**1. Что такое хранилище данных, какие требования к нему предъявляются? В чем отличие ХД от обычной базы данных.**

Очевидно, что подход к поиску знаний «от данных» предъявляет к последним ряд требований, нарушение которых может привести к получению недостоверных знаний и принятию ошибочных управленческих решений. Основными из них являются хронологичность, целостность, полнота, непротиворечивость, актуальность, отсутствие дубликатов, противоречий и аномалий в данных, а также их целостность. Кроме этого, большую роль играет скорость доступа к данным для применения к ним различных методов аналитической обработки.

Обычные СУБД не гарантируют выполнения перечисленных требований, и поэтому их использование для консолидации данных с целью анализа, может потребовать сложных процедур предобработки. Именно для поддержки процесса ИАД (интеллектуальный анализ данных) и поиска знаний в больших массивах данных получила развитие новая концепция консолидации данных – хранилища данных (Data Warehouse)-дейта вейхаус.

***Хранилище данных*** (ХД), как определил автор концепции Б. Инмон, представляет собой «предметно-ориентированный, интегрированный, неизменчивый и поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений».

**2. Каковы причины появления и развития концепции ХД.**

Концепция ХД появилась и начала развиваться не «на пустом месте», а явилась результатом эволюционного процесса поиска конфигураций для систем хранения данных, оптимизированных сточки зрения аналитической обработки в интересах поддержки принятия решений

**3. Дайте понятие метаданных и семантического слоя в ХД.**

Кроме собственно данных, описывающих исследуемые процессы и явления, важнейшим элементом ХД являются ***метаданные*** (от [лат.](https://ru.wikipedia.org/wiki/%D0%9B%D0%B0%D1%82%D0%B8%D0%BD%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *meta* — цель, конечный пункт, предел, край). Метаданные (данные о данных) – это служебные данные, описывающие структуру хранилища, содержащие информацию о принадлежности данных к тому или иному типу (измерение, атрибут или факт). С помощью метаданных формируется *семантический слой*, который обеспечивает визуальное представление хранилища и визуальные средства манипуляции с данными и метаданными. Метаданные в ХД можно разделить на два вида:

- ***технические*** – обеспечивают работу собственно ХД;

- ***бизнес-метаданные*** – описывают структуру данных в рамках заданной бизнес-модели.

**4. Перечислите основные элементы логической структуры ХД, поясните их смысл, приведите примеры.**

В настоящее время существует несколько моделей представления данных в ХД, основными из которых являются многомерная и реляционная

В основе многомерных ХД лежит идея разделения данных на два вида – ***измерения*** и ***факты***. Факты – это данные, количественно описывающие бизнес-процессы (суммы, количества, цены и т.д.), а измерения – качественные данные, формирующие контекст для фактов. Действительно, любые количественные характеристики бессмысленны без контекста, который представляет собой объекты, связанные с этими характеристиками. Например, для количества необходимо указать объекты, для которых оно определяется; для цены – какие-то товары и услуги, и т.д. Раздельное использование измерений и фактов не имеет смысла, поскольку, выражаясь языком статистики, при этом нарушается единство количественной и качественной сторон описываемых явлений и процессов.

Рассмотрим более подробно многомерную модель данных, её свойства и основные элементы.

В ХД могут использоваться две разновидности многомерных моделей данных – «звезда» и «снежинка». В основе обеих моделей лежат следующие понятия:

- ***измерение*** – последовательность значений определенного признака. Например, измерение «Клиент» будет содержать последовательность наименований клиентов, измерение «Товар» - перечень товаров и т.д.;

- ***атрибут*** – свойство измерения, которое может быть полезно для его анализа. Например, для товара это может быть цвет, вес, форма и т.д.

- ***факт*** – значение, определяющее количественное описание объекта, события или явления. Факты всегда связаны с определенным измерением. Например, для измерения «Товар» могут быть указаны следующие факты – цена, количество и сумма, для измерения «Клиент» - возраст, доход, стаж работы и т.д.

- ***процесс*** – совокупность измерений и фактов, связанных с определенным объектом, процессом или явлением. Процессов в ХД может быть несколько и каждый из них образуется отдельной «снежинкой» или «звездой».

**5. Дайте понятие измерения как элемента структуры данных ХД. Приведите примеры измерений. В чём отличие измерений от фактов.**

***Измерение*** (таблица измерений - dimension table) — таблица в структуре многомерной модели данных, которая содержит значения признаков объектов (событий, явлений), сохраненных в таблице фактов. Признаки представляют собой текстовые или иные описания, логически объединенные в одно целое. Например, в таблице измерения «Товары» будут содержаться наименования товаров, а в таблице «Клиенты» - названия организаций или ФИО частных лиц. Очевидно, что кроме качественных характеристик, объекты бизнес-процессов имеют и количественные, например, цена товара и сумма, на которую он был реализован. Для банка может быть важен возраст клиента и его доход. Такие количественные значения ***содержатся в таблице фактов***. Таким образом, в многомерной модели данных используется два типа таблиц: таблица измерений, которая содержит качественные характеристики объектов, процессов и явлений, и таблица фактов, содержащая количественные значения.

**6. Дайте понятие процесса в ХД. Приведите пример.**

***процесс*** – совокупность измерений и фактов, связанных с определенным объектом, процессом или явлением. Процессов в ХД может быть несколько и каждый из них образуется отдельной «снежинкой» или «звездой».

**7. Логическая структура данных на основе схемы «звезда».**

Схема «звезда» включает в себя единственную таблицу фактов и множество таблиц соединенных с ней измерений. Каждая таблица измерений должна находиться в отношении «один к одному» с таблицей фактов.

**8. Логическая структура данных на основе схемы «снежинка».**

Схема «снежинка» отличается от схемы «звезда» тем, что содержит хотя бы одну таблицу измерений, ссылающуюся на таблицу, которая находится по отношению к ней в связи «один ко многим». Иными словами, измерения в «снежинке», могут иметь иерархически подчиненные измерения. Например, измерение «Регион» может быть связано с измерением «Город», т.е. для каждого региона будет содержаться список расположенных в нем городов. В ХД, где скорость доступа к данным является наиболее критическим факторам, оттают предпочтение схеме снежинка.

Таблица фактов является основной таблицей процесса. Она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться. Обычно такая таблица содержит уникальный составной ключ, объединяющий первичные ключи таблиц измерений. Помимо этого, таблица фактов содержит одно или несколько числовых полей, на основании которых в процессе выполнения аналитических запросов вычисляются агрегатные данные.

**9. Что такое агрегирование (агрегация) данных? Приведите пример.**

***Агрегатными (агрегированными)*** называются значения, вычисленные на основе совокупности некоторого набора исходных значений. Например, суммы продаж за неделю (месяц, год) есть результат агрегирования ежедневных продаж с помощью функции суммирования. Агрегирование позволяет управлять детальностью представления явлений в анализируемых данных, что позволяет снизить количество обрабатываемых значений, избавиться от случайных флуктуаций.

Агрегатные данные могут формироваться и содержаться в ХД, что при большом количестве автоматически вычисляемых агрегатов может привести к информационному взрыву хранилища. Другой вариант – вычислять агрегаты «на лету», в процессе выполнения запроса к ХД, что может увеличить время ожидания. Все загружаемые в ХД данные обязательно должны быть определены как измерение, атрибут, либо факт

**10. Опишите последовательность действий при загрузке данных в ХД Destructor Wireless.**

Поэтому, в процессе перемещения данных из источников в ХД предусмотрена специальная процедура, называемая ETL (Extract, Transform, Load) – извлечение, преобразование, загрузка.

***Извлечение*** – процесс получения доступа к данным в источнике, их выгрузка и помещение в специальную область памяти, называемую ***промежуточной областью*** (staged area). Использование промежуточной области необходимо еще и потому, что различные источники, данные из которых должны быть интегрированы, могут располагаться как на локальных носителях, так и на сетевых, что обуславливает различное время их получения. Следовательно, данные, полученные раньше, должны где-то ожидать данные, которые поступят позже. Таким местом и является промежуточная область. Данные в промежуточной области хранятся в виде временных таблиц, которые удаляются после завершения работы с данными. Выгрузка данных из источников может производиться как средствами самого ХД, так и средствами ИС в которой они содержаться (например, системы OLTP, АСУП и т.д.).

***Преобразование*** – производится в промежуточной области и включает в себя преобразование данных различных типов и форматов к единому формату хранилища. Кроме этого, на данном этапе производится первичная предобработка данных с целью их дедупликации (исключения повторяющихся элементов данных), устранения противоречий, поиска ошибок, разбора составных значений (например, адресов) и других операций, определенных разработчиком системы. Эти операции должны обеспечить представление данных в ХД, которое наилучшем образом соответствует бизнес-модели организации и способствовать наиболее быстрому выполнение аналитических запросов к хранилищу. К таким операциям относится агрегирование числовых значений, их округление и т.д.

***Загрузка.*** Заключается в переносе данных из промежуточных таблиц в структуры ХД. При загрузке в хранилище переносится не вся информация из источников, а только та, которая была изменена в течение некоторого промежутка времени, прошедшего с предыдущей загрузки. При этом, как правило, формируется два потока данных: поток добавления и поток обновления. В первом потоке передаются новые, ранее не существовавшие данные, а во втором – данные, которые существовали и ранее, но были изменены со времени предыдущей загрузки. Для организации такого процесса должны быть предусмотрены специальные средства регистрации изменений в данных.

Как правило, в ИС предприятий загрузка данных в корпоративные ХД производится не произвольно, а в соответствии с определенным регламентом. В частности, периодичность загрузки должна соответствовать времени, в течение которого в источниках данных накапливается значительное количество новых данных. Выбор времени загрузки хранилища также имеет значение – оно должно выбираться так, чтобы не перегружать ЛВС организации (например, в обеденный перерыв или ночью).