

Open Edu Analytics Solution Guide

Published: January, 2021

Introduction	1
1) Setup of base architecture and test env	2
Debugging	4
2) Walking through the included example	5
3) Understanding the notebooks	7
4) Power BI dashboard examples	8
5) Connect Power BI workspace	11
6) Privacy and Security	12

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](#) 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](#).

For a step-by-step guide through Synapse Analytics, see: [Get Started with Azure Synapse Analytics](#).

For a detailed e-book on analytics in Azure, see: [Cloud Analytics with Microsoft Azure](#)

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, and a set of AAD security groups
- 2) 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) 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.

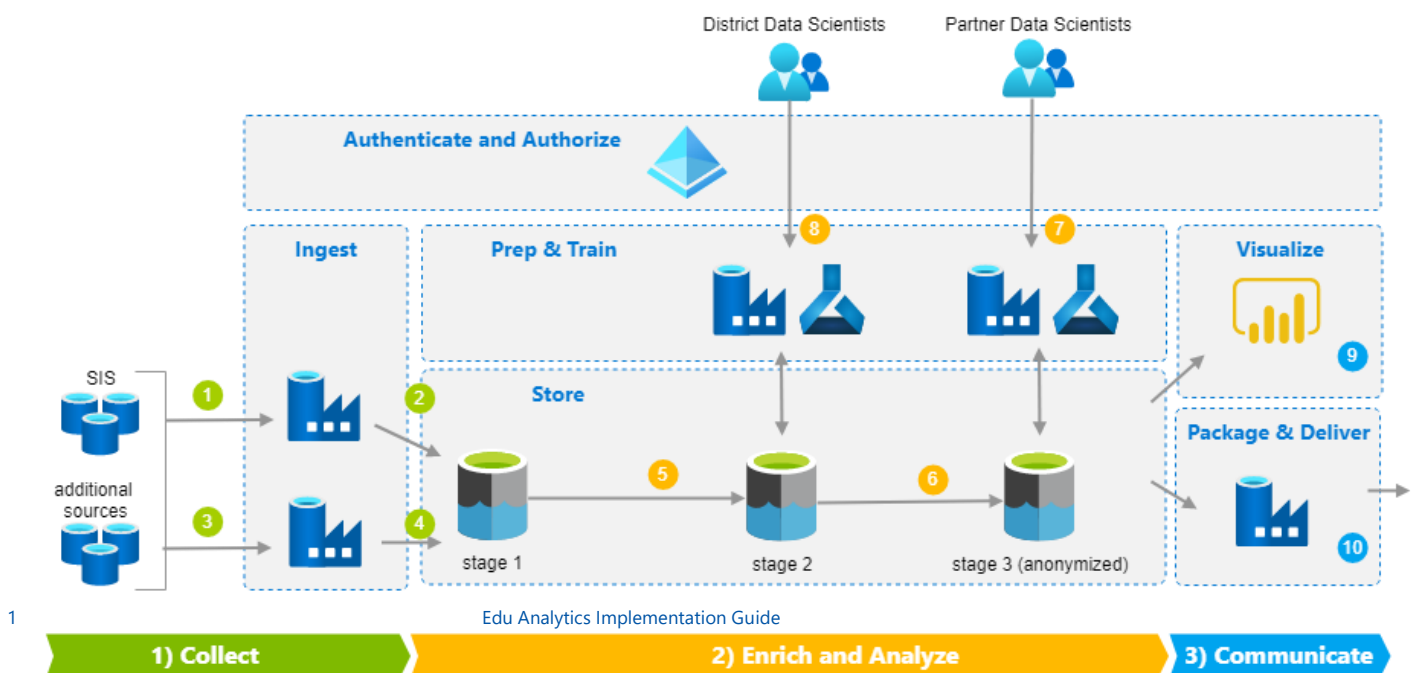
Modules and packages in Open Edu Analytics can contain the same set of assets – the main distinction between the two is that modules are self-contained while packages have dependencies on one or more modules.

Modules and packages have the following standard structure:

1. a readme.md for basic documentation
2. a setup.sh script to be used for automated deployment from [cloud shell](#)
3. a notebooks folder for Synapse notebooks
4. a powerbi folder for Power BI assets (this is optional)

In order to begin the setup of the Open Edu Analytics solution, all you need is an [Azure subscription](#). See the following section for detailed setup instructions.

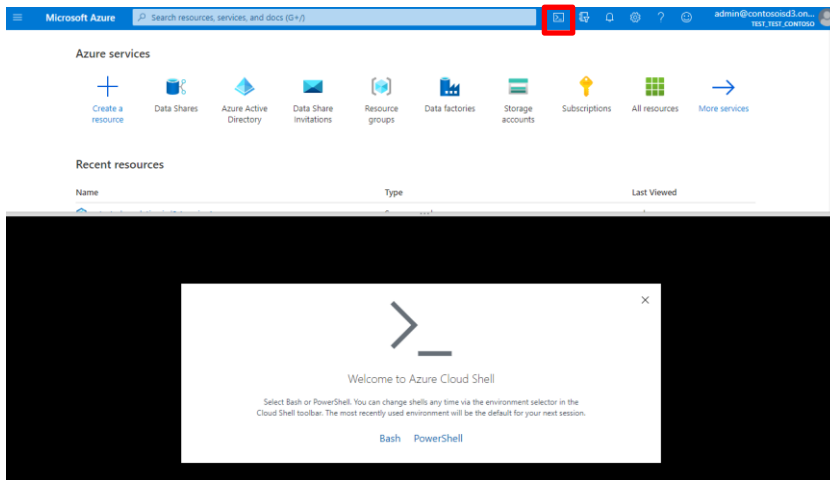
The diagram below provides a high-level overview of the reference architecture.



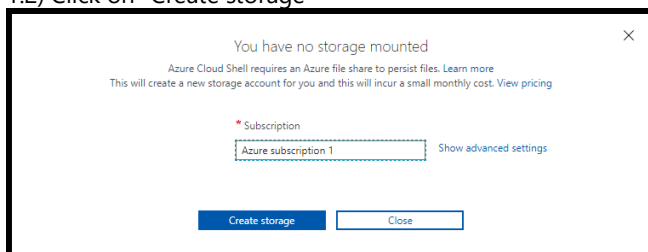
1) Setup of base architecture and test env

In this section you will use a script to provision the Azure resources that comprise the core of this solution, as well as an example solution 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”.

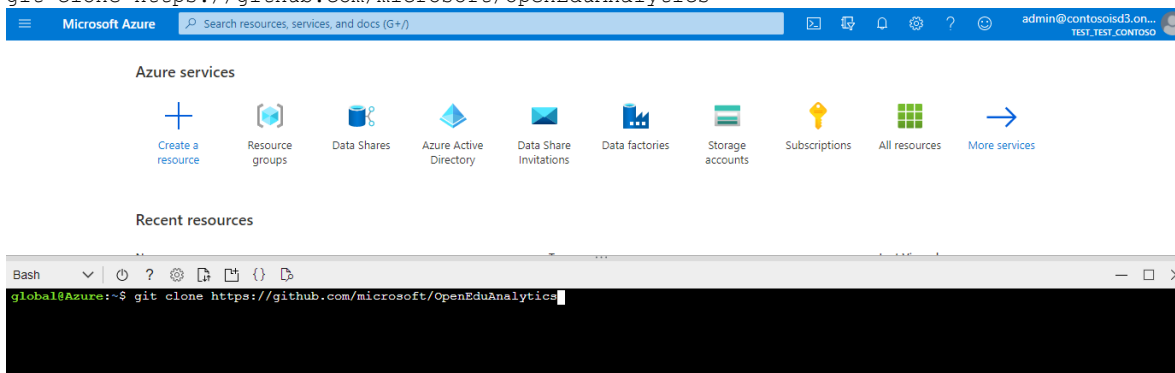


1.2) Click on “Create storage”



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
```



1.4) Now run the setup script found in the root directory of OpenEduAnalytics by running the following commands. Note that for <unique_suffix> in the command below, 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 <unique_suffix>
```

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 <unique_suffix> <location>

For a list of available locations, see: [Products available by region](#)

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, test-env)
3. Azure Synapse workspace
4. Apache Spark pool

Pictured below are screenshots of the created resources:

The screenshot shows the 'EduAnalyticsCISD3a' resource group overview in the Microsoft Azure portal. The left sidebar contains navigation links for Overview, Activity log, Access control (IAM), Tags, Events, Settings (Deployments, Policies, Properties, Locks), and Cost Management. The main content area displays the 'Essentials' section with the following information:

- Subscription: Azure subscription 1
- Subscription ID: 9116e83a-48f0-4e84-80d8-7e73430608df
- Location: East US
- Tags: Click here to add tags

Below this, there is a table of resources with columns for Name, Type, and Location. The table shows three records:

Name	Type	Location
spark1 (syeduanalyticscis3a/spark1)	Apache Spark pool	East US
steduanalyticscis3a	Storage account	East US
syeduanalyticscis3a	Synapse workspace	East US

The screenshot shows the 'steduanalyticscis3a' storage account overview in the Microsoft Azure portal. The left sidebar contains navigation links for Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data transfer, Events, and Storage Explorer (preview). The main content area displays the 'Containers' section with a search bar and a table of containers.

Name	Last modified	Public access level
stage1	12/7/2020, 4:42:57 PM	Private
stage2	12/7/2020, 4:42:58 PM	Private
stage3	12/7/2020, 4:42:59 PM	Private
synapse	12/7/2020, 4:42:56 PM	Private
test-env	12/7/2020, 4:43:00 PM	Private

The screenshot shows the 'test_test_Contoso' group overview in the Microsoft Azure portal. The left sidebar contains navigation links for All groups, Deleted groups, Diagnose and solve problems, Settings (General, Expiration, Naming policy), and Activity. The main content area displays the 'All groups' section with a search bar and a table of groups.

Name	Object Id	Group Type	Membership Ty
Edu Analytics Dat...	2c35a650-0d54-49a5-b1...	Security	Assigned
Edu Analytics Dat...	ece793a4-fd6d-41a1-a5...	Security	Assigned
Edu Analytics Ext...	2cb2cd62-2cbb-4577-8...	Security	Assigned
Edu Analytics Glo...	14fe17d6-1fa3-4eba-85...	Security	Assigned

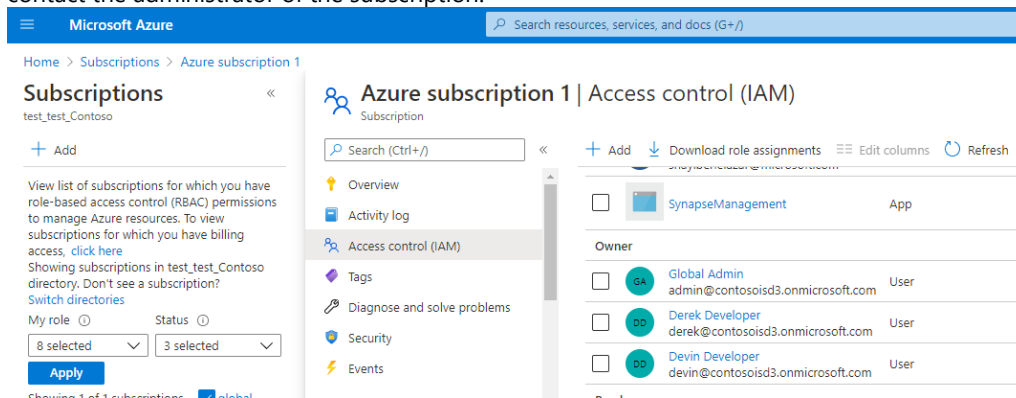
Note too that the automated setup installs test datasets in the test-env storage container. See the section “Walking through the example” for more details on how to use this data to run example notebooks and learn more about Synapse Analytics. You can also choose to have the script create security groups to facilitate the use of role based access control to the data lake. If you are running the setup for an environment in which you have Global Admin permissions on the tenant, and you want to have security groups provisioned, you can invoke the setup script like this:

```
./setup.sh <unique_suffix> <location> true
```

Debugging

Role Assignment of “Owner”

If you got errors when running the setup script, verify that you have the role assignment of “Owner” on your Azure subscription by going to Azure portal, then going to Subscriptions -> Access control (IAM) -> Role assignments. If you do not have the role assignment of “Owner”, you may be able to add the role assignment yourself – otherwise you’ll need to contact the administrator of the subscription.



ValidationFailed

If the setup script fails, scroll up to the first error that occurred. If that error is “BadRequestError: ValidationFailed: Workspace request validation failed, check error details for more information” and it occurs when attempting to create the Synapse workspace, these are some potential causes:

- 1) A resource provisioning timing issue can occur. Wait few minutes then try rerunning the setup script with the same arguments (it will skip over resources already created).
- 2) It’s possible that there was a naming conflict (some of the resources being created must be globally unique). Try rerunning the setup script using a different prefix.
- 3) Your Azure subscription may need to have additional Resource providers registered. See: [Resource providers and resource types - Azure Resource Manager | Microsoft Docs](#)

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

--> Setup of the test environment is complete.

Click on this url to open your Synapse Workspace: <https://web.azuresynapse.net?workspace=%2fsubscriptions%2f9116e83a-48f0-4e84-80d8-7e73430608df>

or you can also launch your Synapse Workspace from Azure portal, as show here:

The screenshot shows the Azure portal interface for a Synapse workspace named 'syeduanalyticscis3a'. The left sidebar contains navigation options like Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings, SQL Active Directory admin, Properties, Locks, Analytics pools, and Security. The main content area displays workspace details such as Resource group (EduAnalyticsCISD3a), Status (Succeeded), Location (East US), Subscription (Azure subscription 1), Subscription ID, Managed virtual network (No), Managed Identity object, Workspace web URL, and Tags. A 'Getting started' section is highlighted with a red box, featuring a button to 'Open Synapse Studio' and a link to 'Read documentation'.

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.

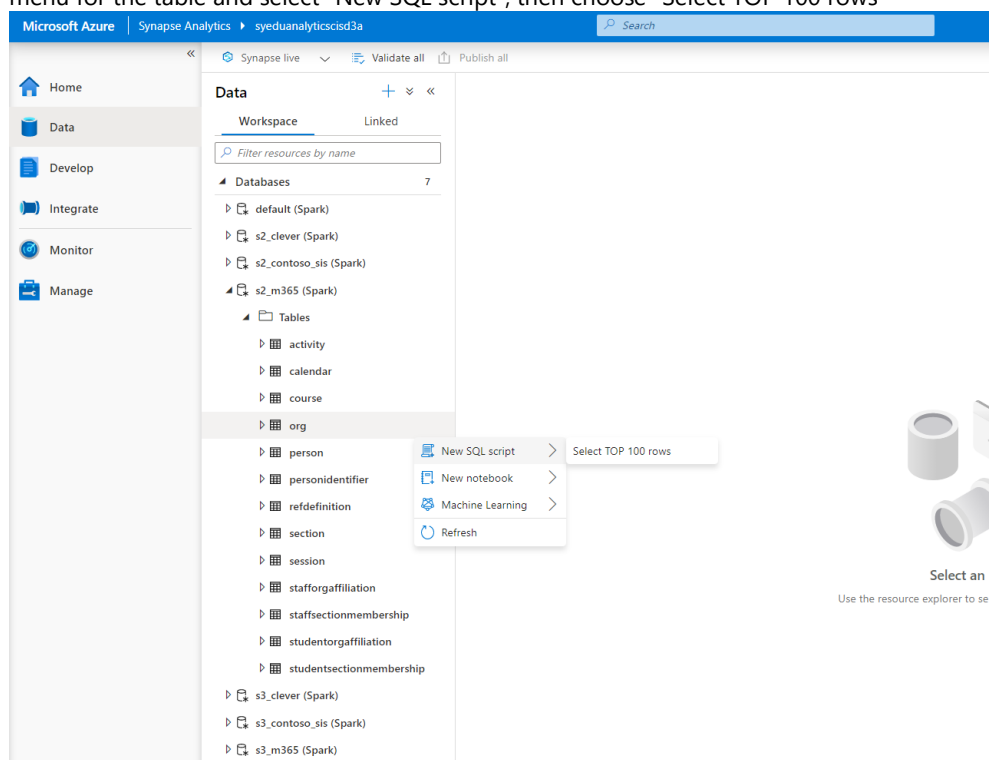
The screenshot shows the Azure Cloud Shell interface. The terminal displays the command to copy the notebook from the cloud drive: `azcopy copy "/usr/cuser/clouddrive/OpenEduAnalytics/tmp/Contoso_ISD_all_in_one.ipynb" "/usr/cuser/clouddrive/OpenEduAnalytics/tmp/Contoso_ISD_all_in_one.ipynb" --recursive`. The output shows the file being scanned and downloaded. A context menu is visible over the terminal, with the 'Download' option highlighted.

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.

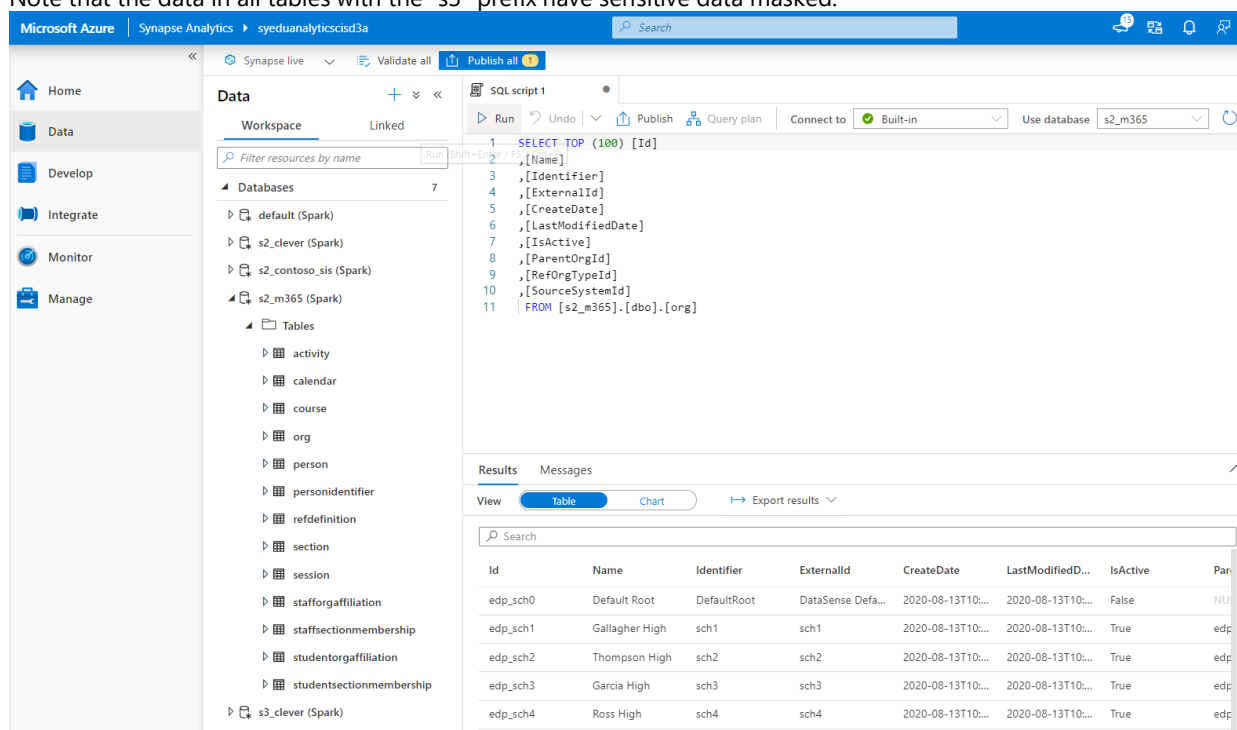
The screenshot shows the Synapse Studio interface. The left sidebar contains navigation options like Home, Data, Develop, Integrate, Monitor, and Manage. The main content area displays the 'Contoso_ISD_all_in_one.ipynb' notebook. The notebook is in the 'Develop' tab, and the 'Cell 2' is selected. The code in Cell 2 sets the storage account and test environment. The 'Cell 3' is also visible, containing code for setting up the environment. The notebook is attached to the 'spark1' pool and the language is set to 'PySpark (Python)'.

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”



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.



3) Understanding the notebooks

[Create external tables backed by Parquet in Spark and query from serverless SQL pool](#)

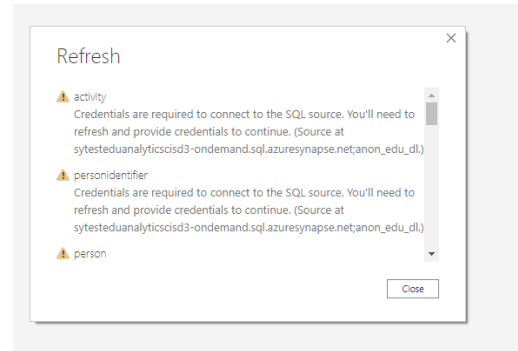
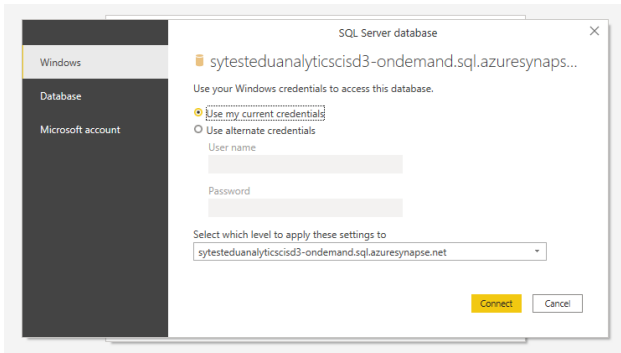
4) Power BI dashboard examples

The previous section demonstrated the steps needed for a complete setup with a test environment and test data.

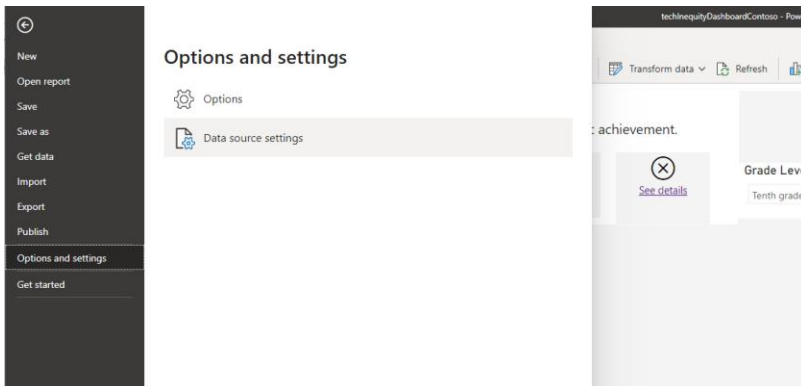
This section will demonstrate how to open the example Power BI dashboards in Power BI desktop and connect to the data lake in your test environment via SQL On-Demand. You will need to have Power BI Desktop installed on your computer to complete this section (Power BI Desktop is free to download and free to use – it can be [downloaded from here](#)).

3.1) Navigate to your local version of the OpenEduAnalytics repository, and double-click on OpenEduAnalytics\packages\ContosoISD\powerbi\techInequityDashboardContoso.pbix to open the Power BI dashboard in Power BI desktop.

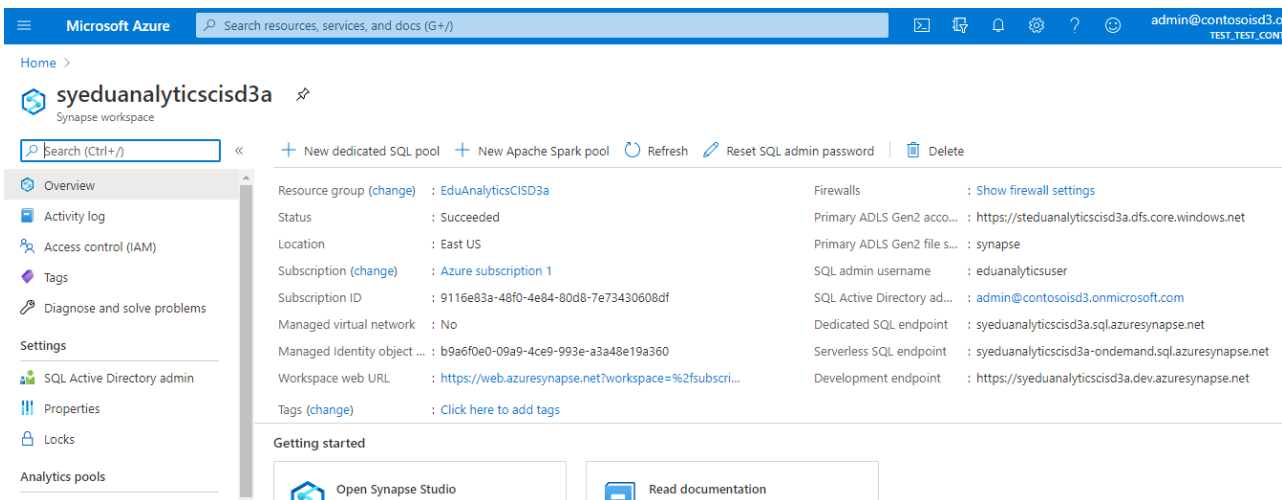
3.2) You will be prompted for credentials to the pre-configured data source, but since you need to specify a new data source, you can just click on “Cancel”, and then click on “Close” on the next dialog.



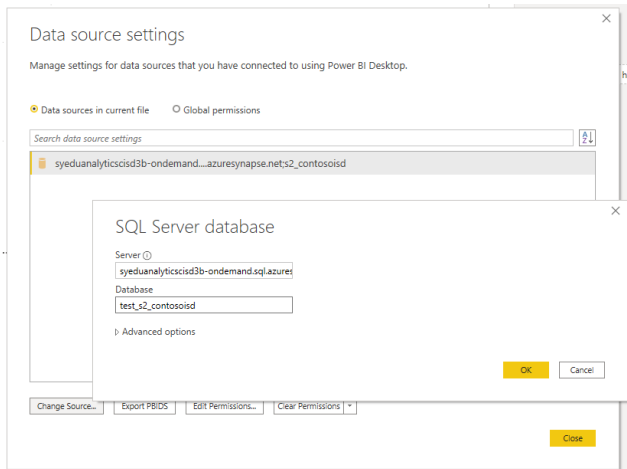
3.3) Now click on File -> Options and Settings -> Data source settings, and on the next screen click on “Change Source”



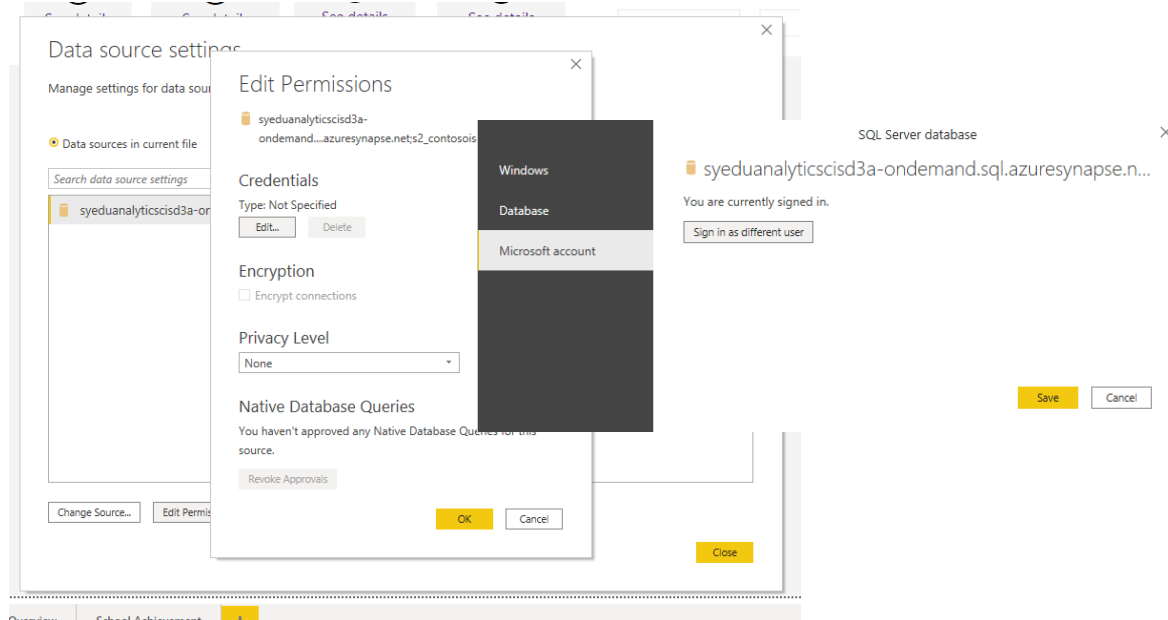
3.4) In order to get the right url for your server, go to portal.azure.com and navigate to your Synapse instance. You need to copy the value for “Serverless SQL endpoint”.



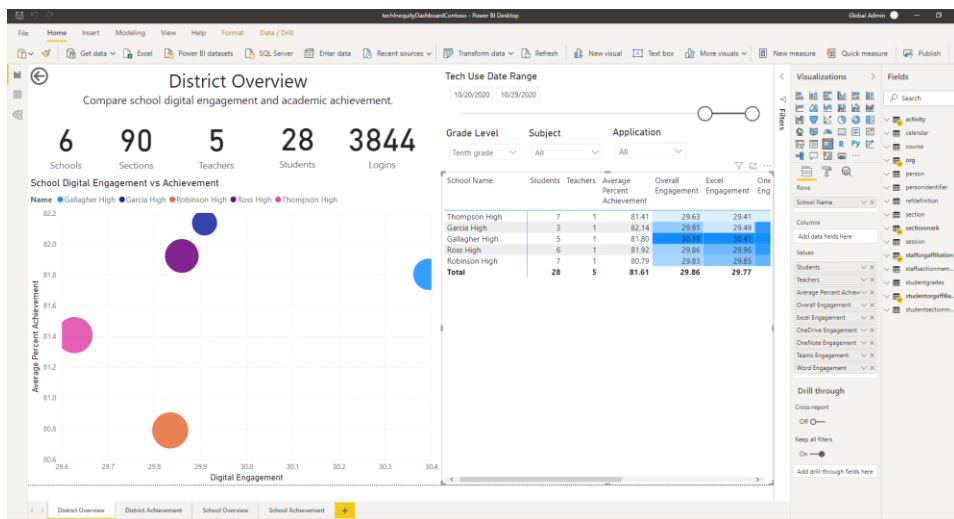
3.5) Enter the value you retrieved in the previous step in the textbox for "Server", and for "Database" enter "test_s2_contosoisd", then click on "OK"



3.6) Click "Edit Permissions", then under the heading of "Credentials" click "Edit", and in the next window click "Microsoft account", then click "Sign in", and complete the sign in process with the credentials for the user that has access to the Synapse workspace. Then click on "Save", followed by "OK", followed by "Close", and then click on "Apply changes".



3.7) You should see a dashboard similar to this screenshot. Click around and test the different tabs and interactive components.

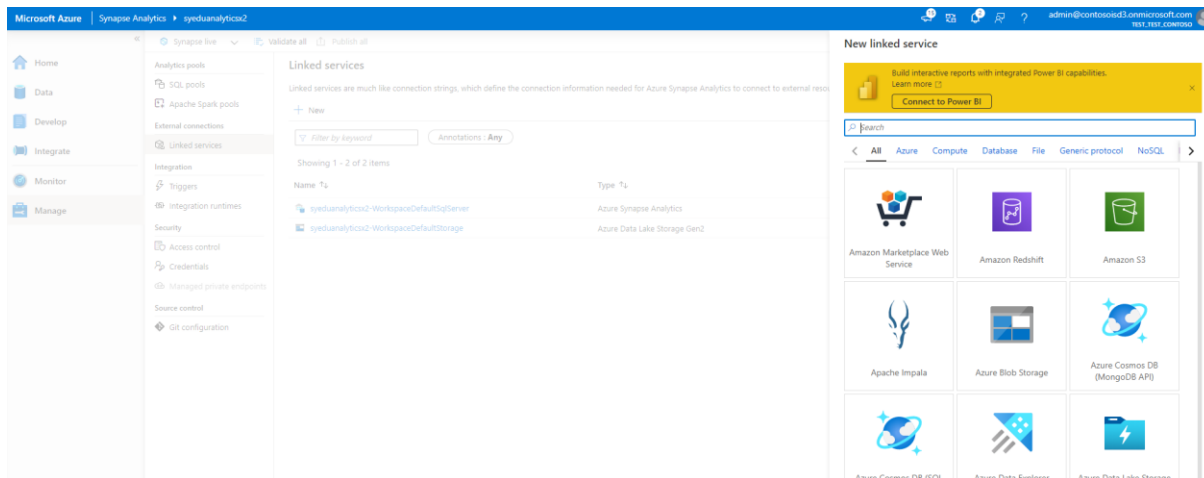


3.8) Following these same steps, you can open the report within the M365 module found at OpenEduAnalytics\modules\M365\powerbi\M365_dashboard.pbix
 The only difference is that when specifying the value for "Database", use the value "test_s2_m365"

5) Connect Power BI workspace

If you have a Power BI cloud account, you have the option of connecting a cloud based Power BI workspace to Synapse.

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.



For more details see: [Linking a Power BI workspace to a Synapse workspace](#)

For more information about what Power BI licenses are needed for a given scenario, see: [Power BI Premium FAQ - Power BI | Microsoft Docs](#)

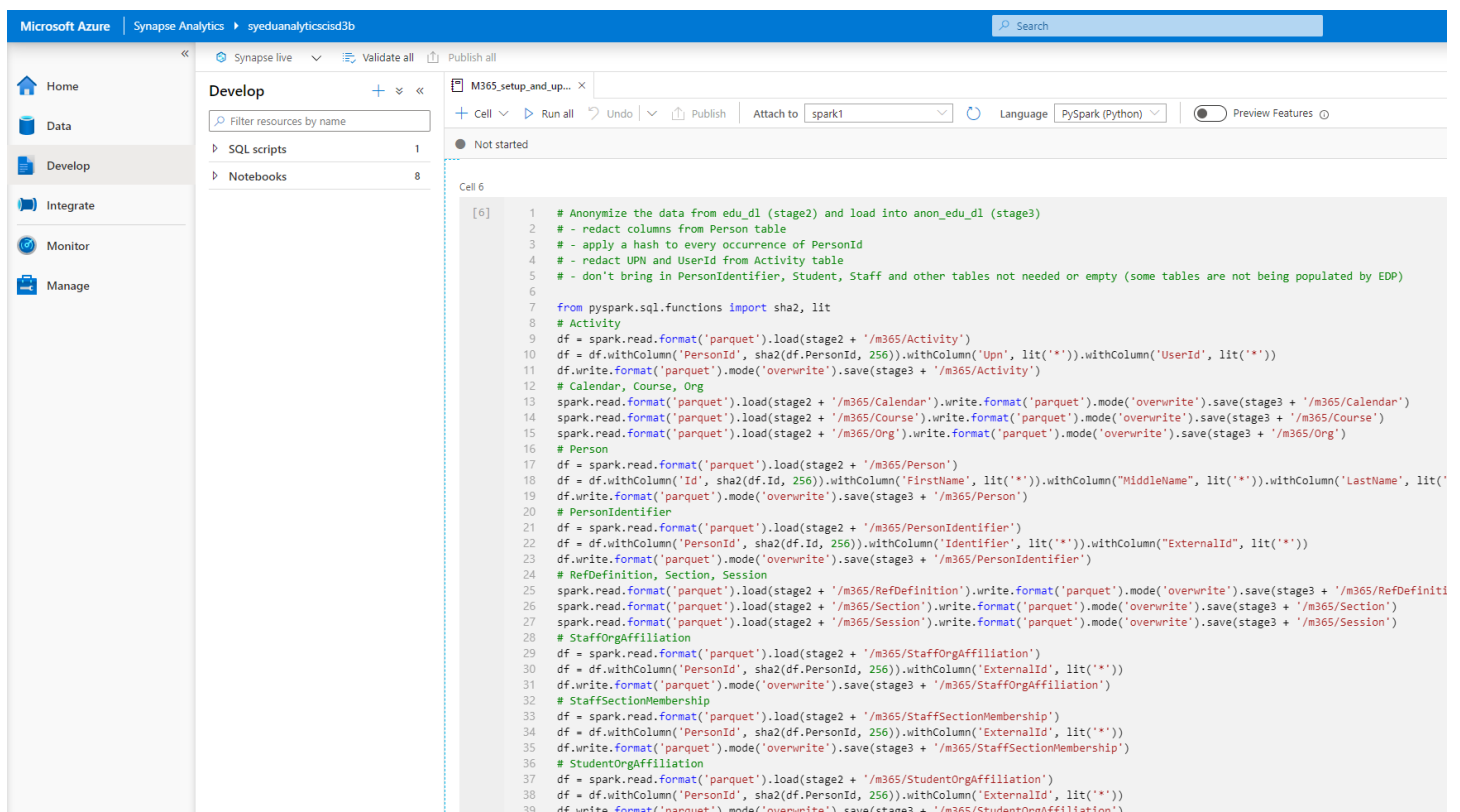
6) Privacy and Security

At the storage level, data protection comes from [Azure's automatic data encryption](#). Data in the data lake is automatically encrypted as it is written to storage using 256-bit AES encryption, and automatically decrypted as it is read from storage from an authorized source.

Security of access is provided at the storage level through the use of [Security Groups within Azure Active Directory](#), allowing the Global Admin to grant the minimum access necessary for specific groups of users to specific zones within the data lake based on the access needed for a given use case. Through this use of role-based access control (RBAC) at the storage level, access is controlled regardless of the tools used to query or analyze the data. Furthermore, using RBAC to set up the minimum access necessary for a new group of users or for a specific use case is straightforward and easily maintained. Additional permissions can be set at the SQL level for finer-grained control over access. See [Securing access to ADLS files using Synapse SQL permission model](#) for more info.

In the Open Edu Analytics architecture, data privacy is guarded by first reducing what data is made available - that is, reducing the data set to that which is needed for a given use case. In addition, the data is either pseudonymized or anonymized to protect personally identifiable information (PII). Pseudonymization is [defined in the GDPR](#) as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately", while anonymization is the processing of personal data such that the personal data can no longer be attributed to a specific data subject, even with the use of additional information.

The process of pseudonymization or anonymization is performed through a pyspark script which reads in non-anonymized data from the data lake and explicitly obscures fields with PII and performs a one-way hash on ID's and writes the anonymized data to a stage3 container in the data lake. An excerpt from the script [OpenEduAnalytics/modules/M365/notebooks/M365_setup_and_update.ipynb](#) that is responsible for the anonymization of the M365 data is shown below.



The screenshot displays the Microsoft Azure Synapse Analytics environment. The left sidebar shows navigation options: Home, Data, Develop, Integrate, Monitor, and Manage. The 'Develop' section is active, showing a list of resources: SQL scripts (1) and Notebooks (8). The main workspace shows a Jupyter Notebook titled 'M365_setup_and_update.ipynb'. The code in the notebook is as follows:

```
[6]
1 # Anonymize the data from edu_d1 (stage2) and load into anon_edu_d1 (stage3)
2 # - redact columns from Person table
3 # - apply a hash to every occurrence of PersonId
4 # - redact UPN and UserId from Activity table
5 # - don't bring in PersonIdentifier, Student, Staff and other tables not needed or empty (some tables are not being populated by EDP)
6
7 from pyspark.sql.functions import sha2, lit
8 # Activity
9 df = spark.read.format('parquet').load(stage2 + '/m365/Activity')
10 df = df.withColumn('PersonId', sha2(df.PersonId, 256)).withColumn('Upn', lit('')).withColumn('UserId', lit(''))
11 df.write.format('parquet').mode('overwrite').save(stage3 + '/m365/Activity')
12 # Calendar, Course, Org
13 spark.read.format('parquet').load(stage2 + '/m365/Calendar').write.format('parquet').mode('overwrite').save(stage3 + '/m365/Calendar')
14 spark.read.format('parquet').load(stage2 + '/m365/Course').write.format('parquet').mode('overwrite').save(stage3 + '/m365/Course')
15 spark.read.format('parquet').load(stage2 + '/m365/Org').write.format('parquet').mode('overwrite').save(stage3 + '/m365/Org')
16 # Person
17 df = spark.read.format('parquet').load(stage2 + '/m365/Person')
18 df = df.withColumn('Id', sha2(df.Id, 256)).withColumn('FirstName', lit('')).withColumn('MiddleName', lit('')).withColumn('LastName', lit(''))
19 df.write.format('parquet').mode('overwrite').save(stage3 + '/m365/Person')
20 # PersonIdentifier
21 df = spark.read.format('parquet').load(stage2 + '/m365/PersonIdentifier')
22 df = df.withColumn('PersonId', sha2(df.Id, 256)).withColumn('Identifier', lit('')).withColumn('ExternalId', lit(''))
23 df.write.format('parquet').mode('overwrite').save(stage3 + '/m365/PersonIdentifier')
24 # RefDefinition, Section, Session
25 spark.read.format('parquet').load(stage2 + '/m365/RefDefinition').write.format('parquet').mode('overwrite').save(stage3 + '/m365/RefDefinition')
26 spark.read.format('parquet').load(stage2 + '/m365/Section').write.format('parquet').mode('overwrite').save(stage3 + '/m365/Section')
27 spark.read.format('parquet').load(stage2 + '/m365/Session').write.format('parquet').mode('overwrite').save(stage3 + '/m365/Session')
28 # StaffOrgAffiliation
29 df = spark.read.format('parquet').load(stage2 + '/m365/StaffOrgAffiliation')
30 df = df.withColumn('PersonId', sha2(df.PersonId, 256)).withColumn('ExternalId', lit(''))
31 df.write.format('parquet').mode('overwrite').save(stage3 + '/m365/StaffOrgAffiliation')
32 # StaffSectionMembership
33 df = spark.read.format('parquet').load(stage2 + '/m365/StaffSectionMembership')
34 df = df.withColumn('PersonId', sha2(df.PersonId, 256)).withColumn('ExternalId', lit(''))
35 df.write.format('parquet').mode('overwrite').save(stage3 + '/m365/StaffSectionMembership')
36 # StudentOrgAffiliation
37 df = spark.read.format('parquet').load(stage2 + '/m365/StudentOrgAffiliation')
38 df = df.withColumn('PersonId', sha2(df.PersonId, 256)).withColumn('ExternalId', lit(''))
39 df.write.format('parquet').mode('overwrite').save(stage3 + '/m365/StudentOrgAffiliation')
```

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