

LOVOO auf Wolke 7

Stefan Weigert
(Head of Data)

Über LOVOO

- 110+ Mitarbeiter
- 2 Büros (B, DD)
- > 50M Nutzer Weltweit
- ~1M req/min
- Android, IOS, Web

Über LOVOO

- 110+ Mitarbeiter
- 2 Büros (B, DD)
- > 50M Nutzer Weltweit
- ~1M req/min
- Android, IOS, Web





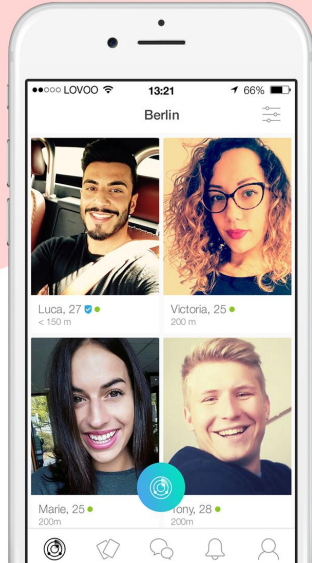
DecompiledD

April 6 – Parkhotel Dresden

#decompiled18

Die App Funktionen

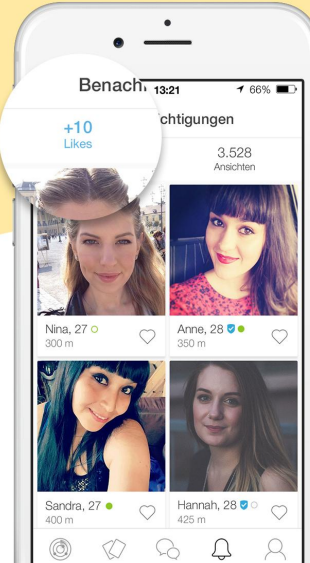
FINDE NEUE LEUTE
aus deiner Umgebung.



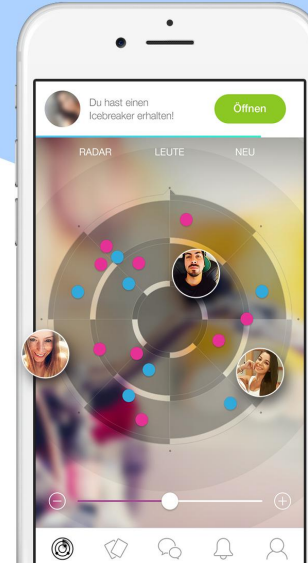
CHATTE OHNE MATCH -
mit dem Icebreaker.



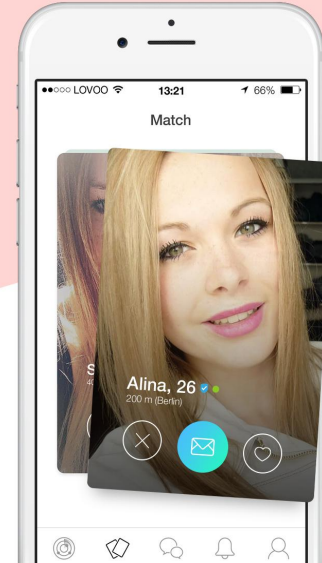
FINDE HERAUS,
wer dich mag.



SIEH LIVE,
wer in deiner Nähe ist.



SPIELE MATCH
und verteile Likes.



Data @ LOVVO

Fragestellungen

- Product Updates
- Real Users vs Spammers
- Recommendations
- Product-/System- Health
- User Insights

Data Team

Das Team

- ~20 Mitarbeiter
- Data Scientists
- Data Engineers
- Analysten
- Machine Learning
- Stream Processing

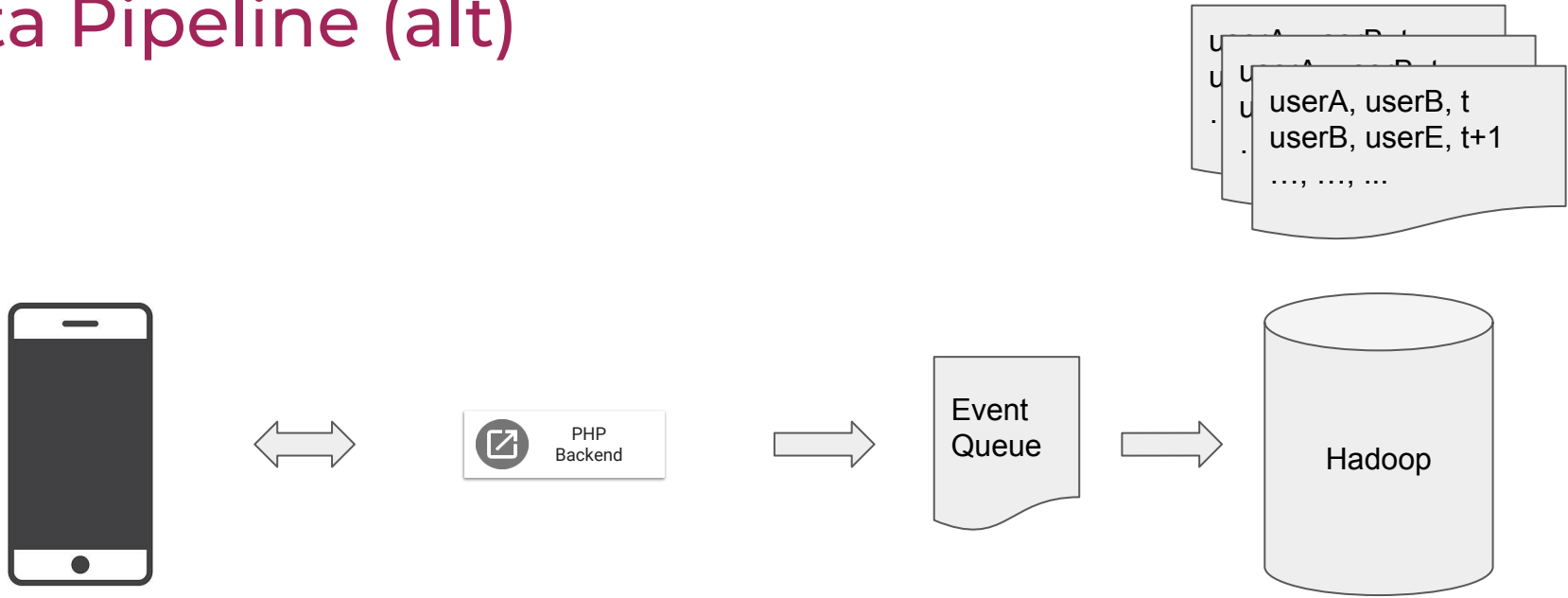


Fragestellungen ...und unsere Antworten

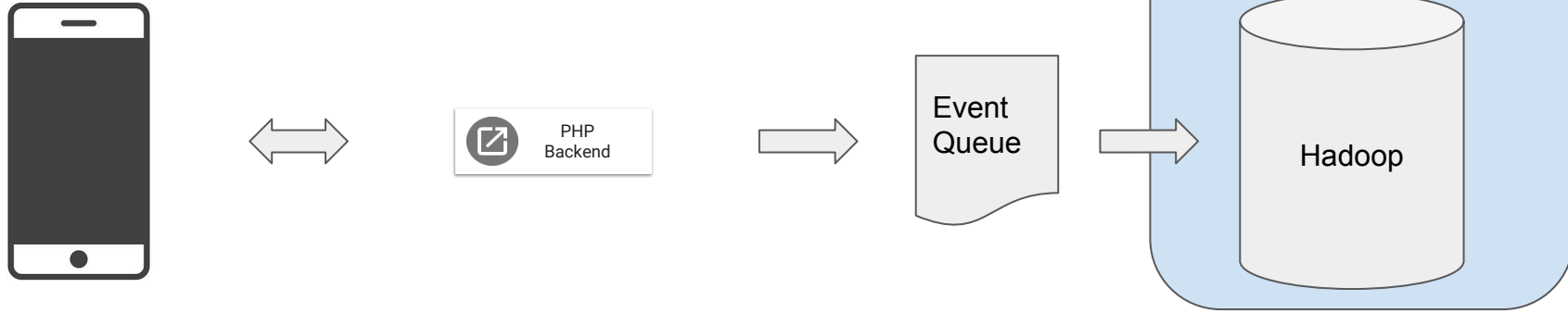
- Product Updates
 - A/B Testing
 - Dashboards: Retention, Engagement, Revenue
- Real Users vs Spammers
 - AntiSpam
- Recommendations
 - Collaborative Filtering
 - Content-Based Filtering
- Product-/System- Health
 - (Online) Anomaly Detection
- User Insights
 - Segmentation

Wie alles beginnt...

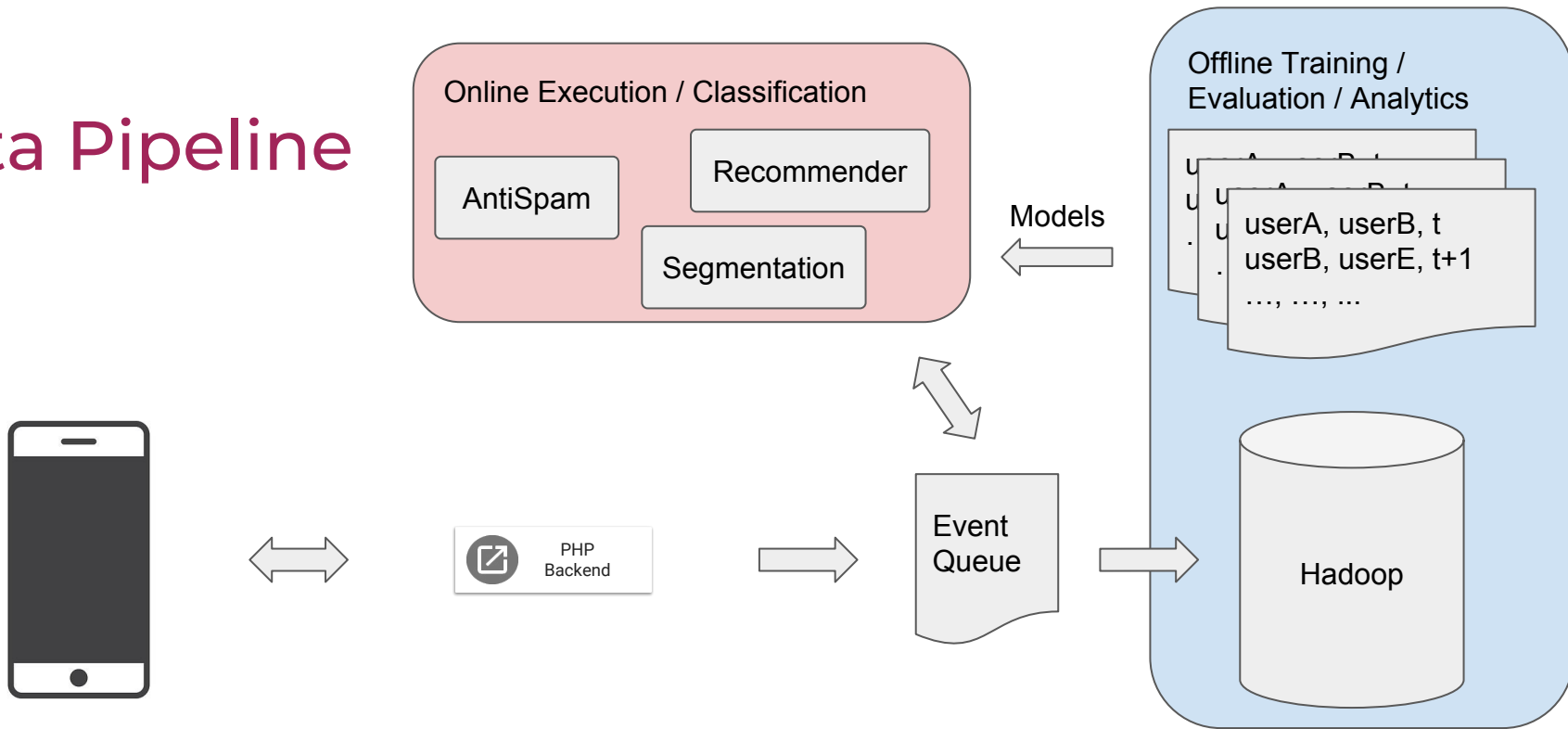
Data Pipeline (alt)



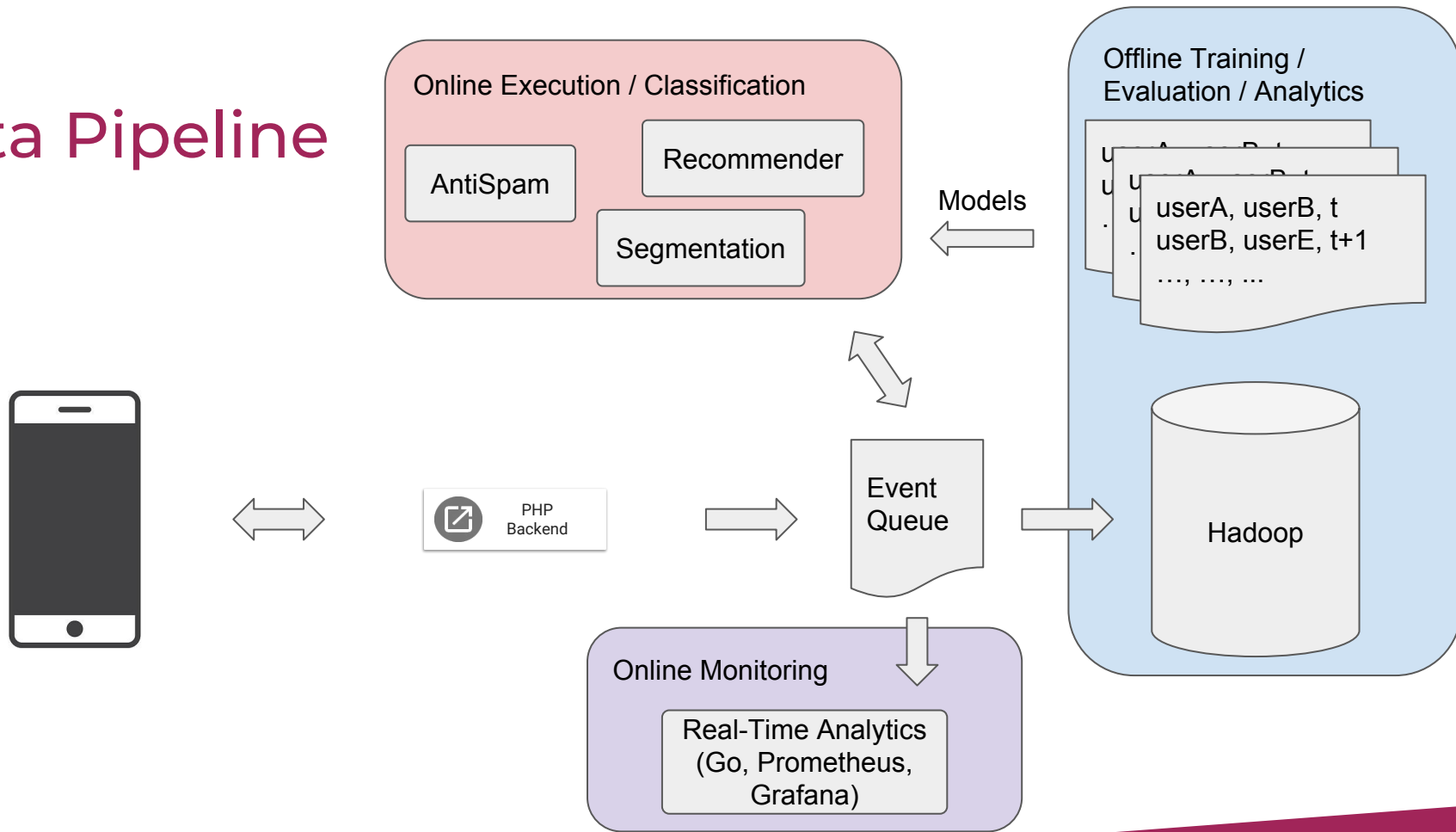
Data Pipeline



Data Pipeline



Data Pipeline



Probleme

- Hadoop Spezifisch
 - Langsam
 - Maintenance
 - Löschen geht nicht
 - Keine Berechtigungen (“alles oder nichts”)
- Generell
 - JSON
 - GDPR
 - Kein CI für ETL pipeline

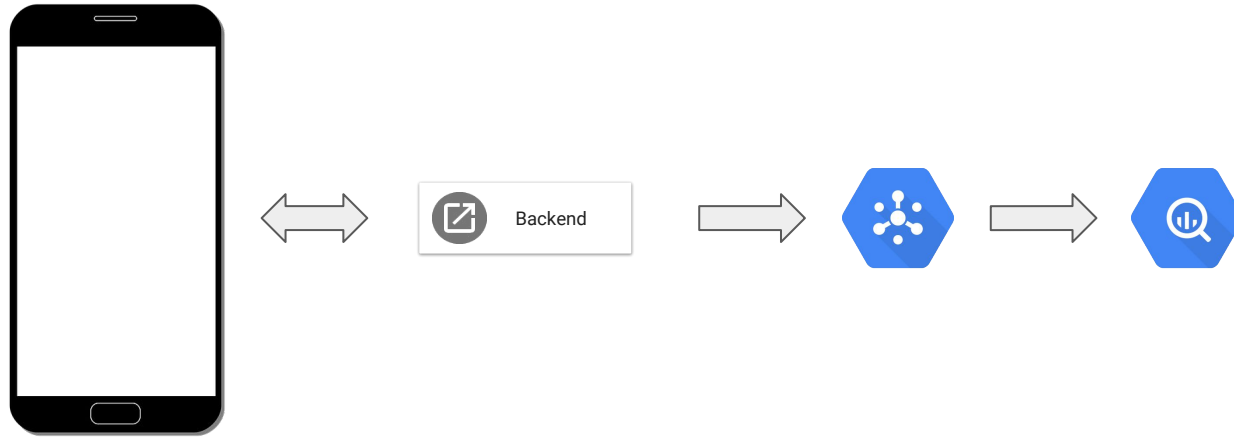


Ab in die cloud

Wie alles begann

- Kurzes “window of opportunity” von ~ 3 Monaten
- “All hands”
- Google Cloud ist teil der tech Strategie
- Neue Event infrastruktur

Grobarchitektur



Langsam

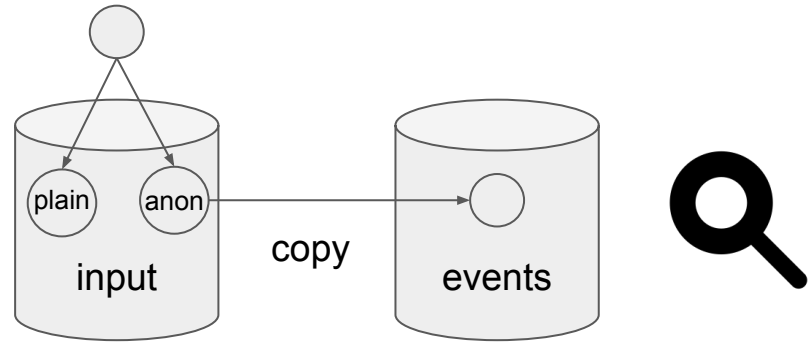
Neue Event Infrastruktur

- Google Protobufs (Verifikation beim kompilieren)
- Zentrales Repository für alle Events
- “Events as code”
 - Code reviews
 - GIT
 - Versioning

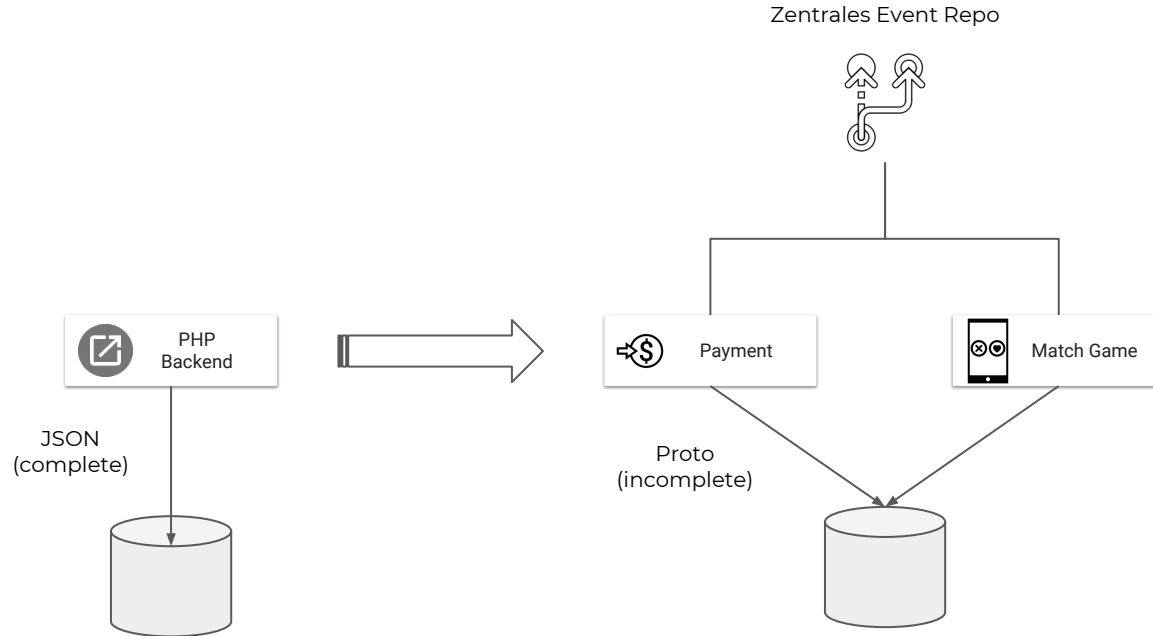
```
// ProfileCreated occurs when a new user profile was created.  
message ProfileCreated {  
    string actor_id = 1;  
    // @anon:omit  
    string name = 2;  
    // @anon:email  
    string email = 3;  
    // @anon:birthday  
    string birthday = 6;  
    ...  
}
```

Anonymisierung

- Stark eingeschränkter Zugriff auf “input”
- Kleine retention für input Tabellen
- “Plain” wird gelöscht wenn Nutzer sein Profil löscht



Monolith → Microservice (am Beispiel)

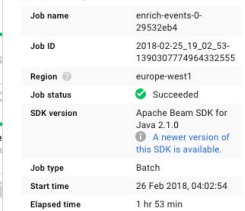


Apache Beam

- Map/Reduce Framework
- Basiert auf Apache Beam
- Kann in Spark/Storm/Google Dataflow ausgeführt werden

 LOGS

Job summary



Workers	0
Current state	Worker pool stopped.



Current vCPUs	180
Total vCPU time	288.098 vCPU hr
Current memory	675 GB
Total memory time	1,080.366 GB hr
Current PD	17.58 TB
Total PD time	28,809.765 GB hr
Current SSD PD	0 B
Total SSD PD time	0 GB hr
Total shuffle usage	—

Filter

```
/write/convert-to-row-out0-ElementCount  
/write/convert-to-row-out0-MeanByteCount
```

Moment...

Hadoop -> BQ

- HDFS/Parquet (2-stelliger Terrabyte bereich)
- Variante 1: Neu strukturierung der Events auf Hadoop
- Variante 2:
 - 1) Hochladen der Parquet files
 - 2) Importierung zu BQ
 - 3) Neu Strukturierung auf BQ

Hadoop -> BQ

- HDFS/Parquet (2-stelliger Terrabyte bereich)
- Variante 1: Neu strukturierung der Events auf Hadoop
- Variante 2:
 - 1) Hochladen der Parquet files
 - 2) Importierung zu BQ
 - 3) Neu Strukturierung auf BQ



Hi Florian und Stefan,

ich möchte freundlich darauf hinweisen, dass das Parquet loading feature noch nicht für produktive Workloads verwendet werden sollte - Hintergrund ist, dass dort täglich einige tausend load jobs zu laufen scheinen (Status vergangene Woche)

Ist dies ein Test von Eurer Seite?

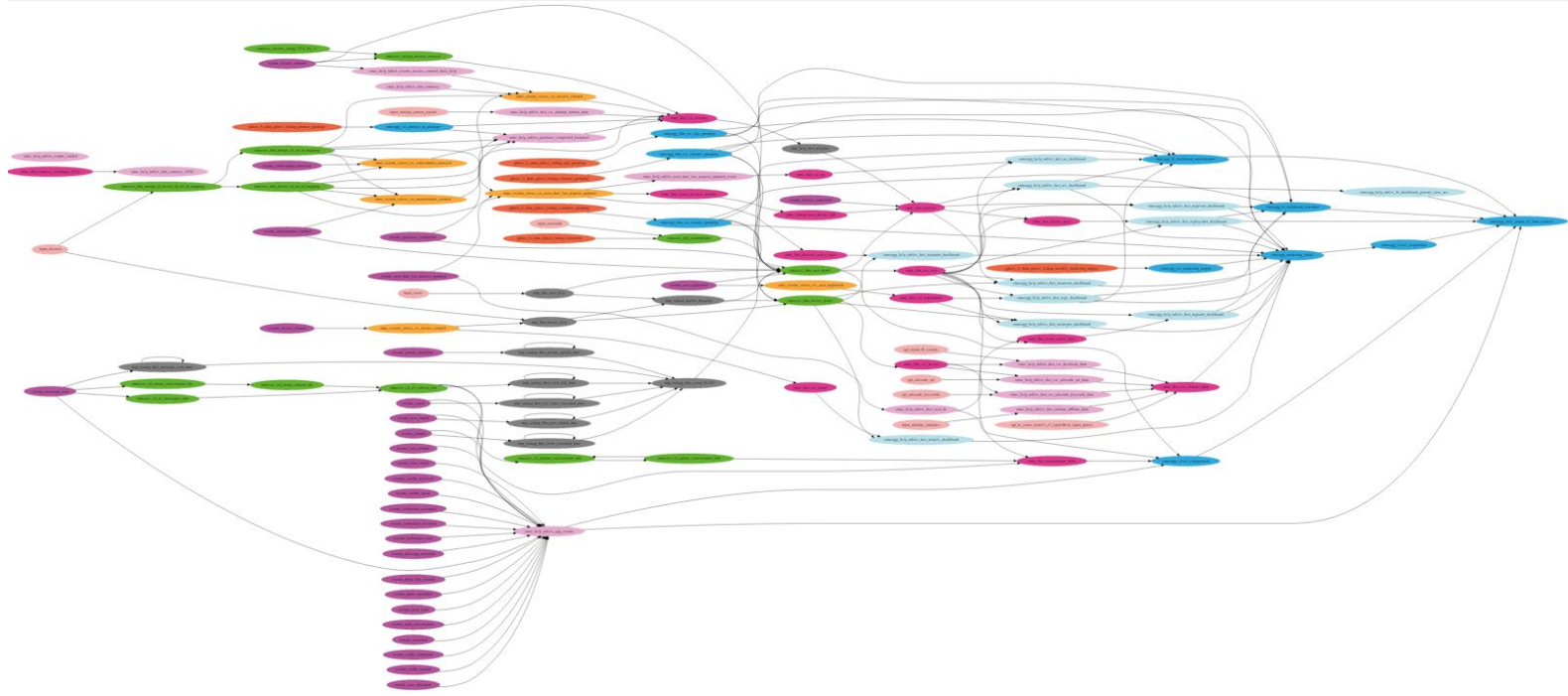
Viele Grüße,
Thomas

ETL @ Google

ETL @ Hadoop

- 114 Tabellen
- 74 Skripte
- Deployment von Hand
- Oozie als Job-Scheduler

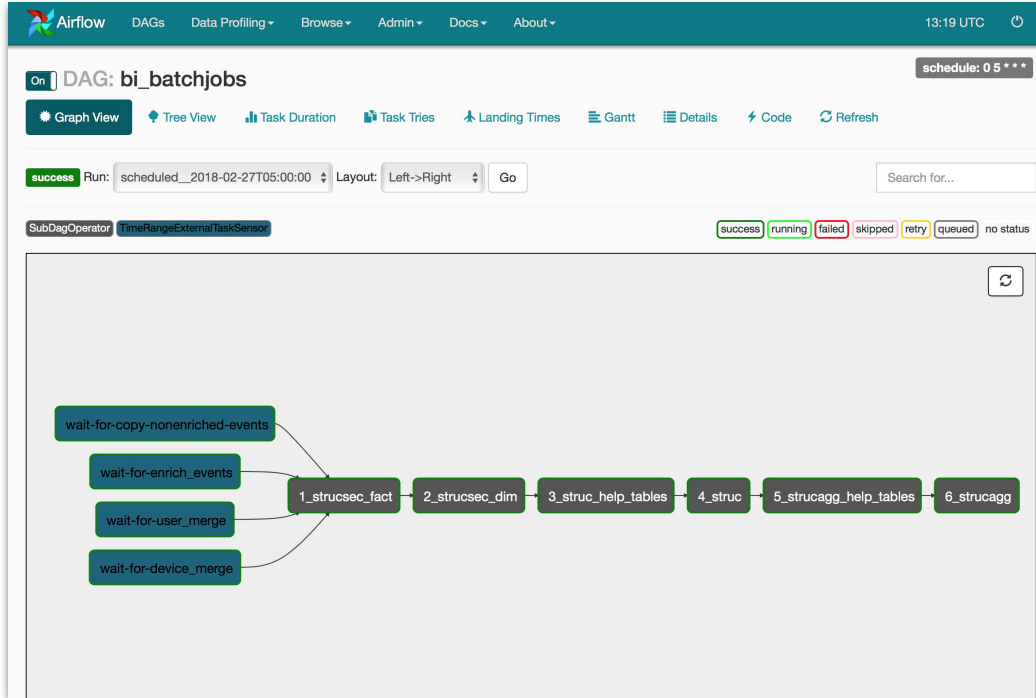
ETL Abhängigkeitsbaum



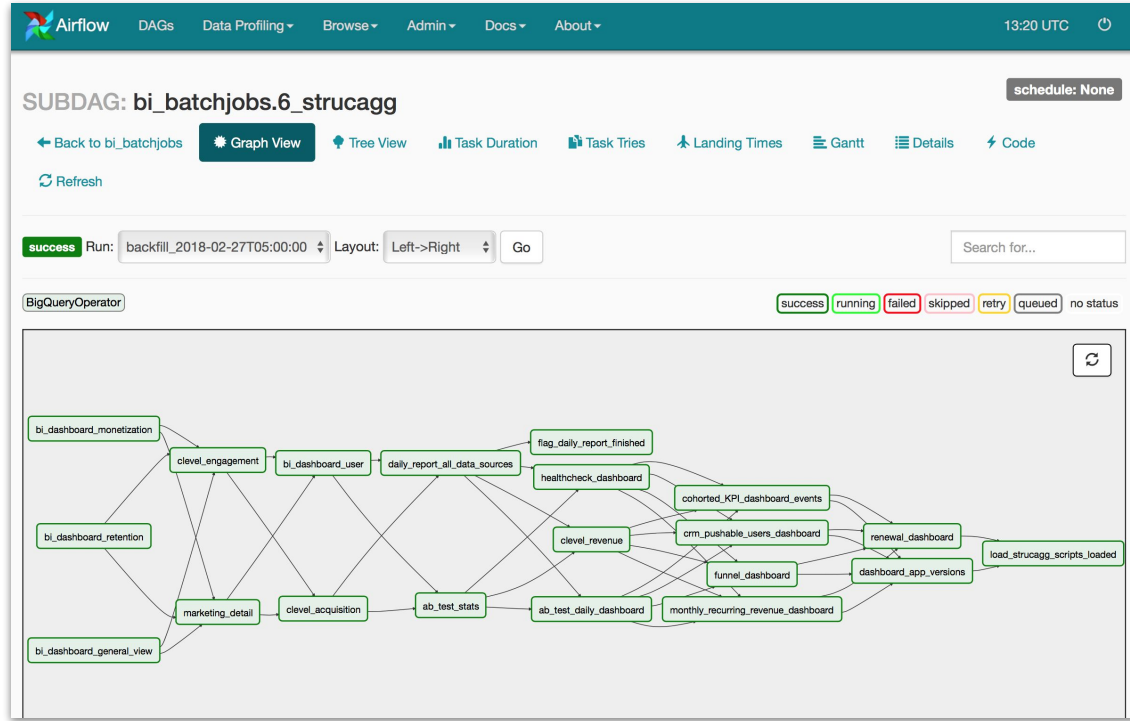
Airflow

- Entwickelt von Airbnb
- Inzwischen Apache Projekt
- DAGs werden in Python geschrieben
- Steuert Dataflows und BigQuery Jobs

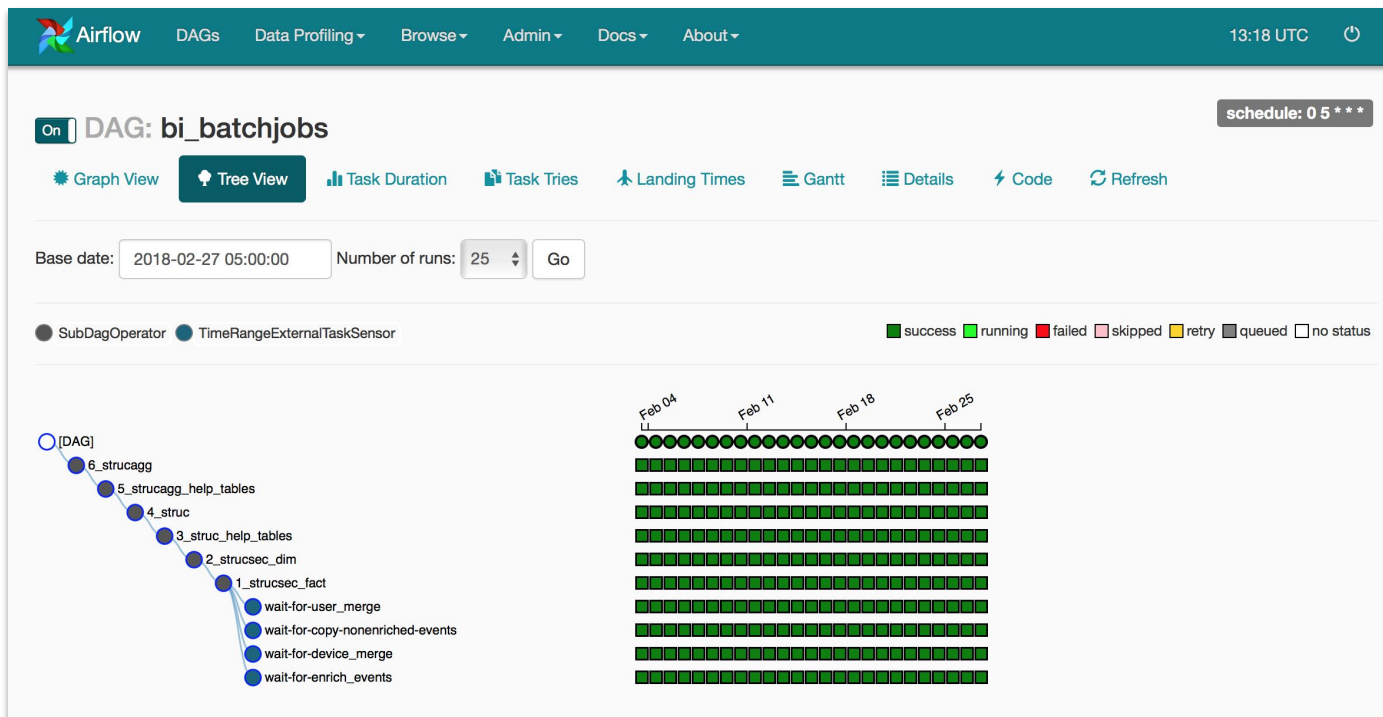
Airflow - ETL DAGs



Airflow - DAG Zoom



Airflow



Versioning for ETL

change dt #512

Merged merged 1 commit into `master` from `dt` 6 days ago

Conversation 0 Commits 1 Files changed 1

 commented 6 days ago + 😊 ✎

No description provided.

  change dt Verified ✓ 024a686

  self-assigned this 6 days ago

  requested a review from 6 days ago

  approved these changes 6 days ago View changes

  merged commit **29092b1** into `master` 6 days ago View details Revert

1 check passed

Deployment

lovoo / bi-scripts > 610

Successful

Merge pull request #512 from lovoo/dt change dt

🕒 6 days ago

🕒 4 minutes, 43 seconds

🔗 29092b1b6a

📁 master

clone	00:02	✅
build	00:38	✅
publish	00:26	✅
run	03:31	✅
notify	00:04	✅

```
1 + git init 0s
2 Initialized empty Git repository in /drone/src/github.com/lovoo/bi-scripts/.git/ 0s
3 + git remote add origin https://github.com/lovoo/bi-scripts.git 0s
4 + git fetch --no-tags origin +refs/heads/master: 0s
5 From https://github.com/lovoo/bi-scripts 1s
6 * branch      master      -> FETCH_HEAD 1s
7 * [new branch] master      -> origin/master 1s
8 + git reset --hard -q 29092b1b6ab450a480be720286c4dc6fe080c97c 1s
9 + git submodule update --init --recursive 1s

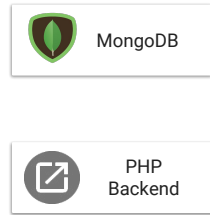
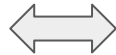
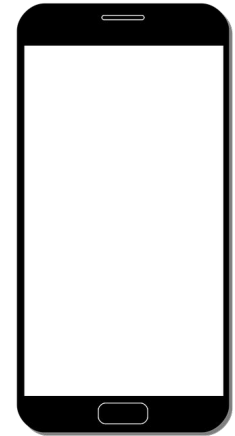
exit code 0
```

⬆

⬇

Am Ziel

Aktuelle Data Architektur



COMPOSE QUERY

Query History

Job History

Filter by ID or label

API Project

No datasets found in this project.

Please create a dataset or select a new project from the menu above.

Public Datasets

- ▶ bigquery-public-data:hacker_news
- ▶ bigquery-public-data:noaa_gsod
- ▶ bigquery-public-data:samples
- ▶ bigquery-public-data:usa_names
- ▶ gdelit-bq:hathitrustbooks
- ▶ gdelit-bq:internetarchivebooks
- ▼ lookerdata:cdc
 - project_tycho_reports
- ▶ nyc-tlc:green
- ▶ nyc-tlc:yellow

New Query ?

Query Editor

UDF Editor



SQL

```
1 SELECT
2   disease,
3   SUM(SAFE_CAST(cases AS INT64)) AS ncases
4 FROM
5   `lookerdata.cdc.project_tycho_reports`
6 GROUP BY
7   disease
```

Valid: This query will process 9.69 MB when run.

Standard SQL Dialect 

RUN QUERY

Save Query

Save View

Format Query

Show Options

Query complete (1.7s elapsed, 9.69 MB processed)



Results

Details

Download as CSV

Download as JSON

Save as Table

Save to Google Sheets

Row	disease	ncases
1	MUMPS	832110
2	POLIO	500809
3	MEASLES	18670998
4	RUBELLA	434966
5	PERTUSSIS	2329020
6	HEPATITIS A	976362
7	SMALLPOX	232828
8	DIPHTHERIA	901240

Table

JSON



← BigQuery



Field Editing in Reports: ON

CANCEL

RECONNECT



FIELDS →

Connectors

File Upload

AdWords

Attribution 360

BigQuery

Cloud SQL

DCM

DFP

Google Cloud Storage

Google Analytics

Google Sheets

MySQL

PostgreSQL

Search Console

YouTube Analytics

Community Connectors

EXPLORE CONNECTORS

DEVELOPERS



BigQuery

BigQuery is Google's fully managed, petabyte scale, low-cost analyticsdata warehouse. BigQuery charges for querying/processing of data.

Those queries are charged to the credit card of the billing project. [LEARN MORE](#)

MY PROJECTS

Project



SHARED PROJECTS

API Project

CUSTOM QUERY

PUBLIC DATASETS

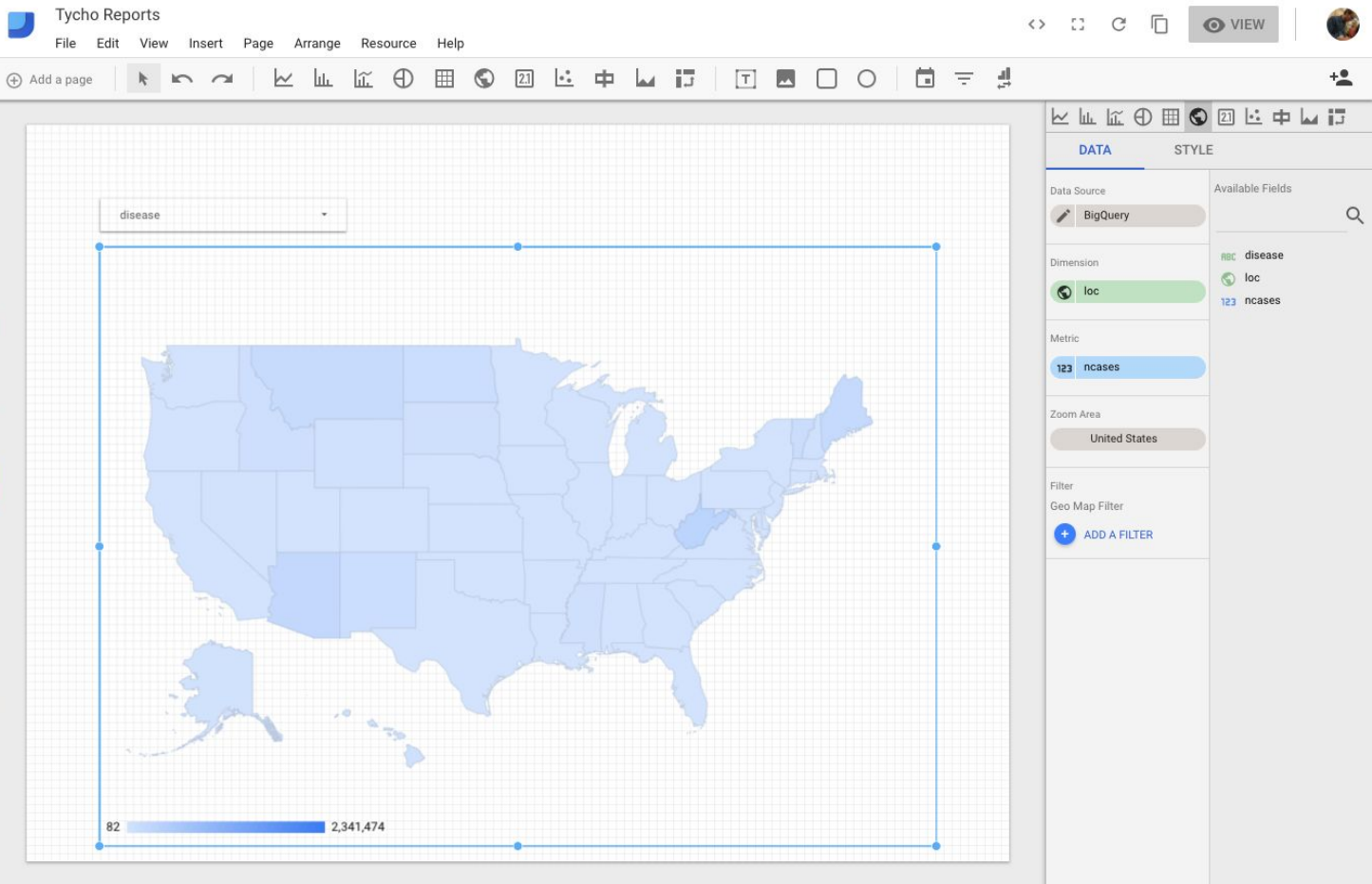
Enter Custom Query

```
1 SELECT
2   disease,
3   loc,
4   SUM(SAFE_CAST(cases AS INT64)) AS ncases
5 FROM
6   `lookerdata.cdc.project_tycho_reports`
7 GROUP BY
8   disease,
9   loc
```

Query Options

☐ Use Legacy SQL ?

Advanced Options ▾

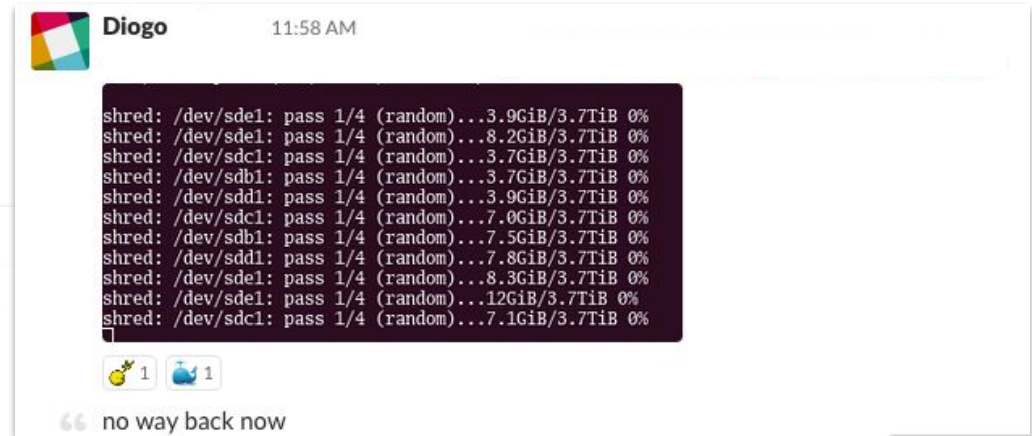
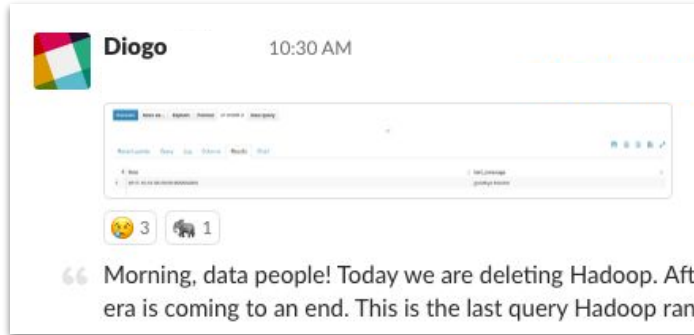


Dinge die wir gelernt haben

- Ein Plan ist gut, aber ...
- Kostenlimit sollte sehr großzügig für Migration gehalten werden
- Fokus, Fokus, Fokus

Dinge die wir gelernt haben

- Ein Plan ist gut, aber ...
- Kostenlimit sollte sehr großzügig für Migration gehalten werden
- Fokus, Fokus, Fokus
- Man braucht Mut



One last thing...

We are hiring!

- Data Scientist
- Test Engineer (Big Data)
- Software Engineer (Big Data)
- Data Analyst
- Software Engineer Frontend and Backend
- Q/A Engineer

<https://inside.lovoo.com/jobs/>

Danke!