**Основные понятия интеллектуального анализа данных**

**Data Mining Concepts**

Интеллектуальный анализ данных представляет собой процесс обнаружения пригодных к использованию сведений в крупных наборах данных. В интеллектуальном анализе данных применяется математический анализ для выявления закономерностей и тенденций, существующих в данных. Обычно такие закономерности нельзя обнаружить при традиционном просмотре данных, поскольку связи слишком сложны, или из-за чрезмерного объема данных.

Эти закономерности и тренды можно собрать вместе и определить как *модель интеллектуального анализа данных*. Модели интеллектуального анализа данных могут применяться к конкретным сценариям, а именно:

* **Прогнозирование**: оценка продаж, прогнозирование нагрузки сервера или времени простоя сервера
* **Риски и вероятности**: Выбор наиболее подходящих заказчиков для целевой рассылки, определение точки равновесия для рискованных сценариев, назначение вероятностей диагнозам или другим результатам
* **Рекомендации**: определение продуктов, которые с высокой долей вероятности могут быть проданы вместе, создание рекомендаций
* **Поиск последовательностей**: анализ выбора заказчиков во время совершения покупок, прогнозирование следующего возможного события
* **Группирование**: разделение заказчиков или событий на кластеры связанных элементов, анализ и прогнозирование общих черт

Построение модели интеллектуального анализа данных является частью более масштабного процесса, в который входят все задачи, от формулировки вопросов относительно данных и создания модели для ответов на эти вопросы до развертывания модели в рабочей среде. Этот процесс можно представить как последовательность следующих шести базовых шагов.

1. [Постановка задачи](https://msdn.microsoft.com/ru-ru/library/ms174949.aspx#DefiningTheProblem)
2. [Подготовка данных](https://msdn.microsoft.com/ru-ru/library/ms174949.aspx#PreparingData)
3. [Просмотр данных](https://msdn.microsoft.com/ru-ru/library/ms174949.aspx#ExploringData)
4. [Построение моделей](https://msdn.microsoft.com/ru-ru/library/ms174949.aspx#BuildingModels)
5. [Исследование и проверка моделей](https://msdn.microsoft.com/ru-ru/library/ms174949.aspx#ValidatingModels)
6. [Развертывание и обновление моделей](https://msdn.microsoft.com/ru-ru/library/ms174949.aspx#DeployingandUpdatingModels)

На следующей диаграмме представлены связи между всеми шагами процесса и технологии Microsoft SQL Server , которые можно использовать для выполнения каждого шага.

Процесс, представленный на следующей диаграмме, является циклическим, то есть создание аналитической модели данных является динамическим и повторяющимся процессом. Выполнив просмотр данных, пользователь может обнаружить, что данных недостаточно для создания требуемых моделей интеллектуального анализа данных, что ведет к необходимости поиска дополнительных данных. Также может возникнуть ситуация, когда после построения нескольких моделей окажется, что они не дают адекватный ответ на поставленную задачу, и поэтому необходимо поставить задачу по-другому. Может возникнуть необходимость в обновлении уже развернутых моделей за счет новых поступивших данных. Для создания хорошей модели может понадобиться многократно повторить каждый шаг процесса.

Интеллектуальный анализ данных Microsoft SQL Server предоставляет интегрированную среду для создания моделей интеллектуального анализа данных и работы с ними. Эта среда включает программу SQL Server Development Studio, которая содержит алгоритмы интеллектуального анализа данных и средства создания запросов, которые облегчают создание полноценного решения для нескольких проектов. Кроме того, эта среда включает компонент Среда SQL Server Management Studio, который содержит средства для поиска моделей и управления объектами интеллектуального анализа данных. Дополнительные сведения см. в разделе [Создание многомерных моделей с помощью средства SQL Server данных & #40; SSDT & #41;](https://msdn.microsoft.com/ru-ru/library/ms174597.aspx).

Пример того, как SQL Server средства могут применяться к бизнес-сценарии см. в разделе [основам интеллектуального анализа данных](https://msdn.microsoft.com/ru-ru/library/ms167167.aspx).

[**Постановка задачи**](javascript:void(0))

Первым шагом процесса интеллектуального анализа данных, как видно из диаграммы ниже, является четкое определение проблемы и рассмотрение способов использования данных для решения проблемы.

Этот шаг включает анализ бизнес-требований, определение области проблемы, метрик, по которым будет выполняться оценка модели, а также определение задач для проекта интеллектуального анализа данных. Эти задачи можно сформулировать в виде следующих вопросов.

* Что необходимо найти? Какие типы связей необходимо найти?
* Отражает ли решаемая задача бизнес-правила или бизнес-процессы?
* Надо ли делать прогнозы на основании модели интеллектуального анализа данных или просто найти содержательные закономерности и взаимосвязи?
* Какой результат или атрибут необходимо спрогнозировать?
* Какие виды данных нужно иметь и какого рода информация находится в каждом столбце? Если существует несколько таблиц, как они связаны? Нужно ли выполнять очистку, статистическую обработку или обработку, чтобы данные стали применимыми?
* Каким образом распределяются данные? Являются ли данные сезонными? Дают ли данные точное представление бизнес-процессов?

Чтобы ответить на эти вопросы, возможно, потребуется исследовать уровень доступности данных, изучить потребности пользователей в отношении доступных данных. Если данные не поддерживают потребности пользователей, то может возникнуть необходимость в изменении определения проекта.

Также необходимо рассмотреть способы для учета результатов модели в ключевых показателях эффективности, которые используются для оценки ведения бизнеса.

[**Подготовка данных**](javascript:void(0))

Вторым шагом процесса интеллектуального анализа данных, как видно из следующей диаграммы, является объединение и очистка данных, определенных во время шага [Постановка задачи](https://msdn.microsoft.com/ru-ru/library/ms174949.aspx#DefiningTheProblem) .

Данные могут находиться в разных частях компании и храниться в различных форматах или содержать такие ошибки согласования, как неверные или отсутствующие записи. Например, согласно данным, может оказаться, что клиент купил товар до того, как тот появился на рынке, или регулярно делает покупки в магазине, расположенном за 3 000 километров от дома.

Очистка данных — это не только удаление недопустимых данных или интерполяция отсутствующих значений, но и поиск в данных скрытых зависимостей, определение источников самых точных данных и подбор столбцов, которые больше всего подходят для использования в анализе. Например, следует ли использовать дату отгрузки или дату заказа? Какой фактор сильнее всего влияет на продажи — количество товара, итоговая цена или цена со скидкой? Неполные данные, ошибочные данные и входные параметры, которые выглядят как независимые, но на самом деле имеют прочную взаимосвязь, могут непредвиденным образом повлиять на результаты модели.

Поэтому перед началом построения моделей интеллектуального анализа данных следует выявить такие проблемы и определить, как их устранить. Во время интеллектуального анализа данных выполняется работа с большими наборами данных и нет возможности проверить каждую транзакцию на предмет качества данных. Поэтому иногда может потребоваться использовать некую форму профилирования данных и средства автоматической очистки и фильтрации данных, например средства, содержащиеся в Службы Integration Services, Службы Microsoft SQL Server 2012 Master Data Servicesили Службы SQL Server Data Quality Services , чтобы исследовать данные и определить несоответствия. Дополнительные сведения см. в разделах:

* [Службы Integration Services в среде Business Intelligence Development Studio](https://technet.microsoft.com/library/ms174181%28v=sql.110%29.aspx)
* [Обзор служб Master Data Services и #40; MDS & #41;](https://msdn.microsoft.com/ru-ru/library/ff487003.aspx)
* [Службы Data Quality Services](https://msdn.microsoft.com/ru-ru/library/ff877925.aspx)

Важно заметить, что данные, используемые для интеллектуального анализа, не обязательно хранить в кубе аналитической обработки в сети (OLAP) или в реляционной базе данных, хотя оба эти типа объектов можно использовать в качестве источника данных. Интеллектуальный анализ данных можно проводить с помощью любого источника, определенного как источник данных служб Службы Analysis Services . Сюда могут относиться текстовые файлы, книги Excel или данные из других внешних поставщиков. Дополнительные сведения см. в разделе [Поддерживаемые источники данных & #40; Службы SSAS — многомерные & #41;](https://msdn.microsoft.com/ru-ru/library/ms175608.aspx).

[**Просмотр данных**](javascript:void(0))

Третьим шагом процесса интеллектуального анализа данных, как видно из диаграммы ниже, является просмотр подготовленных данных.

Для принятия правильных решений при создании моделей интеллектуального анализа данных необходимо понимать данные. Методы исследования данных включают в себя расчет минимальных и максимальных значений, вычисление средневероятного и стандартного отклонения и изучение распределения данных. Например, по максимальному, минимальному и среднему значениям можно заключить, что выборка данных не является репрезентативной для имеющихся клиентов или бизнес-процессов, и поэтому необходимо получить более сбалансированные данные или изменить предположения, лежащие в основе ожидаемых результатов. Стандартное отклонение и другие характеристики распределения могут сообщить полезные сведения о стабильности и точности результатов. Большая величина стандартного отклонения может свидетельствовать о том, что добавление новых данных поможет усовершенствовать модель. Данные, которые сильно отклоняются от стандартного распределения, могут оказаться искаженными или представлять точную картину реальной проблемы, которая делает сложным подбор соответствующей модели для данных.

Изучение данных в свете собственных представлений о бизнес-проблеме может привести к выводу о наличии ошибок в наборе данных, и затем можно выработать стратегию для устранения проблем или получить более глубокое представление о моделях поведения, характерных для бизнеса.

Для просмотра доступных источников данных и определения их доступности для интеллектуального анализа данных используйте средства Службы Master Data Services . Для анализа распределения данных и устранения проблем, таких как неверные или отсутствующие данные, воспользуйтесь таким средством, как Службы SQL Server Data Quality Services, или профилировщиком данных в службах Integration Services.

После определения источников их следует объединить в представлении источников данных с помощью конструктора представлений источников данных в SQL Server Data Tools. Дополнительные сведения см. в разделе [представлений источников данных в многомерных моделях](https://msdn.microsoft.com/ru-ru/library/ms174778.aspx). Конструктор содержит также ряд средств, которые можно использовать для просмотра данных и определения того, подходят ли они для создания модели. Дополнительные сведения см. в разделе [Просмотр данных в представлении источника данных & #40; Службы Analysis Services & #41;](https://msdn.microsoft.com/ru-ru/library/ms174553.aspx).

Обратите внимание, что во время создания модели службы Службы Analysis Services автоматически создают статистические сводки по данным, содержащимся в модели, и эти сводки можно запрашивать для использования в отчетах или при дальнейшем анализе. Дополнительные сведения см. в разделе [запросов интеллектуального анализа данных](https://msdn.microsoft.com/ru-ru/library/bb522459.aspx).

[**Построение моделей**](javascript:void(0))

Четвертым шагом процесса интеллектуального анализа данных, как видно из диаграммы ниже, является построение моделей интеллектуального анализа данных. Знания, полученные при выполнении шага [Просмотр данных](https://msdn.microsoft.com/ru-ru/library/ms174949.aspx#ExploringData) , помогут определить и создать модели.

Пользователь определяет столбцы данных, которые должны быть использованы, путем создания структуры интеллектуального анализа данных. Структура интеллектуального анализа связана с источником данных, но не содержит никаких данных до обработки. Во время обработки структуры интеллектуального анализа службы Службы Analysis Services создают статистические выражения и другие статистические данные, которые могут использоваться в анализе. Эти данные могут использоваться любой моделью интеллектуального анализа данных, которая основана на этой структуре. Дополнительные сведения о структурах интеллектуального анализа данных для модели интеллектуального анализа данных см. в разделе [логической архитектуры & #40; Службы Analysis Services — Интеллектуальный анализ данных & #41;](https://msdn.microsoft.com/ru-ru/library/bb510508.aspx).

Модель интеллектуального анализа данных перед обработкой структуры и модели является просто контейнером, который задает столбцы, используемые для входных данных, прогнозируемый атрибут и параметры, управляющие алгоритмом обработки данных. Обработку модели часто называют *обучением*. Обучение обозначает процесс применения некоторого математического алгоритма к данным в структуре с целью выявить закономерности. Закономерности, обнаруженные в процессе обучения, зависят от выбора обучающих данных, выбранного алгоритма и его конфигурации. SQL Server 2016 содержит множество различных алгоритмов, каждый из которых предназначен для задач различных типов и создает модель, отличную от других. Список алгоритмов, поддерживаемых в SQL Server 2016, в разделе [алгоритмов интеллектуального анализа данных и #40; Службы Analysis Services — Интеллектуальный анализ данных & #41;](https://msdn.microsoft.com/ru-ru/library/ms175595.aspx).

Для настройки каждого алгоритма используются параметры, и кроме того, можно применить фильтры к обучающим данным, чтобы использовать только их подмножество, что приведет к получению других результатов. После прохождения данных через модель объект модели интеллектуального анализа данных будет содержать сводные данные и закономерности, которые можно запрашивать и использовать для прогнозирования.

Можно определить с помощью мастера интеллектуального анализа данных в новую модель SQL Server Data Tools, или при помощи языка расширений интеллектуального анализа данных (DMX). Дополнительные сведения об использовании мастера интеллектуального анализа данных см. в разделе [Мастер интеллектуального анализа данных и #40; Службы Analysis Services — Интеллектуальный анализ данных & #41;](https://msdn.microsoft.com/ru-ru/library/ms175645.aspx). Дополнительные сведения об использовании расширений интеллектуального анализа данных см. в разделе [расширений интеллектуального анализа данных & #40; расширений интеллектуального анализа данных & #41; Справочник по](https://msdn.microsoft.com/ru-ru/library/ms132058.aspx).

Важно помнить, что при любом изменении данных необходимо обновить и структуру, и модель интеллектуального анализа данных. Когда структура интеллектуального анализа данных обновляется путем повторной обработки, службы Службы Analysis Services получают данные из источника, в том числе новые данные, если источник динамически обновляется, и повторно заполняет структуру интеллектуального анализа данных. Если на этой структуре основаны существующие модели, можно обновить эти модели, что будет означать их повторное обучение с новыми данными, или оставить модели без изменений. Дополнительные сведения см. в разделе [обработки требования и вопросы & #40; интеллектуального анализа данных & #41;](https://msdn.microsoft.com/ru-ru/library/ms176135.aspx).

[**Исследование и проверка моделей**](javascript:void(0))

Пятым шагом процесса интеллектуального анализа данных, как видно из диаграммы ниже, является исследование построенных моделей интеллектуального анализа данных и проверка их эффективности.

Перед развертыванием модели в рабочей среде необходимо проверить эффективность работы модели. Кроме того, во время построения модели обычно создается несколько моделей с различной конфигурацией, а затем проверяются все модели, чтобы определить, какая из них обеспечивает лучшие результаты для поставленной задачи и имеющихся данных.

Службы Analysis Services предоставляют средства, упрощающие разделение данных на набор данных для обучения и проверочный набор данных, чтобы можно было точно оценить производительность всех моделей, основанных на одних и тех же данных. Набор данных для обучения используется в ходе построения модели, а набор проверочных данных — для проверки точности модели путем создания прогнозирующих запросов. В службах SQL Server 2016 Analysis Services (SSAS)такое секционирование можно выполнить автоматически во время построения модели интеллектуального анализа данных. Дополнительные сведения см. в разделе [тестирования и проверки & #40; интеллектуального анализа данных & #41;](https://msdn.microsoft.com/ru-ru/library/ms174493.aspx).

Тенденции и закономерности, обнаруживаемые алгоритмами, можно исследовать при помощи средств просмотра в конструкторе интеллектуального анализа данных в среде SQL Server Data Tools. Дополнительные сведения см. в разделе [просмотра моделей интеллектуального анализа данных](https://msdn.microsoft.com/ru-ru/library/ms174580.aspx). Точность прогнозов, создаваемых моделями, можно проверить при помощи таких средств конструктора, как диаграмма точности прогнозов и матрица классификации. Чтобы проверить ли модель зависит от данных, или может использоваться для совершения выводов относительно генеральной совокупности, можно использовать статистический метод, называемый *перекрестной проверки* для автоматического создания подмножества данных и проверить модель по каждому подмножеству. Дополнительные сведения см. в разделе [тестирования и проверки & #40; интеллектуального анализа данных & #41;](https://msdn.microsoft.com/ru-ru/library/ms174493.aspx).

Если ни одна из моделей, созданных при выполнении шага [Построение моделей](https://msdn.microsoft.com/ru-ru/library/ms174949.aspx#BuildingModels) , не обладает нужной эффективностью, может возникнуть необходимость вернуться к предыдущему шагу процесса и либо изменить постановку задачи либо выполнить повторное изучение данных в исходном наборе данных.

[**Развертывание и обновление моделей**](javascript:void(0))

Последним шагом процесса интеллектуального анализа данных, как видно из диаграммы ниже, является развертывание наиболее эффективных моделей в рабочей среде.

После развертывания моделей интеллектуального анализа данных в рабочей среде можно выполнять множество задач, соответствующих потребностям пользователя. Ниже перечислены некоторые задачи, доступные для выполнения.

* Используйте модели для создания прогнозов, которые можно затем использовать для принятия бизнес-решений. SQL Server предоставляет язык расширений интеллектуального анализа данных, который можно использовать для создания запросов прогнозирования, и построитель запросов прогнозирования, облегчающий процесс построения запросов. Дополнительные сведения см. в разделе [расширений интеллектуального анализа данных & #40; расширений интеллектуального анализа данных & #41; Справочник по](https://msdn.microsoft.com/ru-ru/library/ms132058.aspx).
* Создание запросов содержимого для получения статистики, правил или формул из модели. Дополнительные сведения см. в разделе [запросов интеллектуального анализа данных](https://msdn.microsoft.com/ru-ru/library/bb522459.aspx).
* Внедрение функций интеллектуального анализа данных непосредственно в приложение. Можно включать объекты AMO, содержащие набор объектов, которые приложение может использовать для создания, изменения, обработки и удаления структур и моделей интеллектуального анализа данных. Кроме того, можно отправлять сообщения XML для аналитики (XMLA) напрямую в экземпляр служб Службы Analysis Services. Дополнительные сведения см. в разделе [Разработка (службы Analysis Services — Интеллектуальный анализ данных)](https://technet.microsoft.com/library/bb522473%28v=sql.110%29.aspx).
* Использование служб Службы Integration Services для создания пакета, в котором модель интеллектуального анализа данных используется для интеллектуального распределения входящих данных по разным таблицам. Например, если база данных постоянно обновляется за счет добавления потенциальных клиентов, модель интеллектуального анализа данных может использоваться совместно со службами Службы Integration Services для разбиения входящих данных на клиентов, которые с высокой долей вероятности купят товар, и клиентов, которые с высокой долей вероятности не купят товар. Дополнительные сведения см. в статье [Типовое использование служб Integration Services](https://msdn.microsoft.com/ru-ru/library/ms137795.aspx).
* Создание отчета, который позволит пользователям составлять прямые запросы к существующей модели интеллектуального анализа данных. Дополнительные сведения см. в разделе [служб Reporting Services в SQL Server Data Tools & #40; SSDT & #41;](https://msdn.microsoft.com/ru-ru/library/ms173745.aspx).
* Обновление моделей после просмотра и анализа. После любого обновления необходимо выполнить повторную обработку моделей. Дополнительные сведения см. в статье [Processing Data Mining Objects](https://msdn.microsoft.com/ru-ru/library/cc645741.aspx).
* Динамическое обновление моделей по мере поступления в организацию новых данных и постоянные изменения, направленные на повышение эффективности решения, должны быть частью стратегии развертывания. Дополнительные сведения см. в разделе [из данных интеллектуального анализа данных решения для управления и объекты](https://msdn.microsoft.com/ru-ru/library/bb522447.aspx)