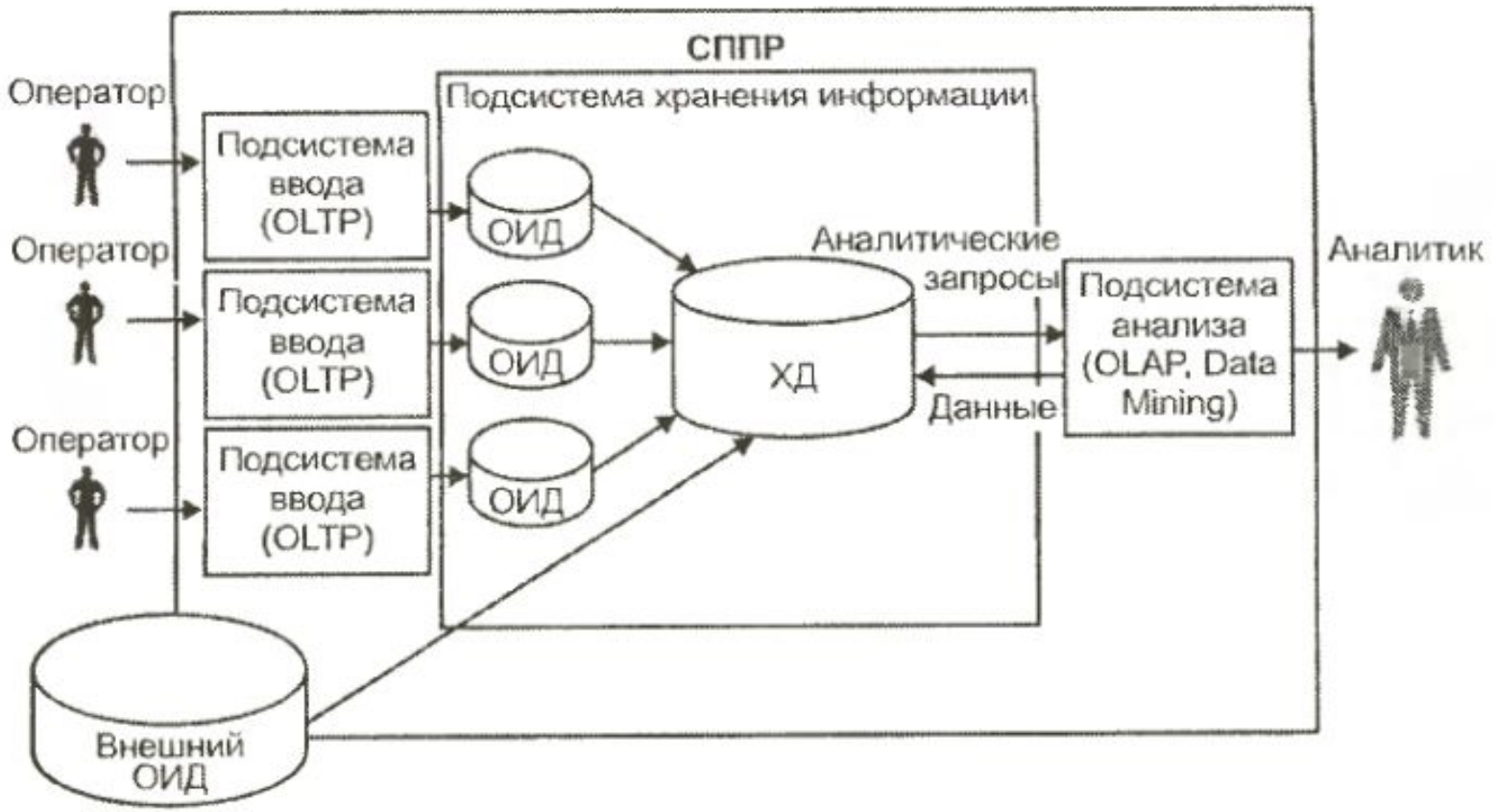
Концепция хранилищ данных

Хранилища данных

В основе концепции ХД лежит идея разделения данных, используемых для оперативной обработки и для решения задач анализа.

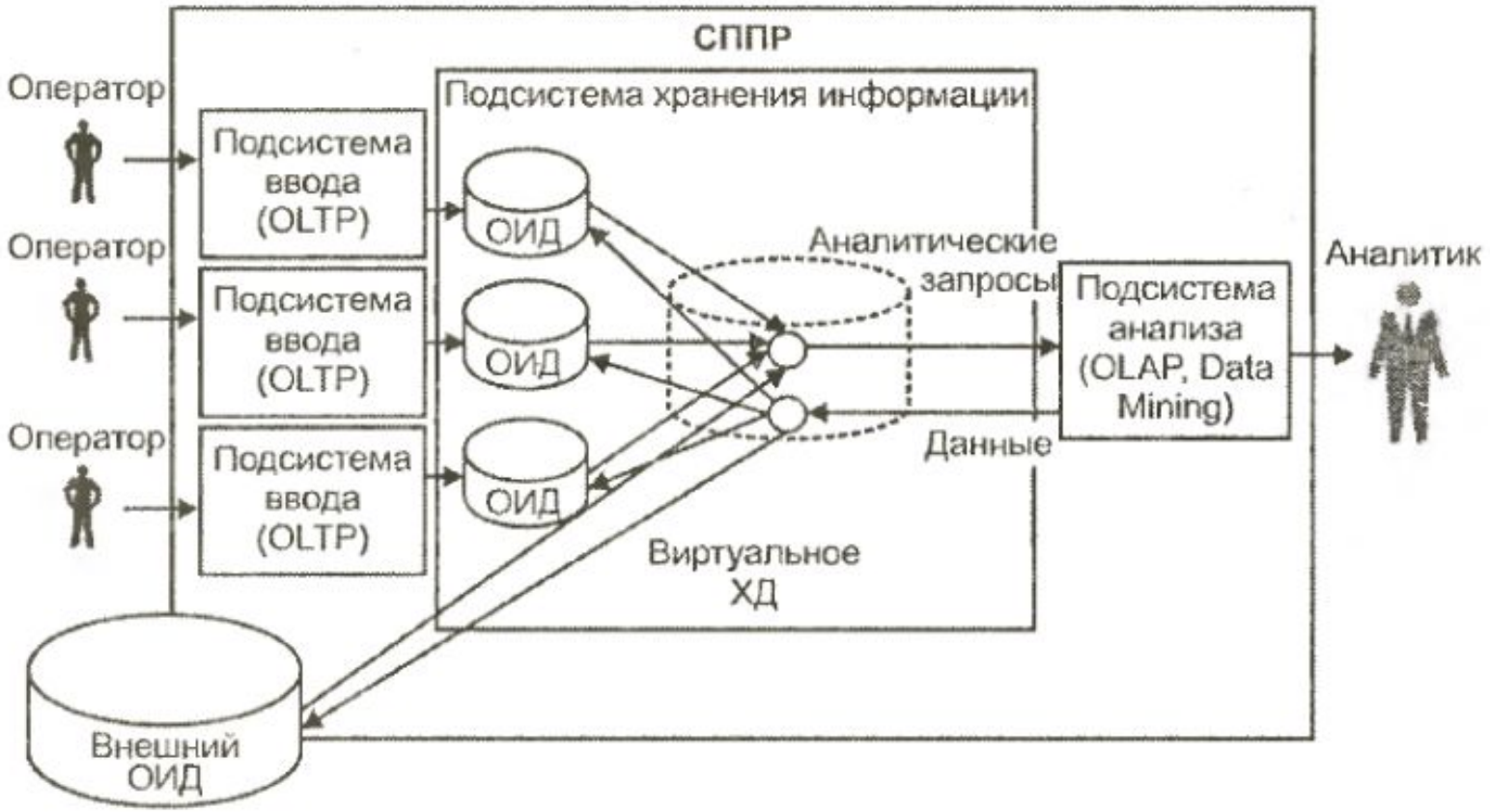
Хранилище данных - предметно ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений.

Структура СППР с физическим ХД



Структура СППР с физическим ХД

Структура СППР с виртуальным ХД

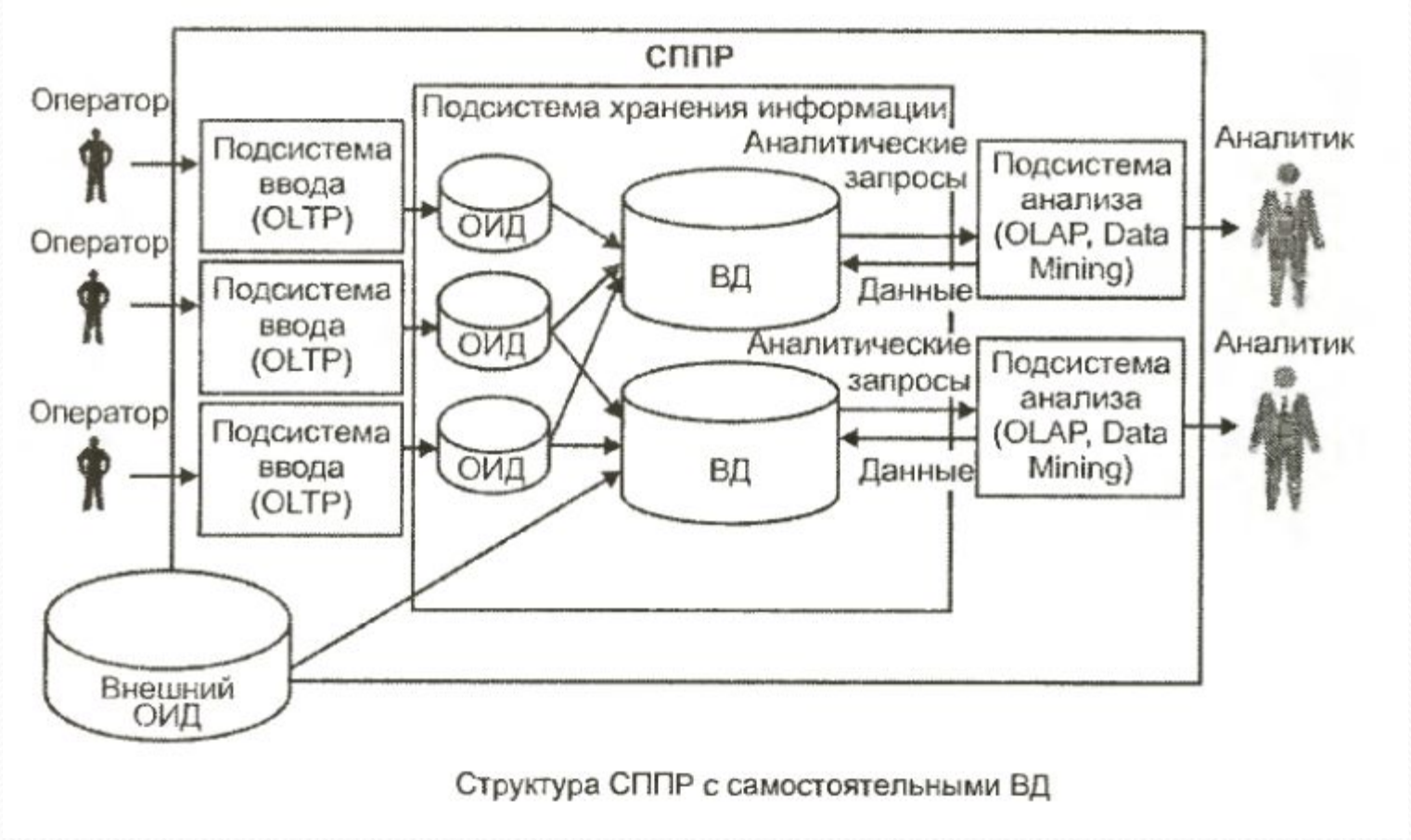


Структура СППР с виртуальным ХД

Проблемы создания физического ХД:

- необходимость интеграции данных из неоднородных источников в распределенной среде;
- потребность в эффективном хранении и обработке очень больших объемов информации;
- необходимость наличия многоуровневых справочников метаданных;
- повышенные требования к безопасности данных.

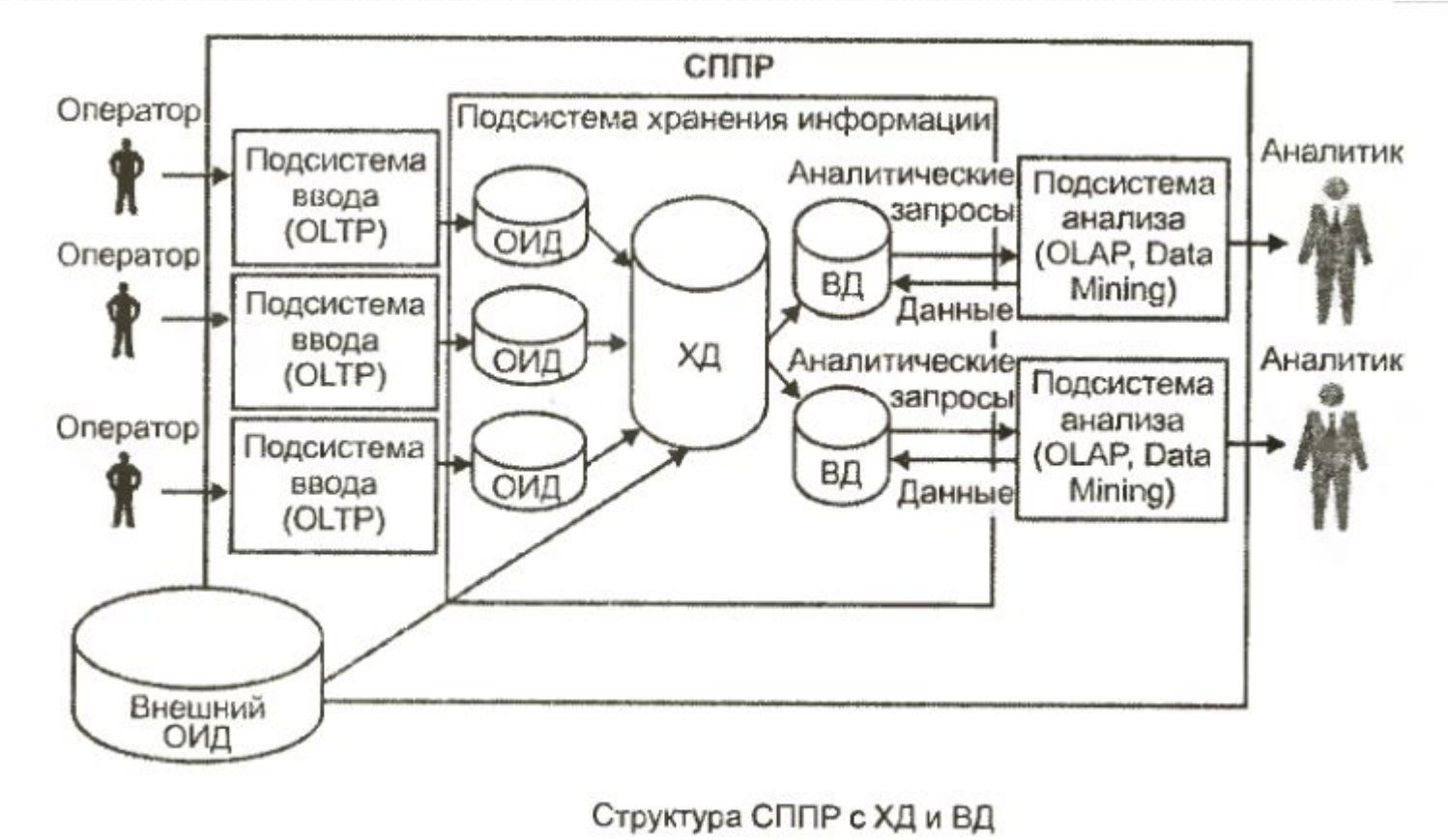
Структура СППР с самостоятельными ВД



Структура СППР с самостоятельными ВД

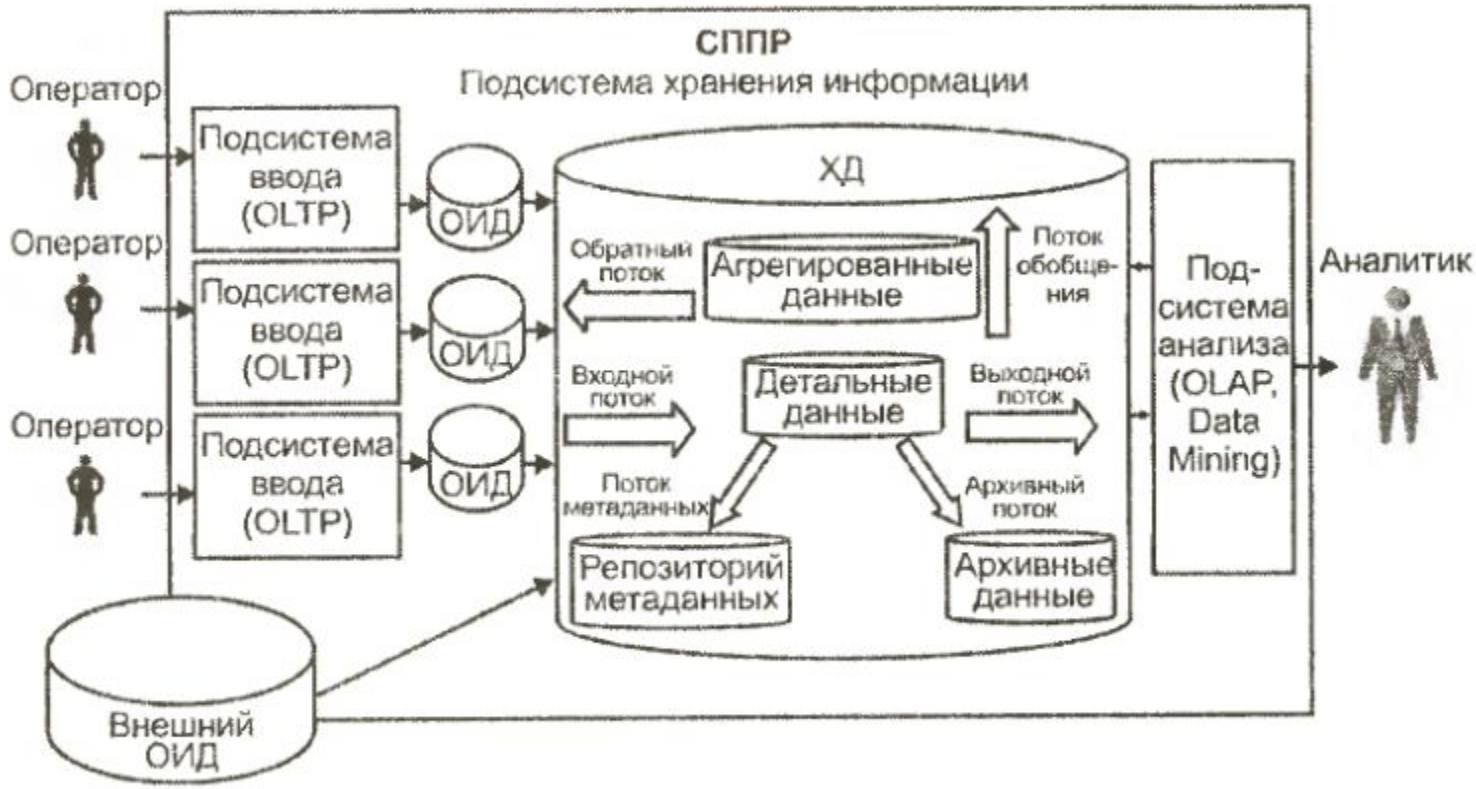
Витрина данных (ВД) - это упрощенный вариант ХД, содержащий только тематически объединенные данные.

Структура СППР с ХД и ВД



Структура СППР с ХД и ВД

Архитектура ХД



Архитектура ХД

Состав ХД

Детальными являются данные, переносимые непосредственно из ОИД. Они соответствуют элементарным событиям, фиксируемым OL TP системами. (Например, продажи, эксперименты и др.). Принято разделять все данные на измерения и факты.

Измерениями называются наборы данных, необходимые для описания событий (например, города, товары, люди и т. п.).

Фактами называются данные, отражающие сущность события (например, количество проданного товара, результаты экспериментов и т. п.).

На основании детальных данных могут быть получены **агрегированные** (обобщенные) данные.

Состав ХД

Для удобства работы с ХД необходима информация о содержащихся в нем данных. Такая информация называется метаданными (данные о данных).

Согласно концепции Дж. Захмана, метаданные должны отвечать на следующие вопросы

- что (описание объектов),
- кто (описание пользователей),
- где (описание места хранения),
- как (описание действий),
- когда (описание времени)
- и почему (описание причин).

Информационные потоки в ХД

- **Входной поток** (Inflow) образуется данными, копируемыми из ОИД в ХД;
- **поток обобщения** (Upflow) образуется агрегированием детальных данных и их сохранением в ХД;
- **архивный поток** (Downflow) образуется перемещением детальных данных, количество обращений к которым снизилось;
- **поток метаданных** (MetaFlow) образуется переносом информации о данных в репозиторий данных;
- **выходной поток** (Outflow) образуется данными, извлекаемыми пользователями;
- **обратный поток** (Feedback Flow) образуется очищенными данными, записываемыми обратно в ОИД.

Оптимизация ХД

Для улучшения производительности ХД используют следующие приемы:

- создание таблиц предварительно агрегированных данных;
- индексирование (чтобы избежать необходимости просматривать слишком большие объемы данных);
- хранение данных в отсортированном виде, устраняющем необходимость в процессе "and sort".
- "денормализация" модели - размещение данных в одной таблице, а не в нескольких, которые необходимо соединять.

Избыточность и денормализация

- **Нисходящая денормализация** – избыточные столбцы из родительской таблицы помещаются в дочернюю таблицу
- **Восходящая ~~денормализация~~ (избыточность)** – данные из дочерней таблицы помещаются в родительскую таблицу.
- **Внутритабличная денормализация** – внутри таблицы создаются избыточные столбцы.
- *(а также Вертикальное и Горизонтальное расщепление.)*

Вопросы

- Что такое хранилище данных?
- Что такое виртуальное и физическое хранилище данных?
- Что такое витрина данных?
- Из чего состоит хранилище данных?
- Какие потоки данных имеются в хранилище данных?
- Какие есть приемы оптимизации хранилищ данных?
- Какие типы денормализации вы знаете?