choose the parameter created at the Pipeline level

**10.Validate, Publish and Trigger the pipeline. Enter the desired value for the parameter.**

**11.Monitor the Pipeline run and take a screenshot of the finished pipeline run.**



The aggregate data flow activities



Connect Project to Github

In this step, connect Azure Data Factory to Github

* Login to your Github account and create a new Repo in Github
* Connect Azure Data Factory to Github
* Select your Github repository in Azure Data Factory
* Publish all objects to the repository in Azure Data Factory

Download your project repository from Github and **submit your repository and screenshots of your pipeline runs**.

Graphical user interface, text, application

Description automatically generated