

# Data & AI Tech Immersion Workshop – Product Review Guide and Lab Instructions

## Day 1, Experience 1 - Handling Big Data with SQL Server 2019 Big Data Clusters

- [Data & AI Tech Immersion Workshop – Product Review Guide and Lab Instructions](#)
  - [Day 1, Experience 1 - Handling Big Data with SQL Server 2019 Big Data Clusters](#)
  - [Technology overview](#)
  - [Scenario overview](#)
  - [Experience requirements](#)
  - [Before the lab: Connecting to SQL Server 2019](#)
    - [Connect with Azure Data Studio](#)
    - [Connect with SQL Server Management Studio](#)
  - [Task 1: Query and join data from flat files, data from external database systems, and SQL Server](#)
  - [Task 2: Train a machine learning model, score and save data as external table](#)
  - [Task 3: Identify PII and GDPR-related compliance issues using Data Discovery & Classification in SSMS](#)
  - [Task 4: Fix compliance issues with dynamic data masking](#)
  - [Wrap-up](#)
  - [Additional resources and more information](#)

### Technology overview

SQL Server 2019 brings innovative security and compliance features, industry leading performance, mission-critical availability, and advanced analytics to all data workloads, now with support for big data built-in.

SQL Server 2019 is a hub for data integration. Data virtualization allows queries across relational and non-relational data without movement or replication. The enhanced PolyBase feature of SQL Server 2019 is able to connect to Hadoop clusters, Oracle, Teradata, MongoDB, and more.

Customers will be able to deliver transformational insights over structured and unstructured data with the power of SQL Server, Hadoop and Spark. SQL Server 2019 big data clusters offer scalable compute and storage composed of SQL Server, Spark and HDFS. Big data clusters will also cache data in scale-out data marts.

SQL Server 2019 is a complete AI platform to train and operationalize R and Python models in SQL Server Machine Learning Services or Spark ML using Azure Data Studio notebooks.

SQL Server 2019 will give customers and ISVs the choice of programming language and platform. They will be able to build modern applications with innovative features using .NET, PHP, Node.js, Java, Python, Ruby, and more – and deploy the application on either Windows, Linux, or containers both on-premises and in the cloud. Application developers are now able to run Java code on SQL Server and store and analyze graph data.

SQL Server 2019 allows customers to run real-time analytics on operational data using HTAP (Hybrid Transactional and Analytical Processing), leverage the in-memory technologies for faster transactions and analytical queries, and get higher concurrency and scale through persistent memory.

Intelligent Query Processing features in SQL Server 2019 improve scaling of queries, and Automatic Plan Correction resolves performance problems.

SQL Server 2019 enables several layers of security including protection of computations in Always Encrypted secure enclaves. Customers can track compliance with sophisticated tools such as Data Discovery & Classification labeling for GDPR and Vulnerability Assessment tool.

For High Availability and Disaster Recovery, SQL Server 2019 now supports up to eight secondary replicas in an Always On Availability Group. Customers can also run Always On Availability Groups on containers using Kubernetes.

SQL Server 2019 also has powerful tools for Business Intelligence including Analysis Services and Power BI Report Server which provide visual data exploration and interactive analysis of business data.

### Scenario overview

Contoso Auto stores data in several data stores, including relational databases, NoSQL databases, data warehouses, and unstructured data stored in a data lake. They have heard of data virtualization in SQL Server 2019, and are interested to see whether this feature will allow them to more easily access their data stored in these disparate locations. They have heard of the new Big Data Clusters that can be scaled out to handle their Big Data workloads, including machine learning tasks and advanced analytics. They are also interested in any performance improvements against their internal SQL tables by moving to 2019, since the overall amount of data is growing at a rapid pace.

This experience will highlight the new features of SQL Server 2019 with a focus on Big Data Clusters and data virtualization. You will gain hands-on experience with querying both structured and unstructured data in a unified way using T-SQL. This capability will be illustrated by joining different data sets, such as product stock data in flat CSV files in Azure Storage, product reviews stored in Azure SQL Database, and transactional data in SQL Server 2019 for exploratory data analysis within Azure Data Studio. This joined data will be prepared into a table used for reporting, highlighting query performance against this table due to intelligent query processing. With the inclusion of Apache Spark packaged with Big Data Clusters, it is now possible to use Spark to train machine learning models over data lakes and use those models in SQL Server in one system. You will learn how to use Azure Data Studio to work with Jupyter notebooks to train a simple model that can predict vehicle battery lifetime, score new data and save the result as an external table. Finally, you will experience the data security and compliance features provided by SQL Server 2019 by using the Data Discovery & Classification tool in SSMS to identify tables and columns with PII and GDPR-related compliance issues, then address the issues by layering on dynamic data masking to identified columns.

### Experience requirements

Before you begin this lab, you need to find the following information on the Tech Immersion Mega Data & AI Workshop On Demand Lab environment details page:

- Server IP address and port number: `SQL SERVER_2019_CLUSTER URL`
- SQL username: `SALES DB USERNAME`
- SQL password: `SALES DB PASSWORD`

- Sales database name (your unique copy): SALES DB
- Azure SQL Database server: AZURE DATABASE SERVER
- Azure SQL Database name: DATABASE NAME
- Azure SQL Database username: DATABASE USER
- Azure SQL Database password: DATABASE PASSWORD



# Tech Immersion Mega Data & AI Workshop

By : Microsoft

Overview

Lab

Your On Demand Lab is ready (47 hour(s), 15 minute(s) remaining)

Environment Details

Virtual Machines

## Azure Credentials

Here are your credentials to login to [Microsoft Azure](#) and access the lab environment.

Username

odl\_user\_58735@cloudbreak

Password

EVENTHUB NAME

tech-immersion-eventhub

AZURE DATABASE SERVER

tech-immersion-sqls

DATABASE NAME

CA\_Commerce

DATABASE USER

SqlReader

DATABASE PASSWORD

MySQLBigData2019

CLUSTER URL

Password.123456

SQL SERVER\_2019\_CLUSTER URL

52.247.206.120,3143

HDFS/KNOX

52.175.197.186,3044

SALES DB

Sales\_58735

SALES DB USERNAME

sa\_58735

SALES DB PASSWORD

Lab Guide : <https://github.com/solliancenet/tech-immersion-data-a>

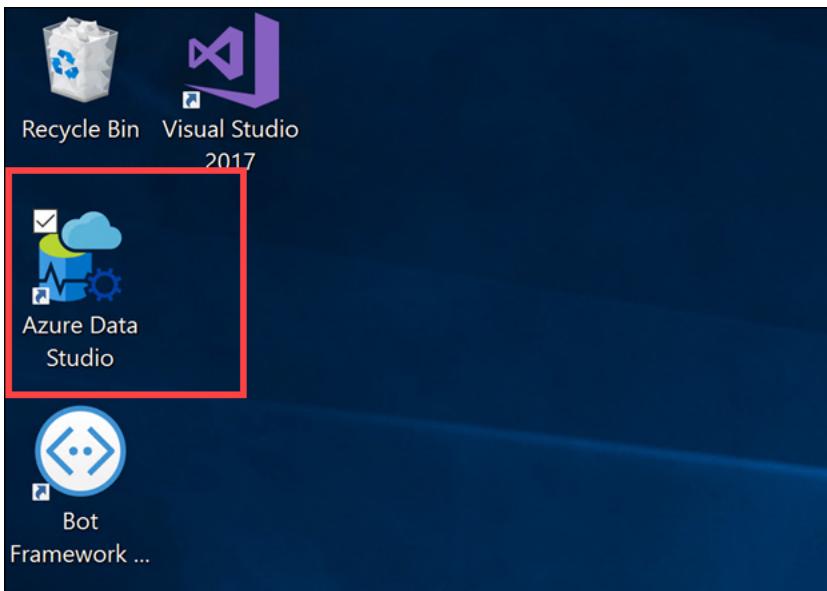
Screenshot of the lab environment details

## Before the lab: Connecting to SQL Server 2019

Follow the steps below to connect to your SQL Server 2019 cluster with both Azure Data Studio and SQL Server Management Studio (SSMS).

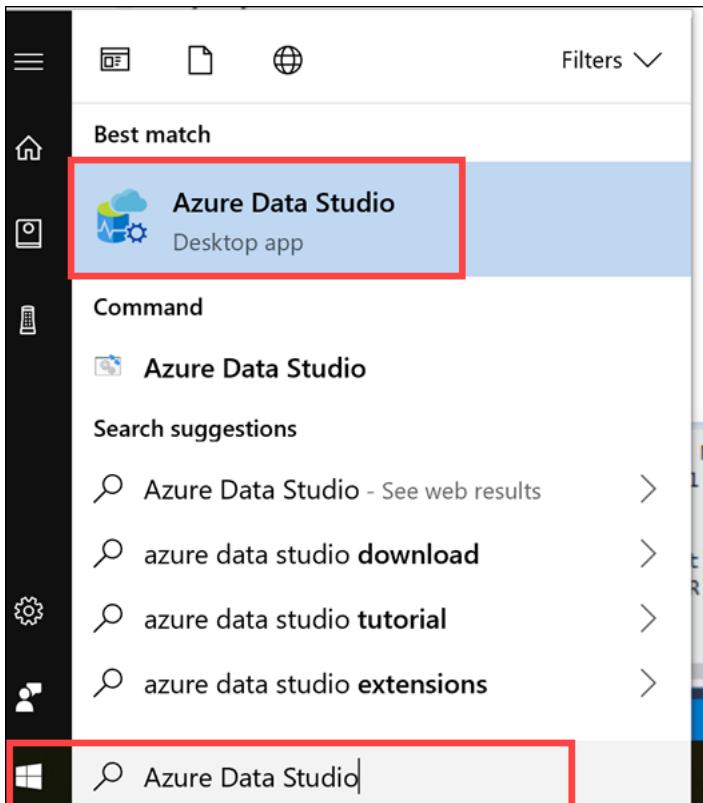
### Connect with Azure Data Studio

A link to Azure Data Studio should already be on the desktop of the VM. If not, follow the instructions in Step 1 below.



Azure Data Studio is highlighted on the desktop.

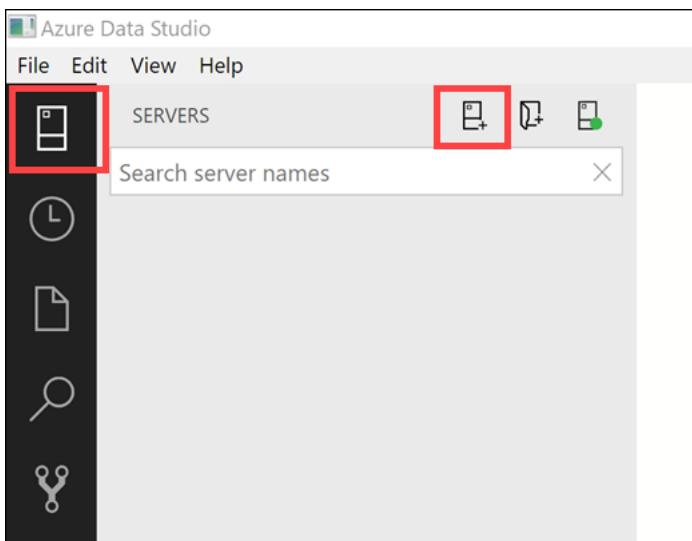
1. On the bottom-left corner of your Windows desktop, locate the search box next to the Start Menu. Type **Azure Data Studio**, then select the Azure Data Studio desktop app in the search results.



The search box has “Azure Data Studio” entered into it and the desktop app is highlighted in the results.

**Please note:** If Azure Data Studio prompts you to update, please **do not apply** the update at this time. The lab has been tested with the software and library versions loaded in the provided environment.

2. Within Azure Data Studio, select **Servers** from the top of the left-hand menu, then select **New Connection** from the top toolbar to the right of the menu.



The Servers menu icon is selected, as well as the new connection icon.

3. Within the Connection dialog, configure the following:

- **Connection type:** Select Microsoft SQL Server.
- **Server:** Enter the IP address, followed by port number 31433. For example: 123.123.123.123,31433.
- **Authentication type:** Select SQL Login.
- **Username:** Enter “sa”.
- **Password:** Enter the password provided to you for this lab.
- **Remember password:** Checked.
- Leave all other options at their default values.

**Connection**

**Recent Connections** **Saved Connections**

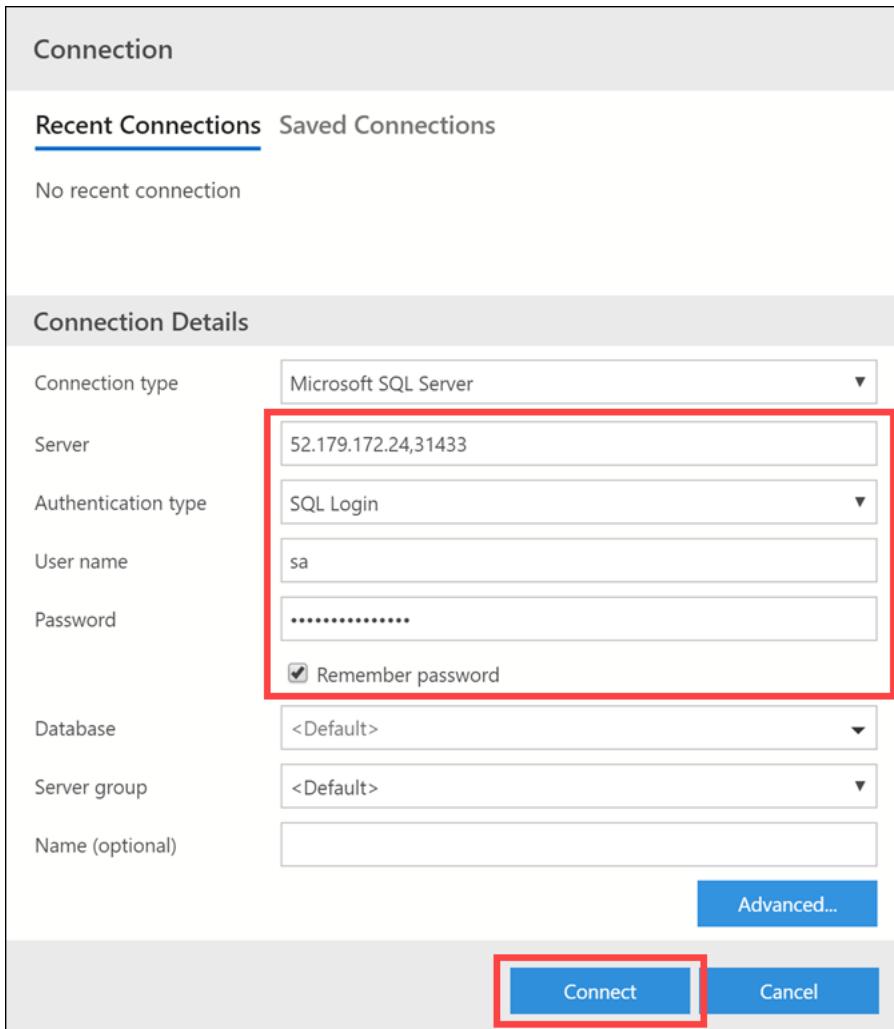
No recent connection

**Connection Details**

Connection type	Microsoft SQL Server
Server	52.179.172.24,31433
Authentication type	SQL Login
User name	sa
Password	*****
<input checked="" type="checkbox"/> Remember password	
Database	<Default>
Server group	<Default>
Name (optional)	

**Advanced...**

**Connect** **Cancel**



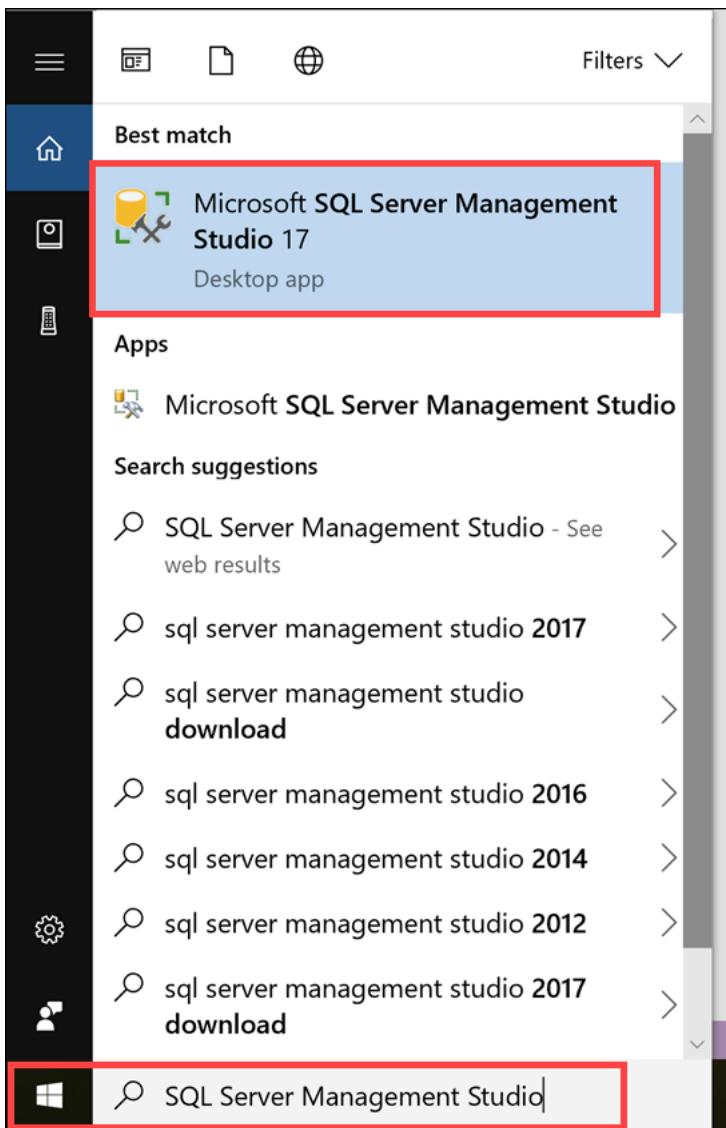
The Connection form is filled out with the previously mentioned settings entered into the appropriate fields.

4. Click **Connect**.

### Connect with SQL Server Management Studio

The version of SQL Server Management Studio (SSMS) used in this lab is v17.x. There is a [newer preview version \(v18\)](#) that includes some SQL Server 2019 features, which is not included in the provided environment.

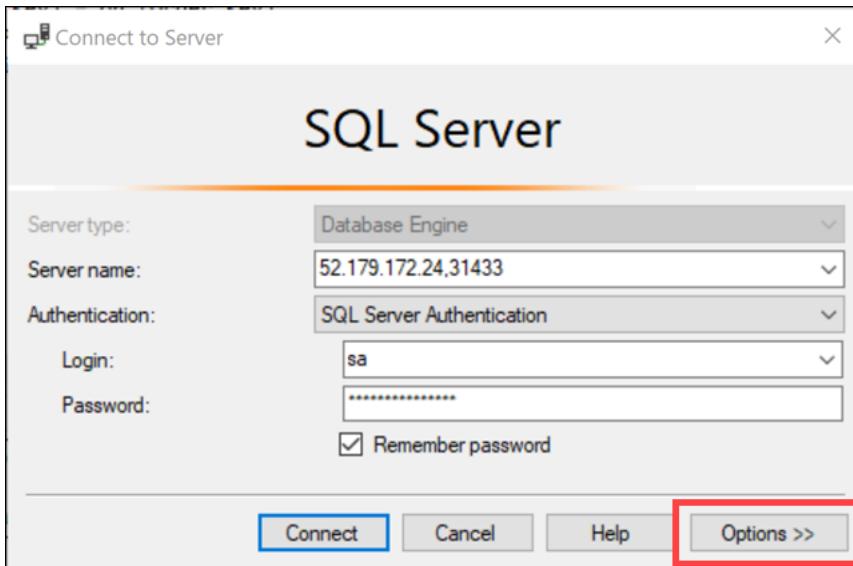
1. On the bottom-left corner of your Windows desktop, locate the search box next to the Start Menu. Type **SQL Server Management Studio**, then select the SQL Server Management Studio desktop app in the search results.



The search box has “SQL Server Management Studio” entered into it and the desktop app is highlighted in the results.

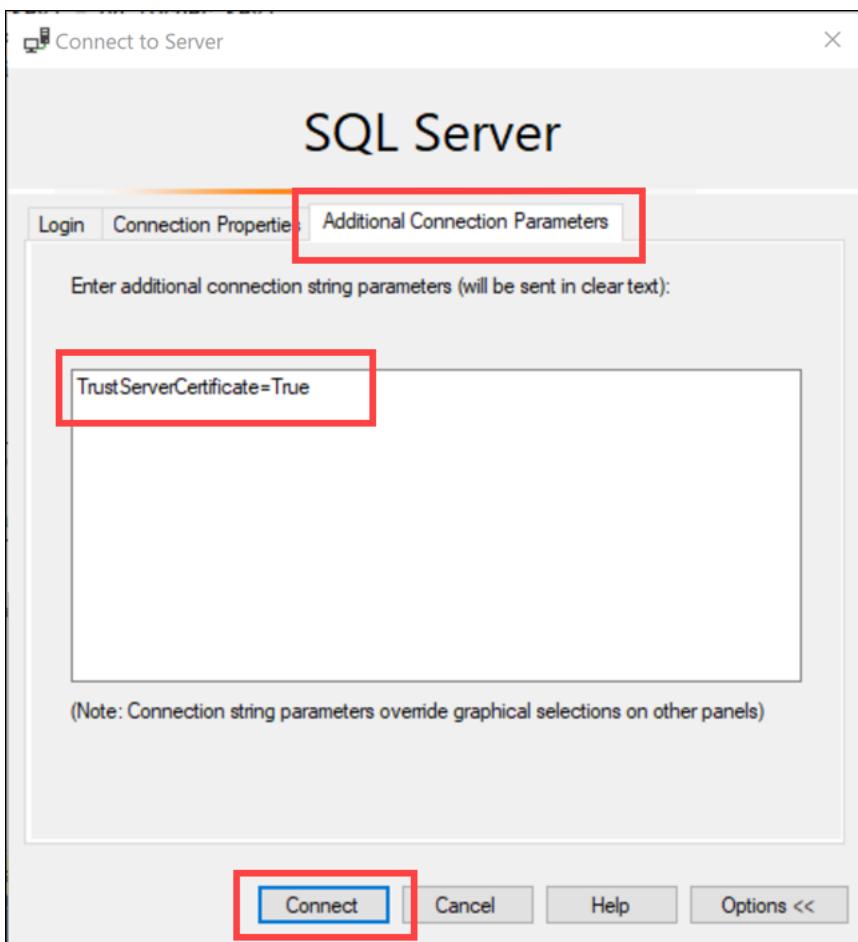
2. Within the Connection dialog that appears, configure the following:

- **Server name:** Enter the IP address, followed by port number 31433. For example: 123.123.123.123,31433.
- **Authentication:** Select SQL Server Authentication.
- **Login:** Enter “sa”.
- **Password:** Enter the password provided to you for this lab.
- **Remember password:** Checked.



The Connect form is filled out with the previously mentioned settings entered into the appropriate fields.

3. Click **Options >>**
4. Select the **Additional Connection Parameters** tab. In the text area below, enter `TrustServerCertificate=True`. This is needed because the server certificates are dynamically generated for the Big Data Clusters, and are self-signed.

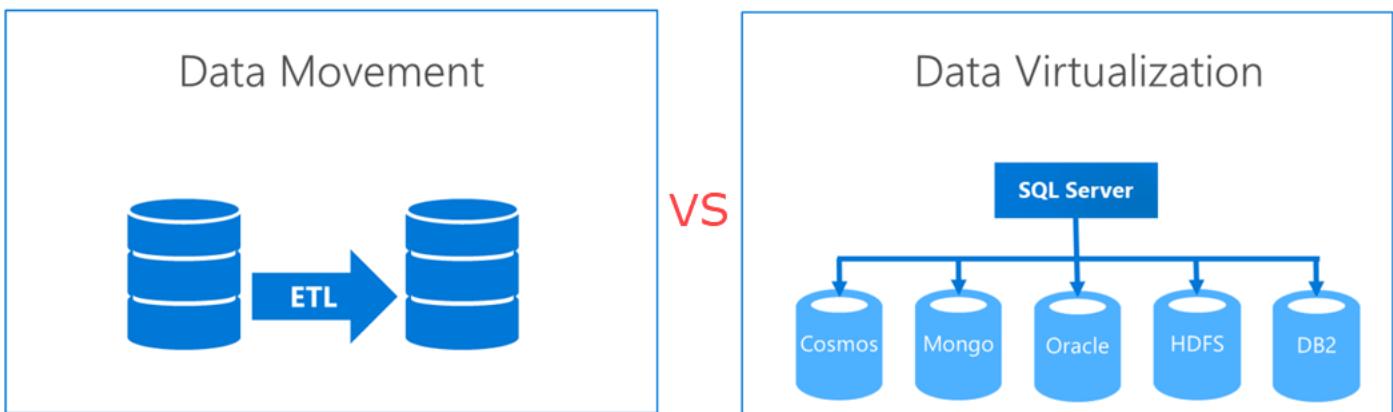


The Additional Connection Parameters tab is selected and the TrustServerCertificate=True value is highlighted.

5. Click **Connect**.

### Task 1: Query and join data from flat files, data from external database systems, and SQL Server

One of the key new features of SQL Server 2019 is data virtualization. What this means is that you can *virtualize* external data in a SQL Server instance, regardless of source, location, and format, so that it can be queried like any other table, or sets of tables, within your SQL Server instance. In essence, data virtualization helps you create a single “virtual” layer of data from these disparate sources, providing unified data services to support multiple applications and users. A more familiar term we could use is data lake, or perhaps data hub. Unlike a typical data lake, however, you do not have to move data out from where it lives, yet you can still query that data through a consistent interface. This is a huge advantage over traditional ETL (extract-transform-load) processes where data must be moved from its original source to a new destination, oftentimes with some data transformation or mapping. This causes delays, extra storage, additional security, and a fair amount of engineering in most cases. With data virtualization, no data movement is required, which means the data sets are up-to-date, and it is possible to query and join these different data sources through these new capabilities, thanks to the use of new [PolyBase](#) connectors. The data sources you can connect to include Cosmos DB, SQL Server (including Azure SQL Database), Oracle, HDFS (for flat files), and DB2.



Data Virtualization compared to traditional ETL.

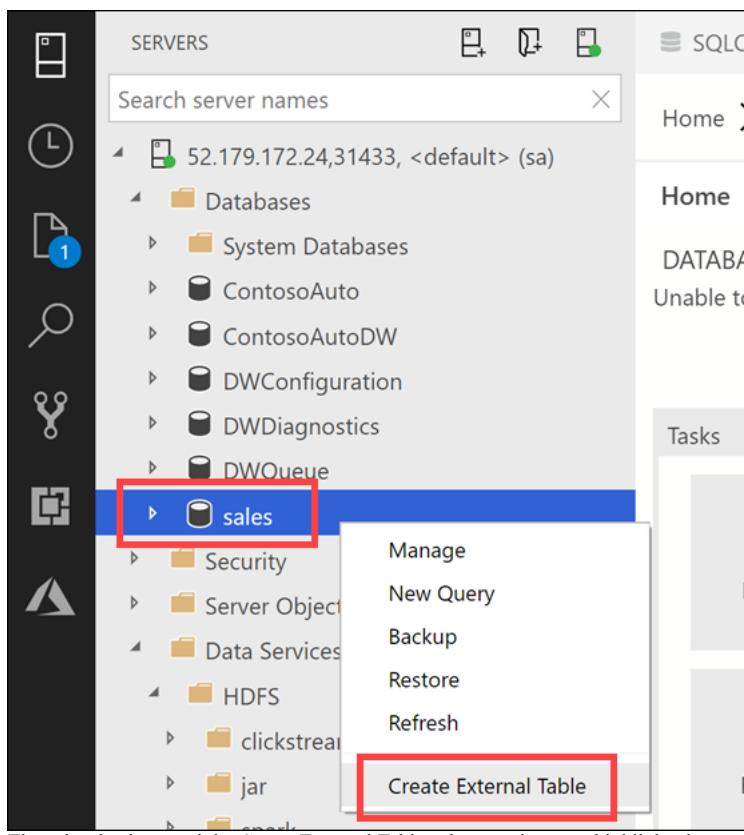
The image to the left represents traditional data movement using ETL. Compare that to data virtualization, which does not require data movement and provides a unified layer over top of existing data sources.

In this task, you will experience how to configure data virtualization in SQL Server 2019 by joining data in a single query from a SQL Server 2019 table, an external file stored in HDFS (Hadoop Filesystem), and an external database.

Learn more about using data virtualization with [relational data sources](#) and [CSV files](#), using the External Table Wizard.

To start, we will use the External Table Wizard in Azure Data Studio to connect to an external Azure SQL Database.

1. Open Azure Data Studio and connect to your SQL Server 2019 cluster, following the [connection steps](#) above.
2. Expand the Databases folder, right-click on the **sales\_YOUR\_UNIQUE\_IDENTIFIER** database, then select **Create External Table**. The **YOUR\_UNIQUE\_IDENTIFIER** portion of the name is the unique identifier assigned to you for this lab.



The sales database and the Create External Table sub-menu item are highlighted.

3. Select the **SQL Server** data source type, then click **Next**.

### Create External Table

Step 1  
Select a Data Source

Select the destination database for your external table [?](#)

▼

Select your data source type

  
SQL Server

  
Oracle

**Next** **Cancel**



The SQL Server data source type is selected.

4. The next step is to create a database master key, if it does not already exist. This secures the credentials used by an external data source. Enter **MySecure@MasterKey1** in the **Password** and **Confirm Password** fields. If you see a message stating that a master key already exists, you may skip this step. Click **Next**.

### Create External Table

**① Information**  Co

A Master Key already exists for the selected database. No action is required on this page.

Step 2  
Create Database Master Key

A master key is required. This secures the credentials used by an External Data Source. Note that you should back up the master key by BACKUP MASTER KEY and store the backup in a secure, off-site location.

**Set the Master Key password.**

Password \*

Confirm Password \*

Strong passwords use a combination of alphanumeric, upper, lower, and special characters.

**Previous** **Next**



The Master Key step is displayed.

5. Now, enter the credentials provided to you for the **CA\_Commerce** Azure SQL Database within the following fields:

- **External Data Source Name:** Enter “SQLReviews”.
- **Server Name:** Enter the value provided to you for the Azure SQL Server name. The name should end with `.database.windows.net` (lab environment details page: AZURE DATABASE SERVER).
- **Database Name:** Enter “CA\_Commerce”.
- **Choose Credential:** Select “`-- Create New Credential --`”.
- **New Credential Name:** Enter “SQLCred”.
- **Username:** Enter the Azure SQL Server username provided to you for this lab (lab environment details page: DATABASE USER).
- **Password:** Enter the Azure SQL Server password provided to you for this lab (lab environment details page: DATABASE PASSWORD).

**Create External Table**

Step 3

Create a connection to your Data Source

External Data Source Name \* ② SQLReviews

Server Connection

Server Name \* contosoauto-server.database.windows.net

Database Name CA\_Commerce

Configure Credential

Choose Credential \* ② -- Create New Credential --

New Credential Name \* SQLCred

Username \* causer

Password \* [REDACTED]

Previous Next

The external data source connection form is filled out with the previously mentioned settings entered into the appropriate fields.

6. Click **Next**. This process will take a few moments while the External Table Wizard attempts to connect to your data source.
7. The next screen allows you to configure external table mapping and select the tables for which you want to create external views. Expand the **CA\_Commerce** database node, then expand Tables, and check the box next to the **dbo.Reviews** table. Click on the table name to highlight it as well. It is here where you can rename the external table if you wish. For now, just click **Next**.

**Create External Table**

The external tables are listed and the Reviews table is checked and selected.

8. In the Summary page that follows, you can see the name of the database scoped credential and external data source objects to be created in the destination database. Here you can click Generate Script to view the SQL script that will run to create the external table. Instead, click **Create**.

**Create External Table**

Step 5

Summary

Destination Database:

sales

The following objects will be created in the destination database:

Object type	Name
Database Scoped Credential	SQLCred
External Data Source	SQLReviews
External Table	dbo.Reviews

[Previous](#) [Generate script](#) [Create](#) [Cancel](#)

A screenshot of the summary is displayed.

9. After a few moments, a “Create External Table succeeded” message will display.

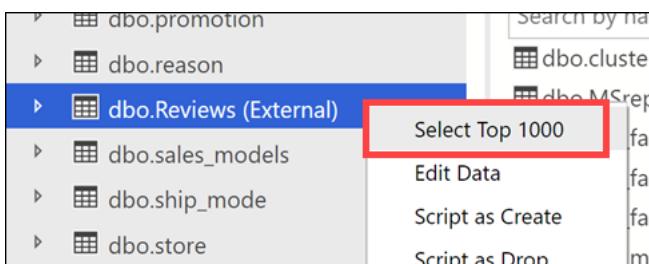


The Create External Table succeeded message is displayed.

10. Select the Servers link (Ctrl+G) on the left-hand menu, then expand the Tables list underneath your **sales\_YOUR\_UNIQUE\_IDENTIFIER** database and find the **dbo.Reviews (External)** table. If you do not see it, right-click on the Tables folder, then select Refresh. The "(External)" portion of the table name denotes that it is a virtual data object that was added as an external table.

The Reviews external table is displayed in the sales tables list.

11. Right-click the **dbo.Reviews (External)** table, then select the **Select Top 1000** menu option to display the table contents.



The Select Top 1000 rows menu item is highlighted.

12. You should see a SQL query selecting the top 1000 records from the Reviews table and its results. The interesting thing to note is that the query selects the table and fields using the same syntax you would use to select from any other table in the sales database. The fact that the Reviews table is external is completely seamless and transparent to the user. This is the power of data virtualization in SQL Server 2019.

The screenshot shows the Azure Data Studio interface. At the top, there are buttons for Run, Cancel, Disconnect, Change Connection, and a dropdown menu set to 'sales'. To the right of the dropdown is an 'Explain' button. Below the toolbar, a code editor displays the following T-SQL query:

```

1  SELECT TOP (1000) [product_id]
2      ,[customer_id]
3      ,[review]
4      ,[date_added]
5  FROM [sales].[dbo].[Reviews]

```

Below the code editor is a 'RESULTS' section containing a table with five rows of data. The columns are labeled: product\_id, customer\_id, review, and date\_added. The data is as follows:

	product_id	customer_id	review	date_added
1	9308	67028	Laguiole kniv...	2019-02-24 1...
2	2430	89770	Good sound t...	2019-02-24 2...
3	16203	84679	AWESOME FE...	2019-02-24 2...
4	5239	84953	love the retro ...	2019-02-24 2...
5	17432	72621	Works fine. Ea...	2019-02-22 0...

On the right side of the results table, there are several icons for data manipulation: a green plus sign, a blue minus sign, a red exclamation mark, a blue arrow, and a yellow bar chart.

Below the results table is a 'MESSAGES' section with the following log entries:

- 11:05:10 PM Started executing query at Line 1
- (20 rows affected)
- Total execution time: 00:00:01.566

At the bottom of the interface, there is a status bar with the following information: 20 rows, 00:00:01, 52.179.172.24,31433 : sales, Ln 5, Col 31, Spaces: 4, UTF-8, CRLF, SQL, MSSQL, a smiley face icon, and a bell icon.

The Reviews query and results are displayed.

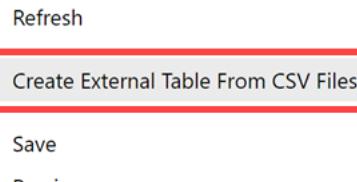
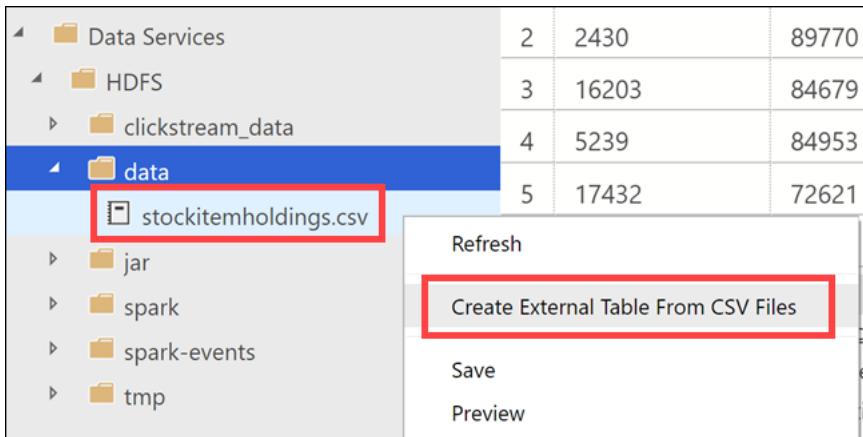
```

SELECT TOP (1000) [product_id]
    ,[customer_id]
    ,[review]
    ,[date_added]
FROM [sales_YOUR_UNIQUE_IDENTIFIER].[dbo].[Reviews]

```

13. The next data source we will be virtualizing is a CSV file that was uploaded to HDFS.

14. Within Azure Data Studio, scroll down below the list of SQL Server 2019 databases to find the **Data Services** folder. Expand that folder, expand the **HDFS** folder, then expand the **data** subfolder. Right-click on the **stockitemholdings.csv** file, then select **Create External Table From CSV Files**.



The CSV file and the Create External Table From CSV Files menu item are highlighted.

15. The first dialog has you select the SQL Server Master instance containing your Big Data Cluster. Select the connection underneath **Active SQL Server connections** that includes the cluster's IP address and the **sales\_YOUR\_UNIQUE\_IDENTIFIER** database name.

### Create External Table From CSV

Step 1  
Connect to the SQL Server Master Instance associated with your SQL Server big data cluster

1 Active SQL Server connections  
52.179.172.24,31433, sales (sa)

2 Server name \*

3 Username \*

4 Password \*

Next Cancel

The active SQL Server connection is highlighted.

16. Click **Next**.

17. In the destination database step, select the **sales** **YOUR\_UNIQUE\_IDENTIFIER** database underneath **Database the external table will be created in**. Leave the name and schema at their defaults, then click **Next**.

### Create External Table From CSV

Step 2  
Select the destination database for your external table

1 Source File  
/data/stockitemholdings.csv

2 Database the external table will be created in  
sales

3 Name for new external table \*  
stockitemholdings

4 Schema for new external table  
dbo

Previous Next Cancel

The sales database is selected and highlighted.

18. The next step displays a preview of the first 50 rows CSV data for validation. Click **Next** to continue.

## Create External Table From CSV



Step 3

### Preview Data

This operation analyzed the input file structure to generate the preview below for up to the first 50 rows.

StockItemID	QuantityOnHa...	BinLocation	LastStocktake...	LastCostPrice	ReorderLevel	TargetStockLe...	LastEd
1	175609	L-1	171341	9.5	20	100	16
2	165538	L-1	161435	9.5	20	100	3
3	253190	L-2	246900	11.25	10	120	3
4	208109	L-3	202964	12	5	100	3
5	199064	L-3	194162	12	5	100	3
6	196995	L-3	192127	12	5	100	3
7	205295	L-3	200201	12	5	100	16
8	412277	L-3	401980	88.5	10	200	16
9	192749	L-3	187968	12	5	100	3

[Previous](#)

[Next](#)

[Cancel](#)

A preview of the CSV data is displayed.

19. In the next step, you will be able to modify the columns of the external table you intend to create. You are able to alter the column name, change the data type, and allow for Nullable rows. For now, leave everything as-is and click **Next**.

## Create External Table From CSV

Step 4

Modify Columns

1  
2  
3  
4  
5

Column Name	Data Type	Allow Nulls
StockItemID	tinyint	<input type="checkbox"/>
QuantityOnHand	int	<input type="checkbox"/>
BinLocation	nvarchar(50)	<input type="checkbox"/>
LastStocktakeQuantity	int	<input type="checkbox"/>
LastCostPrice	float	<input type="checkbox"/>
ReorderLevel	tinyint	<input type="checkbox"/>
TargetStockLevel	smallint	<input type="checkbox"/>
LastEditedBy	tinyint	<input type="checkbox"/>
LastEditedWhen	nvarchar(50)	<input type="checkbox"/>

Previous    Next    Cancel

The Modify Columns step is displayed.

20. Verify that everything looks correct in the Summary step, then click **Create Table**.

## Create External Table From CSV

Step 5

Summary

1  
2  
3

Create External Table information

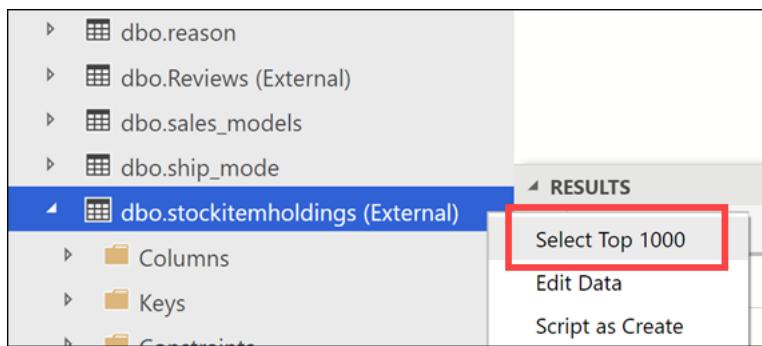
Object type	Name
Server name	52.179.172.24,31433
Database name	sales
Table name	stockitemholdings
Table schema	dbo
Source File	/data/stockitemholdings.csv

Previous    Generate script    Create Table    Cancel

The Summary step is displayed.

21. As with the previous external table you created, a “Create External Table succeeded” dialog will appear under your task history in a few moments. Select the Servers link (Ctrl+G) on the left-hand menu, then expand the Tables list underneath your **sales** database and find the **dbo.stockitemholdings (External)** table. If

you do not see it, right-click on the Tables folder, then select Refresh. **Right-click** the **dbo.stockitemholdings (External)** table, then select **Select Top 1000** from the context menu.



The Select Top 1000 rows menu item is highlighted.

22. Just as before, you should see a SQL query selecting the top 1000 rows and its query results, this time from the stockitemholdings table. Again, the SQL query is the same type of query you would write to select from a table internal to the sales database.

The screenshot shows the SSMS Query Editor window. The query in the editor is:

```
1  SELECT TOP (1000) [StockItemID]
2    ,[QuantityOnHand]
3    ,[BinLocation]
4    ,[LastStocktakeQuantity]
5    ,[LastCostPrice]
6    ,[ReorderLevel]
7    ,[TargetStockLevel]
8    ,[LastEditedBy]
9    ,[LastEditedWhen]
10   FROM [sales].[dbo].[stockitemholdings]
```

The results pane shows the first 7 rows of the query output:

	StockItemID	QuantityOnHand	BinLocation	LastStocktakeQ...	LastCostPrice	ReorderLevel	T...
1	1	175609	L-1	171341	9.5	20	
2	2	165538	L-1	161435	9.5	20	
3	3	253190	L-2	246900	11.25	10	
4	4	208109	L-3	202964	12	5	
5	5	199064	L-3	194162	12	5	
6	6	196995	L-3	192127	12	5	
7	7	205295	L-3	200201	12	5	

The messages pane shows the execution details:

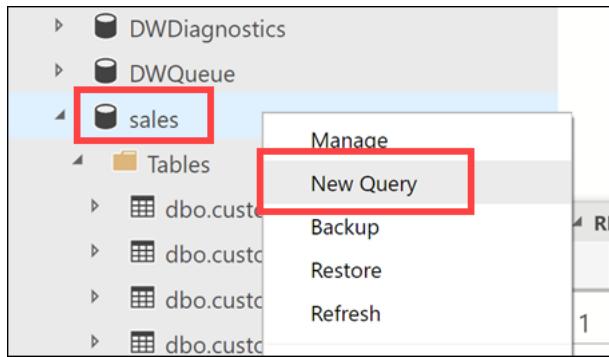
12:20:59 AM    Started executing query at Line 1  
(227 rows affected)  
Total execution time: 00:00:03.487

227 rows  00:00:03  52.179.172.24,31433 : sales  Ln 1, Col 1  Spaces: 4  UTF-8  CRLF  SQL  MSSQL

The stockitemholdings query and results are displayed.

```
SELECT TOP (1000) [StockItemID]
    ,[QuantityOnHand]
    ,[BinLocation]
    ,[LastStocktakeQuantity]
    ,[LastCostPrice]
    ,[ReorderLevel]
    ,[TargetStockLevel]
    ,[LastEditedBy]
    ,[LastEditedWhen]
FROM [sales]_YOUR_UNIQUE_IDENTIFIER].[dbo].[stockitemholdings]
```

23. Now that we have our two external tables added, we will now join those two external tables and two internal tables with a new SQL query to demonstrate how you can seamlessly combine all these data sources without having to copy any files or with separate queries or additional processing of that data. Right-click the sales\_YOUR\_UNIQUE\_IDENTIFIER database, then select New Query.

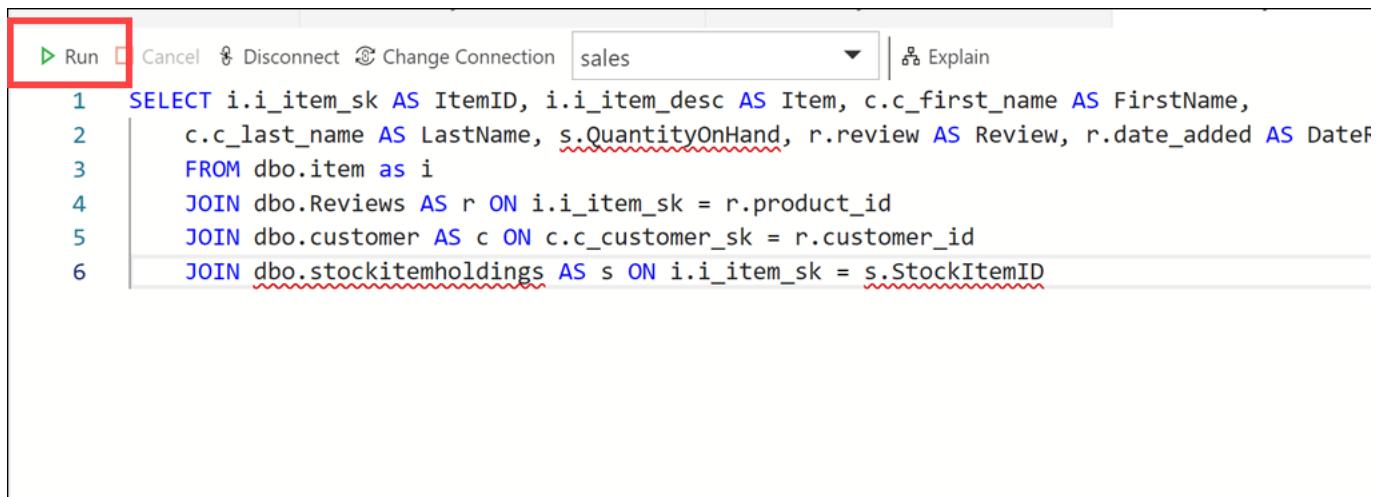


The sales database and New Query menu item are highlighted.

24. Paste the following into the new query window:

```
SELECT i.i_item_sk AS ItemID, i.i_item_desc AS Item, c.c_first_name AS FirstName,
       c.c_last_name AS LastName, s.QuantityOnHand, r.review AS Review, r.date_added AS DateReviewed
  FROM dbo.item as i
 JOIN dbo.Reviews AS r ON i.i_item_sk = r.product_id
 JOIN dbo.customer AS c ON c.c_customer_sk = r.customer_id
 JOIN dbo.stockitemholdings AS s ON i.i_item_sk = s.StockItemID
```

25. Click the Run button above the query window to execute.



The Run button above the query window is highlighted.

26. At the bottom of the query window, you will see results that include columns from the four data sources.

Run Cancel Disconnect Change Connection sales Explain

```

1  SELECT i.i_item_sk AS ItemID, i.i_item_desc AS Item, c.c_first_name AS FirstName
2    | c.c_last_name AS LastName, s.QuantityOnHand, r.review AS Review, r.date_added
3  FROM dbo.item as i
4  JOIN dbo.Reviews AS r ON i.i_item_sk = r.product_id
5  JOIN dbo.customer AS c ON c.c_customer_sk = r.customer_id
6  JOIN dbo.stockitemholdings AS s ON i.i_item_sk = s.StockItemID

```

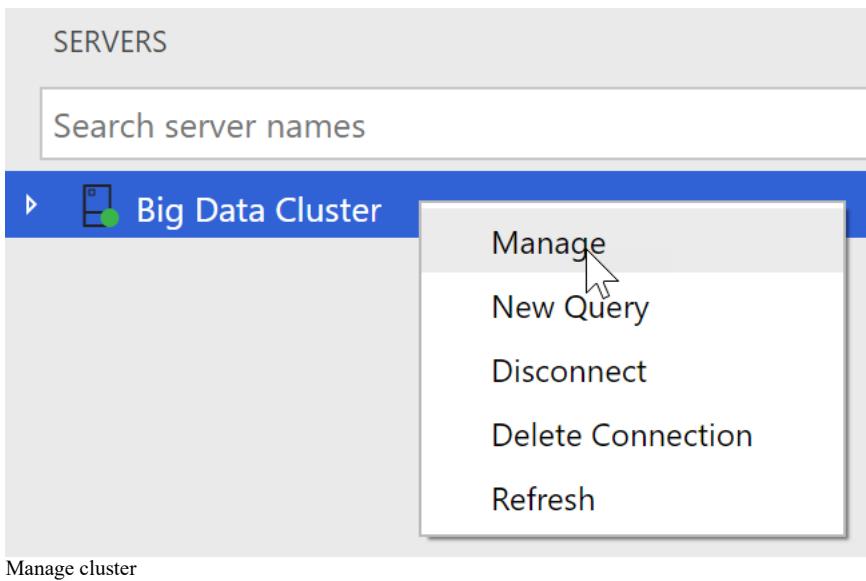
RESULTS			
	ItemID	Item	FirstName
1	17432	final blithe fray...	Karen
2	16816	never blithe wa...	Berneice
3	9342	ruthless final c...	Judy
4	10399	stealthy furious...	Isaac
5	7384	ironic grouch...	Cara
6	5123	patterns amon...	Eric
7	14908	warthogs abou...	Sara
8	11385	orbits must hin...	Virginia
9	6299	quiet permane...	Francis
10	5575	somas play per...	Arianna
11	15031	idle final platel...	Elmer
			LastName
			Bach
			Clifton
			Mendoza
			Coons
			Perez
			Cain
			Ramirez
			Slattery
			Perkins
			Patrick
			Tucker
			QuantityOnHand
			175609
			165538
			253190
			208109
			199064
			196995
			205295
			412277
			192749
			222572
			203148

Query results from the four data sets.

## Task 2: Train a machine learning model, score and save data as external table

In this task, you will use Azure Data Studio to execute a notebook that will enable you to train a model to predict the battery lifetime, apply the model to make batch predictions against a set of vehicle telemetry and save the scored telemetry to an external table that you can query using SQL.

1. Open Azure Data Studio and select Servers.
2. Right click your Big Data Cluster node select Manage.



3. In the window, select the SQL Big Data Cluster tab and then select the Open Notebook tile.

The screenshot shows the 'SQL Server Big Data Cluster' tab in Azure Data Studio. The 'Tasks' section contains five tiles:

- New Notebook (Icon: book with a green plus sign)
- New Spark Job (Icon: Spark logo)
- View Yarn History (Icon: yellow elephant)
- Open Notebook (Icon: book with a blue circular arrow)
- View Spark History (Icon: Spark logo)

Open notebook

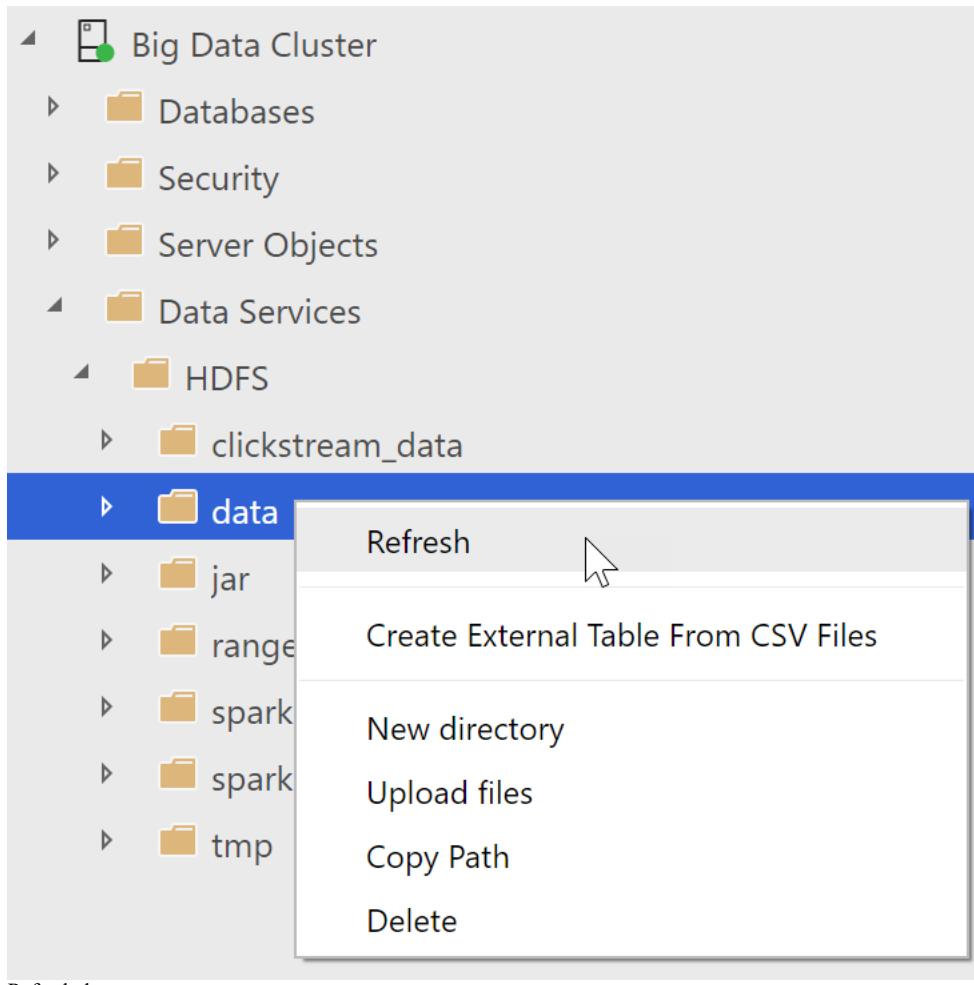
4. Browse to C:\files\1, select **predict-battery-life-with-sqlbdc.ipynb** and select Open.

5. Follow the instructions in the notebook and return to the next step after you have completed the notebook.

There may be a kernel error pertaining to there not being a valid SQL connection when you open the notebook. If this happens, close the notebook and Azure Data Studio, then re-launch, reconnect, then re-open the notebook.

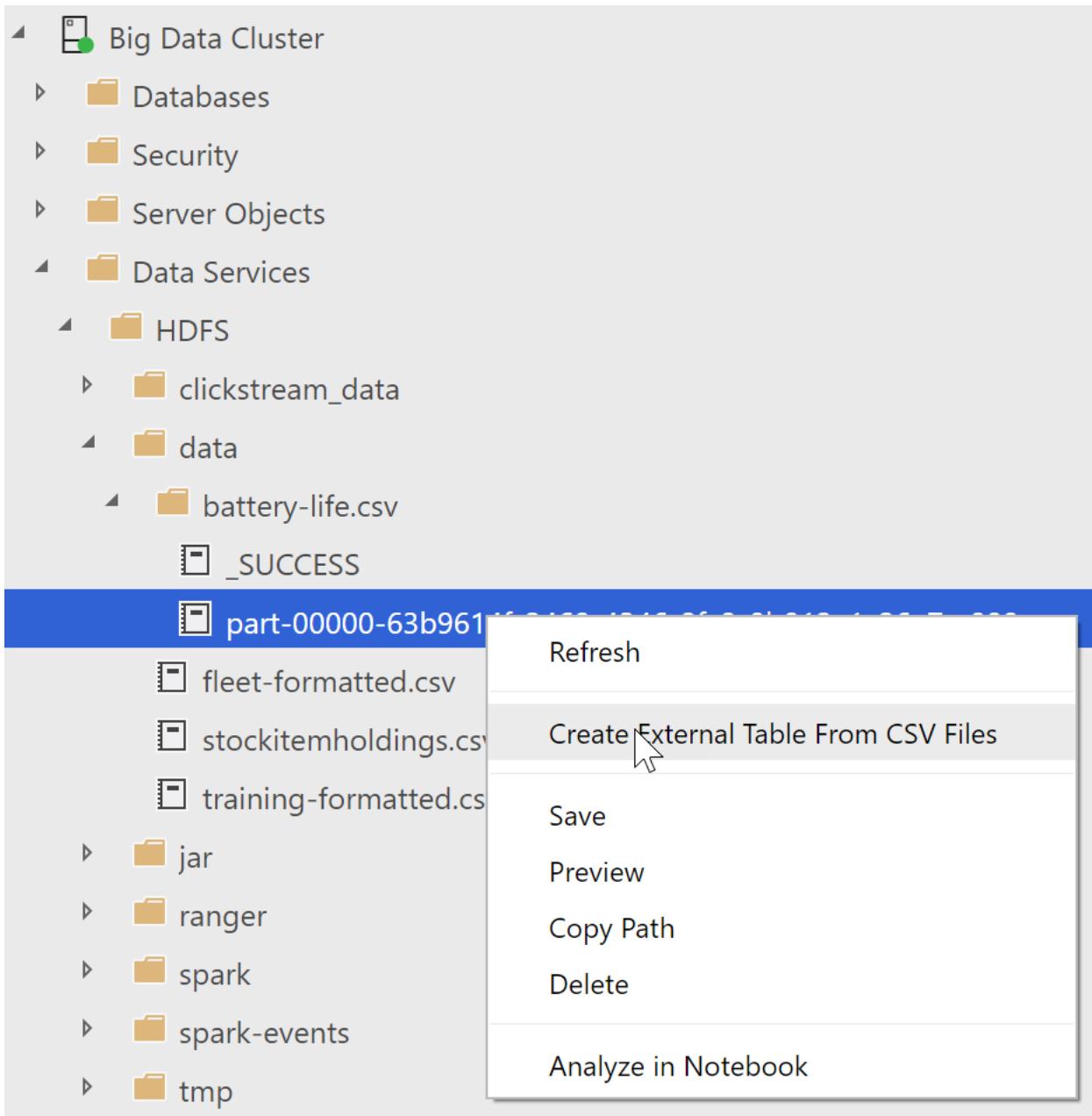
6. In Azure Data Studio, under Servers, expand your Big Data Cluster, Data Services, HDFS, data.

7. Right click the data folder and select Refresh to see the newly created folder.



Refresh data

8. You should see `battery-life-YOUR_UNIQUE_IDENTIFIER.csv` as a folder (where `YOUR_UNIQUE_IDENTIFIER` is your assigned identifier), expand it and then right click on the CSV file whose name starts with `part-00000-` and select `Create External Table From CSV Files`.



Create External Table

9. In Step 1 of the wizard, select your Active SQL Server connection to connect to your Big Data Cluster endpoint and select Next.

## Create External Table From CSV



### Step 1

Connect to the SQL Server Master Instance associated with

Active SQL Server connections

52.179.172.24,31433, <default> (sa)



Server name \*

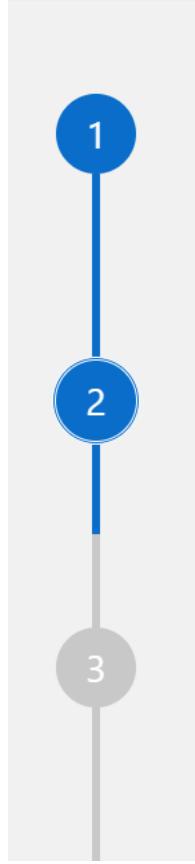
Username \*

Password \*

Select endpoint

10. In Step 2, select the sales\_YOUR\_UNIQUE\_IDENTIFIER database and for the Name for new external table field provide battery-life-predictions. Select Next.

## Create External Table From CSV



Select the destination database for your external table

Source File

/data/battery-life.csv/part-00000-63b9614f-2460-4346-8fc9-0b012a1c26e7-c000.csv

Database the external table will be created in

sales

Name for new external table \*

battery-life-predictions

Schema for new external table

dbo

Step 2

11. On Step 3, select Next.
12. On Step 4, for the column Car\_Has\_EcoStart set the Data Type to char(10). Select Next.

# Create External Table From CSV

Step 4

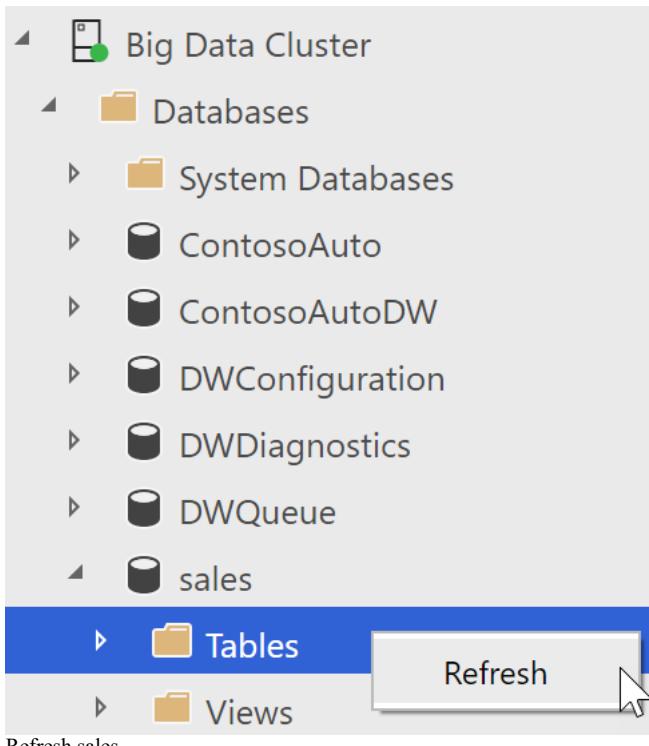
Modify Columns

1  
2  
3  
4

Step 4

Column Name	Data Type	Allow Nulls
Trip_Length_Mean	float	<input type="checkbox"/>
Trip_Length_Sigma	float	<input type="checkbox"/>
Trips_Per_Day_Mean	float	<input type="checkbox"/>
Trips_Per_Day_Sigma	float	<input type="checkbox"/>
Battery_Rated_Cycles	smallint	<input type="checkbox"/>
Alternator_Efficiency	float	<input type="checkbox"/>
Car_Has_EcoStart	char(10)	<input type="checkbox"/>

13. On Step 5, select `Create Table`. Your predictions are now available for SQL querying in the battery-life-predictions table in the sales database.
14. In Azure Data Studio, Servers, expand your Big Data Cluster, Databases, `sales_YOUR-UNIQUE-IDENTIFIER`, right click Tables and then select Refresh.



15. Expand tables, right click `battery-life-prediction` and select query to view the data contained by the external table.

The screenshot shows the 'Tables' node expanded under 'sales'. A context menu is open over the 'dbo.battery-life-predictions' table, with the 'Select Top 1000' option highlighted. Other options in the menu include 'Edit Data', 'Script as Create', 'Script as Drop', and 'Refresh'. A tooltip 'Select Top 1000' is also visible near the cursor.

16. The vehicle telemetry along with predictions will appear. These are queried from the external table which is sourced from the CSV you created using the notebook.

Run Cancel Disconnect Change Connection sales ▾

```

1  SELECT TOP (1000) [Trip_Length_Mean]
2    ,[Trip_Length_Sigma]
3    ,[Trips_Per_Day_Mean]
4    ,[Trips_Per_Day_Sigma]
5    ,[Battery_Rated_Cycles]
6    ,[Alternator_Efficiency]
7    ,[Car_Has_EcoStart]
8    ,[Twelve_hourly_temperature_history_for_last_31_days]

```

#### RESULTS

_Reading...	Sensor_Reading...	Sensor_Reading...	Sensor_Reading...	Sensor_Reading...
082	3.752959	-16.78719	3.178833	-9.794724
807	6.507458	-6.905893	-0.8549166	-5.882981
218	9.494973	5.325149	24.37264	10.20268
998	3.521219	-3.946181	11.73532	-2.694235
218	9.494973	5.325149	24.37264	10.20268
571	19.46408	19.17639	26.87875	15.30772
127	9.416532	6.192503	7.30854	2.94854
429	-0.09202123	-8.994135	1.11721	-11.40551
708	6.618633	-4.793797	0.9704788	-8.445375
603	1.182248	-3.389345	3.795741	-8.844791

[View data](#)

### Task 3: Identify PII and GDPR-related compliance issues using Data Discovery & Classification in SSMS

Contoso Auto has several databases that include tables containing sensitive data, such as personally identifiable information (PII) like phone numbers, social security numbers, financial data, etc. Since some of their personnel and customer data include individuals who reside within the European Union (EU), they need to adhere to the General Data Protection Regulation ([GDPR](#)) as well. Because of this, Contoso Auto is required to provide periodic data auditing reports to identify sensitive and GDPR-related data that reside within their various databases.

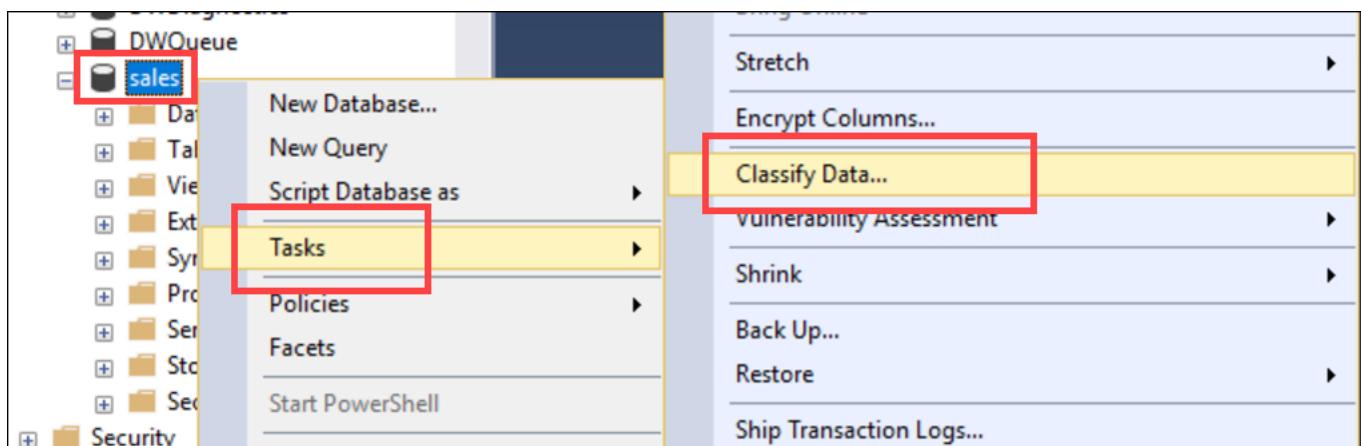
With SQL Server Management Studio, they are able to identify, classify, and generate reports on sensitive and GDPR-related data by using the [SQL Data Discovery & Classification](#) tool. This tool introduces a set of advanced services, forming a new SQL Information Protection paradigm aimed at protecting the data, not just the database:

- **Discovery & recommendations** - The classification engine scans your database and identifies columns containing potentially sensitive data. It then provides you an easy way to review and apply the appropriate classification recommendations, as well as to manually classify columns.
- **Labeling** - Sensitivity classification labels can be persistently tagged on columns.
- **Visibility** - The database classification state can be viewed in a detailed report that can be printed/exported to be used for compliance & auditing purposes, as well as other needs.

In this exercise, you will run the SQL Data Discovery & Classification tool against their customer database, which includes personal, demographic, and sales data.

1. Open SQL Server Management Studio (SSMS) and connect to your SQL Server 2019 cluster.

2. Right-click on the **sales**\_YOUR\_UNIQUE\_IDENTIFIER database, then choose **Tasks > Classify Data....**



3. When the tool runs, it will analyze all of the columns within all of the tables and recommend appropriate data classifications for each. What you should see is the Data Classification dashboard showing no currently classified columns, and a classification recommendations box at the top showing that there are 45 columns that the tool identified as containing sensitive (PII) or GDPR-related data. Click on this classification recommendations box.

The data classification recommendations box is highlighted.

4. The list of recommendations displays the schema, table, column, type of information, and recommended sensitivity label for each identified column. You can change the information type and sensitivity labels for each if desired. In this case, accept all recommendations by **checking the checkbox** in the recommendations table header.

Data Classification - sales

Save Add Classification View Report

45 columns with classification recommendations (click to minimize)

**0 classified columns**

Schema Table Column Information Type Sensitiv

45 columns with classification recommendations (click to minimize)

**Accept selected recommendations**

Schema	Table	Column	Information Type	Sensitivity
dbo	customer	c_email_address	Text	Low
dbo	customer	c_first_name	Text	Low
dbo	customer	c_last_name	Text	Low
dbo	customer_address	ca_address_id	Text	Low
dbo	customer_address	ca_city	Text	Low
dbo	customer_address	ca_street_name	Text	Low
dbo	customer_address	ca_street_number	Text	Low
dbo	customer_address	ca_street_type	Text	Low
dbo	customer_address	ca_zip	Text	Low

The recommendations are shown with each checkbox checked.

5. Click **Accept selected recommendations**.

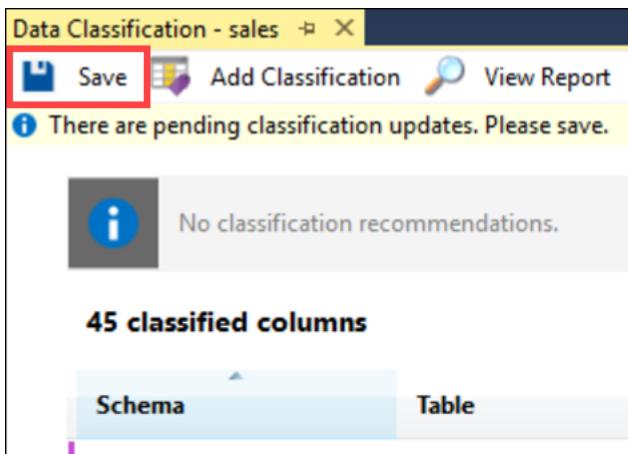
45 columns with classification recommendations (click to minimize)

**Accept selected recommendations**

Schema	Table	Column	Information Type	Sensitivity
dbo	customer	c_email_address	Text	Low
dbo	customer	c_first_name	Text	Low
dbo	customer	c_last_name	Text	Low
dbo	customer_address	ca_address_id	Text	Low
dbo	customer_address	ca_city	Text	Low
dbo	customer_address	ca_street_name	Text	Low
dbo	customer_address	ca_street_number	Text	Low
dbo	customer_address	ca_street_type	Text	Low
dbo	customer_address	ca_zip	Text	Low

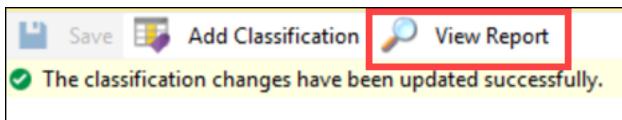
The Accept selected recommendations button is highlighted.

6. Click **Save** in the toolbar above to apply your changes.



The Save button is highlighted.

7. After the changes are saved, click View Report.



The View Report button is highlighted.

8. What you should see is a report with a full summary of the database classification state. When you right-click on the report, you can see options to print or export the report in different formats.

# SQL Data Classification Report

Server name: mssql-master-po

Report generated:

Database name: sales

Classified columns

45 / 483

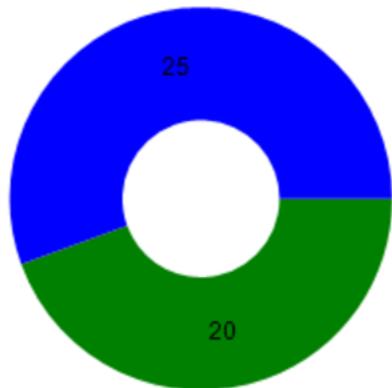
Tables containing sensitive data

11 / 34

Unique information types

4

Label distribution



■ Confidential  
■ Confidential - GDPR

Information Type distribution



Schema	Table	Column	Information Type
dbo	customer	c_first_name	Name
	customer	c_last_name	Name
	customer	c_email_address	Contact Info
	customer_address	ca_address_id	Contact Info
	customer_address	ca_street_number	Contact Info
	customer_address	ca_street_name	Contact Info
	customer_address	ca_street_type	Contact Info

The report is displayed, as well as the context menu showing export options after right-clicking on the report.

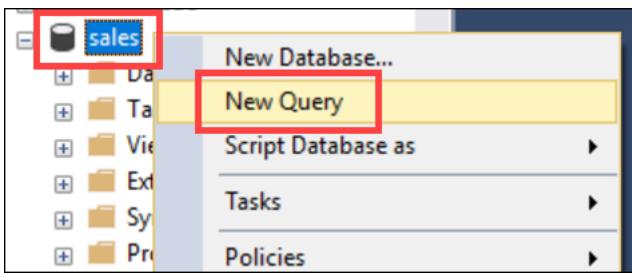
## Task 4: Fix compliance issues with dynamic data masking

Some of the columns identified by the Data Discovery & Classification tool as containing sensitive (PII/GDPR) information include phone numbers, email addresses, billing addresses, and credit card numbers. One way to ensure compliance with various rules and regulations that enforce policies to protect such sensitive data is to prevent those who are not authorized from seeing it. An example would be displaying xxx-xxx-xx95 instead of 123-555-2695 when outputting a phone number within a SQL query result, report, web page, etc. This is commonly called data masking. Traditionally, modifying systems and applications to implement data masking can be challenging. This is especially true when the masking has to apply all the way down to the data source level. Fortunately, SQL Server and its cloud-related product, Azure SQL Database, provides a feature named [dynamic data masking](#) (DDM) to automatically protect this sensitive data from non-privileged users.

Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. DDM can be configured on the database to hide sensitive data in the result sets of queries over designated database fields, while the data in the database is not changed. Dynamic data masking is easy to use with existing applications, since masking rules are applied in the query results. Many applications can mask sensitive data without modifying existing queries.

In this task, you will apply dynamic data masking to one of the database fields so you can see how to address the reported compliance issues. To test the data mask, you will create a test user and query the field as that user.

1. Open SQL Server Management Studio (SSMS) and connect to your SQL Server 2019 cluster.
2. Expand the databases list, right-click on **sales\_YOUR\_UNIQUE\_IDENTIFIER**, then select **New Query**.



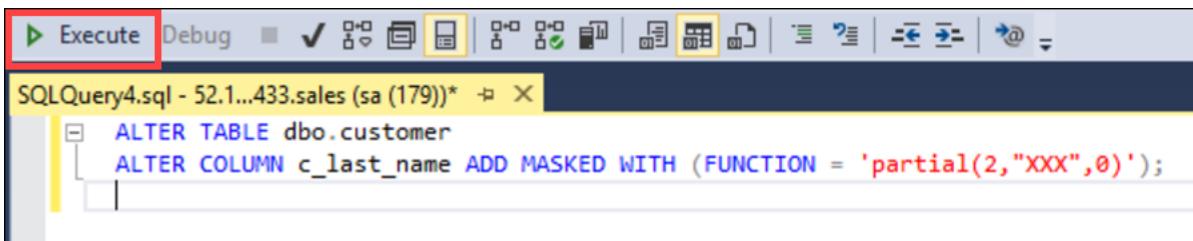
The sales database and New Query menu item are highlighted.

- Add a dynamic data mask to the existing dbo.customer.c\_last\_name field by pasting the below query into the new query window:

```
ALTER TABLE dbo.customer
ALTER COLUMN c_last_name ADD MASKED WITH (FUNCTION = 'partial(2,"XXX",0)');
```

The partial custom string masking method above exposes the first two characters and adds a custom padding string after for the remaining characters. The parameters are: prefix,[padding],suffix

- Execute the query by clicking the **Execute** button above the query window, or enter *F5*.



The dynamic data mask query is shown and the Execute button is highlighted above.

- Clear the query window and replace the previous query with the following to add a dynamic data mask to the dbo.customer.c\_email\_address field:

```
ALTER TABLE dbo.customer
ALTER COLUMN c_email_address ADD MASKED WITH (FUNCTION = 'email()');
```

The email masking method exposes the first letter of an email address and the constant suffix ".com", in the form of an email address: aXXX@XXXX.com.

- Clear the query window and replace the previous query with the following, selecting all rows from the customer table:

```
SELECT * FROM dbo.customer
```

	.sales_date_sk	c_salutation	c_first_name	c_last_name	c_preferred_cust_flag	c_birth_day	c_birth_month	c_birth_year
1		Dr.	Todd	Reynolds	N	25	6	1926
2		Dr.	Timothy	Hill	N	12	6	1941
3		Dr.	Charles	Austin	N	16	6	1984
4		Dr.	Michael	Byme	N	14	6	1942
5		Dr.	Juan	Ogden	N	8	6	1931
6		Dr.	Jennifer	Smithson	N	4	6	1962
7		Dr.	Nancy	Alvarez	N	28	6	1947
8		Dr.	Shannon	Wright	N	20	6	1961
9		Dr.	Gilbert	Allen	N	15	6	1964
10		Dr.	Anna	Ayers	N	17	6	1946
11		Dr.	Christine	Morgan	N	15	6	1973
12	1	Dr.	Raymond	Wheeler	N	1	6	1990

The query results are shown with no mask applied to the Postal Code field.

- Notice that the full last name and email address values are visible. That is because the user you are logged in as a privileged user. Let's create a new user and execute the query again:

```
CREATE USER TestUser WITHOUT LOGIN;
GRANT SELECT ON dbo.customer TO TestUser;

EXECUTE AS USER = 'TestUser';
SELECT * FROM dbo.customer;
REVERT;
```

- Execute the query by clicking the **Execute** button. Notice this time that the Postal Code values are masked (90xxx).

```

CREATE USER TestUser WITHOUT LOGIN;
GRANT SELECT ON dbo.customer TO TestUser;

EXECUTE AS USER = 'TestUser';
SELECT * FROM dbo.customer;
REVERT;

```

100 %

Results Messages

	c_first_name	c_last_name	c_preferred_cust_flag	c_birth_day	c_birth_month	c_birth_year	c_birth_country
1	Todd	ReXXX	N	25	6	1926	GUYANA
2	Timothy	HiXXX	N	12	6	1941	BOTSWANA
3	Charles	AuXXX	N	16	6	1984	TURKMENISTAN
4	Michael	ByXXX	N	14	6	1942	SRI LANKA
5	Juan	OgXXX	N	8	6	1931	KOREA, REPUBLIC OF
6	Jennifer	SmXXX	N	4	6	1962	BENIN
7	Nancy	AIXXX	N	28	6	1947	LESOTHO
8	Shannon	WrXXX	N	20	6	1961	ESTONIA
9	Gilbert	AIXXX	N	15	6	1964	ANDORRA
10	Anna	AyXXX	N	17	6	1946	UZBEKISTAN
11	Christine	MoXXX	N	15	6	1973	JAMAICA
12	Raymond	WhXXX	N	1	6	1990	AMERICAN SAMOA
13	Stephanie	BaXXX	N	5	6	1948	DENMARK
14	Olga	McXXX	N	24	6	1987	TOGO
15	Manual	DrXXX	N	11	6	1959	GUINEA
16	Michael	NuXXX	N	22	6	1988	MARSHALL ISLANDS
17	John	MuXXX	N	11	6	1958	CAYMAN ISLANDS
18	Kevin	JaXXX	N	28	6	1945	SOMALIA
19	Wanda	PiXXX	N	11	6	1967	ITALY
20	Myron	BaXXX	N	12	6	1953	JORDAN
21	Philip	DaXXX	N	21	6	1983	JORDAN
22	Charlotte	ReXXX	N	22	6	1964	GEORGIA
23	Richard	BiXXX	N	9	6	1943	GUINEA

The query results are shown with the mask applied to the Postal Code field.

## Wrap-up

Thank you for participating in the SQL Server 2019 Big Data Clusters experience! We hope you are excited about the new capabilities, and will refer back to this experience to learn more about these features.

To recap, you experienced:

- How to minimize or remove the need for ETL through **data virtualization** with [relational data sources](#) and [CSV files](#), by being able to query against these alongside internal SQL 2019 tables with no data movement required.
- Training a machine learning model by running a Jupyter notebook on the Big Data cluster, then scoring data with the trained model and saving it as an external table for easy access.
- Using the [SQL Data Discovery & Classification](#) tool to identify and tag PII and GDPR-related compliance issues.
- Used dynamic data masking to automatically protect sensitive data from unauthorized users.

## Additional resources and more information

- [What's new in SQL Server 2019 preview](#)
- [SQL Server 2019 big data clusters overview and architecture](#)
- [How to run a sample notebook in Azure Data Studio on a SQL Server 2019 big data cluster, and leverage Spark](#)
- [What is Azure Data Studio?](#)
- [Security Center for SQL Server Database Engine and Azure SQL Database](#)
- [SQL Data Discovery and Classification tool documentation](#)

- [Intelligent query processing in SQL databases](#)
- [What's new in SQL Server Machine Learning Services](#)
- [How to run Java code in SQL Server 2019](#)
- [Learning content in GitHub: SQL Server Workshops](#)
- [SQL Server Samples Repository in GitHub. Feature demos, code samples etc.](#)

SQL Server 2019 has a new database compatibility level: 150. When you set your database to this level, you will gain query performance benefits due to improvements to the family of intelligent query processing (QP) features in SQL Server 2019. These features improve the performance of existing workloads with minimal work on your part to implement.

A few QP features we would like to call out are:

1. [Scalar UDF inlining](#): This feature automatically transforms [scalar UDFs](#) into relational expressions. It embeds them in the calling SQL query. This transformation improves the performance of workloads that take advantage of scalar UDFs. Scalar UDF inlining facilitates cost-based optimization of operations inside UDFs. The results are efficient, set-oriented, and parallel instead of inefficient, iterative, serial execution plans. This feature is enabled by default under database compatibility level 150.
2. [Table variable deferred compilation](#): This feature improves plan quality and overall performance for queries that reference table variables. During optimization and initial compilation, this feature propagates cardinality estimates that are based on actual table variable row counts. This accurate row count information optimizes downstream plan operations. Table variable deferred compilation defers compilation of a statement that references a table variable until the first actual run of the statement. This deferred compilation behavior is the same as that of temporary tables. This change results in the use of actual cardinality instead of the original one-row guess.
3. [Row mode memory grant feedback](#): A query's post-execution plan in SQL Server includes the minimum required memory needed for execution and the ideal memory grant size to have all rows fit in memory. Performance suffers when memory grant sizes are incorrectly sized. Excessive grants result in wasted memory and reduced concurrency. Insufficient memory grants cause expensive spills to disk. By addressing repeating workloads, batch mode memory grant feedback recalculates the actual memory required for a query and then updates the grant value for the cached plan. **When an identical query statement is executed**, the query uses the revised memory grant size, reducing excessive memory grants that impact concurrency and fixing underestimated memory grants that cause expensive spills to disk. Row mode memory grant feedback expands on the batch mode memory grant feedback feature by adjusting memory grant sizes for both batch and row mode operators.

Read more about [intelligent query processing](#) in SQL databases.