Open Education Analytics

Module Creation Kit

Version 1.1

Date Published: January 2022

[1 Introduction 2](#_Toc85669480)

[1.1 What is an OEA module? 3](#_Toc85669481)

[1.2 Why are OEA modules important? 3](#_Toc85669482)

[2 Pre-requisites for OEA module creation 3](#_Toc85669483)

[3 Setting up the OEA Module Creation template on GitHub 4](#_Toc85669484)

[4 Assets in a standard OEA module 5](#_Toc85669485)

[4.1 Pipeline 6](#_Toc85669486)

[4.2 Synapse Notebook 6](#_Toc85669487)

[4.2.1 Additional information about pseudonymization process 6](#_Toc85669488)

[*4.3* PowerBI template 7](#_Toc85669489)

[4.4 Test data 7](#_Toc85669490)

[4.5 Privacy: Scripts for pseudonymization​ and role-based access control 7](#_Toc85669491)

[4.6 Documentation 8](#_Toc85669492)

[4.7 Readme 8](#_Toc85669493)

[4.8 Docs folder 8](#_Toc85669494)

[5 Steps for creating a new OEA module 9](#_Toc85669495)

[6 Module Quality 11](#_Toc85669496)

[6.1 Factors to Consider 11](#_Toc85669497)

[6.2 Quality Rubric and Criteria 12](#_Toc85669498)

[7 Contributing to OEA 14](#_Toc85669499)

[8 Additional Resources 14](#_Toc85669500)

# Introduction

Open Education Analytics (OEA) is an open source community coordinated by Microsoft Education. We collaborate with education systems and analytics organizations across the world to develop modern data intelligence capabilities.

The OEA community is building shared technology resources for implementing the modern data estate in education, based on Azure Synapse Analytics. OEA GitHub contains a set of assets for education systems to implement the OEA reference architecture and start using their data more quickly and effectively.

## What is an OEA module?

In the OEA framework, modules are a set of assets (e.g., notebooks, PowerBI templates, pipelines, test data, etc) for moving a single data source into Azure, preparing it for exploration, and visualizing that data source. An example of an OEA module is the [Microsoft Graph Reports API module](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/Microsoft_Data/Microsoft_Graph) which provides a gateway to data and intelligence in Microsoft 365. This Graph Reports data is freely available to education customers using Office 365 and includes app usage data (Teams, OneDrive, Outlook, etc). This module leverages the OEA reference architecture to enable education systems to bring in the Microsoft Graph data to their own Azure data lake, to combine it with other datasets for analysis (e.g., Learning Management Systems (LMS), Student Information Systems (SIS), Assessment, or other digital learning apps).

## Why are OEA modules important?

Modules are the building blocks that create layers of data on the OEA architecture. Modules form the foundation for OEA packages, which combine multiple data sources for a use case or specific education data scenario. The more modules the OEA Community develops, the more education data sources can easily be ingested into each organization’s modern data estate. This will save valuable time and resources for both education systems and analytics partners. OEA modules eliminate the need for each team to build education data source pipelines ‘from scratch’ and speed up the time to value for analytics investments.

This OEA Module Creation Kit is designed to make the development of high quality OEA Modules easier by defining guidelines and standards so every module will have a consistent set of assets.

# Pre-requisites for OEA module creation

Before creating a new module in OEA, it is important to ensure that all these pre-requisites have been met:

1. Have an active Azure subscription. Check out the [list of Azure offers](https://azure.microsoft.com/en-us/support/legal/offer-details/) or sign up for a [free subscription](https://azure.microsoft.com/en-us/free/).
2. Setup the OEA reference architecture in Azure Synapse Analytics using the steps outlined in the [OEA GitHub repository.](https://github.com/microsoft/OpenEduAnalytics)
3. After the deployment of the OEA architecture, ensure that all storage containers in the Azure Data Lake in Synapse have been provisioned. When done correctly, you should have the following for v0.5 of the OEA architecture: stage1np, stage2np, stage2p, stage3np, stage3p, oea-framework and synapse-workspace.
4. Identify which data entities in the data source are sensitive or include personal data that will require pseudonymization or anonymization.

# Setting up the OEA Module Creation template on GitHub

In the OEA GitHub repository, you will find the folders and files to setup the OEA reference architecture and build modules on top of it. The [modules folder](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/M365) houses existing OEA modules and contains a [Creation Kit folder](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/_Module_Creation_Kit), which has the recommended templates for the OEA module components to get you started. If you already have the OEA GitHub repository cloned but cannot see the Creation Kit folder under modules, you must ‘Fetch upstream’ or ‘Pull’ the most recent version of the repository.

If you do not have the OEA GitHub repository cloned, you can do that through the command line interface, ZIP download, VS Studio Code or GitHub Desktop.

This is the process to clone the repository using GitHub desktop:

1. Go to <https://github.com/microsoft/OpenEduAnalytics>.
2. Click on the Code button and select **Open with GitHub Desktop**.

Graphical user interface, text, application, email

Description automatically generated

1. When GitHub Desktop launches, ensure that you are in the **main** branch and all recent updates have been pulled by clicking the **Fetch origin** button.
2. Right click the repository name from the list of repositories on GitHub Desktop and select **Show in Explorer.**

Graphical user interface, application

Description automatically generated

1. Navigate to modules and then the [Module Creation Kit folder](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/_Module_Creation_Kit) to start creating your new module. Remember to save all your updates.

# Assets in a standard OEA module

Get familiar with the main assets found in a standard OEA module and why they are important. In the Steps for creating a new OEA module section, we provide a list of steps to follow when creating a new OEA module.

* If this is your first time of creating an OEA module and you don’t have any experience working with Synapse, you should first complete the recommended [MS Learning Path for Synapse Analytics](https://docs.microsoft.com/en-us/learn/paths/realize-integrated-analytical-solutions-with-azure-synapse-analytics/). With a basic understanding of how Synapse, OEA and PowerBI works, it may take you at least 80 hours to create a working first version of the new module after completing some of the preliminary Synapse, OEA and PowerBI trainings.
* If you have intermediate or advanced skills in Synapse, OEA architecture and PowerBI, it may take you about 40 hours to have a working first version of a new module, depending on the complexity of the data source.

This is the list of assets and where to upload them in the folder of the new module you are creating:

|  |  |
| --- | --- |
| **Asset** | **Folder to upload to** |
| Test data sets (in CSV, JSON or any other appropriate format) | test\_data |
| Pseudonymization notebook (in. ipynb format) | notebook |
| Preprocessing notebook (in .ipynb format) | notebook |
| PowerBI template, with all pages and visuals properly labelled and with tooltips (in .pbix format) | powerbi |
| Pipeline for ingesting data, triggering periodic activities, etc (in JSON format) | pipeline |
| Documentation for each asset type (in .md format) | readme file associated with each asset’s folder as well as the general readme |
| * Any additional documentation like tutorials and explanations * All images | docs |

## Pipeline

The module should include pipelines for ingesting data from its original source into the Synapse data lake. It is recommended that pipelines are uploaded in JSON format so that module users can easily import the pipeline template into their Synapse environment. Make sure to include any permissions required for the pipeline to work as expected.

Estimated time to create pipeline(s): 5 hours

Link to pipeline template: [*https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/\_Module\_Creation\_Kit/pipeline*](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/_Module_Creation_Kit/pipeline)

## Notebook

A notebook like the [preprocessing notebook](https://github.com/microsoft/OpenEduAnalytics/blob/main/modules/_Module_Creation_Kit/notebook/preprocessing_notebook.ipynb) will be used for aggregating, transforming, enriching and exploring data from the data source to make it ready for reporting or modelling. We strongly encourage the use of PySpark in the notebooks; however, it is okay to use other programming languages.

The steps for processing data in the notebooks will include reading data from the data lake *(N.B: read the raw data from stage1np)*, preprocessing and aggregating the data, and then writing back the data into the lake *(N.B: write back to stage 2p after the data has been pseudonymized).*

For each notebook, list out the databases created, tables created, table purpose, databases used, and tables used.

Estimated time to create notebook(s): 11 hours

Link to notebook template: [*https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/\_Module\_Creation\_Kit/notebook*](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/_Module_Creation_Kit/notebook)

### Additional information about pseudonymization process

The process of ingesting data into the lake and iteratively preparing, cleaning, and optimizing that data is represented in 3 stages that map to Delta Lake's proposed stages of bronze, silver, and gold. Bronze represents the stage when the raw data is ingested in stage1np (np means non-pseudonymized while p means pseudonymized). Silver represents the stage where any columns with personal identifiable information (PII) in the raw data is hashed or masked. Fields like email addresses and identification numbers are usually hashed while fields like names are masked. Gold represents the stage where all the processing and aggregation work has been finalized and the data is now report ready or model ready.

## PowerBI template

It is important for education customers to be able to quickly visualize and explore the ‘ready’ data from the module’s data source in PowerBI. Modules should aim to make the time from data ingestion to visualization rapid. The PowerBI template created as part of the module creation should include visuals that combine and express the different data represented in the data source to better understand the data. It should also include the PowerBI semantic model to explore the relationships between entities. PowerBI visuals should be well labeled and each visual should have a tooltip that provides a brief description. As shown in the PowerBI template example on GitHub, please add a description of the module in the Explanation page.

Estimated time to create PowerBI template: 6 hours

Link to PowerBI template: [*https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/\_Module\_Creation\_Kit/powerbi*](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/_Module_Creation_Kit/powerbi)

N.B: Providing a PowerBI template is preferred but not required.

## Test data

Including a simulated test data set in the module makes it easier for education systems to bring the module to life in their non-production environments. This helps stakeholders get a first-hand experience of how the module works without having an impact on their live environments. We provide some sample data sets and a standard M365 data generator as part of this kit. However, you are free to generate your own test data (including a data dictionary) that will suit the new module you are creating. When generating new test data, it is important to make sure that it aligns with existing data or new data you plan to create so it becomes easier to join the tables and create relationships for PowerBI visualization. For example, using the same students and same schools.

Estimated time to generate test data: 12 hours

Link to test data template: [*https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/\_Module\_Creation\_Kit/test\_data*](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/_Module_Creation_Kit/test_data)

## Privacy: Scripts for pseudonymization​ and role-based access control

All modules created must show how to ensure data is used responsibly and ethically including protecting students’ privacy in accordance with GDPR and CCPA Compliance. To protect students’ identity, it is required that the personal identifiable information of students like names, email addresses, etc., are pseudonymized using the pseudonymization scripts provided as part of the OEA framework. As an OEA contributor, you share in OEA’s commitment to ensuring that student data is always protected. You should also ensure that role-based access control is enforced in the Synapse environment through Azure Active Directory.

The pseudonymization operations provided as part of the OEA deployment include:

* **hash-no-lookup or hnl:** This means that the lookup can be performed against a different table, so no lookup is needed.
* **hash or h:** This will hash the column and create a lookup table as well.
* **mask or m:** This will mask the column and will not create a lookup table.
* **no-op or x:** No operation will be performed so the column will remain as it is.

Estimated time to create pseudonymization scripts: 2 hours

Link to an example of how data is pseudonymized in an existing OEA module: [*https://github.com/microsoft/OpenEduAnalytics/blob/main/modules/Student\_and\_School\_Data\_Systems/notebook/ContosoSIS\_py.ipynb*](https://github.com/microsoft/OpenEduAnalytics/blob/main/modules/Student_and_School_Data_Systems/notebook/ContosoSIS_py.ipynb)

## Documentation

OEA modules should be well documented. Documents and resources for a module should include OEA logo, Creative Commons license, MIT license and “OEA curated” before publishing​.

## Readme

Each asset folder should have a readme file that describes the content of the folder.

The main module readme file should include among other details: ​

* Description of data sources: what it is used for, data available, data format and possible use cases or OEA packages it can be used for.​
* Explanation of how to use the module: prerequisites (like subscriptions), permissions, types of data transfer services that can be used to ingest in OEA​, etc.

## Docs folder

Documents with additional information and instructions can be uploaded into the docs folder in cases where the default readme file on GitHub is insufficient like [this example](https://github.com/microsoft/OpenEduAnalytics/blob/main/modules/Microsoft_Data/Microsoft_Graph/docs/Graph%20Reports%20API%20Module%20Tutorial.pdf). The documents can include PDFs that have screenshots and explanations on the steps for performing tasks. For example, setting up a data pipeline for ingesting data from an API to Synapse, guidelines for using the module in a production environment, the number of days that data can be pulled from a data source, etc. The docs folder can also include files like a module guide, diagrams, etc. All images should be placed within the images folder in the docs folder.

Estimated time for writing documentation: 4 hours

Link to docs template: [*https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/\_Module\_Creation\_Kit/docs*](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/_Module_Creation_Kit/docs)

# Steps for creating a new OEA module

1. Complete the pre-requisites listed in the Pre-requisites for OEA module creation section.
2. Define the problem statement and benefits the module will have on learners and educators. This will help you decide what the best solution will be. For example, will it require some visualization in PowerBI or not? In case you will be creating a PowerBI dashboard, it is suggested that you speak with potential users, if possible, to understand what their pain points are and how this solution will be of help to them.
3. If you are going to be working on a PowerBI dashboard, this will be the time for you to think through what kind of visuals will be most compelling, actionable and tell a story of the data. It is a good idea to ideate the visuals by doing rough sketches that you can come back to iterate upon.
4. Based on the draft visuals, start working through the kind of data that will be needed to create the module. Will it be data from the student information system or data ingested through an API?
5. Once you have defined the kind of data needed and its sources, it is time to generate test data. As part of this module creation kit, we provide sample data sets in CSV format that you can load to the Synapse workspace. We also provide a test generator notebook for M365 data that can be used to create more test data for the M365 data source. This notebook can be modified to create new test data for other data sources.
6. Create a data dictionary for the data that explains each column of the test data sets.
7. List out any columns that contain personal identifiable information of students as they will be pseudonymized in a future step.
8. A pipeline can be created at different stages of the module creation process. For example, if you are planning to ingest data from an API into Synapse, you may want to create a pipeline for that. You can also use a pipeline to ingest data from other data sources into the Synapse data lake.
9. If the test data you have generated doesn’t need a pipeline for ingestion into the data lake, you may upload it directly to the stage1np container in the data lake. You can find this by navigating to the Data tab in Synapse, clicking on Azure Data Lake Storage Gen2, clicking on the storage account associated with your Synapse workspace and then choosing stage1np. It is a best practice to create a separate folder within the stage1np storage container for the new data file being uploaded. Also make sure to name the file appropriately.
10. You will now have to upload a notebook in the Develop tab for pseudonymizing any personal identifiable information of students like names and email addresses. The OEA deployment comes with a pseudonymization script that can be used to pseudonymize columns that contain personal identifiable information of students. Review [an example](https://github.com/microsoft/OpenEduAnalytics/blob/main/modules/Student_and_School_Data_Systems/notebook/ContosoSIS_py.ipynb) of the pseudonymization script that comes with setting up the Contoso\_SIS modules as part the OEA deployment.
11. Next, you will create a notebook for preprocessing the data. This may include cleaning, subsetting and aggregating the data. To do this, you will first have to provision the storage accounts and then load the raw data from the stage1np container in the data lake. You can then modify and implement the functions for processing data into the various stages that come with the OEA deployment. After preprocessing and aggregating, don’t forget to write back to the data lake, specifically the stage2p container and then load to a Spark DB so that the data can be easily queried and connected to PowerBI. We recommend using Synapse for most (and possibly all) of your data transformation before connecting to PowerBI.​
12. To connect the Spark DB to PowerBI, you will need to use the SQL Serverless endpoint. You will find the SQL Serverless endpoint in the Synapse workspace Overview in the resource group on the Azure portal. In the Data source settings in PowerBI, you will have to input the Server details and the Database name to be able to connect the Spark DB to PowerBI.
13. Revisit the draft visuals you came up with in Step 3 and start iterating upon them in PowerBI. It is recommended to name all pages and each visual appropriately. Also, add labels to each visual and tooltips that provide a brief description of the various visuals.
14. Setup any other pipelines to automate the various stages of the process. For example, ingesting the data, copying the data, executing the notebook, making sure to indicate the appropriate parameters. After creating all pipelines, export the pipeline template directly from Synapse and upload to the pipeline folder.
15. Complete this module creation table to ensure that all the important pieces are taken care of:

|  |  |
| --- | --- |
|  | **[NAME OF MODULE]** |
| Contributor name or Company name |  |
| Assets to develop | *\*Select all that apply*  Pipeline  Notebook  PowerBI template  Documentation  Test data |
| Data source | *Which data source are you planning to use for this module?* |
| Privacy and Compliance | *What considerations are you making to ensure that student privacy is always protected?* |
| Columns to ingest | *List all the columns in the chosen data source that must be ingested into the data lake:* |
| Pseudonymization | *List all the columns that contain personal identifiable information that must be pseudonymized:*  Columns to hash:  Columns to mask: |
| References and Resources | *List any references and resources related to this module:* |
| Plans beyond module submission | *What do you plan to do to maintain and improve upon this module after submission?* |

# Module Quality

## Factors to Consider

In assessing a module’s quality, we consider the following among others:

1. Module deployment ease, i.e., the time it takes to deploy the module. In general, for most modules, the goal is for it to be deployable in less than 2 hours.
2. Market tested prior to publishing, with advanced modules having 2 or more customers deploying the module successfully before publishing.
3. Commitment to submit module iterations and improvements to the OEA GitHub repository.

## Quality Rubric and Criteria

Modules contributed to OEA are categorized in 3 levels: **Basic**, **Approved** and **Advanced**.

The benchmark for OEA module creation is the [Microsoft Graph Reports API](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/Microsoft_Data/Microsoft_Graph) module, which meets the requirements for the ‘Advanced’ level. This module can be used as a guideline for creating and submitting a new OEA module.

Check the boxes associated with the requirement of each level to keep track of your progress directly in the [rubric folder](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/_Module_Creation_Kit/rubric) of the OEA GitHub repository.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Component | Basic  *(Meets minimum requirements)​* | Approved ​  *(Curation approved for GitHub)​* | | Advanced ​  *(Rockstar, highlighted)​* |
| Documentation​ | Use general OEA templates for all assets (OEA Logo, Creative Commons License, partner logo - if applicable)​.  Readme file showing where all assets in module are. Assets should be organized in folders.  Description of data source and data dictionary.  List any prerequisites for module (like subscriptions/licenses to data source needed)​. | Use general OEA templates for all assets (OEA Logo, Creative Commons License, partner logo - if applicable)​.  Readme file showing where all assets in module are. Assets should be organized in folders.  Description of data source: what it is used for, data available, data dictionary and possible use cases or OEA packages it can be used for.  Guidance on prerequisites for module (like subscriptions/licenses to data source needed)​. | Everything in the Approved level PLUS:  Documentation gives guidance for transitioning from sample data to production data​.  Deeper “user-guide” to be uploaded in docs folder​.  *Optional: Module roadmap​.* | |
| Collect​ | Sample data set (flat files, eg: CSV).  Scripts to clean all sensitive data from project assets​.  Synapse pipeline demonstrating data extraction from the data source. | Sample data set (flat files, eg: CSV).  Scripts to clean all sensitive data from project assets​.  Synapse pipeline demonstrating data extraction from the data source. | Everything in the Approved level PLUS:  Test data generator. | |
| Compute​ | Define schema for initial data prep and pseudonymization​.  Implement process\_stage1\_into\_stage2()​. | Define schema for initial data prep and pseudonymization​.  Implement process\_stage1\_into\_stage2()​. | Everything in the Approved level PLUS:  Follows OEA framework script.​  Add data validation, cleaning, aggregation and enrichment​.  Implement process\_stage2\_into\_stage3()​. | |
| Communicate​ |  | PowerBI semantic model demonstrating entity relationships​.  PowerBI dashboard with pages and visuals properly labeled. Each visual should also have tooltips with brief descriptions. | Everything in the Approved level. | |
| Quality​ |  | Module deployment takes less than 2 hours​. | Module deployment takes less than 30 minutes​.  Follows coding standards and useful comments in code​.  2 or more customers deployed successfully before publishing​. | |

# Contributing to OEA

Congrats on creating a new OEA module!

The next step is to contribute the module to the [OEA GitHub repository](https://github.com/microsoft/OpenEduAnalytics) by pushing all your updates and creating a pull request. If your module meets the module creation guidelines and quality rubric, it will be published on the OEA GitHub repository within 10 business days. As part of the module curation process, note that the OEA team reserves the right to update or modify any module submitted at any time.

N.B: Before submitting your new module, change the name of the folder from \_Creation\_Kit to the name of the new module.

# Additional Resources

* [Open Education Analytics Website](https://openeducationanalytics.org/)
* [Accelerating Learning Analytics and AI in Education](https://edudownloads.azureedge.net/msdownloads/Microsoft-Accelerating-Learning-Analytics-and-AI-in-Education.pdf)
* [MS Learning Path for Synapse Analytics](https://docs.microsoft.com/en-us/learn/paths/realize-integrated-analytical-solutions-with-azure-synapse-analytics/)
* [Azerbaijani Ministry of Education unlocks the power of data-driven education](https://customers.microsoft.com/en-us/story/862925-ministry-of-education-of-the-republic-of-azerbaijan-government-azure-en-azerbaijan)
* [Data Lakes and Modern Analytics for Education](https://www.bluegranite.com/blog/data-lakes-and-modern-analytics-for-education)
* [E2 2021 | Data-Driven Decision Making with Education Insights in Microsoft Teams](https://www.youtube.com/watch?v=JII0sBMWYFg)