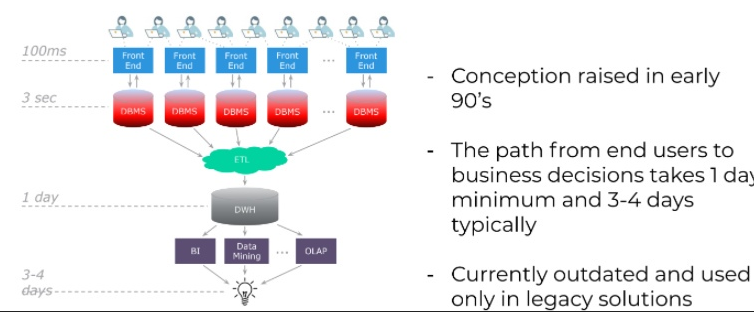
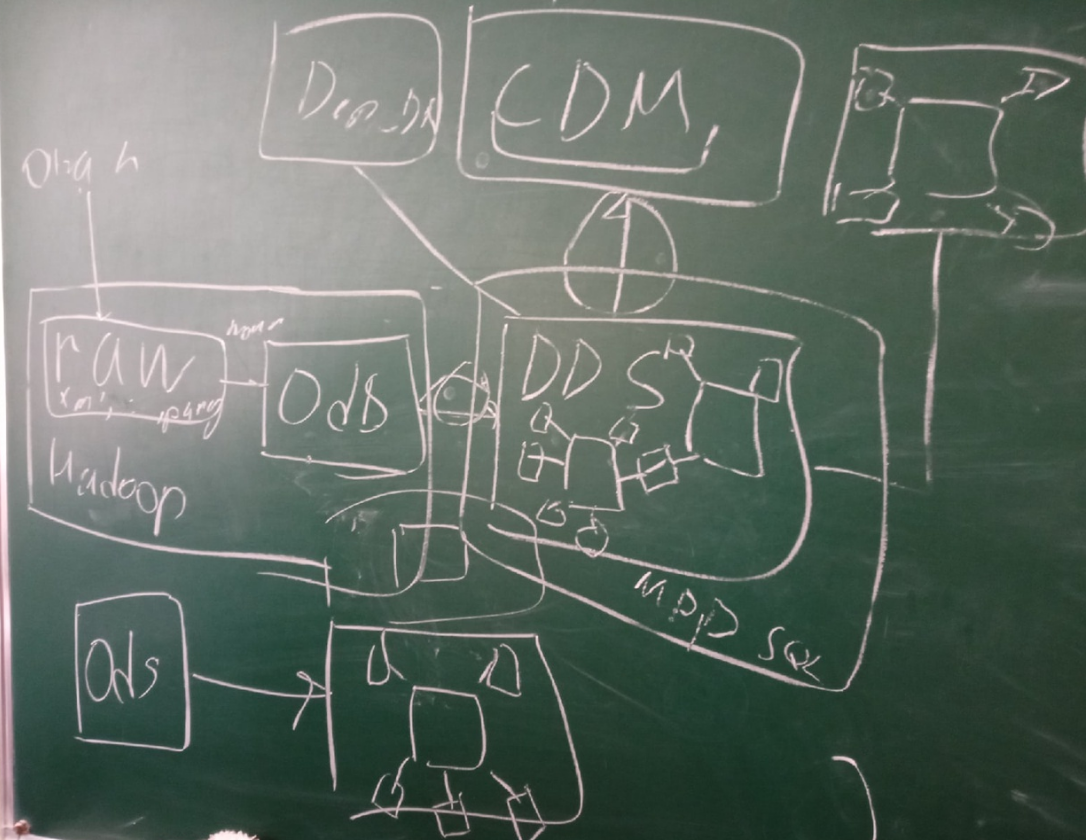
**DWH**

**Храни́лище да́нных** ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *Data Warehouse*) — предметно-ориентированная информационная [база данных](https://ru.wikipedia.org/wiki/%D0%91%D0%B0%D0%B7%D0%B0_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85), специально разработанная и предназначенная для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений в организации. Строится на базе [систем управления базами данных](https://ru.wikipedia.org/wiki/%D0%A1%D0%A3%D0%91%D0%94) и [систем поддержки принятия решений](https://ru.wikipedia.org/wiki/%D0%A1%D0%9F%D0%9F%D0%A0). Данные, поступающие в хранилище данных, как правило, доступны только для чтения.





Принципы организации хранилища

* *Проблемно-предметная ориентация*. Данные объединяются в категории и хранятся в соответствии с областями, которые они описывают, а не с приложениями, которые они используют.
* *Интегрированность*. Данные объединены так, чтобы они удовлетворяли всем требованиям предприятия в целом, а не единственной функции бизнеса.
* *Некорректируемость*. Данные в хранилище данных не создаются: то есть поступают из внешних источников, не корректируются и не удаляются.
* *Зависимость от времени*. Данные в хранилище точны и корректны только в том случае, когда они привязаны к некоторому промежутку или моменту времени.

## Дизайн хранилищ данных

Существуют два основных архитектурных направления — нормализованные хранилища данных и хранилища с измерениями.

**1.** В нормализованных хранилищах, данные находятся в предметно ориентированных таблицах [третьей нормальной формы](https://ru.wikipedia.org/wiki/%D0%A2%D1%80%D0%B5%D1%82%D1%8C%D1%8F_%D0%BD%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D1%84%D0%BE%D1%80%D0%BC%D0%B0). Нормализованные хранилища характеризуются как простые в создании и управлении, недостатки нормализованных хранилищ — большое количество таблиц как следствие нормализации, из-за чего для получения какой-либо информации нужно делать выборку из многих таблиц одновременно, что приводит к ухудшению производительности системы. Для решения этой проблемы используются денормализованные таблицы — [витрины данных](https://ru.wikipedia.org/wiki/%D0%92%D0%B8%D1%82%D1%80%D0%B8%D0%BD%D0%B0_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85), на основе которых уже выводятся отчетные формы. При громадных объемах данных могут использовать несколько уровней «витрин»/«хранилищ».

**2.** Хранилища с измерениями используют [схему «звезда»](https://ru.wikipedia.org/wiki/%D0%A1%D1%85%D0%B5%D0%BC%D0%B0_%D0%B7%D0%B2%D0%B5%D0%B7%D0%B4%D1%8B) или [схему «снежинка»](https://ru.wikipedia.org/wiki/%D0%A1%D1%85%D0%B5%D0%BC%D0%B0_%D1%81%D0%BD%D0%B5%D0%B6%D0%B8%D0%BD%D0%BA%D0%B8). При этом в центре «звезды» находятся данные ([таблица фактов](https://ru.wikipedia.org/wiki/%D0%A2%D0%B0%D0%B1%D0%BB%D0%B8%D1%86%D0%B0_%D1%84%D0%B0%D0%BA%D1%82%D0%BE%D0%B2)), а [измерения](https://ru.wikipedia.org/wiki/%D0%A2%D0%B0%D0%B1%D0%BB%D0%B8%D1%86%D0%B0_%D0%B8%D0%B7%D0%BC%D0%B5%D1%80%D0%B5%D0%BD%D0%B8%D0%B9) образуют лучи звезды. Различные таблицы фактов совместно используют таблицы измерений, что значительно облегчает операции объединения данных из нескольких предметных таблиц фактов (пример — факты продаж и поставок товара). Таблицы данных и соответствующие измерения образуют архитектуру «шина». Измерения часто создаются в третьей нормальной форме, в том числе, для протоколирования изменения в измерениях. Основным достоинством хранилищ с измерениями является простота и понятность для разработчиков и пользователей, также, благодаря более эффективному хранению данных и формализованным измерениям, облегчается и ускоряется доступ к данным, особенно при сложных анализах. Основным недостатком является более сложные процедуры подготовки и загрузки данных, а также управление и изменение измерений данных.

Так же альтернативные концепции:

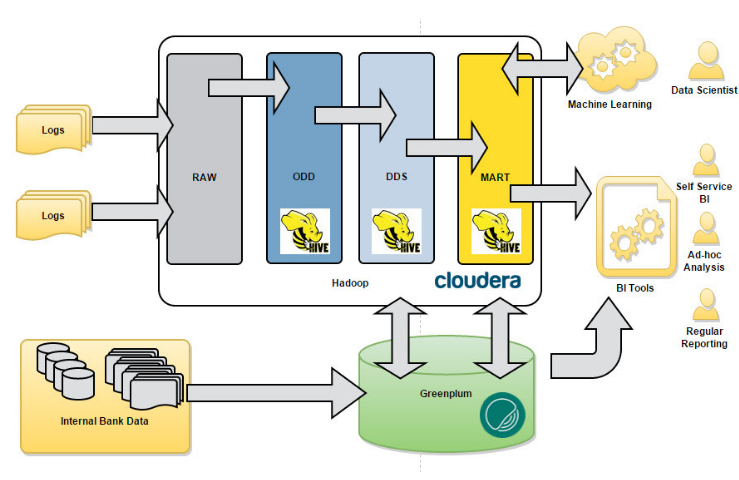
**3**. DataVault

**4**. Anchor model

## Data Lake

В отличие от подхода Data Warehouse, метод Data Lake представляет собой хранилище, где хранятся все необработанные данные в исходном формате без преобразования. Каждый элемент в Data Lake обозначен уникальным идентификатором и набором тегов метаданных. Данные могут быть неструктурированными, полуструктурированными или структурированными, они преобразуются только при запросе на использование. Из-за сложности подхода, Data Lake больше подходит для пользователей, занимающихся глубоким анализом данных и их исследованием.

По затратам средств метод Data Lake обходится намного дешевле, чем хранение данных в Data Warehouse. Также Data Lake быстрее выдает результат, так как у пользователей есть доступ к данным до их преобразования. Data Warehouse же потребует гораздо больше времени для внесения любых изменений в хранилище.



* RAW – слой сырых данных, сюда загружаем файлы, логи, архивы. Форматы могут быть абсолютно различные: tsv, csv, xml, syslog, json и т.д. и т.п.;
* ODD — Operational Data Definition. Сюда мы загружаем данные в формате приближенном к реляционному. Данные здесь могут являться результатом предобработки данных из RAW перед загрузкой в DDS;
* DDS — Detail Data Store. Здесь мы собираем консолидированную модель детальных данных.
* MART – витрины данных. Здесь мы собираем прикладные витрины данных.