Open Education Analytics

Module Creation Kit

Version 1.3

Last Updated: September 2022

[1 Introduction 3](#_Toc115010044)

[1.1 What is an OEA module? 3](#_Toc115010045)

[1.2 Why are OEA modules important? 3](#_Toc115010046)

[2 Pre-requisites for OEA module creation 3](#_Toc115010047)

[3 Setting up the OEA Module Creation template on GitHub 4](#_Toc115010048)

[4 Assets in a standard OEA module 5](#_Toc115010049)

[4.1 Pipeline 6](#_Toc115010050)

[4.2 Notebook 7](#_Toc115010051)

[4.2.1 Additional information about pseudonymization process 7](#_Toc115010052)

[*4.3* Power BI template 7](#_Toc115010053)

[4.4 Test data 8](#_Toc115010054)

[4.5 Privacy: Scripts for pseudonymization​ and role-based access control 8](#_Toc115010055)

[4.6 Documentation 9](#_Toc115010056)

[4.7 Readme 9](#_Toc115010057)

[4.8 Docs folder 9](#_Toc115010058)

[5 Steps for creating a new OEA module 9](#_Toc115010059)

[6 Module Quality 12](#_Toc115010060)

[6.1 Factors to Consider 12](#_Toc115010061)

[6.2 Quality Rubric and Criteria 13](#_Toc115010062)

[7 Contributing to OEA 15](#_Toc115010063)

[8 Additional Resources 15](#_Toc115010064)

# Introduction

Open Education Analytics (OEA) is an open source program 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 quickly and effectively.

## What is an OEA module?

In the OEA framework, modules are a set of assets (e.g., notebooks, Power BI 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 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. This data can be combined with other digital activity data from other applications and platforms used in learning to develop "digital learning insights" across the ecosystem of applications and platforms a student uses.

## 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 for all these pre-requisites to 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 reference architecture in Azure Synapse Analytics using the steps outlined in the [OEA GitHub repository.](https://github.com/microsoft/OpenEduAnalytics)
3. After deploying 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 v0.6 of OEA: 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) 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 to get you started. 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](https://github.com/microsoft/OpenEduAnalytics/tree/main/modules/module_creation_kit) 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.

# 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 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, Power BI and the OEA architecture work, it may take you about 80 hours to create a first working version of the new module after completing some of the preliminary Synapse, OEA and Power BI trainings.
* If you have intermediate or advanced skills in Synapse, Power BI and the OEA architecture, it may take you about 40 hours to have a first working 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) * Data preprocessing, transformation and analysis notebook (in .ipynb format) | notebook |
| Power BI template, with all pages and visuals properly labelled and with tooltips (in .pbix format) | Power BI |
| 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 file |
| * 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.

## Power BI 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 Power BI. Modules should aim to make the time from data ingestion to visualization rapid. The Power BI 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 Power BI semantic model to explore the relationships between entities. Power BI visuals should be well labeled and have a tooltip that provides a brief description of each visual. As shown in the Power BI template example on GitHub, please add a description of the module in the Explanation page.

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.

## 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. This helps stakeholders get a first-hand experience of how the module works without having an impact on their live environments. We provide sample data sets as part of this kit. However, you are free to generate your own test data (including a data, 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.

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/module\_catalog/Student\_and\_School\_Data\_Systems/notebook/ContosoSIS\_py.ipynb*](https://github.com/microsoft/OpenEduAnalytics/blob/main/modules/module_catalog/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.md 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/module_catalog/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 the 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 Power BI or not? In case you will be creating a Power BI dashboard, we suggest 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 will be working on a Power BI dashboard, this will be the time for you to 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.
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.
6. Create a data dictionary 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 like data contained in a file in a GitHub repository.
9. You will now have to upload the notebook(s) in the Develop tab for pseudonymizing any personal identifiable information of students like names and email addresses, as well as notebooks for preprocessing the data including cleaning, subsetting, aggregating the data, etc. 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 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.​
10. To connect the Spark DB to Power BI, you will need to use the SQL Serverless endpoint. This can be found in the Synapse workspace Overview in the resource group on the Azure portal. In the Data source settings in Power BI, input the Server details and the Database name to be able to connect the Spark DB to Power BI.
11. Revisit the draft visuals you came up with in Step 3 and iterate upon them in Power BI. Name all pages and visuals appropriately, adding labels and tooltips that provide a brief description of each visual.
12. 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.
13. 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  Power BI 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:  N.B: The initial 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/module_catalog/Student_and_School_Data_Systems/notebook/ContosoSIS_py.ipynb) of the pseudonymization script that comes with setting up the Student and School Data Systems module as part the initial OEA deployment. |
| 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 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, 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 function to process data from stage 1 into stage 2. | Define schema for initial data prep and pseudonymization​.  Implement function 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 to process data from stage 2 into stage 3. | |
| Communicate​ |  | Power BI semantic 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 semantic model demonstrating entity relationships​.  Power BI dashboard with pages and visuals properly labeled. Each visual should also have tooltips with a brief description. | |
| 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 module\_creation\_kit to the name of the new module.

# Additional Resources

* [Open Education Analytics 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)