

Data Warehouse Analyst

Что такое DWH



Проверить, идет ли запись

Меня хорошо видно && слышно?



Ставим "+", если все хорошо
"-", если есть проблемы



Тема вебинара

Принципы построения DWH



Водовозова Татьяна

Руководитель направления аналитики DWH

Об опыте:

- участвовала в проектах развития хранилищ данных,
- построение хранилищ с нуля,
- миграция данных
- [ментор](#)

Правила вебинара



Активно
участвуем



Off-topic обсуждаем
в Telegram @OTUS DWH-2025-04



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

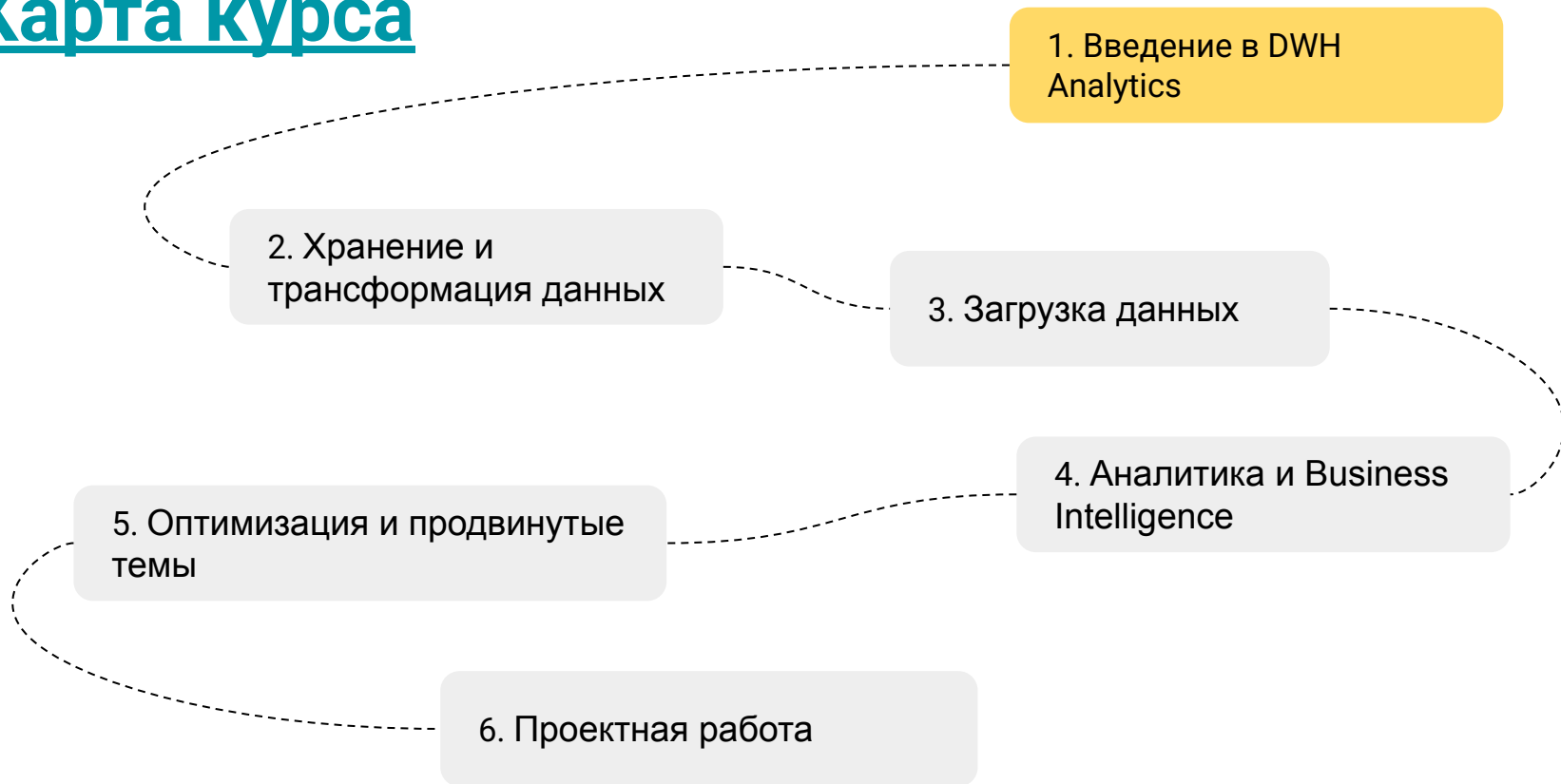


Документ



Ответьте себе или
задайте вопрос

Карта курса



Цели вебинара

К концу занятия вы сможете

1. Узнать какие системы бывают
2. Определить для чего бизнесу нужен DWH
3. Ответить на вопрос, что такое DWH, как понять, что перед нами именно он

Смысл

Зачем вам это уметь

1. Понимать, каким образом DWH нам (бизнесу) помогает
 2. Понимать над чем мы будем работать в рамках этого курса
-

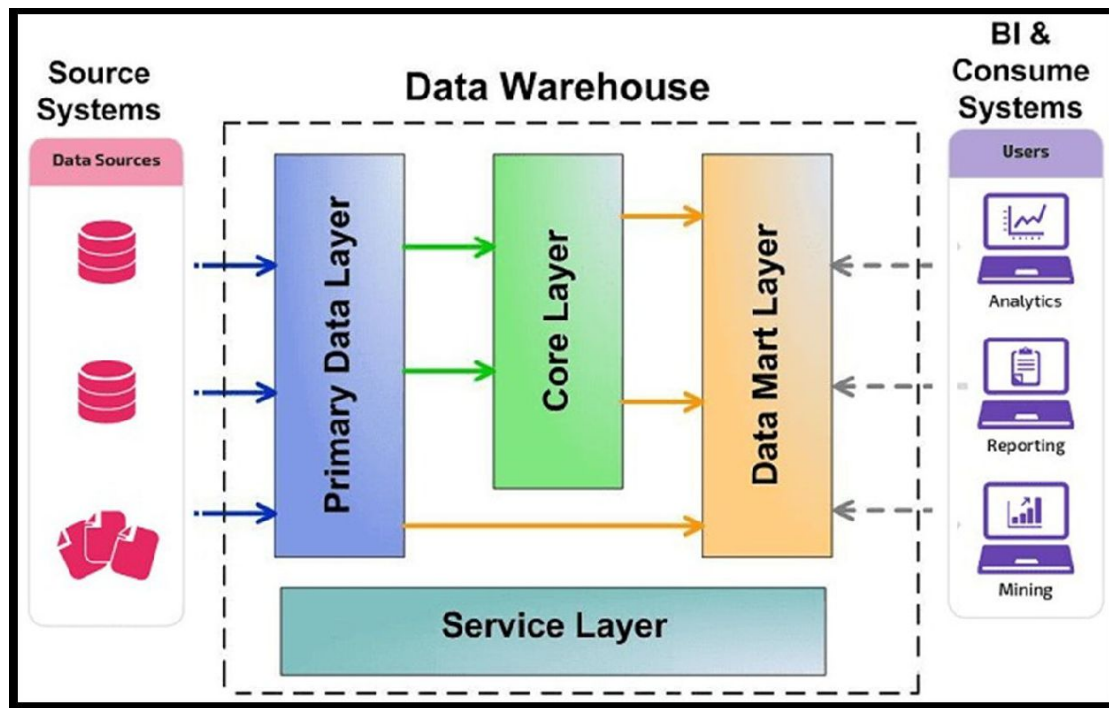
Что же такое DWH?

Data Warehouse - единое корпоративное хранилище данных из разных источников

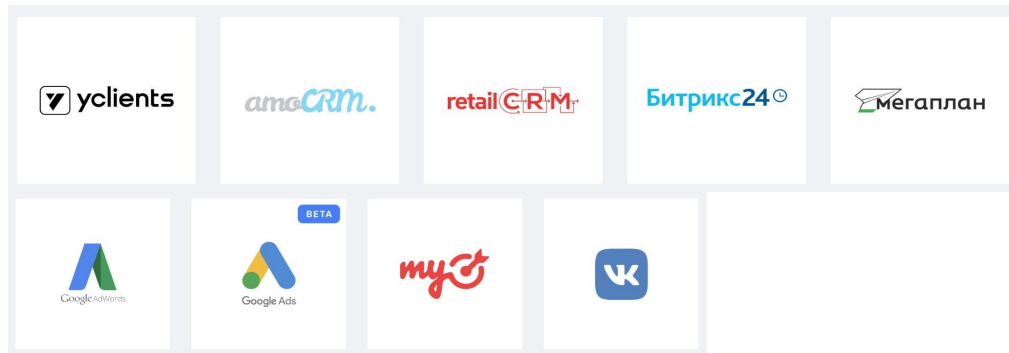
Характеристики:

- Интегрирует в себе данные из различных источников
- Поставляет бизнесу подготовленные для анализа данные для принятия решений
- Позволяет хранить и использовать исторические данные

Основные компоненты DWH



Какие бывают источники



ЯндексМетрика



Типы данных источников

Структурированные данные

- есть описание схемы
- явно описаны типы данных
- необходима передача схемы при её изменении
- явно сломается при несовместимом изменении схемы

Примеры: Thrift, Avro, MsgPack

Неструктурированные данные

- нет описание схемы
- описание типов данных ложится на получателя
- неявно сломается при изменении схемы

Примеры: CSV, JSON, XML

Виды источников

Пассивный источник

- есть интерфейс, через который можно запросить данные
- требует постоянного опроса
- возможно повторное чтение в случае проблем
- позволяет контролировать скорость загрузки

Примеры: веб-сервис с HTTP API, FTPсервер, RDBMS

Активный источник

- сам пишет данные
- в какое-либо хранилище (например,
- очередь сообщений)
- более эффективен с точки зрения ресурсов
- сложно или невозможно повторить загрузку
- скорость поступления данных напрямую зависит от скорости их генерации в источнике

Примеры: веб-сервис, пишущий в Kafka,

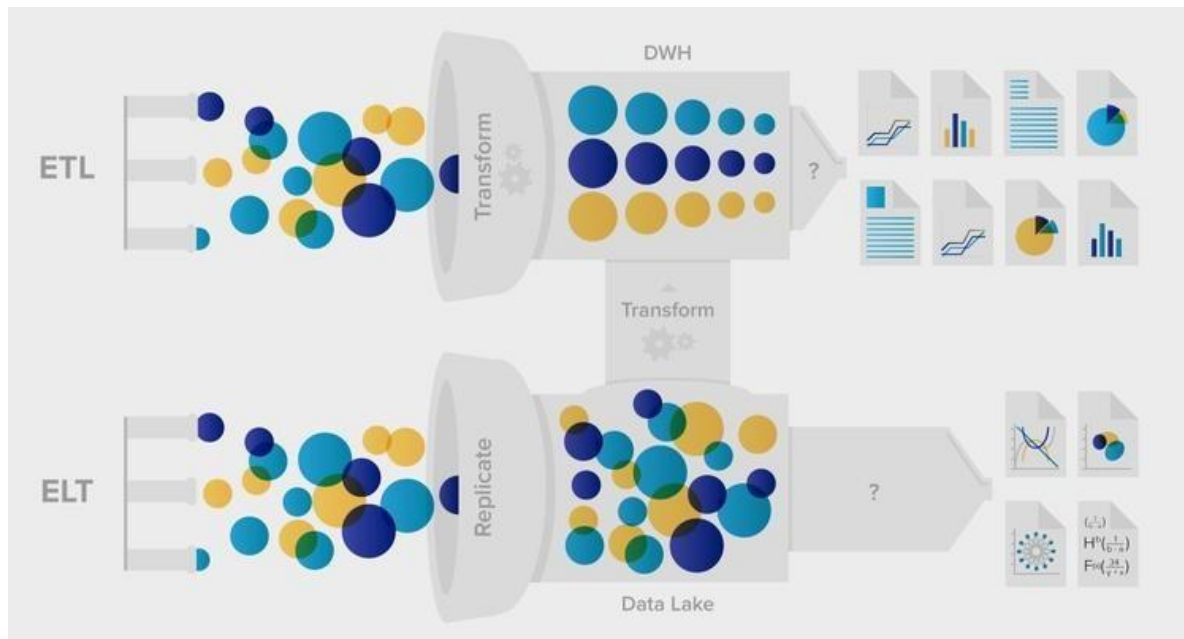
Отличие DWH от других решений

- Транзакционная БД – предназначена для повседневной работы, обновления real time. Уже потом данные с определенной периодичностью поступают в DWH
- Data Lake – данные поступают и хранятся в необработанном виде без какой-либо структуры

Data Lake и Data Warehouse

Характеристика	DL	DWH
Структура данных	Хранит данные в сыром, слабо структурированном формате (тексты, изображения, JSON и т.д.)	Хранит структурированные данные, обычно в табличном формате (например, реляционные базы данных)
Схема	Подход "схема по мере необходимости" (schema-on-read): структура данных определяется в момент извлечения	"схема по мере записи" (schema-on-write): структура данных определяется перед загрузкой
Цели использования	Максимально гибкий анализ данных. Помогает исследователям и аналитикам экспериментировать с данными, тестировать гипотезы	Оптимизирован для обработки и анализа структурированных данных, предназначен для отчетности и бизнес-аналитики
Производительность	Меньшая производительность при выполнении сложных запросов	Высокая производительность для анализа и отчетности, благодаря оптимизации для запросов и модели данных
Пользователи	DS, аналитиками и специалистами по большим данным для глубокого анализа и экспериментов	фокус на бизнес-пользователях и руководителях, которым отчеты нужны регулярно

ETL/ELT



Как понять, что перед нами DWH

- Данные очищены и интегрированы из нескольких источников
- Данные организованы в структуре, подходящей для аналитических расчетов
- Вся необходимая информация по всем направлениям бизнеса в едином месте
- Продолжительная сохранность данных (историчность)
- Бизнес-аналитика не влияет на другие процессы и системы – работаем только с данными в DWH
- Модель данных построена строго в соответствии с сущностями бизнеса
- Бизнес доверяет данным

Вопросы?



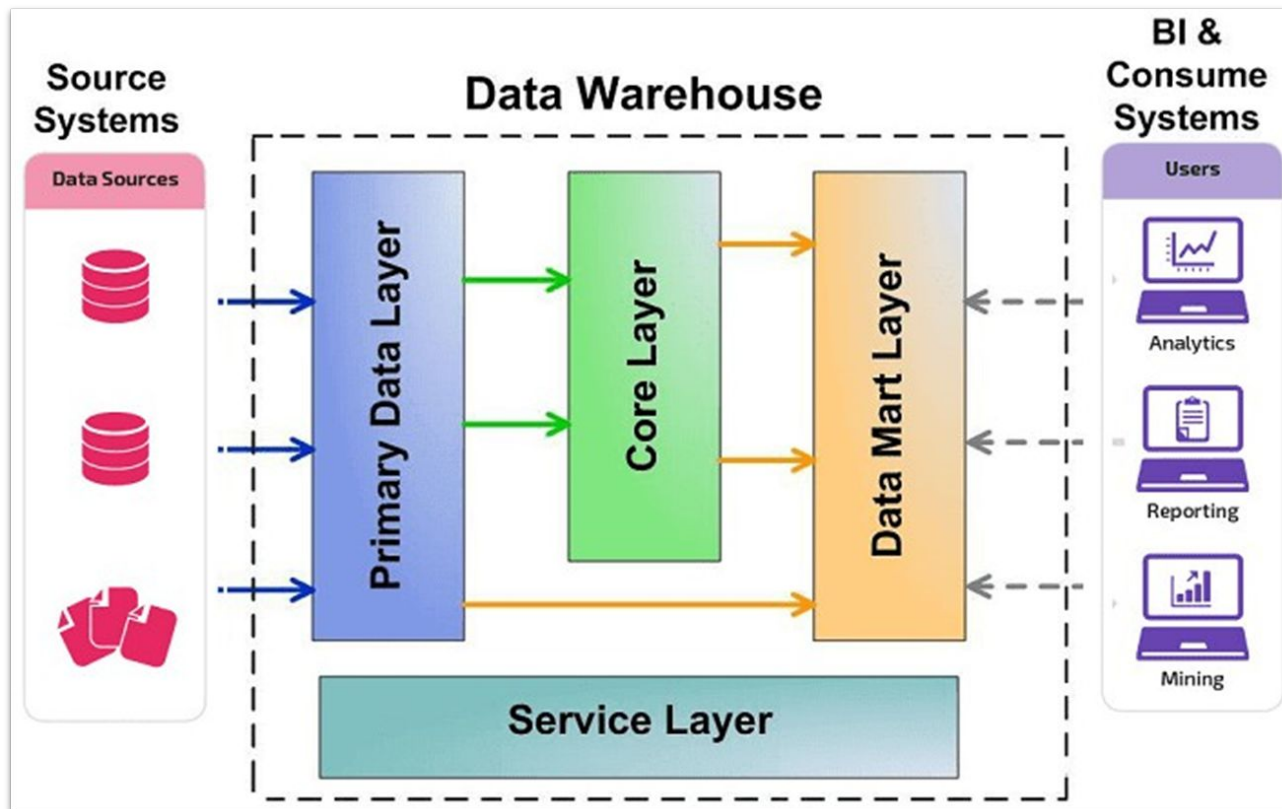
Ставим “+”,
если вопросы есть



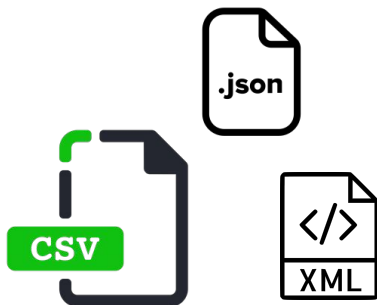
Ставим “-”,
если вопросов нет

Примеры

Структура DWH



Распространенные форматы хранения данных (наши источники)



OLTP Databases (Backend)



MongoDB Backend ⚙️

MongoDB • Ingests in Real-time



financial service ⚙️

PostgreSQL • Ingests in Real-time

🔍 table_name 🔼🔼	📊 used_mb 🔼🔼	📊 num_rows 🔼🔼
quotes	59,337	496,594,319
orders	11,944	12,639,669
device_infos	3,358	15,457,275
status_info	3,204	58,113,878
eta_measurements	2,928	46,471,590
routes	2,808	6,218,755
status_sessions	2,356	37,537,520
users	2,074	1,726,506
transactions	2,040	15,862,806
orders_reload	2,032	3,793

Applications – CRM / ERP

Performance Marketing



BETA




Event collectors



id	ajs-f9e9994bf2cdd88deb096b039cd66b91
received_at	2019-01-31T01:36:04.654+00:00
anonymous_id	2f0d3faa-f996-4c9a-b87d-7595291b4db3
context_page_title	Republic — invest in startups on a leading investment platform
context_user_agent	Mozilla/5.0 (iPhone; CPU iPhone OS 12_1_2 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/12.0 Mobile/15E148 Safari/604.1
sent_at	2019-01-31T01:36:04.000+00:00
context_library_name	analytics.js
design	Offering card
event_text	Click "Offering" link
timestamp	2019-01-31T01:36:04.653+00:00
user_id	∅
context_ip	172.58.227.205
context_library_version	3.7.2
context_page_path	/
context_page_referrer	https://angel.co/investor/new
context_page_url	https://republic.co?utm_source=angellist&utm_medium=promo&utm_campaign=investor_onboarding_unaccredited
original_timestamp	2019-01-31T01:36:04.549+00:00
event	click_offering_link
uuid_ts	2019-01-31T12:21:41.000+00:00
context_page_search	?utm_source=angellist&utm_medium=promo&utm_campaign=investor_onboarding_unaccredited
context_campaign_name	investor_onboarding_unaccredited



Логи

**kibana**

Discover

Visualize

Dashboard

Timelion

Dev Tools

Management

logfmt.addr

logfmt.msg

logfmt.path

logfmt.status

logfmt.user_agent

@timestamp

@version

_id

_index

_score

_type

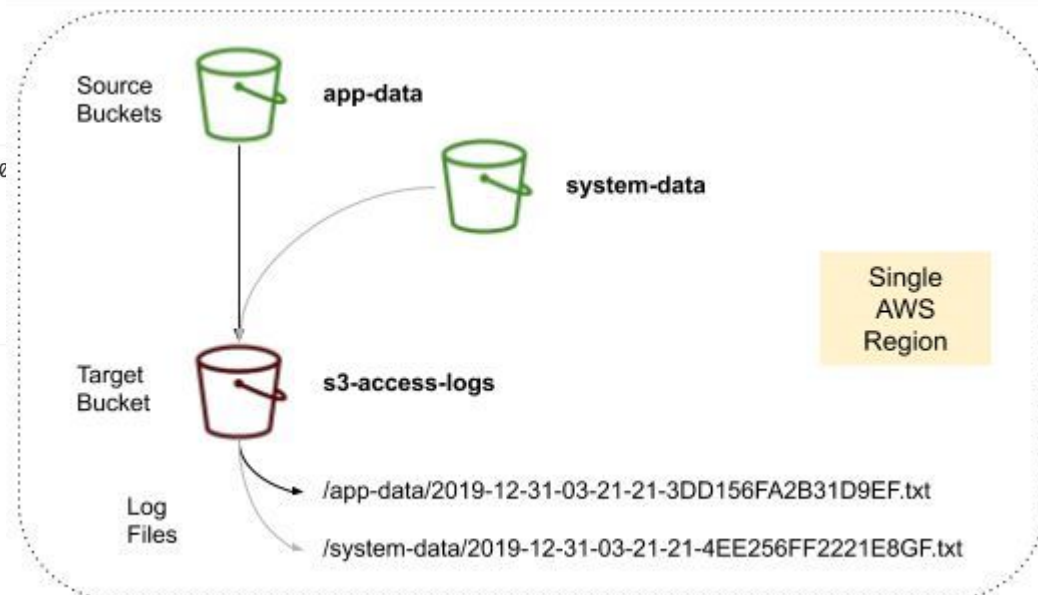
ec2.availability_zone

ec2.dc

ec2.env

Time ^	ec2.name	message	severity_label
2021-07-21 03:00:00.111	kubernetes-staging-worker0	time=2021-07-21T00:00:00.111Z level=info msg= logger=server path=/_health len=0 addr=10.3.10.199:60950 duration=9.052e-06 method=GET host=10.233.77.13:8080 query= status=200	-
2021-07-21 03:00:00.276	kubernetes-staging-	time=2021-07-21T00:00:00.276+00:00 level=info method=PUT path=/v6/auth/auth/activate_saml_status=200 len=2662	-

2021-07-21 03:0



Сторонние API

Examples

HTTP · jQuery

```
https://openexchangerates.org/api/latest.json?app_id=YOUR_APP_ID
```


Result Format

• 200 OK


```
{
  disclaimer: "https://openexchangerates.org/terms/",
  license: "https://openexchangerates.org/license/",
  timestamp: 1449877801,
  base: "USD",
  rates: {
    AED: 3.672538,
    AFN: 66.809999,
    ALL: 125.716501,
    AMD: 484.902502,
    ANG: 1.788575,
    AOA: 135.295998,
    ARS: 9.750101,
    AUD: 1.390866,
    /* ... */
  }
}
```

Files and Object Storage




#33

 **Braze Current – Prod Chauffeurs** ⚙️
S3 • Ingests every 5 minutes

→

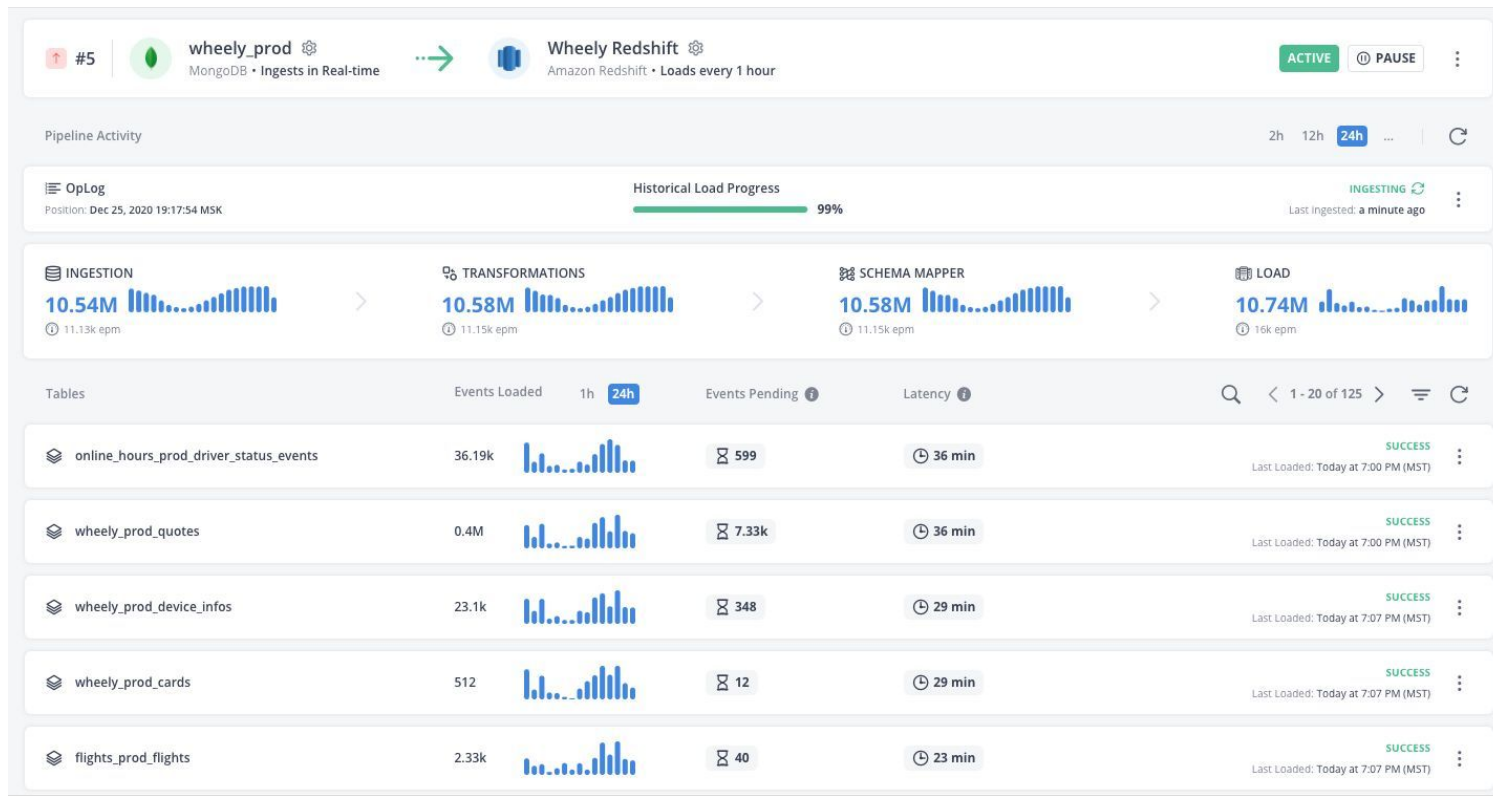
 **Wheely Redshift** ⚙️
Amazon Redshift • Loads every 3 hours

ACTIVE **PAUSE** ⋮

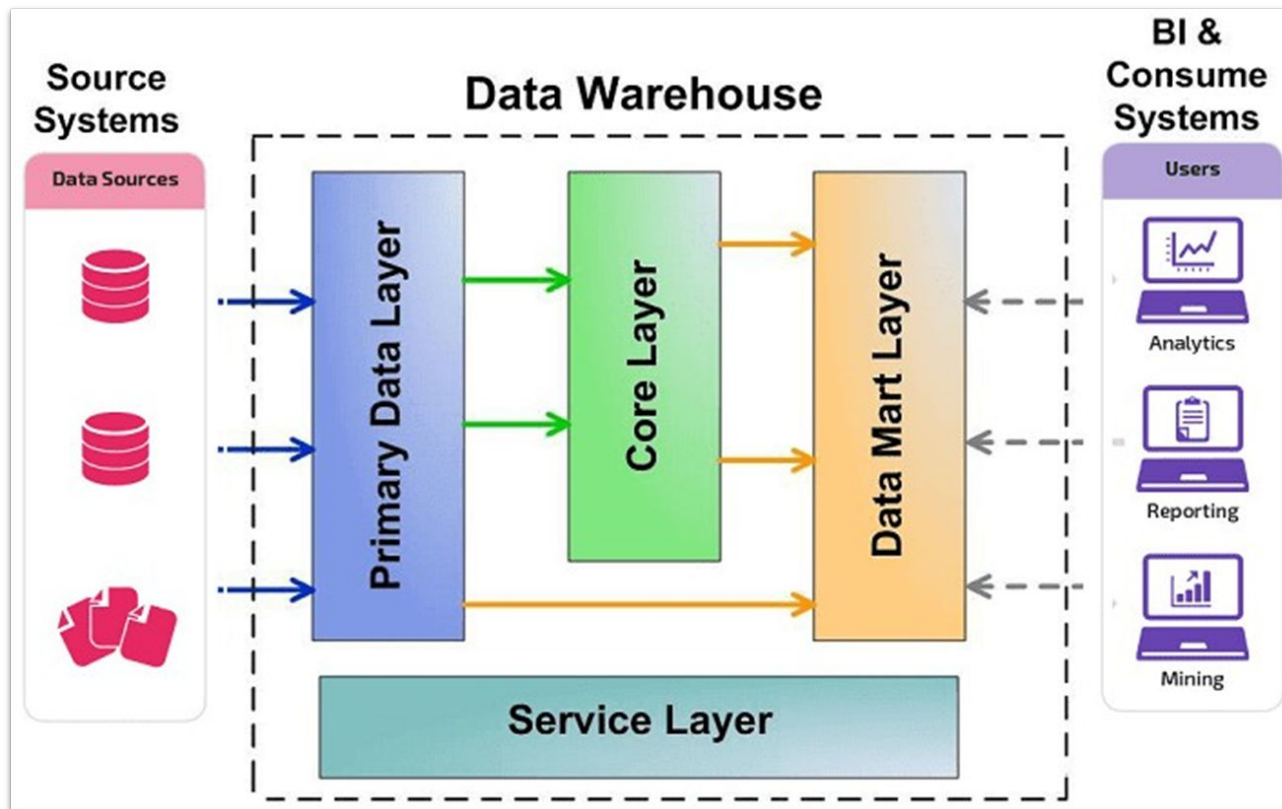
Files	Started At	File Size	
 braze/currents/dataexport.prod-01.S3.integration.60b9...	Today at 4:53 PM (MST)	2.49 KB	INGESTED a few seconds ago
 braze/	Today at 4:53 PM (MST)	3.87 KB	INGESTED a few seconds ago
 braze/currents/dataexport.prod-01.S3.integration.60b9...	Today at 4:53 PM (MST)	1.65 KB	INGESTED a few seconds ago

braze/currents/dataexport.prod-01.S3.integration.60b9ed4edb096721f97a18b4/event_type=users.messages.newsfeedcard.impression/date=2021-07-29-12/262/prod-01/dataexport.prod-01.S3.integration.60b9ed4edb096721f97a18b4+1+0000226693.avro

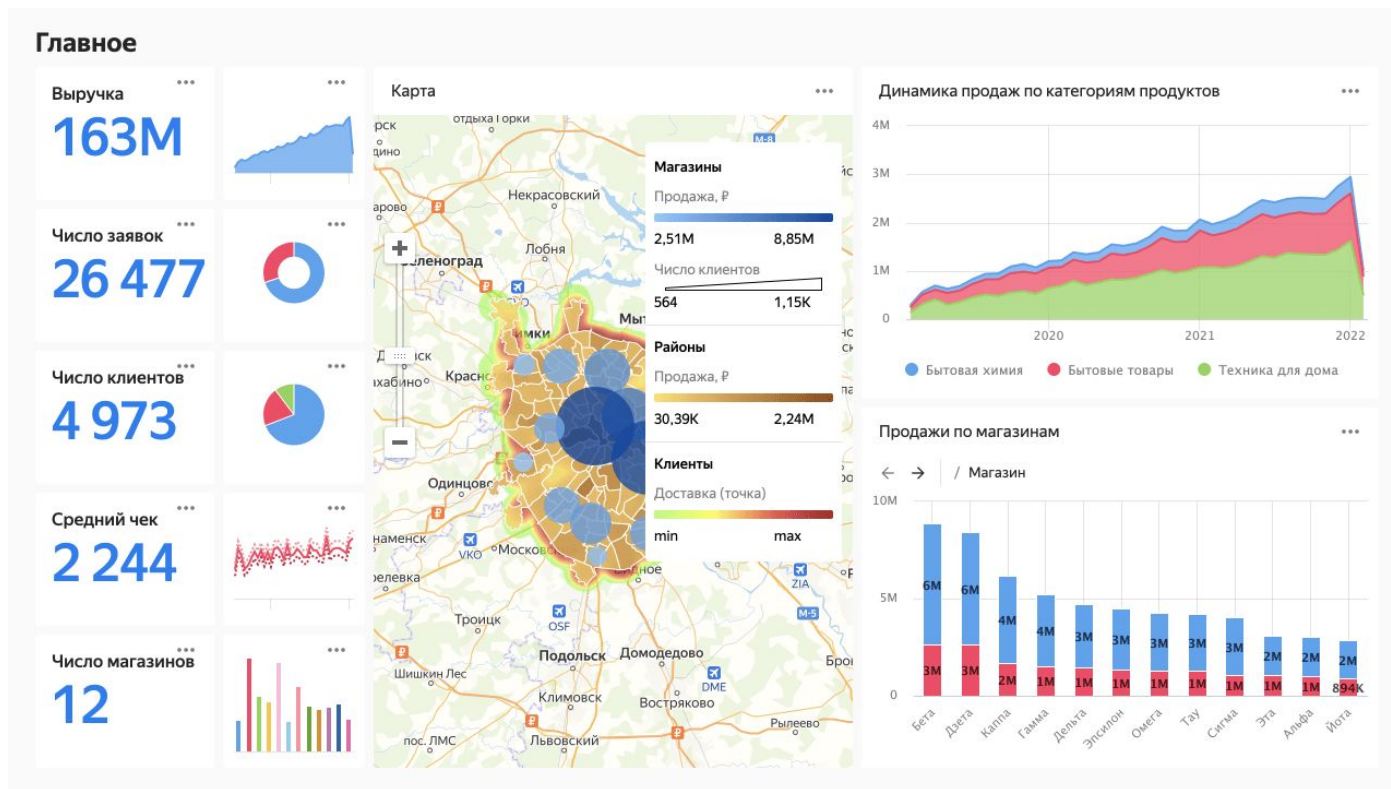
Webhooks



Структура DWH



ВІ решения



Вопросы?



Ставим “+”,
если вопросы есть



Ставим “-”,
если вопросов нет

Для чего нужно хранилище

Как считать метрики

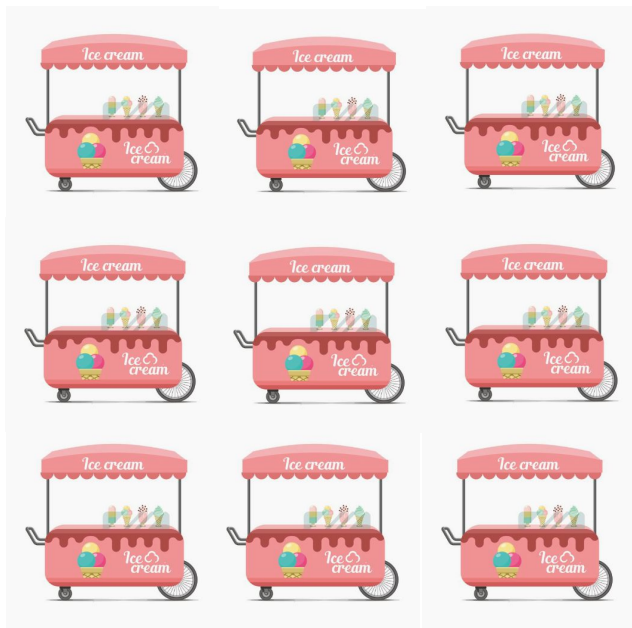


Финансовая модель ларька с мороженым:

Выручка – расходы = прибыль

За всеми операциями можно следить в приложении банка на смартфоне – этого будет достаточно для понимания состояния бизнеса

По мере роста – новые метрики и новые источники данных



Как оценить эффективность маркетинга?

Как оценить эффективность HR?

Как оценить эффективность закупок?

Как оценить эффективность производства?

Невозможно дать комплексную оценку состояния бизнеса только по приходящим и исходящим платежам. Нужно работать с данными.

Если вы – крупная корпорация



- Большое количество отделов
- Сотни систем-источников
- Тысячи бизнес-метрик
- Различные направления деятельности

Что нужно мне, как владельцу бизнеса



- Всегда получаю понятную картину в цифрах
- Знаю , что по мере роста или запуска новых направлений буду понимать что происходит
- Извлекаю прибыль из огромного объема данных который генерирует моя компания

Вопросы?



Ставим “+”,
если вопросы есть



Ставим “-”,
если вопросов нет

Список материалов для изучения

1. <https://www.dasca.org/world-of-data-science/article/what-is-a-data-warehouse-and-why-is-it-important>
2. https://habr.com/ru/companies/smartup_tech/articles/807379/

Цели вебинара

К концу занятия вы сможете

1. Узнать какие системы бывают
2. Определить для чего бизнесу нужен DWH
3. Ответить на вопрос, что такое DWH, как понять, что перед нами именно он

Рефлексия

**Заполните, пожалуйста,
опрос о занятии
по ссылке в чате**

Спасибо за внимание!