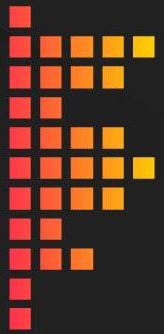


Data Warehouse Analyst

Инструменты для выгрузки данных

otus.ru

Проверить, идет ли запись

Меня хорошо видно && слышно?

Ставим “+”, если все хорошо “-”, если есть проблемы

Тема вебинара

Инструменты для выгрузки данных

Михаил Гаркунов

Senior / Team Lead Marketing Analyst

Опыт:

- 6 лет в дата аналитике / 11 лет в Digital (web/app) аналитике.
- 18 лет опыта работы в маркетинге / продажах.
- 12 лет управленческого опыта.

Телефон / эл. почта / соц. сети:

- https://t.me/mgarkunov

Правила вебинара

Активно
участвуем

Off-topic обсуждаем
в учебной группе

Задаем вопрос
в чат или голосом

Вопросы вижу в чате,
могу ответить не сразу

Условные обозначения

Индивидуально

Время, необходимое
на активность

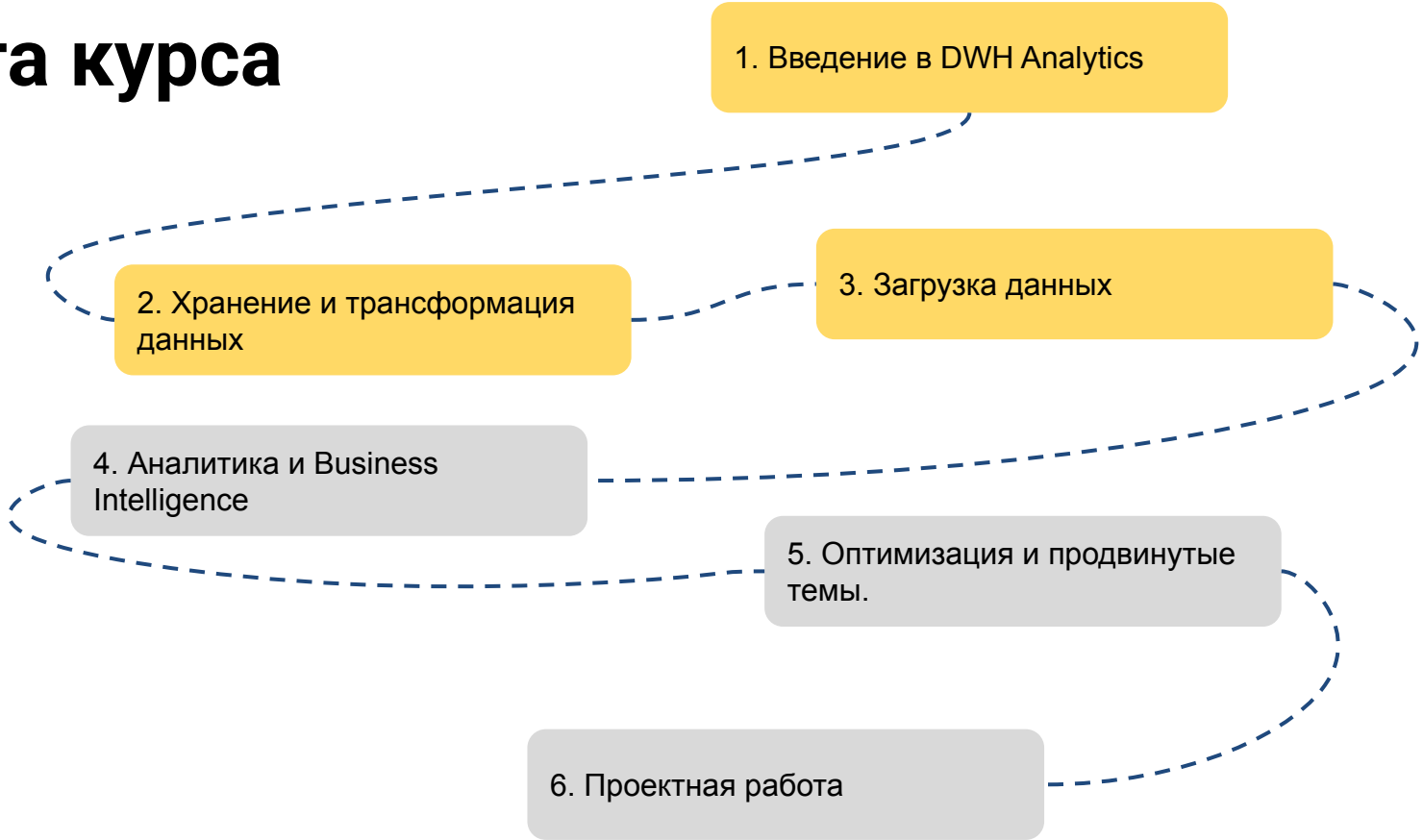
Пишем в чат

Говорим голосом

Документ

Ответьте себе или
задайте вопрос

Карта курса


1. Введение в DWH Analytics
2. Хранение и трансформация данных
3. Загрузка данных
4. Аналитика и Business Intelligence
5. Оптимизация и продвинутые темы.
6. Проектная работа

Маршрут вебинара

Знакомство

Обзор типов источников данных

Инструменты для выгрузки данных

Критерии выбора ETL инструмента

Рефлексия

Цели вебинара

К концу занятия вы сможете

1. Понимать методики загрузки / выгрузки данных.
2. Обзор основных инструментов для загрузки / выгрузки данных.
3. Оценить применение инструментов для своего проекта.

- Какие инструменты выгрузки / загрузки данных вы знаете?
- Есть опыт работы по выгрузке / загрузке данных?

Обзор типов источников данных

Цель DWH

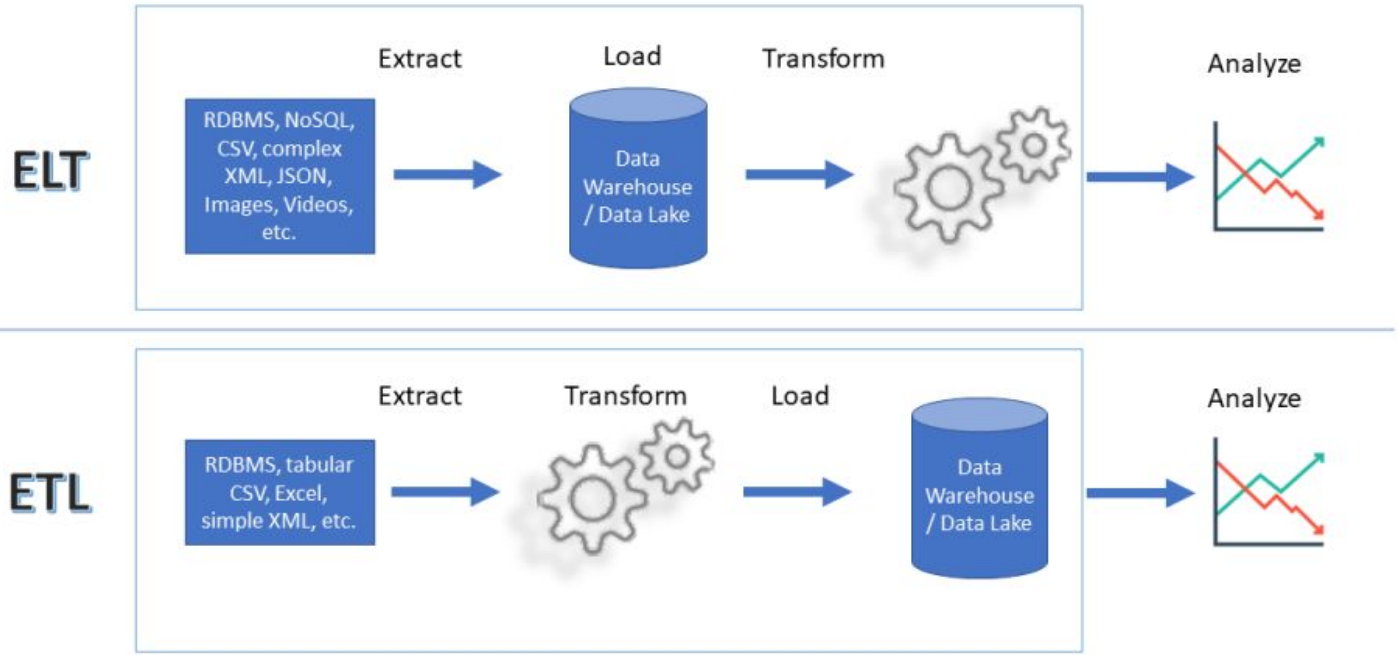
Создать единое хранилище данных для бесшовного / сквозного слоя данных. Источниками данных являются корпоративные информационные системы: ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), SCM (Supply Chain Management), HRM (Human Resource Management), ECM (Enterprise Content Management), EAM (Enterprise Asset Management) и т.д.

IT платформа

- Внутренний контур - программно - аппаратная платформа внутри корпоративной сети компании.
- Внешний контур - программно - аппаратная платформа во вне корпоративной сети компании, т.е. сервисы партнеров / поставщиков и сторонних сервисов.

* Облачные сервисы - часть внутренней корпоративной сети компании, так как обычно используют VPN каналы.

ELT vs ETL


ELT
Extract
Load
Transform
Analyze
RDBMS, NoSQL, CSV, complex XML, JSON, Images, Videos, etc.
Data Warehouse / Data Lake
ETL
Extract
Transform
Load
Analyze
RDBMS, tabular CSV, Excel, simple XML, etc.
Data Warehouse / Data Lake

Типы данных

- Структурированные данные - дамп / снапшот базы данных, выгрузки из КИС / баз данных (csv, электронные таблицы и т.д.)
- Слабо структурированные данные - JSON, XML, Avro, Parquet, данные логов, данные из NoSQL баз данных.
- Неструктурированные данные - Текстовые документы (Word, PDF, электронные письма), изображения, аудиофайлы, видеофайлы, записи голосовых звонков, посты в социальных сетях без явной разметки.

What is the Difference Between

Data Warehouse

Data Lake

Data Lakehouse

ETL для структурированных данных

- Загрузка в реляционные базы: PostgreSQL/Greenplum, Oracle, MS SQL и т.д.
- Загрузка в NoSQL максимально приближенные к реляционным базам: Clickhouse, Greenplum, Vertica и т.д.
- Использование Data Lake как системы холодного хранения.

ETL для слабо структурированных данных

- Загрузка в реляционные базы данные в текстовом формате полей и/или специальных полей при поддержки в СУБД, например: JSON в PostgreSQL.
- Загрузка в NoSQL базы данных (MongoDB, Cassandra, Redis).
- Использование Data Lake как системы горячего / холодного хранения.

ELT для неструктурированных данных

- Загрузка данных в Data Lake. DL - основное хранилище данных для таких файлов.
- Процессинг / трансформация неструктурированных данных в структурированный формат с помощью ML / DS инструментов.
- Загрузка уже структурированных данных в “горячий слой” - базы данных.

Историчность данных в ETL

- Полная история изменений / SCD (Slowly Changing Dimension) - полная история всех изменений от первичного внесения и до текущего момента.
- Частичная история изменений - история изменений за определенный период и/или определенных элементов данных.
- Отсутствие исторических данных, т.е. данные можно получить здесь и сейчас. Завтра данные уже будут на завтра…

Методы получения данных

- **Дамп / снапшот базы данных / Slave / зеркало базы данных. ***
- API (Application Programming Interface) - интерфейс для M2M “коммуникации”.
- Автоматизированные выгрузки на (S)FTP / S3 и другие файловые системы.
- Парсинг данных.
- Ручные выгрузки.

* - не нужно загружать данные из рабочей СУБД, так как обычно это OLTP СУБД. И тем более не стоит строить DWH на рабочей СУБД

Вопросы?

Ставим “+”,
если вопросы есть

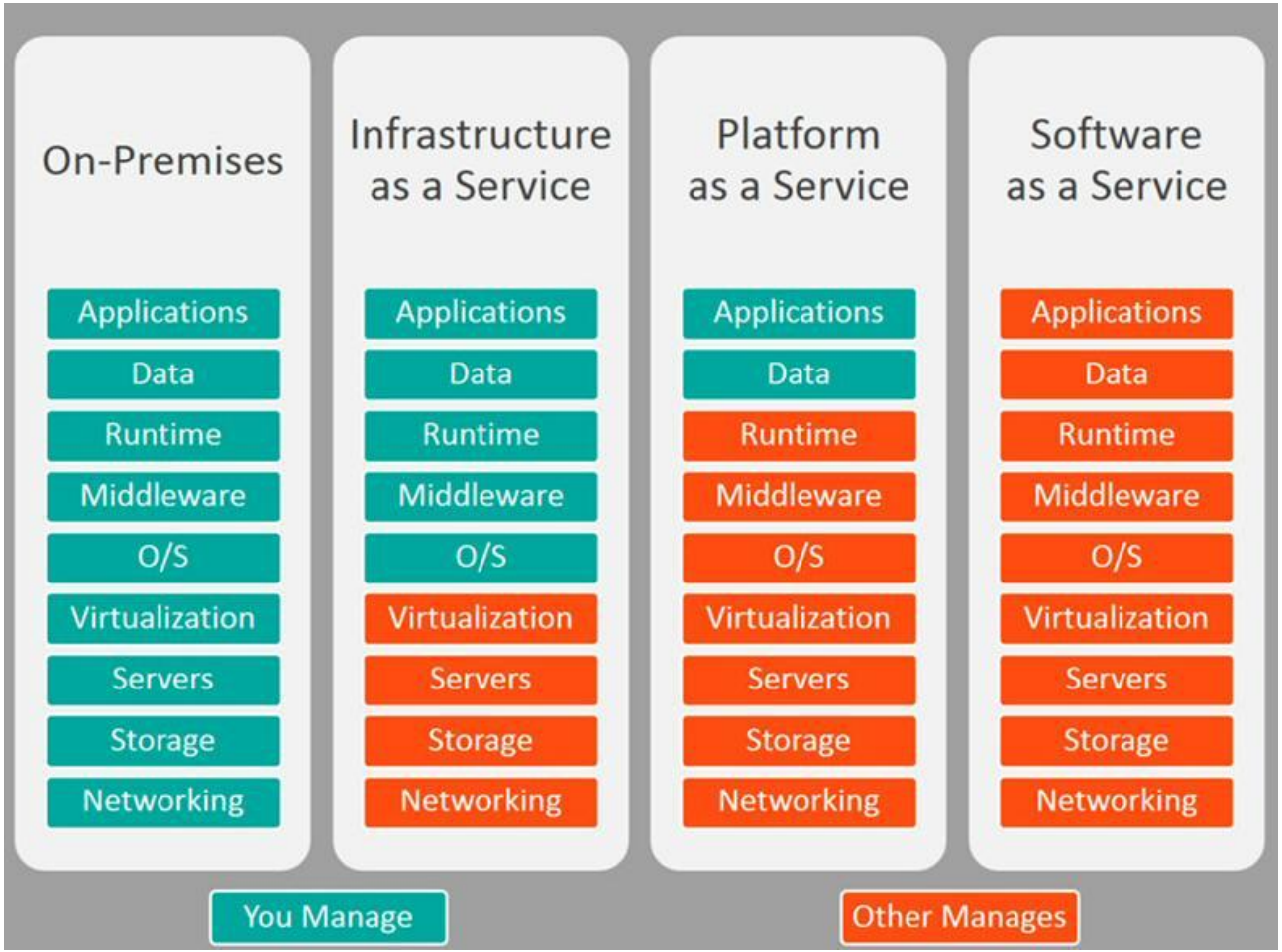
Ставим “–”,
если вопросов нет

Инструменты для выгрузки данных

Категории / По типу развертывания

- On-Premise (Локальные решения) - инструменты устанавливаются и управляются на собственной инфраструктуре компании.
  Например: Oracle Data Integrator, Apache Airflow

- Cloud-Native / IaaS, PaaS, SaaS (Облачные решения) - инструменты полностью управляются провайдером и предоставляются как сервис (SaaS).
  Например: Yandex Managed Service for Apache Airflow, mybi connect.


On-Premises
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
Infrastructure as a Service
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
Platform as a Service
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
Software as a Service
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
You Manage
Other Manages

Категории / По подходу к разработке

- GUI-driven / Low-code / No-code (С графическим интерфейсом) - инструменты позволяют создавать ETL-процессы, используя визуальные дизайнеры, перетаскивание компонентов и минимальное количество кода.
  Например: Informatica PowerCenter, Apache NiFi.

- Code-driven / Scripting (Ориентированные на код) - инструменты требуют написания кода на языках программирования для реализации ETL-логики.
  Например: Apache Airflow, Apache Spark.

Категории / По масштабу и сложности данных

- Batch Processing (Пакетная обработка) - инструменты, которые собирают данные в пакеты и периодически загружают данные. Основной тип ETL инструментов.
  Например: Apache Airflow, Apache NiFi.

- Stream Processing / Real-time (Потоковая обработка / В реальном времени) - инструменты, предназначенные для непрерывной обработки данных по мере их поступления, с минимальной задержкой.
  Например: Apache Kafka, Apache Flink.

- Big Data Processing (Обработка больших данных) - инструменты, оптимизированные для работы с очень большими объемами данных (петабайты) в распределенных средах.
  Например: Apache Spark, Hadoop Ecosystem.

Категории / По открытости и лицензированию

- Open Source (Открытый исходный код) - инструменты бесплатны для использования и дают возможность сообществу вносить вклад в их развитие.
  Например: Apache Spark, Apache NiFi, Apache Airflow, Talend Open Studio, Airbyte, Pentaho Data Integration (Kettle), Debezium.

- Commercial (Коммерческие решения) - инструменты предлагаются вендорами, часто с платными лицензиями и подписками. Обычно включают профессиональную поддержку, регулярные обновления и более широкий набор функций.
  Например: Informatica PowerCenter, IBM DataStage, SSIS, Fivetran, Stitch, Matillion, Azure Data Factory, AWS Glue, Google Cloud Dataflow.

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Критерии выбора ETL инструмента

1. Кто будет работать с ETL инструментом?

- Бизнес заказчики, аналитики данных? >> Выбираем Low-code / No-code.
  Пример: Альбато (albato.ru), Loginom (loginom.ru), mybi connect (connect.mybi.ru) и т.д.

- Опытные аналитики? >> Выбираем GUI системы.
  Пример: Airbyte, Apache Nifi и т.д.

- Есть инженеры данных? >> Выбираем сложные инструменты.
  Пример: Apache Airflow, Apache Spark и т.д.

2. Готовы к покупке?

- Нет? >> Используем Open Source системы.
  Пример: Airbyte, Apache Nifi, Apache Airflow и т.д.

- Да? >> Покупаем российские системы.
  Пример: Arenadata (arenadata.tech)

3. Есть DevOps?

DevOps - это методология разработки программного обеспечения, которая объединяет процессы разработки (Development) и эксплуатации (Operations) для обеспечения более быстрой, надежной и автоматизированной поставки программных продуктов

- Есть DevOps? >> Разворачиваем на собственной платформе.

- Нет DevOps? >> Используем облачные сервисы и “Managed Service”.
  Пример: Yandex Managed Service for Apache Airflow, VK Cloud Streams, VK Cloud Kafka и т.д.

4. Требуется обработка больших данных?

- Да? >> Требуется опытные DE / DS / ML специалисты.
- Нет? >> Можно выбирать простые сервисы / инструменты.

Итоговый выбор

Тип ETL инструментов	Условно бесплатные системы	Платные системы
Простые сервисы		• Альбато (albato.ru), • Loginom (loginom.ru), • mybi connect (connect.mybi.ru)
Продвинутые системы	• AirByte (airbyte.com) • Talend Open Studio (talend.com)	• Cloud Managed Service севрисы
Профессиональные системы	• Apache Airflow (airflow.apache.org) • Apache NiFi (nifi.apache.org) • Apache Flink (flink.apache.org)	• Cloud Managed Service севрисы • Arenadata (arenadata.tech)
Big Data системы	• Apache Spark (spark.apache.org) • Фзфсру Hadoop (hadoop.apache.org)	• Cloud Managed Service севрисы • Arenadata (arenadata.tech)

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Рефлексия

Ключевые тезисы

1. Понимать методики загрузки / выгрузки данных.
2. Обзор основных инструментов для загрузки / выгрузки данных.
3. Оценить применение инструментов для своего проекта.

Рефлексия

С какими впечатлениями уходите с вебинара?

Как будете применять на практике
то, что узнали на вебинаре?

Следующий вебинар

19 июня 2025

Введение в оркестрацию

- Ссылка на вебинар будет в ЛК за 15 минут
- Материалы к занятию в ЛК — можно изучать
- Обязательный материал обозначен красной лентой

Заполните, пожалуйста, опрос о занятии по ссылке в чате

Спасибо за внимание!

Приходите на следующие вебинары

Михаил Гаркунов

Senior / Team Lead Marketing Analyst

Опыт:

- 6 лет в дата аналитике / 11 лет в Digital (web/app) аналитике.
- 18 лет опыта работы в маркетинге / продажах.
- 12 лет управленческого опыта.

Телефон / эл. почта / соц. сети:

- https://t.me/mgarkunov