Universidad Europea VALENCIA

FINAL PROJECT: "ADVENTURE WORKS DASHBOARD AND EXPLORATORY DATA ANALYSIS"

Dmitriy Kirichenko

Ignacio Beneyto

Luis Llobell

Data Science Degree

Project Big Data II

Professor: Alexander PERDIGUERO O´LEARY

Index

Table of contents

About the project

The project was assigned to an entire group with the goal of achieving certain set results. We would need to achieve an advanced level of quality, work ethics, advanced theoretical and practical knowledge applications.

The path of making this project should look as follows:

1. Creation of suitable environment for parallel work as a team.
2. Data extraction from the corresponding sources.
3. Data wrangling.
4. Exloratory Data Analysis.
5. KPI’s and questions assignment.
6. Creation of Data Warehouse.
7. Data streaming.
8. Predictive Model Training.
9. Creation of user-friendly Power Bi dashboard.
10. Final documentation commitment.

Problems

When using GitHub to collaborate on a project that involves connecting from three different computers, performing Exploratory Data Analysis (EDA), running SQL queries, and working in Power BI, several challenges and potential issues could arise:

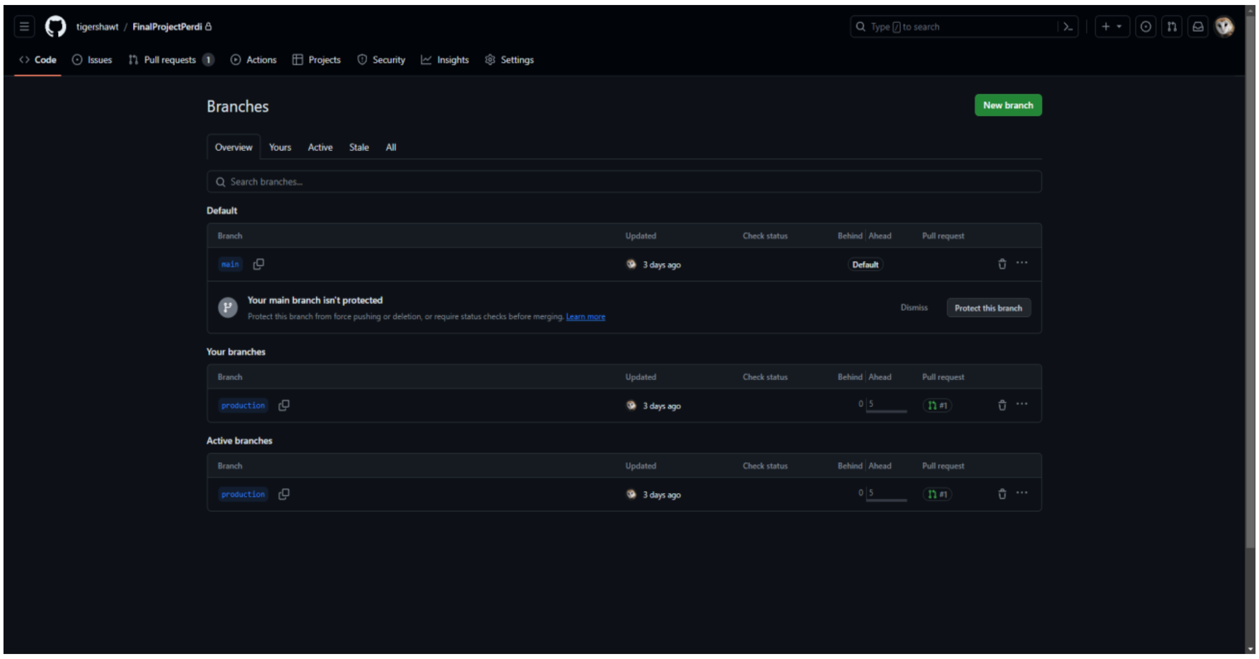
- Version synchronization: It is essential to ensure that all team members are working with the same version of code and files. Version conflicts can arise if changes are not synchronized correctly.
- Merge conflicts: When multiple people are working simultaneously on various parts of the project, conflicts may occur when merging changes. This may require manual resolution and communication between team members.
- Concurrent access to shared resources: If multiple people try to access and modify the same files simultaneously, locking and conflict issues can arise. It is important to establish protocols to avoid such situations.
- Data security and privacy: When working with sensitive data, such as SQL queries that could contain sensitive information, it is vital to ensure data security and privacy. Appropriate security and encryption measures must be implemented.
- Performance and scalability: Depending on the size of the project and the amount of data involved, system performance and scalability can become issues. It is important to monitor and optimize system performance, as necessary.
- Tool and version compatibility: Ensuring that all team members are using compatible versions of necessary tools, such as SQL and Power BI, can avoid incompatibility issues.
- Communication and coordination: Effective communication between team members is key to ensuring everyone is aware of changes, issues, and pending tasks. Clear communication channels must be established, and a log of activities must be kept up to date.
- Training and technical knowledge: Some team members may not be familiar with all the tools and technologies used in the project. Providing training and technical support can help mitigate this problem.

Creation of working environment

Git repository creation

First, it is vital to create sharable working space with logs and version controls to be able to do work simultaneously and test it. For this reason, we have chosen GitHub (https://github.com/) since it is the most advanced and reliable resource available on the internet for free. The material part of this project is not sponsored at any means, so we must work with null budget.

We have created the repository called "FinalProjectPerdi" where every piece of work from members of the team would come together and be version controlled in order if someone would commit any mistakes. For this reason, we have created additional to a main brunch called "production" where all the changes would be revised before merging the changed or added files into the main branch.



Branches

Default

Branch	Updated	Check status	Behind | Ahead	Pull request
main	3 days ago		Default	

Your main branch isn't protected
Protect this branch from force pushing or deletion, or require status checks before merging. Learn more

Your branches

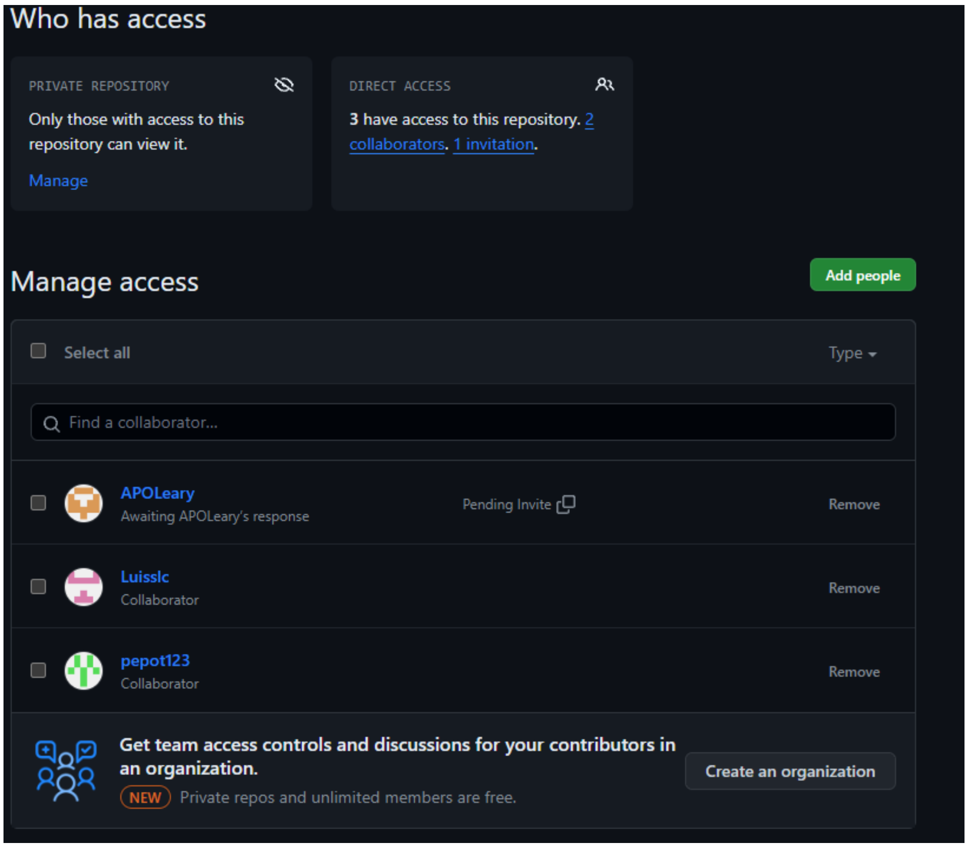
Branch	Updated	Check status	Behind | Ahead	Pull request
production	3 days ago		0 | 5	#1

Active branches

Branch	Updated	Check status	Behind | Ahead	Pull request
production	3 days ago		0 | 5	#1

Git repository access

We have made the repository the closed one, and only team members with permissions would have access to it. All the members have the same rights. They need to identify themselves from a local instance to proceed with the git functionality in our repository (Later will be explained in detail).



Who has access

PRIVATE REPOSITORY
Only those with access to this repository can view it.
Manage

DIRECT ACCESS
3 have access to this repository. 2 collaborators. 1 invitation.

Manage access

Add people

Select all — Type

Find a collaborator...

APOLeary — Awaiting APOLeary's response — Pending Invite — Remove

Luisslc — Collaborator — Remove

pepot123 — Collaborator — Remove

Get team access controls and discussions for your contributors in an organization.
NEW Private repos and unlimited members are free.
Create an organization

Repository roles:

Dmitriy Kirichenko - tigershawt – owner

Ignacio Beneyto – pepot123 – collaborator

Luis Llobell – Luisslc – collaborator

Alexander PERDIGUERO O´LEARY - APOLeary – collaborator

Data extraction

Finding the source data

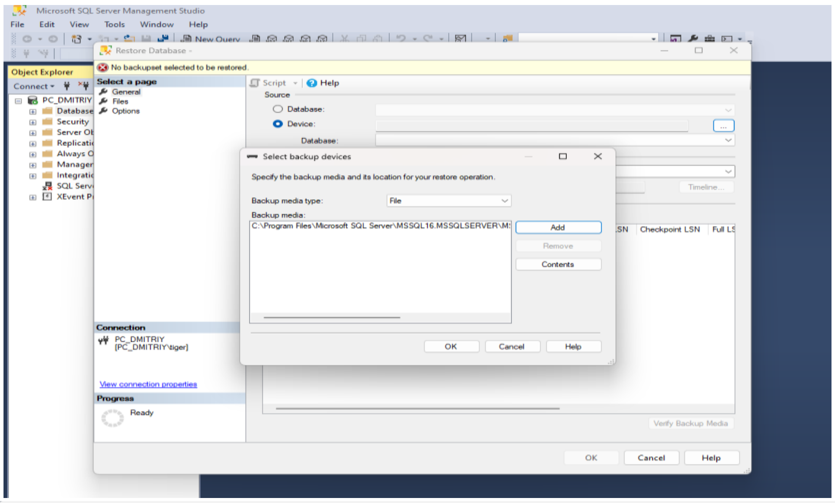
Finding a source data was a trivial task. First search in a google would give the needed result. Adventure Works database is available as a backup file in many places, we have downloaded it through GitHub. The version we have chosen would be the latest since we hope it would be the most saturated with data insights. The file was "AdventureWorks2022.bak" of (OLTP) full database backup (https://github.com/Microsoft/sql-server-samples/releases/tag/adventureworks).

Restoring the backup version

We have followed the guide from Microsoft official website (https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms#download-backup-files) that suggests using "Microsoft SQL Server" to restore the backup version. We have chosen the local SQL environment (https://www.microsoft.com/es-es/sql-server/sql-server-downloads) to do not use any cloud solutions which might be costly. This software is needed to launch local SQL server to proceed with the backup.

To complete the backup, we have installed SSMS to be able to finally restore the DB, manipulate with the data, and clearly see the internal structure for better understanding. All these steps are described in the same guide from the closest link above.

Universidad Europea VALENCIA



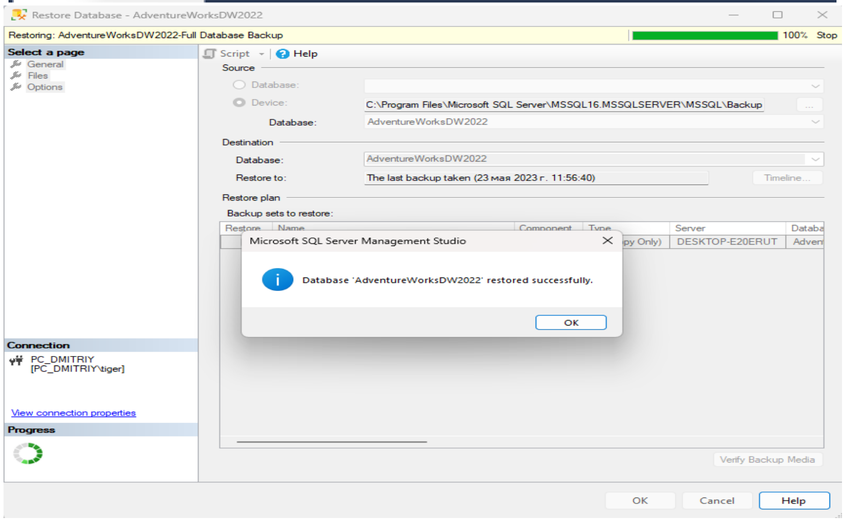
No backupset selected to be restored.

Specify the backup media and its location for your restore operation.

Backup media type: File

Backup media: C:\Program Files\Microsoft SQL Server\MSSQL16.MSSQLSERVER\M

Connection: PC_DMITRIY [PC_DMITRIY\tiger]



Restoring: AdventureWorksDW2022-Full Database Backup

Device: C:\Program Files\Microsoft SQL Server\MSSQL16.MSSQLSERVER\MSSQL\Backup

Database: AdventureWorksDW2022

Destination Database: AdventureWorksDW2022

Restore to: The last backup taken (23 мая 2023 г. 11:56:40)

py Only) DESKTOP-E20ERUT Adven

Database 'AdventureWorksDW2022' restored successfully.

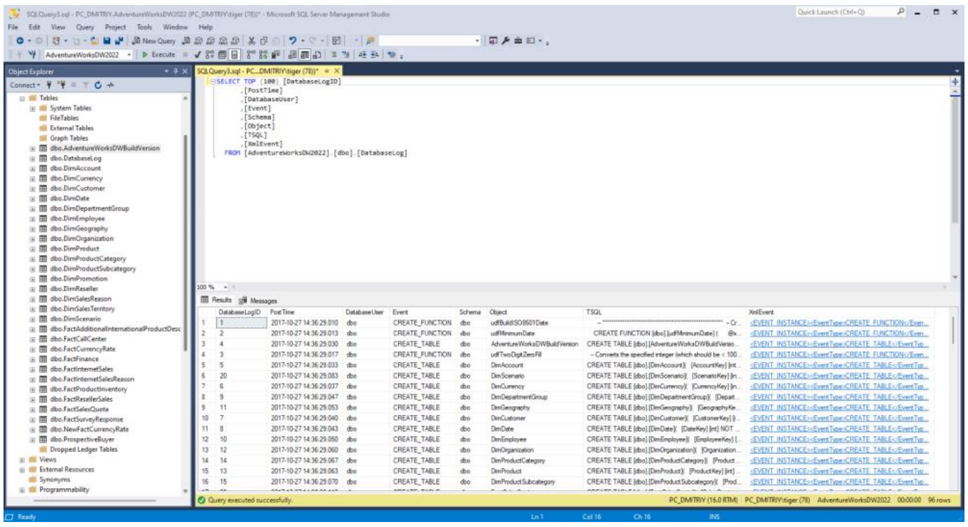
Connection: PC_DMITRIY [PC_DMITRIY\tiger]

DB first look

After making the backup, we see the entire database, since we only need to data to work with, we do exploratory visual analysis of all the data tables and see what we might need in our project and what we would not. However, we have taken all the tables except the two:

AdventureWorksDWBuildVersion – The table with 1 parameter stating the version of the database and the time it was patched.

DatabaseLog – History of changes made to various parts of the database stating time, action, and user.



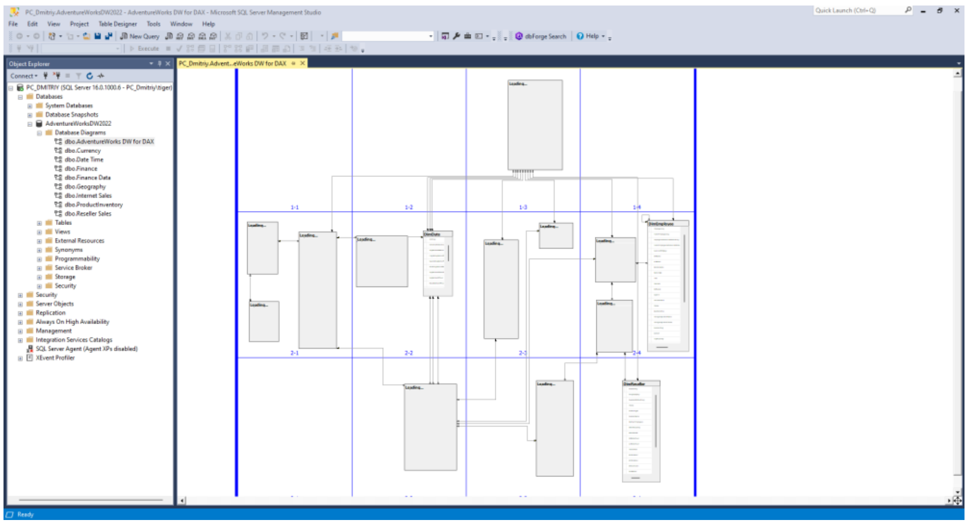
```
SELECT TOP (100) [DatabaseLogID]
      ,[PostTime]
      ,[DatabaseUser]
      ,[Event]
      ,[Schema]
      ,[Object]
      ,[TSQL]
      ,[XmlEvent]
  FROM [AdventureWorksDW2022].[dbo].[DatabaseLog]
```

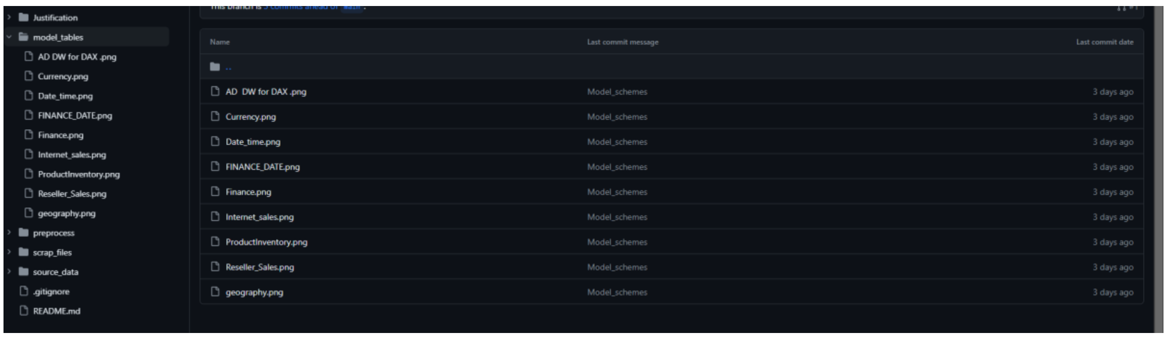
	DatabaseLogID	PostTime	DatabaseUser	Event	Schema	Object	TSQL	XmlEvent
1	1	2017-10-27 14:36:29.010	dbo	CREATE_FUNCTION	dbo	udfBuildISO8601Date	-- ************************************ -- Cr...	<EVENT_INSTANCE><EventType>CREATE_FUNCTION</Even...
2	2	2017-10-27 14:36:29.013	dbo	CREATE_FUNCTION	dbo	udfMinimumDate	CREATE FUNCTION [dbo].[udfMinimumDate] (@x...	<EVENT_INSTANCE><EventType>CREATE_FUNCTION</Even...
3	4	2017-10-27 14:36:29.030	dbo	CREATE_TABLE	dbo	AdventureWorksDWBuildVersion	CREATE TABLE [dbo].[AdventureWorksDWBuildVersio...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
4	3	2017-10-27 14:36:29.017	dbo	CREATE_FUNCTION	dbo	udfTwoDigitZeroFill	-- Converts the specified integer (which should be < 100...	<EVENT_INSTANCE><EventType>CREATE_FUNCTION</Even...
5	5	2017-10-27 14:36:29.033	dbo	CREATE_TABLE	dbo	DimAccount	CREATE TABLE [dbo].[DimAccount]([AccountKey] [int...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
6	20	2017-10-27 14:36:29.083	dbo	CREATE_TABLE	dbo	DimScenario	CREATE TABLE [dbo].[DimScenario]([ScenarioKey] [in...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
7	6	2017-10-27 14:36:29.037	dbo	CREATE_TABLE	dbo	DimCurrency	CREATE TABLE [dbo].[DimCurrency]([CurrencyKey] [in...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
8	9	2017-10-27 14:36:29.047	dbo	CREATE_TABLE	dbo	DimDepartmentGroup	CREATE TABLE [dbo].[DimDepartmentGroup]([Depart...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
9	11	2017-10-27 14:36:29.053	dbo	CREATE_TABLE	dbo	DimGeography	CREATE TABLE [dbo].[DimGeography]([GeographyKe...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
10	7	2017-10-27 14:36:29.040	dbo	CREATE_TABLE	dbo	DimCustomer	CREATE TABLE [dbo].[DimCustomer]([CustomerKey] [i...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
11	8	2017-10-27 14:36:29.043	dbo	CREATE_TABLE	dbo	DimDate	CREATE TABLE [dbo].[DimDate]([DateKey] [int] NOT ...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
12	10	2017-10-27 14:36:29.050	dbo	CREATE_TABLE	dbo	DimEmployee	CREATE TABLE [dbo].[DimEmployee]([EmployeeKey] [...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
13	12	2017-10-27 14:36:29.060	dbo	CREATE_TABLE	dbo	DimOrganization	CREATE TABLE [dbo].[DimOrganization]([Organization...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
14	14	2017-10-27 14:36:29.067	dbo	CREATE_TABLE	dbo	DimProductCategory	CREATE TABLE [dbo].[DimProductCategory]([Product...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
15	13	2017-10-27 14:36:29.063	dbo	CREATE_TABLE	dbo	DimProduct	CREATE TABLE [dbo].[DimProduct]([ProductKey] [int] ...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...
16	15	2017-10-27 14:36:29.070	dbo	CREATE_TABLE	dbo	DimProductSubcategory	CREATE TABLE [dbo].[DimProductSubcategory]([Prod...	<EVENT_INSTANCE><EventType>CREATE_TABLE</EventTyp...

Query executed successfully. PC_DMITRIY (16.0 RTM) PC_DMITRIY\tiger (78) AdventureWorksDW2022 00:00:00 96 rows

Extracting the table's relationships

To understand data at an appropriate level, after analysing all the database, we have found the relational scheme, where all the primary and foreign keys relationships are clearly stated. This will be useful in future EDA operations such as "table join". We have captured all the available diagrams with windows "Snipping Tool" and have placed them into out GitHub repository.





```
Justification
model_tables
  AD DW for DAX .png
  Currency.png
  Date_time.png
  FINANCE_DATE.png
  Finance.png
  Internet_sales.png
  ProductInventory.png
  Reseller_Sales.png
  geography.png
preprocess
scrap_files
source_data
.gitignore
README.md
```

Name	Last commit message	Last commit date
..		
AD DW for DAX .png	Model_schemes	3 days ago
Currency.png	Model_schemes	3 days ago
Date_time.png	Model_schemes	3 days ago
FINANCE_DATE.png	Model_schemes	3 days ago
Finance.png	Model_schemes	3 days ago
Internet_sales.png	Model_schemes	3 days ago
ProductInventory.png	Model_schemes	3 days ago
Reseller_Sales.png	Model_schemes	3 days ago
geography.png	Model_schemes	3 days ago

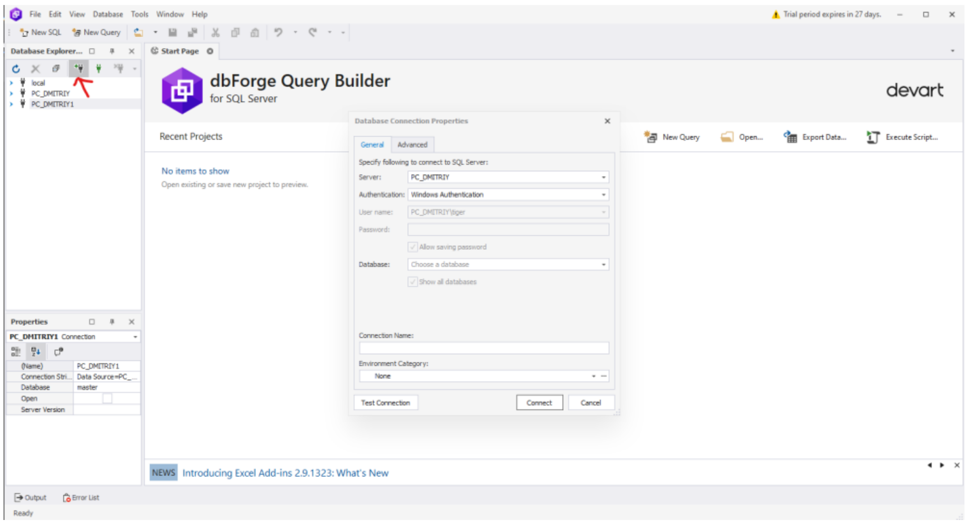
Desired format

We are Data Scientist and primarily work with python and csv files, we have investigated ways of extracting the tables into csv files. We tend to believe, that we have found the most appropriate way of doing it.

Since SSMS does not have the default option of converting the SQL relational database into CSV with keeping all the header and rows in place. We have found the plugin that significantly helped us and saved lots of time alternatively parsing through the available on GitHub csv (https://github.com/microsoft/sql-server-samples/tree/master/samples/databases/adventure-works/oltp-install-script) containing data and rules for SQL server.

The tool, which is called Data Pump from Devart, was easy to use and of quality. We have followed the following documentation from the developer (https://docs.devart.com/data-pump/exporting-data/export-to-csv.html). Attention this documentation was outdated since it does not show the integrated buttons in SSMS itself. So, we had to find alternative way of using this Data Pump.

After downloading and installing the free trial version of "dbForge SQL Tools Standard Trial" we need to find the application in our system called "dbForge Query Builder for SQL Server". After opening the app, we would need to connect to our running SQL server by choosing the credentials from the image below. (red arrow shows the button to initialize new connection).



Database Connection Properties

General | Advanced

Specify following to connect to SQL Server:

Server: PC_DMITRIY

Authentication: Windows Authentication

User name: PC_DMITRIY\tiger

Password:

Allow saving password

Database: Choose a database

Show all databases

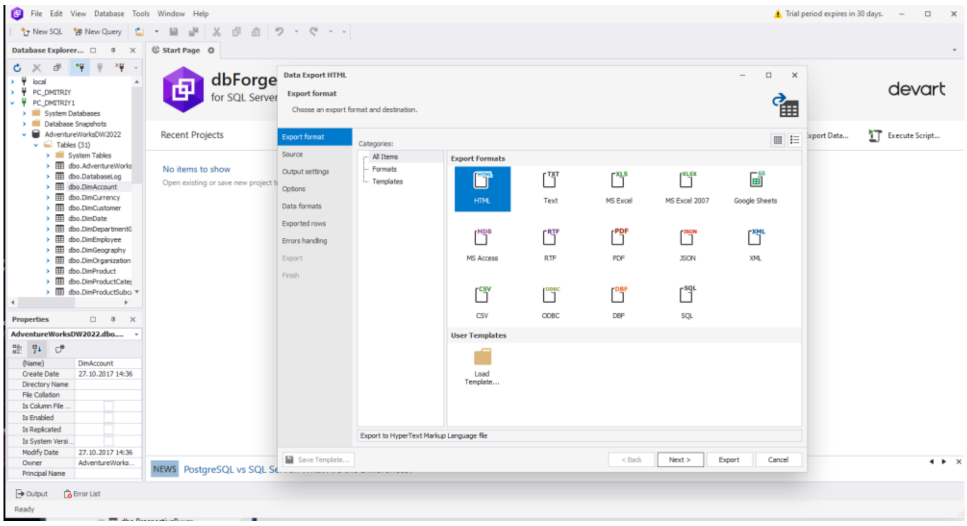
Connection Name:

Environment Category: None

PC_DMITRIY1 Connection

(Name)	PC_DMITRIY1
Connection Stri...	Data Source=PC_...
Database	master
Open	
Server Version	

After we have established connection, we can keep up with developers' documentation and using the Data Pump.



Data Export HTML

Export format

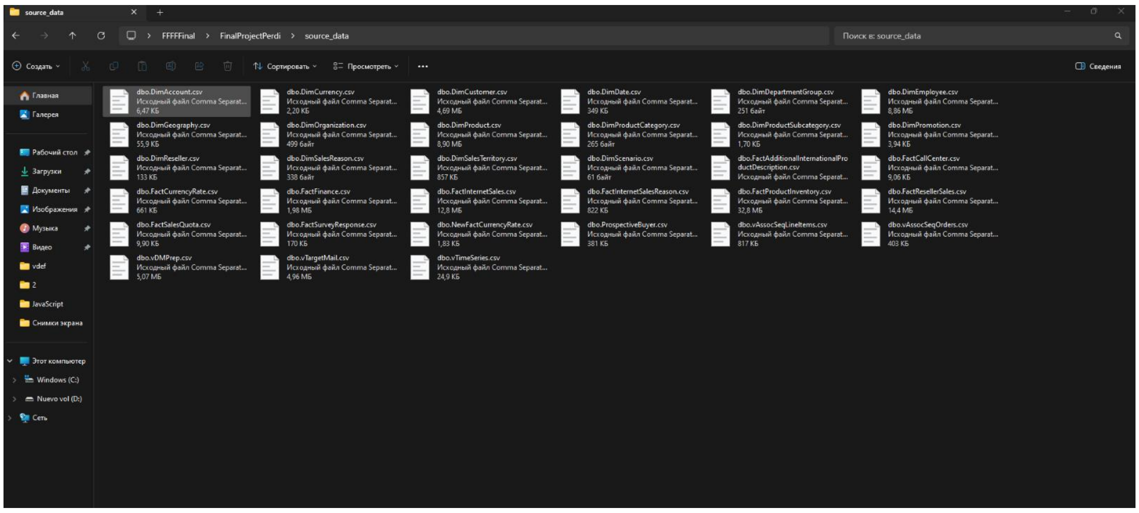
Choose an export format and destination.

Export to HyperText Markup Language file

AdventureWorksDW2022.dbo....

(Name)	DimAccount
Create Date	27.10.2017 14:36
Directory Name	
File Collation	
Is Column File ...	
Is Enabled	
Is Replicated	
Is System Versi...	
Modify Date	27.10.2017 14:36
Owner	AdventureWorks...
Principal Name	

Extracting all the necessary tables into our git repository.



```
dbo.DimAccount.csv            dbo.DimCurrency.csv           dbo.DimCustomer.csv           dbo.DimDate.csv               dbo.DimDepartmentGroup.csv    dbo.DimEmployee.csv
Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...
6,47 КБ                       2,20 КБ                       4,69 МБ                       349 КБ                        251 байт                      8,86 МБ
dbo.DimGeography.csv          dbo.DimOrganization.csv       dbo.DimProduct.csv            dbo.DimProductCategory.csv    dbo.DimProductSubcategory.csv dbo.DimPromotion.csv
Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...
55,9 КБ                       499 байт                      8,90 МБ                       265 байт                      1,70 КБ                       3,94 КБ
dbo.DimReseller.csv           dbo.DimSalesReason.csv        dbo.DimSalesTerritory.csv     dbo.DimScenario.csv           dbo.FactAdditionalInternationalPro dbo.FactCallCenter.csv
                                                                                                                        ductDescription.csv
Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...
133 КБ                        338 байт                      857 КБ                        61 байт                                                     9,06 КБ
dbo.FactCurrencyRate.csv      dbo.FactFinance.csv           dbo.FactInternetSales.csv     dbo.FactInternetSalesReason.csv dbo.FactProductInventory.csv dbo.FactResellerSales.csv
Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...
661 КБ                        1,98 МБ                       12,8 МБ                       822 КБ                        32,8 МБ                       14,4 МБ
dbo.FactSalesQuota.csv        dbo.FactSurveyResponse.csv    dbo.NewFactCurrencyRate.csv   dbo.ProspectiveBuyer.csv      dbo.vAssocSeqLineItems.csv    dbo.vAssocSeqOrders.csv
Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...
9,90 КБ                       170 КБ                        1,83 КБ                       381 КБ                        817 КБ                        403 КБ
dbo.vDMPrep.csv               dbo.vTargetMail.csv           dbo.vTimeSeries.csv
Исходный файл Comma Separat...Исходный файл Comma Separat...Исходный файл Comma Separat...
5,07 МБ                       4,96 МБ                       24,9 КБ
```