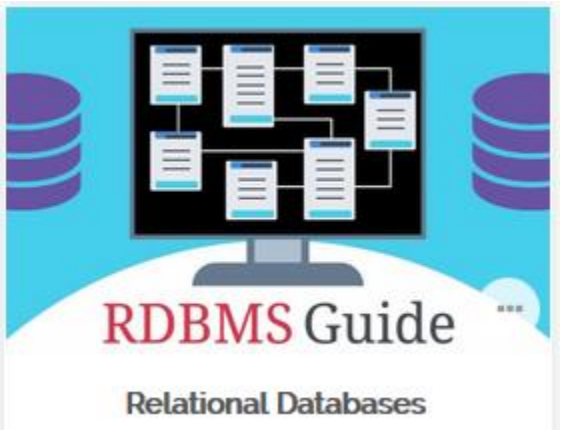
DATA SCIENCE TECH INSTITUTE

Relational Databases

Exam Project Document



Team:

- Stephane Durfort (Lead)
- Vigneshwar Gurunathan
- Olivier Degardin
- Hassan Oumar Idriss
- Nguyen Vinh Binh

Table of contents

1. Project Overview

Implement a project about relational databases go a bit further into databases concepts with ETL, Datawarehouse, Business Intelligent model…

2. Project Goal

Use Microsoft ETL : SSIS. An ETL is a tool vastly used in Business Intelligence projects to create data flows. Starting from the source files, the model of Data warehouse (DWH) will be deployed with previous steps STA (Staging) and ODS (Operational Data Store).

Create a datawarehouse (DWH) using two spreadsheets:

- Crimes_Delits_France_depuis_1996.xls
- Mapping départements.csv

3. Deliverables

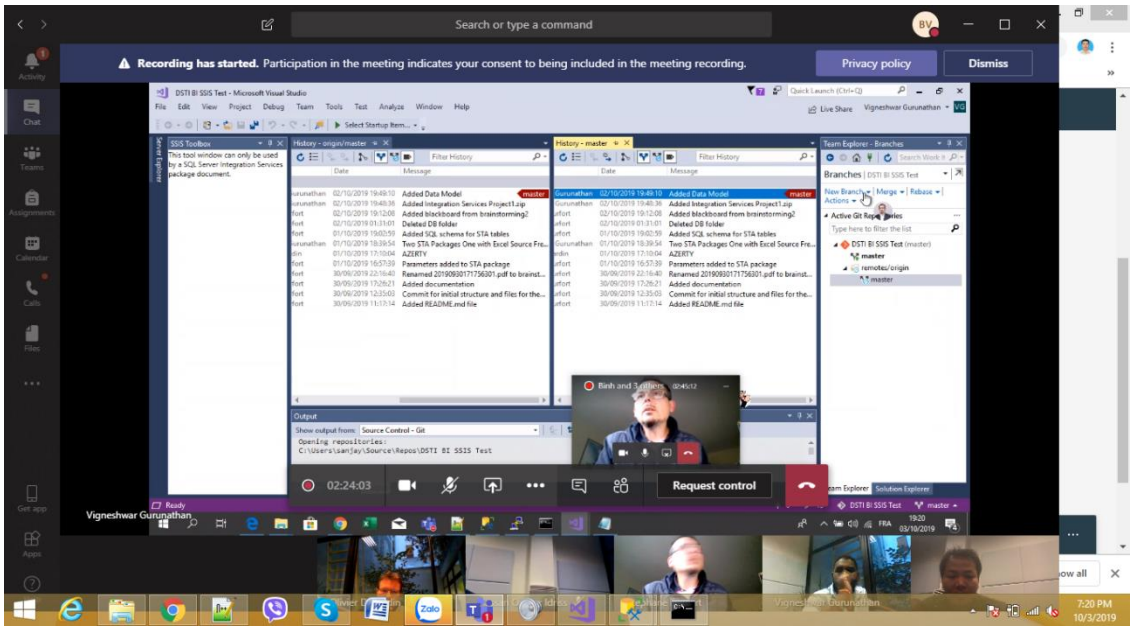
- Documentation (this guide)
- Archive of the Git repository used for the project with SSIS Solution and Packages as well as other resources used during the project implementation.

4. Method of Working & Environment

In order to implement this project as a team, we choose to rely on standard DevOps and Collaboration tools.

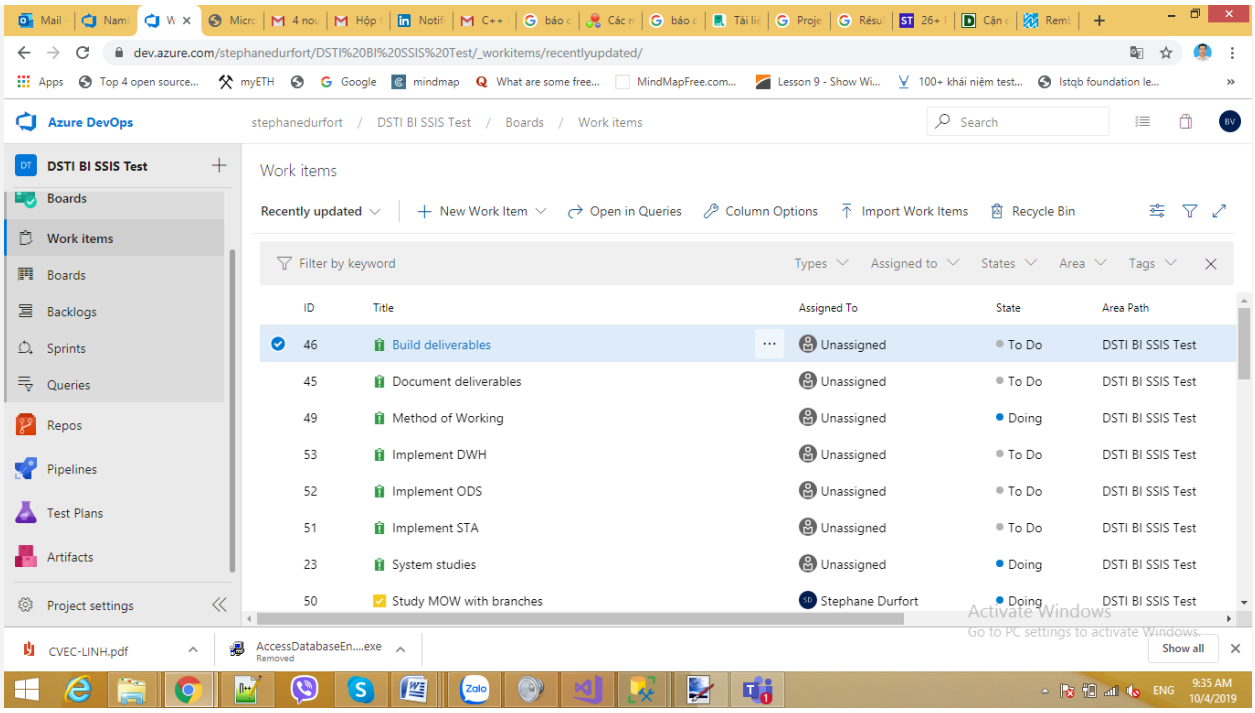
4.1 Working team tool

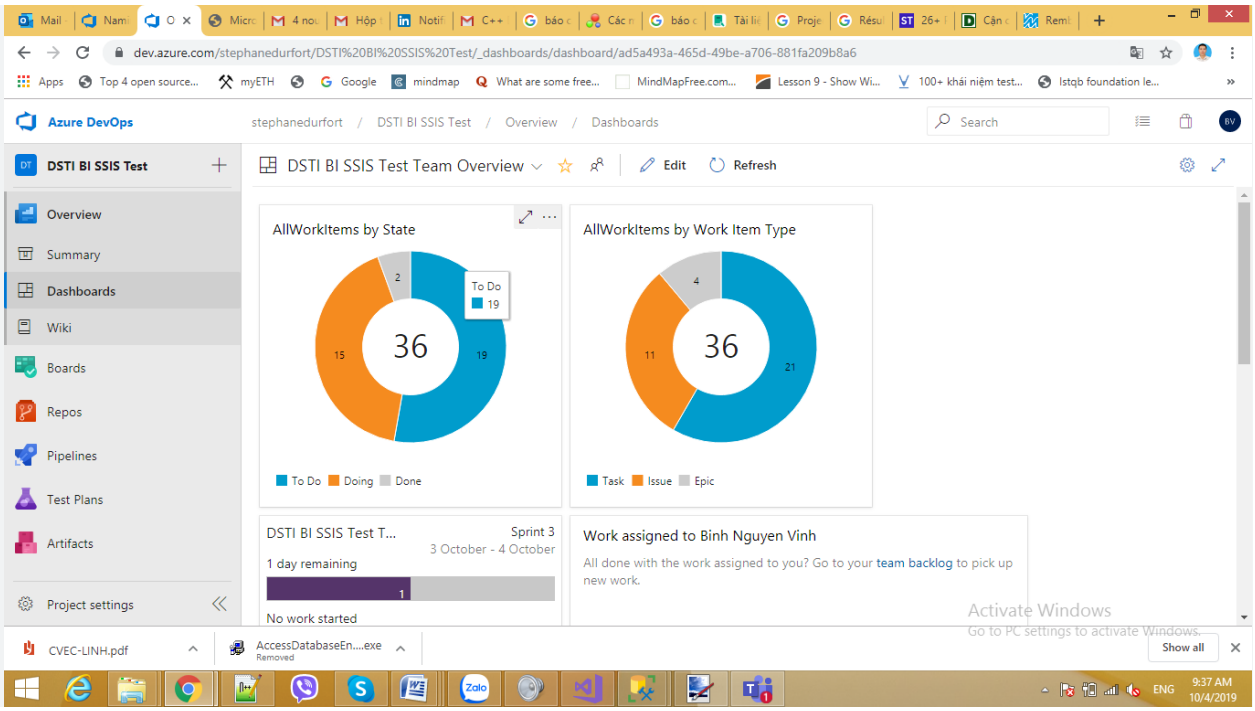
Microsoft Teams: to use video call group, share screen, use screen remote control, exchange and communicate (questions, progress, …) via the 'Chat' process.



4.2 Development environment

- Microsoft Azure for source code server + project management





- IDE : Visual Studio 2017 + SSDT
- SQL server 2017
- Git

5. Design Overview

5.1 Design Approach

- A single SSIS Solution: BI_IS_Project
- A single Database: BI_CrimesInFrance
- A DataSource repository in the root of C: drive for the Excel and CSV files
- An initial SSIS package to initialize environment and database, plus utility packages to run all packages according to their dependencies
- All imported record have a RecordCreatedDate field to track initial import date

5.2 Limitations / Potential improvements

- No historical retention in DWH: data is fully reloaded for both facts and dimension in DWH by default for sake of simplicity.
- Limited Technical/Functional validation rejects

5.3 Requirements

- In order to import Excel data and depending on what is installed in your environment, you may need to install either the 32-bit or 64-bit version of Microsoft Access Database Engine 2010 or 2016 (DatabaseEngine.exe or AccessDatabaseEngine_X64.exe)
  - Microsoft Access Database Engine 2016 Redistributable
  - Microsoft Access Database Engine 2010 Redistributable

Note: Running the solution in release mode has not been tested and may require to have 32-bit version of SSIS Runtime installed

5.4 Usage

Configuration

- Update the GitSourceDir project parameter to match with your configuration.

Packages Execution

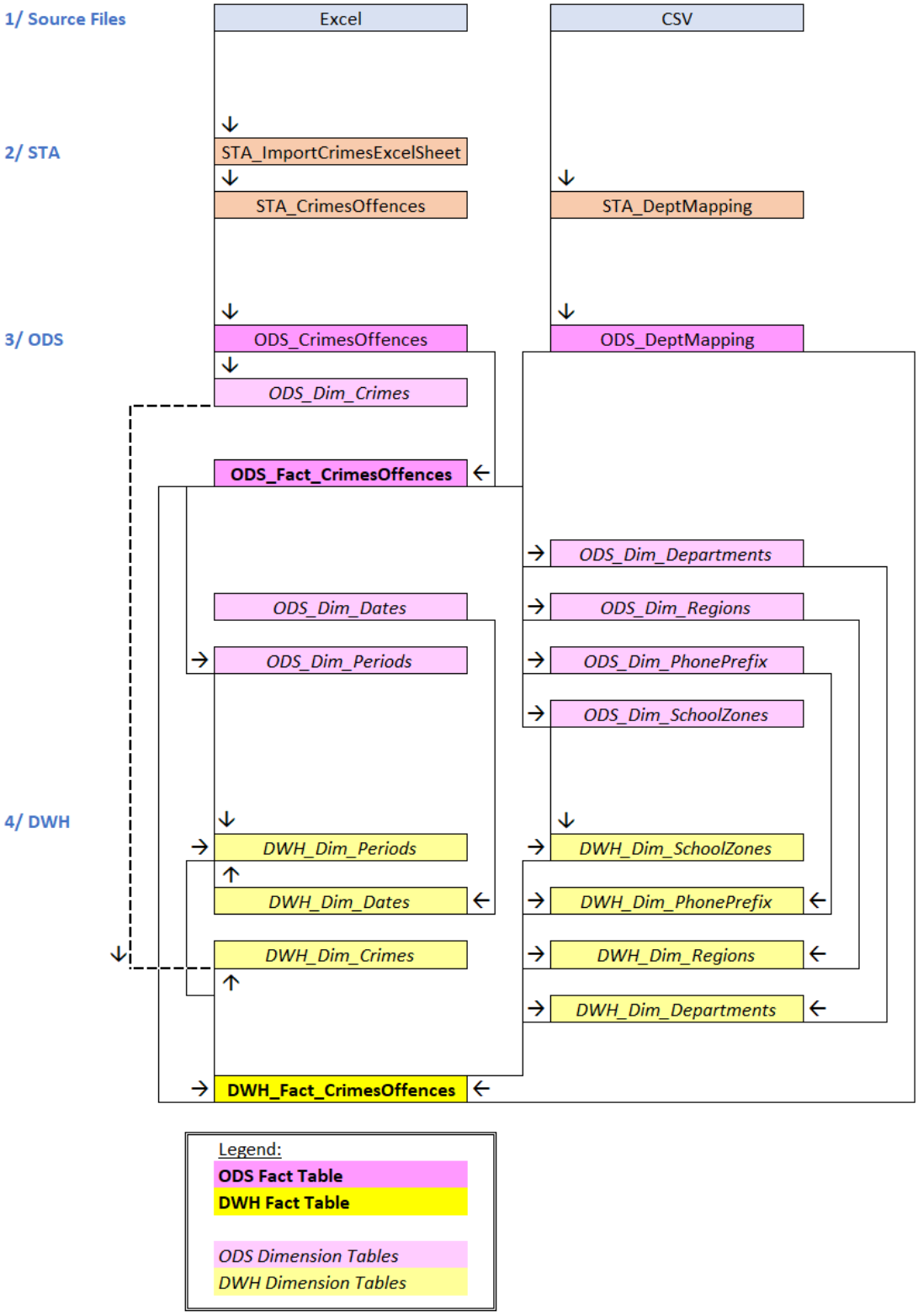
Package	Description
_Project_InitEnv	Execute this package to create the proper folder/file structure on your machine as well as database and tables. It should be executed only once.
_Project_RunSTA	Execute this package to run both the Excel & CSV imports into STA
_Project_RunODS	Execute this package to run all packages in ODS according to their dependencies
_Project_RunDWH	Execute this package to run all packages in DWH according to their dependencies

To verify all data is properly imported, you can use a script from the Resources\DBSchema folder (BI_CrimesInFranceDW-ShowTables.sql) to dump the number of records in all tables:

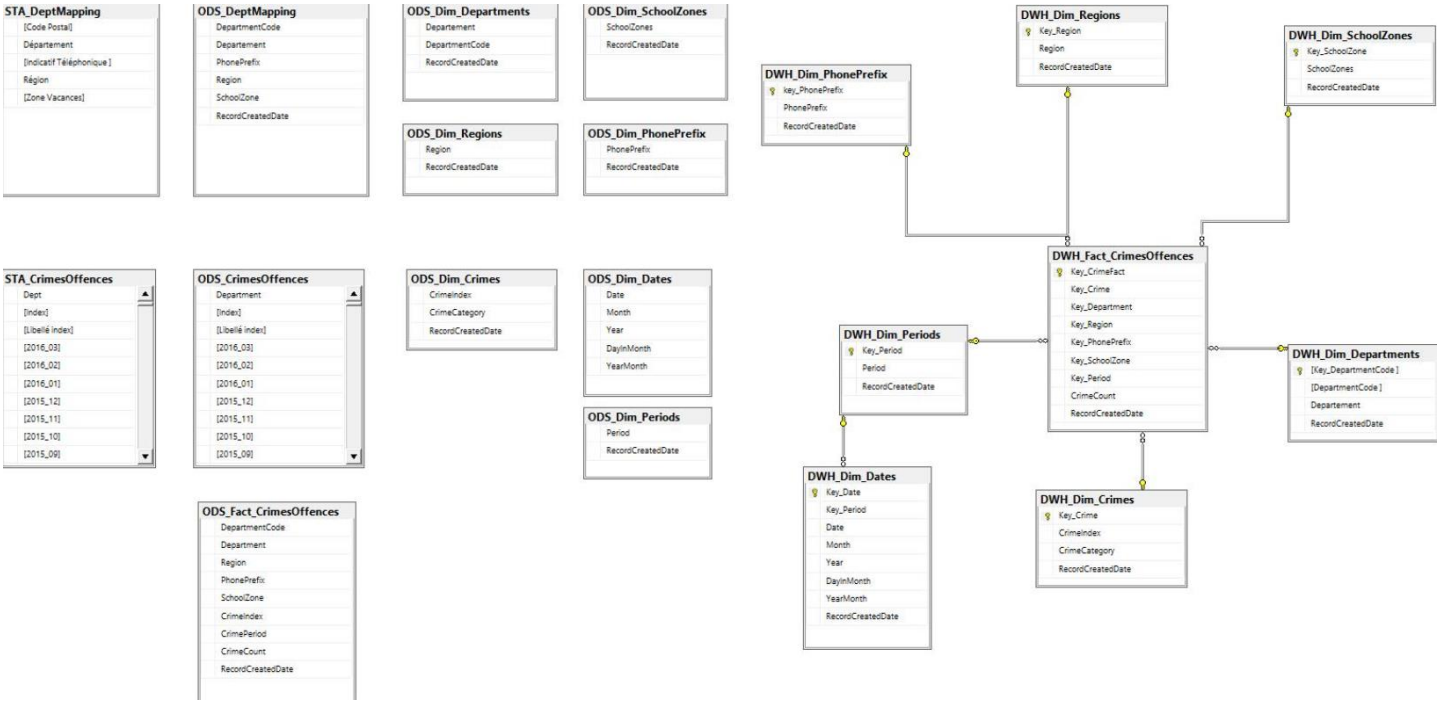
Résultats | Messages

	Line	DBTable	RecordCount
1	1	-+- STA -+-	NULL
2	2	[BI_CrimesInFranceDW].[dbo].[STA_CrimesOffences]	10272
3	3	[BI_CrimesInFranceDW].[dbo].[STA_DeptMapping]	95
4	4	-+- ODS -+-	NULL
5	5	- Level-1	NULL
6	6	[BI_CrimesInFranceDW].[dbo].[ODS_CrimesOffences]	10272
7	7	[BI_CrimesInFranceDW].[dbo].[ODS_DeptMapping]	95
8	8	- Level-2	NULL
9	9	[BI_CrimesInFranceDW].[dbo].[ODS_Fact_CrimesOffen...	2470093
10	10	[BI_CrimesInFranceDW].[dbo].[ODS_Dim_Crimes]	107
11	11	[BI_CrimesInFranceDW].[dbo].[ODS_Dim_Periods]	243
12	12	[BI_CrimesInFranceDW].[dbo].[ODS_Dim_Dates]	8767
13	13	[BI_CrimesInFranceDW].[dbo].[ODS_Dim_Departments]	95
14	14	[BI_CrimesInFranceDW].[dbo].[ODS_Dim_Regions]	13
15	15	[BI_CrimesInFranceDW].[dbo].[ODS_Dim_SchoolZones]	4
16	16	[BI_CrimesInFranceDW].[dbo].[ODS_Dim_PhonePrefix]	5
17	17	-+- DWH -+-	NULL
18	18	- Facts	NULL
19	19	[BI_CrimesInFranceDW].[dbo].[DWH_Fact_CrimesOffe...	2470093
20	20	- Dimensions	NULL
21	21	[BI_CrimesInFranceDW].[dbo].[DWH_Dim_Crimes]	35175
22	22	[BI_CrimesInFranceDW].[dbo].[DWH_Dim_Periods]	243
23	23	[BI_CrimesInFranceDW].[dbo].[DWH_Dim_Dates]	8767
24	24	[BI_CrimesInFranceDW].[dbo].[DWH_Dim_Departments]	95
25	25	[BI_CrimesInFranceDW].[dbo].[DWH_Dim_Regions]	13
26	26	[BI_CrimesInFranceDW].[dbo].[DWH_Dim_PhonePrefix]	5
27	27	[BI_CrimesInFranceDW].[dbo].[DWH_Dim_SchoolZones]	4

6. Data Workflow

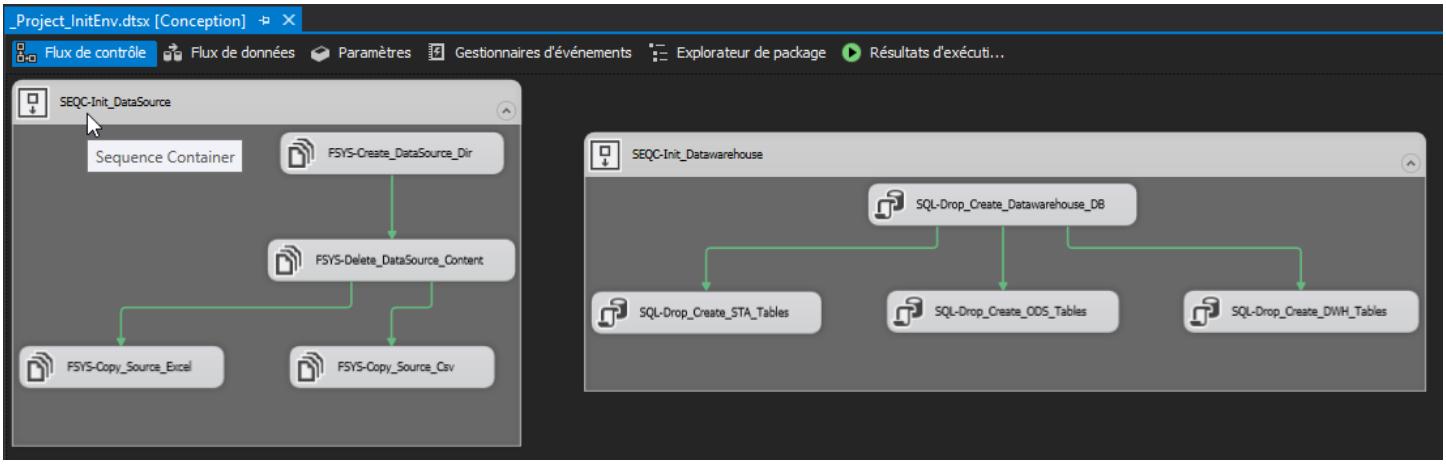


6.1 Database schema – diagram from SQL

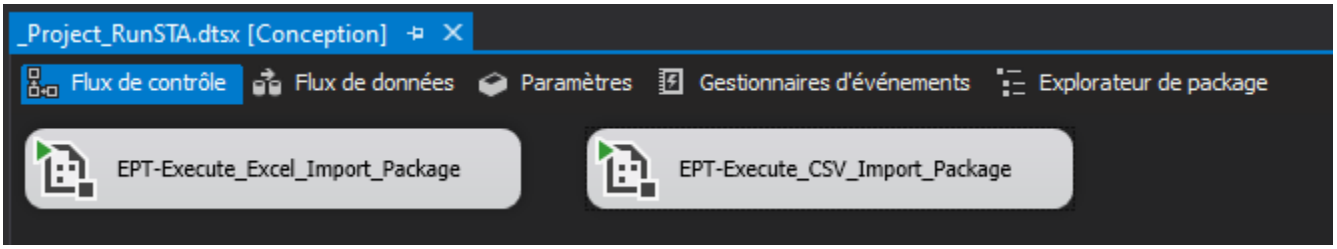


6.2 Packages' Sequencing

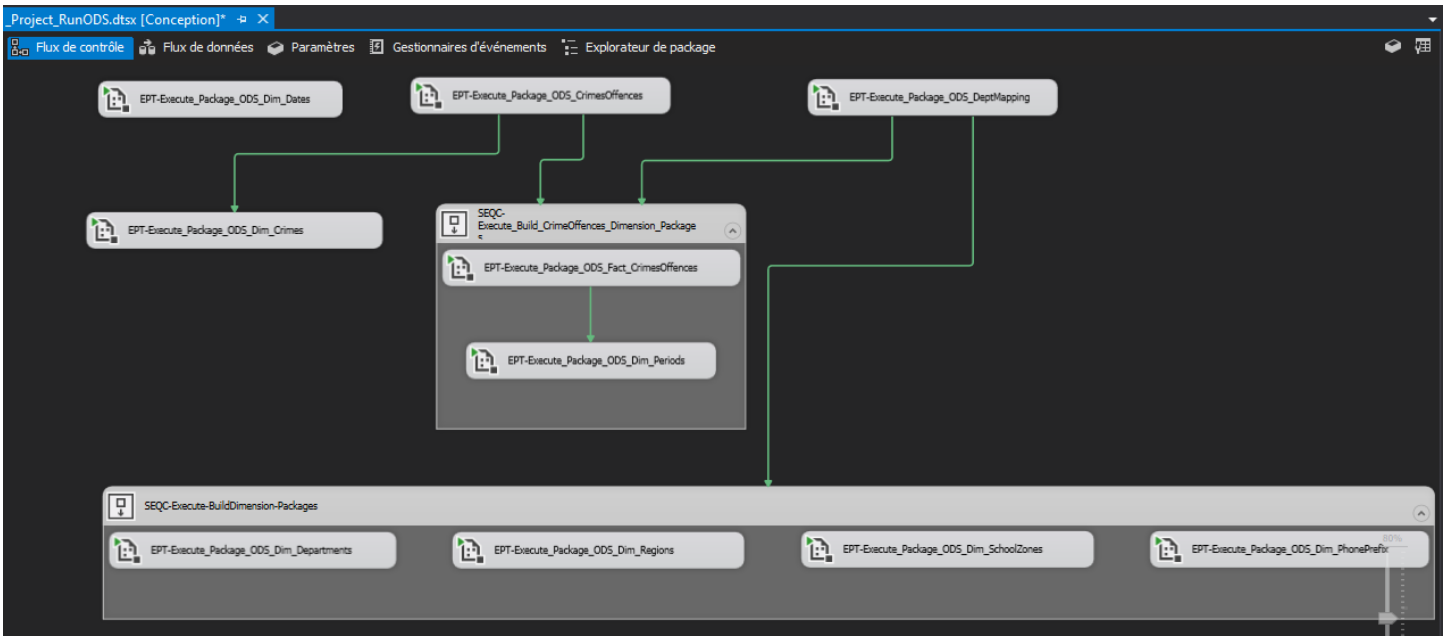
_Project_InitEnv.dtsx : (should only be run to initialize the environment)



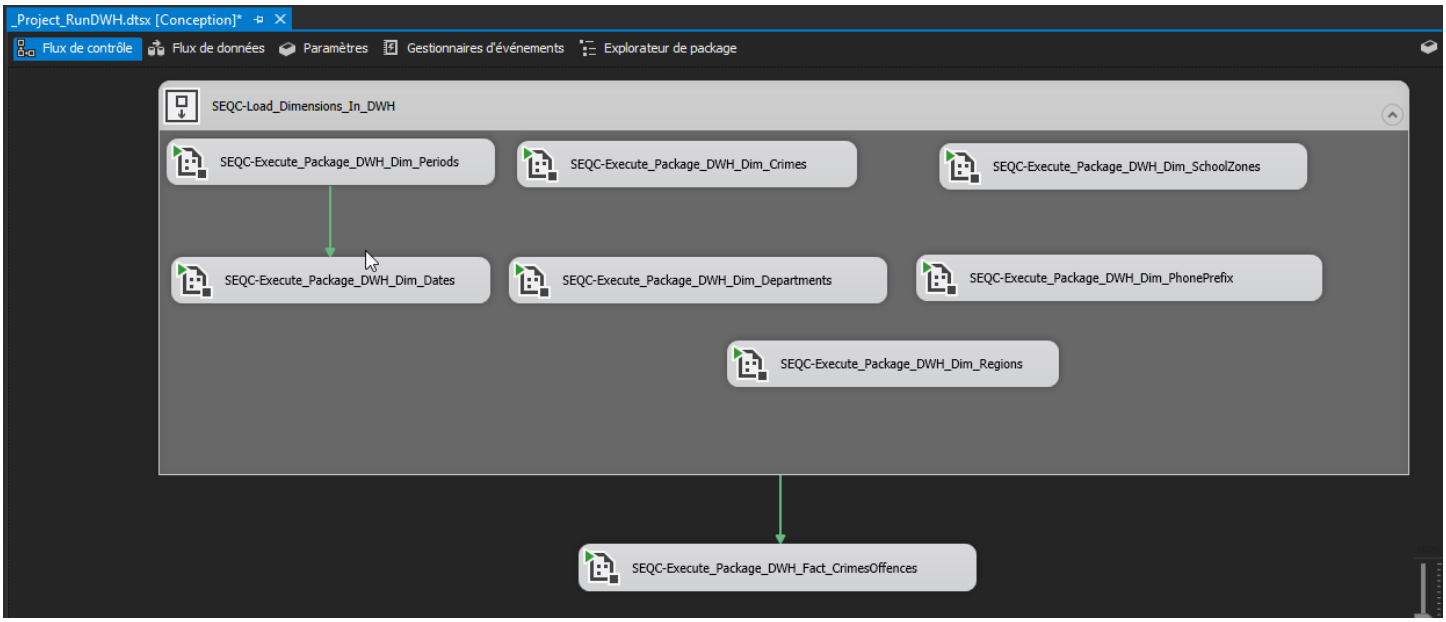
_Project_RunSTA.dtsx :



_Project_RunODS.dtsx :



_Project_RunDWH.dtsx :



6.3 Data source anomalies

Some anomalies have to be managed in the technical or functional validation stages

Crimes_Delits_France_depuis_1996.xls

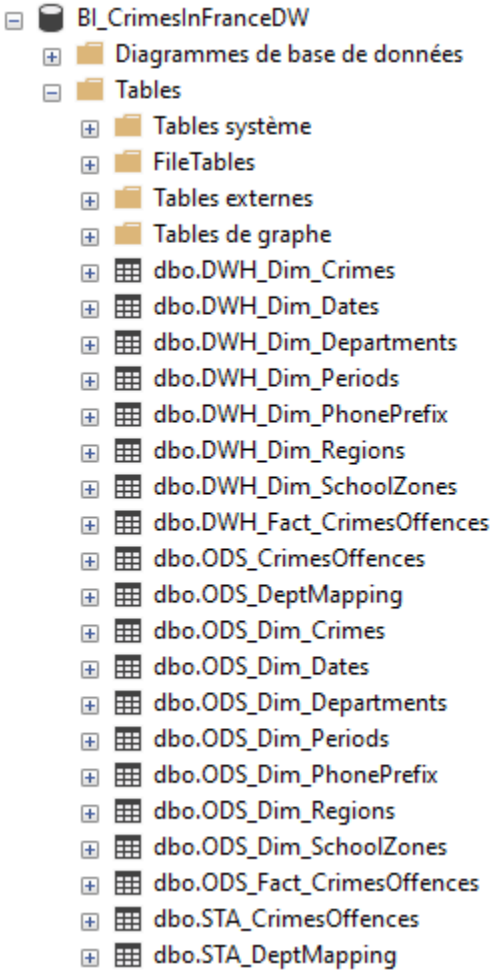
- Department sheet - the sequence of DepXX sheets has some specific issues
  - Departments are alphanumeric (Dep2A, Dep2B are present)
  - Duplicate label index in "Libellé index" = "Index non utilisé"

Mapping départements.csv

- 'Code Postal' column - name to be modified, it is not a zipcode, it is a department code:

        - Department 87 is missing
        - Departments 2A and 2B: not integer
    - 'Zone vacances' column:
        - Departments 2A and 2B: result = 'Spécial' instead of A, B or C for the other departments

6.4 Database and Tables Naming Convention



A/ Database's name = **BI_CrimesInFranceDW**

B/ Tables' names

- For the STA's section:

[dbo].[STA_CrimesOffences]
[dbo].[STA_DeptMapping]

- For the ODS's section:

[dbo].[ODS_CrimesOffences]	
[dbo].[ODS_DeptMapping]	
[dbo].[ODS_Dim_Crimes]	*Dimension Table*
[dbo].[ODS_Dim_Dates]	*Dimension Table*
[dbo].[ODS_Dim_Departments]	*Dimension Table*
[dbo].[ODS_Dim_Periods]	*Dimension Table*
[dbo].[ODS_Dim_PhonePrefix]	*Dimension Table*
[dbo].[ODS_Dim_Regions]	*Dimension Table*
[dbo].[ODS_Dim_SchoolZones]	*Dimension Table*
[dbo].[ODS_Fact_CrimesOffences]	**Fact Table**

- For the DWH's section:

[dbo].[DWH_Dim_Crimes]	*Dimension Table*
[dbo].[DWH_Dim_Dates]	*Dimension Table*
[dbo].[DWH_Dim_Departments]	*Dimension Table*
[dbo].[DWH_Dim_Periods]	*Dimension Table*
[dbo].[DWH_Dim_PhonePrefix]	*Dimension Table*
[dbo].[DWH_Dim_Regions]	*Dimension Table*
[dbo].[DWH_Dim_SchoolZones]	*Dimension Table*
[dbo].[DWH_Fact_CrimesOffences]	**Fact Table**

7. Project Implementation Details

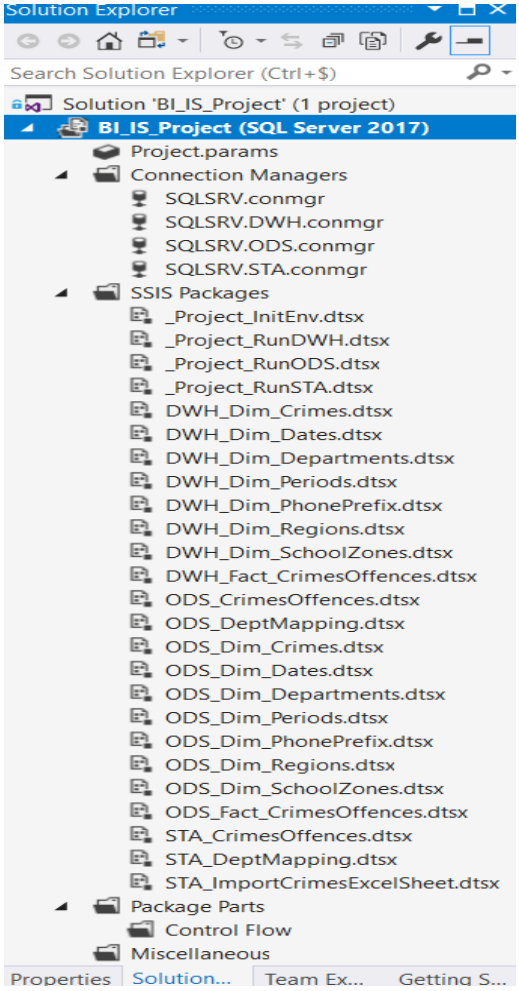
7.1 Project Parameters

- DataSourceDir: Path to local import folder when running SSIS packages
- GitSourceDir: Path to Excel & CSV files in Local Git Repos
- GitCSVDeptMappingFilename: CSV filename in Git
- GitExcelCrimesDataFilename: Excel filename in Git

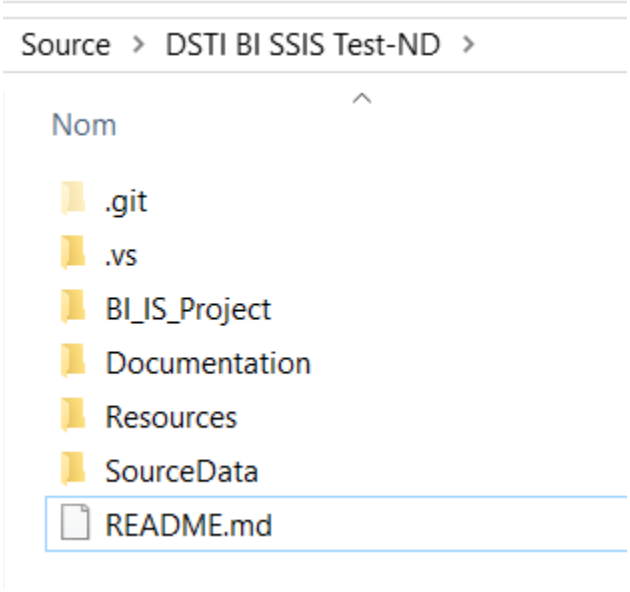
7.2 Connections

Scope	Connection Name	Type	Server	Catalog	Description
Package	DeptMappingCVS	FlatFile	Mapping départements.csv		
Package	CrimesOffencesXLS	Excel	Crimes_Delits_France_depuis_1996.xls		
Project	SQLSRV	OLEDB	localhost		connection to master on localhost using SSPI
Project	SQLSRV.STA	OLEDB	localhost	BI_CrimesInFranceDW	connection to BI_CrimesInFranceDW on localhost using SSPI
Project	SQLSRV.ODS	OLEDB	localhost	BI_CrimesInFranceDW	connection to BI_CrimesInFranceDW on localhost using SSPI
Project	SQLSRV.DWH	OLEDB	localhost	BI_CrimesInFranceDW	connection to BI_CrimesInFranceDW on localhost using SSPI

7.3 Solutions Explorer Overview



7.4 SSIS Source Structure



The above structure is detailed as below for each repository with those contents:

- .git: global files for a configuration which allows to fetch, pull, commit, etc.
- .vs: json structure for the implementation of the project in Visual Studio
- BI_IS_Project: the packages ('.dtsx' files)
- Documentation: the used/working/final documents
- Resources: SQL queries to drop and create the database and the tables + our brainstorming documentation
- DataSource: the both data sets (Excel and CSV)
- README.md: used work in progress/implementations' documentation for this project

7.5 Packages

Package Project_InitEnv

This package should be executed the first time the project is clone on a new machine. It will create the proper folder structure to import data and the database and related tables.

Variables

- Excel_CrimesData_Git
- Excel_CrimesData
- Csv_DeptMapping_Git
- Csv_DeptMapping

Tasks

- Create SourceData at the root of the C: drive and
- Copy Excel (and rename?)
- Copy CSV (and rename?)
- Drop and create Project Database
- (WIP) Drop and create STA Tables
- Drop and create ODS Tables
- Drop and create DWH Tables

Package STA_CrimesOffences (Excel import)

ISSUE with Microsoft Excel.

Due to stability issues implementing a Foreach container calling data flow task in order to import every tab of the Excel sheet, we moved to a similar implementation using execute package task instead of data flow task for the inner task that imports a single Excel sheet. This approach seems more stable in our environments from our testing. TODO: investigate reasons for "For each + data flow" option failures.

Variables

- FLC_SheetName

Connection

- IN: CrimesOffencesXLS: ExcelSource = C:\DataSource\Crimes_Delits_France_depuis_1996.xls

Tasks

- Truncate STA table
- For each Sheets in Excel sheet call Package `STA_ImportCrimesExcelSheet` with `User::FLC_SheetName`

Package STA_ImportCrimesExcelSheet (Excel import single sheet)

Parameter

- SheetName

Connection

- IN: CrimesOffencesXLS: ExcelSource = C:\DataSource\Crimes_Delits_France_depuis_1996.xls
- OUT: SQLSRV.STA

Tasks

- Load Excel Sheet `$Package::SheetName`
- Add Derived Column `Dept` as `$Package::SheetName`
- Import to STA

Package STA_DeptMapping (CSV import)

Connection

- IN: DeptMAppingCSV: FlatFile = C:\DataSource\Mapping départements.csv
- OUT: SQLSRV.STA

Tasks

- Truncate destination STA table
- Load CSV file
- Convert data to unicode string
- Import to STA

Package ODS_CrimesOffences

Connection

- IN: SQLSRV.STA
- OUT: SQLSRV.ODS

Tasks

- Truncate destination ODS table
- Load CrimesData from STA
- Convert index to integer
- Convert sheetname to Department code (2 digit alpha-numeric code)
- Add import timestamp
- Import to ODS

Package ODS_Fact_CrimesOffences

TODO: Technical Rejects

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.ODS

Tasks

- Truncate destination ODS table
- Load CrimesData from ODS_CrimesOffences
- Unpivot date periods
- Convert Crimes count value to integers
- Lookup department data
  - Unmatched results to Technical reject
- Store to ODS

Package ODS_Dim_Crimes (Build Crimes category dimension from ODS_CrimesOffences)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.ODS

Tasks

- Truncate destination ODS table
- Load Crimes from ODS_CrimesOffences
- Sort & remove duplicates
- Update timestamp
- Import to ODS_Dim_Crimes

Package ODS_DeptMapping (Transfert to ODS and technical validation)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.ODS

Tasks

- Truncate destination STA table
- Load Department Mapping data from STA
- Convert Tel Prefix to integer
- Transform CodePostal to 2 digit strings (i.e 1 becomes 01)
- Add timestamp
- Save to ODS_DeptMapping

Package ODS_Dim_Departments (Build Department dimension from ODS_DeptMapping)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.ODS

Tasks

- Truncate destination ODS table
- Load data from ODS_DeptMapping
- Sort & remove duplicates
- Convert DepartmentCode to Zip code (i.e. 75 becomes 75000)
- Store in ODS_Dim_Departments

Package ODS_Dim_Regions (Build Region dimension from ODS_DeptMapping)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.ODS

Tasks

- Truncate destination ODS table
- Load data from ODS_DeptMapping
- Sort & remove duplicates
- Store in ODS_Dim_Regions

Package ODS_Dim_PhonePrefix (Build PhonePrefix dimension from ODS_DeptMapping)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.ODS

Tasks

- Truncate destination ODS table
- Load data from ODS_DeptMapping
- Sort & remove duplicates
- Store in ODS_Dim_PhonePrefix

Package ODS_Dim_SchoolZones (Build SchoolZones dimension from ODS_DeptMapping)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.ODS

Tasks

- Truncate destination ODS table
- Load data from ODS_DeptMapping
- Sort & remove duplicates
- Store in ODS_Dim_SchoolZones

Package ODS_Dim_PhonePrefix (Build PhonePrefix dimension from ODS_DeptMapping)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.ODS

Tasks

- Truncate destination ODS table
- Load data from ODS_DeptMapping
- Sort & remove duplicates
- Store in ODS_Dim_PhonePrefix

Package ODS_Dim_Dates (Automaticaly generate Dates dimension)

Variables

- StartDate
- EndDate

Tasks

- Truncate destination ODS table
- Generate date records between StartDate and EndDate
- Add timestamp
- Store to ODS_Dim_Dates

Package ODS_Dim_Periods (Build Periods dimension from ODS_Fact_CrimesOffences)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.ODS

Tasks

- Truncate destination ODS table
- Load data from ODS_Fact_CrimesOffences
- Sort and remove duplicate periods
- Add timestamp
- Store to ODS_Dim_Periods

Package DWH_Dim_Crimes (Load Crimes dimension into DWH)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.DWH

Tasks

- Truncate destination DWH table
- Load Crimes from ODS_Dim_Crimes
- Import to ODS_Dim_Crimes

Package DWH_Dim_Periods (Load Periods dimension into DWH)

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.DWH

Tasks

- Truncate destination DWH table
- Load Crimes from ODS_Dim_Periods
- Import to DWH_Dim_Periods

Package DWH_Dim_Dates (Load Dates dimension into DWH)

Depends on DWH_Dim_Periods

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.DWH

Tasks

- Truncate destination DWH table
- Load Crimes from ODS_Dim_Dates
- Lookup respective Period from DWH_Dim_Periods
- Import to DWH_Dim_Dates

Package DWH_Dim_Dates (Load Dates dimension into DWH)

Depends on DWH_Dim_Periods

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.DWH

Tasks

- Truncate destination DWH table
- Load Crimes from ODS_Dim_Dates
- Lookup respective Period from DWH_Dim_Periods
- Import to DWH_Dim_Dates

Packages DWH_Dim_Departments/DWH_Dim_Regions/DWH_Dim_PhonePrefix/DWH_Dim_SchoolZones

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.DWH

Tasks

- Truncate destination DWH table
- Load data from repective tables in ODS
- Import to respective tables in DWH

Package DWH_Fact_CrimesOffences (Load Crimes Fact table and lookup all fields from DWH)

TODO: Functional validation

Connection

- IN: SQLSRV.ODS
- OUT: SQLSRV.DWH

Tasks

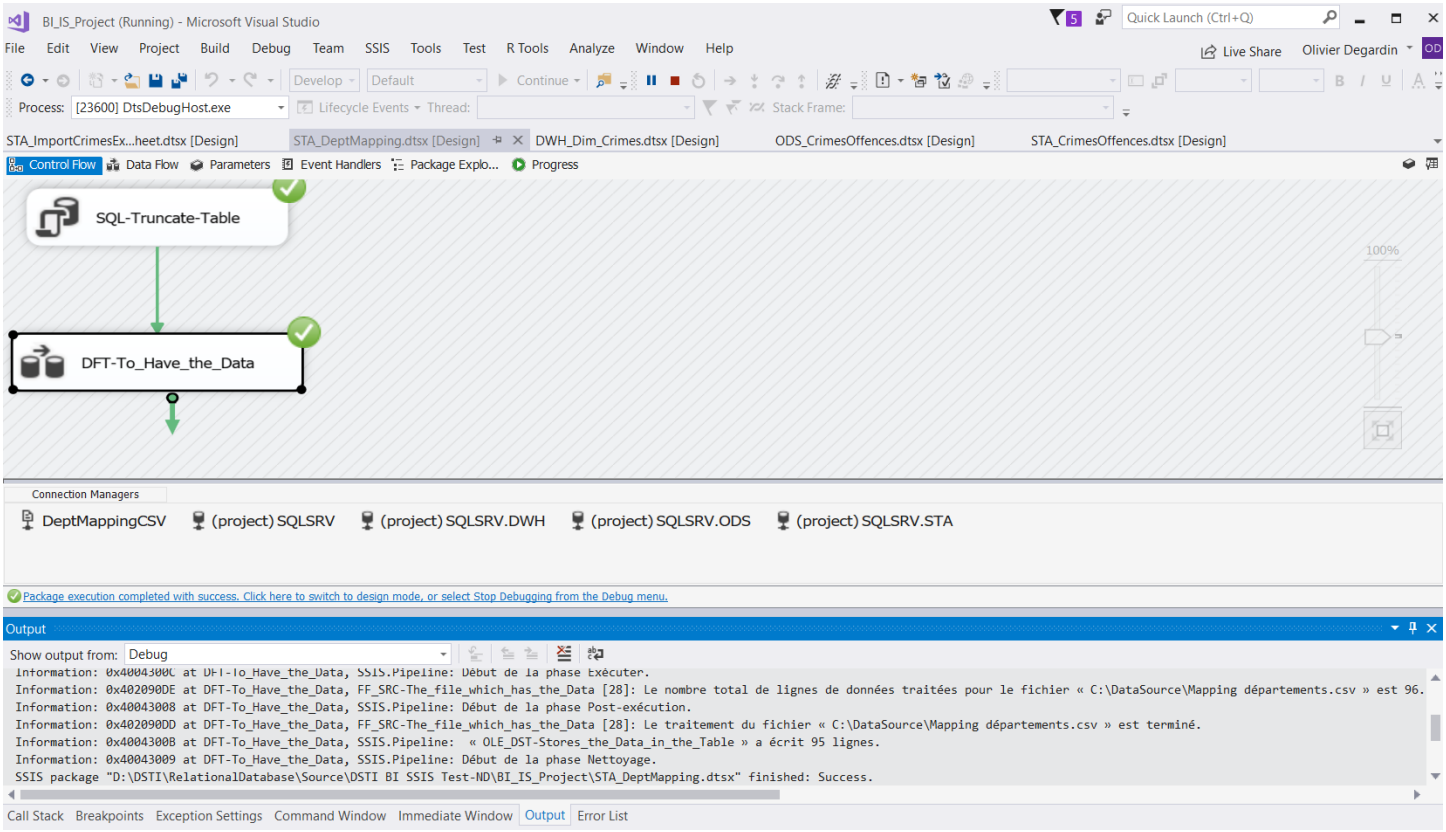
- Truncate destination DWH table
- Load data from ODS_Fact_CrimesOffences
- Lookup Crime Key
  - Unmatched results to Functional reject
- Lookup Period Key
  - Unmatched results to Functional reject
- Lookup Department Key
  - Unmatched results to Functional reject
- Lookup Region Key
  - Unmatched results to Functional reject
- Lookup PhonePrefix Key
  - Unmatched results to Functional reject
- Lookup SchoolZones Key
  - Unmatched results to Functional reject
- Import to DWH_Fact_CrimesOffences

8. SSIS Package process – single sample

8.1 Description of a package

'STA_DeptMapping' package (STA_DeptMapping.dtsx) - displayed schema via Visual Studio.

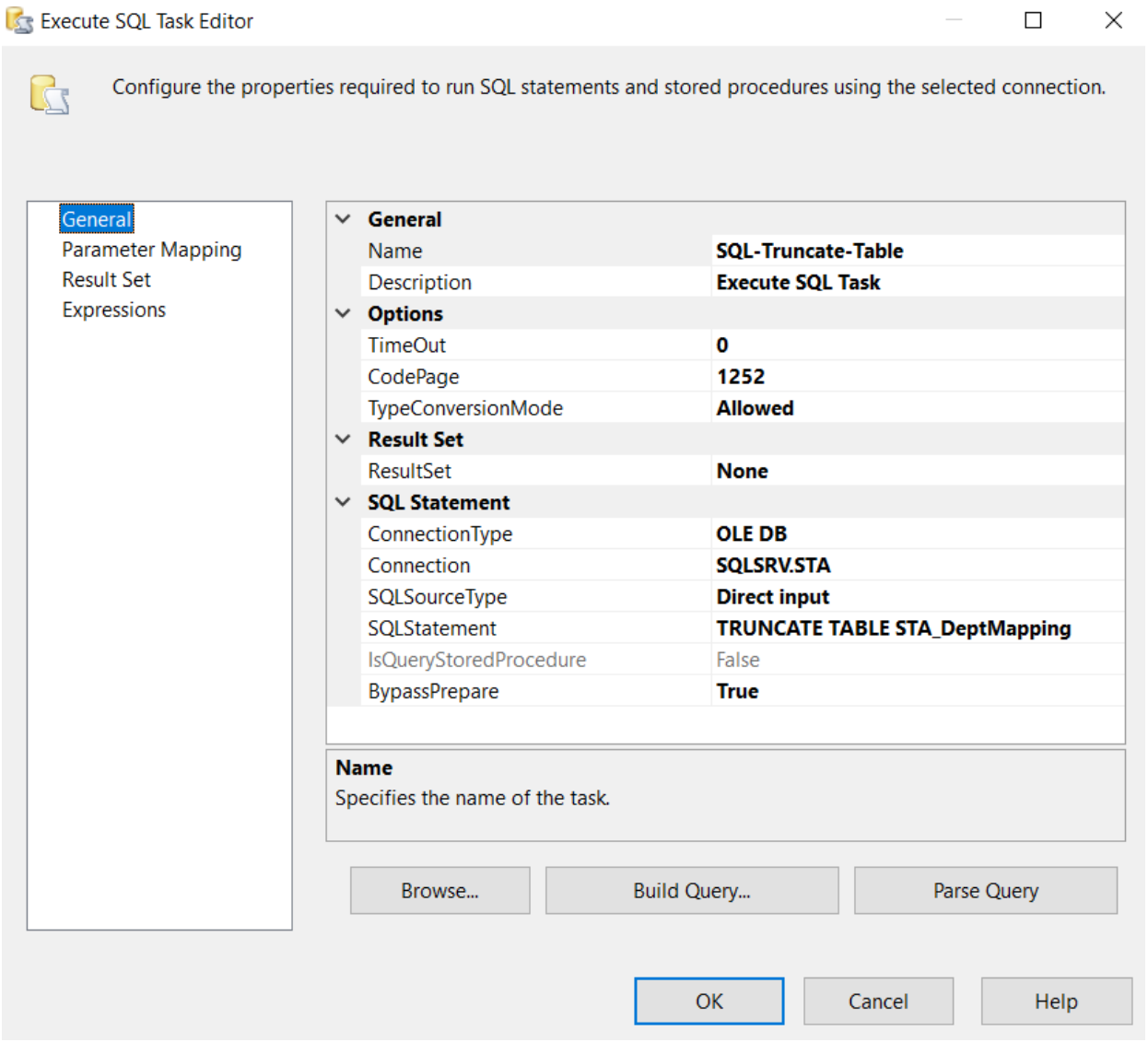
A/ 'Control Flow' section



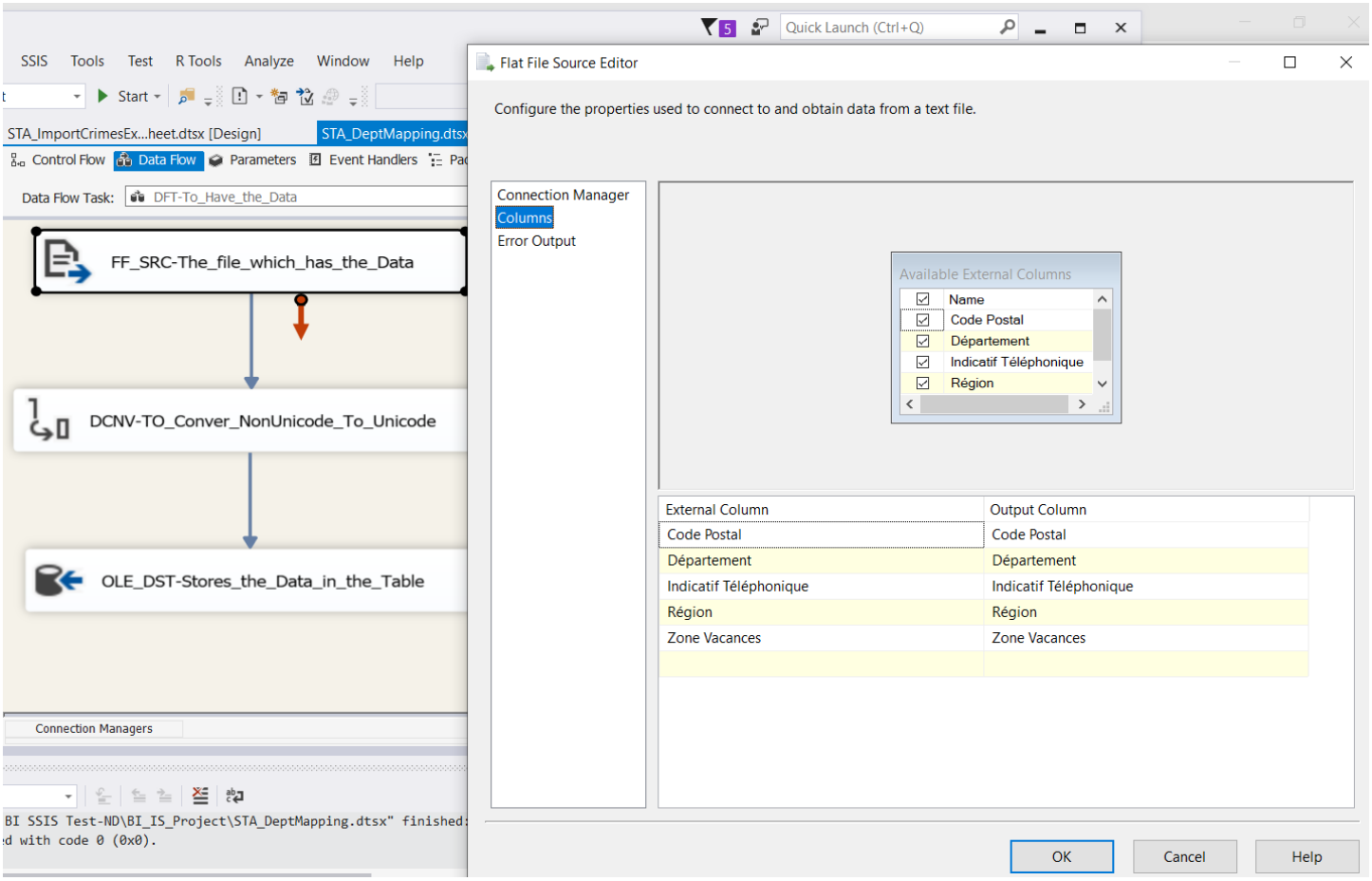
8.2 'Execute SQL Task Editor' process

Edit of the first task 'SQL-Truncate table': it removes the content of the table STA_DeptMapping which was created using Microsoft SQL Server Management Studio – SQL request for the table's creation below.

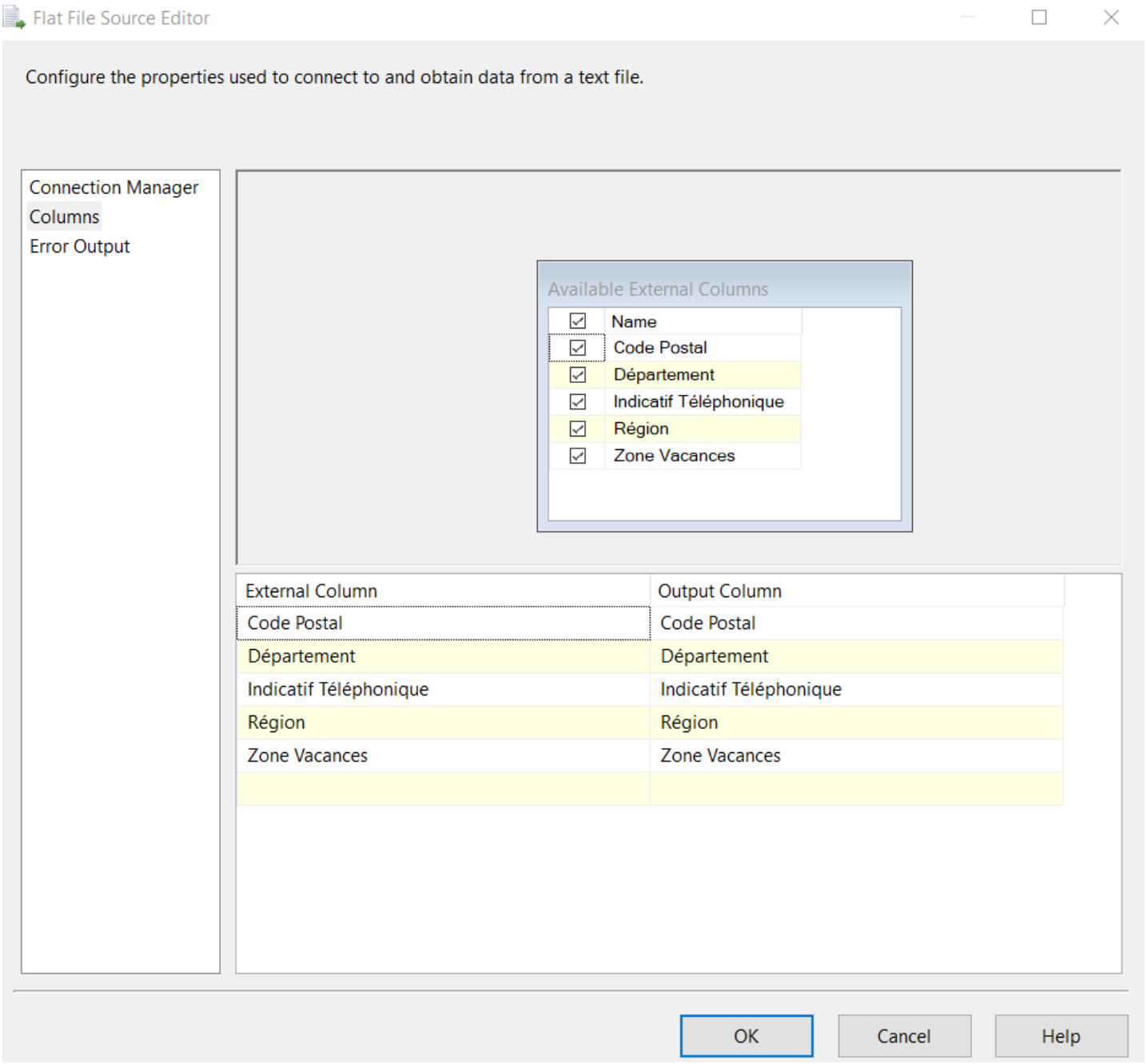
```
CREATE TABLE [dbo].[STA_DeptMapping](
    [Code Postal] [nvarchar](255) NULL,
    [Département] [nvarchar](255) NULL,
    [Indicatif Téléphonique ] [nvarchar](255) NULL,
    [Région] [nvarchar](255) NULL,
    [Zone Vacances] [nvarchar](255) NULL
) ON [PRIMARY]
GO
```



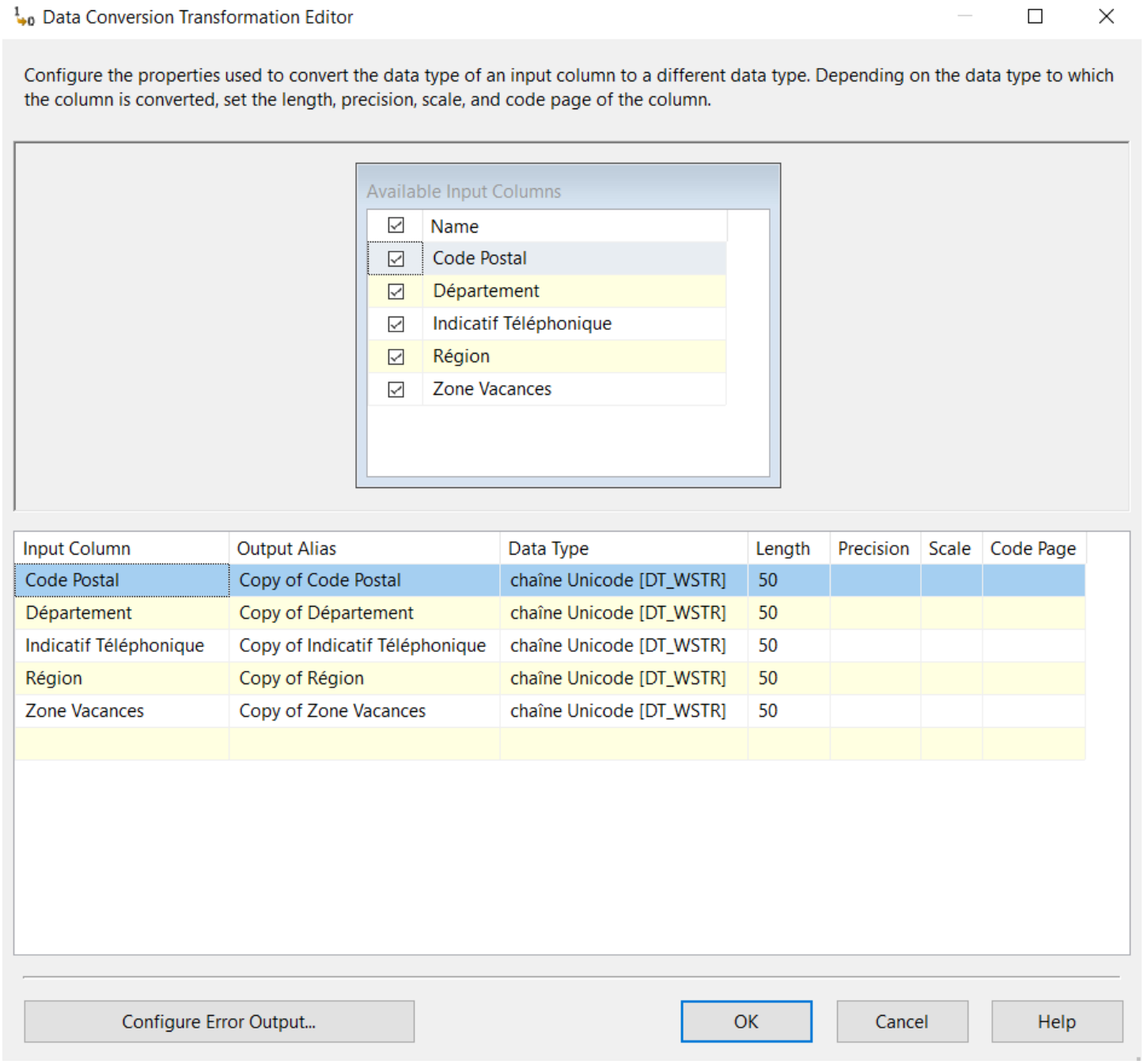
B/ ‘Data Flow’ section



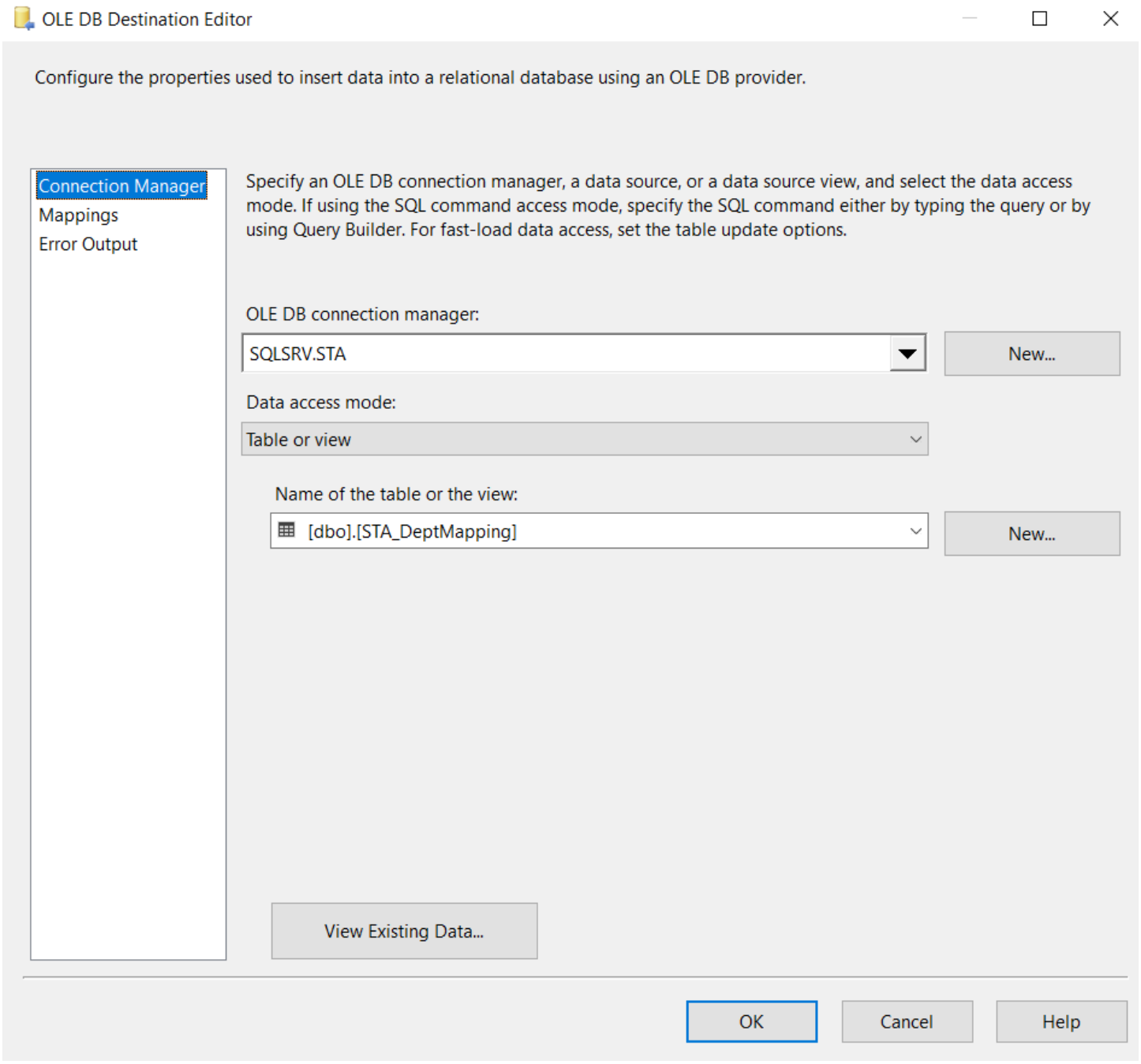
8.3 Flat File Source Editor description



8.4 Data Conversion Transformation Editor



8.5 OLE DB Destination Editor



9. Conclusion

After one week working together in brainstorming ideas, discussing model …, we initially understood the development of a BI project through steps such as STA, ODS and then DWH starting from untreated natural data sources. We have also operated the team with the setting up as a real project execution environment in a company.