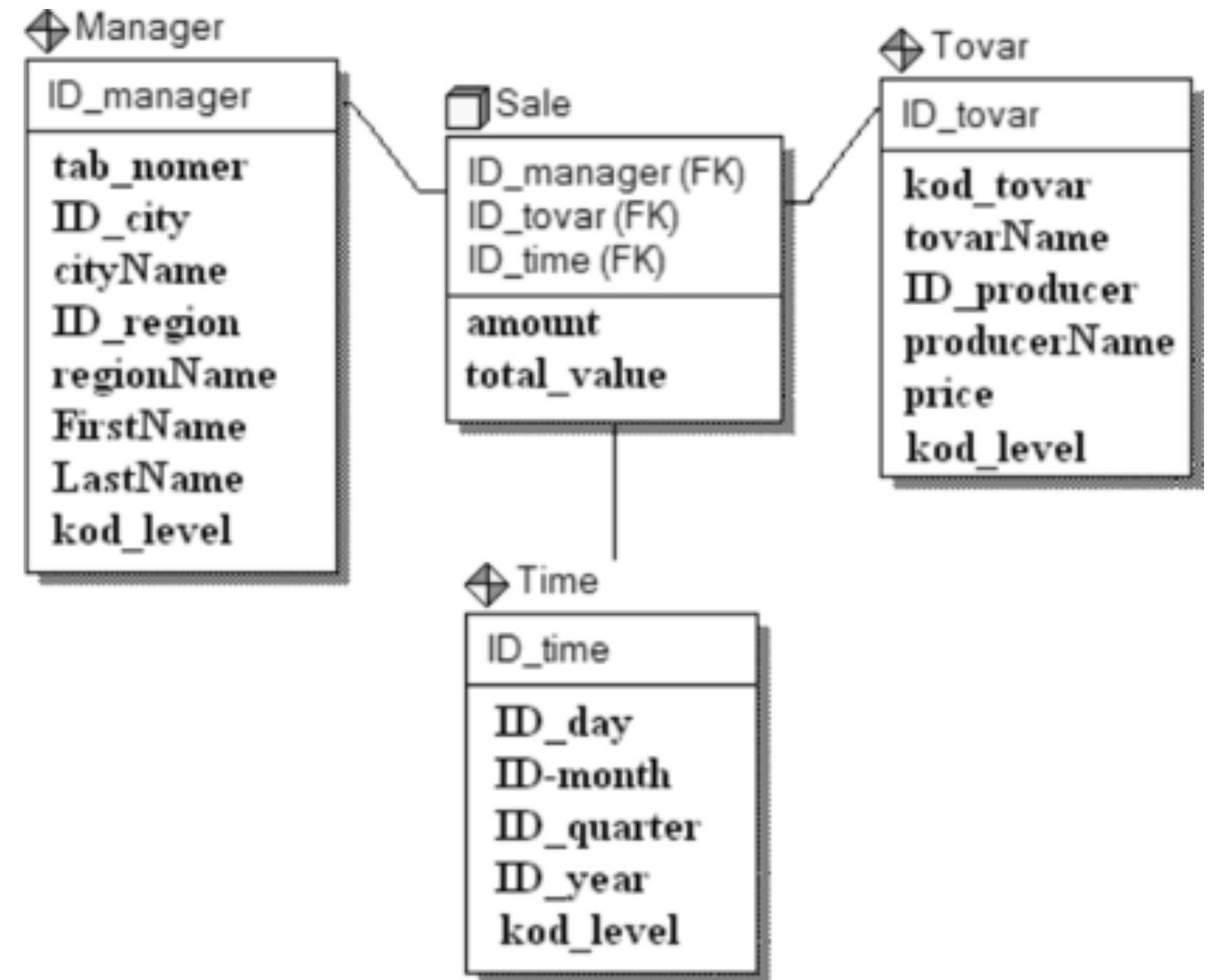


**Аналитические хранилища**  
**Пространственная модель**  
**ETL и ELT**

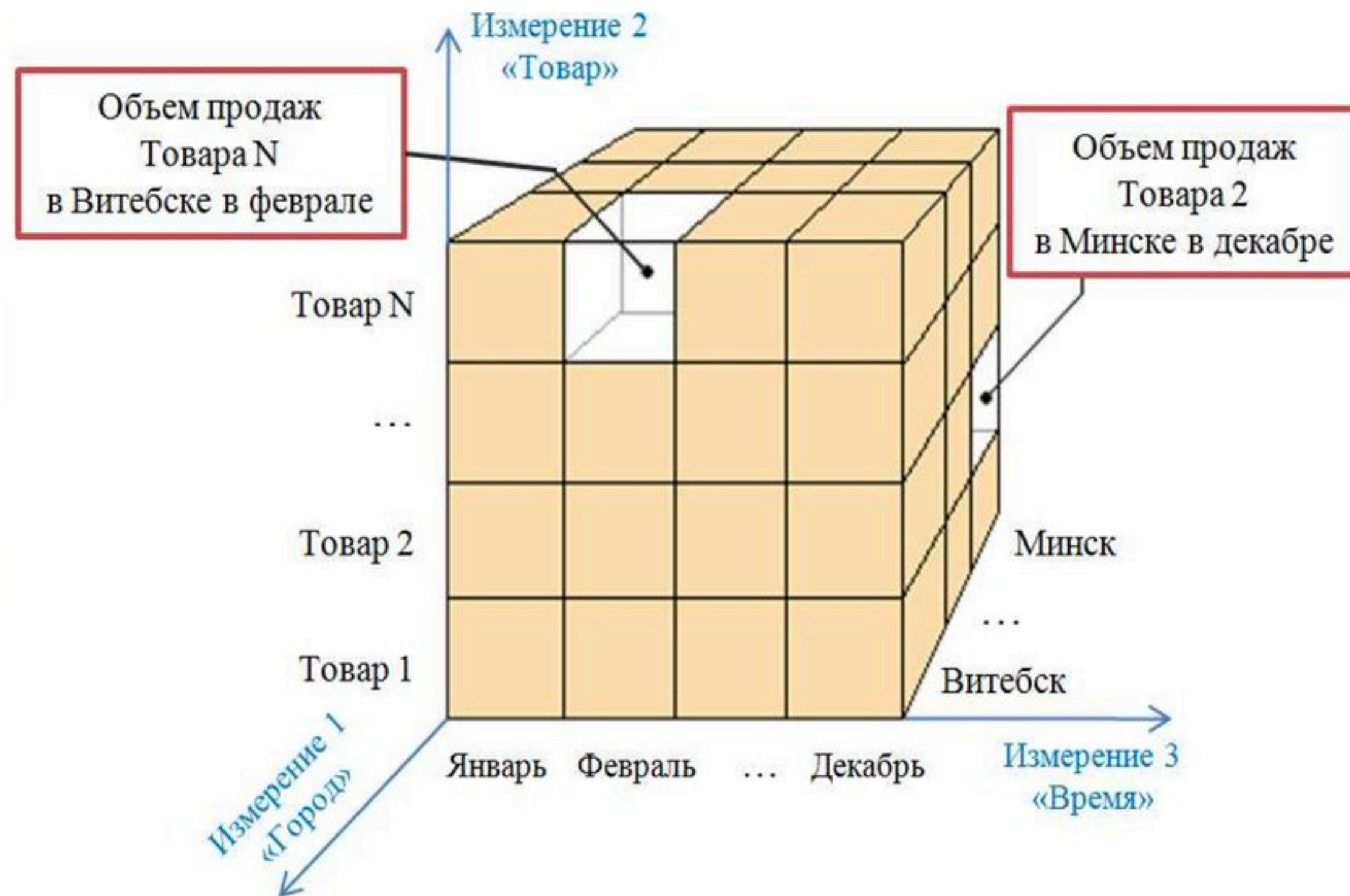
# Пространственная модель

Если реляционная модель акцентируется на целостности и эффективности ввода данных, то **размерная модель (Dimensional)** ориентирована в первую очередь на выполнение сложных запросов к БД.

В размерном моделировании для ХД принят стандарт модели, называемый **схемой звезда** (star schema), которая обеспечивает высокую скорость выполнения запроса посредством денормализации и разделения данных.



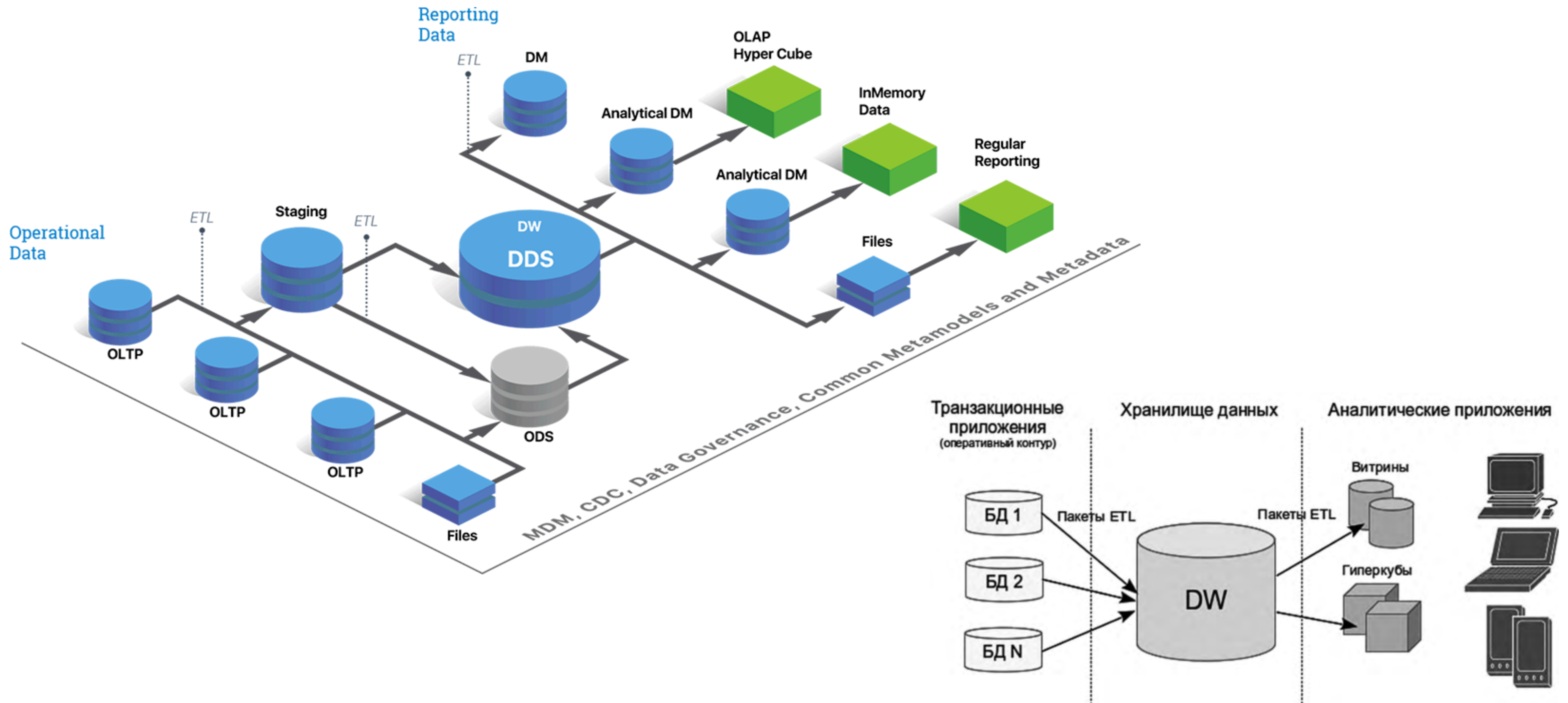
# OLAP-куб



Осями многомерной системы координат служат основные атрибуты анализируемого бизнес-процесса.

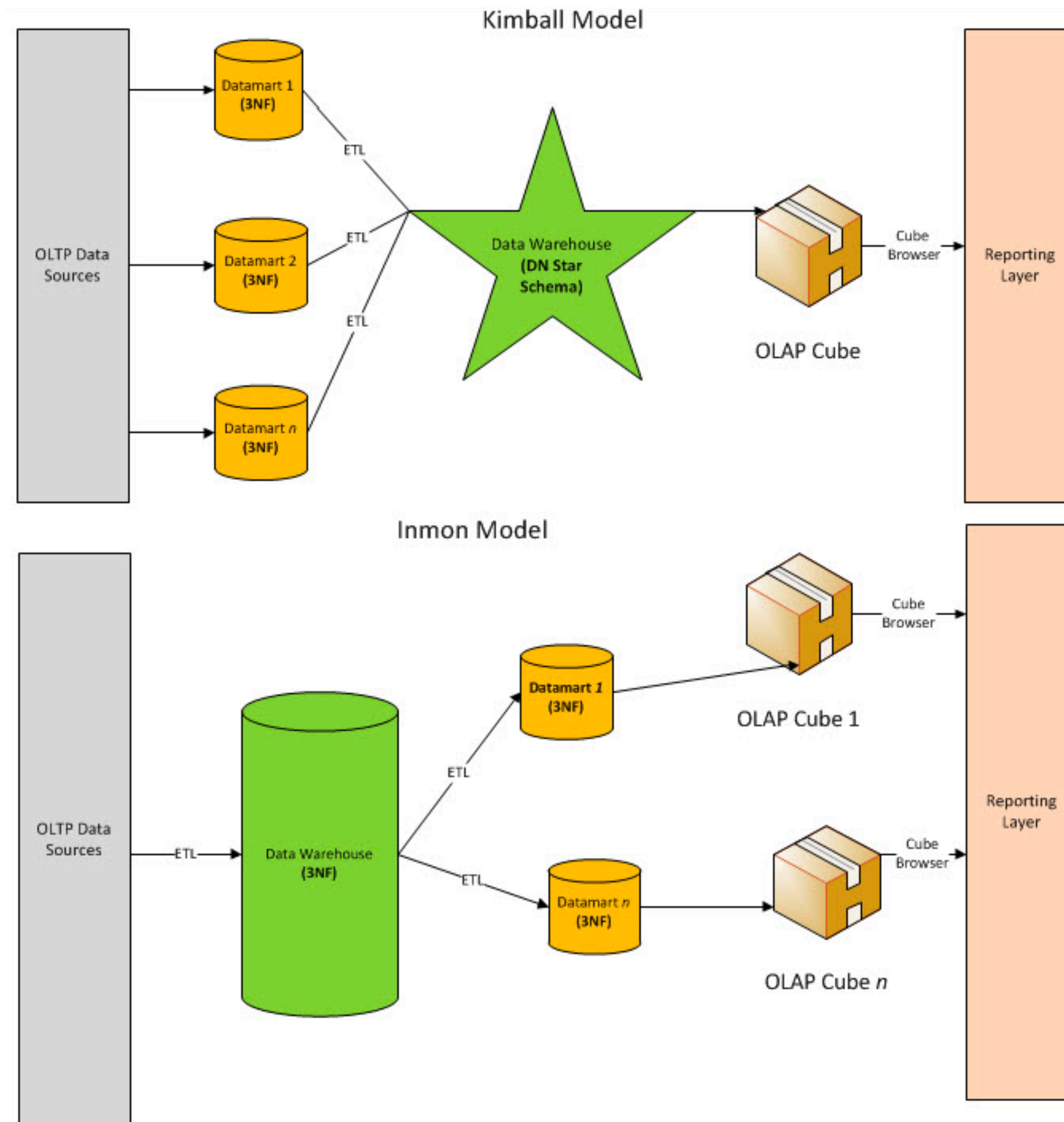
Например, для продаж это могут быть товар, регион, тип покупателя. На пересечениях осей измерений (Dimensions) находятся данные, количественно характеризующие процесс — меры (Measures).

# Аналитические хранилища данных





# Kimball vs. Inmon



## Отличия двух подходов:

1. Архитектура хранилища
  1. у Кимболла - пространственная организации
  2. у Инмона - нормализованная.
2. Физическая организации хранилища.
  1. у Инмона - это физически целостный реально существующий объект
  2. у Кимболла - скорее "виртуальный" объект. Коллекция витрин данных, которые могут быть пространственно разобщенными.

# ETL – аббревиатура от Extract, Transform, Load

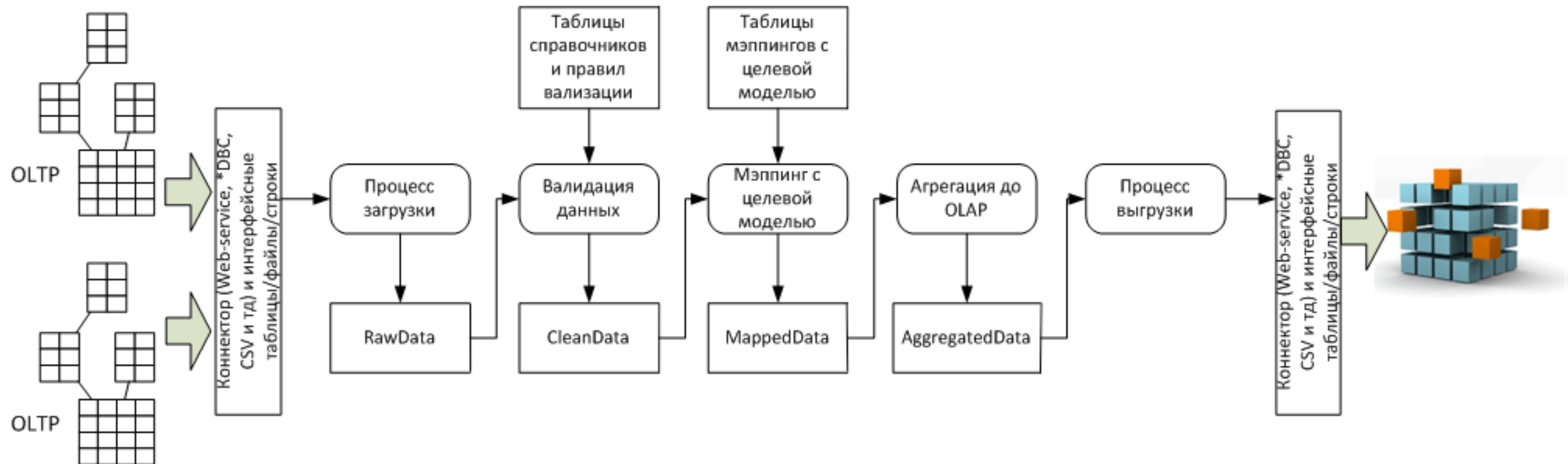
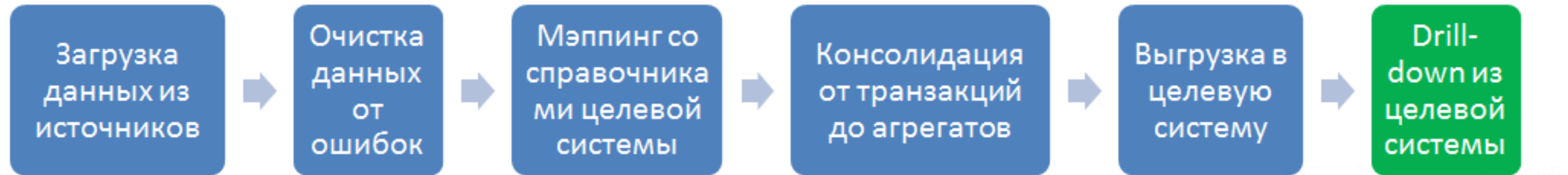
Какие задачи решает:

- Привести все данные к единой системе значений и детализации, попутно обеспечив их качество и надежность;
- Обеспечить аудиторский след при преобразовании (Transform) данных, чтобы после преобразования можно было понять, из каких именно исходных данных и сумм собралась каждая строка преобразованных данных.

Виды ошибок данных:

- Как случайные ошибки, возникшие на уровне ввода, переноса данных, или из-за багов;
- Как различия в справочниках и детализации данных между смежными ИТ- системами.

# Как работает ETL система



# Процесс загрузки данных

Задача этого этапа - затянуть в ETL данные произвольного качества для дальнейшей обработки. При проектировании процесса загрузки данных необходимо помнить о том что:

- Надо учитывать требования бизнеса по длительности всего процесса;
- Данные могут загружаться набегающей волной;
- Данные могут перегружаться много раз;
- Данные всегда содержат ошибки;
- Надо учитывать возможность обогащения данных.



# Процесс валидации

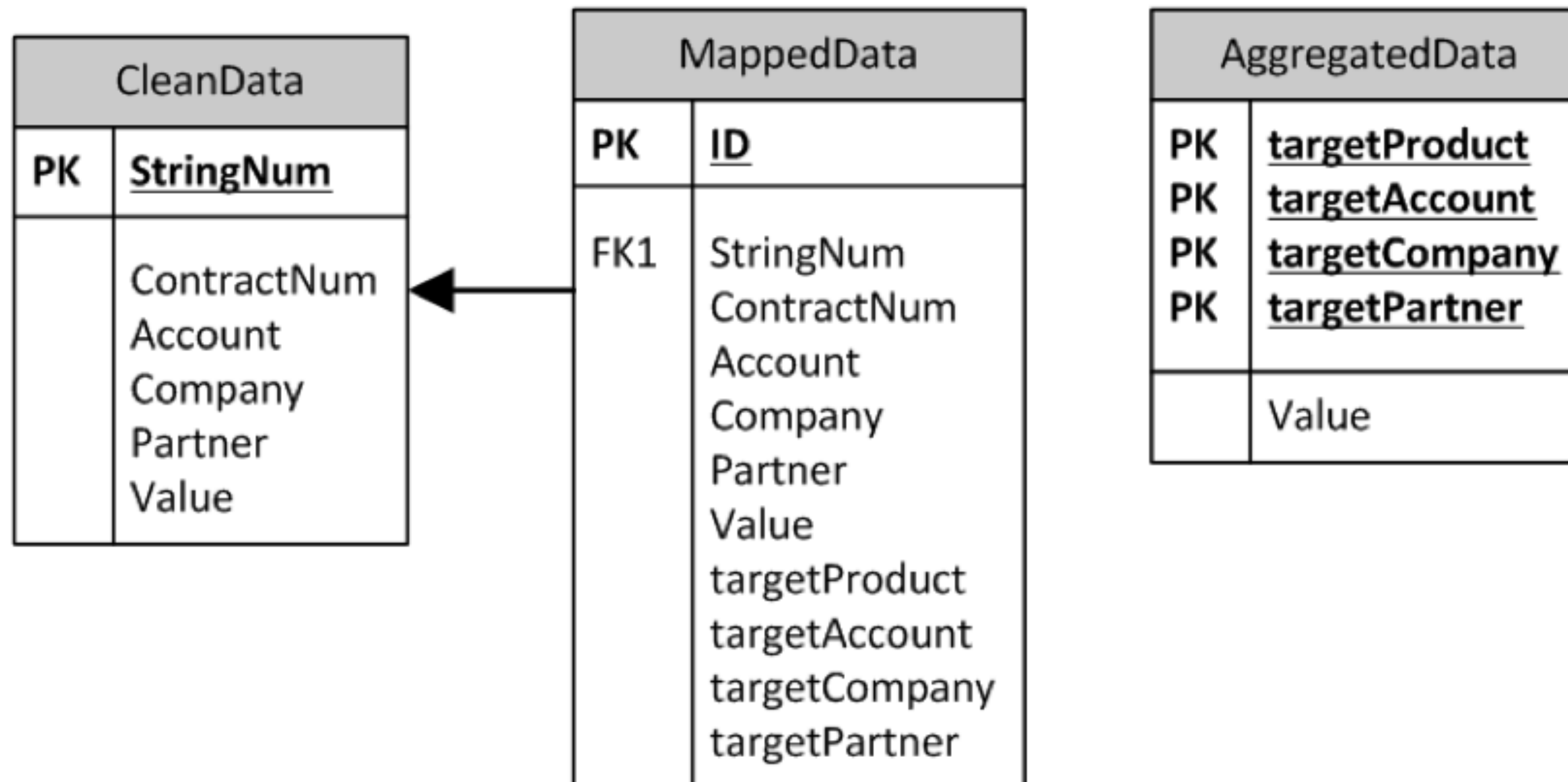
Данный процесс отвечает за выявление ошибок и пробелов в данных, переданных в ETL.

Главный вопрос – как вычислить возможные виды ошибок в данных, и по каким признакам их идентифицировать?

Типы данных	Внутри поля	По отношению к другим полям	Совместимость форматов при передачи между системами
Перечисление и текст	<ul style="list-style-type: none"> <li>• Не из списка разрешенных значений</li> <li>• Отсутствие обязательных значений</li> <li>• Не соответствие формату (например, все договора должны нумероваться «ДГВчччччч»)</li> </ul>	<ul style="list-style-type: none"> <li>• Не из списка разрешенных значений для связанного элемента</li> <li>• Отсутствие обязательных элементов для связанного элемента</li> <li>• Не соответствие формату для заданного элемента (например, для продукта «АИС» все договора должны нумероваться «АИСxxxxx»)</li> </ul>	<ul style="list-style-type: none"> <li>• Символы допустимые в одном формате, недопустимы в другом</li> <li>• Кодировка</li> <li>• Обратная совместимость</li> <li>• Новые значения (нет в мэппингах)</li> <li>• Устаревшие значения (не из списка разрешенных в целевой системе)</li> </ul>
Числа и порядки	<ul style="list-style-type: none"> <li>• Не число</li> <li>• Не в границах разрешенного интервала значений</li> <li>• Пропущено порядковое</li> </ul>	<ul style="list-style-type: none"> <li>• Не выполняется отношение</li> <li>• Присвоен неправильный порядковый номер</li> <li>• Разницы за счет разных правил округления значений</li> </ul>	<ul style="list-style-type: none"> <li>• Переполнение</li> <li>• Потеря точности и знаков</li> <li>• Несовместимость форматов при конвертации не в число</li> </ul>
Даты и периоды		<ul style="list-style-type: none"> <li>• День недели не соответствует дате</li> <li>• Сумма единиц времени не соответствует из-за разницы рабочие/не рабочие/праздничные/сокращенные дни</li> </ul>	<ul style="list-style-type: none"> <li>• Несовместимость форматов даты при передаче текстом</li> <li>• Ошибка точности отсчета и точности при передаче числом</li> </ul>

# Процесс мэппинга данных

Таблица замэпленных данных должна включать одновременно два набора полей – старых и новых аналитик, чтобы можно было сделать select по исходным аналитикам и посмотреть, какие целевые аналитики им присвоены, и наоборот:



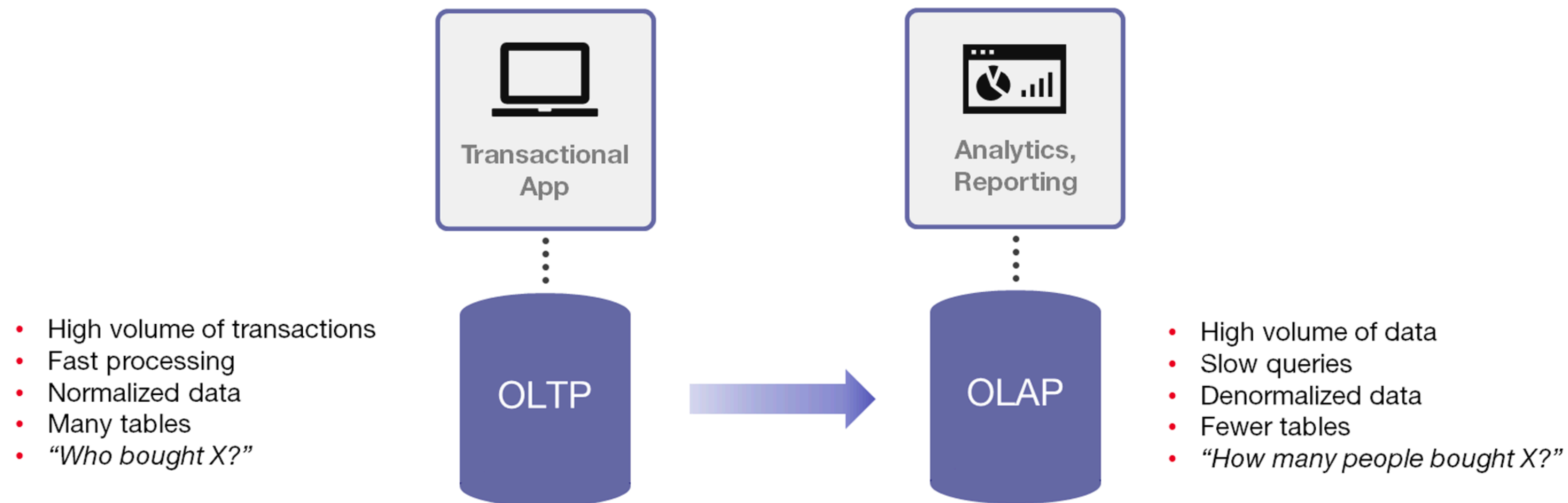
# Процесс агрегации данных

Этот процесс нужен из-за разности детализации данных в OLTP и OLAP системах.

OLTP система может содержать несколько сумм для одного и того же набора элементов справочников.

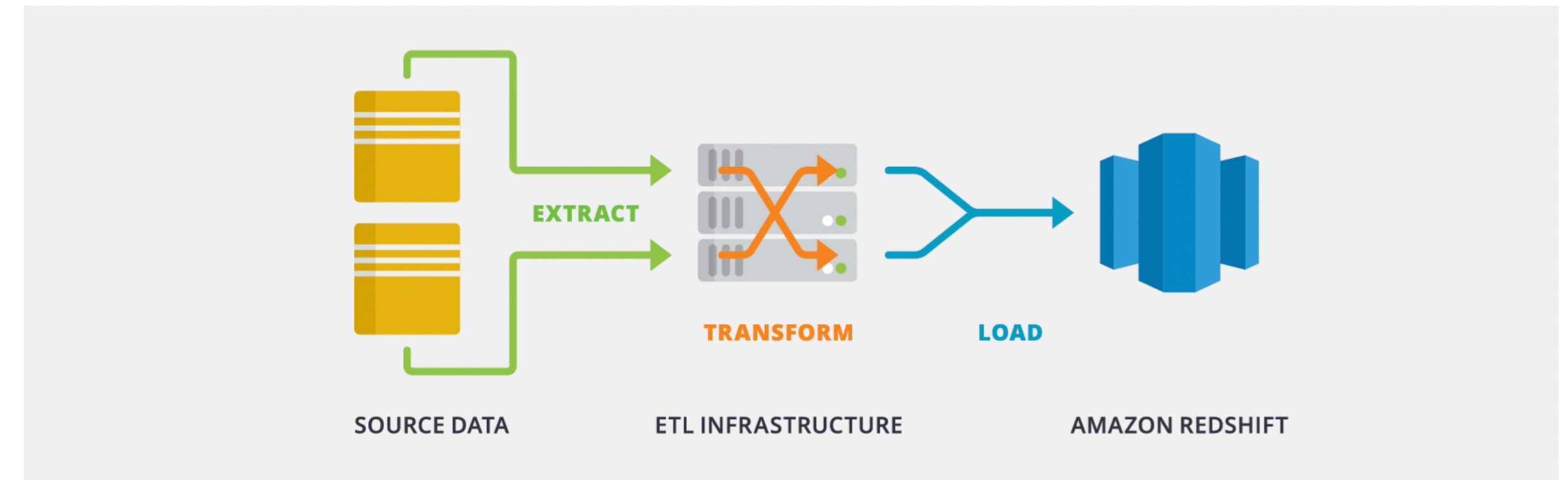
OLAP-системы — это, по сути, полностью денормализованная таблица фактов и окружающие ее таблицы справочников.

## OLTP vs OLAP



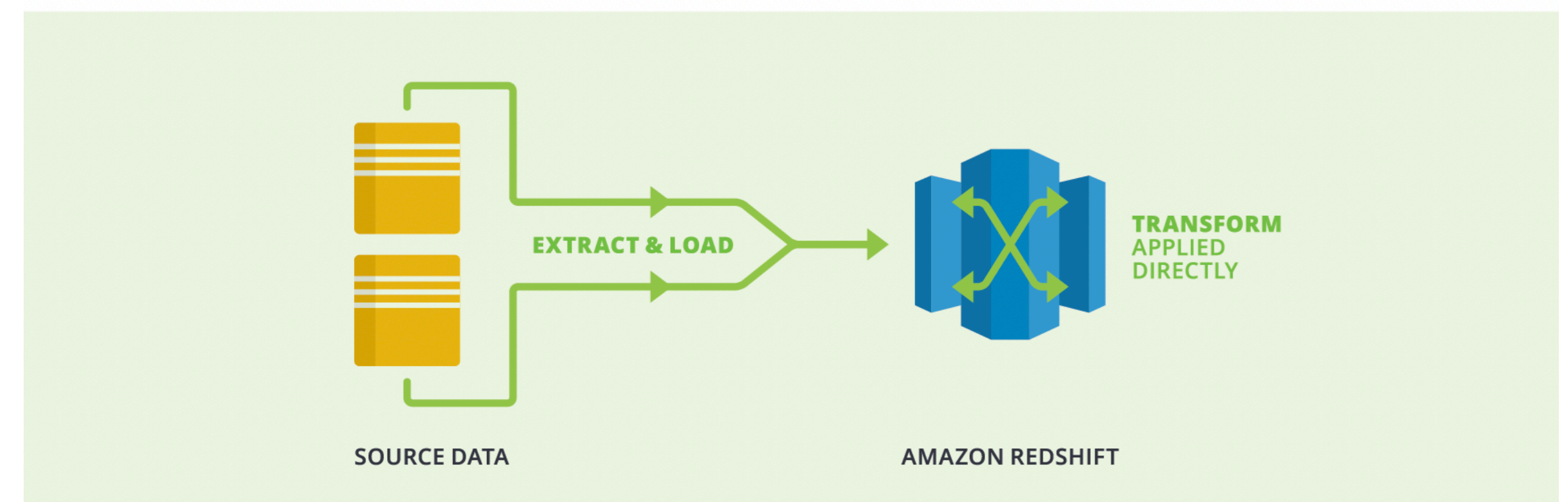
# ETL и ELT

**ETL (Extract, Transform, Load)** сначала извлекают данные из пула источников данных. Данные хранятся во временной промежуточной базе данных. Затем выполняются операции преобразования, чтобы структурировать и преобразовать данные в подходящую форму для целевой системы хранилища данных. Затем структурированные данные загружаются в хранилище и готовы к анализу.



В случае **ELT (Extract, Load, Transform)** данные сразу же загружаются после извлечения из исходных пулов данных. Промежуточная база данных отсутствует, что означает, что данные немедленно загружаются в единый централизованный репозиторий.

Данные преобразуются в системе хранилища данных для использования с инструментами бизнес-аналитики и аналитики.





# Data Lake

Data Lake — это хранилище данных, которое может хранить большое количество структурированных, полуструктурированных и неструктурированных данных. Это место для хранения всех типов данных в собственном формате без фиксированных ограничений на размер учетной записи или файл. Он предлагает большое количество данных для повышения аналитической производительности и встроенной интеграции.

Data Lake похожа на настоящее озеро и реки. Точно так же, как в озере есть несколько притоков, озеро данных содержит структурированные данные, неструктурированные данные, от машины к машине, журналы, проходящие в режиме реального времени.

