DATA ENGINEERING

DWH – корпоративное хранилище данных

АЛЕКСАНДРОВ АНТОН

Head of Big Data Platform /
Детский мир

СОДЕРЖАНИЕ

О СЕБЕ

1. Обучение
   МИИТ Прикладная математика и Информатика
   МФТИ Прикладная математика и физика
2. Предыдущие проекты
   BEELINE - Performance Management платформа
   200 GB raw data /day
   Работа с MapReduce API
   LUXOFT - Озеро данных Почты России
   Отчеты для команды Логистики
   SEVERGROUP - собственная платформа ClickStream аналитики
   Использование ClickHouse как основного хранилища
   РОСТЕЛЕКОМ – Создание клиентского профиля
   1.5 TB/day
   Uid matching
   Внедрение ML моделей в Production
   Решение проблем производительности
3. Суммарный опыт работы с Big Data tools 7 лет
4. **ДЕТСКИЙ МИР** / Head of Big Data Platform

DWH — ОСНОВНЫЕ ПОНЯТИЯ

DWH — ОПРЕДЕЛЕНИЕ

DWH — это система, которая консолидирует и хранит корпоративную информацию из различных источников в форме, подходящей для аналитических запросов и отчетов для поддержки инициатив бизнес-аналитики и анализа данных.

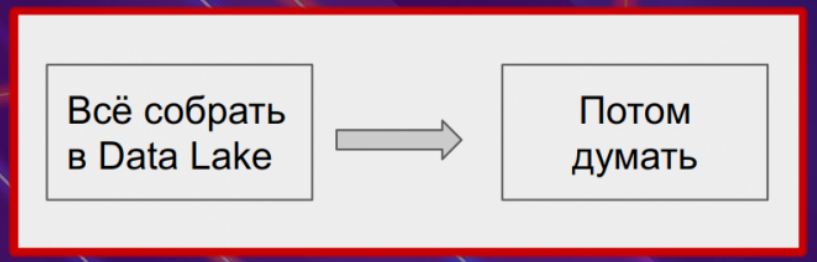
ЗАЧЕМ НУЖЕН DWH?

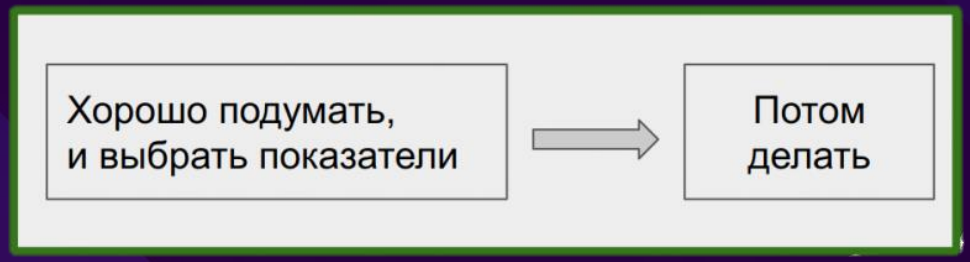
- Единое хранилище аккумулирующее данные из разных отделов одного предприятия
- Единая справочная система
- Аналитика и отчетность на уровне всего предприятия
- Накопление исторических данных – возможности для анализа с использованием временных рядов
- Широкий выбор методов анализа собранных данных
- Высокое качество данных
- Высокая скорость анализа и формирования отчетности

С ЧЕМ МОЖНО СПУТАТЬ DWH?

Data Lake – озеро данных

- Data Lake ориентированы на накопление большИх объемов неструктурированных данных, бОльшая часть из которых может не быть востребована здесь и сейчас. Data lake, как правило, не имеют четкой архитектуры по слоям, а также имеют значительно более низкие требования к модели данных.





С ЧЕМ ЕЩЕ МОЖНО СПУТАТЬ DWH?

Data mart – витрины данных

- Витрины данных – представляют собой срез агрегированных эталонных данных и предназначены в первую очередь для конечного пользователя(аналитика или бизнес-аналитика). Витрины могут являться частью DWH, но не могут выполнять функции хранилища.

База данных

- Классические транзакционные БД имеют, как правило, четкую ориентированность на конкретное бизнес-подразделение компании, например, отдел продаж или склад. БД может выступать в качестве источника данных для DWH, но не подходит для анализа в масштабе всей организации.

OLAP-куб

- Структура данных главной целью которой является многомерный анализ данных. OLAP куб содержит в себе лишь единовременный слепок данных, но не может выступать в качестве консолидированного хранилища, каким является DWH

КЕЙСЫ КОМПАНИЙ

ЛОГИСТИКА

- Соблюдение SLA
- Остатки в объекте/в пути
- Выявление зацикливаний
- Предсказание сроков доставки
- Трассировка

ТЕЛЕКОММУНИКАЦИИ

- Выявление оттока клиента
- Проактивная поддержка
- Анализ в режиме реального времени
- Целевой маркетинг

ФИНАНСОВЫЙ СЕКТОР

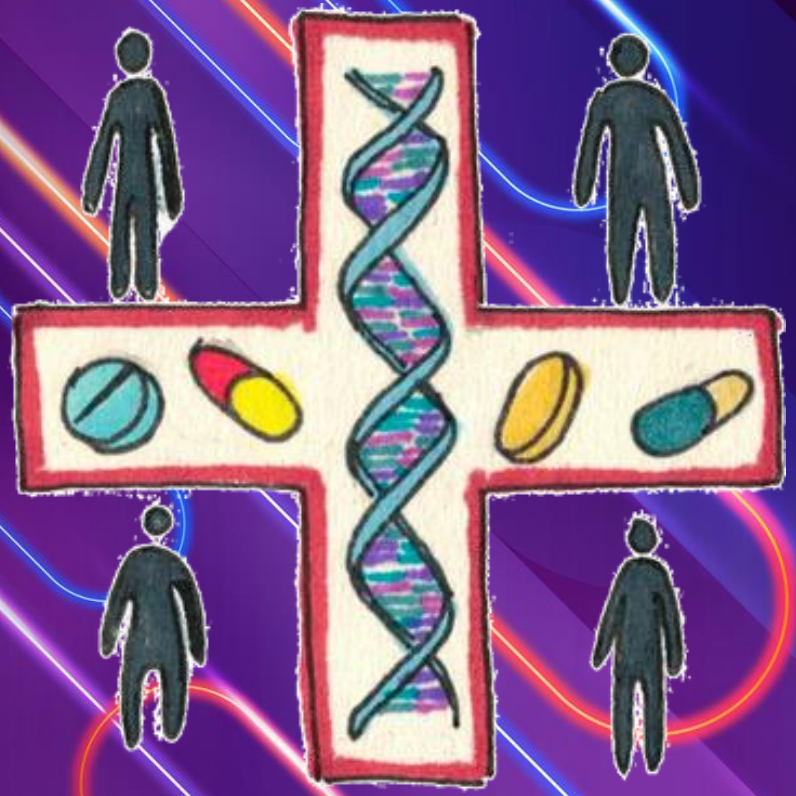
- Скоринг клиентов
- Анализ рынка
- Создание персонализированных предложений

СОЦИАЛЬНЫЕ СЕТИ

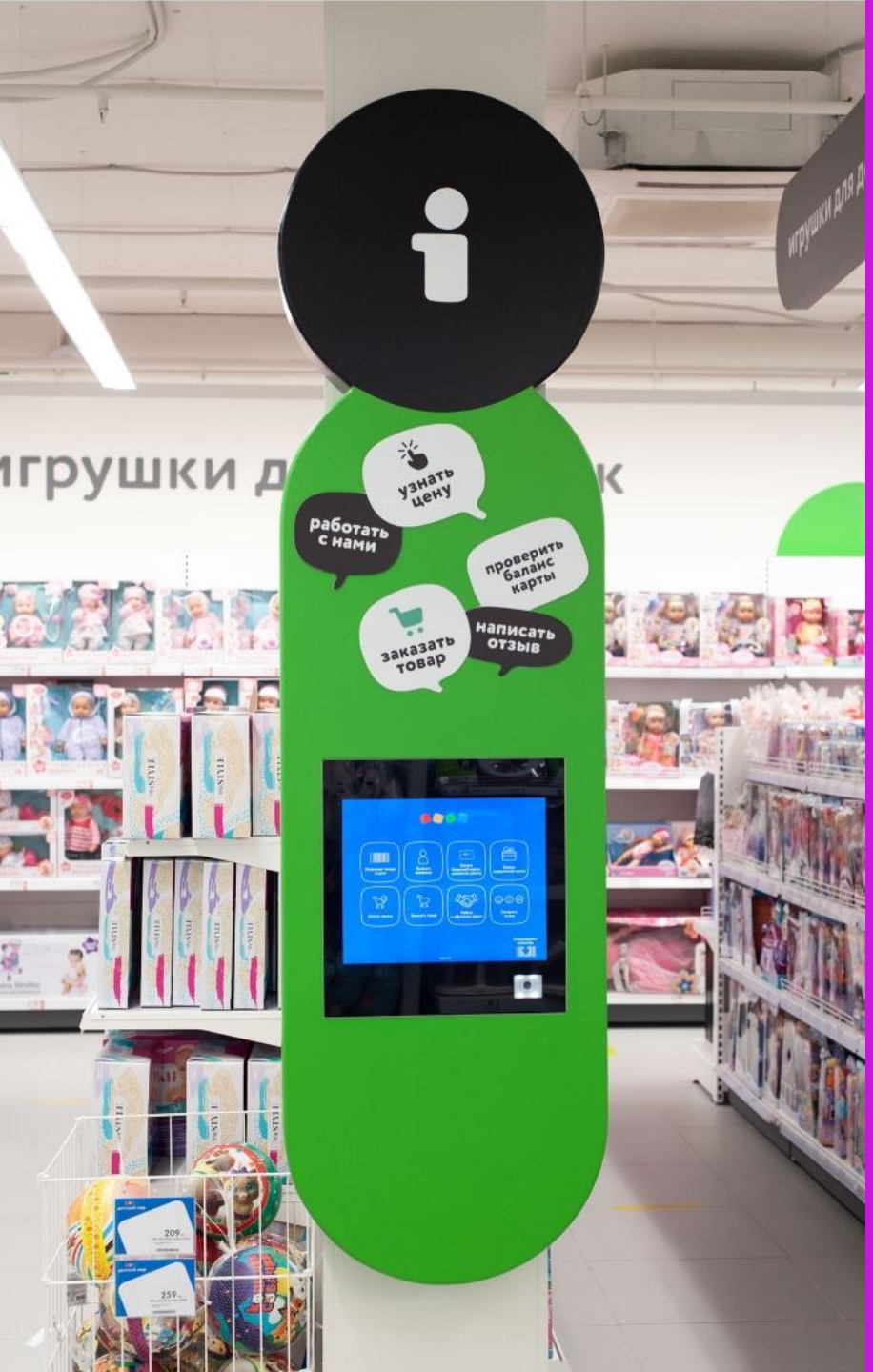
- Показ рекламы
- Анализ графа связей
- Рекомендации
- Анализ взаимодействия с контентом

* «Facebook/Instagram — проект Meta Platforms Inc., деятельность которой в России запрещена»

МЕДИЦИНА



- Анализ эффективности препаратов
- Проактивный анализ состояния здоровья

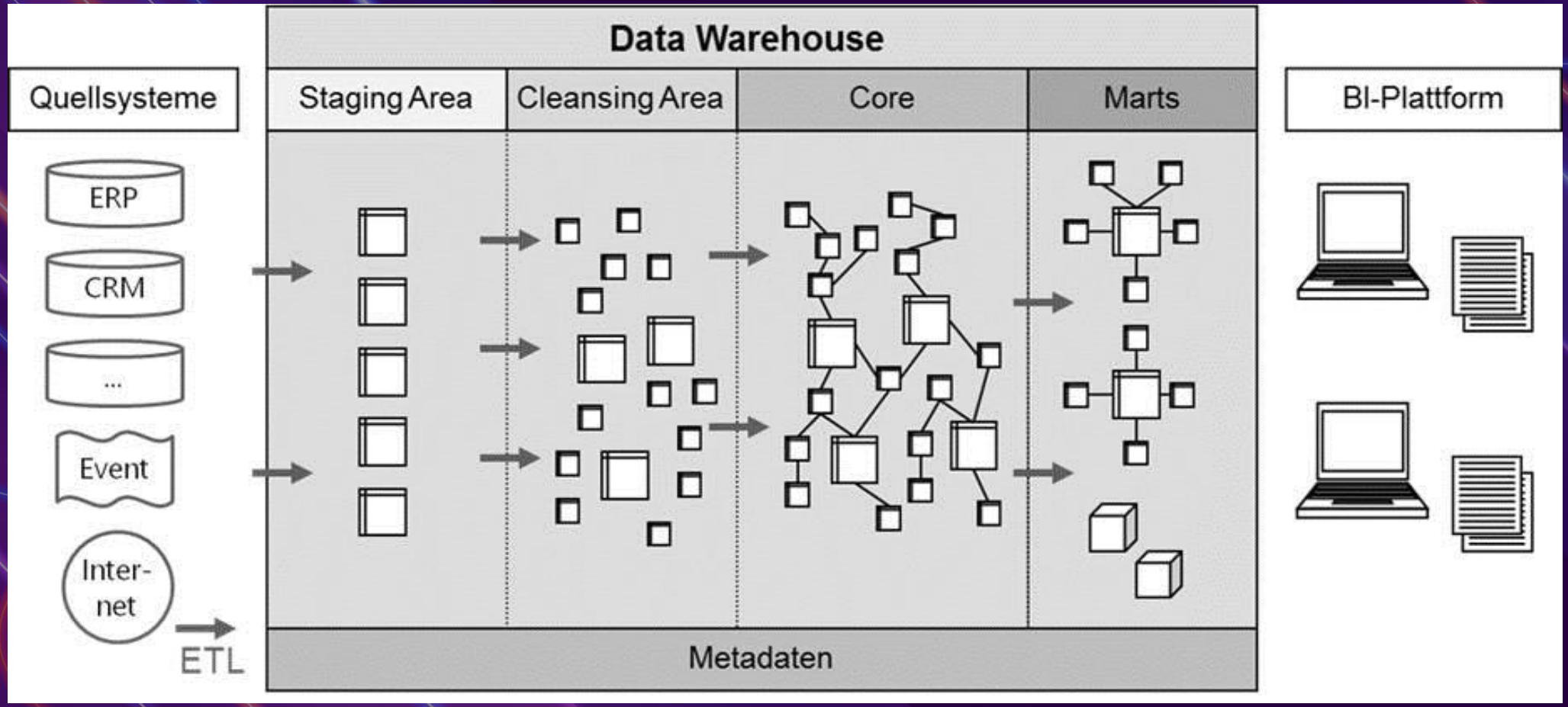


РИТЕЙЛ

- Понимание текущего состояния компании
- Анализ пользовательской активности
- Рекомендации в учетом потребления
- A/B тесты при продажах на сайте
- Расчеты сложных финансовых показателей

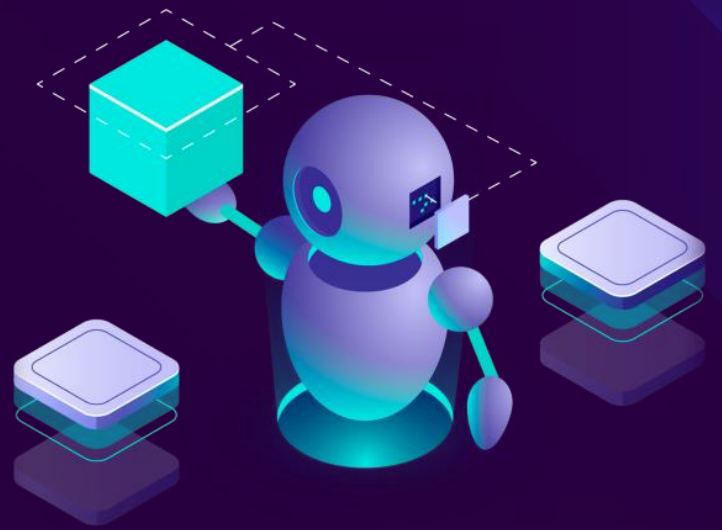
АРХИТЕКТУРА DWH

КАК ВЫГЛЯДИТ ТИПИЧНОЕ DWH?



Из чего состоит DWH?

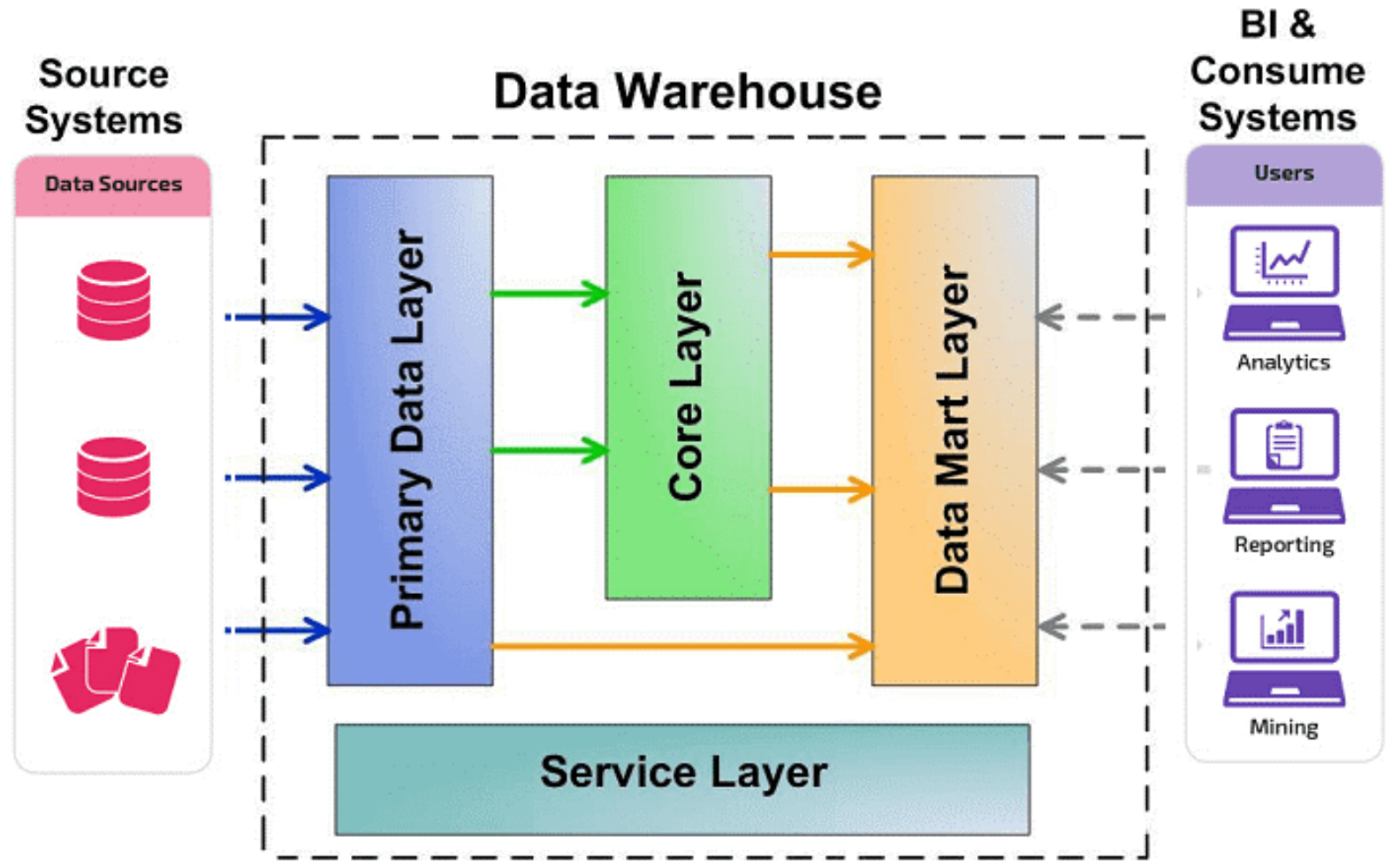
- **Источники данных**
  - БД
  - Data lake
  - Все что угодно, что можно привести к соответствующей структуре
- **Слои данных – LSA(Layered Scalable Architecture)**
  - RAW
  - CORE
    - Модель данных
    - SCD
  - MART
- **ETL - движок**
  - Набор программных инструментов для реализации работы с данными
- **Справочники и метаданные**



ИСТОЧНИКИ ДАННЫХ ДЛЯ DWH

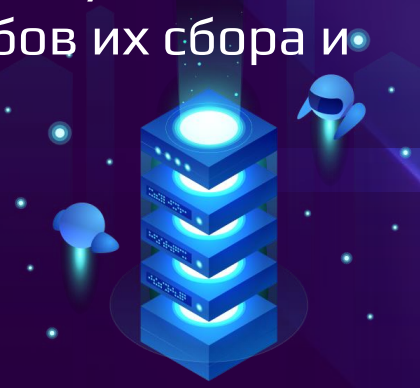
- Реляционные хранилища – Postgres, Oracle, MS SQL
- Нереляционные хранилища – Hadoop, Kafka
- ERP-системы – SAP, 1С
- Файлы со структурой – csv, json, xml
- Файлы без структуры – word, текстовые файлы
- API в сторонние сервисы, в т.ч. Web-сервисы

АРХИТЕКТУРА DWH



ПРОМЕЖУТОЧНАЯ ОБЛАСТЬ (RAW ИЛИ STG LAYER)

Загрузка информации из систем-источников в исходном качестве и сохранением полной истории изменений. Здесь происходит абстрагирование следующих слоев хранилища от физического устройства источников данных, способов их сбора и методов выделения изменений.



Ключевые задачи:

- Сохранить данные с источников «как есть»
- Обеспечить возможность перезагрузки данных без обращения к источнику
- Устойчивость к изменениям

Ядро хранилища (Core Layer)

Центральный компонент, который выполняет консолидацию данных из разных источников, приводя их к единым структурам и ключам. Именно здесь происходит основная работа с качеством данных и общие трансформации, чтобы абстрагировать потребителей от особенностей логического устройства источников данных и необходимости их взаимного сопоставления. Так решается задача обеспечения целостности и качества данных.

Ключевые задачи:

- Предобработка, очистка и трансформация данных
- Консолидация данных между источниками
- Разработка модели данных на уровне бизнес-сущностей и связей между ними
- Устойчивость к изменениям
- Масштабируемость
- Предоставление стандартного интерфейса доступа к данным

МОДЕЛИРОВАНИЕ ДАННЫХ

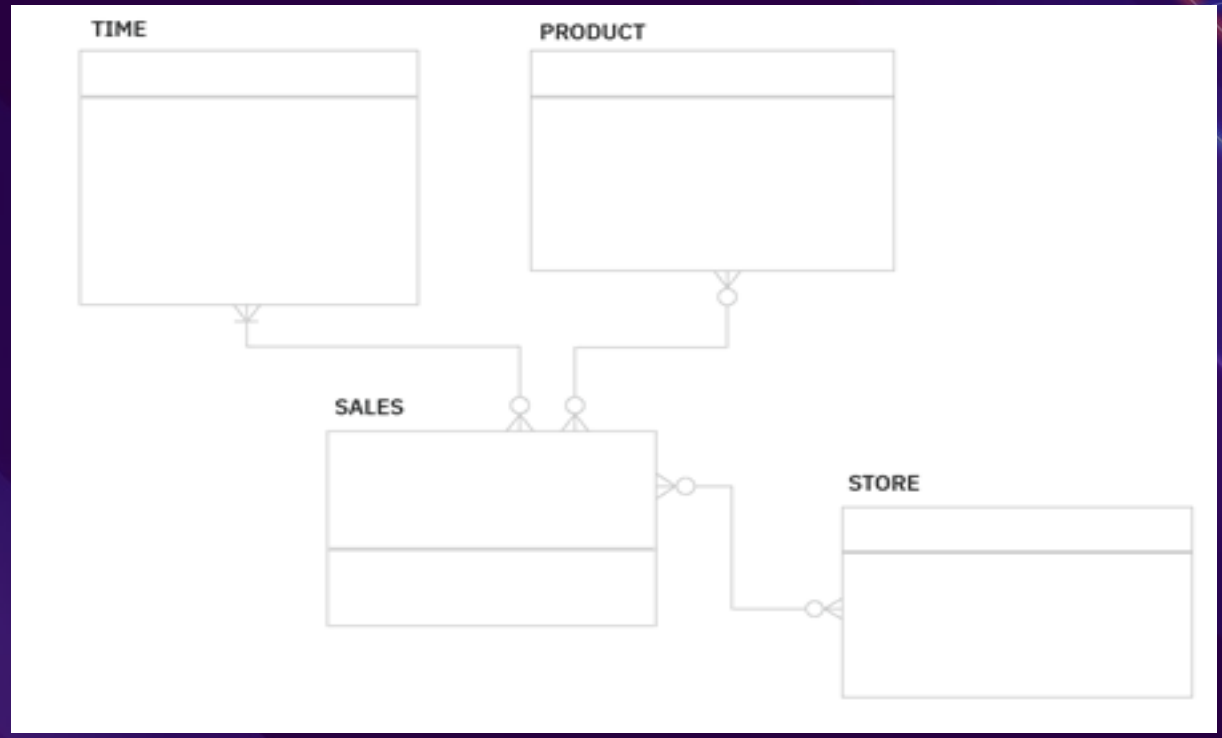
Модель данных корпоративного хранилища представляет собой ER-модель (Entity-relationship model — модель «сущность-связь»), описывающую на нескольких уровнях набор взаимосвязанных сущностей, которые сгруппированы по функциональным областям и отражают потребности бизнеса в аналитическом анализе и отчетности.

Общая модель данных корпоративного хранилища разрабатывается последовательно и состоит из:

- концептуальной модели данных;
- логической модели данных;
- физической модели данных.

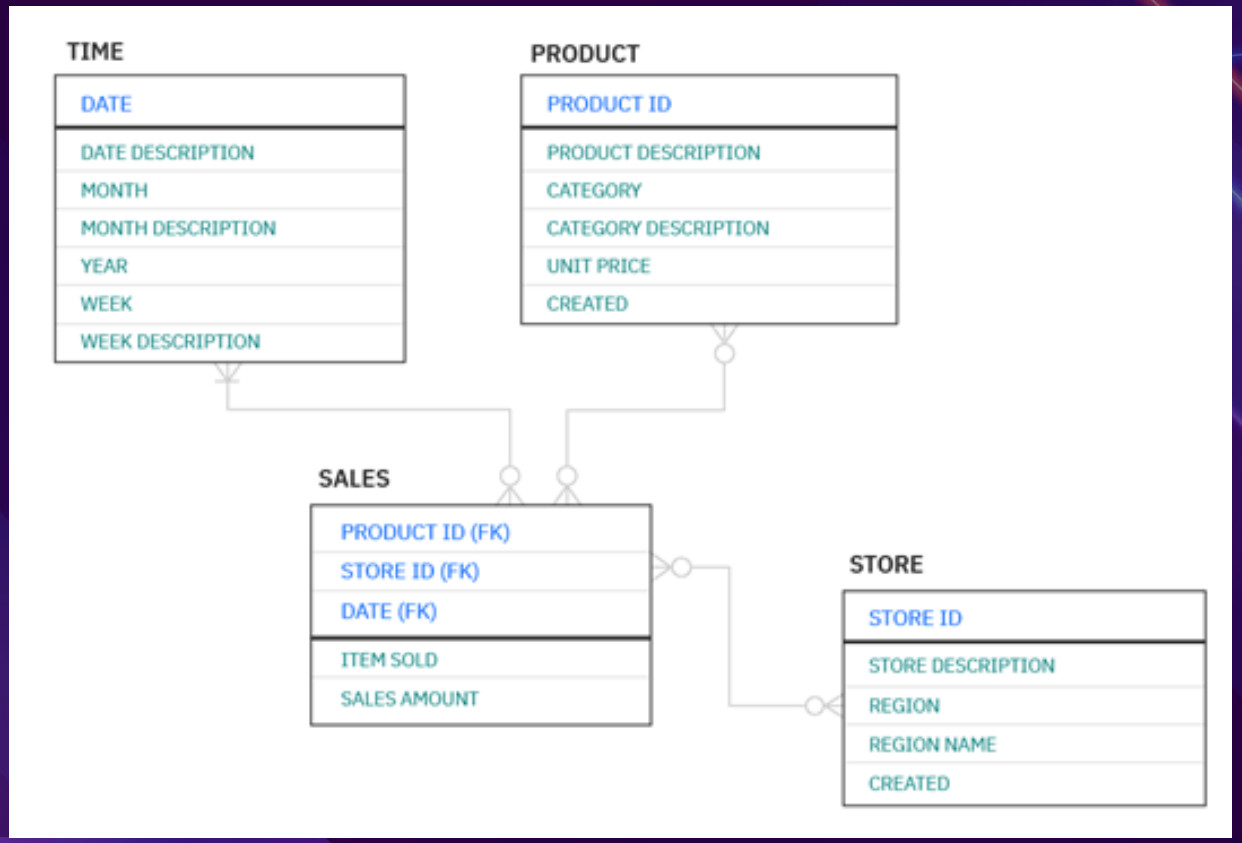
КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ

Концептуальная модель хранилища данных представляет собой описание главных (основных) сущностей и отношений между ними. Концептуальная модель является отражением предметных областей, в рамках которых планируется построение хранилища данных.



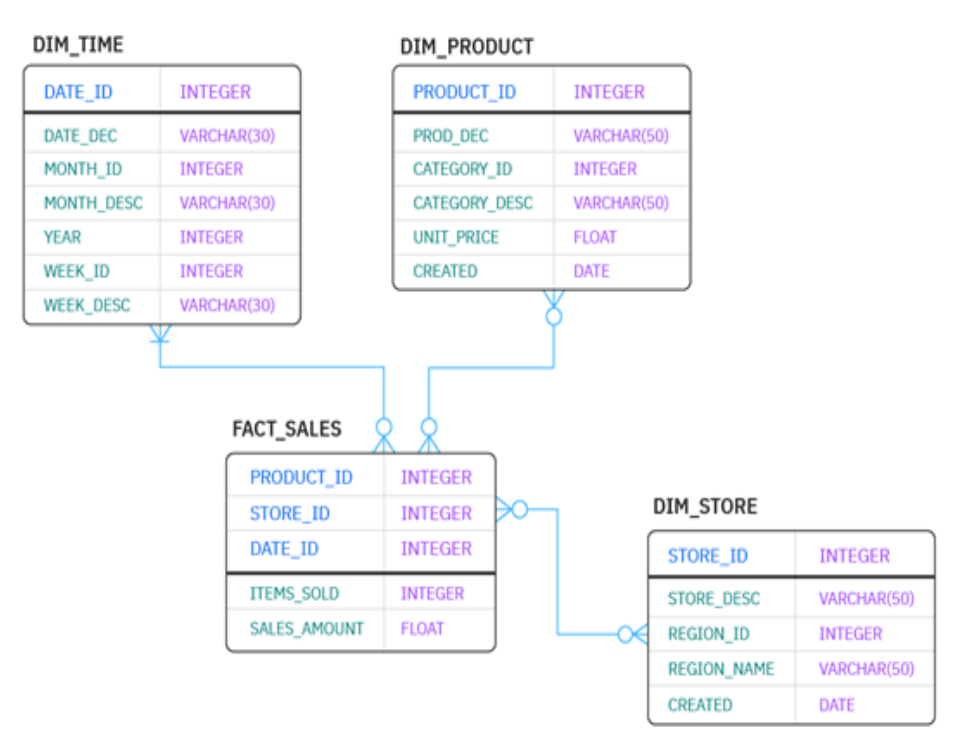
ЛОГИЧЕСКАЯ МОДЕЛЬ

Логическая модель расширяет концептуальную путем определения для сущностей их атрибутов, описаний и ограничений, уточняет состав сущностей и взаимосвязи между ними.



ФИЗИЧЕСКАЯ МОДЕЛЬ

Физическая модель данных описывает реализацию объектов логической модели на уровне объектов конкретной базы данных.



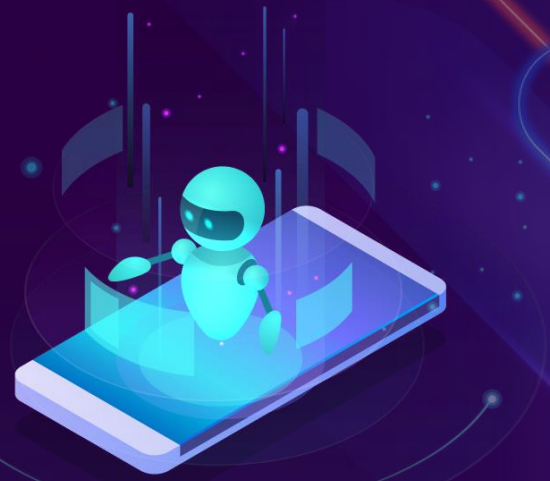

```
CREATE TABLE [Product] (
    [ID_Product] Integer Not Null Primary Key,
    [Name] VarChar(100) ,
    [Price] Decimal(15, 2),
    [Count] Integer,
    [Date] Date,
    [Note] VarChar(200)
)
```

СРАВНЕНИЕ МОДЕЛЕЙ РАЗЛИЧНЫХ УРОВНЕЙ

Объекты модели	Концептуальная модель	Логическая модель	Физическая модель
Предметная область (Subject Area)	X		
Сущности (Entitys)	X	X	
Взаимосвязи между сущностями (Entity Relationships)	X	X	
Атрибуты (Attributes)		X	
Первичные ключи (Primary Keys)		X	X
Внешние ключи (Foreign Keys)		X	X
Наименование таблиц (Table Names)			X
Наименование колонок (Column Names)			X
Типы данных (Column Data Types)			X

ГДЕ РИСОВАТЬ ER-ДИАГРАММЫ?

- **Draw.io** - https://app.diagrams.net/
- **Miro** - https://miro.com/ru/
- **PlantUML** - https://plantuml.com/ru/ie-diagram
- **Paint** – шутка, нет



АНАЛИТИЧЕСКИЕ ВИТРИНЫ (DATA MART LAYER)

Слой, на котором данные преобразуются к структурам, удобным для анализа и использования в BI-дашбордах или других системах-потребителях.

Ключевые задачи:

- Формировать данные в контексте бизнес-потребностей
- Оптимизировать доступ на чтение

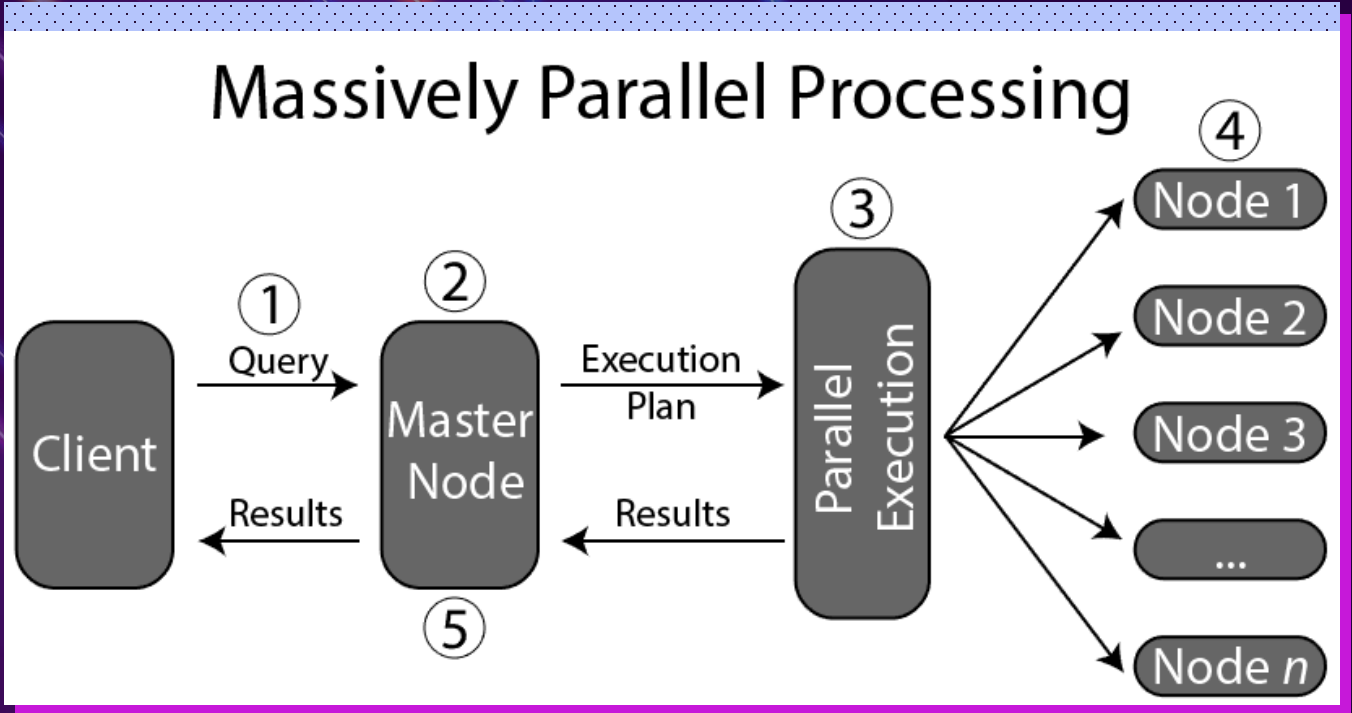
СЕРВИСНЫЙ СЛОЙ (SERVICE LAYER)

На данном слое обеспечивается управление всеми вышеописанными уровнями. Он не содержит бизнес-данных, но оперирует метаданными и другими структурами для работы с качеством данных, Также здесь доступны средства мониторинга и диагностики ошибок, что ускоряет решение проблем.

Ключевые задачи:

- Хранение метаданных;
- Анализ метаданных;
- Обеспечение качества и целостности загружаемых данных.

НЕ МНОГО О MPP



Массово-параллельная архитектура

класс архитектур параллельных вычислительных систем. Особенность архитектуры состоит в том, что память физически разделена.

Плюсы

Горизонтальная масштабируемость

Отказоустойчивость

Минусы

Сложность эксплуатации

Не правильное распределение данных влечет большой штраф по производительности

ПРОГРАММНЫЕ СРЕДСТВА ХРАНЕНИЯ ДАННЫХ DWH

Классические реляционные хранилища

- Postgres
- Oracle
- MS SQL Server

MPP-системы

- Vertica
- Green Plum
- Teradata

Облачные

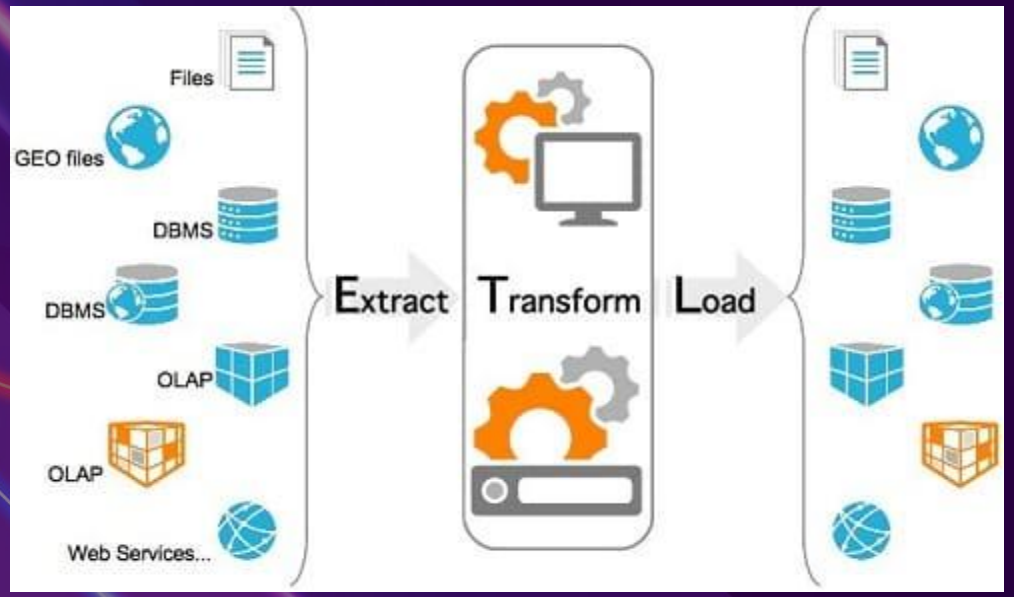
- AWS
- Google cloud platform
- Microsoft Azure

ETL-ПРОЦЕССЫ В DWH

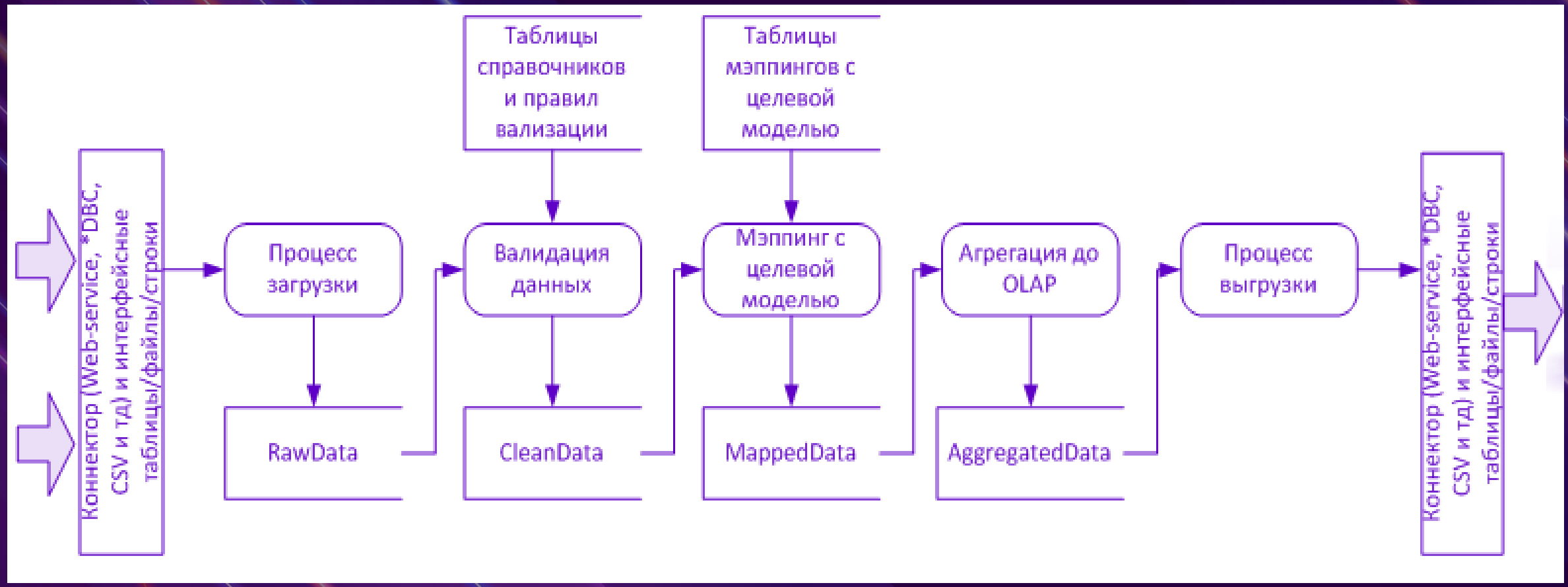
ETL - ПРОЦЕССЫ

ETL (Extract, Transform, Load) – это совокупность процессов управления хранилищами данных, включая:

- извлечение данных из внешних источников;
- преобразование и очистка данных согласно бизнес-потребностям
- загрузка обработанной информации в корпоративное хранилище данных



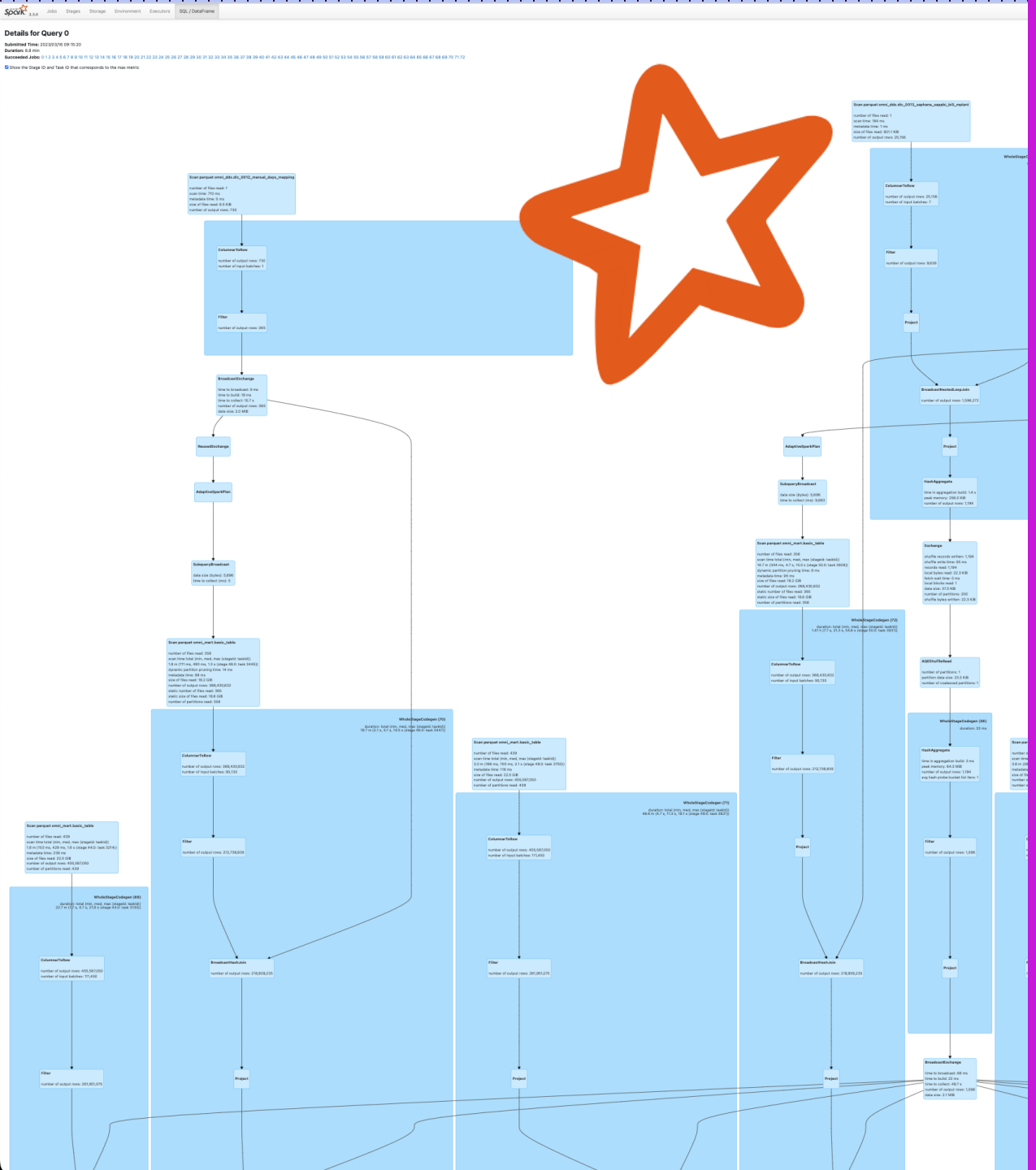
Более подробно про ETL



БОЛЕЕ ПОДРОБНО ПРО ETL

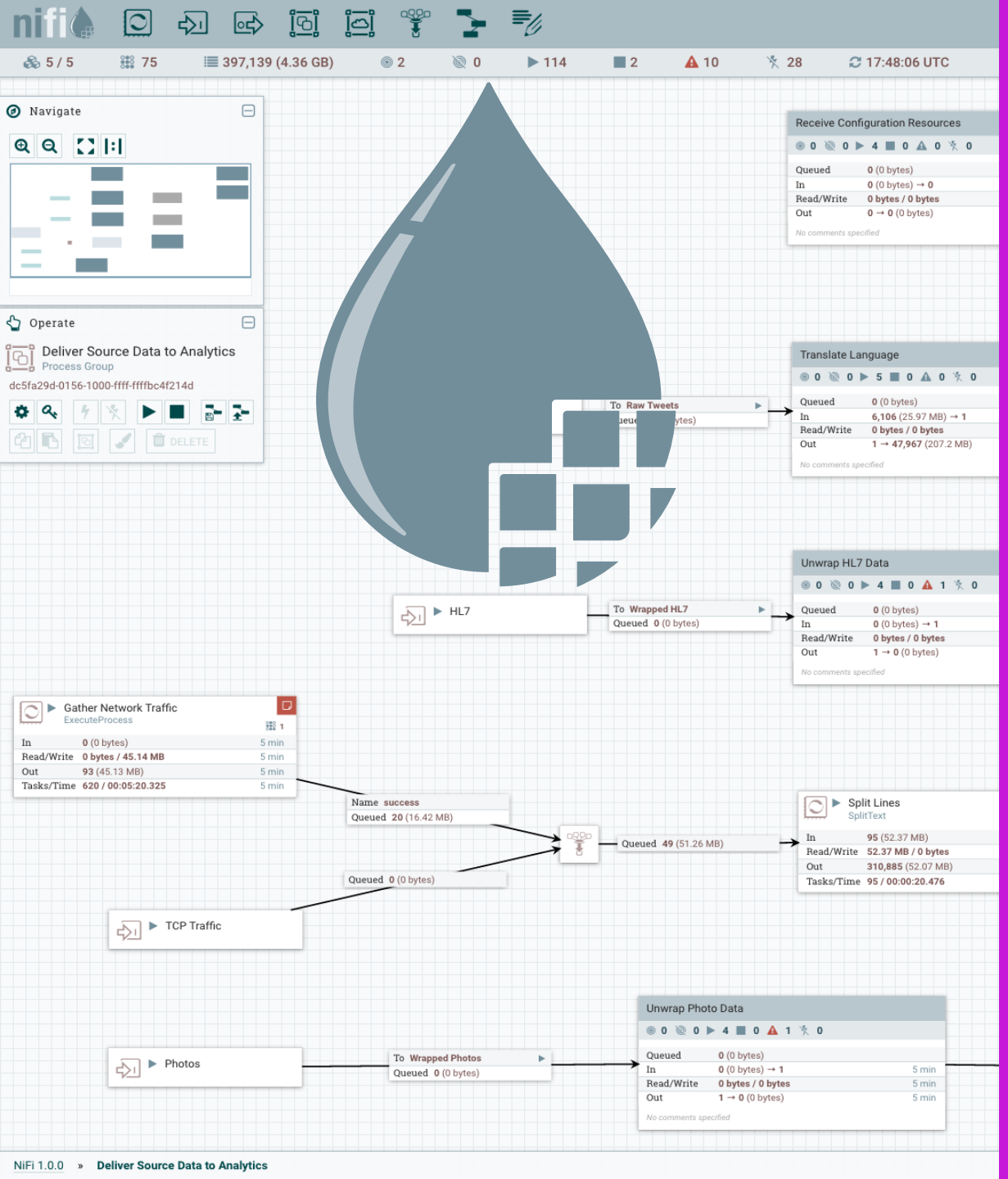
1. **Процесс загрузки** – на данном этапе данные загружаются «как есть» без дополнительных проверок качества. На этом шаге важно оценить корректность данных с точки зрения их объема. Также важно учитывать особенности систем источников и систем приемников данных.
2. **Процесс валидации данных** – на этом этапе данные очищаются от дублей, шумов, выбросов и прочих артефактов. Оценка соответствия типам данных. Оценка корректности и адекватности загружаемых данных. Важно составлять отчет по каждой выгрузке данных о наличии тех или иных ошибок.
3. **Процесс мэппинга данных с целевой моделью** – на этом этапе происходит процесс трансформации данных таким образом, чтобы их можно было встроить в соответствующую модель данных. Т.е. данные «распиливают» по сущностям, проставляют нужные ключи, добавляют технические атрибуты.
4. **Процесс агрегации данных** – процесс предподготовки данных перед формированием витрин. Ключевая задача – упростить и укоротить время разработки витрин данных.
5. **Выгрузка в целевую систему** —финальный этап. Тут происходит формирование витрин данных и передача их либо в чистом виде в систему-приемник, либо трансформация в тот формат, который необходим для успешной интеграции с системой-источника.

ИНСТРУМЕНТЫ



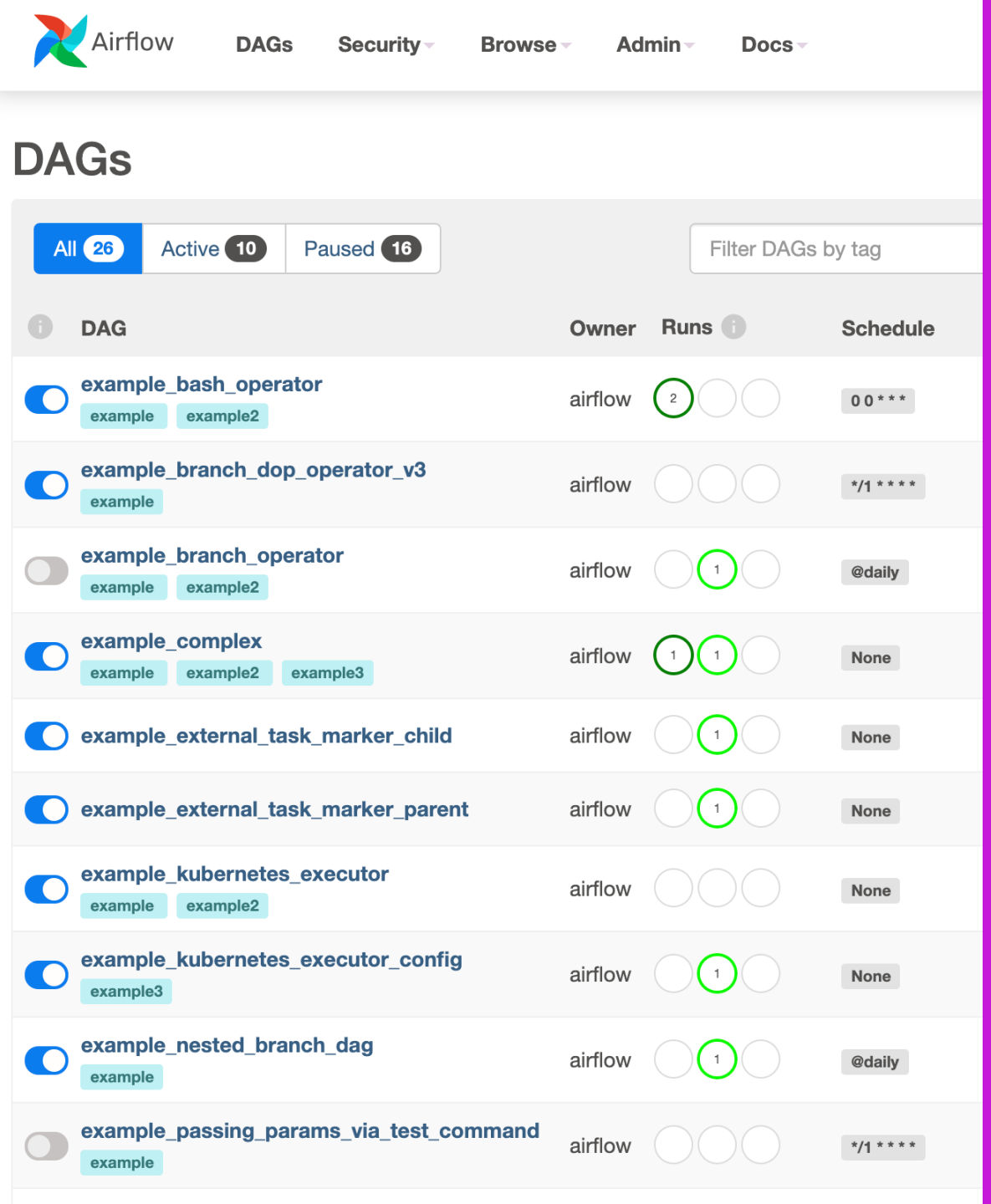
APACHE SPARK

- Написан на Scala
- Универсальный для обработки данных
- Есть коннекторы ко всем популярным БД



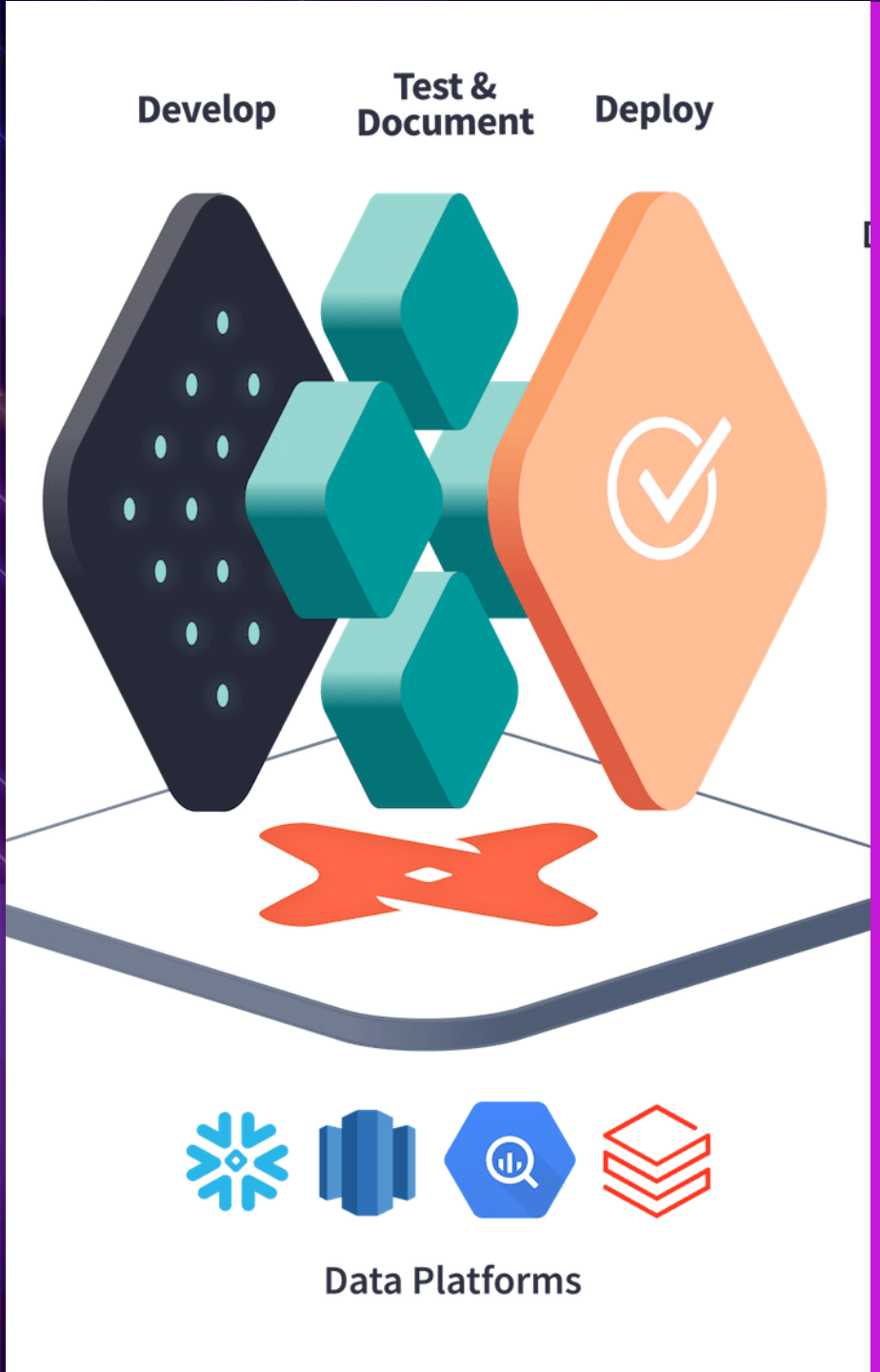
APACHE NIFI

- Написан на Java
- Программирование с помощью соединения «Процессоров»
- Удобен для перетаскивания данных, но без манипуляций



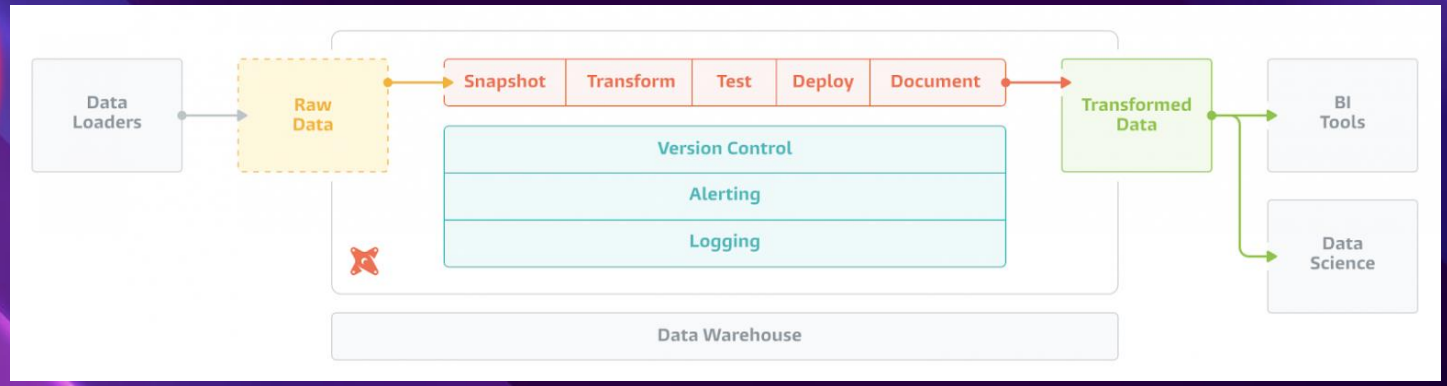
APACHE AIRFLOW

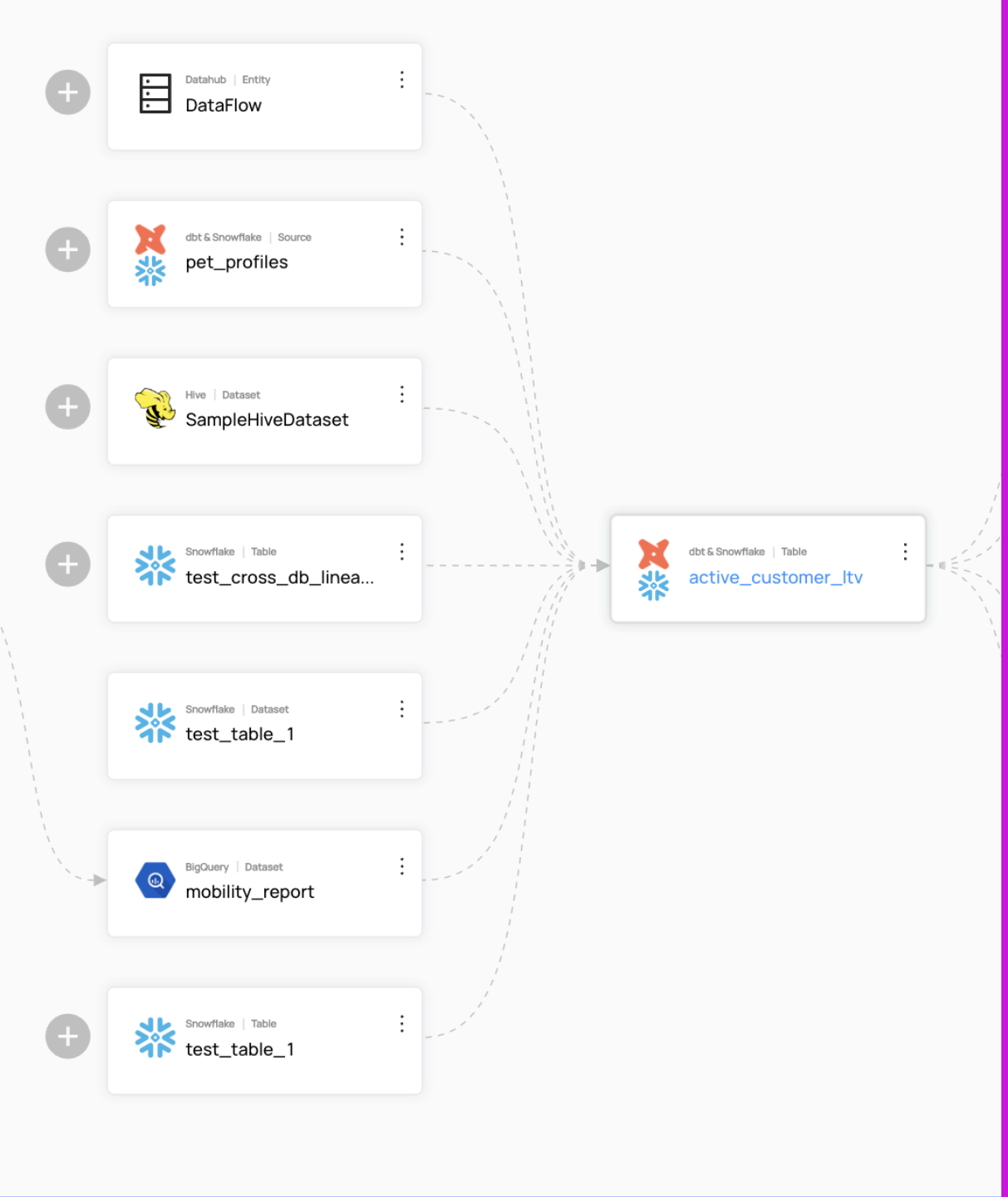
- Был рожден в AirBnb
- Код пишется на Python
- Множество готовых операторов
- Стандарт индустрии для оркестрации



DBT (DATA BUILD TOOL)

- Витрины данных по методологии с GitOps
- Порядок из коробки
- Ускорение разработки тестирования и доставки витрин
- Поиск зависимостей в витринах данных





LINKEDIN DATAHUB

- Понимание как строятся витрины
- Интеграция с многими БД
- Интеграция с Airflow, DBT
- Data Lineage

ЗАДАНИЕ

Задача — спроектировать корпоративное хранилище данных на примере кейса с рекомендательной системой онлайн-кинотеатра.

Подробное описание смотрите под QR

Решения — отправляйте на почту **stagi@1t.ru**

СПАСИБО ЗА ВНИМАНИЕ

АЛЕКСАНДРОВ АНТОН

Head of Big Data Platform

Telegram : @BioQwer