Open Edu Analytics

Solution Guide

Published: November, 2020

[Introduction 1](#_Toc58257496)

[1) Setup of base architecture and test env 2](#_Toc58257497)

[2) Walking through the included example 4](#_Toc58257498)

[2) Setup of modules and/or packages 6](#_Toc58257499)

[3) Connect Power BI workspace 6](#_Toc58257500)

­­

|  |
| --- |
| Introduction |
| This document provides step by step instructions for the setup of the Open Edu Analytics base solution as well as information on how to deploy modules and packages on the base architecture.  Open Edu Analytics is an open source modern data warehouse solution for education, built on [Synapse Analytics](https://azure.microsoft.com/services/synapse-analytics) and the powerful set of Azure platform data services.  For a set of brief introductory videos on Synapse Analytics see: [Azure Synapse Analytics demo videos](https://azure.microsoft.com/en-us/services/synapse-analytics/video-demos).  For a step-by-step guide through Synapse Analytics, see: [Get Started with Azure Synapse Analytics](https://docs.microsoft.com/en-us/azure/synapse-analytics/get-started).  All scripts and documentation for the Open Edu Analytics solution can be found at: <https://github.com/microsoft/OpenEduAnalytics>  The Open Edu Analytics solution is comprised of:   1. The core Open Edu Analytics solution architecture – an Azure storage account, a Synapse workspace, an Apache Spark Pool, AAD security groups 2. Data modules – Apache Spark notebooks for the processing of source data from a specific source system. Data modules can be seen as data silos, bringing in data from a single system, with no dependencies. 3. Solution packages – a package of assets such as Apache Spark notebooks for provisioning a comprehensive view over multiple data sets, Power BI reports, and Machine Learning models. A solution package utilizes one or more data modules for providing the source data utilized in the solution.   In order to begin the setup of the Open Edu Analytics solution, all you need is an [Azure subscription](https://azure.microsoft.com/en-us/features/azure-portal/).  The diagram below provides a high-level overview of the reference architecture. |

Timeline

Description automatically generated

# 1) Setup of base architecture and test env

In this section you will use a script to provision the Azure resources that comprise the core of this solution, as well as a test package that provides example datasets and notebooks to use for further exploring the capabilities of Synapse Analytics.

1.1) In Azure portal, click on the Cloud Shell icon, then select “Bash”.

Graphical user interface, text, application, PowerPoint

Description automatically generated

1.2) Click on “Create storage”

Graphical user interface, text, application

Description automatically generated

1.3) At the bash shell prompt, enter the following commands to download the contents of the OpenEduAnalytics repository to your Azure cloud drive.  
cd clouddrive

git clone https://github.com/microsoft/OpenEduAnalytics

A screenshot of a computer screen

Description automatically generated

1.4) Now run the setup script found in the root directory of OpenEduAnalytics by running the following commands. Note that for <orgid> you should enter an ID for your org which will be used as a suffix of Azure resources that must be unique. For example, a school district named Contoso Independent School District might choose an org ID of “CISD” or “ContosoISD”.

cd OpenEduAnalytics

./setup.sh <orgid>

The Azure resources will be created in the East US region by default. In order to have the resources created in a different location, specify the desired location as the second argument to the script:

./setup.sh <orgid> <location>

For a list of available locations, see: [Products available by region](https://azure.microsoft.com/en-us/global-infrastructure/services/?products=synapse-analytics&regions=all)

The installation script will then take several minutes to complete, as it provisions the following Azure resources:

1. EduAnalytics resource group
2. storage account with 4 storage containers (named synapse, stage1, stage2, stage3, testdata)
3. Azure Synapse workspace
4. Apache Spark pool
5. Azure AD security groups, with appropriate role assignments to access the storage account via Synapse studio

Pictured below are screenshots of the created resources:

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

Note too that the automated setup installs test datasets in the test-env storage container. See the section “Walking through the example” for more details on how to use this data to run example notebooks and learn more about Synapse Analytics.

# 2) Walking through the included example

2.1) Open your new Synapse Workspace by clicking on the url at the end of the setup script



or you also launch your Synapse Workspace from Azure portal:

Graphical user interface, text, application, email

Description automatically generated

2.2) Download the notebook “Contoso\_ISD\_all\_in\_one.ipynb” found in clouddrive/OpenEduAnalytics/tmp

This can be done by clicking on the upload/download icon in Cloud Shell as shown below.

Graphical user interface, text, application

Description automatically generated

2.3) In Synapse Workspace, click on Develop then click on “+” and select Import and choose the Contoso\_ISD\_all\_in\_one.ipynb downloaded in the last step.

Graphical user interface, text, application, email

Description automatically generated

2.4) Click on “Run all”. This will startup a spark cluster and then execute each of the cells in the notebook.

By default, this notebook is configured to use the test data included in the initial setup.

2.5) Once the notebook execution has completed, navigate to “Data”, expand s2\_m365, hover on the table named “org”, click on the menu for the table and select “New SQL script”, then choose “Select TOP 100 rows”

Graphical user interface, text

Description automatically generated

2.6) In the “SQL script 1” tab that opens, click on “Run”. Note the list of schools that were included in the test data set.

Perform other queries across the various databases to get a feel for the test data and the Synapse interface.

Note that the data in all tables with the “s3” prefix have sensitive data masked.

Graphical user interface, text, application, email

Description automatically generated

# 2) Setup of modules and/or packages

The Open Edu Analytics solution provides a base architecture as well as individual modules and packages.

Modules provide assets for the processing and usage of data from a single data source, while packages provide a set of assets that incorporate data from one or more data sources. Modules and packages can contain the same set of assets – the only distinction between the two is that modules are self-contained while packages have dependencies on one or more modules.

Modules and packages follow a standard structure which includes:

1. a readme.md for basic documentation
2. a setup.sh script to be used for automated deployment from [cloud shell](https://azure.microsoft.com/en-us/features/cloud-shell/#overview)
3. a notebooks folder for Synapse notebooks
4. a powerbi folder for Power BI assets (this is optional)

# 3) Connect Power BI workspace

To connect your Power BI workspace to Synapse so that it is accessible from with Synapse studio, login to Synapse studio and click on “Manage”, then select “Linked services”, then click on “Connect to Power BI” and complete the form with the connection info to your Power BI workspace.

Graphical user interface, text, application, email

Description automatically generated

For more details see: [Linking a Power BI workspace to a Synapse workspace](https://docs.microsoft.com/en-us/azure/synapse-analytics/quickstart-power-bi)

*This is a preliminary document and may be changed substantially prior to final commercial release of the software described herein.*

*The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.*

*This white paper is for informational purposes only. Microsoft makes no warranties, express or implied, in this document.*

*Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in, or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.*

*Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.*

*© 2020 Microsoft Corporation. All rights reserved*