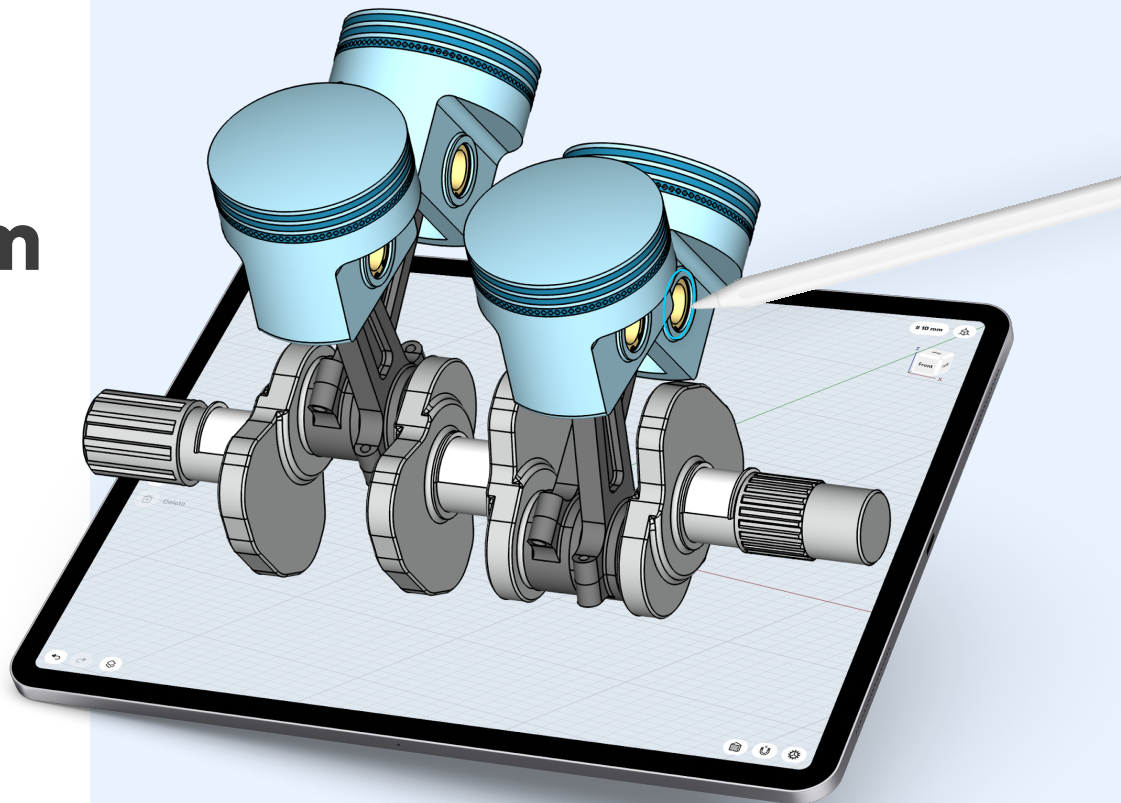




Az adatplatform skálázása

A Shapr3D-nél



Bemutatókozás

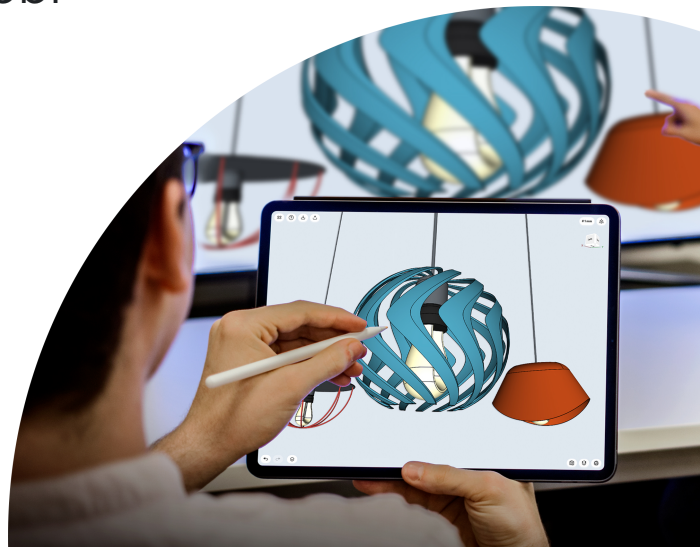


- Haller Bálint
- Senior Data Engineer
- Szoftvermérnök és Data Science háttér
 - BME Mérnök Informatika
 - University of Glasgow Data Science
 - Ex-Skyscanner
- Twitter - @balinthaller

A Shapr3D-ről

A Shapr3D-ről

- CAD **alkalmazás** pontos **3D tervezéshez és gyártáshoz**
- **iPad**-en indult, sokkal egyszerűbb mint a többi CAD
- 2014-ben bootstrappelve, az iPad Pro-t már ezzel mutatták be
- Apple Design Award 2020



A Shapr3D-ről

- Már **Mac**-en, és hamarosan **Windows**-on is
- Kb 70-edikként jöttem tavaly októberben, most
100 körül vagyunk
- Data csapat 1-ről 7-re nőtt < 1 év alatt



Az előadás részei

🗄️ A Shapr3D
adatplatformja

» Streaming
infrastruktúra

📖 Adatkatalógus

A Shapr3D adatplatformja

Motiváció

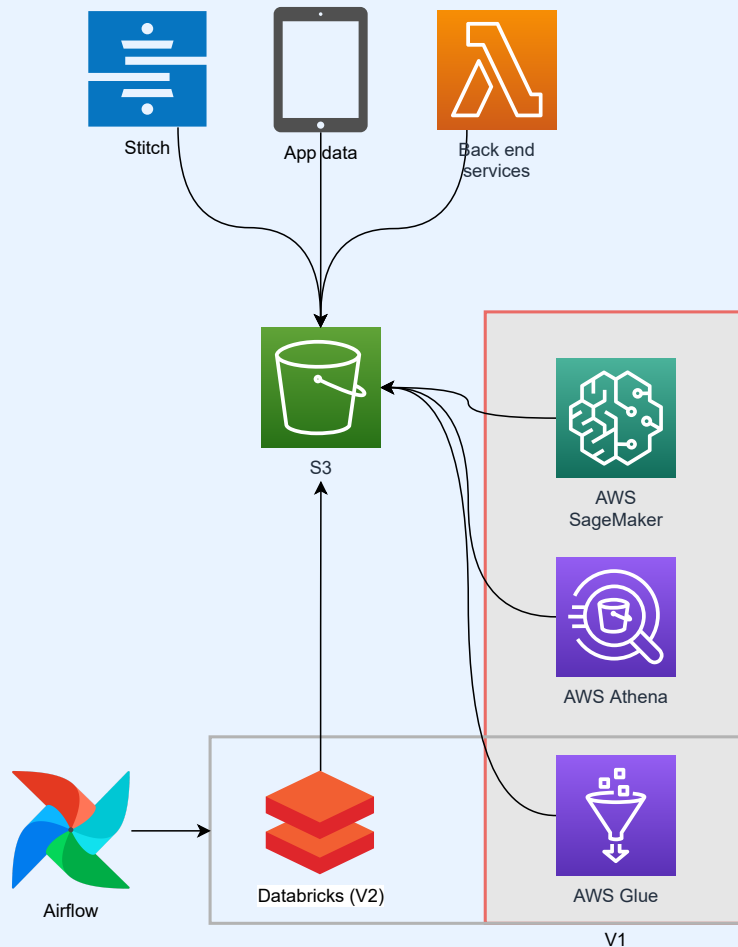
- Hogyan kapcsolunk össze adatokat különböző forrásokból?
- CSV + Lokális Jupyter hamar eléri a korlátait
 - **Adatméret**, lekérdezések **gyorsasága**
- **Manuális munka + hibalehetőségek lecsökkentése**
 - Data Science ne repetitív riportokat gyártson

Architektúra

- S3 alapú data lake

- DB snapshotok is itt
- Stitch a külső adatszettekhez
- SageMaker helyett már Databricks **notebookok**

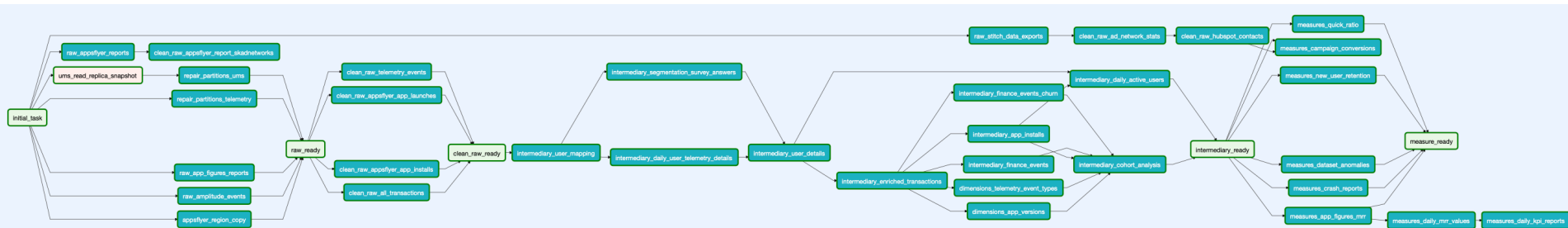
- Spark ETL



Orchesztráció

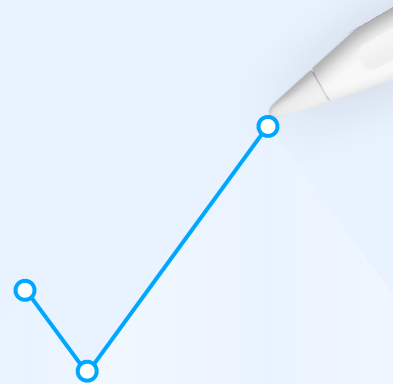
- Rétegzett struktúra
- Sync taskok

- A legtöbb job csak SQL + unit tesztek + validáció



Tanulságok

- Eszközök > Emberek
- A korai optimalizáció különösen drága
- Kicsi csapat lehet gyorsabb
 - De figyelni kell a silókra



» Streaming infrastruktúra

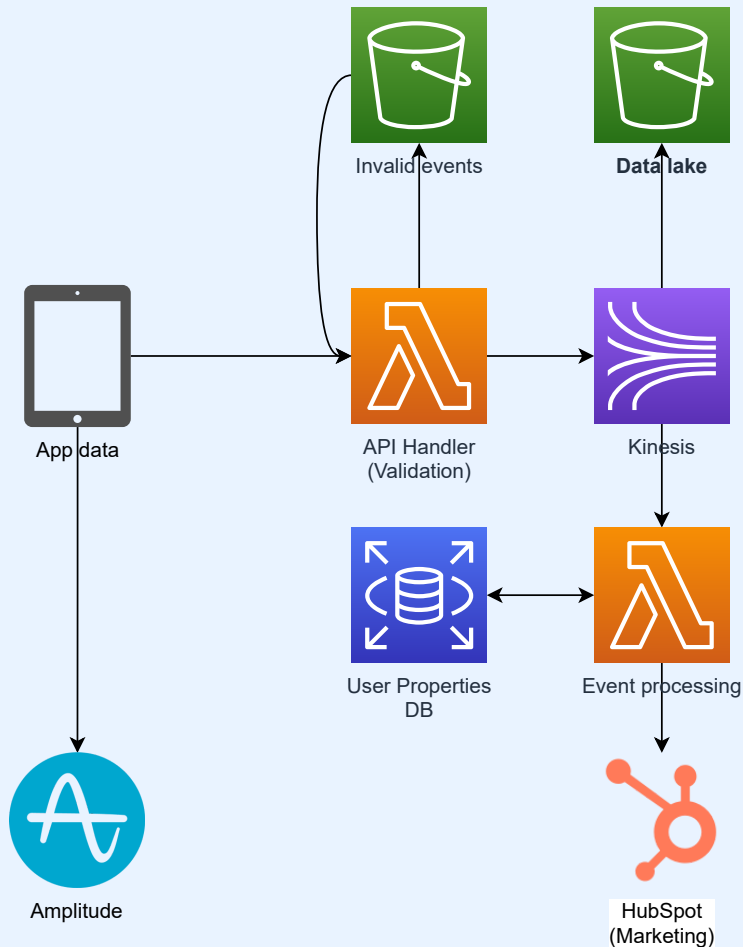
Motiváció

- Real-time analitika - kezdetben nem jelentős
- **Magunknál akarjuk tudni az adatainkat**
(app, web, back end, logging)
 - Újraértelmezhetőség
 - **Definíciókat mi találjuk ki**, nem más diktálja

Streaming infrastruktúra

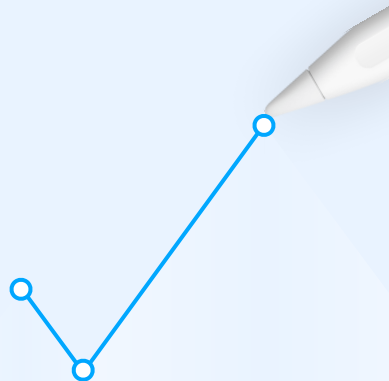
Architektúra

- JSON Schema (backfillel)
 - App is abból dolgozik
- Valós idejű felhasználói adatok
- S3-ba batchelve kerülnek



Tanulságok

- **Perzisztencia** > minden (kis túlzással)
- A **streaming nehéz** – kell use case
 - Operációs és logikai szempontból is
 - Sokszor **egyszerűbbek a kis batchek**
- Nagy szabadságot ad ha birtoklod a saját adataid
- Folyamatos egyszerűsítés kell – pl. Message bus



Adatkatalógus

Motiváció

- Van adatunk, de ki használja és hogyan?
- Ha manuálisan dokumentáljuk az adatokat szinte **mindig** el lesz maradva
- Több helyen él az adat, de egységesen kereshető dokumentáció kell
- Hova menjek kérdésekkel? Friss ez egyáltalán? 🤔

Adatkatalógus

- Ez végülis csak dokumentáció
- Miért ne használjunk erre kitalált technológiákat?

carte-demo

Home

Postgres

Pagila

actor

address

category

city

country

customer

film

film_actor

film_category

inventory

language

payment_p2020_01

payment_p2020_02

payment_p2020_03

payment_p2020_04

payment_p2020_05

payment_p2020_06

rental

staff

actor

Edit

Type: table

Database: pagila

Location: postgres://pagila.public/actor

Description

The actor table lists information for all actors.

The actor table is joined to the film table by means of the film_actor table.

Columns

ACTOR_ID integer

A surrogate primary key used to uniquely identify each actor in the table.

FIRST_NAME text

The actor first name.

LAST_NAME text

Table of contents

actor

Type:

Database:

Location:

Description

Columns

actor_id

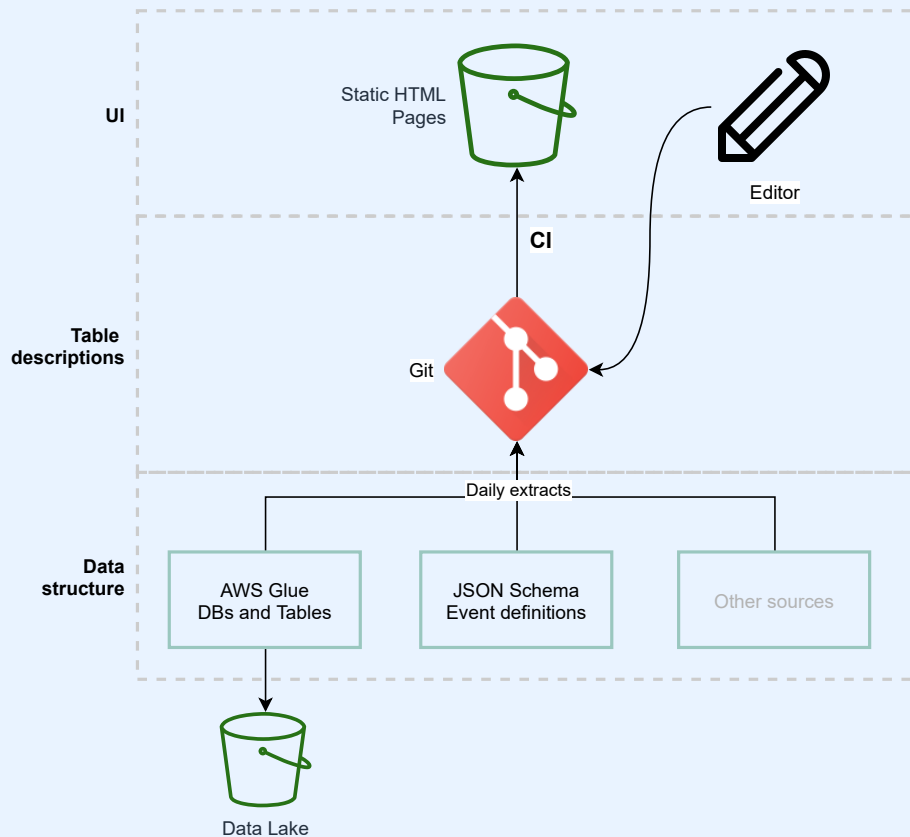
first_name

last_name

last_update

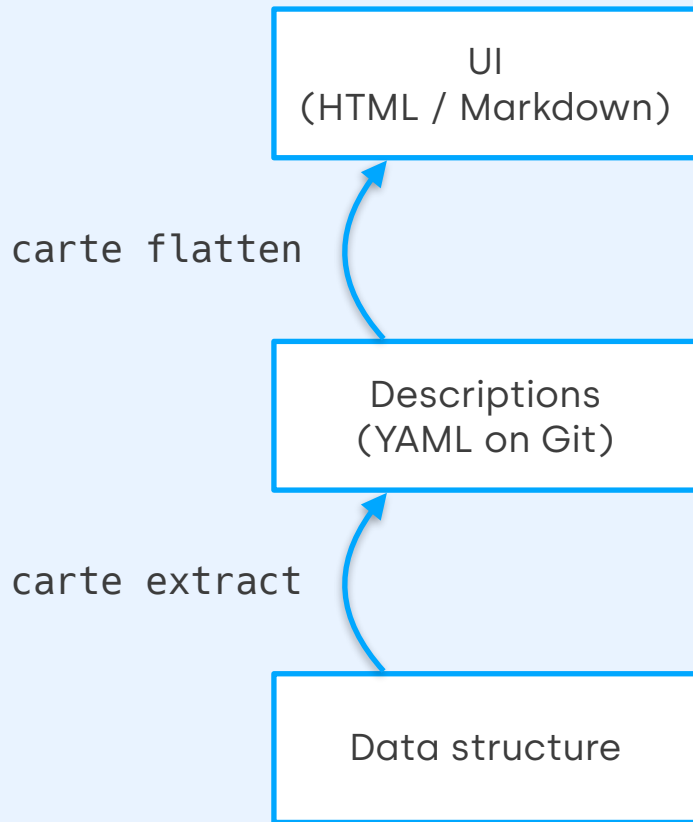
Architektúra

- Struktúra automatikusan jön a forrásokból
- Leírások Giten (verziózott, rollbackelhető)
- Front-end szerkesztő, ami visszacommitol



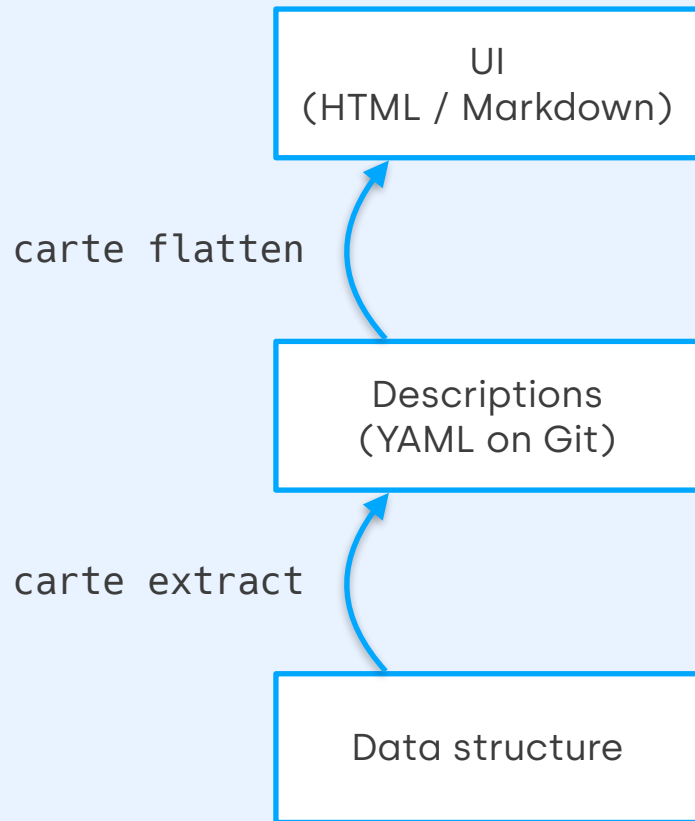
Carte

- Open source CLI
- Nem kell hozzá back end, statikusan hosztolható
- Minden csak YAML + Markdown
- Rengeteg adatforrást támogat (az Amundsen-nek köszönhetően)



Carte

github.com/carte-data/carte
demo.cartedata.com



Tanulságok

- Minden nagy cég épít magának, de a kisebbeknek nincs megfelelő megoldás
- A kereshetőség elengedhetetlen
- Minél kevesebb manuális munka, annál jobb (logikus, nem?)
- Közös definíciók ereje



Záró gondolatok

- Az adat hatalmas érték startupoknak is
 - Maradjunk birtokában
- Egyre inkább mindenki ugyanazokat az eszközöket és technológiákat használja
 - A felhasználók / vásárlók definíciója és viselkedése viszont mindenhol más
 - Ezt fontos felismerni



Kérdések