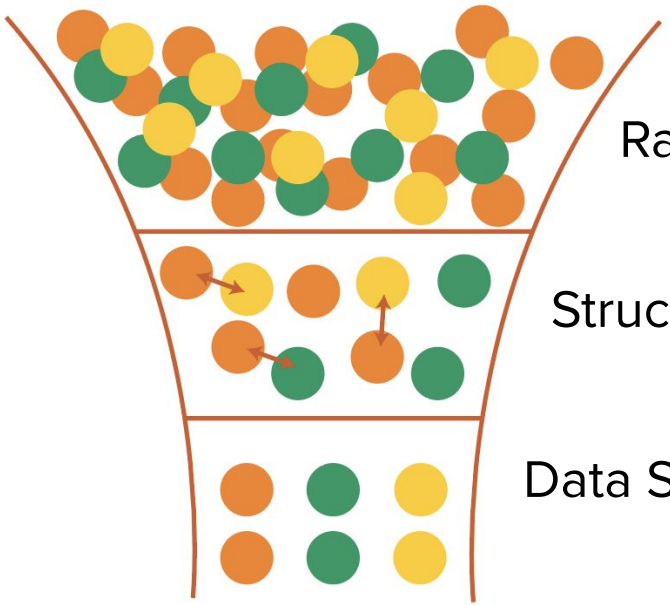
Rethinking Data @ SoundCloud

Ana Pereira
Janette Müller-Lehmann
Mahnaz Namazizavareh

Berlin Women in Machine Learning & Data Science Meetup

November 05, 2020

Agenda



Raw Data

Structured Data

Data Science

Data

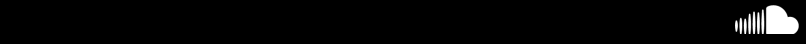
Google

Cloud

Platform

Technology

SoundCloud & Data Science



SOUNDCLOUD
THE WORLD'S LARGEST OPEN AUDIO PLATFORM

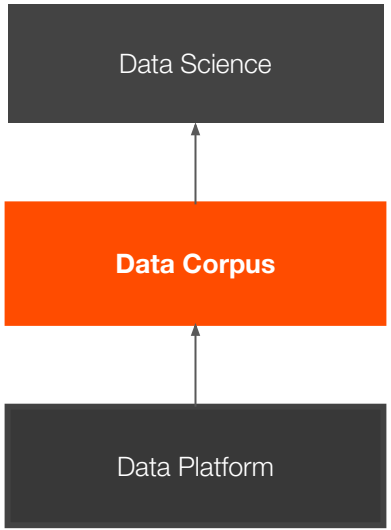
SoundCloud **empowers creators to connect directly** with young, influential listeners driving culture.

A rich audio discovery experience, **combining new music, DJ sets, and remixes direct from creators,** feeding vibrant local scenes worldwide.

SoundCloud is a global accelerator of audio culture:
WHAT'S NEXT IN MUSIC IS FIRST ON SOUNDCLOUD.

200+ million tracks | 25+ million creators | available in 190 countries and territories

Data @ SoundCloud



Data reporting, analytics & modelling.
[Mahnaz]

Data governance for the most critical datasets.
[Ana & Janette]

Data infrastructure tools and capabilities.

Data Science @ SoundCloud

Reporting

How are we performing?
How many plays does my track have (creator)?
...

Audience Intelligence

How is our influencer community structured?
Who are up-and-coming creators to look out for?
...

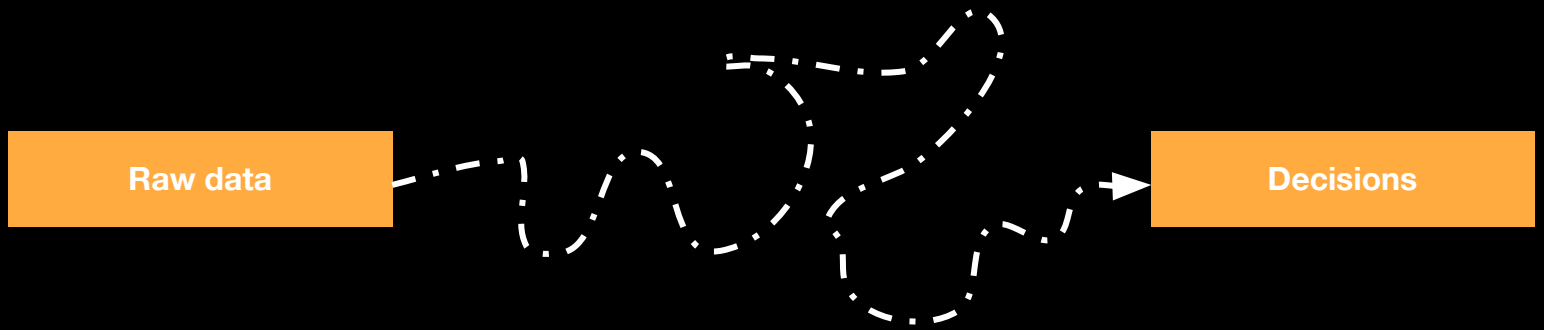
Product Exploration

How is our new feature performing?
Where should we launch next?
...

Data Products

Which are the most interesting tracks for a user?
How do I find relevant hip hop tracks?
...

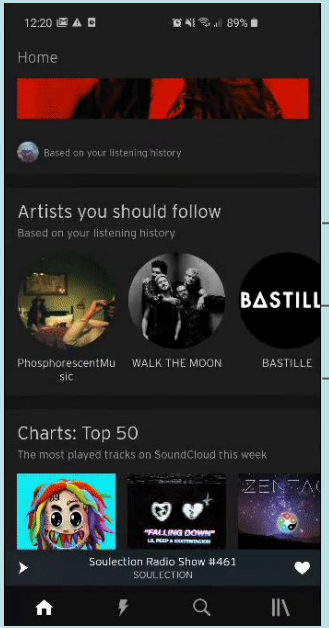
The Journey of Data



Raw Data

Engineering teams (front and back end)

Data platform team (data engineers)



MySQL DB

Server logs

Event Tracking

Anonymization

Data Warehouse [4.4 PT]

Users interact with SoundCloud

Ingestion

Storage

Problem 1
The Infrastructure

Data Science in the Past - In General

Collect	Ingest	Store	Process	Explore & Analyze
Client event tracking	Kafka Streams	HDFS	Spark	Jupyter Notebooks
SQL databases	Ingestion jobs	Hive	HiveQL	Tableau
Server logs		MySQL Database	MapReduce	R
...		CSV files on a cluster	Pig	MS Access / MySQL
				Excel

Various Clusters

distributed computing

Data Science in the Past @ SoundCloud

Collect	Ingest	Store	Process	Explore & Analyze
Client event tracking	Kafka Streams	HDFS	Spark	Jupyter Notebooks
SQL databases	Ingestion jobs	Redshift	Redshift SQL	Tableau
Server logs				R
...				Excel

Various Clusters

distributed computing

Data Science in the Past

Collect

Client event tracking

SQL databases

Server logs

...

High **configuration and maintenance costs** of various clusters, systems and applications.

Batch processing frameworks are **not suitable for ad hoc and trial-and-error data analysis**.

A lot of **copy & paste** between different systems and services.

Not an option for people **without coding experiences**.

Explore & Analyze

Jupyter Notebooks

Tableau

R

MS Access / MySQL

Excel

Various Clusters

distributed computing

Data Science in the Cloud @ SoundCloud

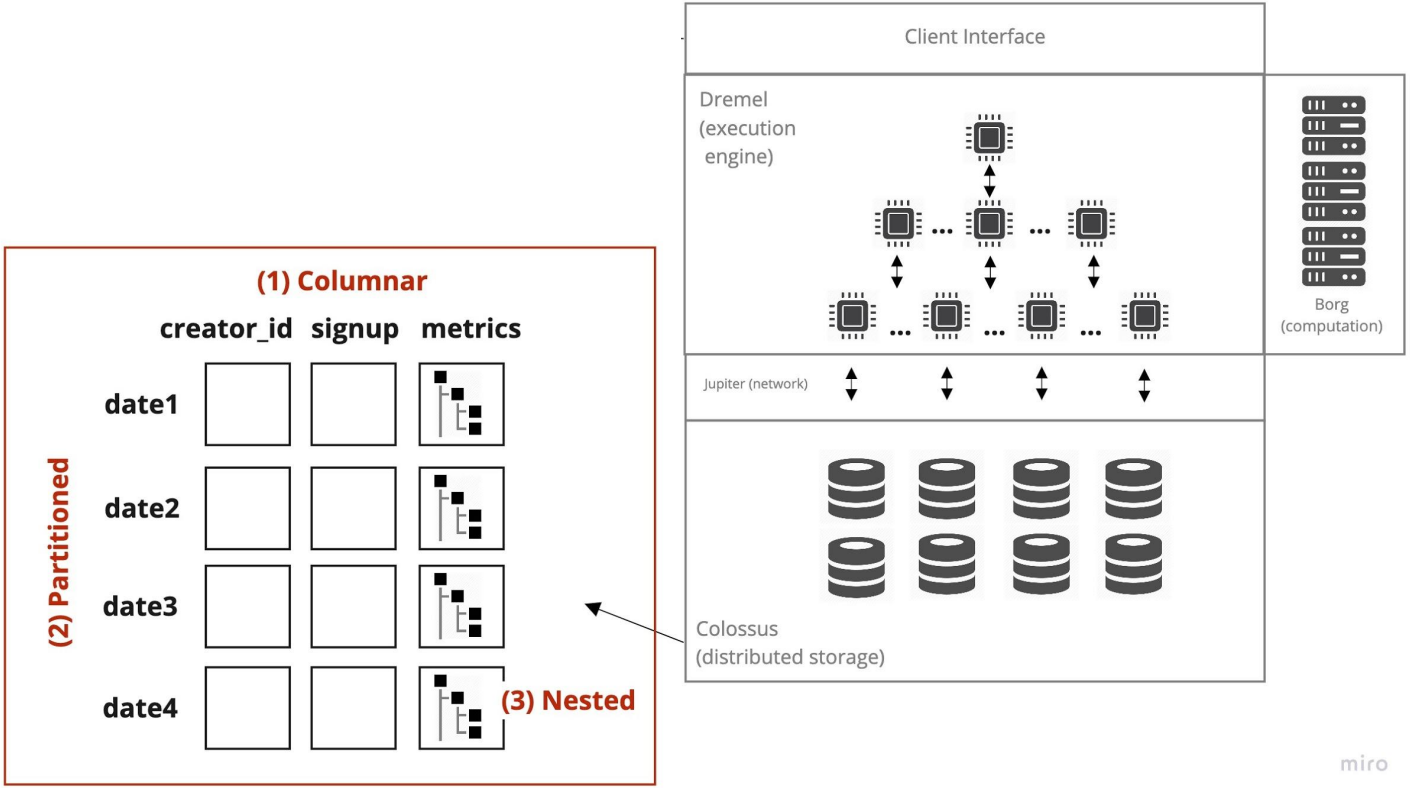
Collect	Ingest	Store	(Process)	Explore & Analyze
Client event tracking	Kafka Streams	BigQuery Storage	BigQuery SQL	BigQuery Web UI
Backlogs	Ingestion jobs	BigTable		Data Studio
...	(YARN/HDFS)			AI Platform

Google Cloud Platform

Kafka Cluster | Google Cloud Architecture

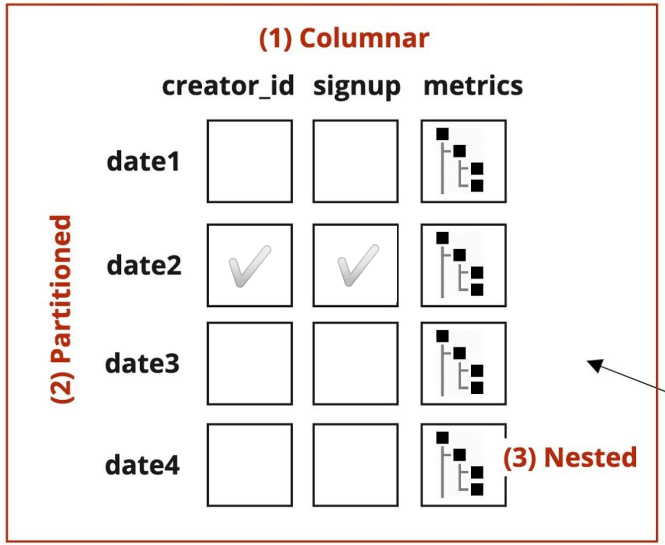
distributed computing

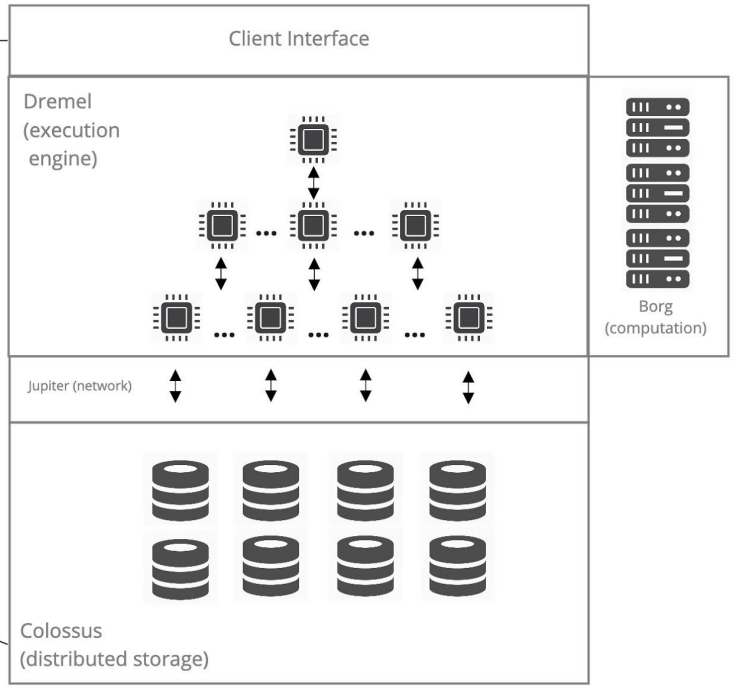
BigQuery Architecture



BigQuery Architecture

```
BigQuery SQL
SELECT
  creator_id,
  signup
FROM table
WHERE partition = 'date2'
```





Problem 2
The Data Itself

The Lake of Raw Data

Raw data has **different schemas**, analysing the data is challenging and error-prone.

Raw data from different systems can be **inconsistent**.

Raw data is **complex** and has a high level of detail.

Raw data is large, even with distributed computing it **takes some time** to get results.

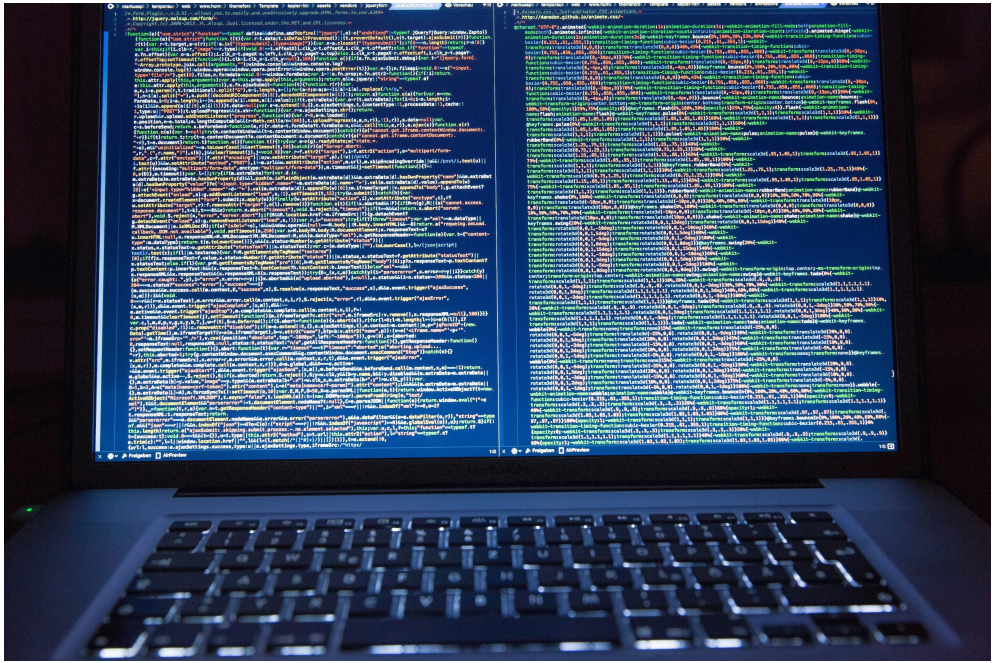
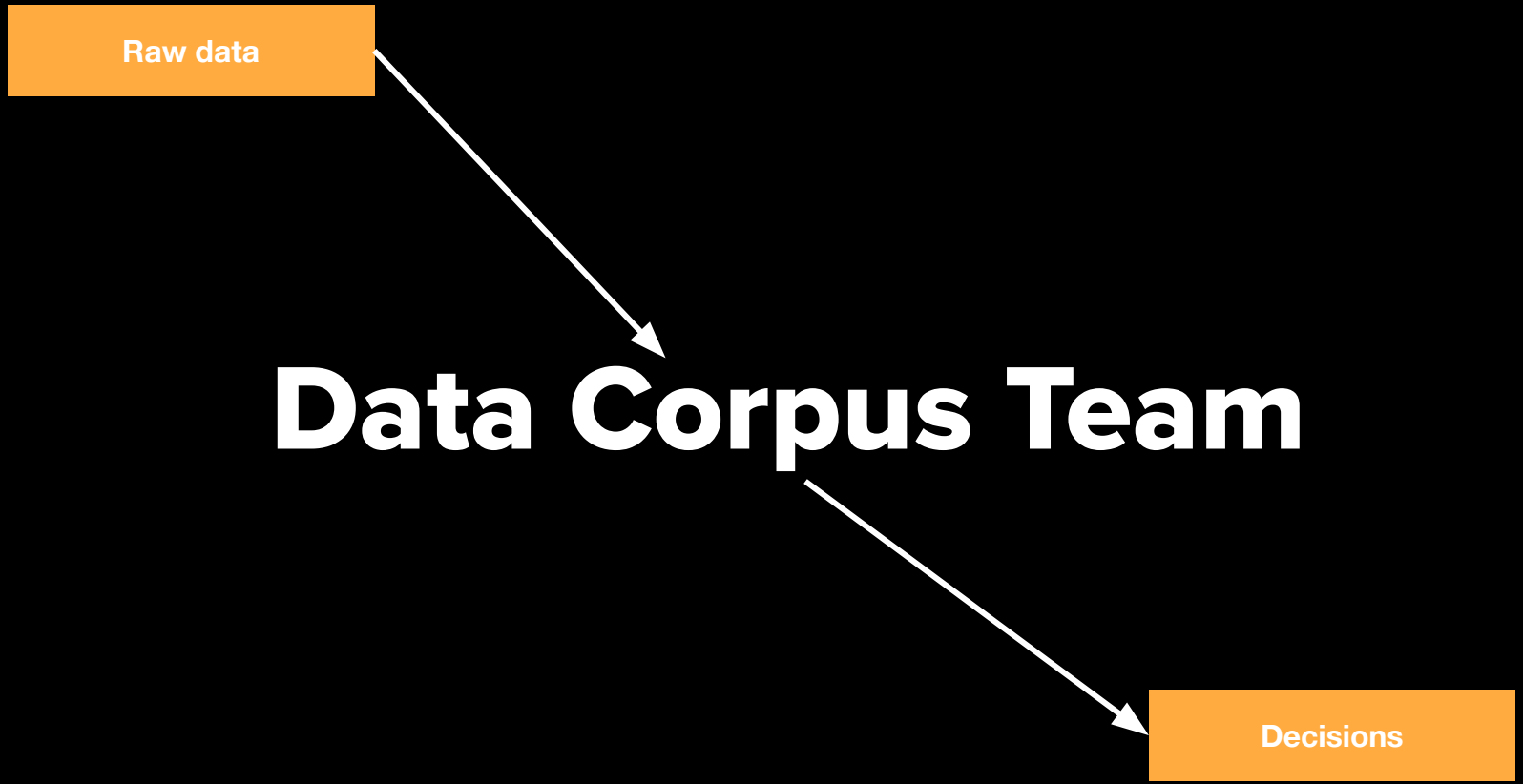


Photo by Markus Spiske from Pexels


Raw data
Data Corpus Team
Decisions

Mission Statement

Implement **data governance** on the datasets that are **most critical** for product and business decision making.

Definition of data governance

What characterizes high data quality?

1. Integrity
2. Compliance
3. Timeliness
4. Usability
5. Efficiency
6. Maintainability

The data quality principle are based on the ISO/IEC 25012 Data Quality model which defines a general data quality model for data retained in a structured format within a computer system. It focuses on the quality of the data as part of a computer system and defines quality characteristics for target data used by humans and systems. Integrity is a combination of accuracy and consistency. Maintainability is an additional principle originating from engineering.

What characterizes high data quality?

1. **Integrity**: Providing an unambiguous and validated source of truth.

- ✔️ Manual validations (when building something new)
- ✔️ Data quality checks that run daily
- ✔️ Anomaly detection
- ✔️ Unit testing

What characterizes high data quality?

2. **Compliance**: Ensuring compliance of data with GDPR.

What characterizes high data quality?

3. **Timeliness**: Daily updates are ready by 10am CET.

✔️ Efficiency of ETLs
✔️ Timeliness of dependencies

What characterizes high data quality?

4. **Usability**: Providing easily accessible and consumable data.

- ✔️ BQ table and field descriptions
- ✔️ Design documents
- ✔️ Ease of use of resources
- ✔️ Providing Training

What characterizes high data quality?

5. Efficiency: Optimize execution time and storage regarding performance and costs.

- ✔️ Storage
- ✔️ Execution time
- ✔️ Make the most of BQ features (sketches, nesting, UDFs)

What characterizes high data quality?

6. **Maintainability**: The code should be easy to read, extend and maintain.

- ✔️ 80/20 principle
- ✔️ Best practices
- ✔️ Scalability

Our tech stack

- BigQuery - data warehouse
- BigQuery SQL - querying language
- Airflow - scheduler
- terraform - resource management
- github - version control

Ready to use out of the box!

Learning to work with BQ (some examples)

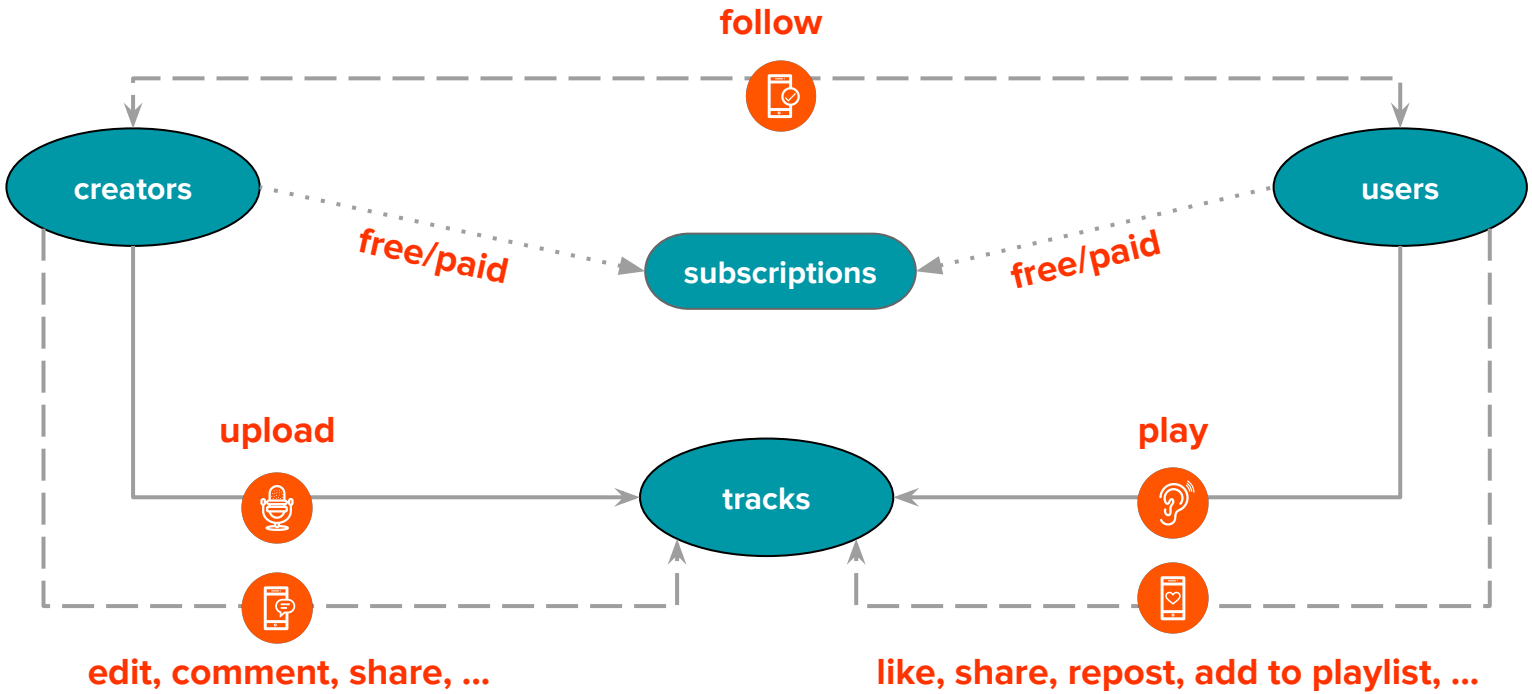
- The storage location of the data is relevant- we use EU, not US
- Make sure to partition your data, if possible
- SELECT only the columns you need
- Nested and repeated fields reduce storage and computational time
- UDFs (User Defined Functions) allow you to abstract away common coding patterns
- Wide tables perform better

Data Corpus

size: ~ 456 TB

90% less storage than raw data!

Data Modelling - Our Main Entities



daily.creator_metrics table

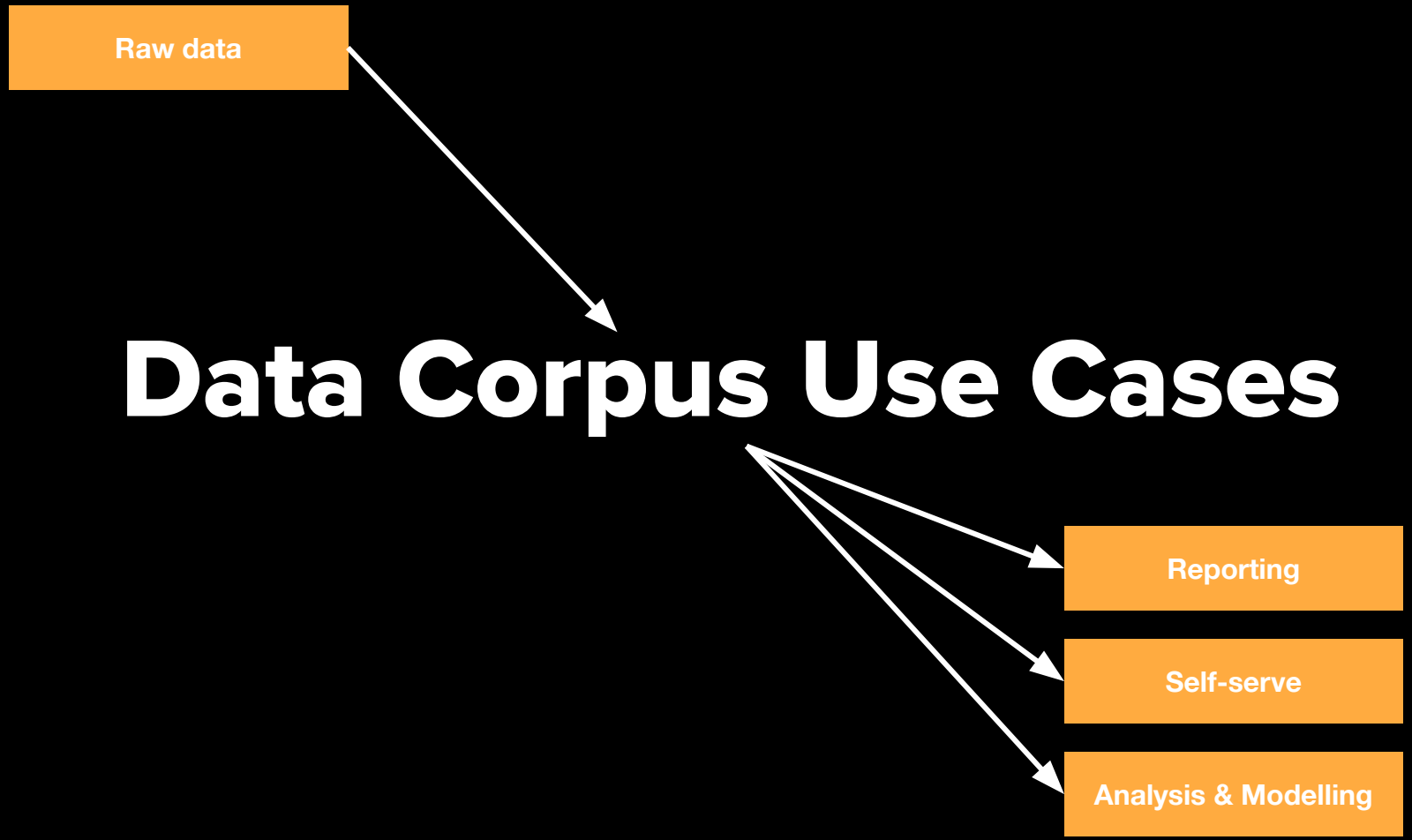
Each entity has a table with its main metrics.

nested and repeated fields

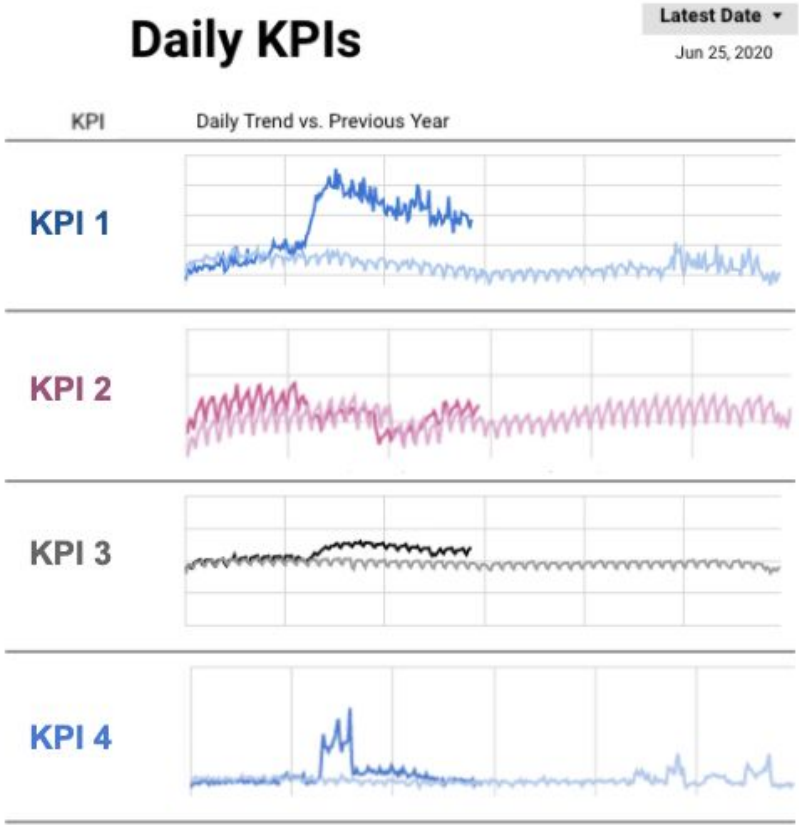
partition column

date	creator_id	has_record_deal	actions_received. is_logged_in	actions_received. metric_name	actions_received. metric_value
2020-05-20	soundcloud:creator:2	TRUE	FALSE	n_plays	10
			TRUE	n_plays	484
			TRUE	n_shares	1
			FALSE	listening_time_ms	75829
2020-05-20	soundcloud:creator:3	FALSE	TRUE	n_plays	199
			TRUE	listening_time_ms	29733
2020-05-20	soundcloud:creator:4	FALSE	FALSE	n_plays	9
			FALSE	listening_time_ms	28485

one row per creator (on each day, if they received an action)


Raw data
Data Corpus Use Cases
Reporting
Self-serve
Analysis & Modelling

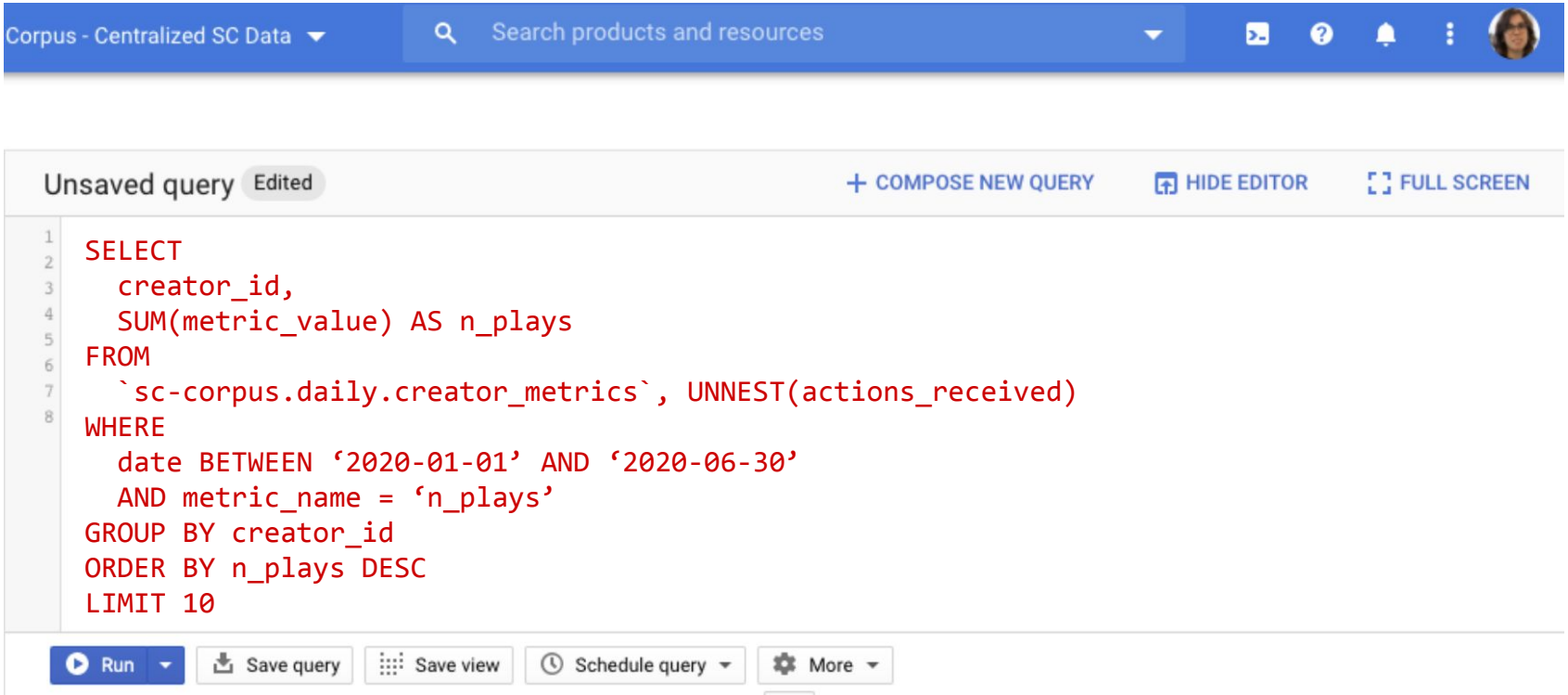
Reporting → Data Studio



- High level reporting
- Everyone in the company can view
- Connects directly to BQ
- KPI- Key Performance Indicator

Self-Serve → BigQuery Web UI

What were the **top 10 creators** for the first half of 2020 (in terms of number of plays)?



Analysis & Modelling → AI Platform Notebooks

- Managed JupyterLab notebook instances
- For the more advanced users
- You can use whatever environment you'd like
- You can also define how powerful your machine should be machine
- Connects seamlessly to BQ

Want to join?

- For this and more job opportunities, check out our jobs page.
- **Engineering trainee program**: DevBridge (Blog post about current edition).

Questions?