Open Education Analytics

Module Creation Kit

Version 1.4

Last Updated: October 2022

[1 Introduction 3](#_Toc117141382)

[1.1 What is an OEA module? 3](#_Toc117141383)

[1.2 Why are OEA modules important? 3](#_Toc117141384)

[2 Setting up the OEA Module Creation template on GitHub 4](#_Toc117141385)

[3 Steps for creating a new OEA module 5](#_Toc117141386)

[4 Assets in a standard OEA module 6](#_Toc117141387)

[4.1 Pipeline 6](#_Toc117141388)

[4.1.1 Instructions for exporting/uploading pipeline templates 6](#_Toc117141389)

[4.2 Notebook 7](#_Toc117141390)

[4.2.1 Privacy: Scripts for pseudonymization​ and role-based access control 7](#_Toc117141391)

[4.3 Test data 8](#_Toc117141392)

[4.4 Documentation 9](#_Toc117141393)

[4.4.1 README.md 9](#_Toc117141394)

[4.4.2 Docs folder 9](#_Toc117141395)

[*4.5* Power BI Template 9](#_Toc117141396)

[5 Module Quality 10](#_Toc117141397)

[5.1 Factors to Consider 10](#_Toc117141398)

[5.2 Quality Rubric and Criteria 10](#_Toc117141399)

[6 Contributing to OEA 13](#_Toc117141400)

[7 Additional Resources 13](#_Toc117141401)

# Introduction

Open Education Analytics (OEA) is an open-source program coordinated by Microsoft Education collaborating with education systems and tech partners across the world to develop modern data intelligence capabilities.

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

## What is an OEA module?

In OEA, modules are a set of assets (e.g., notebooks, pipelines, test data, Power BI templates, 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 Education Insights module](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_catalog/Microsoft_Education_Insights) which brings in data from education-specific applications in O365.

## Why are OEA modules important?

Modules are the building blocks that create layers of data on the OEA framework. 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 tech 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.

Prior to creating a new module in OEA, the following pre-requisites must be 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 framework in Azure Synapse Analytics using the steps outlined in the [OEA GitHub repository.](https://github.com/microsoft/OpenEduAnalytics)
3. After deploying the OEA framework, ensure that all storage containers in the Azure Data Lake in Synapse have been provisioned. When done correctly, you should have the following for v0.6.1 of OEA: stage1np, stage2np, stage2p, stage3np, stage3p, oea-framework and synapse-workspace.

# Setting up the OEA Module Creation template on GitHub

The [modules folder](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules) on the OEA GitHub houses existing OEA modules and contains a [Module Creation Kit folder](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_creation_kit), which has the recommended templates for the OEA module components. If you already have the OEA GitHub repository cloned, first ‘Fetch upstream’ or ‘Pull’ the most recent version of the repository so that all the repository files are up to date.

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, email

Description automatically generated

1. Navigate to modules and then the module\_creation\_kit folder to start creating your new module. You may want to duplicate the module\_creation\_kit folder, rename it to the name of your new module and add your module assets to the respective folders. Remember to save all your updates.

# 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 the benefits the module will have on learners and educators.
3. Identify the data source you will be using for the module.
4. Generate test data based on the data source.
5. List out any columns that contain End User Identifiable Information of students as they will be pseudonymized in a future step.
6. A pipeline can be created at different stages of the module creation process. For example, for ingesting data from an API into Synapse, copying data, executing notebooks, etc.
7. Upload the notebooks for the module data ingestion and module Python class. Don’t forget to write back to stage 2 of the data lake, and then load to the SQL Serverless db so that the data can be easily queried and connected to Power BI. You can add the notebooks as Activities in the pipeline you created in the preceding step. We recommend using Synapse for most of your data transformation before connecting to Power BI.​
8. To connect to Power BI, you will need to use the SQL Serverless endpoint. This can be found in Manage tab in Synapse (under SQL pools). In the Data source settings in Power BI, input the Server details and the Database name.

If you plan to submit a Power BI template as part of this module, we suggest you speak with potential users, if possible, to understand what their pain points are and how they plan to use the Power BI dashboard. Think through the kind of visuals that will be most compelling, actionable and tell the right 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. Also name all pages and visuals in Power BI appropriately, adding labels and tooltips that provide a brief description of each visual.

# Assets in a standard OEA module

Get familiar with the assets in a standard OEA module:

## 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 as a pipeline template .zip file (refer to instructions below) so that module users can easily import the pipeline into their Synapse environment. Also provide a set of instructions for using the module pipeline for test data and migrating the module pipeline to production data.

Estimated time to create pipeline(s): 6 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)

### Instructions for exporting/uploading pipeline templates

Navigate to the module pipeline template you have created and open the pipeline. On the right-hand side you’ll see 3 dots – click on this and select “Export template”. The pipeline template will be downloaded locally after which you can upload it to the pipeline folder of the module.

Graphical user interface, application

Description automatically generated

## Notebook

Typically, there are 2 notebooks in a module:

1. **Module Data Ingestion Notebook** for executing ingestion from the module pipeline. This notebook depends on the module class. The pipeline template incorporates this notebook and is automatically uploaded upon importing the module pipeline template.
2. **Module Python Class Notebook** which contains the module class, defining the functions used in the ingestion notebook. It defines the data schemas and pseudonymization. Basic functions for data ingestion and processing from Stage 1 to Stage 2 data lakes are also included.

Both notebooks depend on the [OEA Python class](https://github.com/microsoft/OpenEduAnalytics/blob/main/framework/synapse/notebook/OEA_py.ipynb) which is a part of the [OEA framework](https://github.com/microsoft/OpenEduAnalytics/tree/main/framework). We strongly encourage the use of PySpark in notebooks; however, it is okay to use other programming languages.

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

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 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 End User Identifiable Information 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.

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 End User 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.

[Example](https://github.com/microsoft/OpenEduAnalytics/blob/main/modules/module_catalog/Student_and_School_Data_Systems/notebook/ContosoSIS_py.ipynb) of how data is pseudonymized in an existing OEA module.

Estimated time to create notebook(s): 6 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)

## Test data

Including sample test data sets in the module makes it easier for education systems to bring the module to life in their non-production environments. We provide sample data sets as part of this kit. However, you are free to generate your own test data sets. Please 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 Power BI visualization. For example, using the same students and same schools. To make this process seamless, we recommend integrating OEA standardized schemas: common education data standards like Caliper, Ed-Fi and SIF allow for data solutions to be built on a common analytical foundation and for a ‘plug and play’ approach to combining data from multiple sources. [Learn more about how to integrate OEA schemas](https://github.com/microsoft/OpenEduAnalytics/tree/main/schemas) in your new module. If relevant, test data sets can be created using tools like [SDV](https://sdv.dev/SDV/index.html). In subsequent iterations of this kit, we will provide best practices for test data generation.

Estimated time to generate test data: 8 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)

## 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”​.

### README.md

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

The main module README.md 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​, simple overview of implementation, etc.

Kindly use the [OEA Module Creation Kit Visuals PowerPoint templates](https://github.com/microsoft/OpenEduAnalytics/blob/main/modules/module_creation_kit/docs/OEA%20Module%20Creation%20Kit%20Visuals.pptx) to create the visuals for the module documentation. When done, screenshots of the visuals for the Module Overview and Module Setup Instructions should be added to the README.md file.

### Docs folder

Documents with additional information and instructions can be uploaded into the docs folder in cases where the default README.md file on GitHub is insufficient. The documents can include screenshots, guides, 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)

## Power BI Template

Education data analysts want to quickly visualize and explore the ‘ready’ data from the module’s data source in Power BI. Modules should aim to make the time from data ingestion to visualization rapid. It should also include the Power BI data model to explore the relationships between entities, and its visuals should be well labeled and have tooltips that provide a brief description. As shown in the Power BI template example on GitHub, please add a description of the module in the Explanation page.

For ease of use, we recommend submitting 2 versions of the Power BI template: one with module test data imported locally and the other connected to a Synapse workspace data source via DirectQuery.

Estimated time to create Power BI template: 6 hours

Link to Power BI template:

[*https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module\_creation\_kit/Power BI*](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_creation_kit/powerbi)

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

# 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 Education Insights](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_catalog/Microsoft_Education_Insights) 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 this kit.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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, contributor’s logo - if applicable)​.  README.md 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, contributor’s logo - if applicable)​.  README.md file showing where all assets in module are. Assets should be organized in folders. Module Overview and Module Setup Instructions visuals are included.  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​ | 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). that's anonymized.  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 function(s) to process data from stage 1 into stage 2. | Define schema for initial data prep and pseudonymization​.  Implement function(s) to process data from stage 1 into stage 2. | Everything in the Approved level PLUS:  Follows OEA framework script.​  Add data validation, cleaning, aggregation and enrichment​.  Implement function(s) to process data from stage 2 into stage 3. | |
| Communicate​ |  | Power BI dashboard with pages and visuals properly labeled. Each visual should also have tooltips with a brief description.  Power BI data model demonstrating entity relationships​. | Power BI dashboard with pages and visuals properly labeled. Each visual should also have tooltips with a brief description.  Power BI data model demonstrating entity relationships​. | |
| 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 modules submitted at any time.

Before submitting your new module, change the name of the folder from module\_creation\_kit to the name of the new module.

If this is your first time 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/).

# Additional Resources

* [OEA Website](https://openeducationanalytics.org/)
* [OEA Recommended Learning Resources](https://openeducationanalytics.org/skills-and-training/)
* [Introduction to OEA Schemas](https://www.youtube.com/watch?v=CKCViioIE0M&t=318s)
* [E2 2021 | Data-Driven Decision Making with Education Insights in Microsoft Teams](https://www.youtube.com/watch?v=JII0sBMWYFg)
* [Accelerating Learning Analytics and AI in Education](https://edudownloads.azureedge.net/msdownloads/Microsoft-Accelerating-Learning-Analytics-and-AI-in-Education.pdf)