

Учебный курс

Хранилища данных

Лекция 10

**Понятия о MDS. Аналитические
службы MS SQL Server**

Лекции читает

Кандидат технических наук, доцент

Перминов Геннадий Иванович

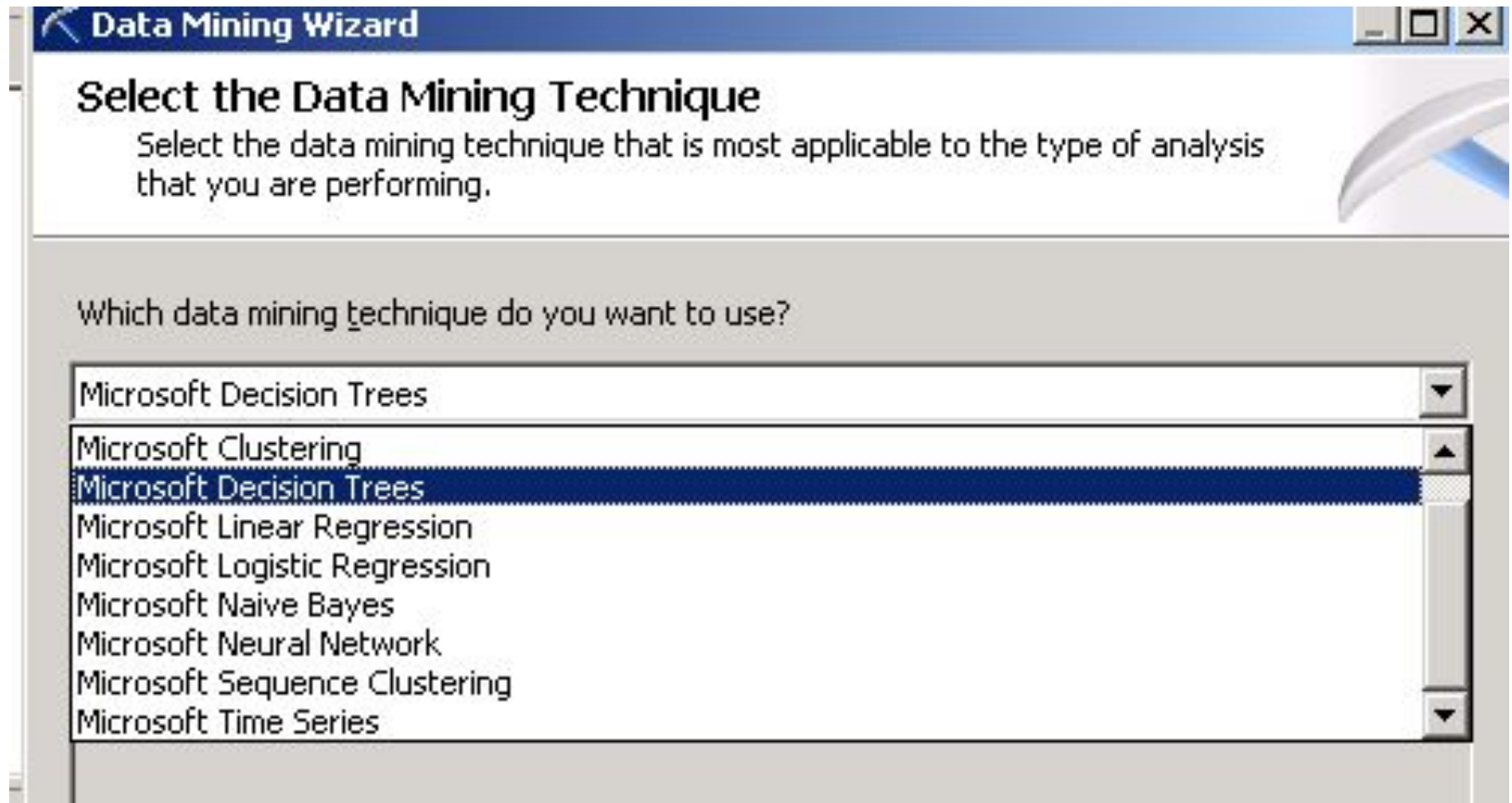
Содержание

- 1. Цель использования аналитических служб
- 2. Модели добычи данных 2. Модели добычи данных (DataMining)
- 3. Алгоритмы добычи данных
 - 3.1. Метод дерева решений
 - 3.2. Кластеризация
- 4. Построение модели добычи данных
 - 4.1. Построение модели Дерево решений (на примере технологии 4.1. Построение модели Дерево решений (на примере технологии MS SQL Server 4.1. Построение модели Дерево решений (на примере технологии MS SQL Server 2000))

1. Цель использования аналитических служб

- Информация, которую вы ищете, уже находится в вашей базе данных. Но она спрятана достаточно глубоко, поэтому найти ее, просто просмотрев данные, будет довольно сложно.
- Добыча данных (data mining) извлекает намного больше информации, содержащейся в ваших данных, позволяя вам обнаруживать скрытые взаимосвязи в данных, находить тенденции, наблюдать за причинами конкретных событий либо даже предсказывать производительность или направление для отдельных аспектов данных.

Аналитические модели в SQL Server 2005



2. Модели добычи данных (DataMining)

- Модель добычи данных представляет собой виртуальную структуру, хранящую данные, используемые при выполнении добычи данных на SQL Server. Информация в модели хранится в том же виде, что и в базе данных, но вместо реальных данных в ней находятся правила и шаблоны для данных, хранимых в модели. Эти правила и шаблоны являются интерпретациями многомерных данных в виде статистической информации, которая в дальнейшем используется для предсказания будущего поведения и изменения определенных аспектов данных.

3. Алгоритмы добычи данных

- **Microsoft Association Rules.** Правила ассоциаций ищут элементы, которые наиболее вероятно появляются вместе в транзакциях и могут быть использованы для прогнозирования присутствия элемента на основании существования других транзакций.
- **Microsoft Clustering.** Кластеризация ищет естественные группировки данных. Это особенно полезно, если вы хотите видеть условия, которые имеют тенденцию появляться вместе.
- **Microsoft Decision Tree.** Дерево решений позволяет создавать прогнозы и виртуальные измерения на основе результатов анализа.
- **Microsoft Linear Regression.** Использование статистики линейной регрессии, позволяющей на основании существующих данных прогнозировать будущее.
- **Microsoft Logistic Regression.** Использование статистики логистической регрессии, позволяющей на основании существующих данных прогнозировать будущее.
- **Microsoft Naive Bayes.** Naive Bayes представляет собой алгоритм классификации, который хорошо себя зарекомендовал в моделях прогнозирования если только атрибуты дискретны или дискретизированы. Предполагается что все входные атрибуты не зависят от прогнозируемого.
- **Microsoft Neural Network.** Нейронные сети лучше всего использовать для классификации дискретных атрибутов, а также регрессии непрерывных атрибутов, если вы заинтересованы в прогнозировании множества атрибутов.
- **Microsoft Sequence Clustering.** Кластеризация последовательностей позволяет прогнозировать наиболее вероятный порядок событий в последовательности, основанной на известных характеристиках.
- **Microsoft Time Series.** Временные ряды позволяют прогнозировать будущие события, зависящие от времени на основе известных или обнаруженных шаблонов.

3.1. Метод дерева решений

- Деревья решений применяются уже довольно долгое время для поиска набора определенных характеристик и правил, а также определения влияния этих правил на искомую переменную. Например, когда вам понадобится определить характерные признаки клиента, который с наибольшей вероятности ответит на рассылку образцов товара, вам придется перевести эти характеристики в набор правил.

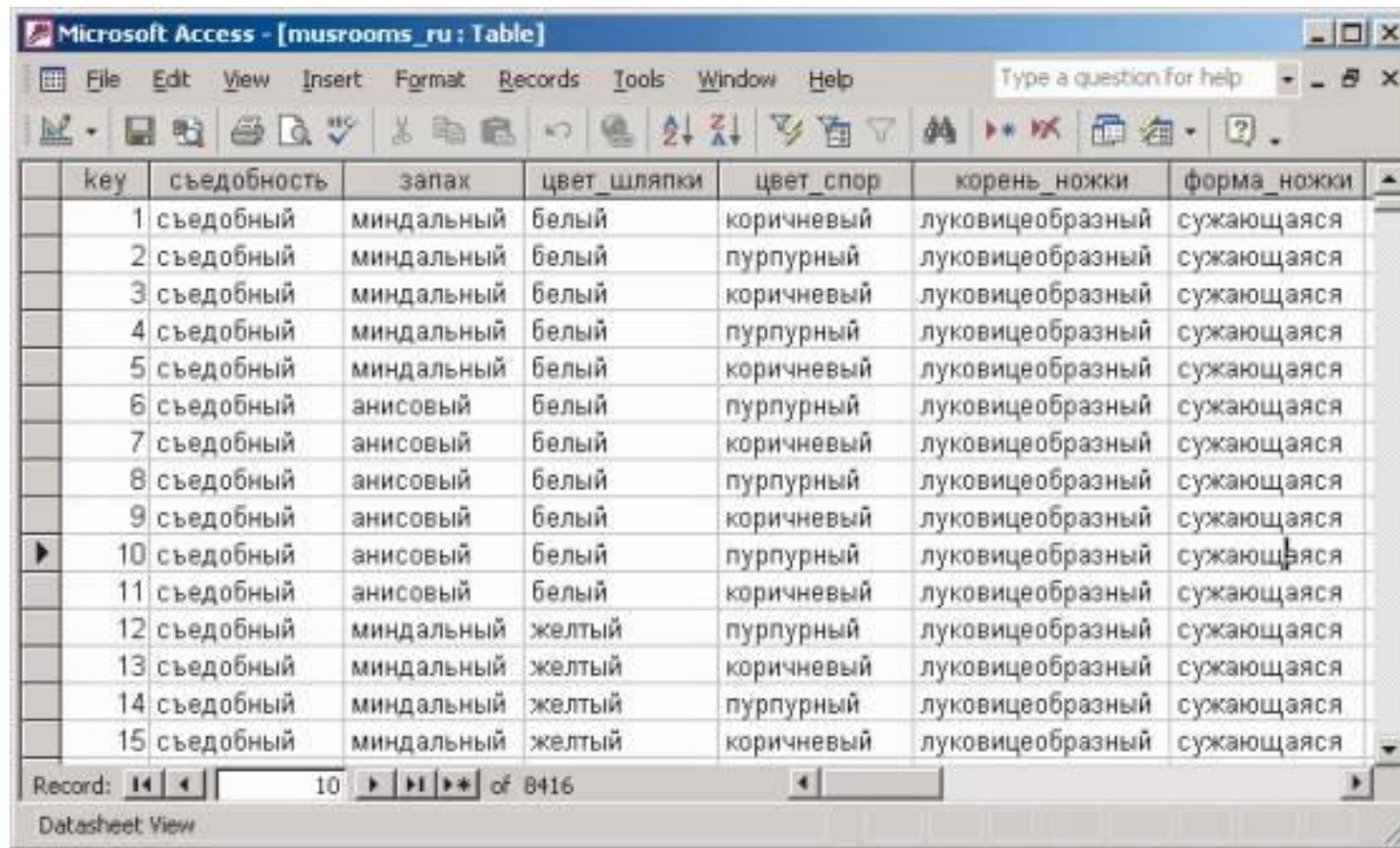
Задача о грибах. Data Mining, будет состоять из двух процессов:

- обучение модели (которое выполняется однократно и требует относительно много времени)
- и принятие решения о том, относится ли конкретный гриб к категории съедобных (что происходит неоднократно).

Исходные данные

- В качестве исходных данных для обучения модели мы воспользуемся набором данных в 8416 грибов, доступных в виде файла в формате CSV по адресу <http://www.ics.uci.edu/~mlearn/MLRepository.html>, который содержит таблицу, где имеется колонка Edibility с двумя возможными значениями (edible - съедобный и poisonous - ядовитый). Файл содержит таблицу, состоящую из 24 столбцов: 22 признака, таких, как форма и цвет шляпки, ножки и т. д., столбец Edibility (съедобность) с двумя возможными значениями (Edible — съедобный и poisonous — ядовитый) и ключевое поле Number, содержащее уникальные ключи наблюдений от 1 до 8416).

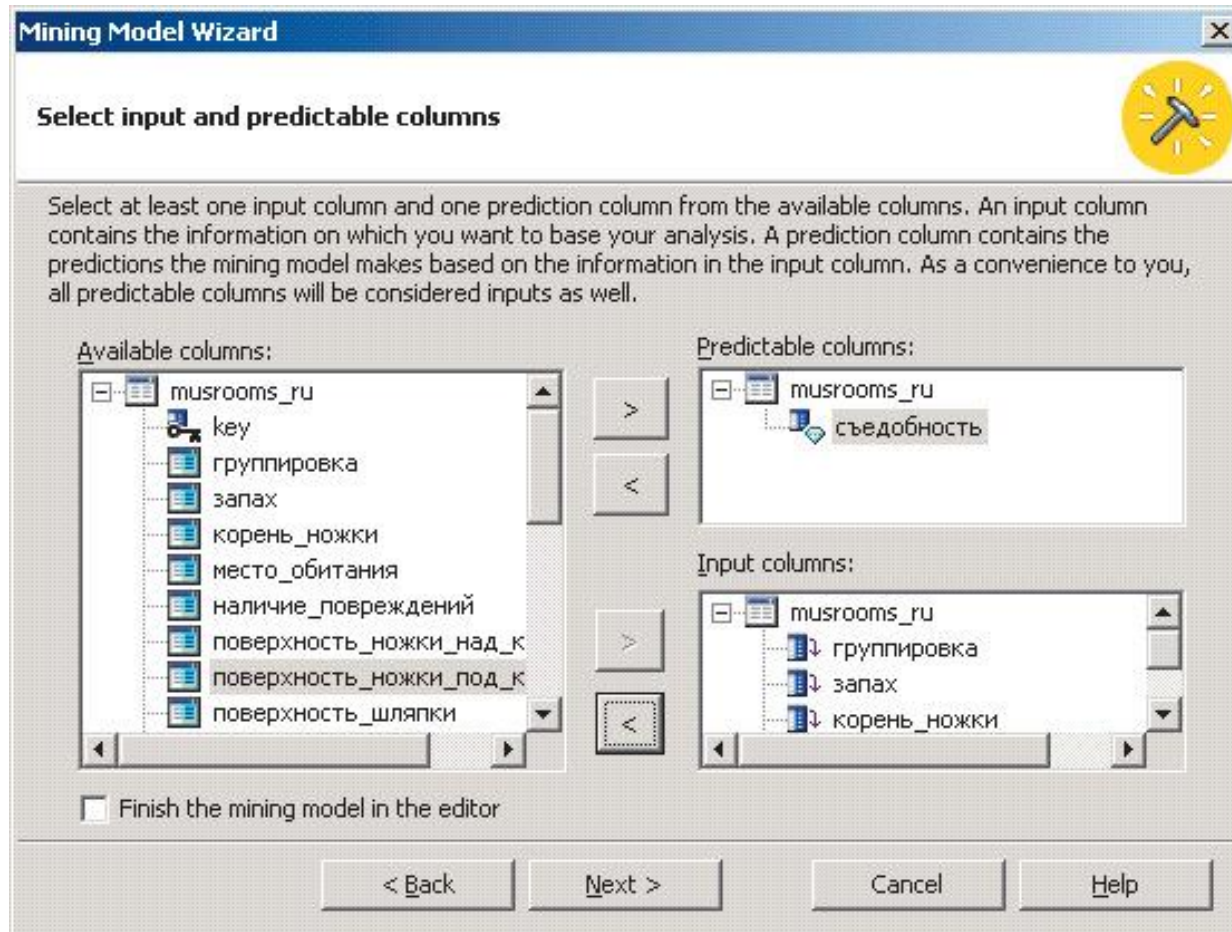
Исходные данные к определению признаков съедобности грибов



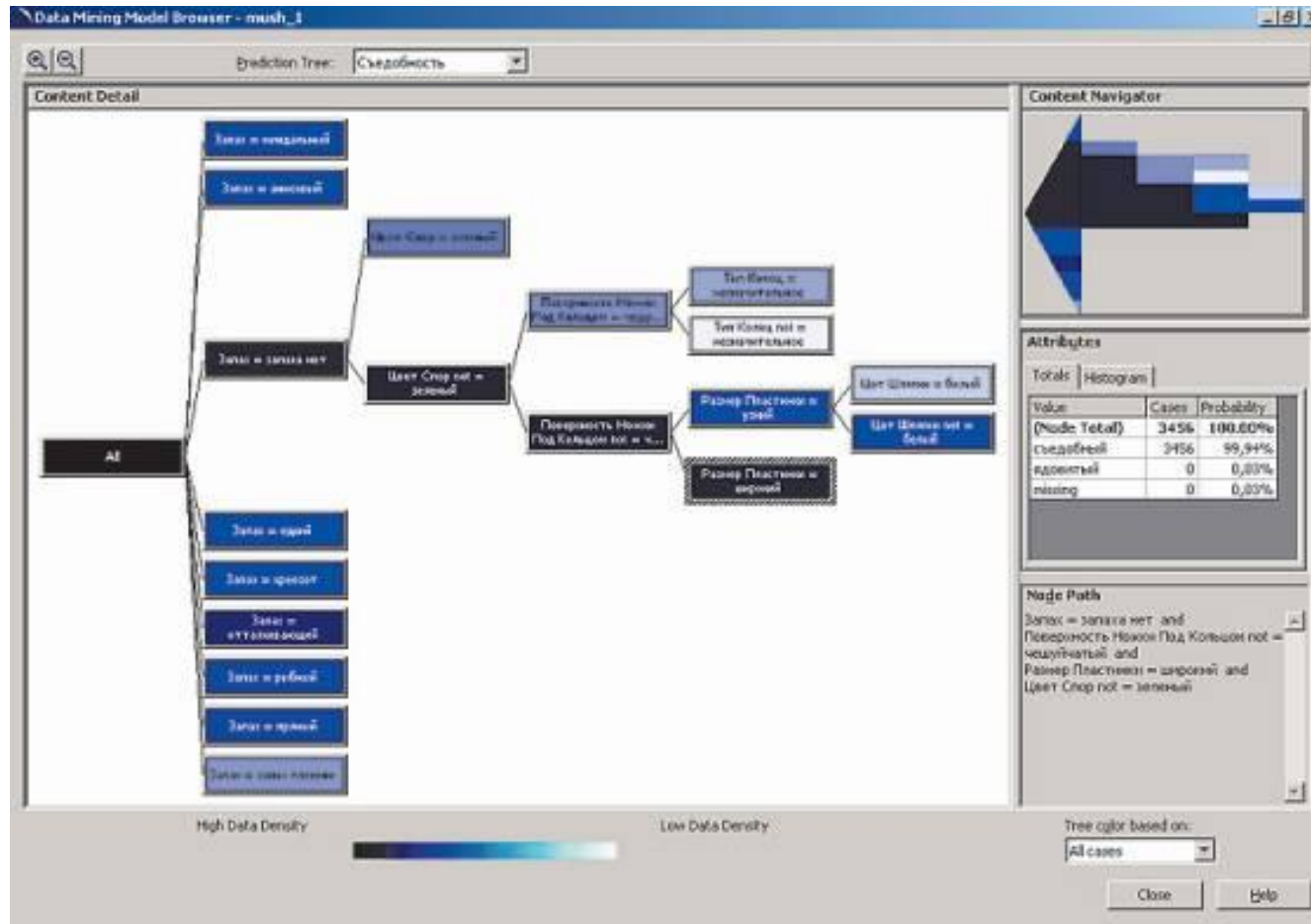
The screenshot shows a Microsoft Access window titled "Microsoft Access - [musrooms_ru : Table]". The window displays a table with 8 columns and 15 rows of data. The columns are: key, съедобность, запах, цвет_шляпки, цвет_спор, корень_ножки, and форма_ножки. The data rows show various mushroom characteristics, such as "съедобный", "миндальный", "белый", "коричневый", "пурпурный", "луковицеобразный", and "сужающаяся". The status bar at the bottom indicates "Record: 10 of 8416" and "Datasheet View".

key	съедобность	запах	цвет_шляпки	цвет_спор	корень_ножки	форма_ножки
1	съедобный	миндальный	белый	коричневый	луковицеобразный	сужающаяся
2	съедобный	миндальный	белый	пурпурный	луковицеобразный	сужающаяся
3	съедобный	миндальный	белый	коричневый	луковицеобразный	сужающаяся
4	съедобный	миндальный	белый	пурпурный	луковицеобразный	сужающаяся
5	съедобный	миндальный	белый	коричневый	луковицеобразный	сужающаяся
6	съедобный	анисовый	белый	пурпурный	луковицеобразный	сужающаяся
7	съедобный	анисовый	белый	коричневый	луковицеобразный	сужающаяся
8	съедобный	анисовый	белый	пурпурный	луковицеобразный	сужающаяся
9	съедобный	анисовый	белый	коричневый	луковицеобразный	сужающаяся
10	съедобный	анисовый	белый	пурпурный	луковицеобразный	сужающаяся
11	съедобный	анисовый	белый	коричневый	луковицеобразный	сужающаяся
12	съедобный	миндальный	желтый	пурпурный	луковицеобразный	сужающаяся
13	съедобный	миндальный	желтый	коричневый	луковицеобразный	сужающаяся
14	съедобный	миндальный	желтый	пурпурный	луковицеобразный	сужающаяся
15	съедобный	миндальный	желтый	коричневый	луковицеобразный	сужающаяся

Выбор полей для исследования



Пример отчета дерева решений



Правила классификации данных

выглядят так:

- если запах гриба миндальный или анисовый (Odor = ALMOND или Odor = ANIS), то гриб съедобный (EDIBLE);
- если запах другой, то гриб ядовитый (POISONOUS);
- если запаха нет (Odor = NONE), то вопрос требует дальнейшего изучения.

Второй уровень иерархии

- Второй уровень иерархии доступен только для ветви, содержащей данные о грибах без запаха, поэтому очередным параметром в этом случае оказывается цвет спор (если споры зеленые (Spore Print Color = GREEN), то гриб ядовитый), если другой (Spore print color not = GREEN), то, опять же, необходим анализ еще какого-то параметра.

Третий уровень иерархии

- Третий уровень иерархии - поверхность ножки под кольцом (Stalk Surface Below Ring), далее для грибов с чешуйчатymi ножками (Stalk Surface Below Ring = SCALY) - анализируем тип колец (Ring Type), а для остальных - размер пластинки (Gill Size) и цвет шляпки (Cap color).

Область применения

- Таким образом, алгоритм построения деревьев решений позволяет определить набор значений характеристик, позволяющих отделить одну категорию данных от другой (в данной ситуации - съедобные грибы от несъедобных); этот процесс называют сегментацией.

3.2. Кластеризация

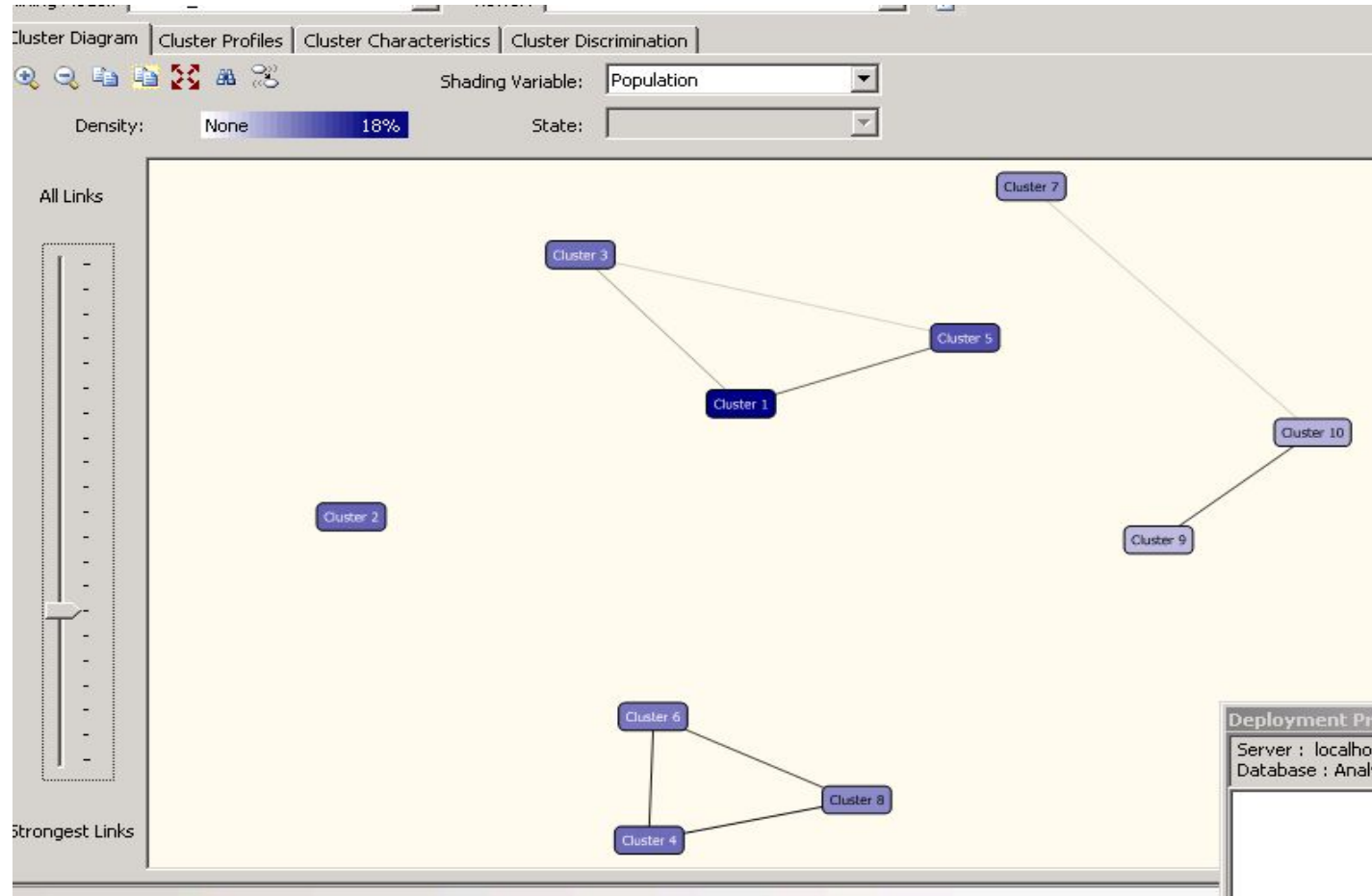
- При помощи кластеризации информация группируется по схожим признакам. В результате применения данной методики мы получаем сегменты данных, каждый из которых состоит из элементов, схожих по определенным признакам.

Рассмотрим многофакторную модель анализа индекса РТС

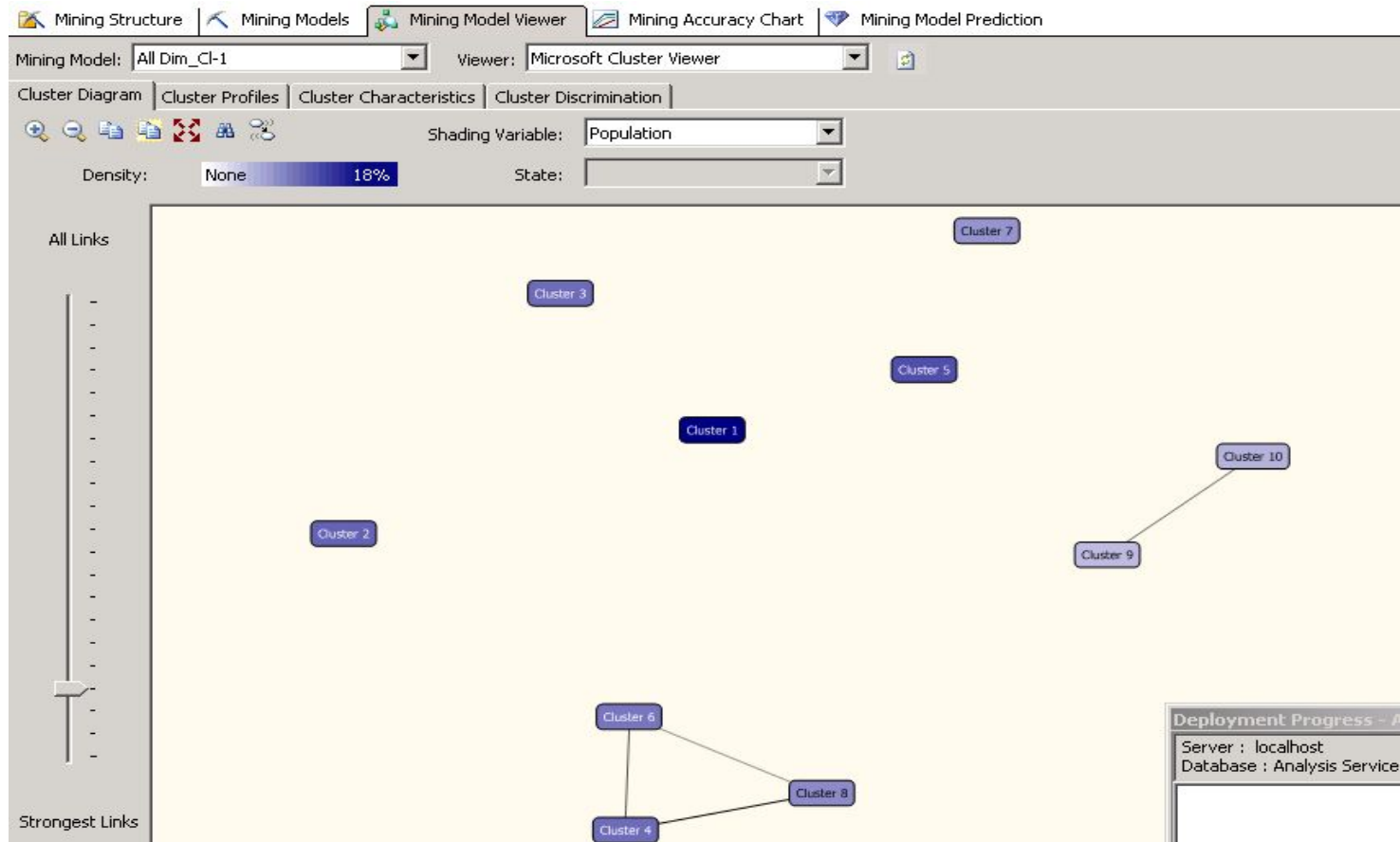
Задачи, решаемые в процессе кластеризации выглядят следующим образом:

1. построить кластерную модель индекса РТС;
 - 1.1. произвести подсоединение к многомерному хранилищу (кубу) как к источнику исходных данных;
 - 1.2. определить необходимые измерения;
 - 1.3. рассчитать кластерную модель для индекса РТС;
 - 1.4. построить визуальную графическую модель кластеров индекса РТС;
2. произвести анализ построенной кластерной модели индекса РТС;
 - 2.1. выяснить наличие и силу связи между кластерами;
 - 2.2. построить графическое представление содержимого кластеров;
 - 2.3. получить вероятности того, что значение входного атрибута попадет в кластер;
 - 2.4. выявить различия между кластерами с низким и высоким индексом РТС.

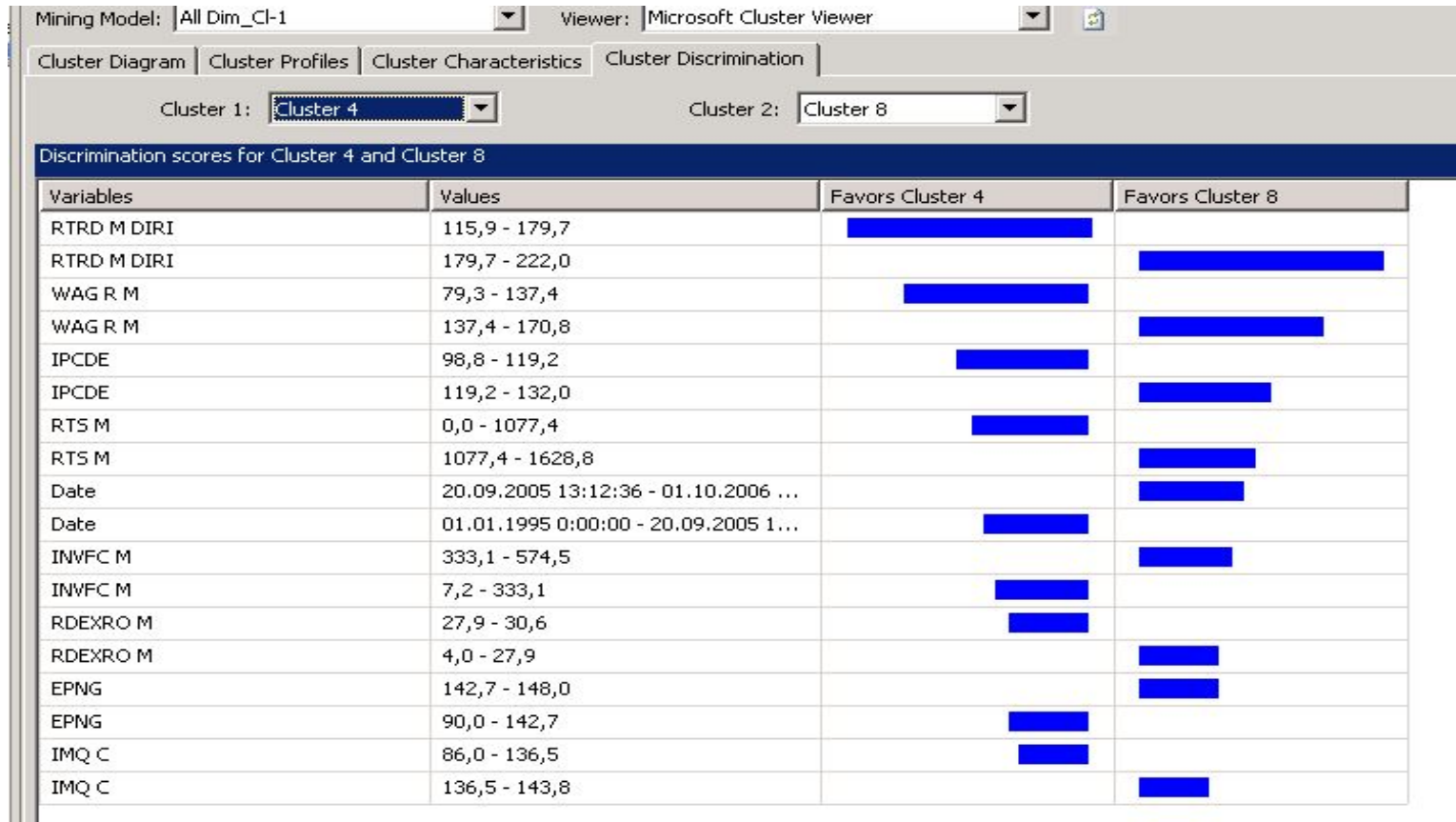
Пример кластеризации



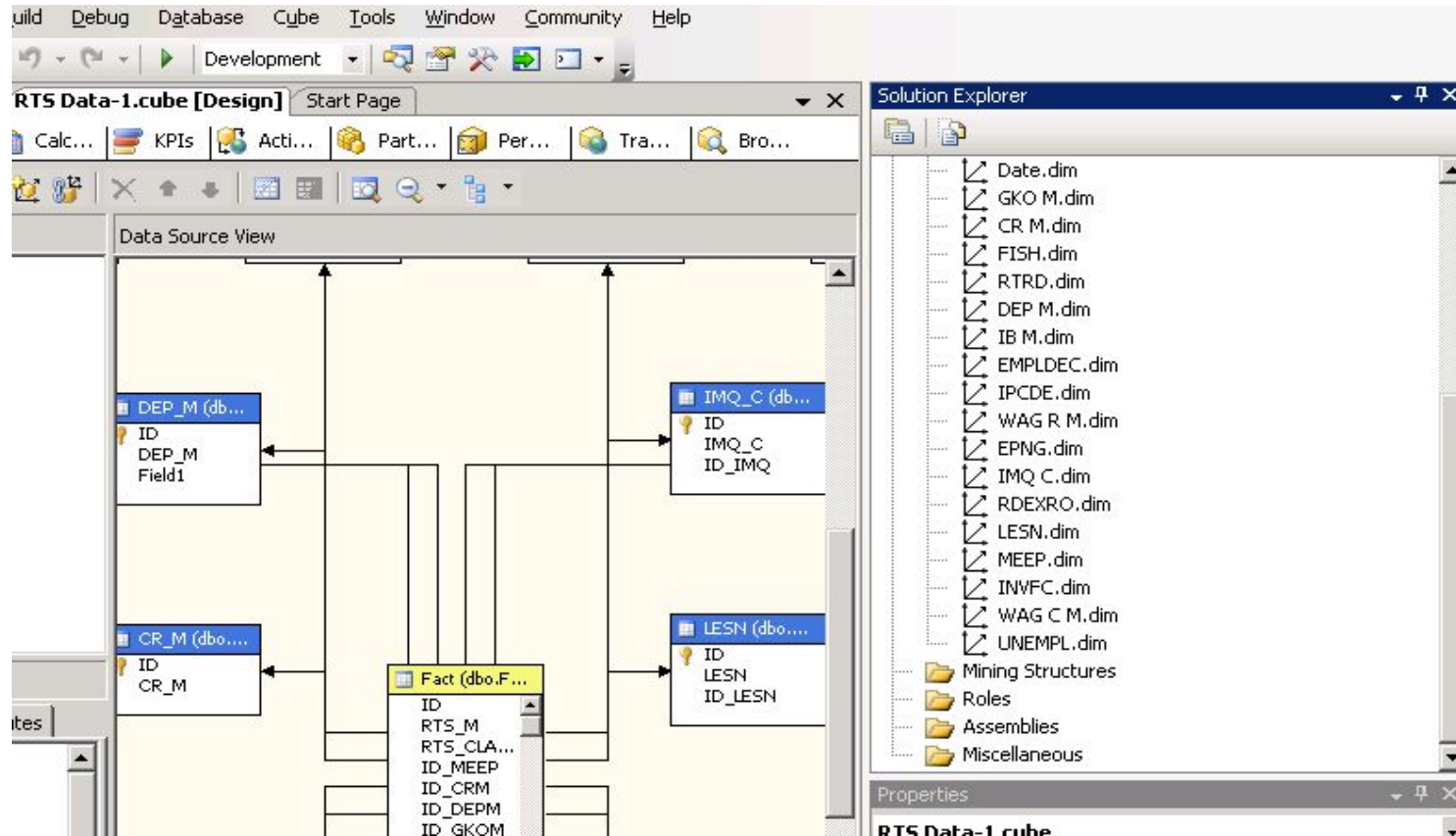
Перемещая движок All Links вверх или вниз, можно просмотреть наличие сильных и слабых связей между кластерами



Панель Cluster Discrimination



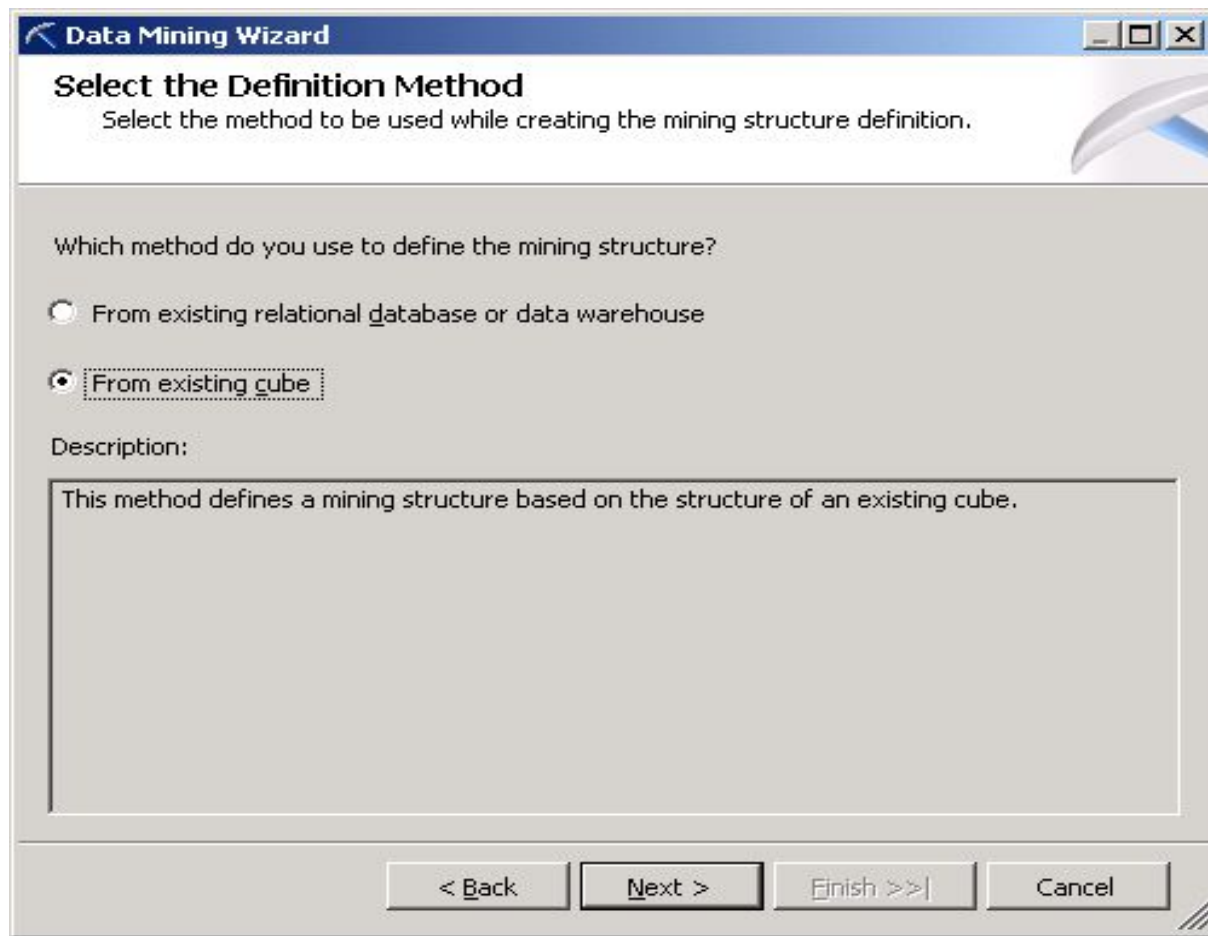
4. Построение модели добычи данных



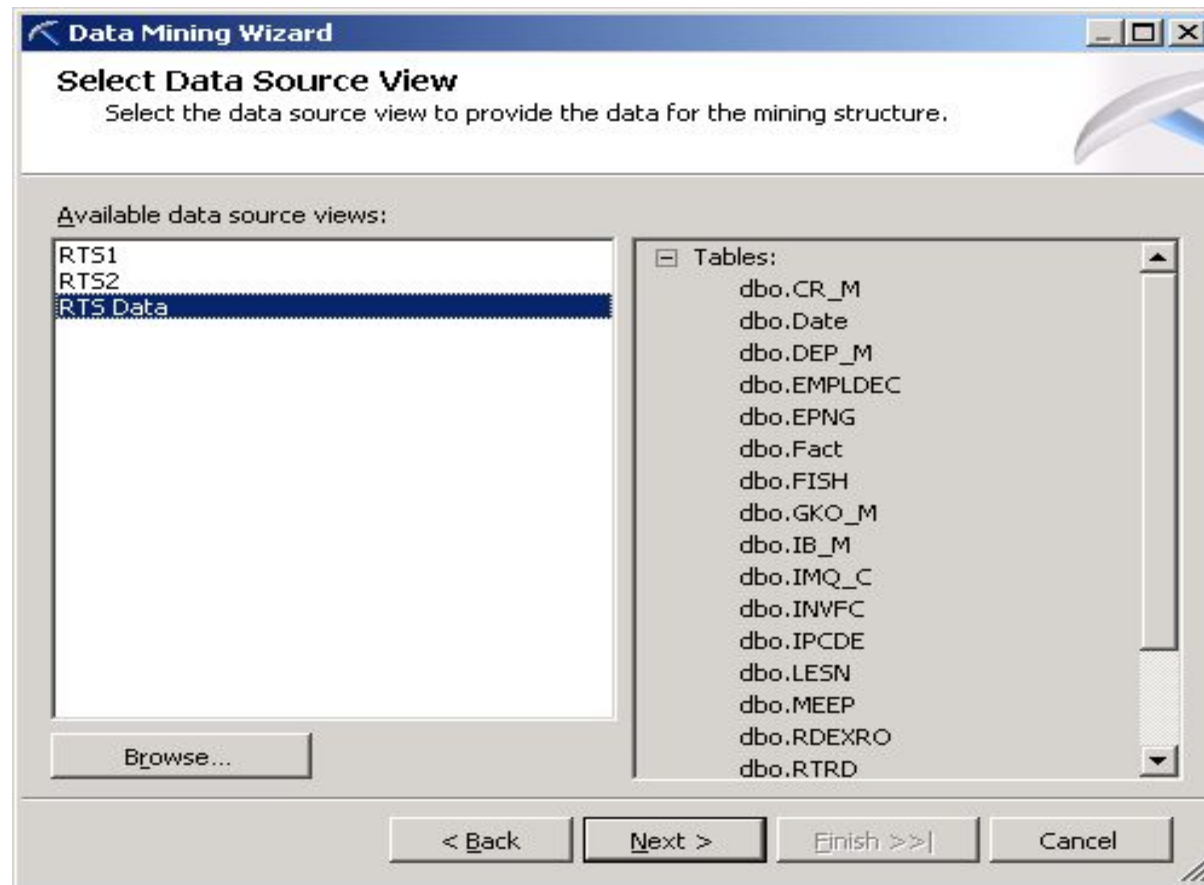
Выбор команды «Новая модель данных»



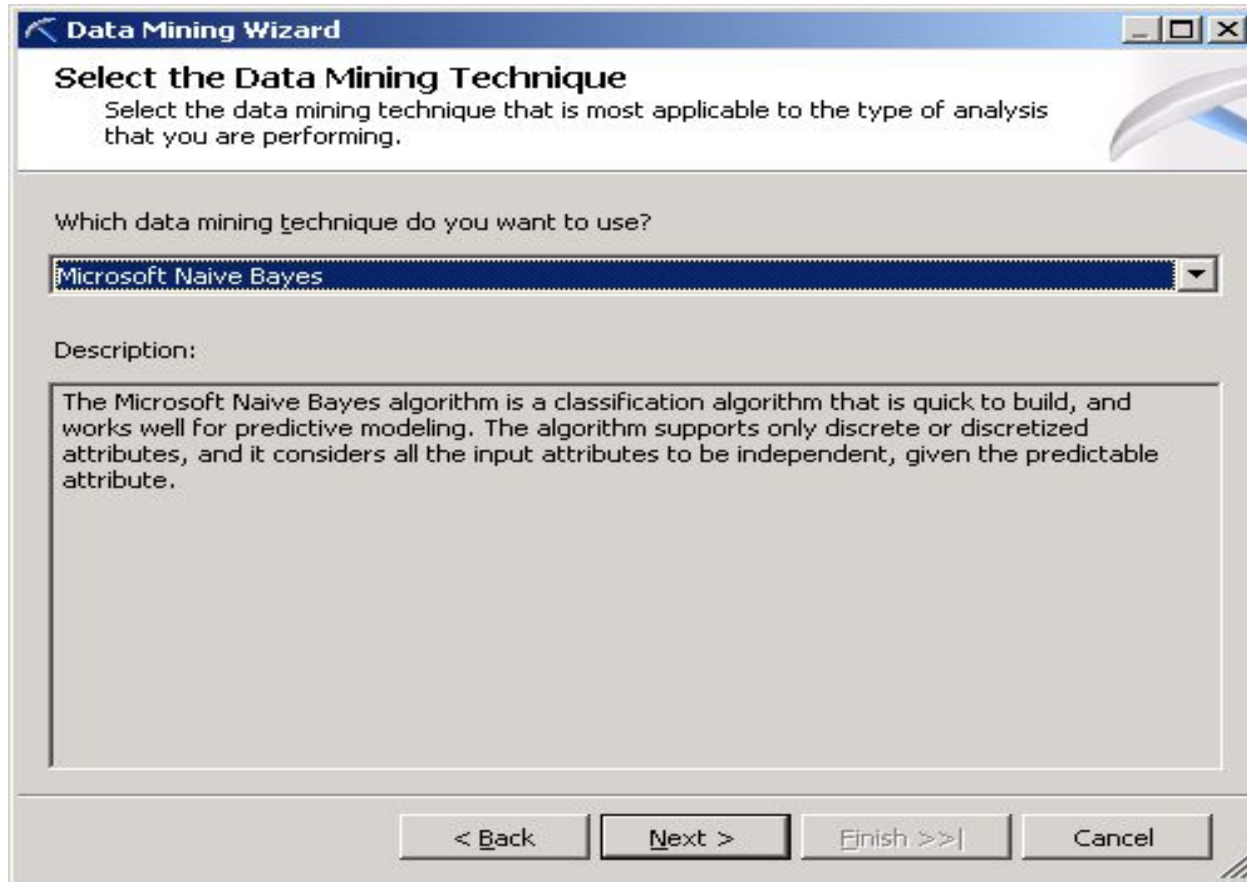
Выбор способа хранения исходных данных



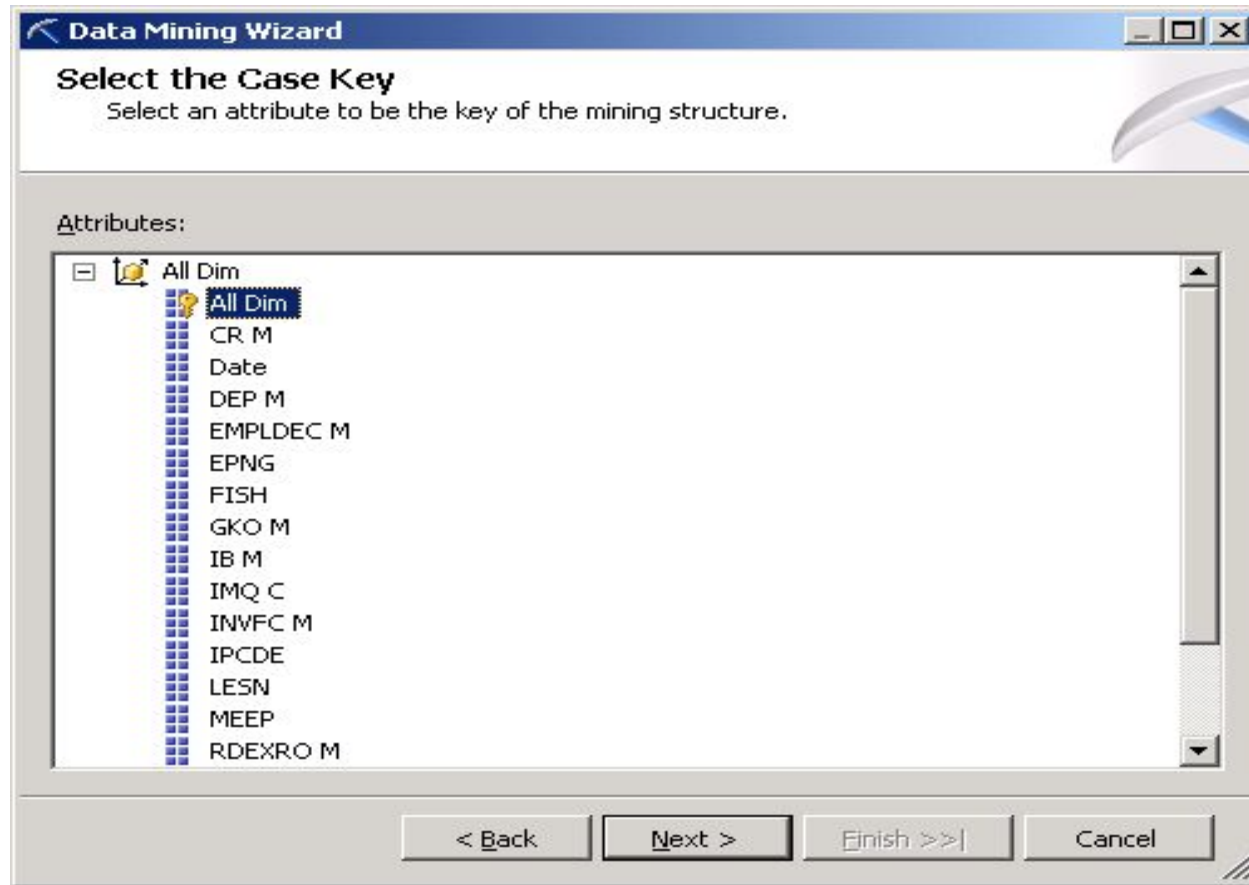
Выбор куба для дальнейшего анализа



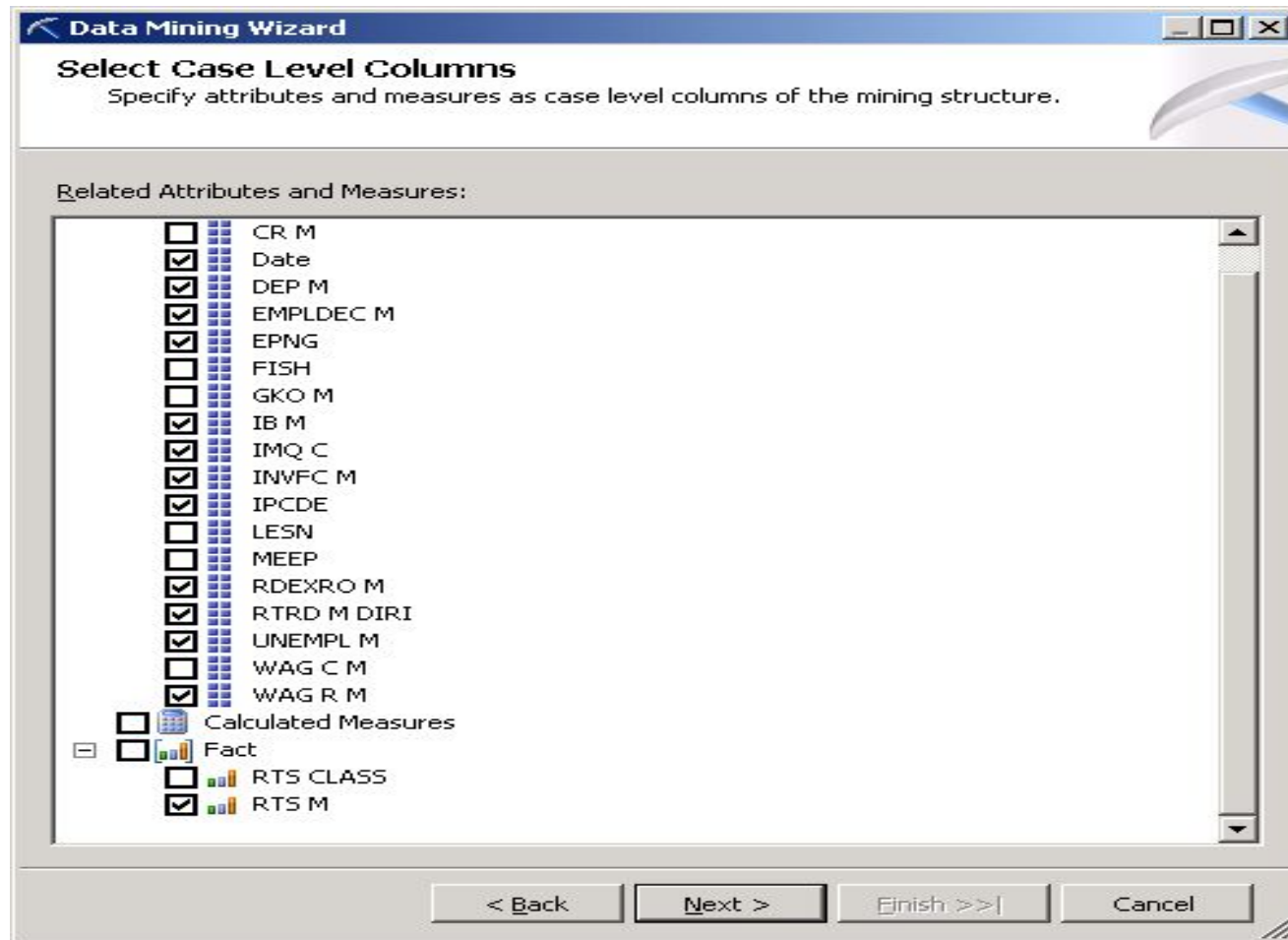
Выбор алгоритма DataMining



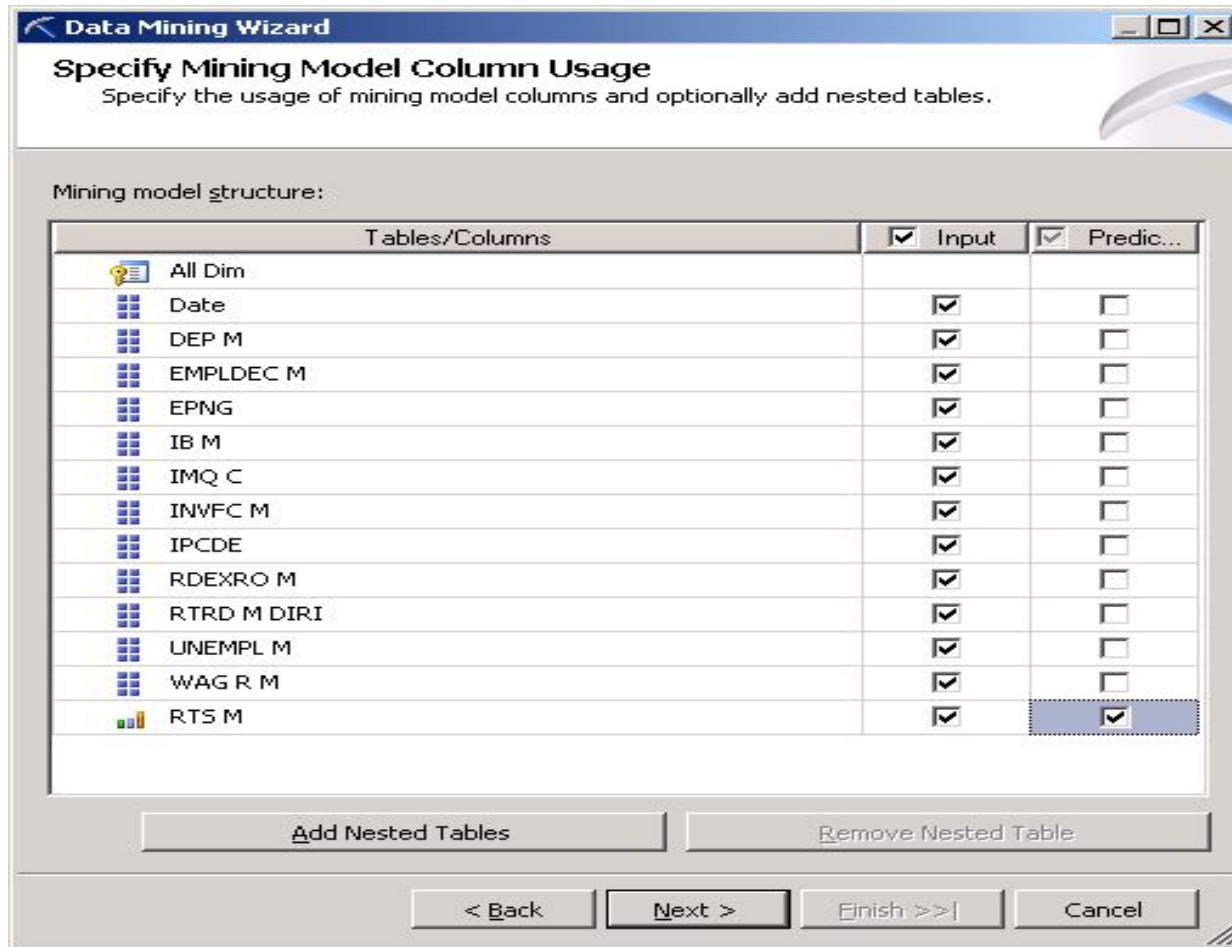
Выбор уровня измерения для многомерной модели



Выбор атрибутов



Разделение атрибутов



Ввод имени модели

Data Mining Wizard

Completing the Wizard
Completing the Data Mining Wizard by providing a name for the mining structure.

Mining structure name:

Mining model name:
 Allow drill through

Preview:

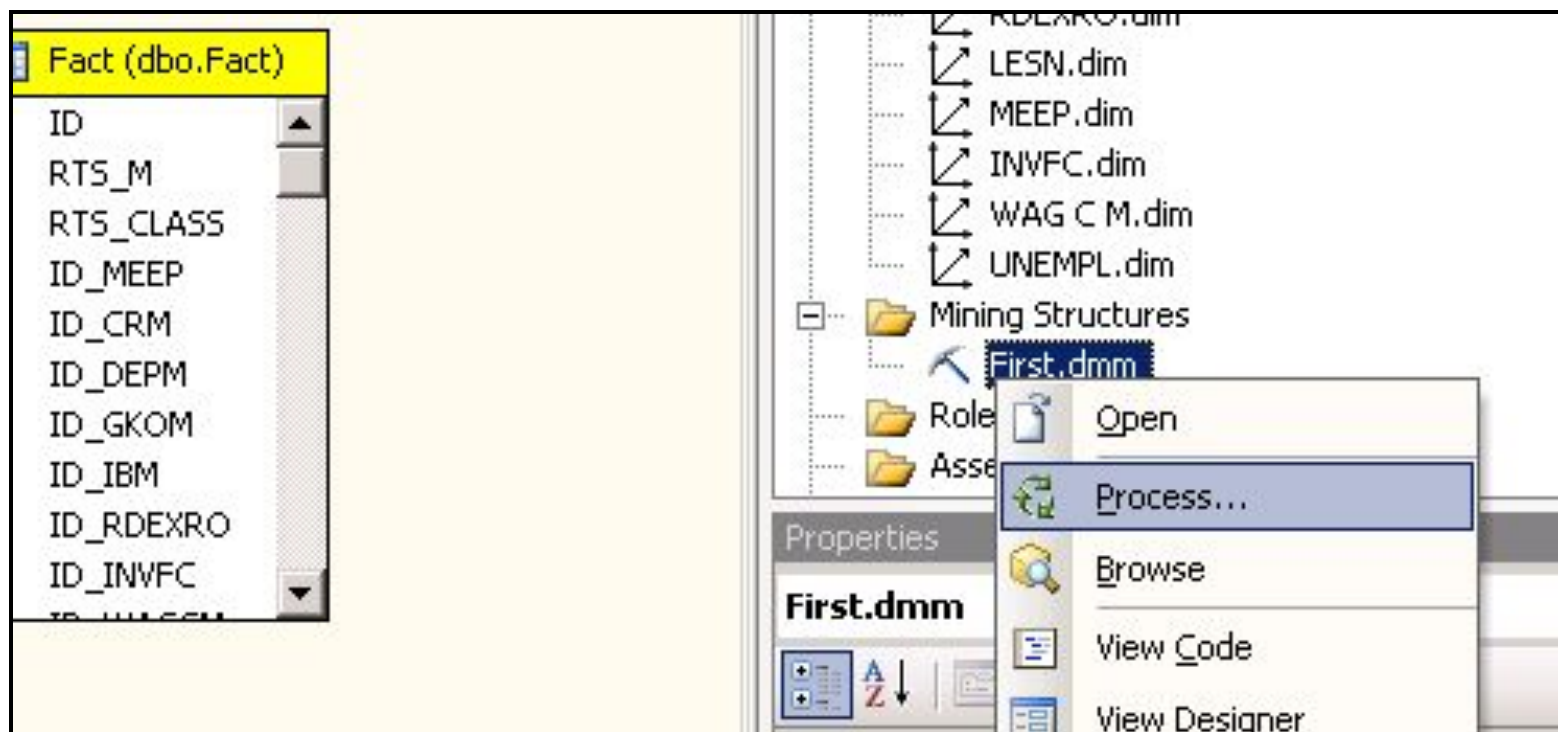
- All Dim_NavBayes-1
 - Columns
 - All Dim
 - Date
 - DEP M
 - EMPLDEC M
 - EPNG
 - IB M
 - IMQ C
 - INVFC M
 - IPCDE
 - RDEXRO M
 - RTD M DDT

Create mining model dimension

Create cube using mining model dimension

< Back Next > Finish Cancel

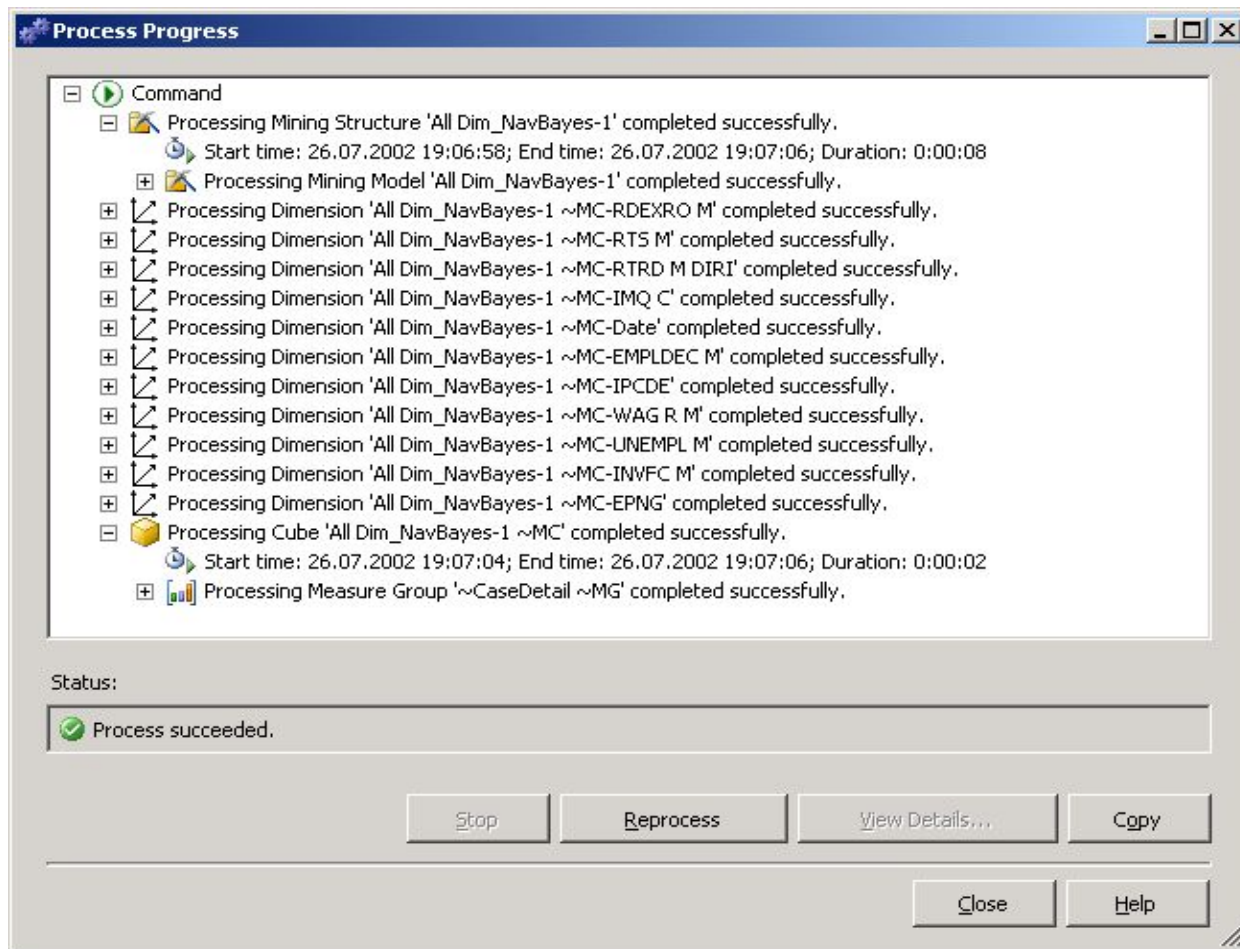
Список моделей Data Mining



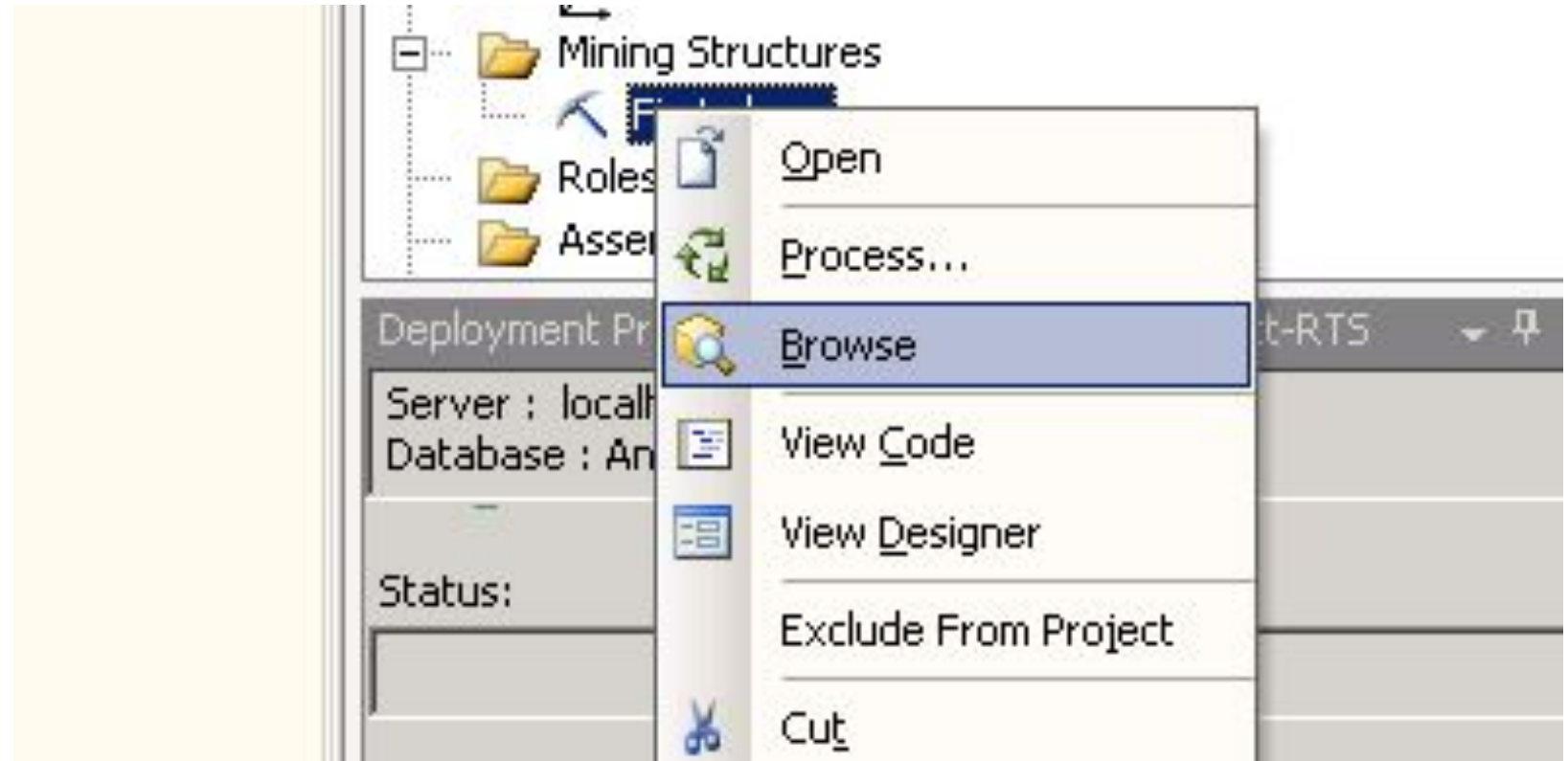
Расчет модели

- В панели **MS_Solution Explorer** щелкните правой кнопкой мыши на имени модели и из контекстного меню выберите **Process Model...** (Процессинг модели...)

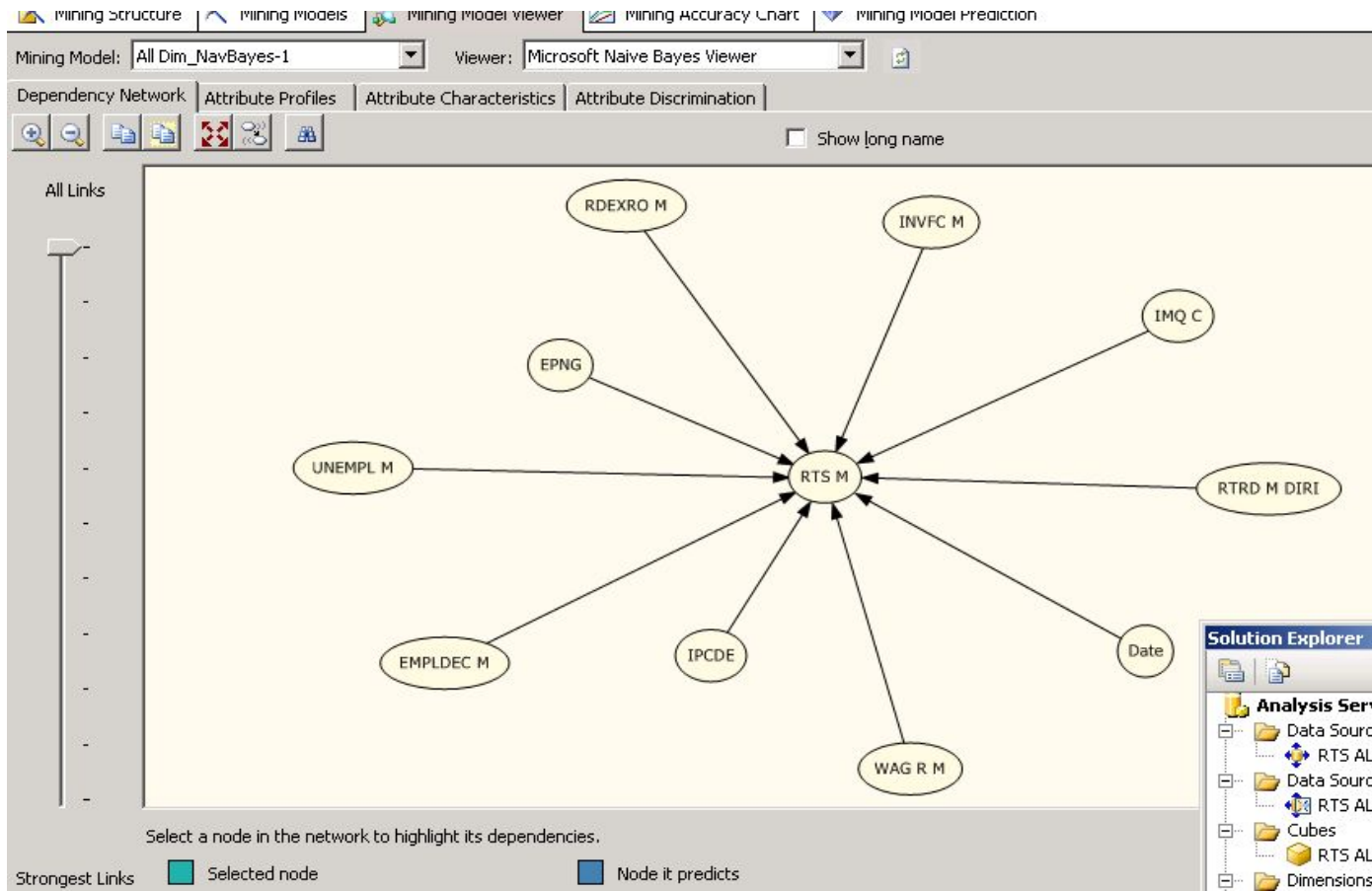
Сообщение об окончании процессинга



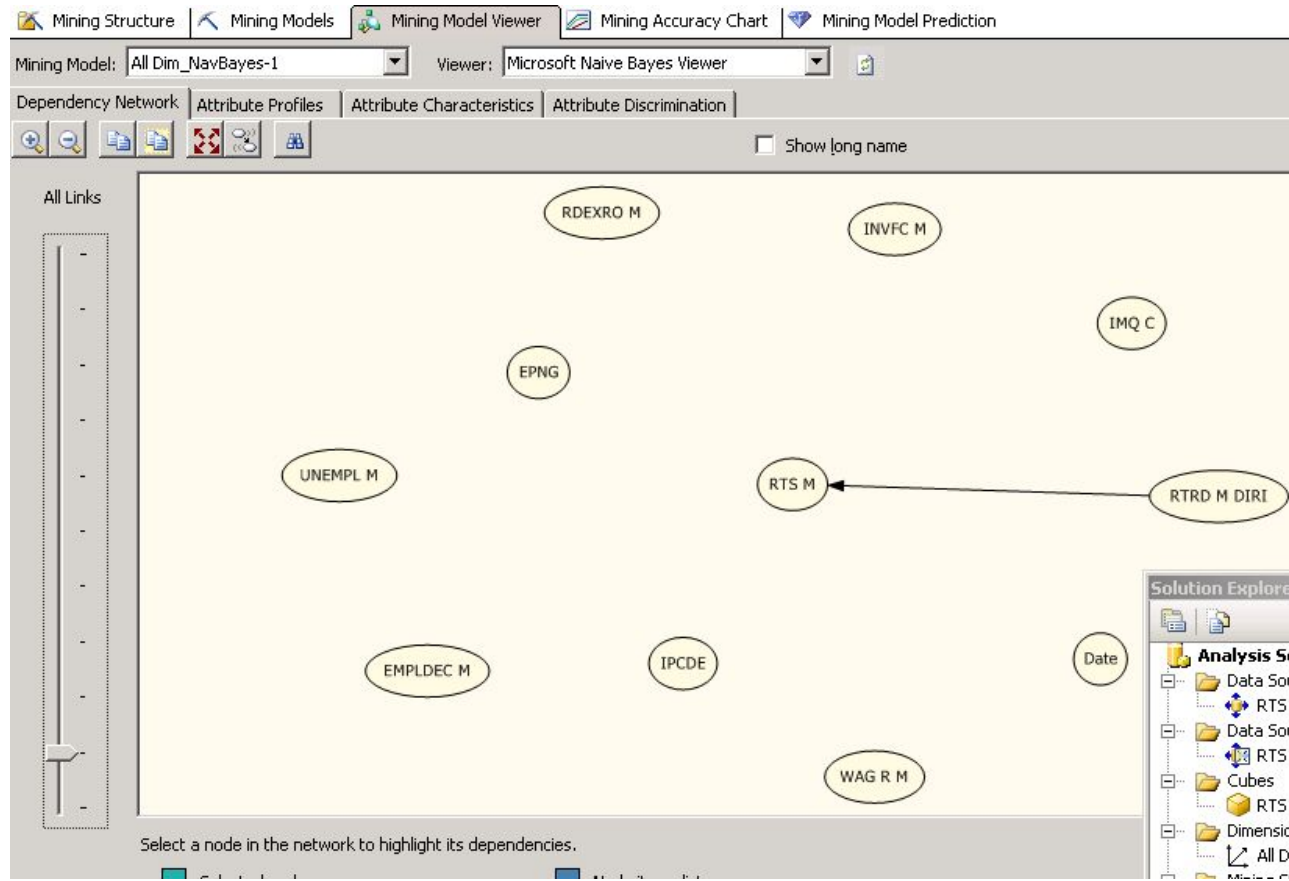
Запуск на иллюстрацию результатов построения модели



Вид модели Naive Bayes в виде графа

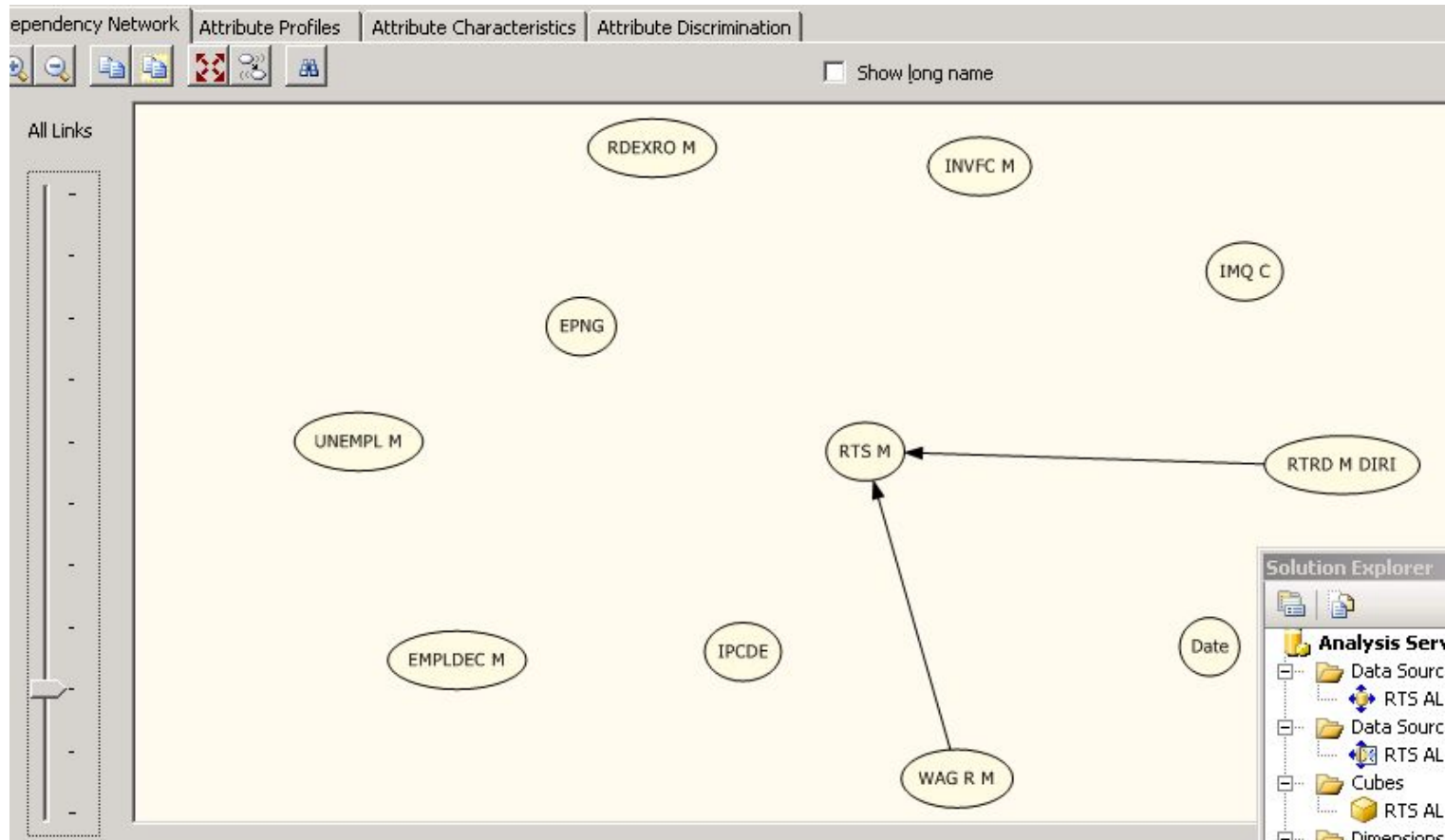


5. Анализ модели Microsoft Naive Bayes

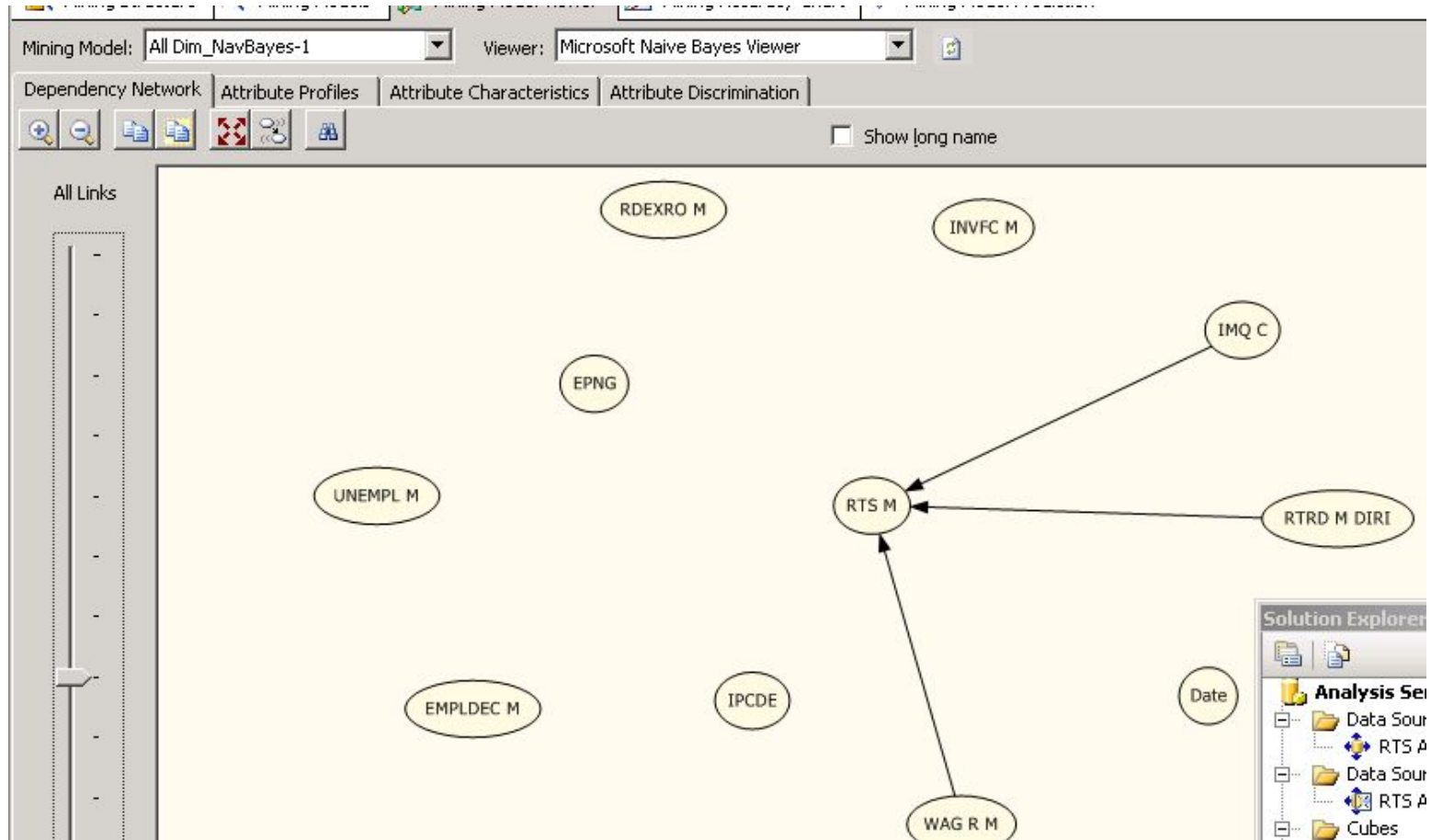


- Появление первой связи между кластерами

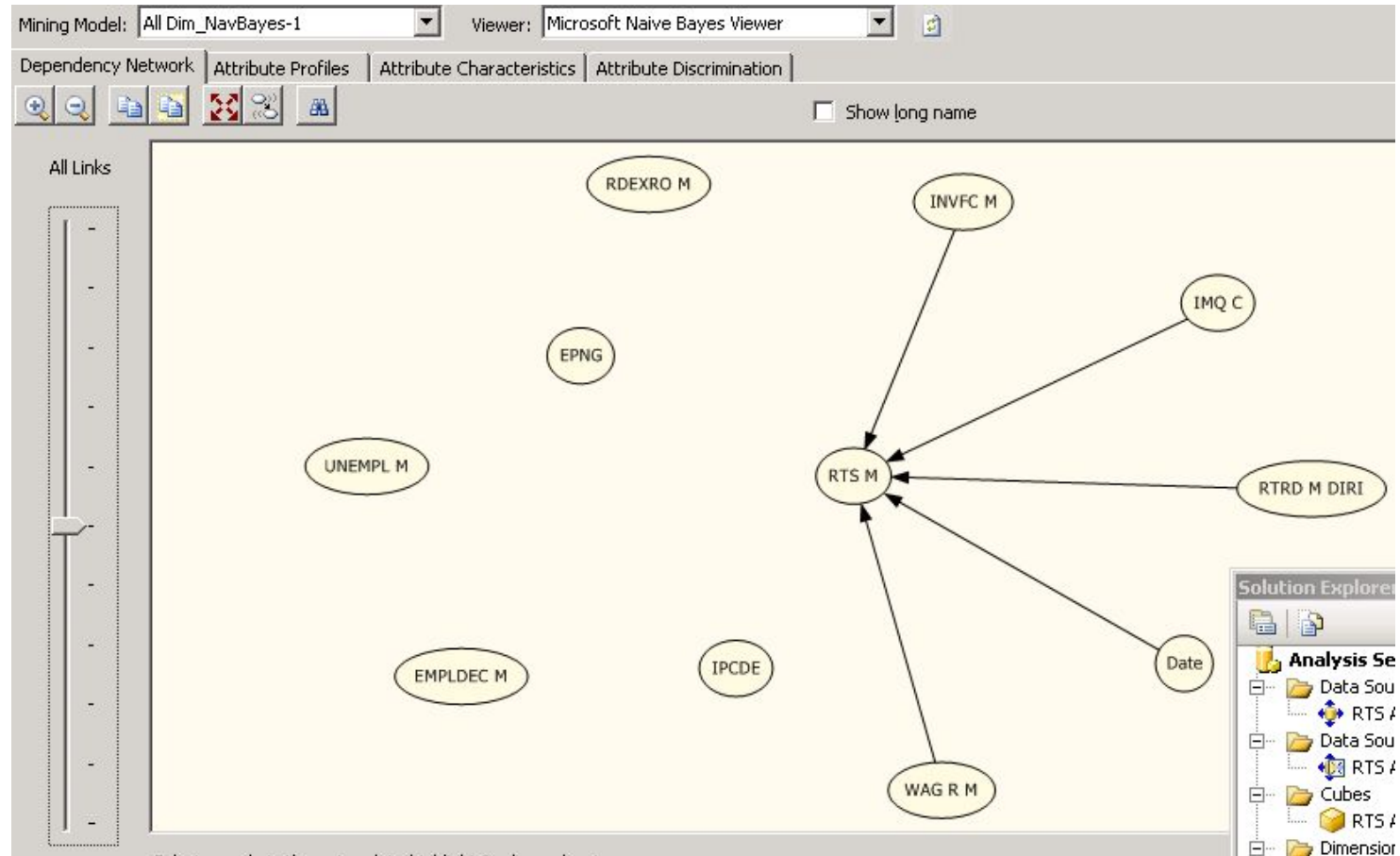
Во вторую очередь появляется влияние уровня средней зарплаты (WAG_R_M)



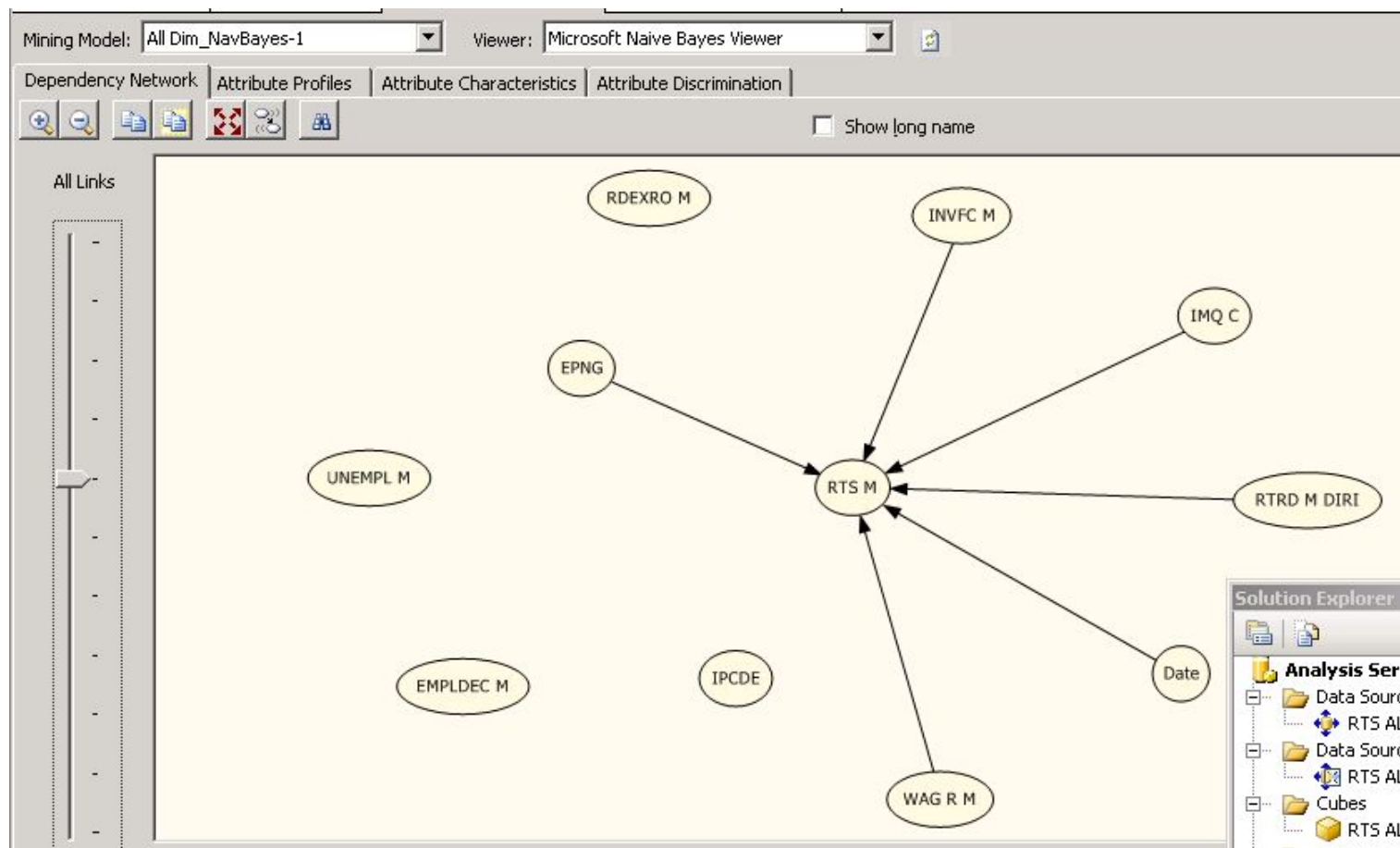
В-третьих, появляется влияние добычи полезных ископаемых (в % к январю 1995 года) IMQ C



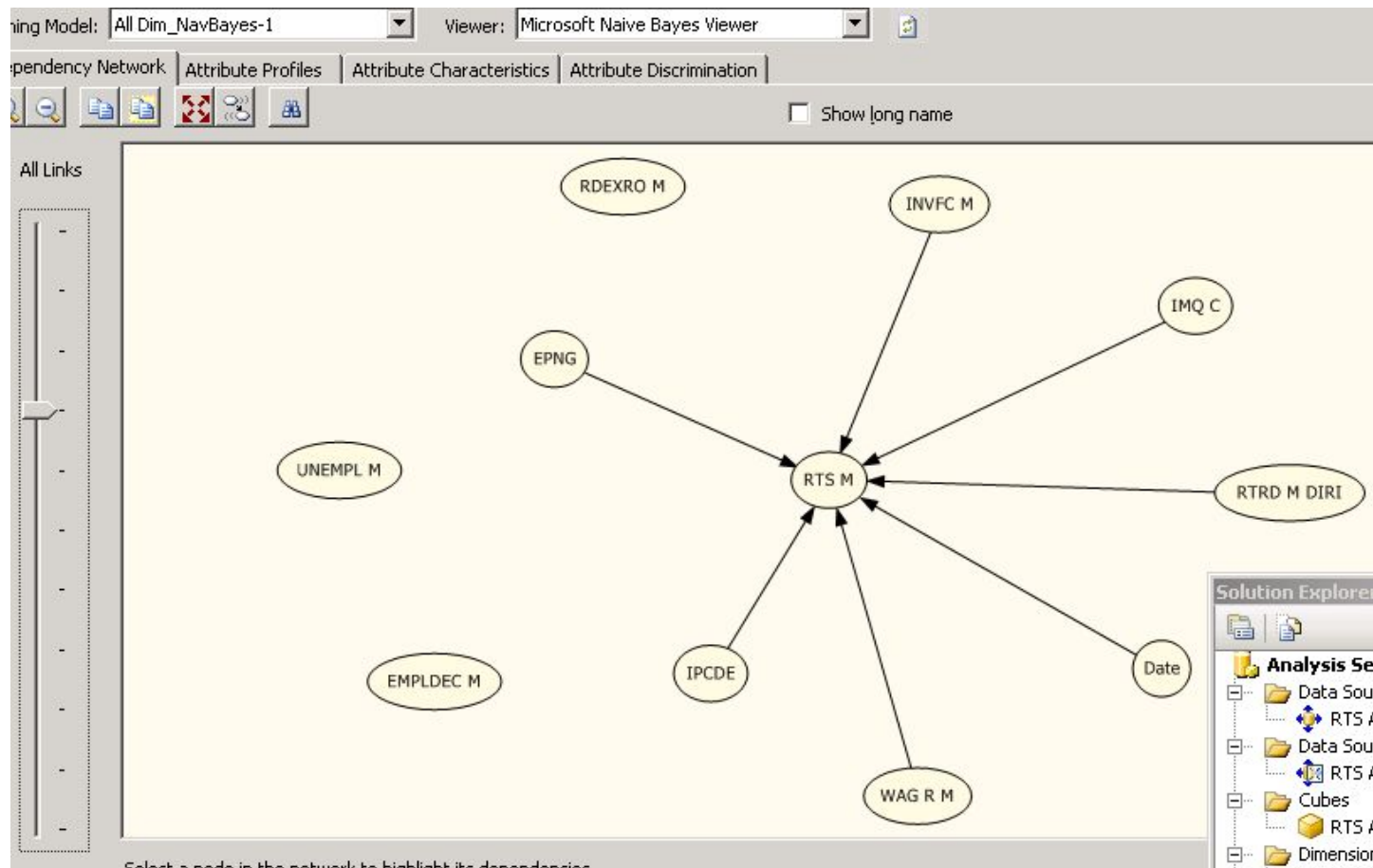
В-четвертых, появляется влияние временного роста и инвестиций в основной капитал в млрд. руб. (Date, INFC M)



Влияние показателя «Добыча сырой нефти и газа»



Влияние показателя «Индекс промышленности».



Построение модели DataMining по технологии MS SQL Server 2000

