**Лабораторная работа № 13**

Построение хранилища данных

**Цель работы:** построить хранилище данных.

**Теоретическая база для проектирования хранилища данных**

**Бизнес-аналитика** (BI) – это область технологий баз данных, направленная на извлечение значимой информации из данных для поддержки принятия бизнес-решений.

Основана на интеграции данных из различных источников в единое хранилище.

Позволяет конечным пользователям выполнять нерегламентированные запросы для анализа данных и поиска закономерностей.

**Цель**: улучшение бизнес-решений за счет анализа данных.

**OLTP** (Online Transaction Processing): Системы, предназначенные для оперативной обработки транзакций (например, онлайн-заказы, банковские операции). Характеристики:

* Реляционные базы данных.
* Множество коротких транзакций.
* Большое количество пользователей.
* Частые операции чтения и записи.

**OLAP** (Online Analytical Processing): Системы, предназначенные для анализа данных и поддержки принятия решений. Характеристики:

* Данные собираются из различных источников и приводятся к согласованному виду.
* Регулярная загрузка данных в базу данных (хранилище или витрину данных).
* Преимущественно режим чтения.
* Относительно небольшое количество одновременных пользователей.
* Создание отчетов и выполнение сложных запросов.

**Хранилище данных** (Data Warehouse) – централизованная база данных, содержащая все данные организации, предназначенная для анализа.

Обеспечивает единообразный доступ к данным для всех пользователей.

Проблемы: Большой объем данных, данные из разных источников и в разное время.

Решение: Консолидация данных (ETL-процесс).

**ETL-процесс** (Extract, Transform, Load):

* Extract (извлечение): Сбор данных из различных источников.
* Transform (преобразование): Очистка, преобразование и приведение данных к единому формату.
* Load (загрузка): Загрузка преобразованных данных в хранилище данных.

**Витрина данных** (Data Mart) – подмножество хранилища данных, ориентированное на конкретное подразделение или бизнес-функцию организации.

Обеспечивает более быстрый доступ к релевантным данным для конкретных пользователей.

**Таблица фактов** (Fact Table) содержит ключевые показатели (меры) и внешние ключи к таблицам измерений. Обычно содержит большое количество строк.

* Ключевые столбцы – внешние ключи, связывающие таблицу фактов с таблицами измерений.
* Неключевые столбцы (меры) – числовые значения, представляющие бизнес-показатели (например, продажи, прибыль, количество).
* Задача таблицы фактов – дать сумму, максимальное, минимальное и пр. значения.

**Таблица измерений** (Dimension Table) – содержит атрибуты, описывающие бизнес-объекты (например, продукт, клиент, дата).

* Содержит относительно небольшое количество строк.
* Ключевой столбец – первичный ключ.
* Имена столбцов используются в качестве заголовков в отчетах.

Разреженность измерений – наличие или отсутствие данных для определенного измерения (например, отсутствие продаж по определенной дате).

**Меры** (Measures) – числовые столбцы в таблице фактов, представляющие бизнес-показатели (например, продажи, количество заказов).

**Куб** – это многомерное представление данных, которое позволяет анализировать информацию с разных сторон и под разными углами. Он предназначен для быстрого анализа больших объемов информации.

**Измерения** (Dimensions): Это категории, по которым можно группировать и фильтровать данные (например, время, продукты, регионы, клиенты). В Excel-аналогии это столбцы.

**Факты** (Facts): Это числовые данные, которые мы анализируем (например, продажи, количество заказов, прибыль). Это значения внутри ячеек таблицы.

**Меры** (Measures): Это то, что мы хотим измерить, обычно это агрегированные факты (например, сумма продаж, средняя цена, количество клиентов).

Иерархии (Hierarchies): Отношения между уровнями в измерении, например: год > квартал > месяц > день.

**Проектирование хранилища данных для «Университет»**

Бизнес-задачи хранилища данных:

1. Анализ успеваемости студентов:

* Определение динамики успеваемости студентов с течением времени.
* Выявление факторов, влияющих на успеваемость (выбор курса, преподаватель, посещаемость).
* Прогнозирование риска отчисления студентов на основе успеваемости.

1. Анализ использования ресурсов:

* Оптимизация использования учебных аудиторий и лабораторий.
* Определение курсов, требующих наибольшего количества ресурсов (материалов, оборудования).

1. Финансовый анализ:

* Анализ структуры источников финансирования университета (гранты, государственное финансирование, плата за обучение).
* Распределение финансовых ресурсов между факультетами и кафедрами.

1. Анализ вовлеченности студентов:

* Определение уровня вовлеченности студентов во внеучебные мероприятия.
* Выявление связей между вовлеченностью студентов и успеваемостью.

**Таблица фактов, измерения и меры.**

1. **Fact\_Grades** (Фактические данные по оценкам)

* ID – Уникальный идентификатор записи.
* StudentID – Внешний ключ на Dim\_Student.
* CourseID – Внешний ключ на Dim\_Course.
* SemesterID – Семестр.
* Grade – Оценка студента за курс.

2. **Dim\_Student** (Измерение студентов)

* StudentID – Уникальный идентификатор студента.
* Name – ФИО студента.
* Faculty – Факультет.
* YearOfEnrollment – Год поступления.

3. **Dim\_Course** (Измерение курсов)

* CourseID – Уникальный идентификатор курса.
* CourseName – Название курса.
* Department – Кафедра или департамент.
* Credits – Количество кредитов за курс.

|  |
| --- |
| use DataWarehouse;  CREATE TABLE Dim\_Student (  StudentID INT PRIMARY KEY,  Name NVARCHAR(100) NOT NULL,  Faculty NVARCHAR(50) NOT NULL,  YearOfEnrollment INT NOT NULL  );  CREATE TABLE Dim\_Course (  CourseID INT PRIMARY KEY,  CourseName NVARCHAR(100) NOT NULL,  Department NVARCHAR(50) NOT NULL,  Credits INT NOT NULL  );  CREATE TABLE Fact\_Grades (  ID INT IDENTITY(1,1) PRIMARY KEY,  StudentID INT NOT NULL,  CourseID INT NOT NULL,  SemesterID INT NOT NULL,  Grade DECIMAL(5,2) NOT NULL,  FOREIGN KEY (StudentID) REFERENCES Dim\_Student(StudentID),  FOREIGN KEY (CourseID) REFERENCES Dim\_Course(CourseID)  );  INSERT INTO Dim\_Student (StudentID, Name, Faculty, YearOfEnrollment) VALUES  (101, 'Иван Иванов', 'ФКТИ', 2020),  (102, 'Анна Смирнова', 'ФЭУ', 2019);  INSERT INTO Dim\_Course (CourseID, CourseName, Department, Credits) VALUES  (301, 'Алгебра', 'Математика', 3),  (302, 'Программирование', 'ФКТИ', 4);  INSERT INTO Fact\_Grades (StudentID, CourseID, SemesterID, Grade) VALUES  (101, 301, 202301, 90),  (102, 302, 202301, 85); |

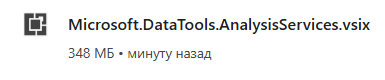
Описание построения куба

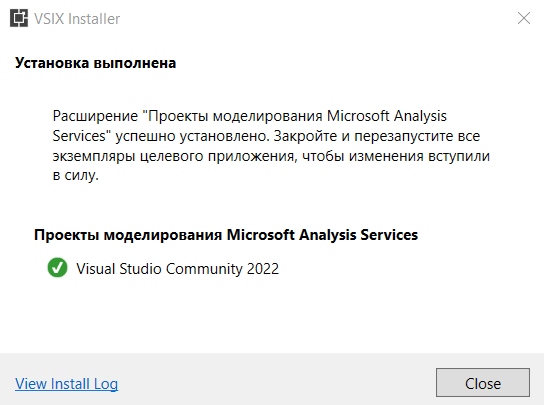
SQL Server Analysis Services (SSAS) — это служба аналитики и платформа для построения хранилищ данных и бизнес-аналитики от Microsoft.

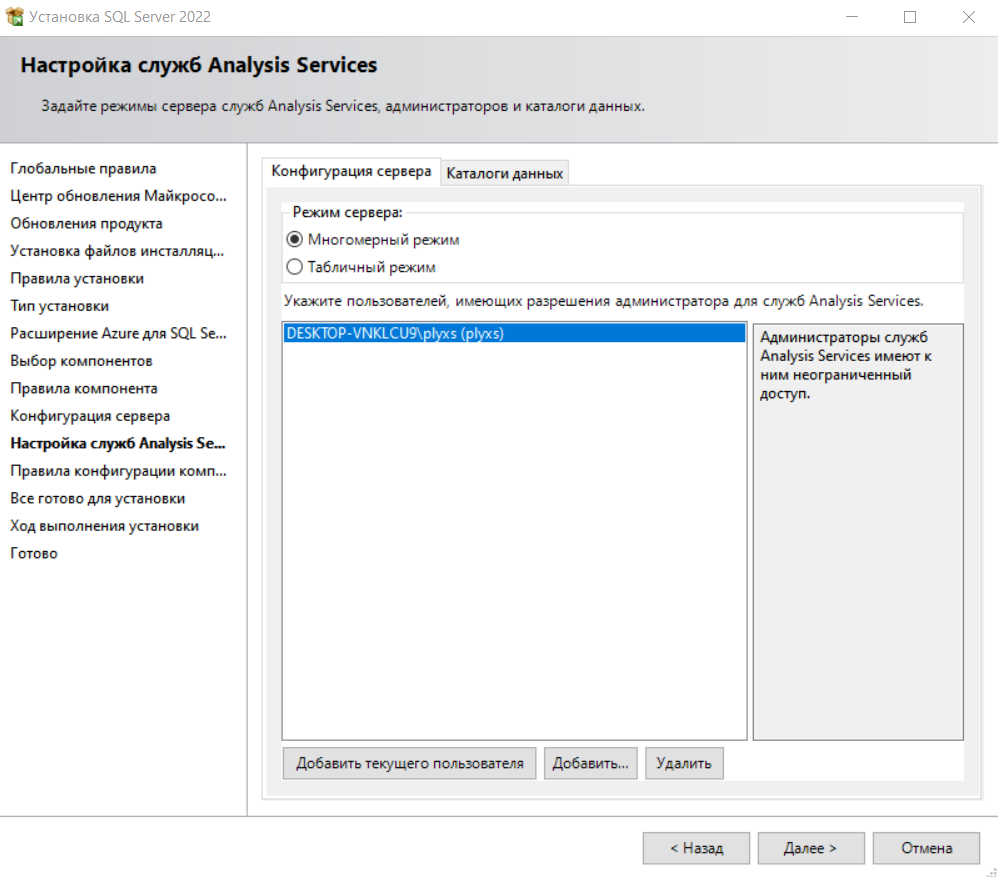
SSAS позволяет создавать многомерные OLAP-кубы (Online Analytical Processing). Куб — это структура данных, оптимизированная для быстрого анализа больших объёмов информации. Данные в кубе предварительно агрегируются (суммируются, усредняются и т.д.), что значительно ускоряет выполнение запросов.

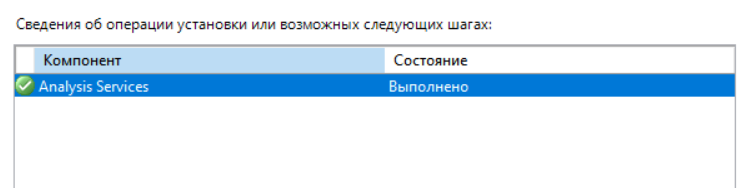
Для запросов к кубам используется язык MDX (Multidimensional Expressions). Он отличается от SQL и предназначен для работы с многомерными данными.

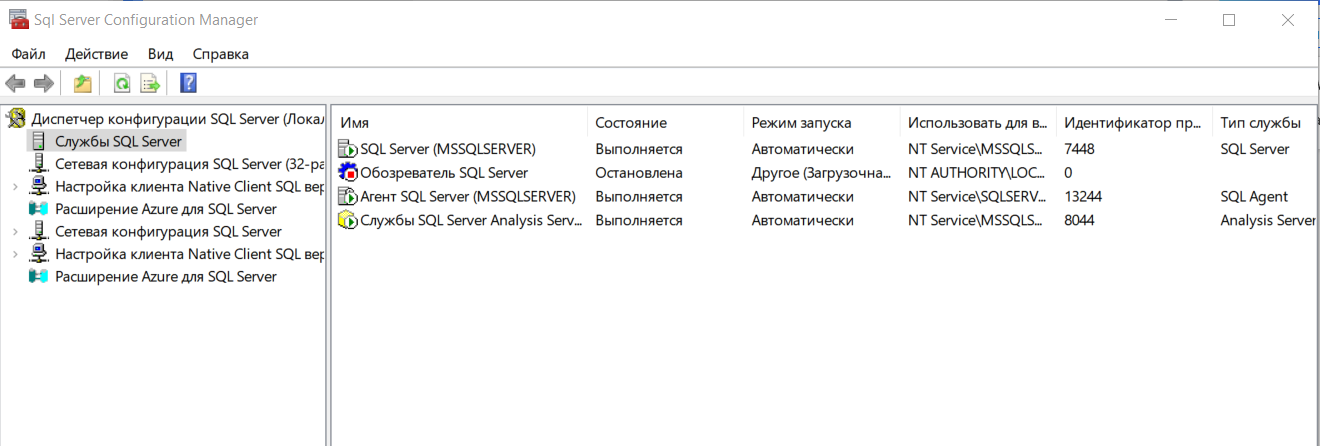


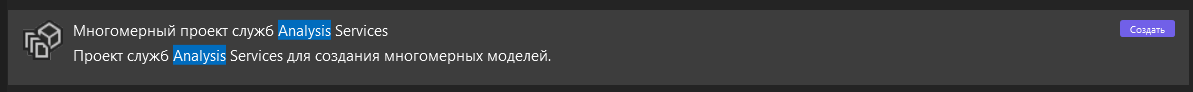


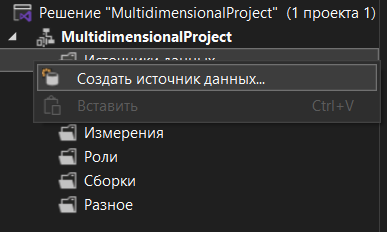


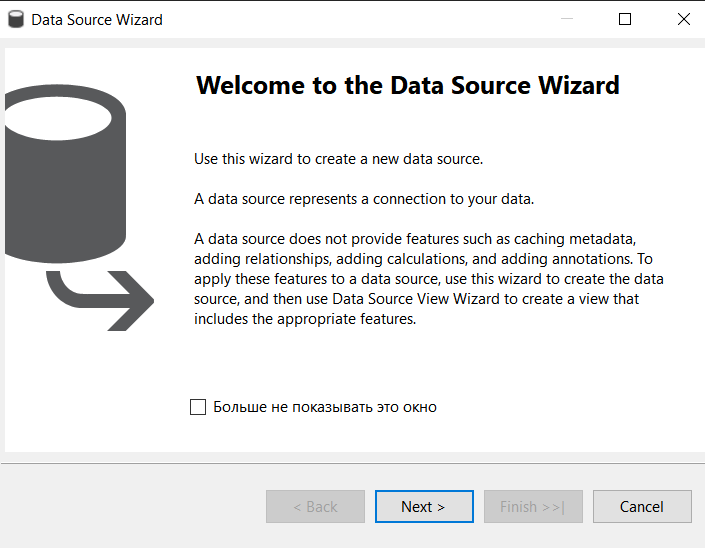


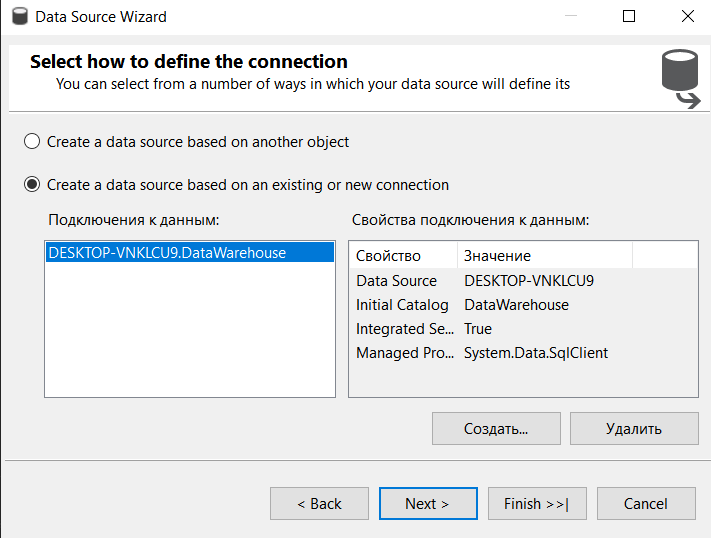


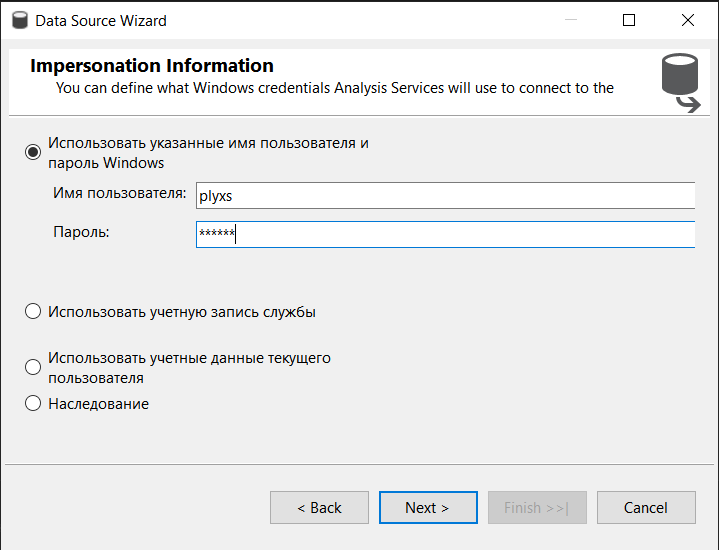


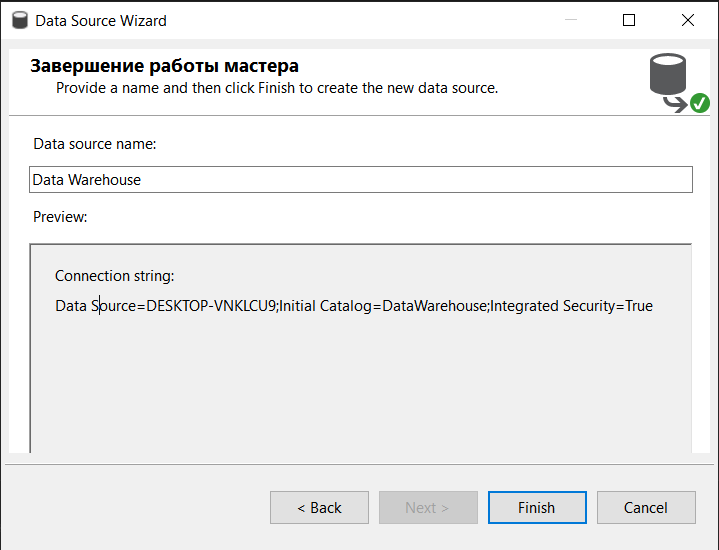


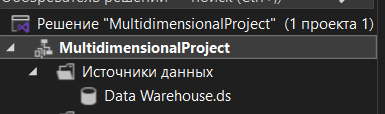




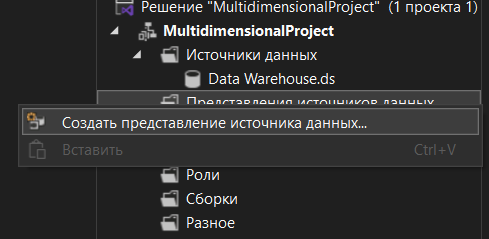


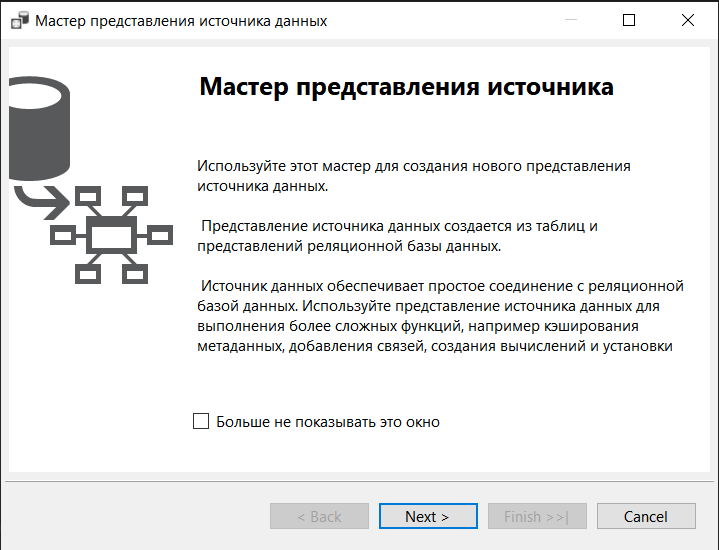


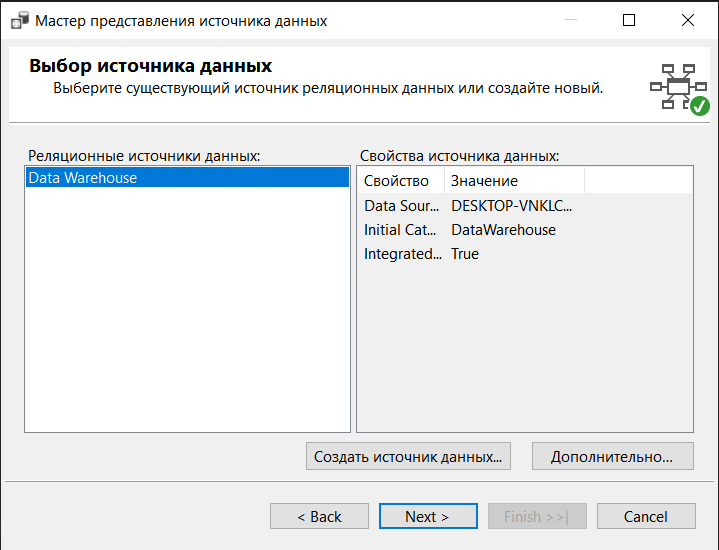


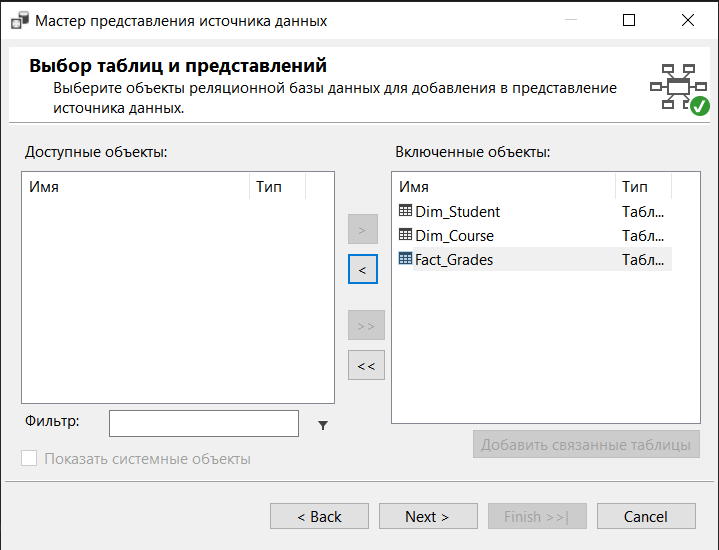


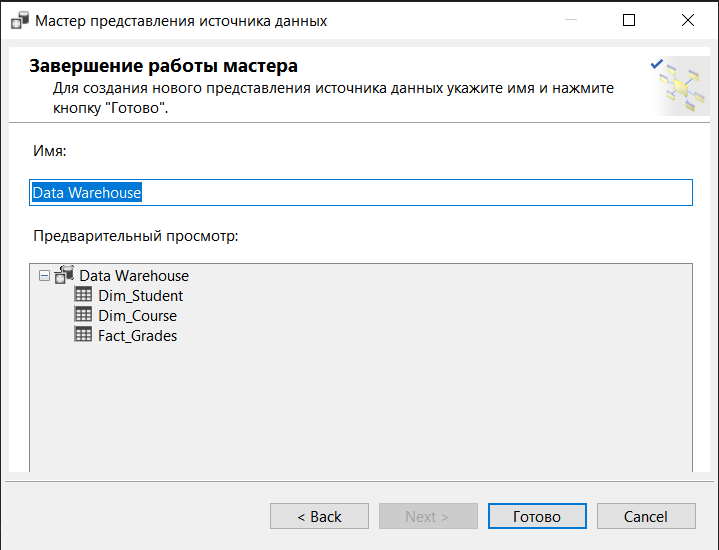
Представление источника данных (Data Source View, DSV) – это абстрактное представление данных из одного или нескольких источников данных. Это логический слой между источником данных и кубом.

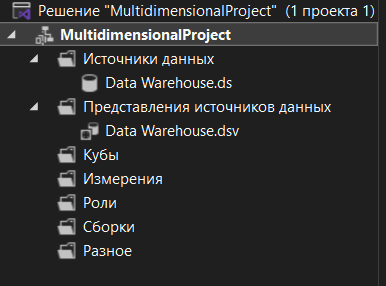


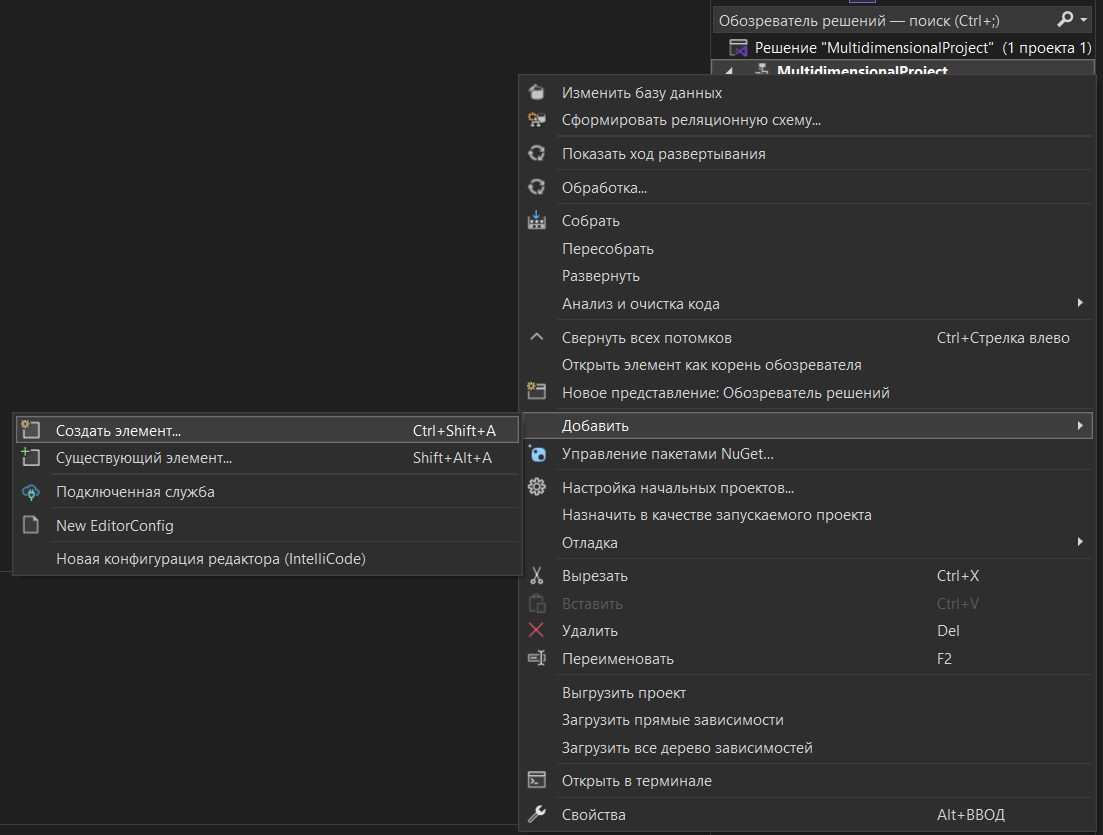


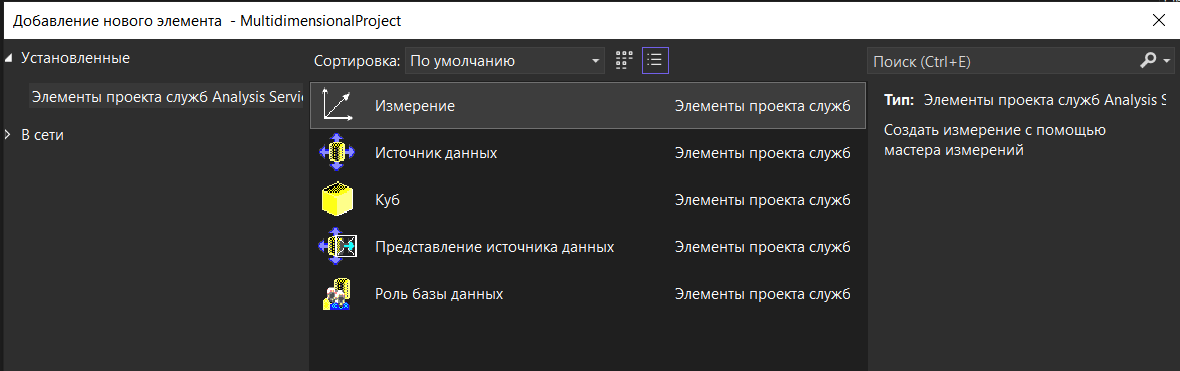


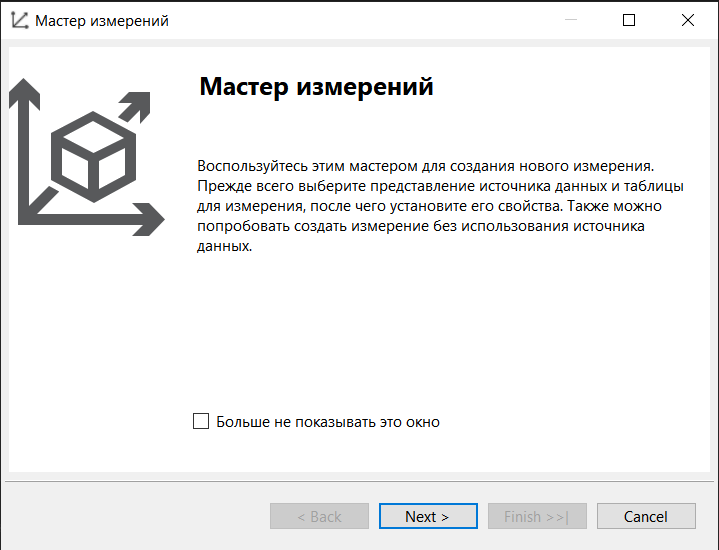


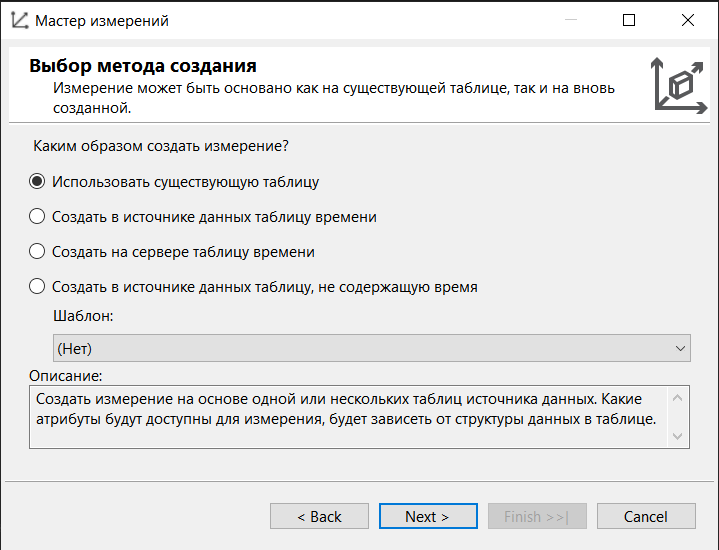


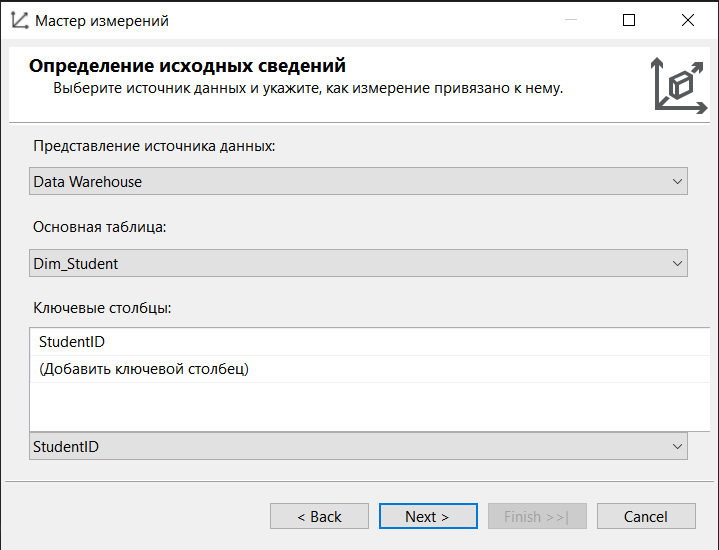


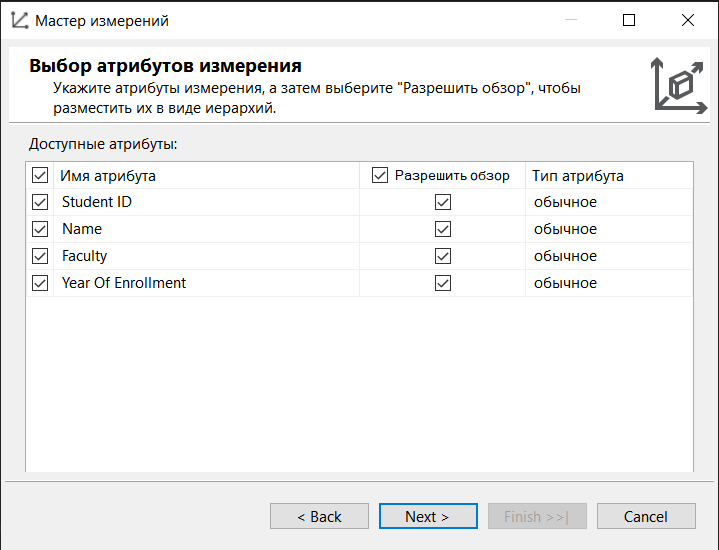


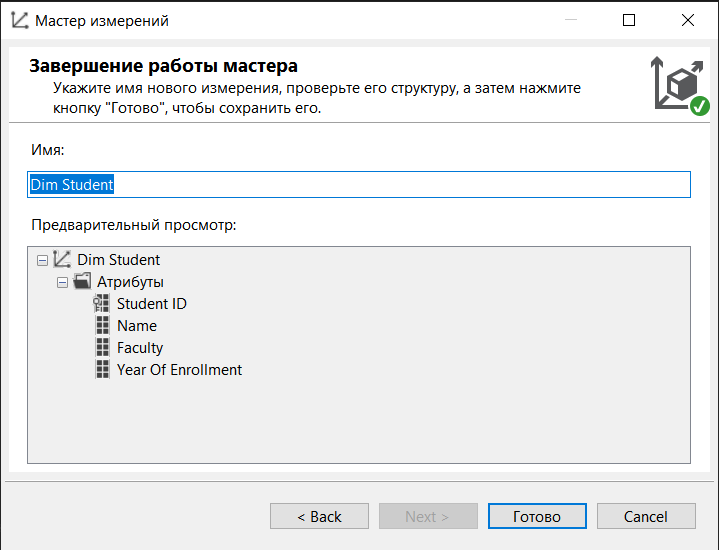


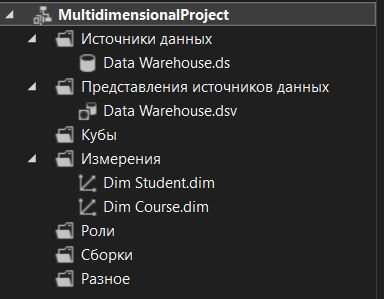


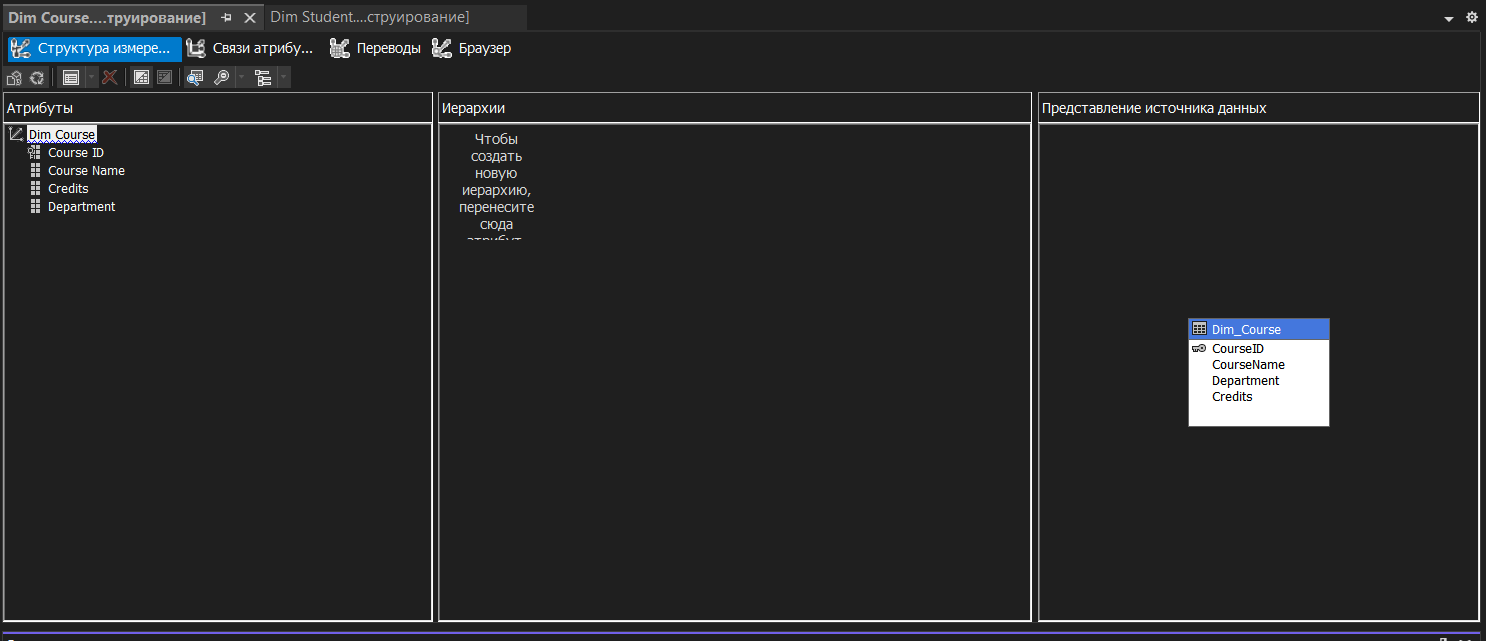


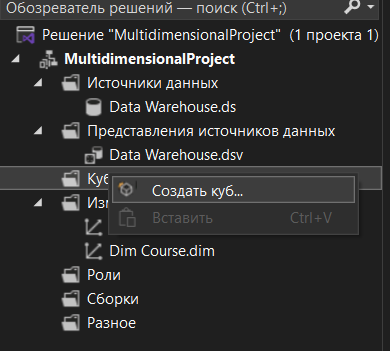


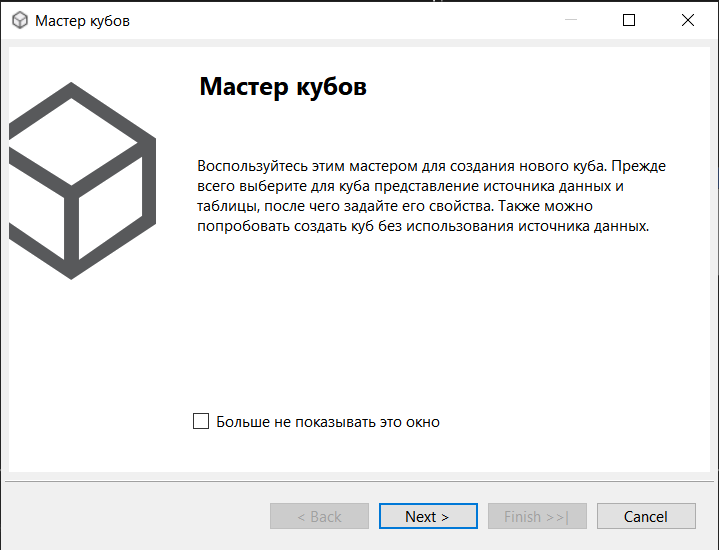


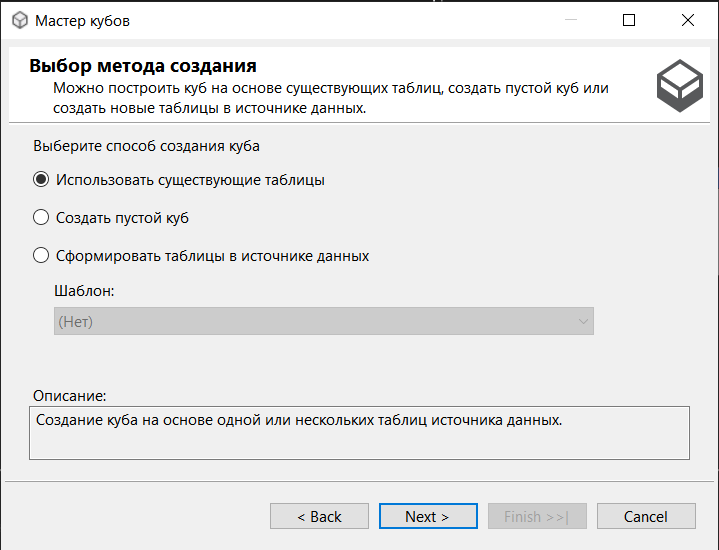


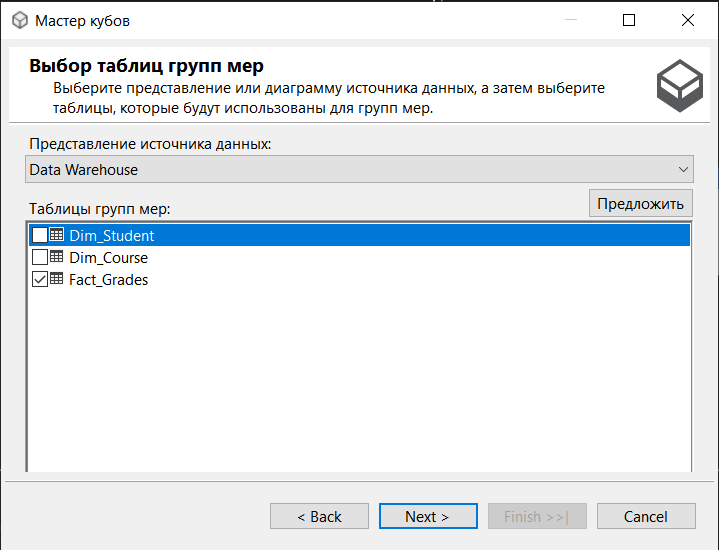


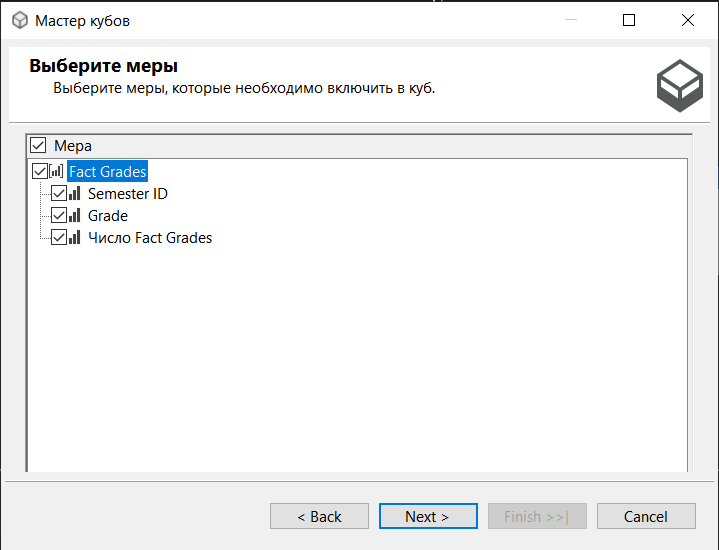


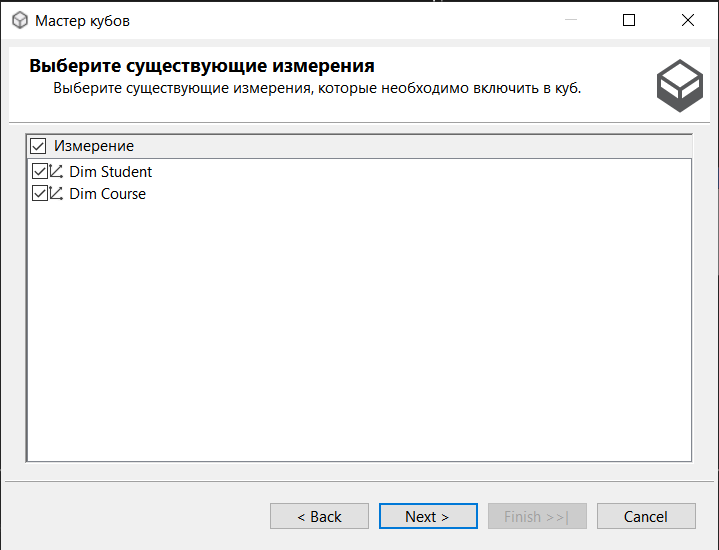


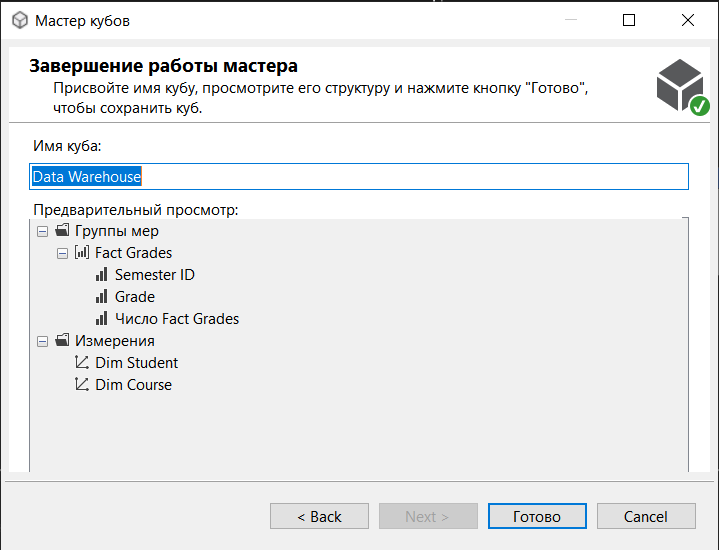






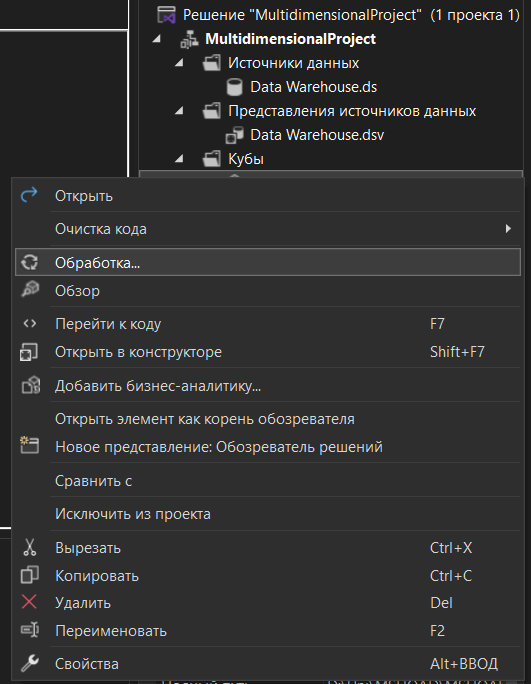


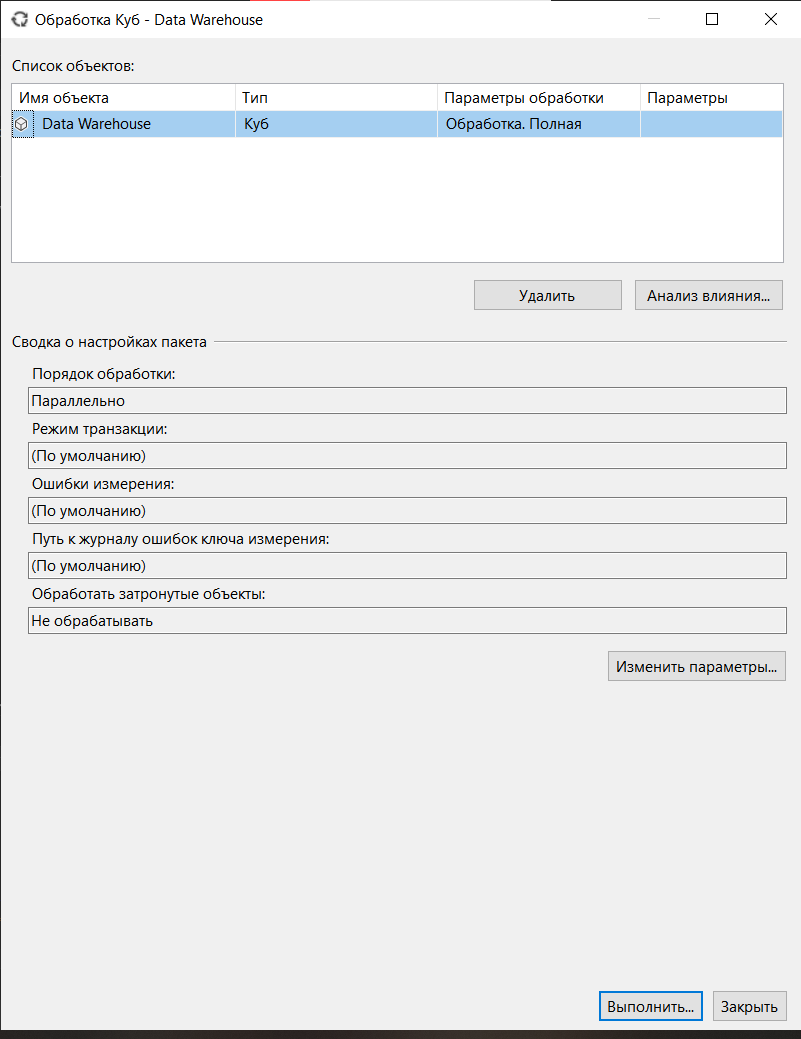


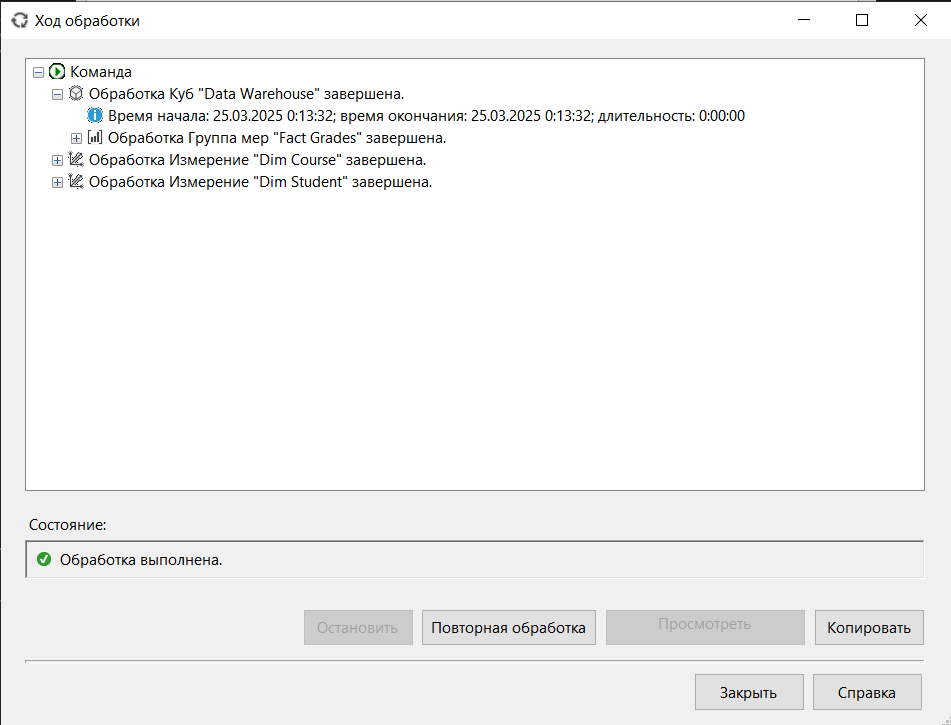


После создания структуры куба (определения измерений, мер, иерархий) в SQL Server Analysis Services (SSAS), его необходимо обработать (Process).

Обработка куба (Processing) – это процесс загрузки данных из источника данных (через представление источника данных) в многомерную структуру куба и выполнения необходимых вычислений.



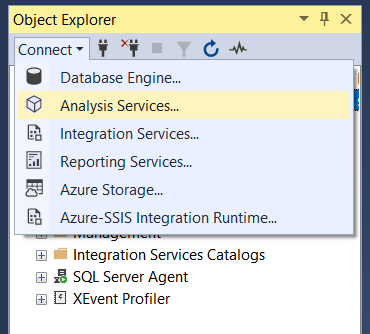


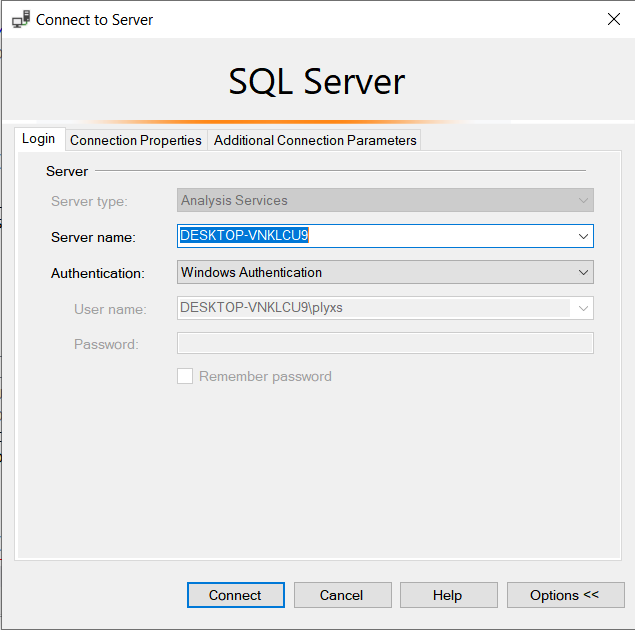


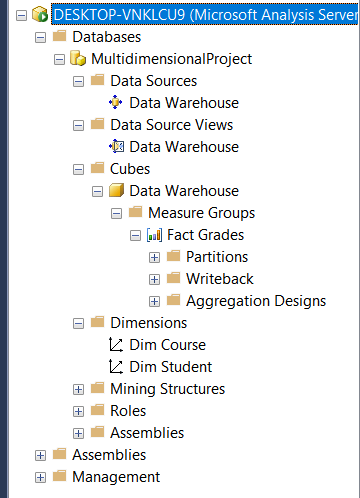
Мы работаем с проектом Analysis Services. Этот проект содержит определения куба (источники данных, DSV, измерения, меры, иерархии и т.д.).

Чтобы куб стал доступен для использования, его нужно развернуть (deploy) на экземпляре (instance) Analysis Services.

При развертывании мы указываем сервер SSAS, на котором нужно создать куб, и базу данных Analysis Services, в которой он будет размещен.







Выполнить 5 запросов к данным в кубе, используя следующие функции языка MDX:

1. Числовые

|  |
| --- |
| WITH  MEMBER [Measures].[Max Grade] AS  Max(  [Dim Student].[Student ID].MEMBERS,  [Measures].[Grade]  )  SELECT  [Measures].[Max Grade] ON COLUMNS,  [Dim Course].[Course Name].MEMBERS ON ROWS  FROM [Data Warehouse] |

1. Обработки строковых значений

|  |
| --- |
| WITH  MEMBER [Measures].[Фамилия] AS  Mid(  [Dim Student].[Name].CurrentMember.Name,  InStr([Dim Student].[Name].CurrentMember.Name, " ") + 1,  100  ),  FORMAT\_STRING = "String"  SELECT  [Measures].[Фамилия] ON COLUMNS,  [Dim Student].[Name].MEMBERS ON ROWS  FROM [Data Warehouse] |

1. Для наборов

|  |
| --- |
| WITH  SET [Top Students] AS  TopCount(  [Dim Student].[Name].MEMBERS,  3,  Avg(  [Dim Course].[Course Name].MEMBERS,  [Measures].[Grade]  )  )  SELECT  [Measures].[Grade] ON COLUMNS,  [Top Students] ON ROWS  FROM [Data Warehouse]; |

1. Для элемента измерения

|  |
| --- |
| WITH  MEMBER [Measures].[Course Name] AS  [Dim Course].[Course Name].CURRENTMEMBER.NAME  SELECT  [Measures].[Course Name] ON COLUMNS,  [Dim Course].[Course ID].MEMBERS ON ROWS  FROM [Data Warehouse]; |

1. Для иерархий

|  |
| --- |
| WITH  MEMBER [Measures].[HierarchyLevelCount] AS  [Dim Student].[Faculty].LEVELS.Count  SELECT  { [Measures].[HierarchyLevelCount] } ON COLUMNS  FROM [Data Warehouse]; |