

Проектирование Data Warehouse (DWH) - основы

ETL-процессы

ETL (Extract, Transform, Load) — это процесс, который включает извлечение данных из различных источников, их преобразование для соответствия бизнес-правилам и требованиям, и загрузку в целевую систему, обычно в хранилище данных (DWH).

Процесс настройки ETL

1. **Определение источников данных (Extract):** Начните с определения источников данных, откуда будут извлекаться данные. Это могут быть базы данных, CRM-системы, файлы Excel, лог-файлы и т.д.
2. **Определение целевой структуры данных:** Определите, как должны выглядеть данные в хранилище. Это включает в себя модели данных слоев STG, DDS и CDM.
3. **Проектирование процесса трансформации (Transform):** Преобразование данных может включать очистку, дедупликацию, сопоставление, агрегацию, расчеты и применение бизнес-правил. Для каждого типа данных определите необходимые преобразования.
4. **Настройка загрузки данных (Load):** Планируйте, как и в каком порядке данные будут загружаться в хранилище. Для DDS и CDM важен порядок из-за зависимостей и ключей.

Как на деле работает ETL

- **Использование ETL-инструментов:** Существуют специализированные программные решения, которые предоставляют графические интерфейсы и наборы инструментов для упрощения процесса ETL. Например, инструменты, такие как Talend, Informatica или SSIS (SQL Server Integration Services), позволяют вам "нарисовать" ETL-процесс и автоматизировать его исполнение.

- Скрипты и программирование: Можно написать скрипты на SQL или использовать языки программирования, такие как Python, для выполнения более сложных операций ETL.

Давайте рассмотрим примеры

- Staging (STG): Вы можете настроить задание, которое периодически извлекает данные из CRM-системы и загружает их в STG. Здесь данные остаются в "сыром" виде без преобразований.

```
INSERT INTO stg_customers SELECT * FROM crm_customers;
```

- Detail Data Store (DDS): После STG данные трансформируются для DDS. Если вам нужно загрузить данные о продажах, убедитесь, что они соответствуют схеме вашей DDS, и что клиенты и продукты уже загружены в соответствующие таблицы. В DDS очень важен порядок миграции данных, здесь с этим строго.

```
INSERT INTO dds_sales (sale_id, customer_id, product_id, sale_date, amount) SELECT source_sale_id, customer_key, product_key, sale_date, amount FROM stg_sales JOIN dds_customers ON stg_sales.customer_id = dds_customers.source_customer_id JOIN dds_products ON stg_sales.product_id = dds_products.source_product_id;
```

- Common Data Marts (CDM): Для CDM данные могут быть агрегированы на основе бизнес-логики. Например, если вам нужно создать витрину для отчета о ежемесячных продажах по продуктам, вы можете использовать агрегирующие функции SQL для расчета общих сумм продаж.

```
INSERT INTO cdm_monthly_sales (month, product_id, total_sales) SELECT DATE_TRUNC('month', sale_date), product_id, SUM(amount) FROM dds_sales GROUP BY DATE_TRUNC('month', sale_date), product_id;
```