

# Проектирование Data Warehouse (DWH) - основы

## Ещё раз про слои

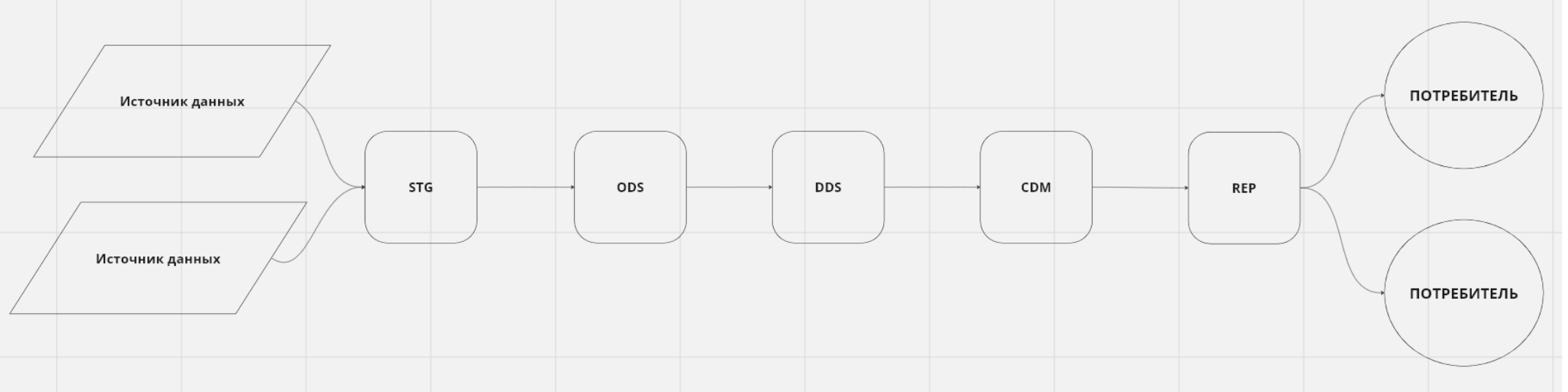
В мире обработки и анализа больших данных, разработка хранилища данных (DWH) часто включает в себя создание многоуровневой, или многослойной, архитектуры. Эта структура, как и сложная механика часов, состоит из ряда взаимосвязанных компонентов, каждый из которых выполняет свою уникальную функцию. Для понимания и успешного проектирования DWH крайне важно разобраться в роли и назначении каждого из этих слоев.

У этого есть преимущества:

- 1. Специализация: Как каждый сотрудник специализируется на определенной задаче, так и каждый слой DWH оптимизирован для выполнения своих специфических функций.
- 2. Изоляция ошибок: Ошибки в одном слое не распространяются на другие, что повышает надежность всей системы.
- 3. Доступность для разных пользователей: Различные слои могут быть доступны для разных групп пользователей в зависимости от их квалификации и потребностей.

В корпоративных хранилищах данных (DWH) слои или уровни представляют собой различные компоненты с уникальной архитектурой, такие как базы данных, системы управления базами данных (СУБД) или файловые системы. Эти слои часто создаются с использованием разных технологий. Обычно в DWH выделяют пять ключевых слоев:

- 1. STG (Staging): Слой для консолидации данных из различных источников.
- 2. ODS (Operational Data Store): Хранилище для операционных данных.
- 3. DDS (Detail Data Store): Хранилище для детальных исторических данных.
- 4. CDM (Common Data Marts): Слой, содержащий широкие витрины данных.
- 5. REP (Reporting): Слой для детальных витрин данных или отчетности.



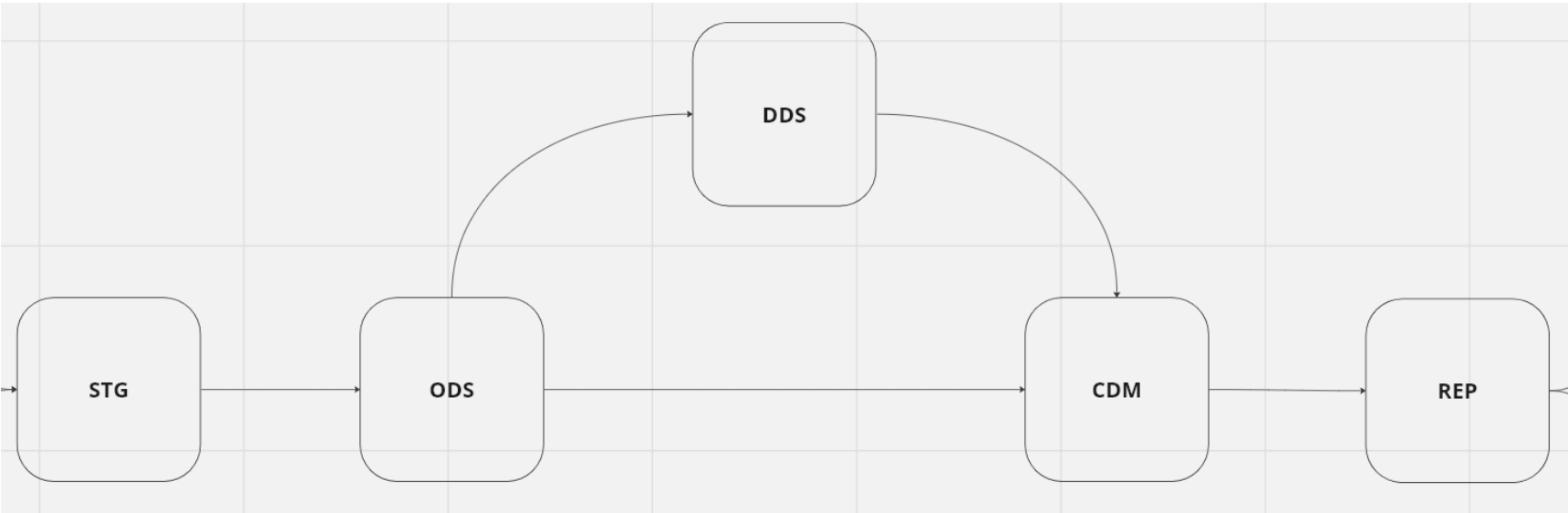
Не все хранилища включают все эти слои. Основной минимальный набор обычно состоит из STG, DDS и CDM. В зависимости от сложности и объема данных в системе, разные слои могут использовать различные типы СУБД. В более простых системах с меньшим объемом данных можно использовать классические реляционные базы данных, например, PostgreSQL, который часто применяется для создания DWH.

**Функции слоёв**

Staging, или промежуточное хранение, это начальный уровень в архитектуре хранилища данных, куда поступает первичная информация из разнообразных источников. Этот слой часто упоминается под разными названиями, такими как STG, staging area или RAW, поскольку он содержит данные в их исходном, необработанном виде. Важно отметить, что этот слой не предназначен для построения отчетов или витрин данных, так как он может быть огромным и неоптимизированным для таких задач.

- Сбор данных: Принимает сырые данные из различных источников.
- Поддержка as is: Хранит данные в их оригинальном формате без изменений.
- Нагрузка: Снимает нагрузку с источников путем переноса данных.
- Архивация: Служит для архивации и устранения ошибок в данных.

Operational Data Store (ODS) хранит операционные, часто текущие детальные данные, которые еще не агрегированы. Этот слой служит источником оперативной информации для создания отчетов и может напрямую использоваться для построения более высокоуровневых слоев. На основе ODS может сразу строится CDM:



В ODS существуют два основных подхода к управлению данными:

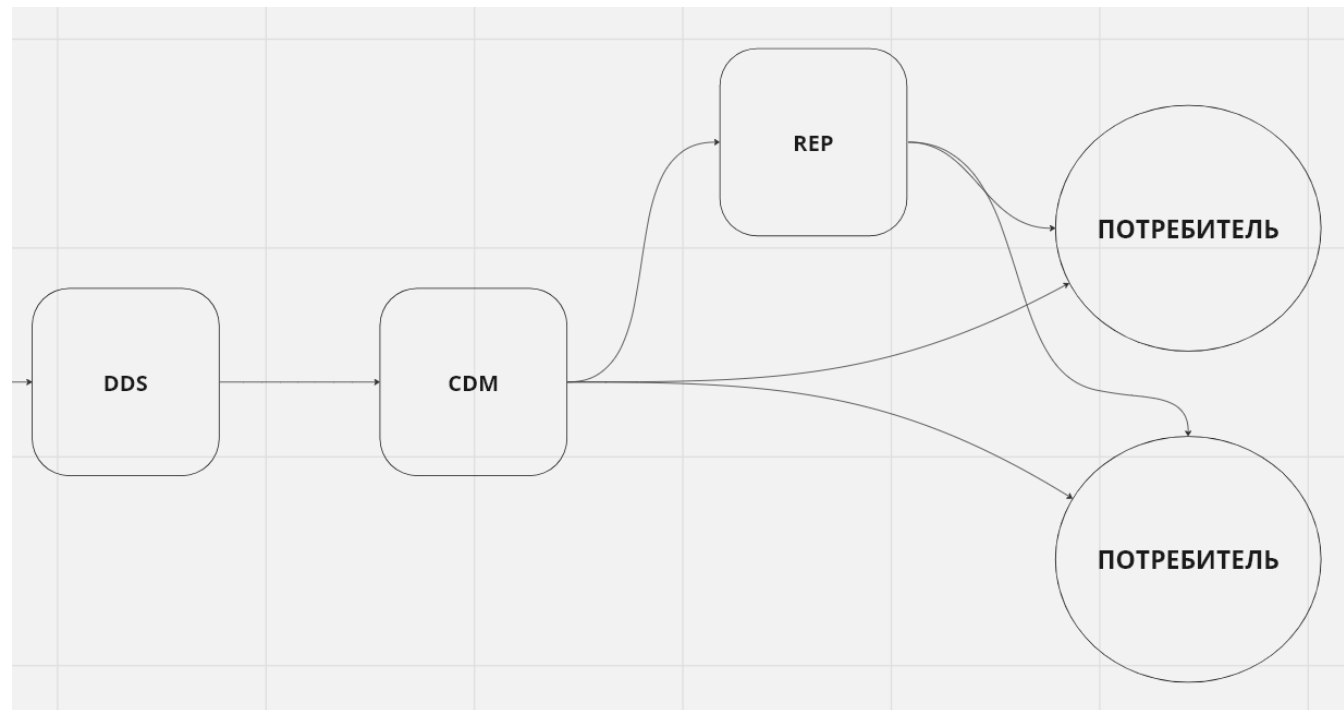
1. Текущие данные: ODS может хранить исключительно актуальные данные без учета их истории. В таком случае ODS выступает в роли промежуточного этапа в цепочке ETL, где происходит нормализация и очистка данных, или как источник сведений для CDM.
  2. Интегрированные данные: В другом подходе данные в ODS полностью интегрированы и могут быть связаны друг с другом, иногда даже с соблюдением третьей нормальной формы (3НФ), при этом сохраняется полная история изменений данных.
- Интеграция: Объединяет данные из разных источников в реляционную модель.
  - Очистка: Удаляет или обновляет неверные данные.
  - Актуальность: Предоставляет текущие операционные данные для оперативных отчетов.

Detail Data Store (DDS) является ядром DWH и содержит полную историю данных. Этот слой может быть представлен в разной степени нормализации и служит для проведения исторического анализа и выявления тенденций.

- Историчность: Хранит полную историю изменений данных.
- Анализ: Позволяет проводить исторический анализ и выявлять тенденции.

Common Data Marts (CDM) представляет собой слой ключевых витрин данных, где информация структурирована и готова к использованию бизнес-аналитиками. Эти данные могут быть детализированы до уровня отдельных пользователей и часто используются для принятия бизнес-решений.

Пользователи могут напрямую забирать данные из Common Data Marts (CDM), однако это может быть не самым удобным способом. Обычно они предпочитают использовать слой отчетности (REP), который специально настроен для их нужд. Если в REP чего-то не хватает, тогда пользователи могут вернуться к CDM за более детальной информацией:



- Информативность: Обеспечивает доступ к детальным и агрегированным данным для анализа.
- Детализация: Предоставляет информацию с высокой степенью детализации.

Reporting Layer (REP) — это конечный продукт, где данные подготовлены специально для пользовательских отчетов. Это полностью готовые ДАННЫЕ, представленные в удобном для визуализации и анализа виде. Этот слой не всегда присутствует в архитектуре DWH, но когда он есть, он обеспечивает быстрый и наглядный доступ к информации для конечных пользователей.

- Отчеты: Генерирует данные для конкретных отчетов и аналитических запросов.
- Визуализация: Предоставляет данные в формате, удобном для визуализации в BI-инструментах.

В процессе выявления требований к хранилищу данных вы можете определить, какие слои будут необходимы для вашей задачи. Не всегда требуется создавать все пять слоев; минимально необходимый набор для функционирования DWH обычно включает в себя STG (промежуточное хранение сырых данных), DDS (детальное хранение исторических данных) и CDM (витрины данных для аналитики).