Data Quality? Yes Please!

Bovard Doerschuk-Tiberi

PNSQC 2017

AKA “More Data Quality Please!”

What is Data Quality?

What is Data Quality?

There are many definitions of data quality but data is generally considered high quality if it is "fit for [its] intended uses in operations, decision making and planning."

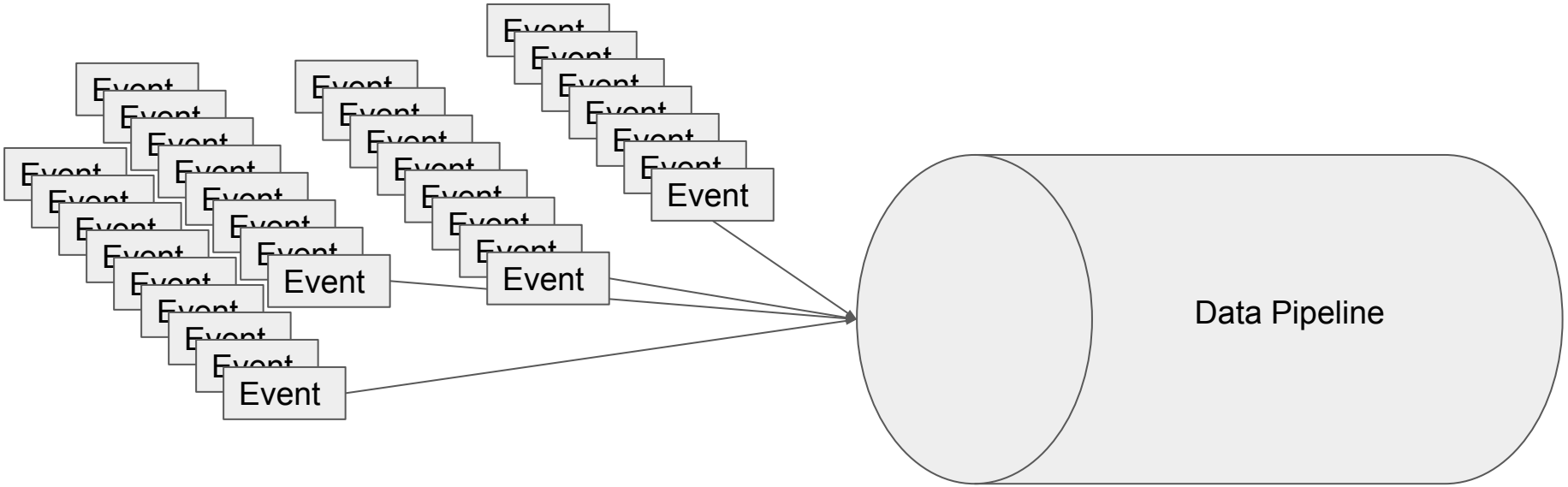
- Wikipedia

Our Focus

Ensuring data quality in an event driven architecture

Our Focus

Ensuring data quality in an event driven architecture

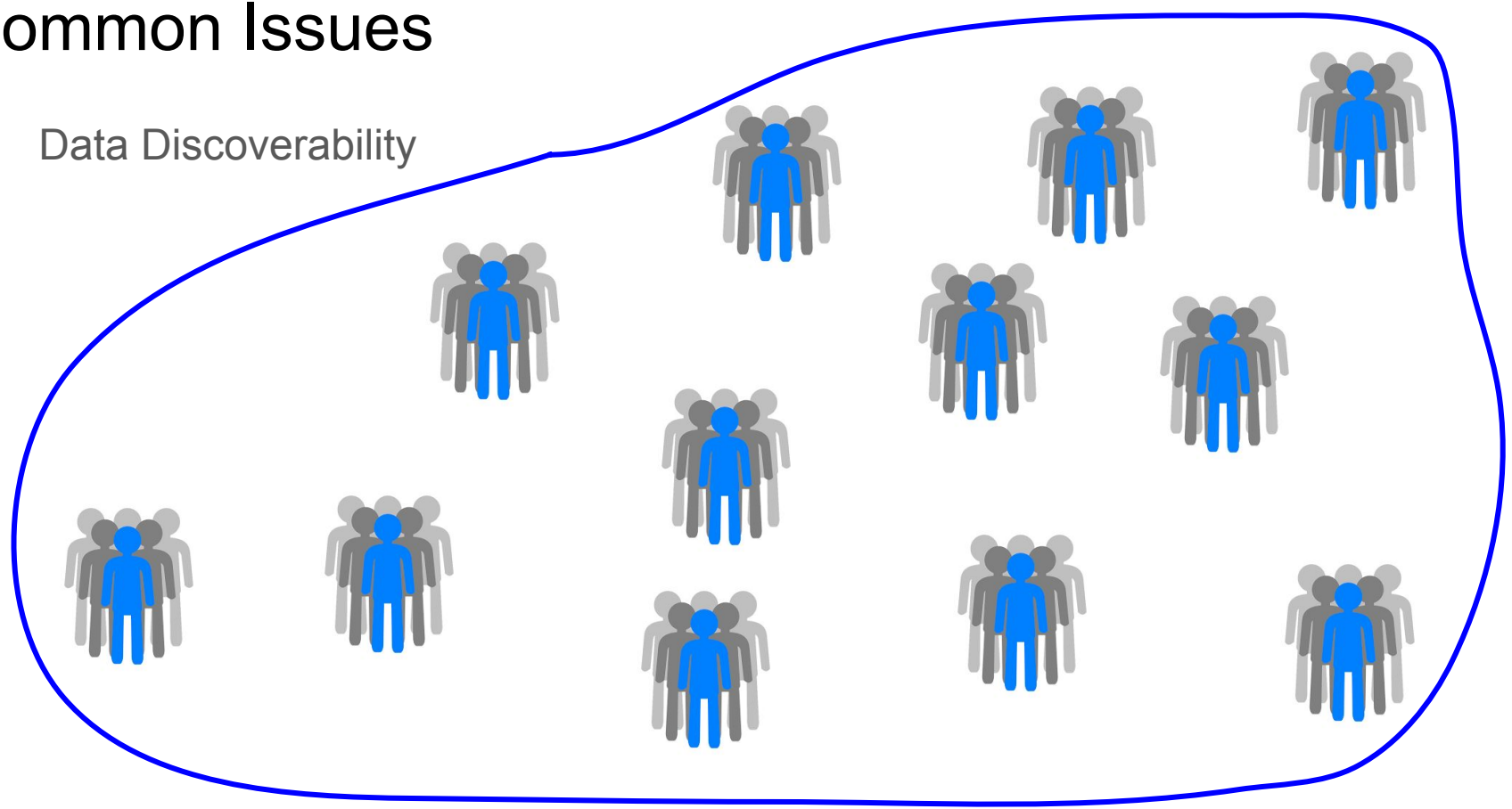


Common Issues

- Data Discoverability

Common Issues

- Data Discoverability

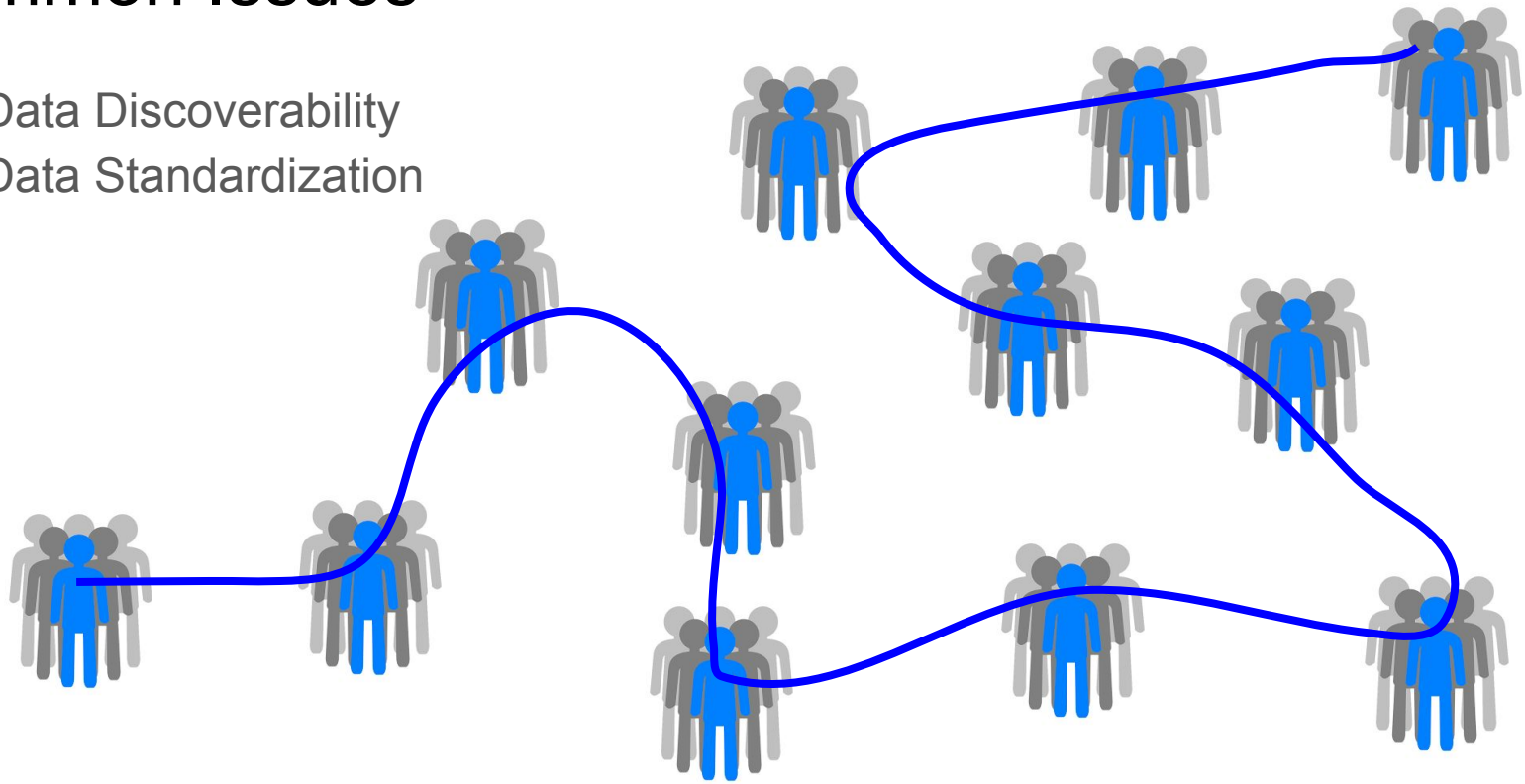


Common Issues

- Data Discoverability
- Data Standardization

Common Issues

- Data Discoverability
- Data Standardization



Common Issues

- Data Discoverability
- Data Standardization
- Data Completeness

Data Completeness

Complete

```
accout_id: 312353

user_id: 189678

document_id: 235789

timestamp: 1499486132

action: “save”
```

Not Complete

```
accout_id: null

user_id: “189678”

document_id: “TODO”

timestamp: -1499486132

action: [“save”]

extra_dimension: “lol!”
```

Common Issues

- Data Discoverability
- Data Standardization
- Data Completeness

Solution!

A central place…

Solution!

A central place…

with a list of all events...

Solution!

A central place…

with a list of all events…

a common set of dimensions...

Solution!

A central place…

with a list of all events…

a common set of dimensions…

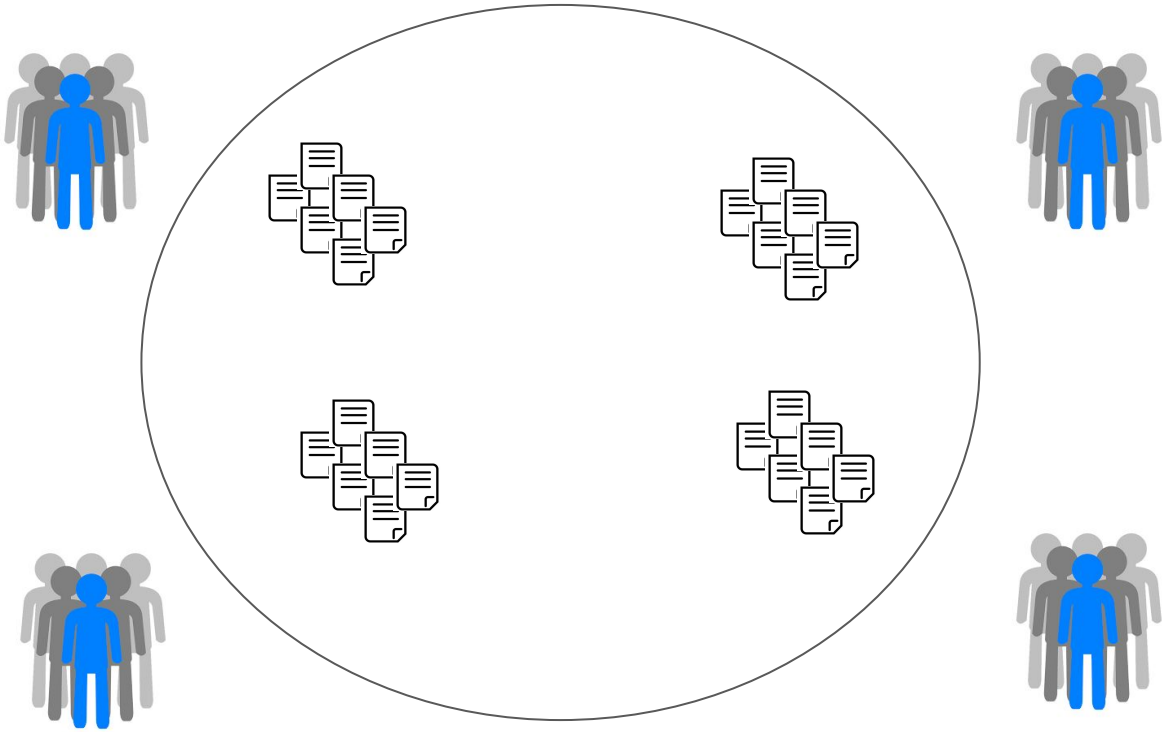
and a well defined event definition...

Strongly `Typed` Data

Strongly `Typed` Data

(all aboard the strongly typed hype train!)

A central location for event schema



Event Schema

A list of dimension definitions

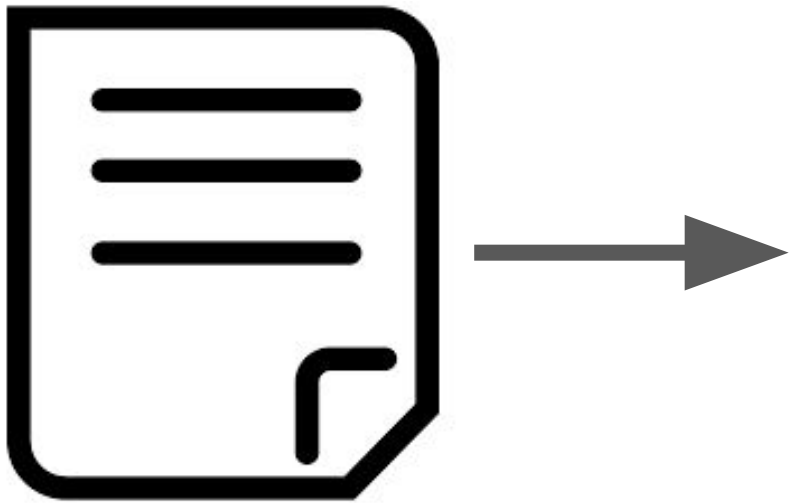
```
dimension: user_id

    type: string

    required: true

    description: id of the user
```

Event Schema

Event: Document Save

- **accout_id**: required, int
- **user_id**: required, int
- **document_id**: required, int
- **timestamp**: required, int

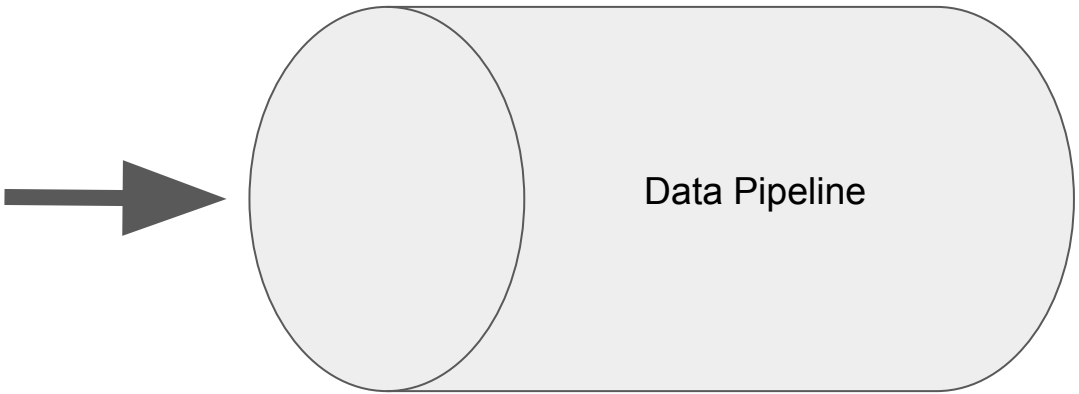
Event Schema Compilation



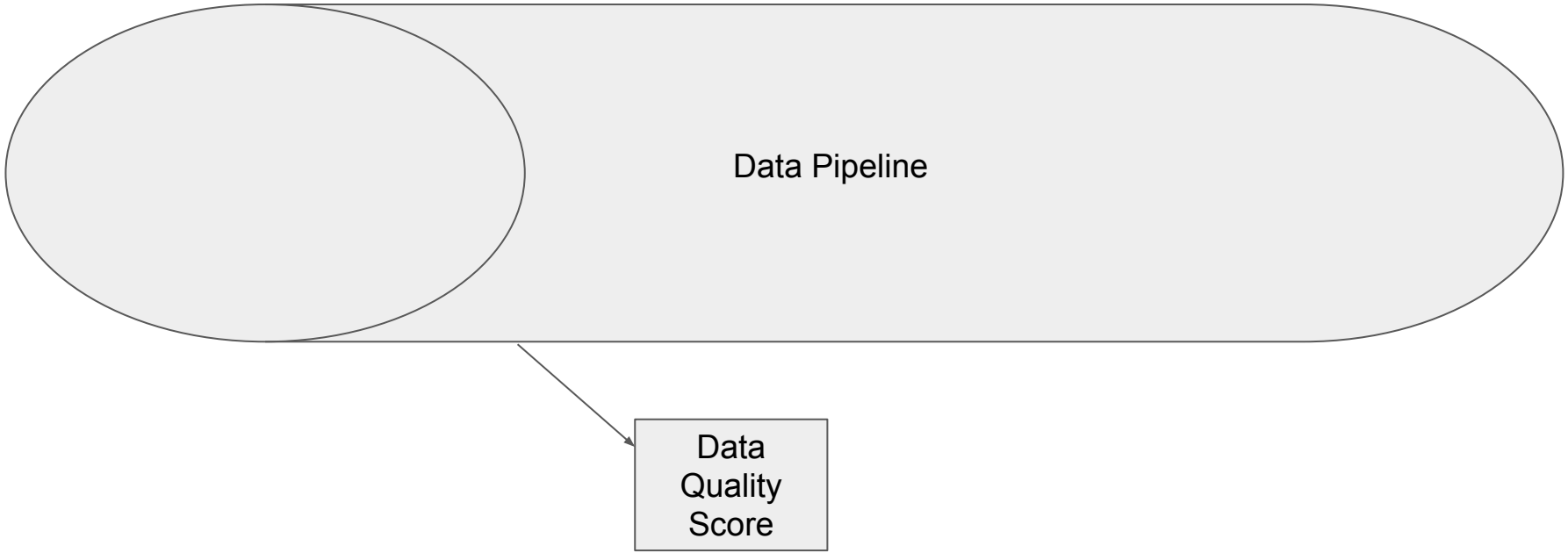
```
public DocumentSaveEvent(
    int account_id,
    int user_id,
    int timestamp,
    int doc_id
)
```

Sending Events

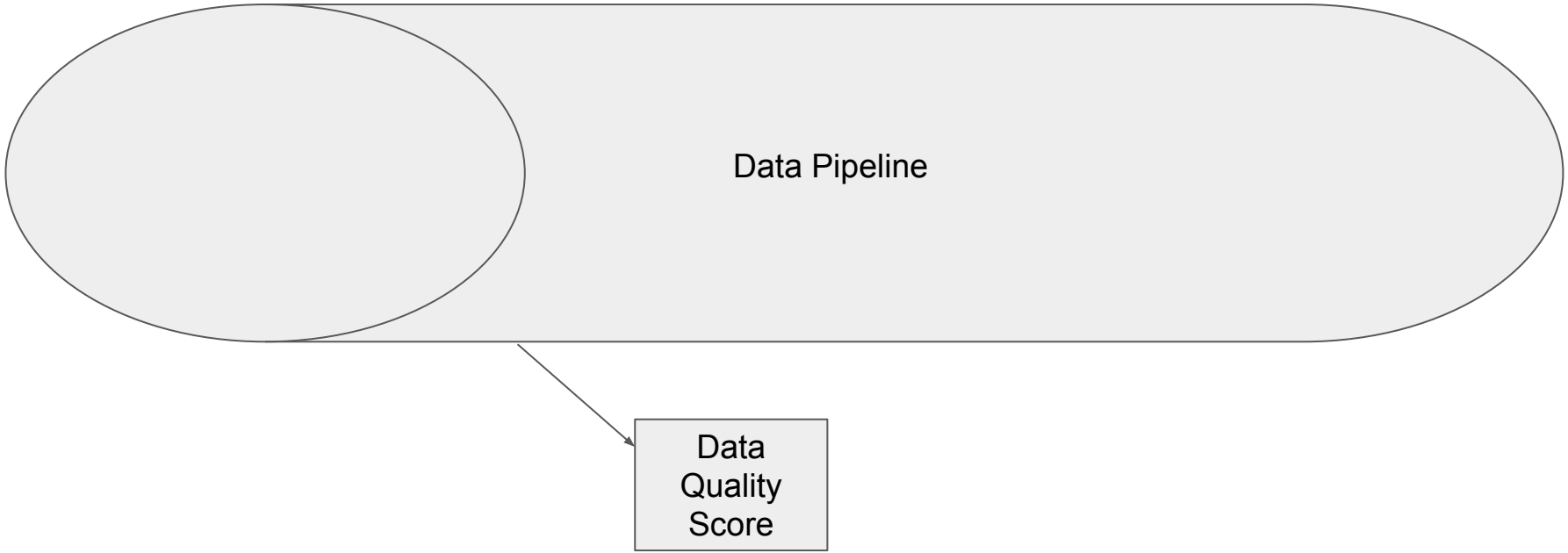
```
client.send(DocumentSaveEvent(
    account_id=689345,
    user_id=250983,
    timestamp=date.now(),
    doc_id=233584
))
```



Sample and check



Sample and check



Data Quality Score

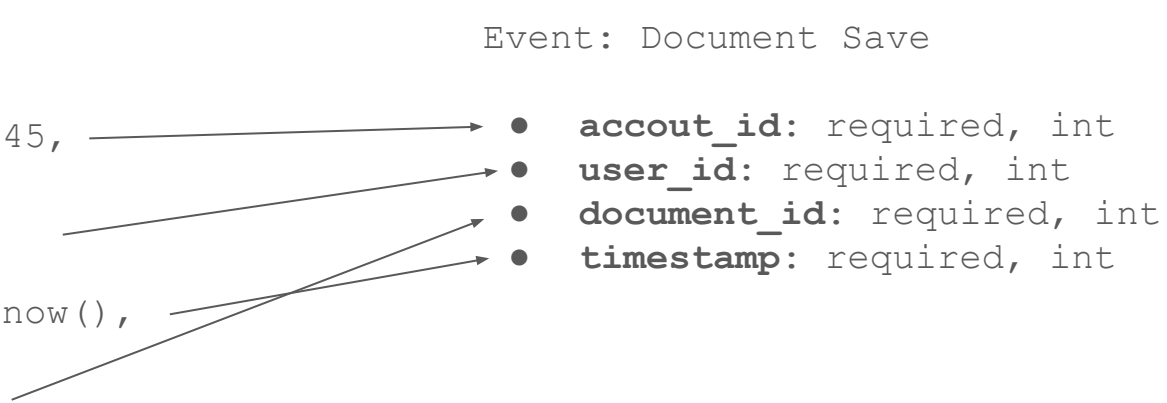
```
DocumentSaveEvent(

    account_id=689345,

    user_id=250983,

    timestamp=date.now(),

    doc_id=233584

)
```

Event: Document Save

- **accout_id**: required, int
- **user_id**: required, int
- **document_id**: required, int
- **timestamp**: required, int

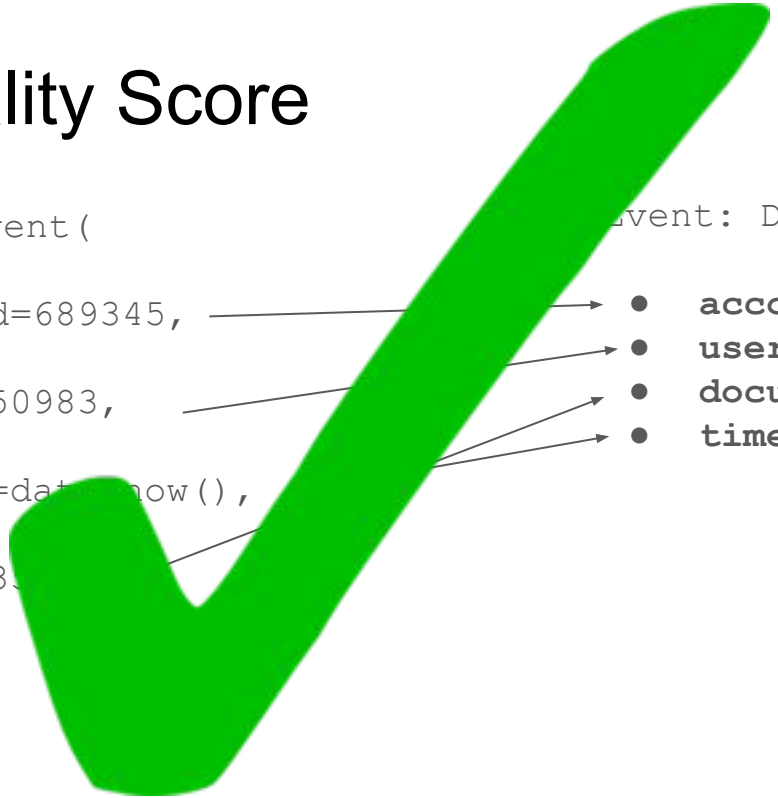
Data Quality Score

```
DocumentSaveEvent(

    account_id=689345,

    user_id=250983,

    timestamp=date.now(),

    doc_id=233584

)
```



Data Quality Score

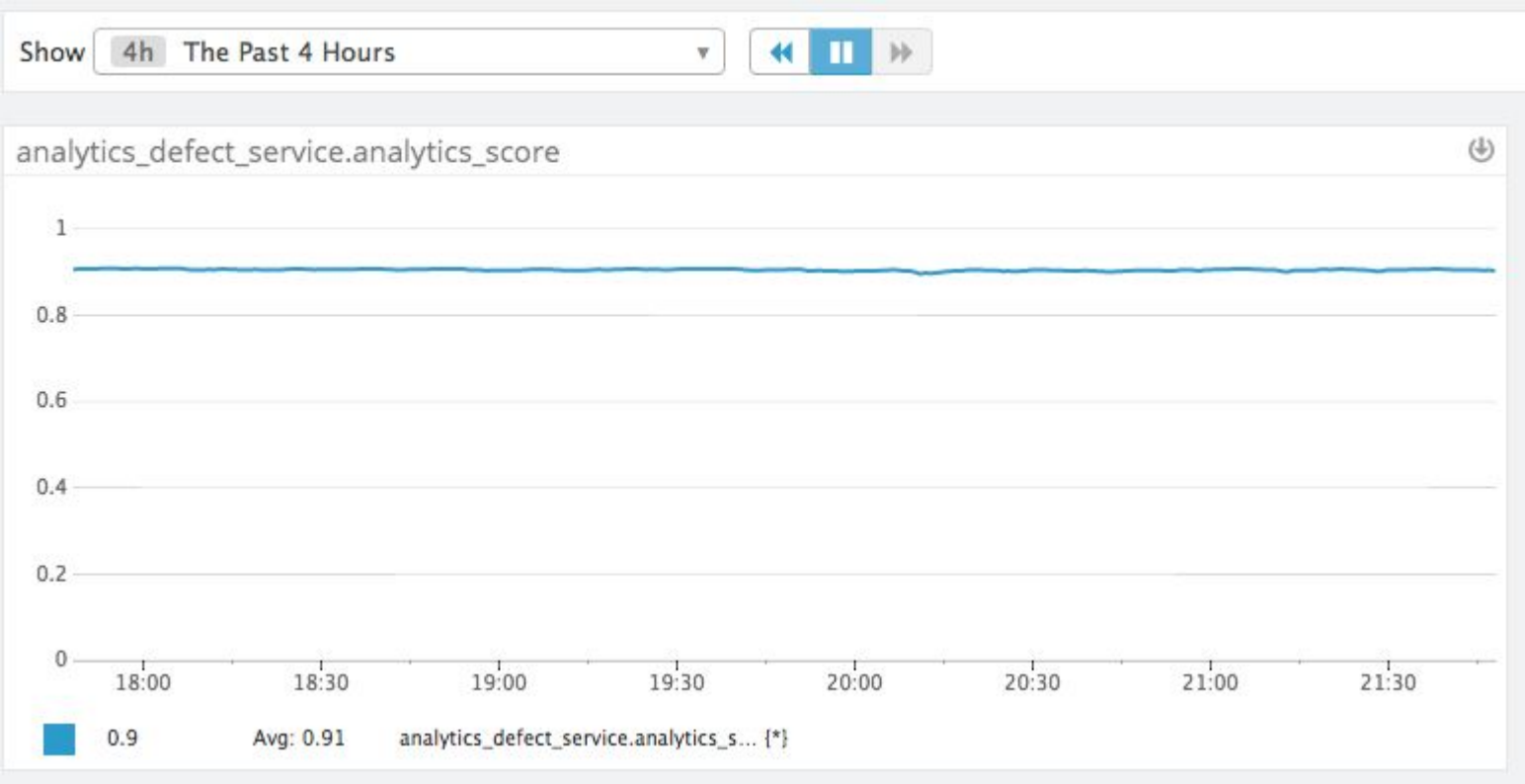
```
DocumentSaveEvent(

    account_id=689345,

    user_id=250983,

    timestamp=dat    now(),

    doc_id=233

)
```



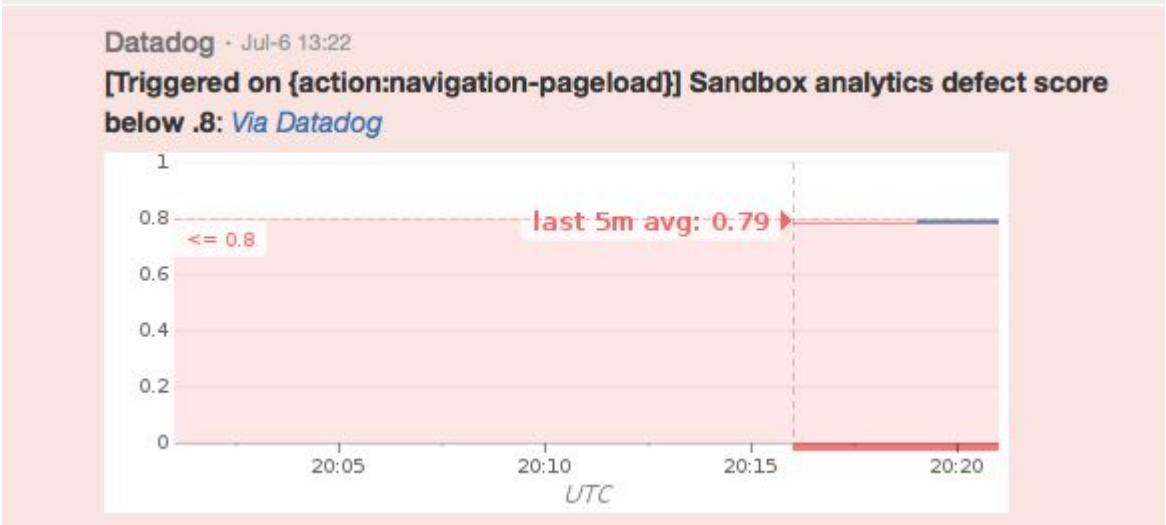
Event: Document Save

- **accout_id**: required, int
- **user_id**: required, int
- **document_id**: required, int
- **timestamp**: required, int

Track this score



Alert on this score



Schema Registry

- An accessible, central, searchable, versioned center of all analytics
- Schema’s compile to class definitions
- Clients send these
- Check them for quality

Demo Time!

Time for a demo!