# Качество тестовых данных

#synthesized #data #data-quality #tools #sql #greatexpectations

Vsevolod Brekelov 04.09.2021 #rndtechconf

# Познакомимся

🙂

# Привет, я Сева! 👋

## Engineering Lead @Synthesized

@volekerb

@brekelov

# Привет, я Сева! 👋

Тестирование - Разработка - Менеджмент > 10 лет

Курс по прикладному тестированию в ИТМО

https://github.com/volekerb/testing-lectures

Канал с инженерными статьями, книгами, видео

https://t.me/engineerreadings

 @volekerb

 @brekelov

Конференции Heisenbug, JPoint, Joker


https://www.youtube.com/playlist?list=PLwvQQeADNQWcA0NdtL6_AEXl58Gf7ERuw

# 3 вопроса

# Для кого?

# О чем?


Зачем?

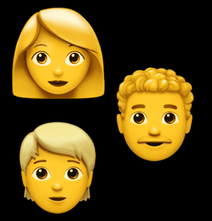
# Почему Data Quality кого-то волнует?

И что это?

Инженер данных

Исследователь данных

Аналитик данных



Бизнес-пользователи

Запросы

- Большие наборы данных
- Бизнес правила

Запросы

- Вопросы
- Обновления

# Companies lose an average of $15M per year due to bad data

https://hbr.org/2020/02/data-driven-decisions-start-with-these-4-que

“With the goal of building and achieving data quality standards across Uber, we have supported over 2,000 critical datasets on this platform, and detected around 90% of data quality incidents.”

https://eng.uber.com/operational-excellence-data-quality/

https://eng.uber.com/operational-excellence-data-quality/

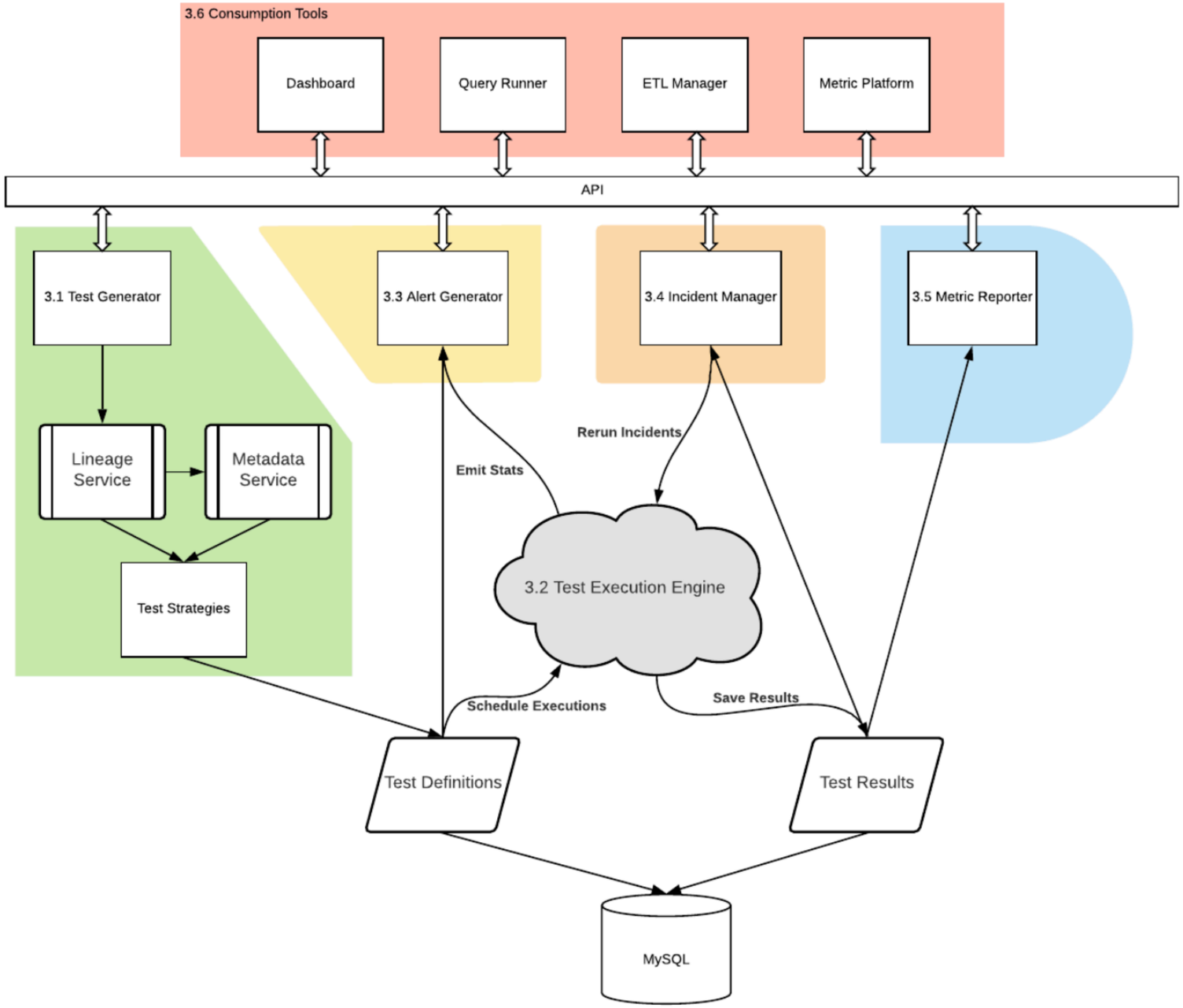
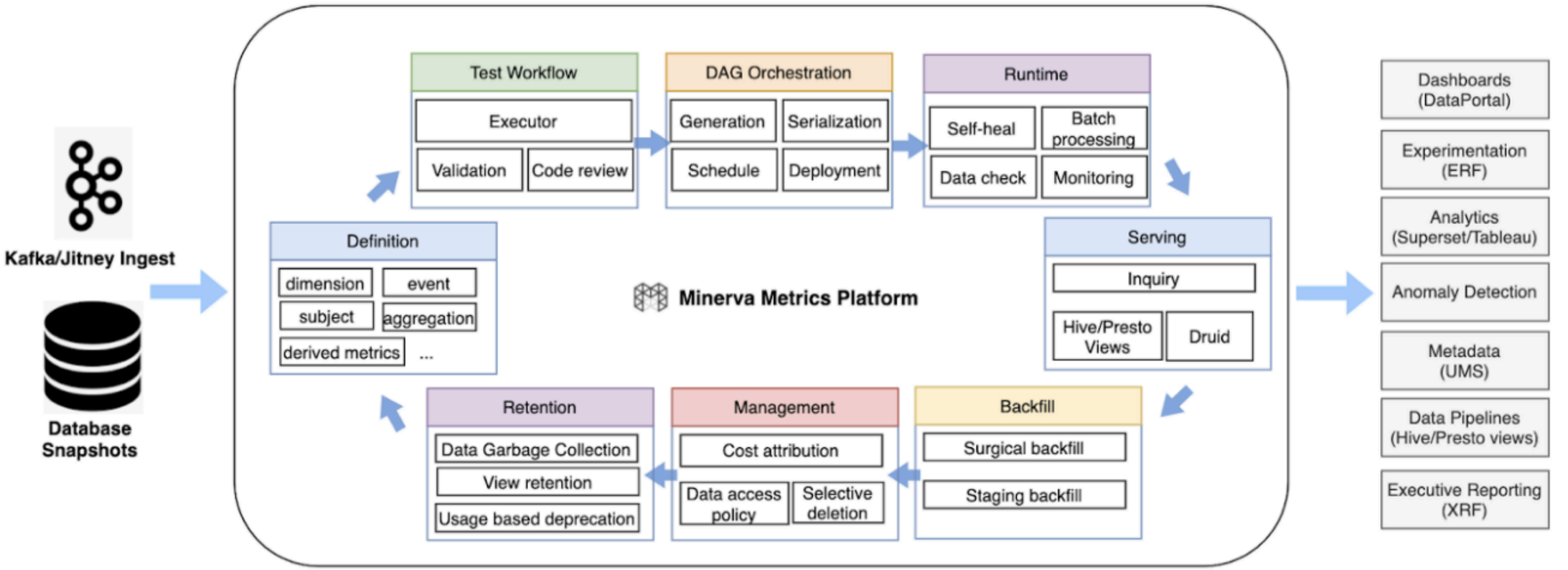



*Figure 1: Data Quality Platform architecture*

“For data consumption, we heard complaints from decision makers that different teams reported different numbers for very simple business questions, and there was no easy way to know which number was correct.”

https://medium.com/airbnb-engineering/how-airbnb-achieved-metric-consistency-at-scale-f23cc53dea70




Minerva manages the entire lifecycle of metrics at Airbnb.

# Data Quality

## Что это?

- **Accuracy**: Is the data correct?
- **Consistency**: Is everybody looking at the same data?
- **Usability**: Is data easy to access?
- **Timeliness**: Is data refreshed on time, and on the right cadence?
- **Cost Efficiency**: Are we spending on data efficiently?
- **Availability**: Do we have all the data we need?

- **Accuracy**: Is the data correct?

  “Корректность” может быть у каждого разной

- **Consistency**: Is everybody looking at the same data?

Скорее речь о том, насколько данные консистенты с теми, что были в моменте “до” или в прошлом

- **Usability**: Is data easy to access?

  Хорошо. Но с качеством данных не очень поможет =)

- **Timeliness**: Is data refreshed on time, and on the right cadence?

Звучит очень хорошо!

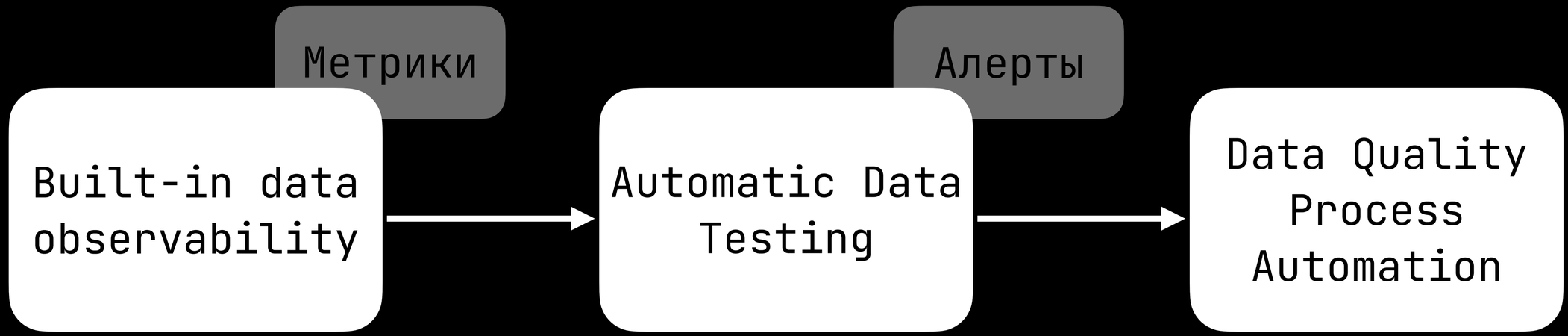
- **Cost Efficiency**: Are we spending on data efficiently?

Можно годами обсуждать с финансовыми отделами

- **Availability**: Do we have all the data we need?

  Это скорее про то, что доступны ли данные нам

# #DataOps or #DataObservability



# Решения в больших компаниях

| | |
|---|---|
| Amundsen (Lyft) | Hive, Redshift, Druit, RDBMS, Presto, |
| Datahub (LinkedIn) | Hive, Kafka, RDBMS |
| Metacat (Netflix) | Hive, RDS, Teradata, Redshift, S3, Ca |
| Atlas (Apache) | HBase, Hive, Sqoop, Kafka, Storm |
| Marquez (Wework) | S3, Kafka |
| Databook (Uber) | Hive, Vertica, MySQL, Postgress, Cass |
| Dataportal (Airbnb) | Unknown |
| Data Access Layer (Twitter) | HDFS, Vertica, MySQL |
| Lexikon (Spotify) | Unknown |

# А что если нам это все сразу не надо?

# По-порядку!

- Приватность
- Битые данные
- Недостаточные данные
- Тестирование данных (Может есть что-то типа JUnit?)

# По-порядку!

- Приватность
- Битые данные
- Недостаточные данные
- Тестирование данных (Может есть что-то типа JUnit?)

Продакшн данные

Мне нужны реальные данные,
тк я тут такую машин лернинг модель сделал УУУХХ!

Мне нужны исторические данные,
хочу угадать, когда надо предлагать кредит!

Мне нужно протестировать систему и точно понять,
что все кейсы ок проходят!

Продакшн данные

Мне нужны реальные данные,
тк я тут такую машин лернинг модель сделал УУУХХ!

# Это все хорошо

Мне нужны исторические данные,
хочу угадать, когда надо предлагать кредит!

# Но доступ мы, конечно, вам не дадим

Мне нужно протестировать систему и точно понять,
что все кейсы ок проходят!

Продакшн данные

Анонимизируем

Данные, но уже не оч похожие на продакшн

Данные, но уже не оч похожие на продакшн

Моя машин лернинг модель чет плохо работает =(

Чет я звоню, а этим людям не нужен кредит =(

Чет тут не все кейсы проходят.
А можно заново сделать, у нас там схема поменялась?

MOSTLY·AI

SYNTHESIZED

SYNTHO

hazy

Synthesis.ai

# Синтетические данные

Продакшн данные

Синтезируем

Данные ведут себя как продакшн данные

О! Все заработало!

Данные ведут себя как продакшн данные

Во! Теперь можно хорошо угадать!

Все быстро сгенерилось и могу сам обновить данные!

Теперь не найти реальные телефоны и адреса и номера кредиток.
Как же я вам дозвонюсь?

# По-порядку!

- Приватность
- Битые данные
- Недостаточные данные
- Тестирование данных (Может есть что-то типа JUnit?)

# По-порядку!

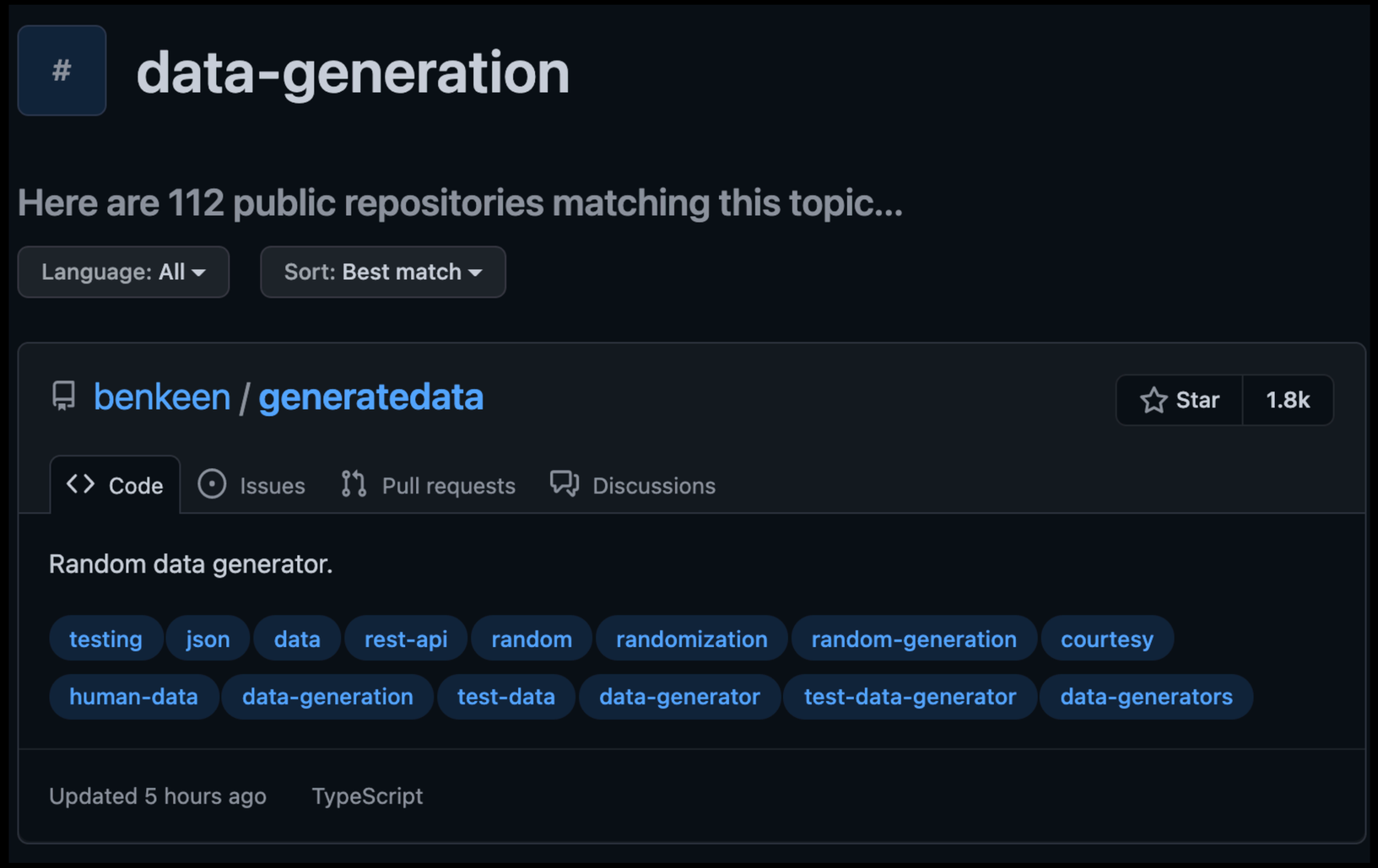
- Приватность
- Битые данные
- Недостаточные данные
- Тестирование данных (Может есть что-то типа JUnit?)

# Когда нужно генерировать данные?

- Юнит/компонентные/интеграционные тесты
- Интеграционные/Системные тесты с production-like системой

- Переехали на новый стек технологий (не было тестов, требований. Все как обычно)
- Только стартуем проект и нет продакшна

# https://github.com/topics/data-generation

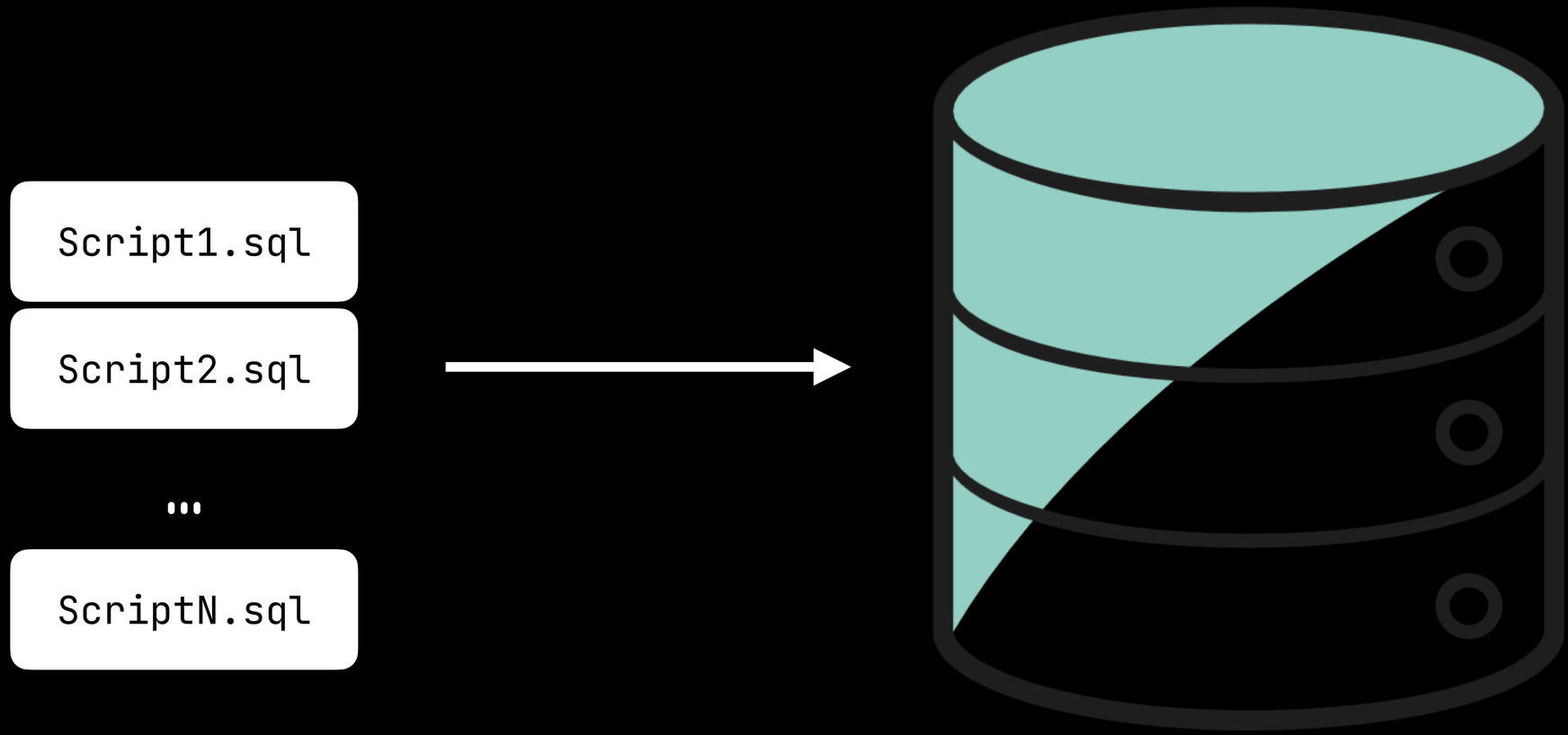


# Юнит тесты

```
@Test

void givenDefaultConfiguration_thenGenerateSingleObject() {

    EasyRandom generator = new EasyRandom();

    Person person = generator.nextObject(Person.class);



    assertNotNull(person.getAge());

    assertNotNull(person.getFirstName());

    assertNotNull(person.getLastName());

}
```

```
Person[firstName='eOMtThyhVNLWUZNRcBaQKxI',
lastName='yedUsFwdkelQbxeTeQOvaScfqIOOmaa', age=-1188957731]
```

# Системные / Интеграционные тесты



Написать

Накатить

Поддерживать

Script1.sql

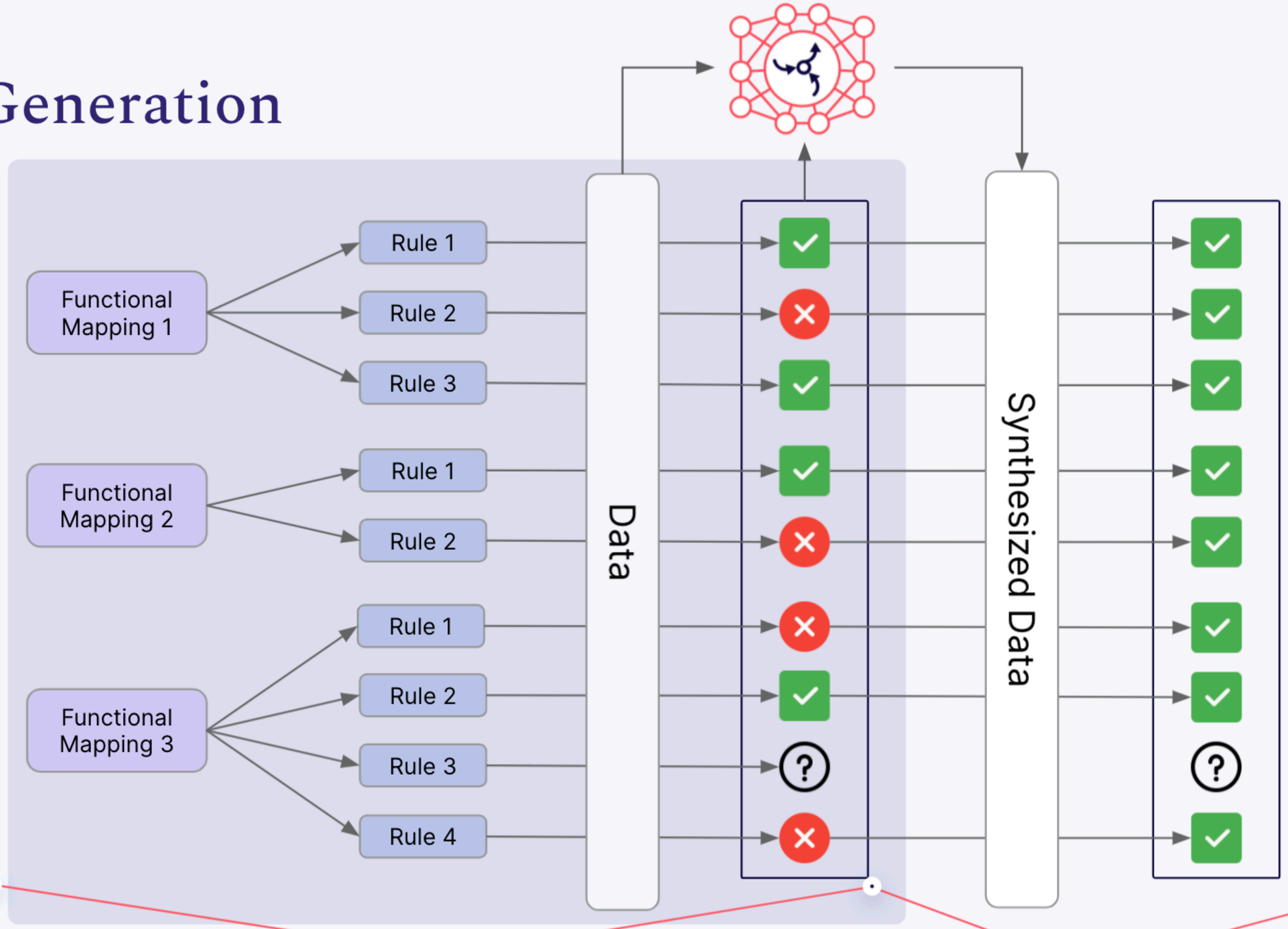
Script2.sql

…

ScriptN.sql

Угадать все сценарии =)

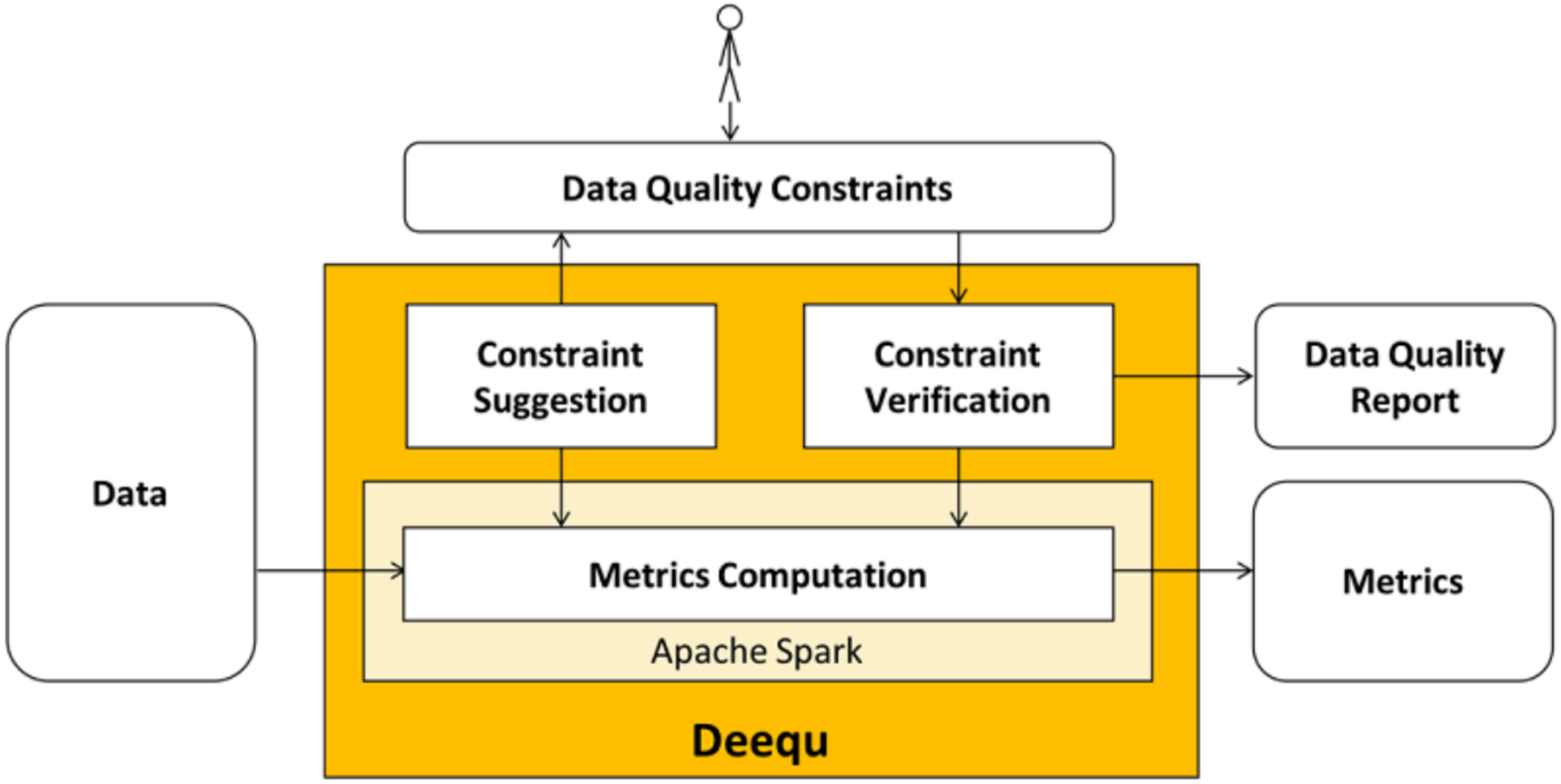
# Data Generation



# По-порядку!

- Приватность
- Битые данные
- Недостаточные данные
- Тестирование данных (Может есть что-то типа JUnit?)

# AWS Deequ




Source — https://aws.amazon.com/blogs/big-data/test-data-quality-at-scale-with-deequ/

```
val verificationResult = VerificationSuite()
  .onData(data)
  .addCheck(Check(CheckLevel.Error, "unit testing my data")
  .hasSize(_ == 5) // we expect 5 rows
  .isComplete("id") // should never be NULL
  .isUnique("id") // should not contain duplicates
  .isComplete(“productName") // should never be NULL
  .isContainedIn(“priority", Array("high", "low"))
  .isNonNegative("numViews") // should not contain negative values
  .containsURL("description", _ >= 0.5)
  .hasApproxQuantile("numViews", 0.5, _ <= 10))
.run()
```

- Persistence and querying of computed metrics of the data with a MetricsRepository
- Data profiling of large data sets
- Anomaly detection on data quality metrics over time
- Automatic suggestion of constraints for large datasets
- Incremental metrics computation on growing data and metric updates on partitioned data (advanced)

# Apache griffin

The DQ config file: dq.json

```
{
  "name": "batch_accu",
  "process.type": "batch",
  "data.sources": [
    {
      "name": "src",
      "baseline": true,
      "connectors": [
        {
          "type": "hive",
          "version": "1.2",
          "config": {
            "database": "default",
            "table.name": "demo_src"
          }
        }
      ]
    }, {
      "name": "tgt",
      "connectors": [
        {

          "type": "hive",
          "version": "1.2",
          "config": {
            "database": "default",
            "table.name": "demo_tgt"
          }
        }
      ]
    }
```

https://griffin.apache.org/

# Great Expectations

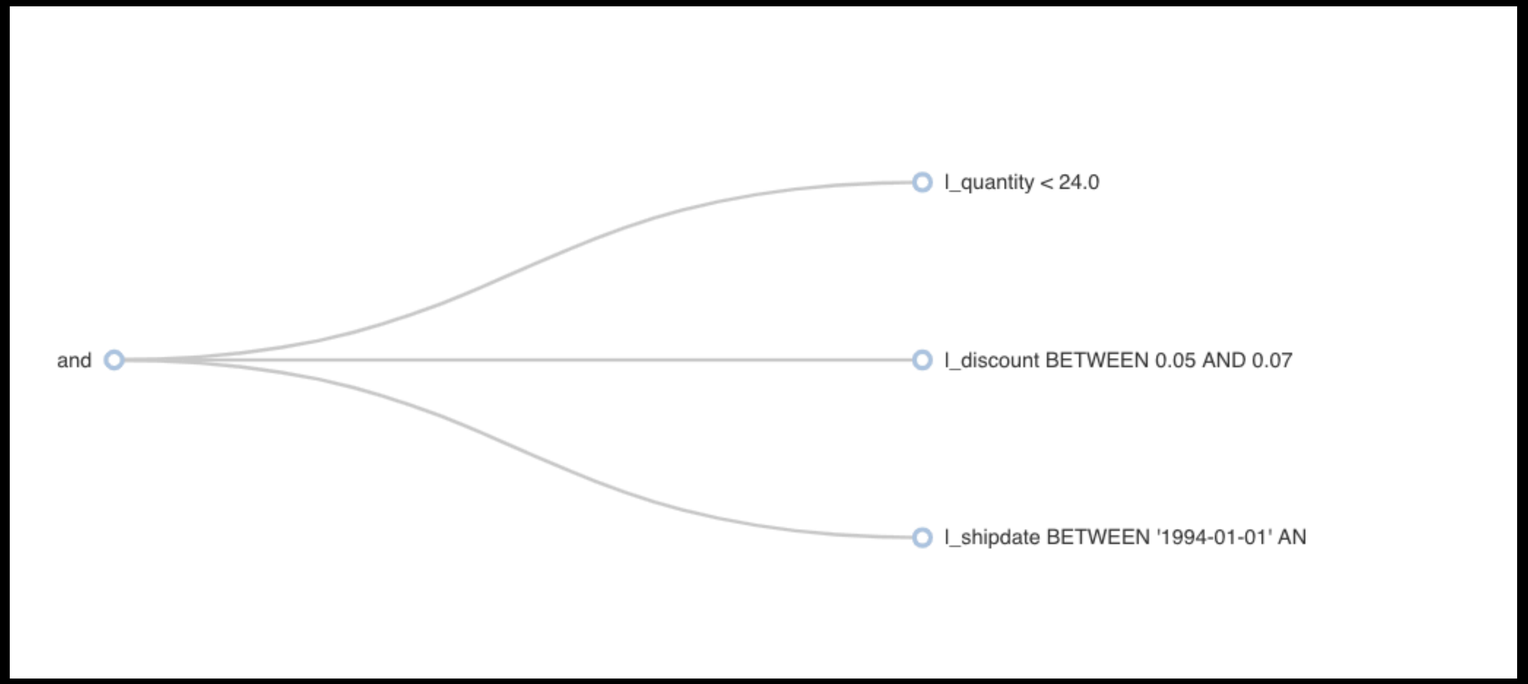
- expect_column_values_to_not_be_null
- expect_column_values_to_match_regex
- expect_column_values_to_be_unique
- expect_column_values_to_match_strftime_format
- expect_table_row_count_to_be_between
- expect_column_median_to_be_between

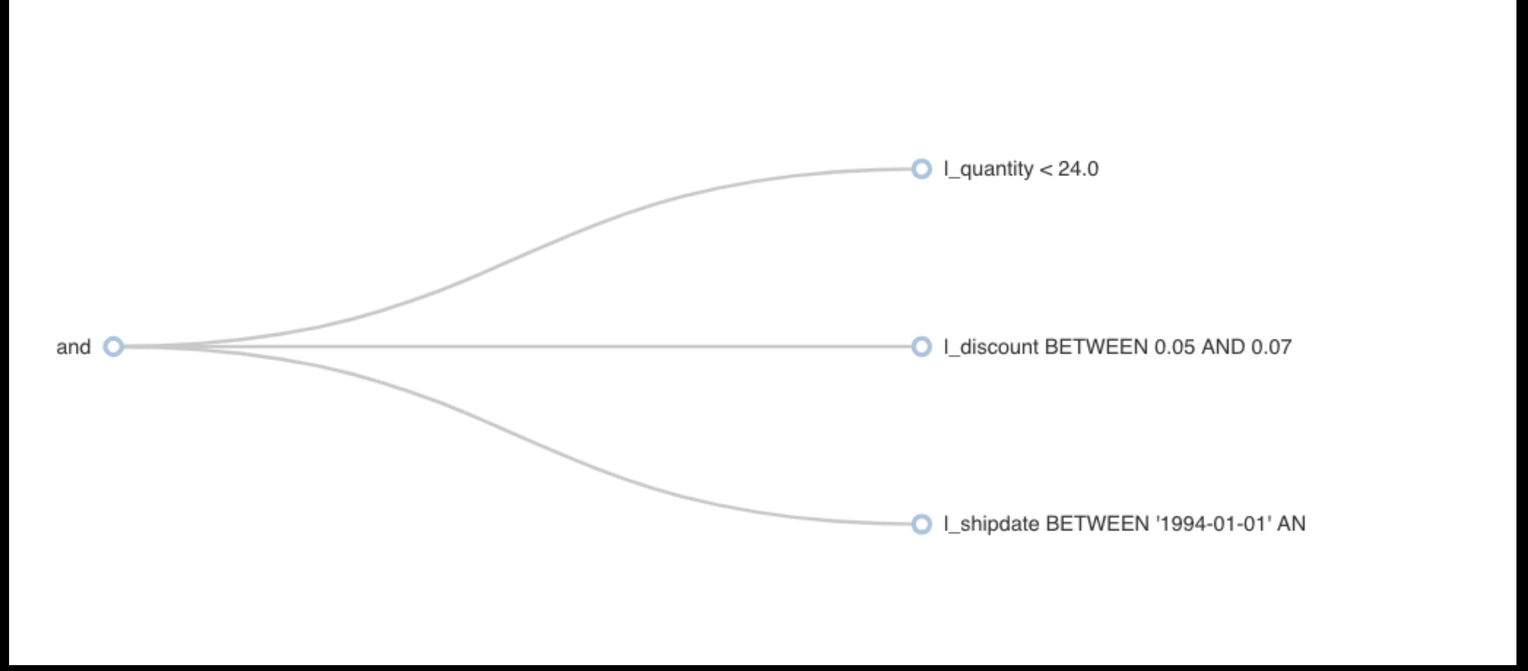
https://github.com/great-expectations/great_expectations

# Бонус

# SQL анализ

```sql
SELECT
    *
FROM
    lineitem
WHERE
    l_shipdate BETWEEN '1994-01-01' AND '1995-01-01'
    AND l_discount BETWEEN 0.05 AND 0.07
    AND l_quantity < 24;
```






+

Датасет / Данные

→ найти и догенерить недостающие


https://holistic.dev/playground/f9a504e1-6bd4-4464-a951-87b3b9db6e69

https://dwh.dev/report/summary

# Надеюсь, достаточно пищи для ума

# Приходите к нам работать! 😉

https://www.synthesized.io/open-positions/kotlin-engineer

# Спасибо! 👋

 @volekerb

 @brekelov

# На почитать

- https://www.vldb.org/pvldb/vol11/p1781-schelter.pdf
- http://giis.uniovi.es/testing/papers/stvr-2010-sqlfpc.pdf
- http://giis.uniovi.es/testing/#stvr10sqlfpc
- https://francois-nguyen.blog/2021/03/07/towards-a-data-mesh-part-1-data-domains-and-teams-topologies/
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.1676&rep=rep1&type=pdf
- https://databricks.com/blog/2020/03/04/how-to-monitor-data-stream-quality-using-spark-streaming-and-delta-lake.html
-