La mia prima data pipeline

Torino Coding Society

Riccardo Magliocchetti

whoami

- Sw developer @ Maieutical Labs / Consultant
- @rmistaken

Menu della serata

Creare una data pipeline che:

- consumi una api
- salvi i dati in db
- ci permetta di fare delle analisi

Cos’è una data pipeline?

Perchè questo talk?

Cliente: *Dobbiamo mostrare ai nostri clienti statistiche dell'uso del nostro prodotto da parte degli utenti*

Criticità

- auth
- dati filtrati per cliente
- gli utenti devono potersi fare le proprio visualizzazioni in autonomia

OTOH

Un progetto senza *legacy* \o/

Come lo implemento?

- Web app fatta in casa
- Elasticsearch / Kibana / Logstash / Beats
- Time-series db + Grafana

Web app fatta in casa

- un'altra app da scrivere e da mantere

Elastic Stack

- altri due servizi complessi da mantenere
- auth a pagamento
- pacchetti distro non aggiornati
- java :P

Time-series db + Grafana

- permessi fine-grained non disponibili
- visualizzazioni Grafana troppo semplici

KISS! aka Boring Tech

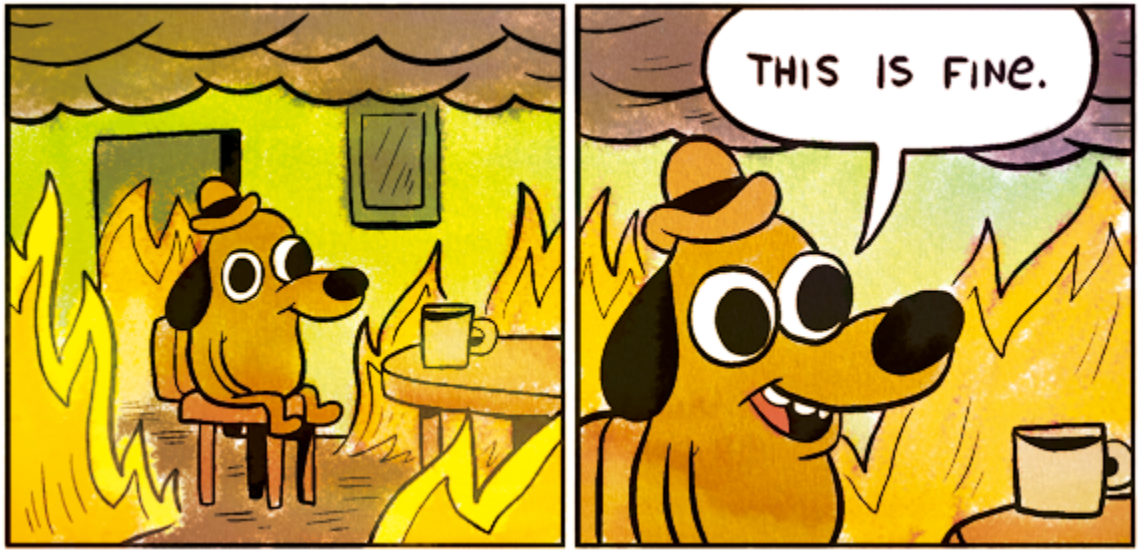
Punti saldi

- storage: db relazionale (Postgres)
- API lato applicazione per esportare i dati
- Luigi per il plumbing
- per il frontend prendiamo tempo :)

Luigi

- creato in Spotify
- scritto in Python
- usato da molti
- secondo miglior idraulico al mondo!

Mi evita di scriptare



Concetti principali

- **Task**: *run()*, *output()*, *requires()*
- **Target**: un file su disco / S3 / HDFS, una riga in un database

Batch vs Realtime

- fallisce, guardo i log e rifaccio ripartire
- voglio cambiare il formato dati? rifaccio le query alle api e mi ricostruisco il db
- granularità richiesta una settimana se non mese

30 marzo 2016

@mistercrunch annuncia:

As a vector for data exploration, discovery, and collaborative analytics, we have built and are now ***open sourcing****, a* ***data exploration*** *and* ***dashboarding platform*** *named* ***Caravel.***

Caravel

- creato e **mantenuto** da un **team** di airbnb
- permette di creare visualizzazioni e dashboard in autonomia

Caravel Tech

- hackable! ~8 KLOC di python vs ES 1.6 MLOC java
- flask + flask app builder
- grafici d3.js / nvd3.js , frontend passaggio a react in corso
- legge dati da dialetto sqlalchemy (hive / impala) o druid.io

DEMO

Thanks!

- @rmistaken
- github
- Contatti