СЛЁРМ Обучение для инженеров и технических лидеров

Дата-инженер

Облачные хранилища, Modern Data Stack

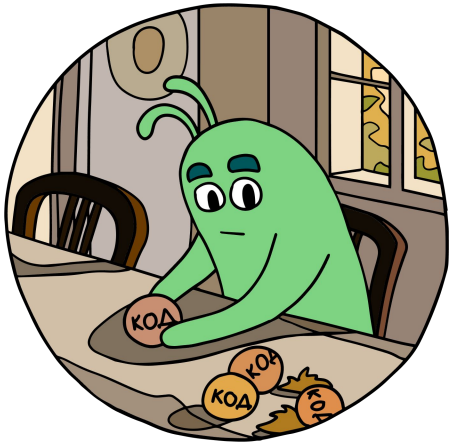
Николай Марков

Цели урока. Что вы узнаете:

1. Что такое *aaS
2. Чем именно нам помогают облака
3. Что такое Modern Data Stack
4. Как работает dbt

slurm.io

Что такое *aaS?



slurm.io

Использование облаков, как IaaS

- Дергаем ползунки – выделяем ресурсы
- Разрезаем облако на кусочки в виде виртуалок
- Виртуалки надо кому-то поддерживать, точно так же, как и bare metal (железо)
- Видите проблему?

slurm.io

Объектные хранилища

- Часто надо что-то хранить в облаке - документы, аналитические данные или просто бинарные объекты
- Плюс, из коробки часто есть интеграция с CDN, что позволяет удобнее масштабироваться

slurm.io

Безопасность и аудит

- Автозапуск тестов
- Эмуляция высокой нагрузки
- Пентестинг
- Сборка пакетов
- Выкатка новых версий

slurm.io

IoT и умный дом

- Отправка сообщений и уведомлений
- Сборка данных с группы устройств
- Триггер событий прямо на устройствах
- Кросс-платформенные приложения и прошивки

slurm.io

Что угодно as a Service

slurm.io

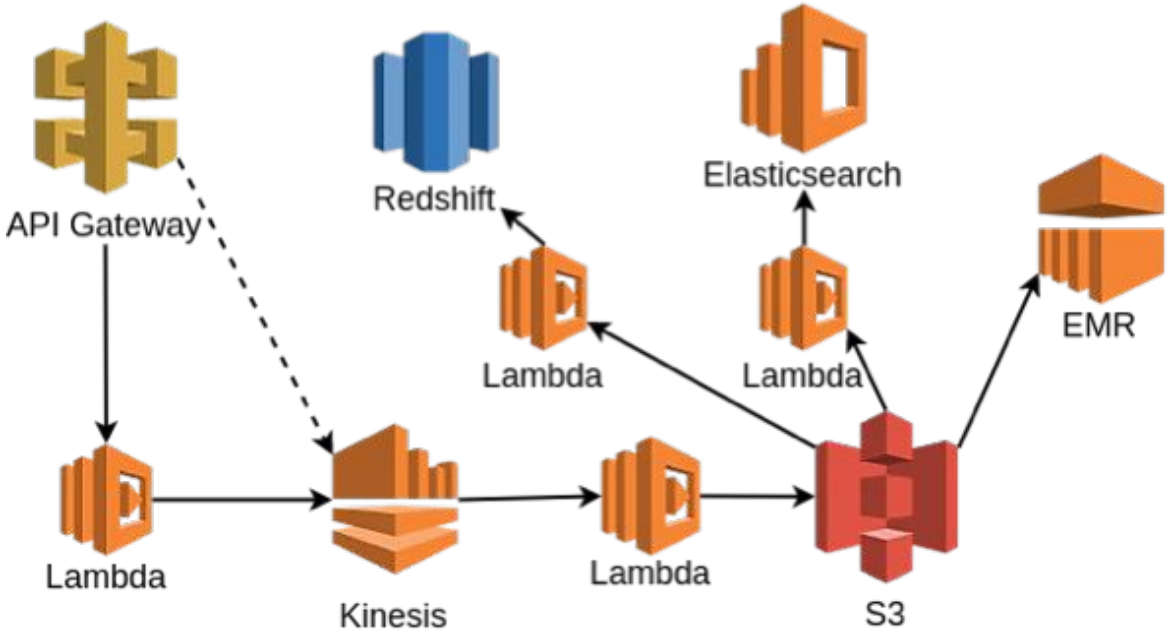
Без серверов?

Блог Мартина Фаулера

- Приложение, которое значительно или полностью **зависит от другого** стороннего **облачного** приложения или сервиса **для исполнения части бизнес-логики и управления состоянием (Backend as a Service)**.
- Кусочки бизнес-логики запускаются в **stateless контейнерах**, которые создаются только как **реакция на событие**, существуют **короткий промежуток времени** (возможно, создаются разово) и полностью управляются **сторонним сервисом (Function as a Service)**.

slurm.io

AWS Lambda



slurm.io

Как облака облегчают жизнь

slurm.io

Поддержка и бюджет

- В облаке можно платить **только за то, что вы потребляете**, что стимулирует эффективнее утилизировать ресурсы
- Соглашение с облачным провайдером подразумевает подписание **SLA (соглашение об уровне предоставления услуги)**, по которому в случае большинства проблем провайдер **обязан возместить вам убытки**
- **Time To Market** гораздо быстрее
- Крупные провайдеры предоставляют разные **программы сертификаций и гранты** для стартапов

slurm.io

Безопасность и контроль доступа

- **RBAC (Role-Based Access Control)** – механизм разделения **прав доступа** не только для людей, но и между сервисами внутри облака
- Облачные платформы имеют **сертификацию по безопасности** и регулярно проходят аудит
- Довольно популярный сервис – **выпуск и управление сертификатами**

slurm.io

Технические нюансы

Автоматическое масштабирование

Бэкапы и снятие образов

Обновления и документация

slurm.io

Выводы

Масштабирование

Скорость развертывания

Гибкость интеграции

Поддержка

Модель биллинга

Обновление и инновации

Отказоустойчивость и SLA

Нужно ОЧЕНЬ пристально следить за тратами

Пересылка данных из/в облако может быть внезапно трудоемкой и дорогой

Задержки по Latency на гибридных инфраструктурах

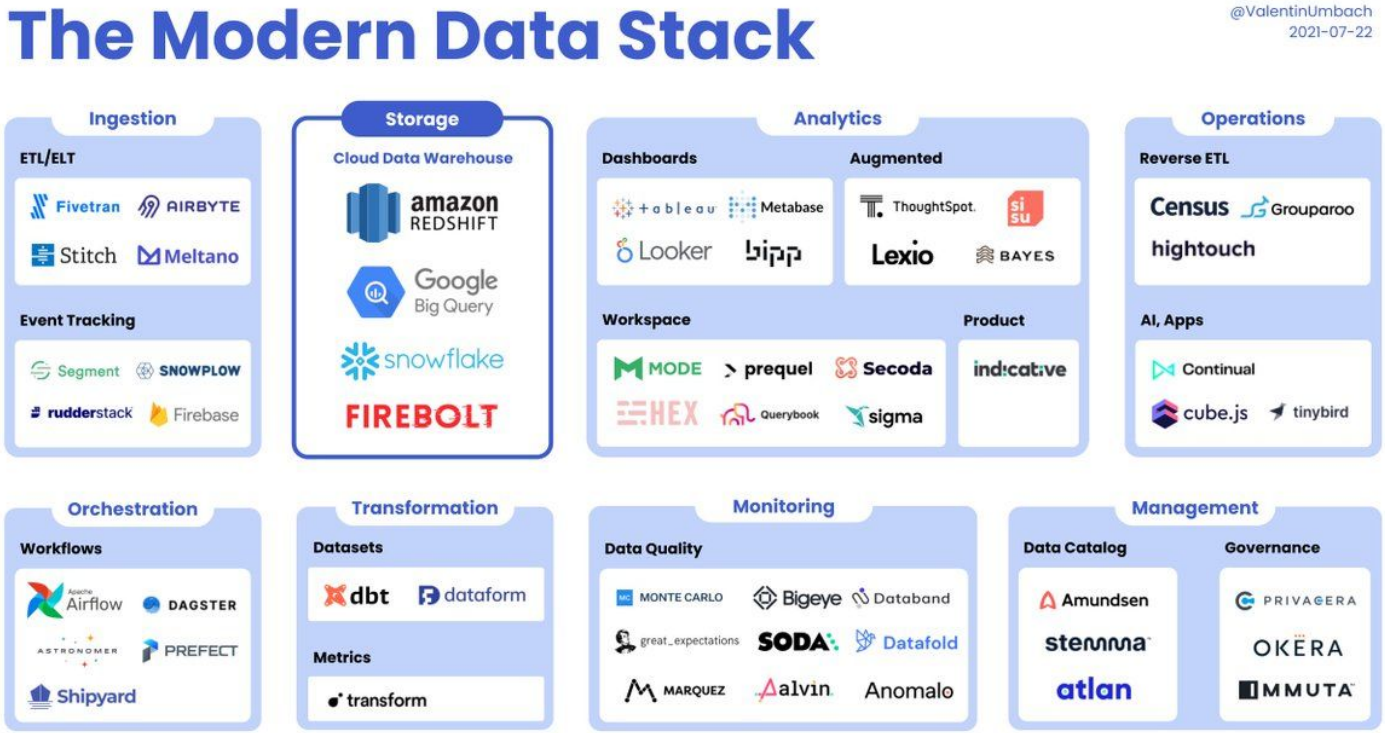
У отдела безопасности могут быть вопросы

slurm.io

Modern Data Stack

slurm.io

Тысячи их



slurm.io

Заливка данных

slurm.io

Корпоративные хранилища

- Данные в S3/ HDFS
- Масштабируется практически неограниченно
- Поддерживает SQL

slurm.io

Data Quality & BI

great_expectations

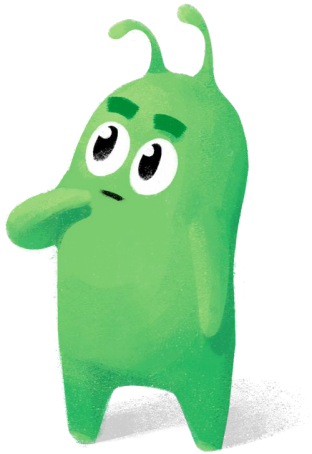
Yandex DataLens

looker

slurm.io

Так что же Modern Data Stack, а что нет?

- Картинки в интернете включают в MDS примерно все подряд, поэтому разделение как минимум спорно. Но чаще все-таки подразумеваются именно SaaS-решения
- Основная задача инструмента MDS – дать сфокусироваться на бизнес-задаче, а не на написании кода. Да-да, nocode-решения тоже часто попадают в MDS
- Как ни странно, MDS решение может быть SaaS, но при этом не заточенное под конкретного облачного вендора



slurm.io

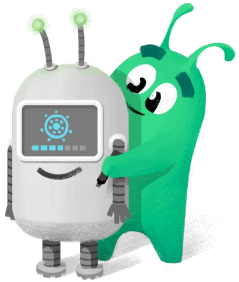
Зачем нужен dbt

slurm.io

Что такое dbt

Модели сущностей в хранилище приходится документировать, визуализировать и отслеживать этапы перекладывания данных (lineage)

Трансформации



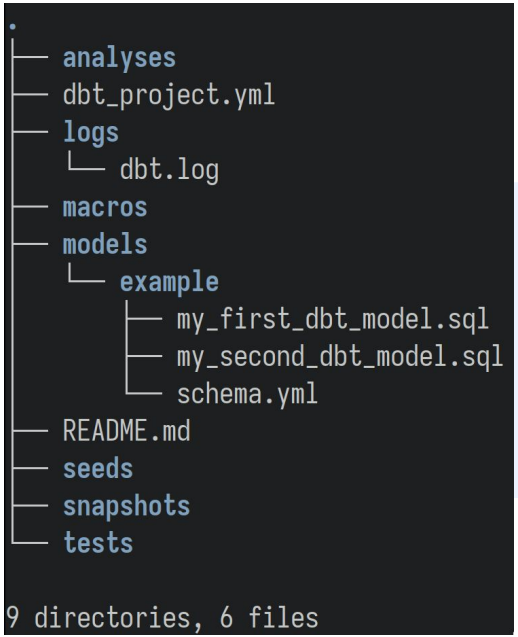
Инкрементальные обновления

Тестирование результатов

slurm.io

Первые шаги с dbt

```
~$ pip install dbt dbt-core dbt-postgres
~$ dbt init
```




```
# настройка реквизитов
~$ vim ~/.dbt/profiles.yml

# проверка связи
~$ dbt debug
```

slurm.io

Агентство путешествий во времени

```
{{ config(materialized='table') }}

SELECT
    'TARDIS' AS machine_name,
    'Gallifrey' AS origin_planet,
    5 AS passenger_capacity,
    '1963-11-23'::DATE AS debut_date
UNION ALL
SELECT
    'DeLorean',
    'Earth',
    2,
    '1985-07-03'::DATE
UNION ALL
SELECT
    'Hot Tub',
    'Earth',
    4,
    '2010-03-26'::DATE
```

time_machines.sql

slurm.io

Агентство путешествий во времени

```
{{ config(materialized='incremental') }}

WITH base AS (
    SELECT
        'TARDIS' AS machine_name,
        'Ancient Rome' AS destination_era,
        '44 BC' AS target_year,
        4 AS passengers,
        '2023-01-01'::DATE AS tour_date
    UNION ALL
    SELECT
        'DeLorean',
        'Wild West',
        '1885 AD',
        2,
        '2023-01-10'::DATE
    UNION ALL
    SELECT
        'Hot Tub',
        'Renaissance Italy',
        '1503 AD',
        4,
        '2023-02-14'::DATE
)
```

```
SELECT *
FROM base
{% if is_incremental() %}
WHERE tour_date > (SELECT
MAX(tour_date) FROM {{ this }})
{% endif %}
```

time_tours.sql

slurm.io

Базовое Data Quality

```
~$ dbt test
```

```
{% macro test_passenger_capacity(model, column_name) %}

SELECT machine_name, {{ column_name }}
FROM {{ model }}
WHERE {{ column_name }} > (
    SELECT passenger_capacity
    FROM {{ ref('time_machines') }}
    WHERE machine_name = {{ model }}.machine_name
)

{% endmacro %}
```

tests.sql + schema.yml

slurm.io

Итоги. О чем поговорили:

1. Использование облаков только для запуска виртуальных машин может быть не очень оправдано
2. Существует гигантское количество *aaS-решений для самых разных бизнес-сценариев
3. Modern Data Stack - не очень осмысленное, но красивое название
4. dbt стал одним из стандартов в мире трансформации данных в хранилище

slurm.io

Спасибо
за внимание!

slurm.io