

ETL (*Extract, Transform, Load*) Tools

(Интеграционные услуги
в MS SQL Server)

ETL (*extract, transform, load*)

- *Възниква във връзка с концепцията за складове от данни*
- *Три функции, които са комбинирани в един инструмент за извличане на данни от източник (напр. база от данни) и поставянето им в склад от данни (приемник).*

Какво е ETL?

- **Extract** - Извличане на данни от източника
- **Transform** - Трансформиране на данните
- **Load** - Зареждане в дестинацията



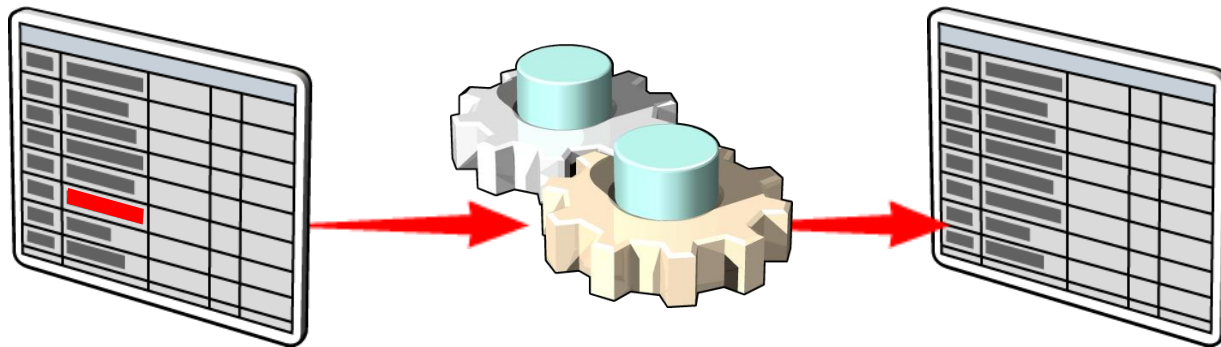
Extract - Извличане на данни

- Повечето проекти за складиране на данни консолидират данни от различни системи-източници.
- Всяка отделна система може да използва различна организация на данните и/или различни формати.
- Източници на данни могат да бъдат: релационни бази от данни, XML и плоски файлове (flat files); могат да се включват и нерелационни бази от данни и др. формати.

Transform

Трансформация на данните включва:

- Почистване на данните
- Промяна формата на данните
- Агрегиране на данните
- Правене на данните съвместими
- Валидиране на данните



Load - зареждане на данните в дестинацията

- Дестинацията-приемник на данните може да бъде склад от данни, база от данни, плоски файлове и други формати.
- Взависимост от изискванията процесът може да включва:
 - Препокриване на съществуващите данни;
 - Обновяване на съществуващите данни периодично –ежедневно, седмично, месечно;
 - Добавяне на нови данни.
- При зареждане на данните в дестинацията върху тях се прилагат правилата за интегритет на данните.

Индустриални ETL инструменти

- Oracle Data Integrator (ODI)
- Microsoft SQL Server Integration Services (SSIS)
- Pentaho Data Integration или Kettle (www.pentaho.com/)
- Microstrategy (www.microstrategy.com)
- Informatica PowerCenter (etl-tools.info/informatica/tutorial.html)
- IBM InfoSphere Datastage (www.ibm.com)
- Ab Initio (www.abinitio.com/)

MS SQL Server

интеграционни услуги

- **SQL Server Integration Services (SSIS)** – корпоративно решение за извличане, трансформиране и интеграция на данни
- SSIS включват
 - богат набор от вградени задачи и трансформации;
 - инструменти за конструиране на пакети;
 - услуги за стартиране и управление на пакетите.
- SSIS съдържа графични инструменти и съветници за извличане, трансформиране и зареждане на данни

SSIS терминология

- Конекции (Connections) – съхраняват информация за източника или дестинацията (получателя) на данните.
- Пакети (Packages) – единици от работа, които могат да се съхраняват и изпълняват.
- Задачи (Tasks) – изпълняват някаква работа в пакетите.

Задачи, които могат да се изпълняват в SSIS

- ❑ Преместване на данни между хетерогенни системи (напр., Oracle в SQL Server или обратно).
- ❑ Преместване на данни между две системи, използващи SQL Server.
- ❑ Преместване на данни от MS Access или MS Excel в SQL Server или обратно.
- ❑ Извличане на данни, трансформирането им чрез асоцииране на колони, попълване на липсващи стойности, преобразуване формата на данните и др. и импортирането им в крайната система.

Пакети на Integration Services

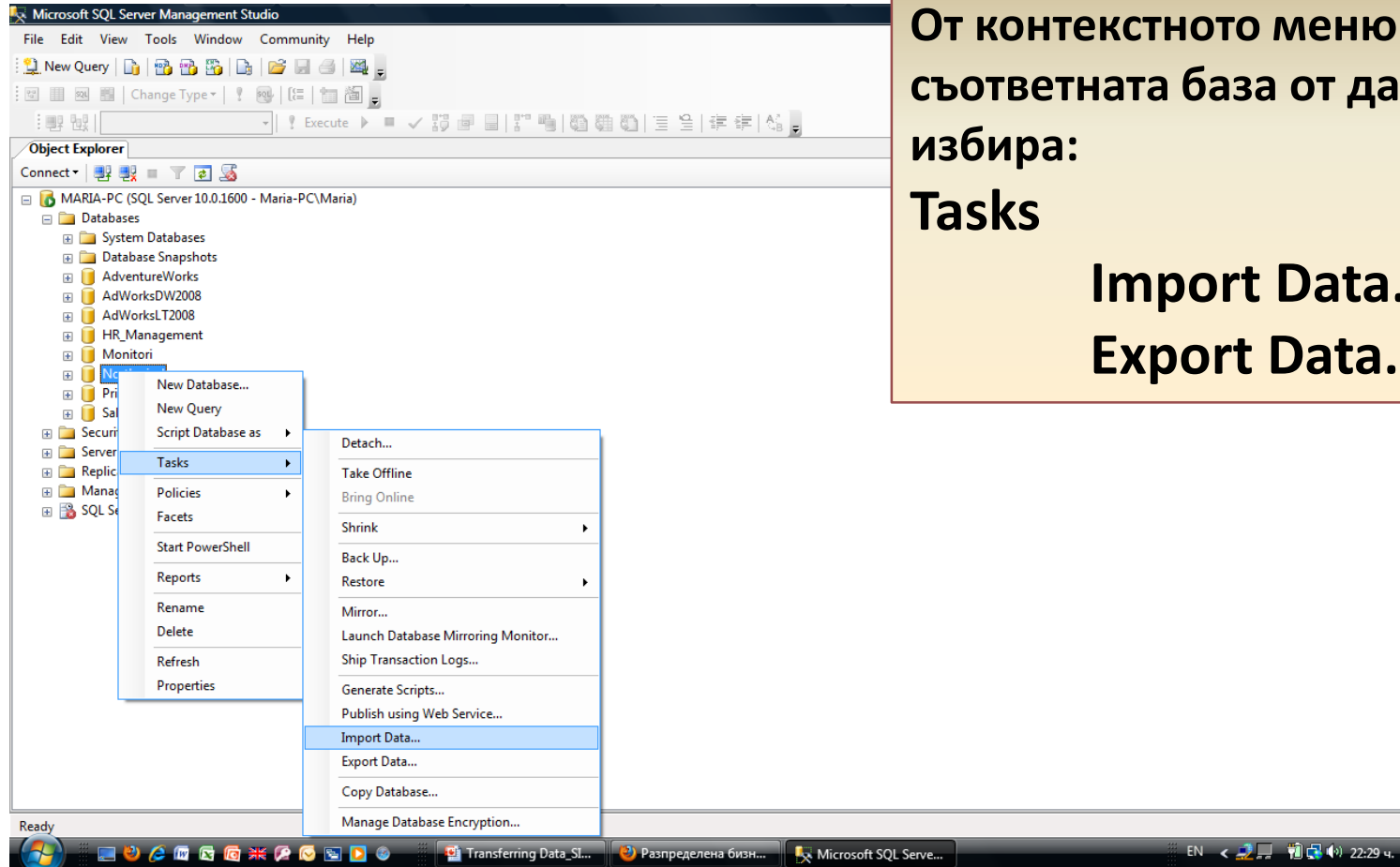
□ Пакетите са набори от задачи за импортиране, трансформиране и експортиране на данни, които могат да се

- Да се използват повторно;
- Да се стартират по график толкова често, колкото е необходимо.

□ Пакетите могат да бъдат:

- Съхранени в системната база от данни msdb на локален или отдалечен сървър;
- Записани като DTSX файлове (полезно при копиране, преместване на пакетите на друго място или изпращане по email).

Съветник за създаване на пакети Import and Export Wizard



От контекстното меню за
съответната база от данни се
избира:

Tasks

Import Data...

Export Data...

Използване на Import and Export Wizard за създаване на пакети



Етап 1: Конфигуриране на
източника и дестинацията

Етап 2: Копиране или
изграждане на на заявка

Етап 3: Форматиране и
трансформация

Етап 4: Записване и
изпълнение

Етап 1: Конфигуриране на източника и дестинацията

- Избор на източника на данни
 - Зависи от източника на данните и включва няколко типа конекции:
 - Flat File Source (Плоски файлове)
 - Microsoft Access (Microsoft Access Database Engine)
 - Microsoft Excel
 - Сървърно-базирани конекции до SQL Server
 - Сървърно-базирани конекции до бази от данни, различни от SQL Server
 - NET Framework за ODBC, Oracle и SQL Server
 - Задава се допълнителна информация, необходима за осъществяване на връзката (зависи от типа на източника)

Етап 1: Конфигуриране на източника и дестинацията

- Избор на дестинация (получател на данните)
 - SQL база от данни
 - MS Access база от данни
 - MS Excel таблици
 - Текстови файлове
 -
- Задаване на допълнителна информация, която зависи от дестинацията

Връзки със SQL Server

- Връзката се съществява със
 - **SQL Native Client**
 - или **Microsoft OLE DB Provider for SQL Server**
- Конфигурирането на връзката включва задаването на:
 - **Server Name** – име на сървъра
 - **Метод на удостоверяване** при достъп до сървъра (ако е SQL Server Authentication допълнително се задава потребителско име и парола)
 - **Database** – име на базата от данни

Конекции за файлово-базирани данни (напр. MS Access и Excel)

- За MS Access се задава:
 - File Name – пълна спецификация на файла (маршрут)
 - User name
 - Password
- За MS Excel се задава:
 - Excel File Path – пълна спецификация на файла (маршрут)
 - Excel Version – версия
 - Опцията First Row Has Column Names се изключва, ако първият ред на електронната таблица не съдържа етикети на колони

Импортиране на плоски файлове

- Ако източникът е плосък файл се избира Flat File Source
- Задава се пълното име на файла и пътя за достъп до него
- Страницата на съветника се актуализира и в нея се задава допълнителна информация:
 - Кодова страница (Code Page)
 - Тип на файла (Format) – определя използвания разделител между колоните във файла (напр. Delimited, ако колоните са разграничени със запетая или друг уникален знак)
 - Определител за текста (Text Qualifier) – напр. Double Quote(“) и др.

Етап 2: Копиране на данни или изграждане на заявка

SQL Server Import and Export Wizard

Specify Table Copy or Query
Specify whether to copy one or more tables and views or to copy the data by using a query from the data source.

☒ **Copy data from one or more tables or views**
Use this option to copy all the data from the existing tables or views in the source database.

☐ **Write a query to specify the data to transfer**
Use this option to write an SQL query to manipulate or to restrict the source data for the copy operation.

Help < Back Next > Finish >>| Cancel

Избор на таблици и изгледи, които ще бъдат копирани

Изграждане на заявка, определяща обектите за прехвърляне

Етап 2: Копиране или изграждане на заявка

Копиране на данни:







- Избор на таблици и изгледи, които ще бъдат копирани
 - Маркират се таблиците, които ще се копират;
 - Приемащите таблици по подразбиране приемат същото име (могат да се променят);
 - С бутон Edit Mappings може да се зададе как да се обработят колоните на таблицата (етап 3).

Select Source Tables and Views

Choose one or more tables and views to copy.



Tables and views:

<input checked="" type="checkbox"/>	Source: C:\Samples\DATA\Customers.xls	Destination: WIN-6C20A8VN2U0
<input checked="" type="checkbox"/>	 'Customers'	 [dbo].[Customers]
<input type="checkbox"/>	 'Customers\$'	
<input type="checkbox"/>	 'Data_klienti'	
<input type="checkbox"/>	 'Orders'	
<input type="checkbox"/>	 'Orders\$'	

Edit Mappings...

Preview...

Help

< Back

Next >

Finish >>|

Cancel

Етап 2: Копиране или изграждане на заявка

- Изграждане на заявка, определяща обектите за прехвърляне
 - Може да се напише директно в прозореца Provide A Source Query;
 - Да се отвори по-рано създадена заявка;
 - Да се създаде заявка в SQL Server Management Studio:
 - * Писане и тестване в New Query или използване на Query Designer;
 - * Копиране на заявката в прозореца Provide A Source Query.

Provide a Source Query

Type the SQL statement that will select data from the source database.



SQL statement:

```
SELECT C.CompanyName, OD.ProductID  
      , P.ProductName, P.UnitPrice  
      , OD.Quantity, O.OrderID  
      , convert(varchar(20),O.OrderDate, 102) AS OrdersDate  
FROM Products AS P JOIN [Order Details] AS OD  
ON P.ProductID = OD.ProductID  
INNER JOIN Orders AS O  
ON OD.OrderID = O.OrderID  
INNER JOIN Customers AS C  
ON O.CustomerID = C.CustomerID  
ORDER BY C.CompanyName,O.OrderDate
```

Parse

Browse...

Help

< Back

Next >

Finish >>

Cancel

Етап 3: Форматиране и трансформация

- Трансформацията е процеса на обработване на изходните данни и форматирането им за посочената дестинация.
- По подразбиране асоциирането на таблиците от източника и дестинацията следва определени правила:
 - Всички колони се копират в изходната таблица;
 - Пренасят се оригиналните имена на таблиците и колоните, типът на данните, точността, порядъка и правилото за Null;
 - Изходните данни се добавят към приемащата таблица или ако няма такава се създава.

Етап 3: Форматиране и трансформация

- Имената на колоните и типа на данните в таблицата-приемник се определят въз основа на асоциациите с колоните на източника по подразбиране. Могат да се разгледат и модифицират (Edit Mappings):
 - **Source** – колона от източника;
 - **Destination** – приемаща колона; <ignore>, за да не се създава колоната източник в целевата таблица;
 - **Type** – тип на данните за целевата колона;
 - **Nullable** – дали позволява стойност Null;
 - **Size** – размер на целевата колона;
 - **Precision** – максимален брой цифри;
 - **Scale** – брой цифри след десетичната запетая.

Column Mappings

Source: `Customers`
 Destination: [dbo].[Customers]

- ☐ Create destination table Edit SQL...
- ☐ Delete rows in destination table ☐ Drop and re-create destination table
- ☒ Append rows to the destination table ☐ Enable identity insert

Mappings:

Source	Destination	Type	Nullable	Size	Precision	Scale
CustomerID	CustomerID	nchar	<input type="checkbox"/>	5		
CompanyName	CompanyName	nvarchar	<input type="checkbox"/>	40		
ContactName	ContactName	nvarchar	<input checked="" type="checkbox"/>	30		
ContactTitle	ContactTitle	nvarchar	<input checked="" type="checkbox"/>	30		
Address	Address	nvarchar	<input checked="" type="checkbox"/>	60		
City	City	nvarchar	<input checked="" type="checkbox"/>	15		
Region	Region	nvarchar	<input checked="" type="checkbox"/>	15		
PostalCode	PostalCode	nvarchar	<input checked="" type="checkbox"/>	10		
Country	Country	nvarchar	<input checked="" type="checkbox"/>	15		

Source column: CustomerID VarChar (255)

OK

Cancel

Етап 3: Форматиране и трансформация

- Правилата за асоцииране за всяка от таблиците могат да се предефинират:
 - *Create Destination Table* - създаване на приемащата таблица
 - *Delete Rows In Destination Table* – изтриване на редовете от приемащата таблица
 - *Append Rows In Destination Table* – добавяне на редове в приемащата таблица
 - *Drop and Re-create Destination Table* – премахване и създаване отново на приемащата таблица

Етап 4: Записване и изпълнение

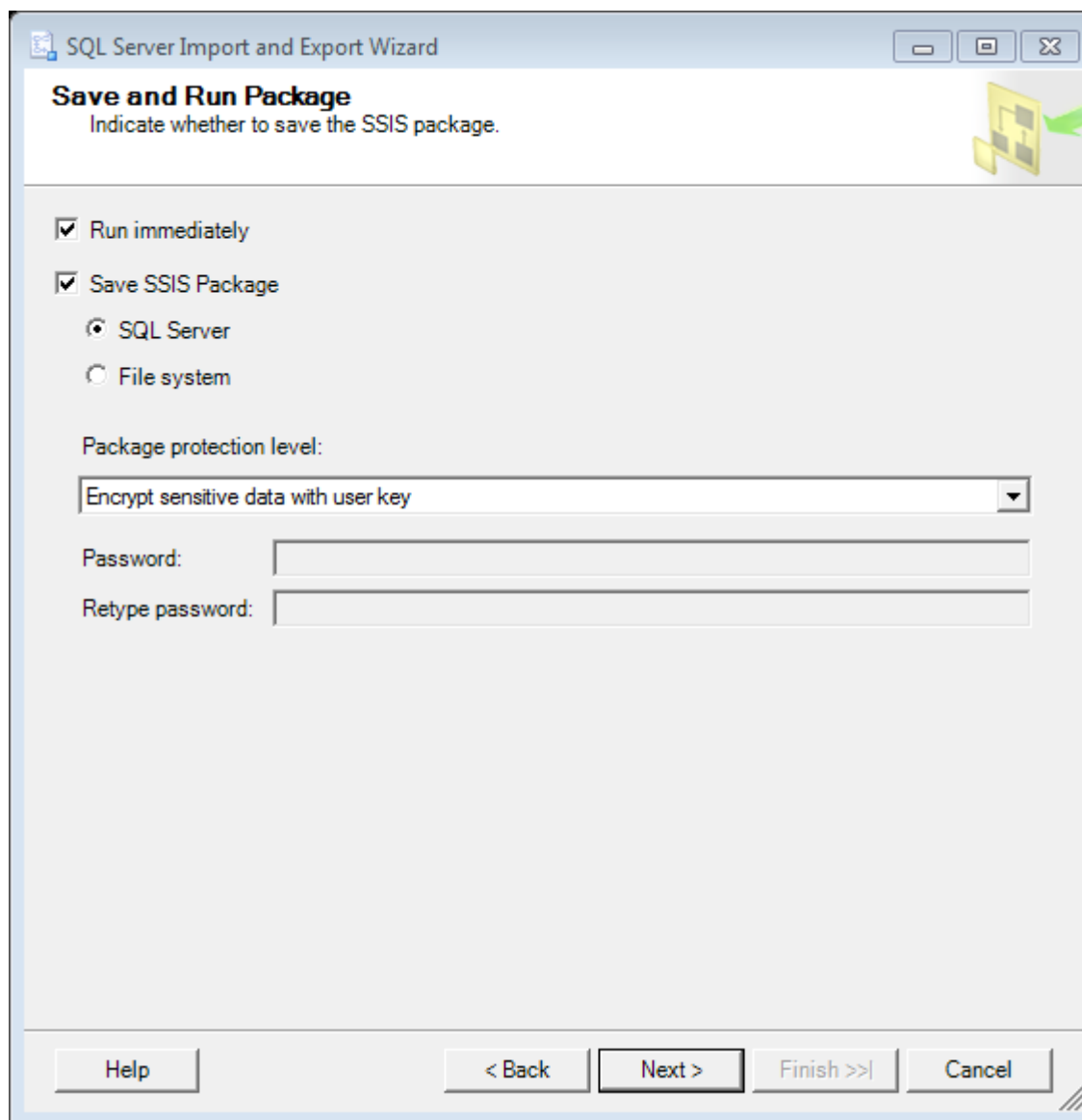
- Задава се:
 - Кога да се изпълнява създадения пакет:
Execute Immediately (Незабавно изпълнение) опцията по подразбиране е включена.
 - Дали да се съхрани за по-нататъшна употреба.

В областта Save се задават опциите как да се съхрани пакетът:

SQL Server – пакетът се записва като локален в базата от данни msdb

File System – пакетът се записва като .dtsx файл

Записване и изпълнение



The screenshot shows the 'Save and Run Package' step of the SQL Server Import and Export Wizard. The window title is 'SQL Server Import and Export Wizard'. The main heading is 'Save and Run Package' with the instruction 'Indicate whether to save the SSIS package.' and a yellow folder icon with a green arrow.

Options:

- ☒ Run immediately
- ☒ Save SSIS Package
 - ☒ SQL Server
 - ☐ File system

Package protection level:

Encrypt sensitive data with user key (selected in dropdown)

Password: [text box]

Retype password: [text box]

Navigation buttons at the bottom: Help, < Back, Next > (highlighted), Finish >>, Cancel.

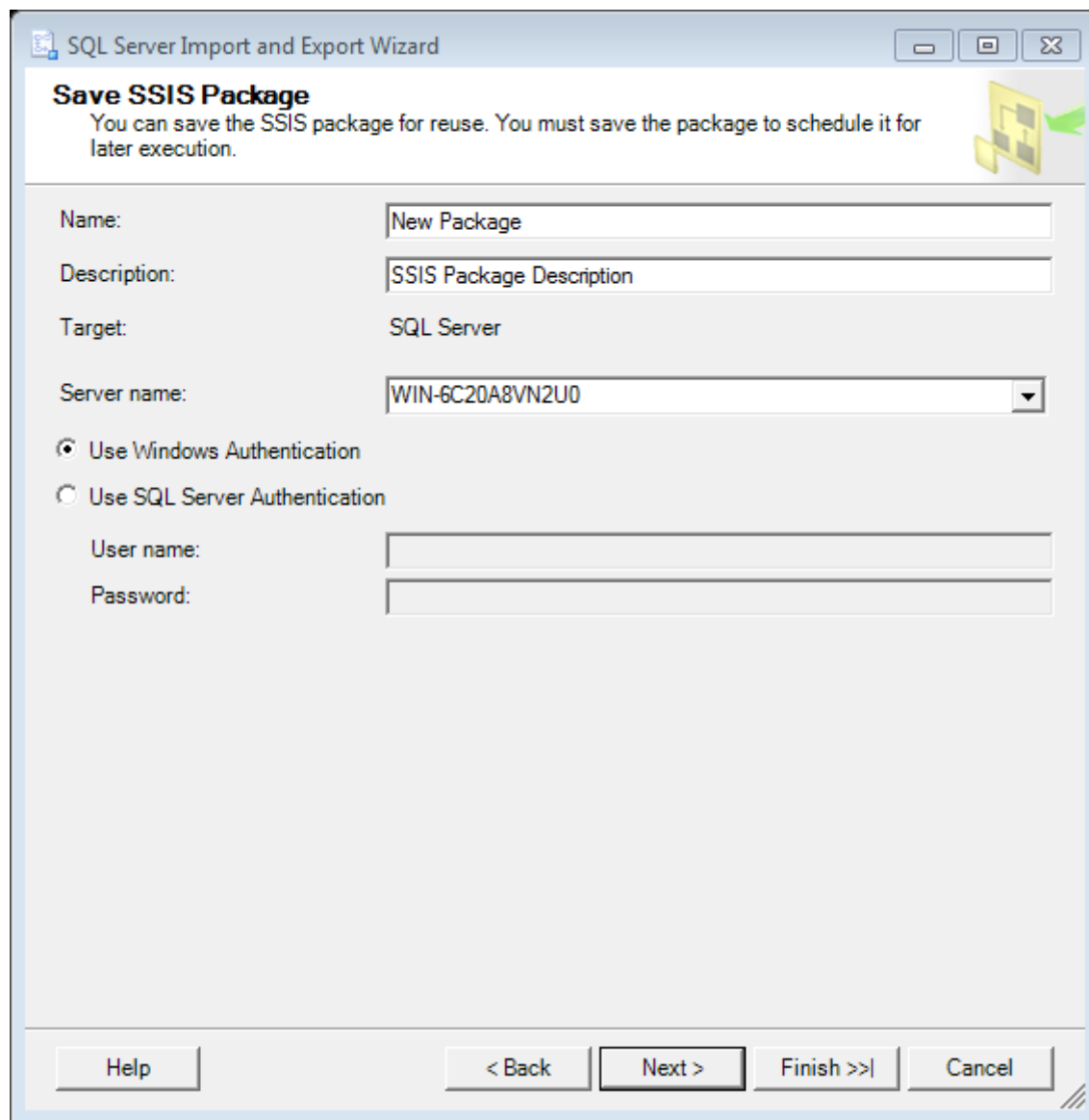
Опции за криптиране на пакета

- *Do Not Save Sensitive Data* – в пакета не се записват конфиденциални данни;
- *Encrypt Sensitive Data With User Key* - в пакета се записват криптирани конфиденциални данни; може да се отваря и изпълнява само от потребителя, който го е създадал;
- *Encrypt Sensitive Data With Password* - в пакета се записват криптирани конфиденциални данни; може да се отваря и изпълнява само при въвеждане на предварително зададената паролна дума;

Опции за криптиране на пакета

- *Encrypt All Data With User Key* – всички данни в пакета се криптират; може да се отваря и изпълнява само от потребителя, който го е създадал
- *Encrypt All Data With Password* - всички данни в пакета се криптират; може да се отваря и изпълнява само при въвеждане на предварително зададената паролна дума;
- *Rely On Server Storage And Roles For Access Control* – само при записване пакета в SQL Server (използват се права и роли на SQL Server за контрол на достъпа).

Съхраняване в SQL Server



The screenshot shows the 'Save SSIS Package' step of the SQL Server Import and Export Wizard. The window title is 'SQL Server Import and Export Wizard'. The main heading is 'Save SSIS Package' with a sub-message: 'You can save the SSIS package for reuse. You must save the package to schedule it for later execution.' There is a yellow folder icon with a green arrow pointing to it.

The form contains the following fields and options:

- Name:** Text box containing 'New Package'.
- Description:** Text box containing 'SSIS Package Description'.
- Target:** Text box containing 'SQL Server'.
- Server name:** Dropdown menu showing 'WIN-6C20A8VN2U0'.
- Authentication:** Two radio buttons: 'Use Windows Authentication' (selected) and 'Use SQL Server Authentication'.
- User name:** Text box (empty).
- Password:** Text box (empty).

At the bottom, there are five buttons: 'Help', '< Back', 'Next >', 'Finish >>', and 'Cancel'.

Съхраняване във файловата система

SQL Server Import and Export Wizard

Save SSIS Package
You can save the SSIS package for reuse. You must save the package to schedule it for later execution.

Name: New Package

Description: SSIS Package Description

Target: File System

File name: C:\Users\MK\Documents\New Package.dtsx

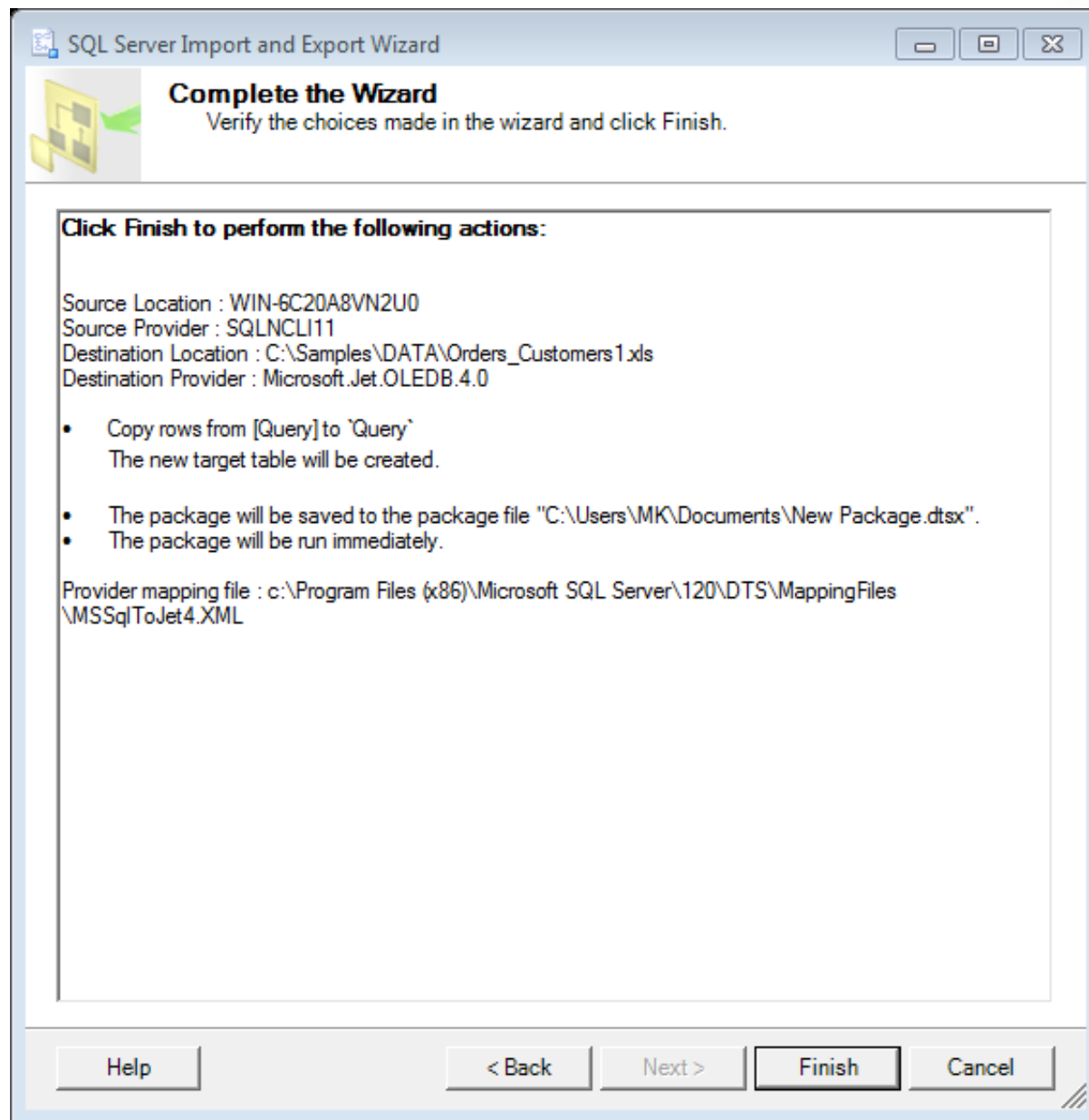
Browse...

Help < Back Next > Finish >> Cancel

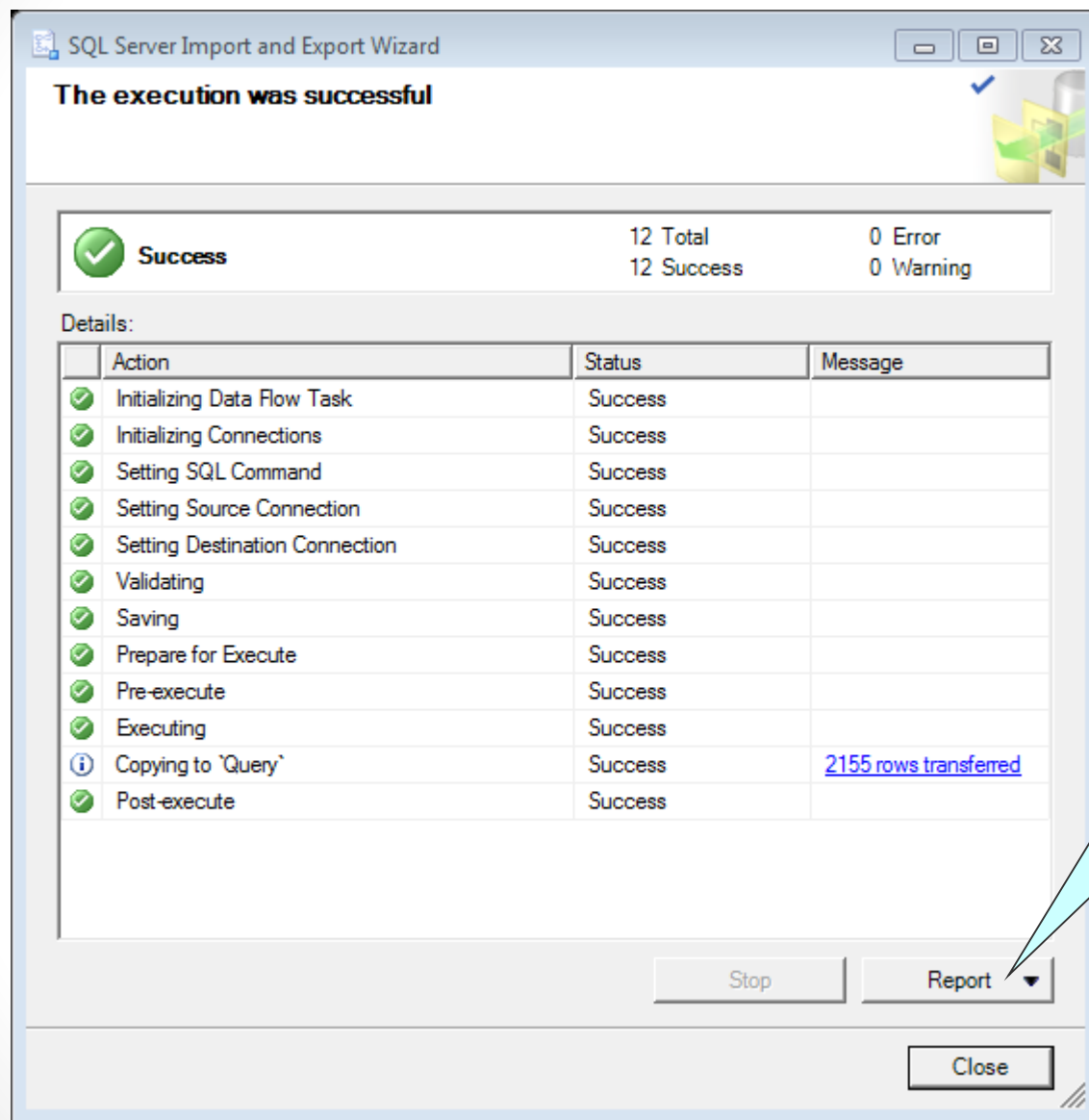
Име на пакета

Име на файла и път до него

Информация
за пакета
(при
кликване на
Next в
прозореца за
Save SSIS
Package)



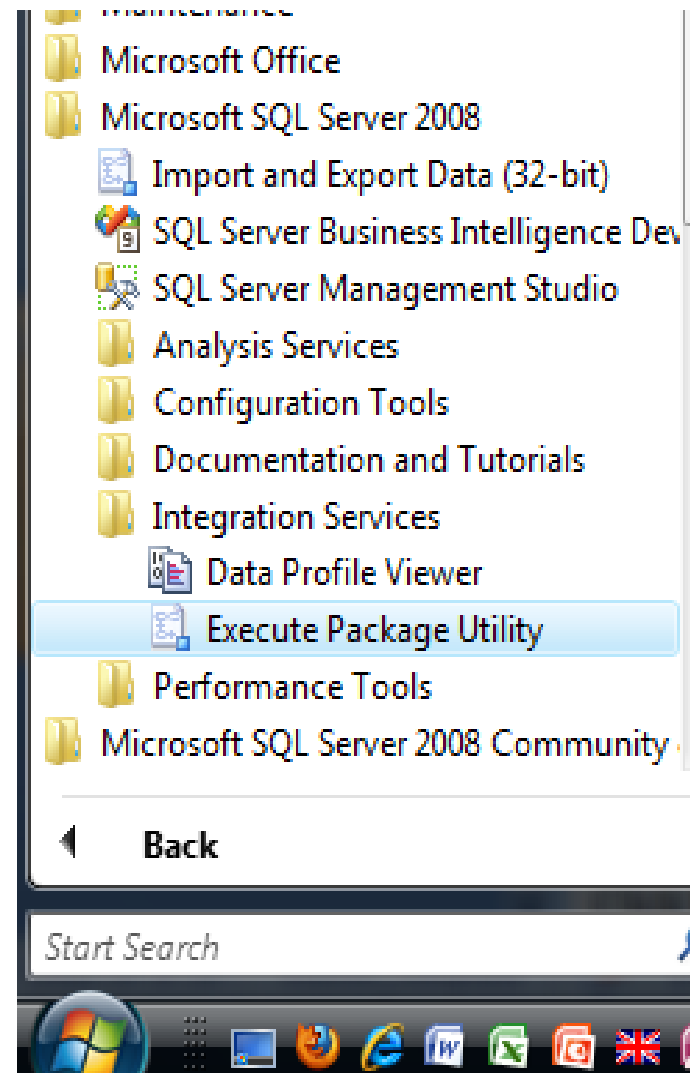
Изпълнение на пакета



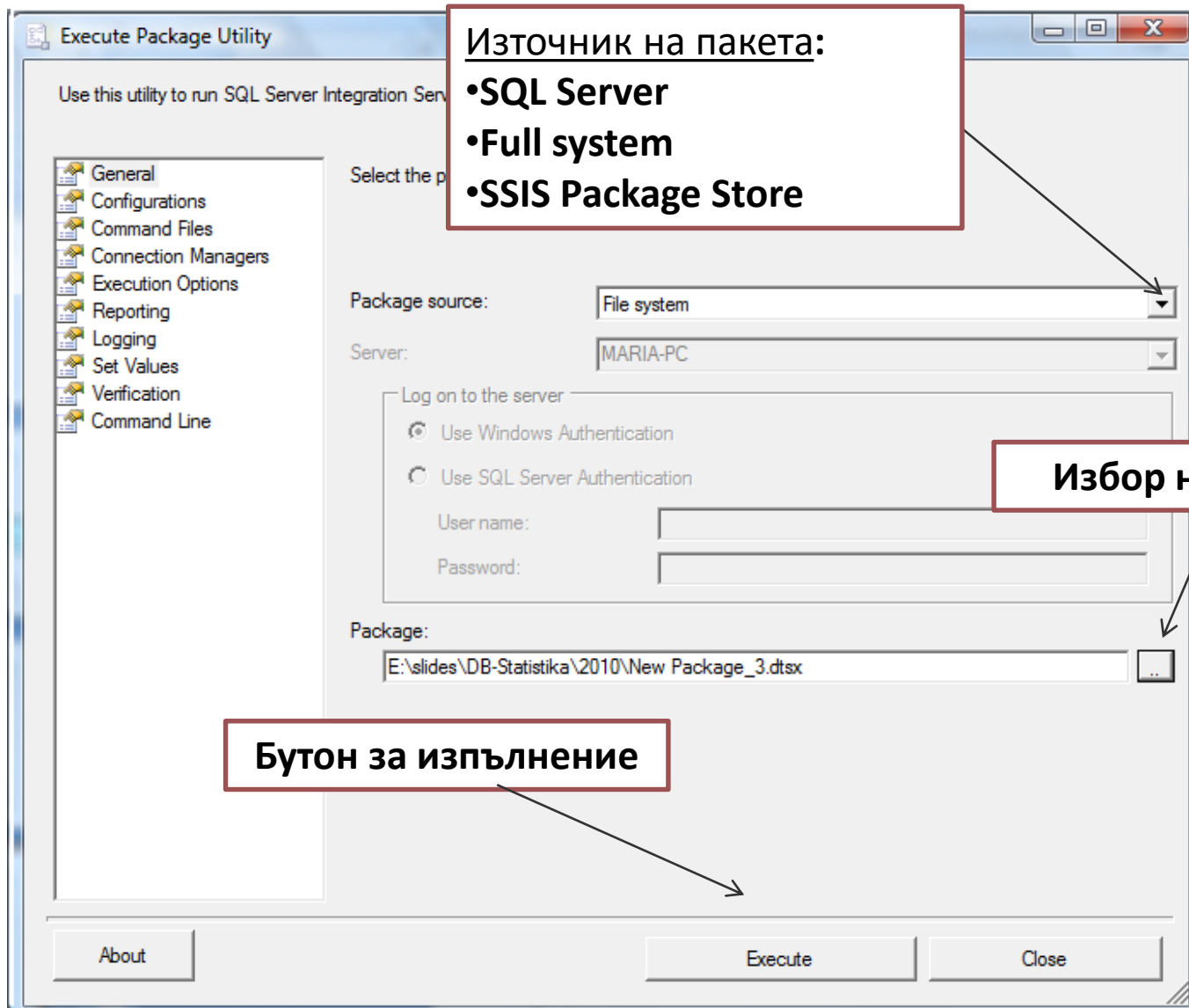
Получаване на
отчет
за изпълнението
на пакета

Стартиране на съхранените пакети за изпълнение

- Съхранените пакети за трансфер на данни могат да се стартират чрез услугата **Execute Package Utility** на **Integration Services**

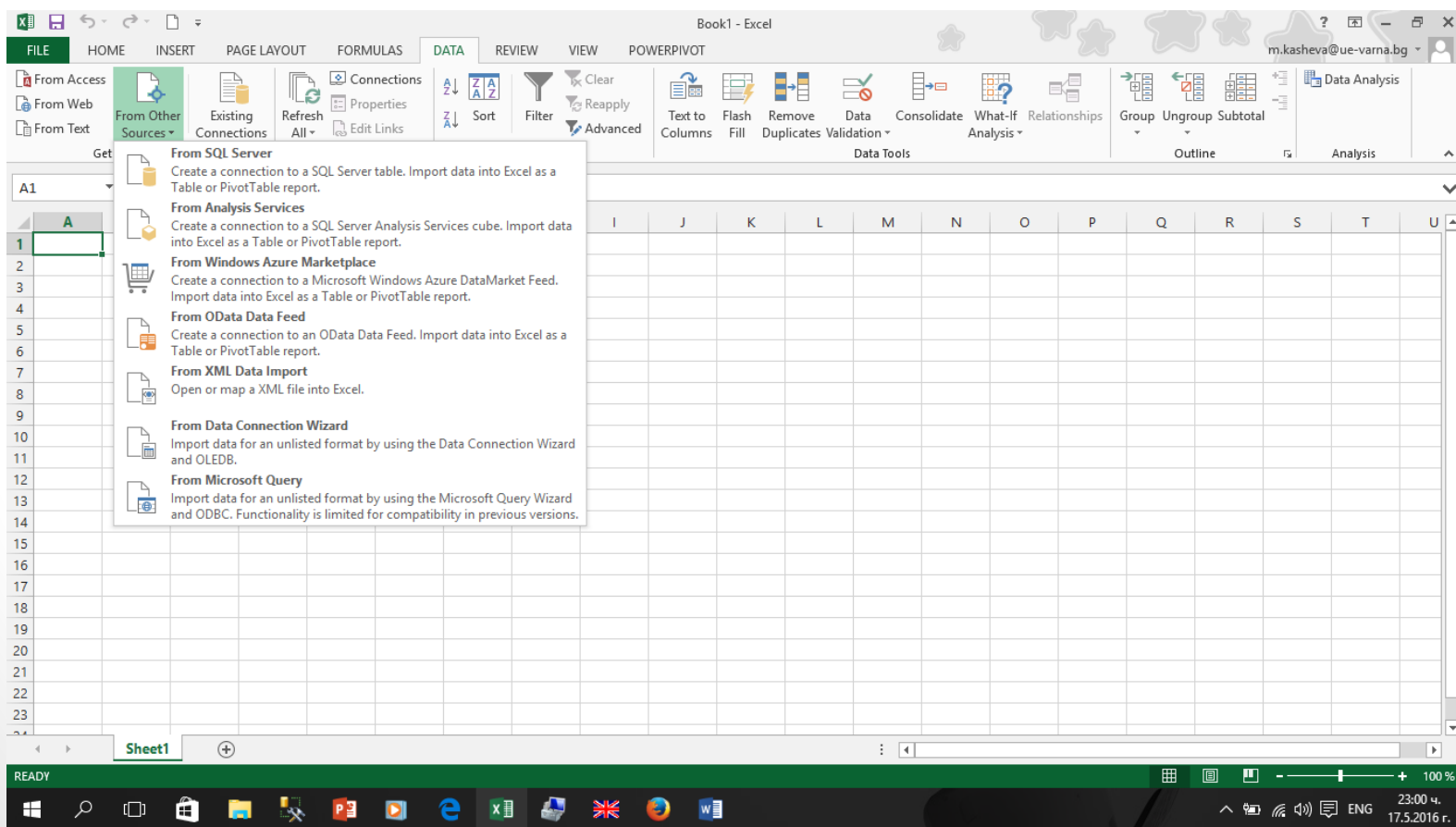


Задаване на съхранения пакет за изпълнение



Импорт на данни от SQL база от данни в Excel

- Отваряне на нова таблица в Excel
- От Data избор от Get External Data на Fro SQL Server



- Да се създадат пакети, които включват:
 - В БД OrdersDW създаване и зареждане на данни от БД Northwind на таблиците Customers, Products, Categories, при което да се премахнат излишните данни.
 - Създаване на таблица Total_Orders и зареждане на данни в нея чрез изгледа със същото име от БД Northwind
 - Създаване и зареждане на данни от Excel таблица Suppliers в аналогична таблица в базата от данни
- Пакетите да се съхранят като файлове .dtsx