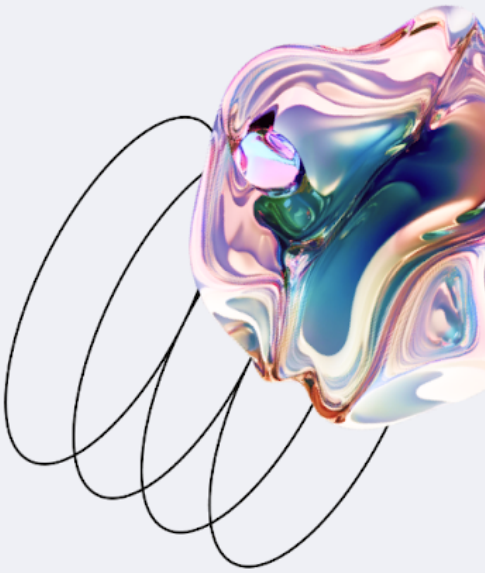
GeekBrains



Введение в подготовку данных для аналитиков. Таблицы фактов и таблицы измерений

ETL: автоматизация подготовки данных

Оглавление

Введение

Всем привет! Это наша вторая лекция на курсе «ETL: автоматизация подготовки данных».

На прошлом уроке мы рассмотрели основные аспекты процесса ETL и поговорили о лучших практиках применения. Обсудили существующие модели данных и немного углубились в реляционную модель. Разобрали процесс нормализации таблиц, поговорили о нормальных формах и способах перехода из одной формы в другую. А также узнали, что такое хранилище данных и затронули схему хранилища «Звезда».

На сегодняшнем уроке мы узнаем, что такое бизнес-аналитика и для чего она нужна. Затем разберем, какие этапы нужно пройти для качественного анализа данных. А в завершении поговорим о таблицах фактов и измерений.

Термины, используемые в лекции

Бизнес-аналитика (Business intelligence, BI) — это набор процессов, программ и услуг, цель которых — преобразовать необработанные данные в значимую информацию, которая может принести бизнесу прибыль.

Таблица фактов — основная таблица хранилища данных. Как правило, содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться.

Таблица измерений — таблица, которая содержит атрибуты событий, сохраненных в таблице фактов.

Введение в подготовку данных для аналитиков

Бизнес-аналитика (Business intelligence, BI) — это набор процессов, программ и услуг, цель которых — преобразовать необработанные данные в значимую информацию, которая может принести бизнесу прибыль.

Бизнес-аналитика напрямую влияет на стратегические, тактические и оперативные решения компании. Помогает принимать решения на основе фактов и исторических данных, а не предположений и интуиции.

Инструменты бизнес-аналитики исследуют данные и создают отчеты, сводки, информационные панели, карты, графики и диаграммы, чтобы дать пользователям подробные сведения о бизнесе.

Для чего нужна бизнес-аналитика?

- Чтобы создать KPI (ключевые показатели эффективности) на основе исторических данных.
- Чтобы определить и установить критерии для процессов.
- Чтобы выявить тенденции рынка и проблемы, которые бизнесу нужно решить.
- Чтобы визуализировать данные — повысить их качество и качество принятия решений.

Системы BI подходят не только крупным, но также средним и малым предприятиям (МСП).

Рассмотрим несколько примеров использования бизнес-аналитики.

Пример 1:

Владелец отеля использует аналитические BI-приложения, чтобы собирать статистику о средней загруженности и стоимости номера. Это решение поможет найти совокупный доход за номер.

Владелец также собирает статистику о доле рынка и данные опросов клиентов других отелей, чтобы определить свою конкурентную позицию.

Анализируя эти тенденции из года в год, месяц за месяцем и день за днем, владелец отеля может предлагать скидки на аренду или другие маркетинговые акции для привлечения клиентов.

Пример 2:

Банк предоставляет менеджерам филиалов доступ к BI-приложениям. С их помощью менеджеры филиалов могут определить, какие клиенты более прибыльные и как с ними работать.

Инструменты BI освобождают ИТ-сотрудников от составления аналитических отчетов для разных отделов и дают персоналу доступ к богатым источникам данных.

Тенденции в бизнес-аналитике

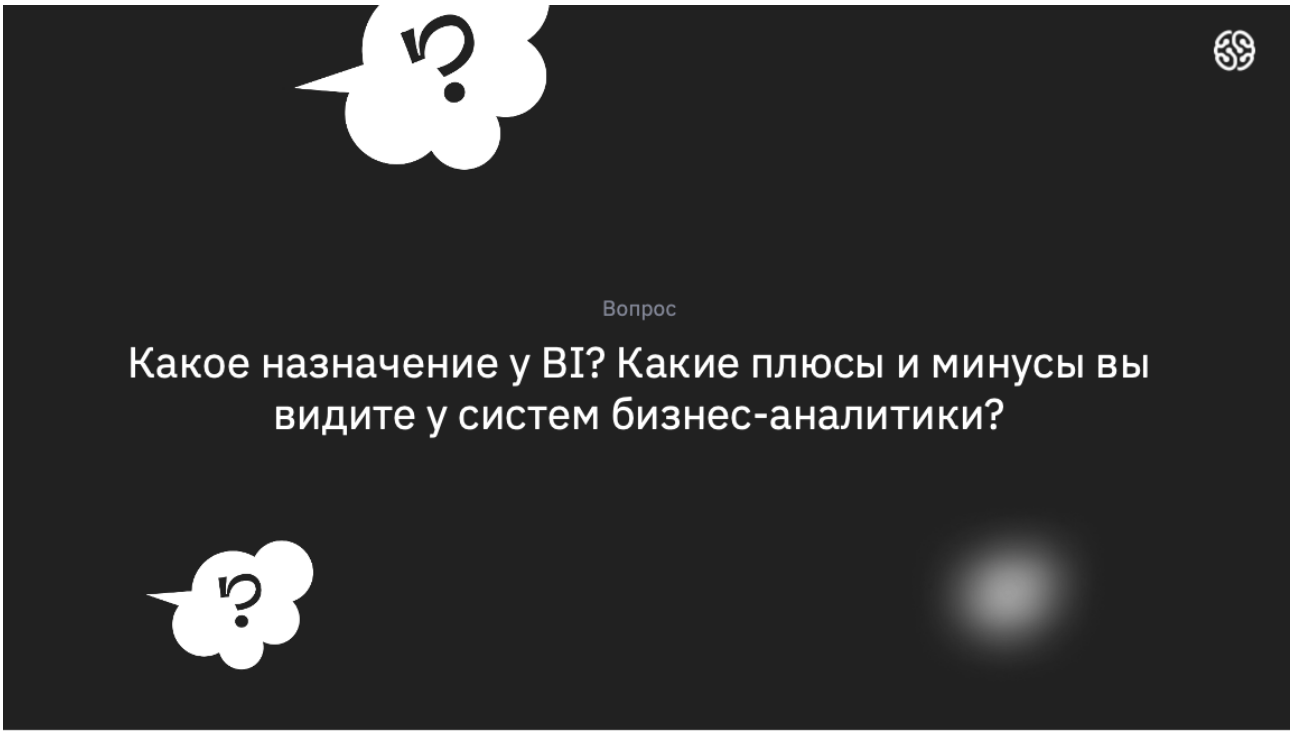
Искусственный интеллект. По отчету Gartner, ИИ и машинное обучение справляются со сложными задачами, которые раньше были под силу только людям. Например, их используют для анализа данных в реальном времени и для создания отчетов.

Collaborative BI. BI-приложения с инструментами для совместной работы расширяют возможности для обмена информацией и для принятия решений в команде.

Embedded BI. Программное обеспечение BI или некоторые его функции можно интегрировать в другие бизнес-приложения, чтобы расширить их возможности отчетности.

Облачная аналитика. BI-приложения скоро будут доступны в облаке и многие компании перейдут на эту технологию. Согласно прогнозам, в течение пары лет расходы на облачную аналитику будут расти в 4,5 раза быстрее.

Давайте подведем итоги и ответим на вопрос:



Ответ: BI-системы помогают предприятиям определить тенденции рынка и выявить проблемы, которые нужно решить.

BI-технологии будут полезны аналитикам данных, ИТ-специалистам, бизнес-пользователям и главам компаний.

Недостатки BI в том, что это дорогостоящий и очень сложный процесс.

Анализ данных

Анализ данных — это процесс из нескольких шагов: сперва сырые данные нужно обработать, чтобы затем создать из них визуализации и сделать предсказания на основе математической модели.

Каждый шаг в анализе данных играет ключевую роль для следующих шагов:

1. Определение проблемы.
2. Извлечение данных.
3. Подготовка данных:
    - очистка,
    - преобразование.
4. Исследование и визуализация данных.
5. Моделирование.
6. Оценка (проверка) модели.

7. Развертывание:
    - визуализация и интерпретация результатов,
    - развертывание решения.

Процесс анализа данных на схеме:

Давайте подробнее изучим процесс анализа данных.

Понимание бизнеса

На этом этапе устанавливаются цели бизнеса и добычи данных.

Задачи:

1. Понять цели бизнеса и клиента. Определить, чего хотят клиенты (часто они сами этого не знают).
2. Подвести итоги текущего сценария добычи данных и текущих бизнес-процессов. Найти ресурсы для добычи данных, проанализировать ограничения и другие факторы, которые могут повлиять на оценку.

3. Определить цели анализа данных на основе бизнес-целей и текущего сценария.
4. Разработать подробный план анализа данных. Фокус — на достижении целей как бизнеса, так и данных.

Понимание данных

На этом этапе проверяется работоспособность данных. Цель — убедиться, подходят ли они для анализа.

Задачи:

1. Собрать данные из нескольких источников, доступных в организации (БД, файлов, хранилищ данных). Это сложный процесс, так как данные из разных источников нелегко сопоставить.

   Например, в таблице A может быть сущность с именем «Номер клиента», а в таблице B — «Идентификационный номер клиента». Трудно гарантировать, что оба объекта ссылаются на одно и то же значение.

   Чтобы уменьшить ошибки в процессе интеграции, нужно использовать метаданные.

2. Выполнить поиск свойств полученных данных. Хороший способ исследовать данные — ответить на вопросы интеллектуального анализа данных (решенные в бизнес-фазе), используя инструменты запросов, отчетов и визуализации.
3. Установить качество данных. Если каких-то данных нет, их нужно получить.

Подготовка данных

На этом этапе данные готовят к анализу. Подготовка занимает около 90% времени проекта.

Задачи — отобрать, очистить, преобразовать, отформатировать и, если нужно, анонимизировать данные из разных источников.

Нужно сгладить зашумленные данные и заполнить пропущенные значения. Например, если в демографическом профиле клиента нет данных о возрасте, данные считаются неполными, их нужно заполнить.

В некоторых случаях могут быть выбросы данных, например, значение 300 в возрасте. Данные могут быть противоречивыми, например, когда имя клиента отличается в разных таблицах.

Операции преобразования данных:

- Сглаживание — удаление шума из данных.
- Агрегация (сводка) — компоновка данных. Например, еженедельные или ежемесячные отчеты группируются, чтобы получить сумму продаж за год.
- Обобщение — замена данных низкого уровня концепциями более высокого на основе предопределенной иерархии. Например, город заменяется областью.
- Нормализация — сглаживание цифровых данных. Например, у нас цифровые значения в большом разбросе диапазонов. Чтобы нормализовать данные, мы можем привести их к значениям от 0 до 1.
- Построение атрибутов — атрибуты включают в себя заданный набор полезных для анализа данных.

Результат процесса — окончательный набор данных, который можно использовать при моделировании.

Исследование и визуализация данных

Изучение данных — это их анализ в графической или статистической форме с целью поиска моделей или взаимосвязей.

> 💡 Визуализация — лучший инструмент для выделения моделей.

За последние годы визуализация данных развилась так сильно, что стала независимой дисциплиной. Некоторые технологии только отображают данные, другие работают так, чтобы пользователь получил лучшую информацию из набора данных.

Этап состоит из шагов:

1. Обобщение данных.
2. Группировка данных.
3. Исследование отношений между разными атрибутами.

4. Определение моделей и тенденций.
5. Построение моделей регрессионного анализа.
6. Построение моделей классификации.

Обобщение — это процесс, при котором количество данных для интерпретации уменьшают без потери важной информации.

Группировка или кластерный анализ — это метод поиска групп, схожих между собой, объединенных общими атрибутами.

Идентификация отношений, тенденций и аномалий в данных — еще один важный этап анализа. Для поиска такой информации часто нужно использовать инструменты и проводить дополнительные этапы анализа, но уже на визуализациях.

Другие методы поиска данных (деревья решений и ассоциативные правила) автоматически извлекают важные факты или правила из данных. Эти подходы используются параллельно с визуализацией для поиска взаимоотношений данных.

Моделирование

Предсказательная аналитика — это процесс в анализе данных, нужный для создания или поиска подходящей статистической модели для предсказания вероятности наступления определенного результата.

После того как мы изучим данные, у нас будет вся информация для построения математической модели. Эта модель свяжет отношения и тенденции, которые просматриваются в данных.

Модели полезны для понимания изучаемой системы и используются в двух направлениях:

1. **Предсказания о значениях данных**, которые создает система. В этом случае речь идет о регрессионных моделях.
2. **Классификация новых продуктов**. Это модели классификации или модели кластерного анализа. Можно разделить модели в соответствии с типом результатов, к которым те приводят:
    - **Модели классификации**, если результат — качественная переменная.

    - **Регрессионные модели**, если результат числовой.
    - **Кластерные модели**, если результат описательный.

Простые методы генерации этих моделей включают техники:

- линейная регрессия,
- логистическая регрессия,
- классификация,
- дерево решений,
- метод k-ближайших соседей.

Таких методов много, у каждого свои характеристики, благодаря которым они подходят для определенных типов данных и анализа. Каждый из них приводит к появлению определенной модели, а их выбор соответствует природе модели продукта.

Некоторые из методов будут предоставлять значения, относящиеся к реальной системе. Они смогут объяснить некоторые характеристики изучаемой системы простым способом. Другие будут делать хорошие предсказания, но их структура будет оставаться «черным ящиком» с ограниченной способностью объяснить характеристики системы.

Оценка

На этом этапе зависимости оцениваются в соответствии с бизнес-целями.

Задачи:

- Сопоставить результаты интеллектуального анализа данных с бизнес-целями. Нужно понять, насколько они соответствуют задаче.
- Определиться, нужен ли дополнительный анализ. Понимание бизнеса — итеративный процесс. Иногда на основе открывшихся показаний может быть принято решение о дополнительном анализе.
- Принять решение о переходе модели на этап развертывания.

Развертывание

На этапе развертывания открытия интеллектуального анализа данных отправляются в повседневные бизнес-операции.

Задачи:

- Оформить полученные знания и информацию так, чтобы она была понятна нетехническим заинтересованным специалистам.
- Создать подробный план развертывания для доставки, обслуживания и мониторинга результатов анализа данных.
- Подготовить окончательный отчет с учетом уроков и ключевых событий проекта.

Таблицы фактов и таблицы измерений

Мы уже немного касались таблиц фактов и измерений, когда говорили о схеме «Звезда»: в ее центре находится таблица фактов, а на лучах — таблицы измерений. В этой главе поговорим о них подробнее и поймем, чем они отличаются.

Таблицы фактов и измерений содержат основные данные для детального анализа.

- **Таблица фактов** позволяет пользователю анализировать верхний уровень, бизнес-аспекты, которые помогают принимать решения.
- **Таблицы измерений** помогают таблице фактов собирать уникальные размерности, то есть измерения показателей, по которым должны быть приняты решения.

> 💡 Запись таблицы фактов представляет собой комбинацию атрибутов из разных таблиц измерений.

Разберем на примере. В розничной организации могут быть таблицы фактов, связанные с покупками клиентов, телефонными звонками по их обслуживанию и возвратами. В качестве таблиц измерений могут быть таблицы, характеризующие товары, или расширенная информация о клиентах.

Информация в таблице фактов — обычно числовые данные. Часто данными можно легко манипулировать, в частности, суммировать тысячи строк. Например, розничной организации может понадобиться отчет о прибыли для определенного магазина, линейки продуктов или сегмента клиентов. В этом случае из таблицы фактов можно извлечь информацию, которая относится к этим транзакциям и соответствует критериям, а затем суммировать информацию.

Таблица фактов

Таблица фактов — основная таблица хранилища данных. Как правило, содержит сведения об объектах или событиях, совокупность которых будет анализироваться.

Выделяют четыре часто встречающихся типа фактов:

1. **Факты, связанные с транзакциями** (Transaction facts) основаны на отдельных событиях. Типичный пример — телефонный звонок или снятие денег со счета с помощью банкомата.

2. **Факты, связанные с «моментальными снимками»** (Snapshot facts) основаны на состоянии объекта (например, банковского счета) в определенный момент, например на конец дня или месяца. Типичный пример — дневная выручка.

3. **Факты, связанные с элементами документа** (Line-item facts) основаны на документе (например, счете за товар или услуги) и содержат подробную информацию о его элементах: например, о количестве, цене, проценте скидки.

4. **Факты, связанные с событиями или состоянием объекта** (Event or state facts) представляют возникшие события без подробностей о них: например, просто факт продажи или факт ее отсутствия без дополнительных объяснений.

Таблица фактов содержит объединенный ключ, который является объединенным первичным ключом всех таблиц измерений. Объединенный ключ должен однозначно идентифицировать строку в таблице фактов.

При разработке таблицы фактов внимание нужно уделить зернистости таблицы — уровню детализации.

Если мы будем разрабатывать таблицу фактов покупки для розничной организации, нужно будет решить, что является зерном таблицы — транзакция клиента или покупка отдельного товара. Если зерно — покупка товара, каждая транзакция клиента будет генерировать несколько записей в таблицу фактов, соответствующих каждому купленному товару. Выбор зерна — фундаментальное решение в процессе проектирования.

Атрибуты таблицы фактов могут быть аддитивными или полуаддитивными.

- **Полностью аддитивные меры** — те, которые можно легко суммировать для всех измерений в таблице фактов. Например, количество заказов — это атрибут, который можно суммировать для всех измерений. Мы можем вывести общее количество по порядку для конкретного клиента, региона, даты, бренда и так далее.
- **Полуаддитивные меры** — те, которые можно суммировать только по некоторым измерениям таблицы фактов. Например, сумма баланса не может быть суммирована по времени, потому что она изменяется со временем.

Иногда в таблице фактов могут быть записи, имеющие атрибуты с нулевыми мерами. Например, не может быть заказа в праздничный день — у атрибутов для этой даты будут нулевые показатели. Нам не нужно хранить измерения для такого рода записей, поскольку они не предоставляют информации.

Иногда вы можете встретить измерения в таблице фактов, которые вообще не аддитивны. Например, идентификационный номер клиента нельзя суммировать ни по каким показателям. Однако, если нам нужно найти заказ, сделанный конкретным клиентом в этом месяце, этот номер нам понадобится. Такие типы атрибутов или измерения таблицы фактов называются **вырожденным измерением**.

Таблица измерений

Таблицы измерений содержат неизменяемые либо редко изменяемые данные. В большинстве случаев эти данные представляют собой по одной записи для каждого члена нижнего уровня иерархии в измерении.

Таблицы измерений также содержат как минимум одно описательное поле (обычно с именем члена измерения) и, как правило, целочисленное ключевое поле (первичный ключ) для однозначной идентификации члена измерения.

Если будущее измерение, основанное на таблице измерений, содержит иерархию, то таблица измерений также может содержать поля, указывающие на «родителя» этого члена в иерархии.

Нередко таблица измерений может содержать поля, указывающие на «прародителей» и других «предков» в иерархии (это характерно для сбалансированных иерархий), а также дополнительные атрибуты членов измерений, содержавшихся в исходной оперативной БД (например, адреса и телефоны клиентов).

Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов.

Скорость роста таблиц измерений должна быть незначительной по сравнению со скоростью роста таблицы фактов. Например, в таблицу измерений новая запись о товаре добавляется только в том случае, если это новый товар, который не продавался раньше. А в таблице фактов новая запись появляется при каждой транзакции с этим товаром. Вернемся к примеру с розничной торговлей. Покупки, возвраты и звонки — это факты. А клиенты, сотрудники, предметы и магазины — это измерения, они должны содержаться в таблицах измерений.

Таблицы измерений содержат сведения о каждом экземпляре объекта. Например, таблица измерения элементов будет содержать запись для каждого предмета, проданного в магазине. Она может содержать информацию о стоимость товара, поставщике, цвете, размерах и так далее.

Таблицы фактов и таблицы измерений связаны друг с другом. Разберем на примере розничной модели: в таблице измерений есть информация о товаре и его свойствах, а в таблице фактов — информация о покупке этого товара. Записи в таблицах связаны ключами — первичным (в таблице измерений) и внешним (в таблице фактов). Благодаря ним можно переходить из одной таблицы в другую.

Типы размеров (измерений)

Типы измерения	**Определение**
Согласованные (соответствующие, базовые) размеры	Характеристики объекта. Например, в таблице фактов есть яблоки. Согласованные размеры — сорт, вес, цвет и другие характеристики, которые относятся к этому факту

Размеры аутригера	Измерение может иметь ссылку на другую таблицу измерений. Эти вторичные размеры называются размерами аутригеров. Их следует использовать осторожно
Уменьшенные измерения	Подразделением строк и столбцов базового измерения. Полезны для разработки агрегированных таблиц фактов
Соединения таблицы размеров	Размеры могут иметь ссылки на другие размеры. Эти отношения могут быть смоделированы с размерами аутригера
Ролевые измерения	Одно ролевое измерение помогает ссылаться несколько раз в таблице фактов, поскольку каждая ссылка связана с логически отдельной ролью измерения
Размеры барахла	Коллекция случайных транзакционных кодов, флагов или текстовых атрибутов. Не может логически принадлежать какому-либо определенному измерению
Вырожденные размеры	Измерение без соответствующего измерения. Используется в транзакциях, для сбора таблиц фактов, а также их моментальных снимков. У этого вида измерения нет своего размера, поскольку он получен из таблицы фактов
Заменяемые размеры	Используются, когда одна и та же таблица фактов связана с разными версиями одного и того же измерения
Шаг измерения (dimension step)	Последовательные процессы, такие как события веб-страниц. В большинстве случаев имеют отдельную строку в таблице фактов для каждого шага процесса. Сообщают, где конкретный шаг следует использовать в общем сеансе

	Измерение не будет повторятся для разных сущностей в таблице фактов. Вернемся к примеру с яблоками: красный цвет будет повторятся для разных яблок, а какое-то действие может быть уникальным, относящимся только к конкретному факту

Различия между таблицами фактов и таблицами измерений.

Параметры	**Таблица фактов**	**Таблица измерений**
Определение	Таблица, в которой хранятся измерения, метрики или факты о бизнес-процессе	Таблица для сопоставления с таблицей фактов. Содержит описательные атрибуты, которые будут использоваться в качестве ограничения запроса
Характеристика	Расположена в центре схемы «Звезда» или «Снежинка», окружена размерами	Подключена к таблице фактов и расположена по краям схемы «Звезда» или «Снежинка»
Дизайн	Определяется зерном или самым атомарным уровнем	Определяется объемом описательных элементов. Проектируется с точки зрения гарантии качества размеров
Задача	Таблица фактов измеряет событие, для которого собираются данные. Описывает конкретное событие, которое произошло с товаром или другим объектом	Таблица измерений собирает и хранит справочную информацию об объекте
Тип данных	Таблицы фактов, как правила содержат информацию о параметрах «Продукт» и «Дата»	Таблица измерения содержит атрибуты, которые описывают детали измерения. Например,

		идентификатор продукта, его категорию и так далее
Ключ	В таблице фактов первичный ключ отображается как внешний ключ для измерений	Таблица измерений имеет конкретный первичный ключ, которые однозначно определяет каждое измерение
Хранимые данные	Помогает хранить метки отчетов и фильтровать значения доменов в таблицах измерений	Внесите подробные данные об объекте, его размере, структуре и характеристиках
Иерархия	Не содержит иерархии	Может содержать или не содержать иерархию. Например, «Местоположение» может содержать страну, область, город и так далее

Ключевая разница

- Таблица фактов содержит измерения, метрики и факты о бизнес-процессе. Таблица измерений — это дополнение к таблице фактов, она содержит описательные атрибуты, которые будут использоваться в качестве ограничения запроса.
- Таблица фактов расположена в центре схемы «Звезда» или «Снежинка», а таблица измерений — по краям.
- Таблица фактов определяется наиболее атомарным уровнем (зерном). Таблица измерений должна быть многословной, описательной, полной и с гарантированным качеством.
- Таблица фактов помогает хранить метки отчетов. Таблица измерений содержит подробные данные.
- Таблица фактов не содержит иерархии, а таблица измерений содержит.

Заключение

Сегодня мы узнали, что такое бизнес-аналитика и для чего она нужна. Поговорили о том, какие этапы нужно пройти для качественной аналитики и какие цели должен ставить перед собой анализ данных. Так же мы рассмотрели таблицы фактов и измерений, узнали их особенности и различия.

Что можно почитать еще?

1. «Бизнес-процессы. Языки моделирования, методы, инструменты», Франк Шёнталер, Готфрид Фоссен, Андреас Обервайс, Томас Карле.
2. «Путь аналитика. Практическое руководство IT-специалиста», Андрей Перерва, Вера Иванова.