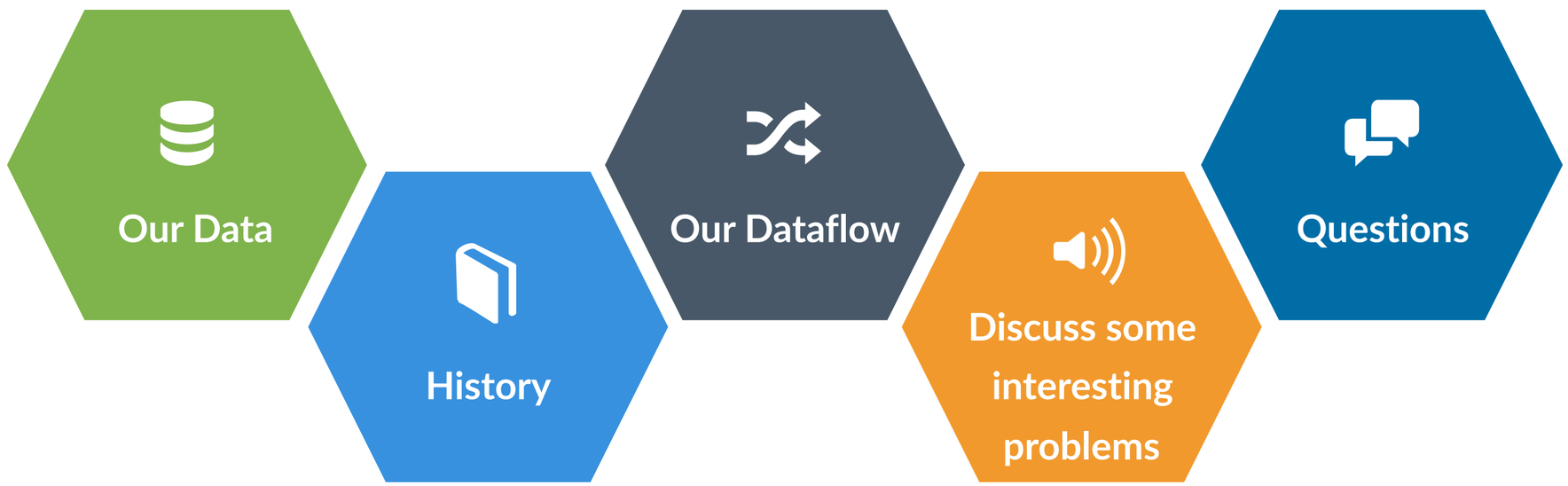
Standardized Data Management

Data Ingestion Platform

Arun Manivannan
Senior Data Engineer

Standard Chartered

So, what do we do now?



Data in SCB context

1. Hundreds of applications (~180 data lake source apps)
2. 50+ countries
3. Variety of applications - Web, Batch, Mainframes
4. Variety of consumption patterns
5. Multi-regulatory guided data storage

Data in SCB context

1. Hundreds of applications (~180 data lake source apps)
2. 50+ countries
3. Variety of applications - Web, Batch, Mainframes
4. Variety of consumption patterns
5. Multi-regulatory guided data storage

History

<2012

Traditional group-wise Data Warehouses

2012

Enterprise Data Management on Teradata

2014

EDM on Hadoop

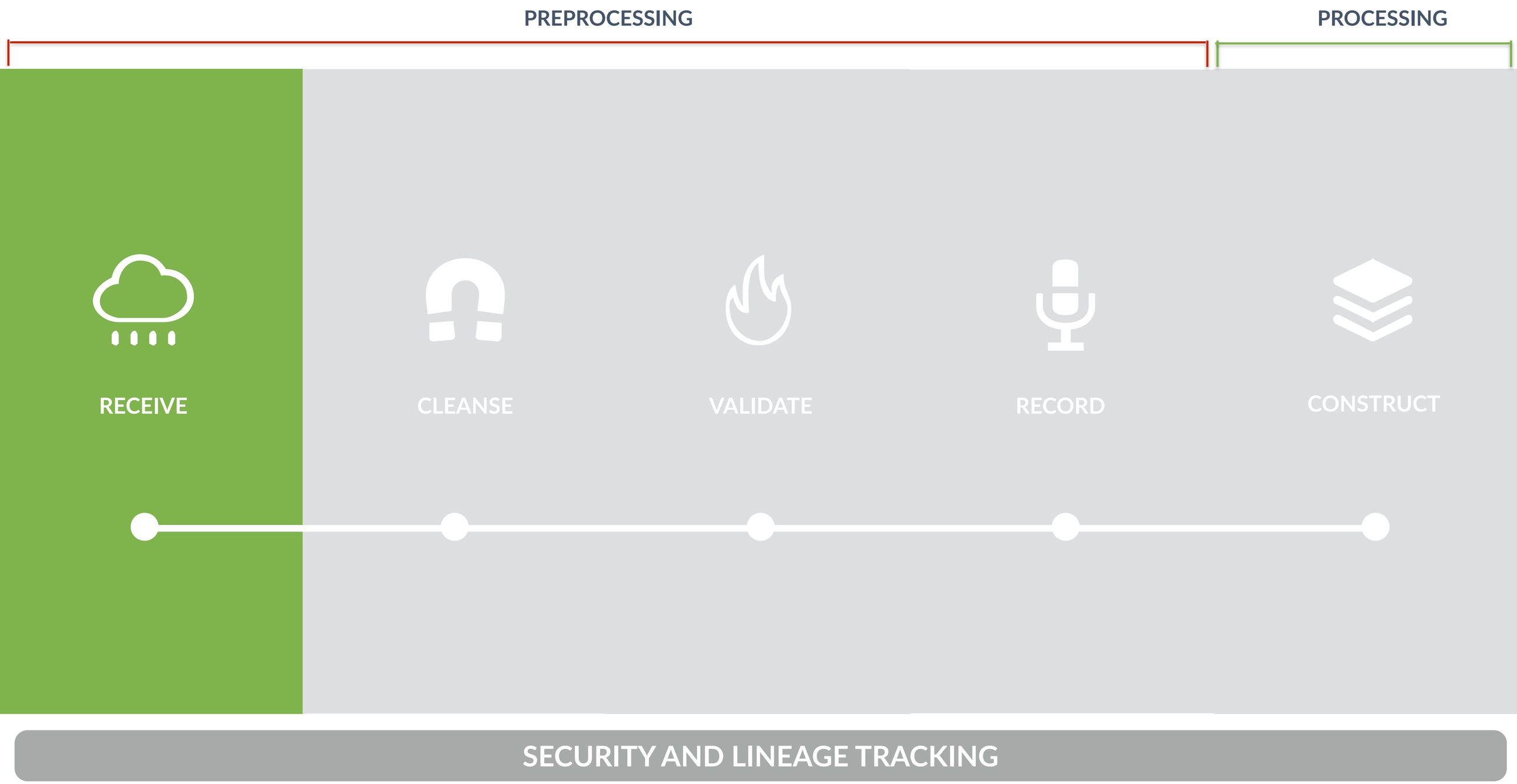
Until a year ago

Legacy Ingestion framework

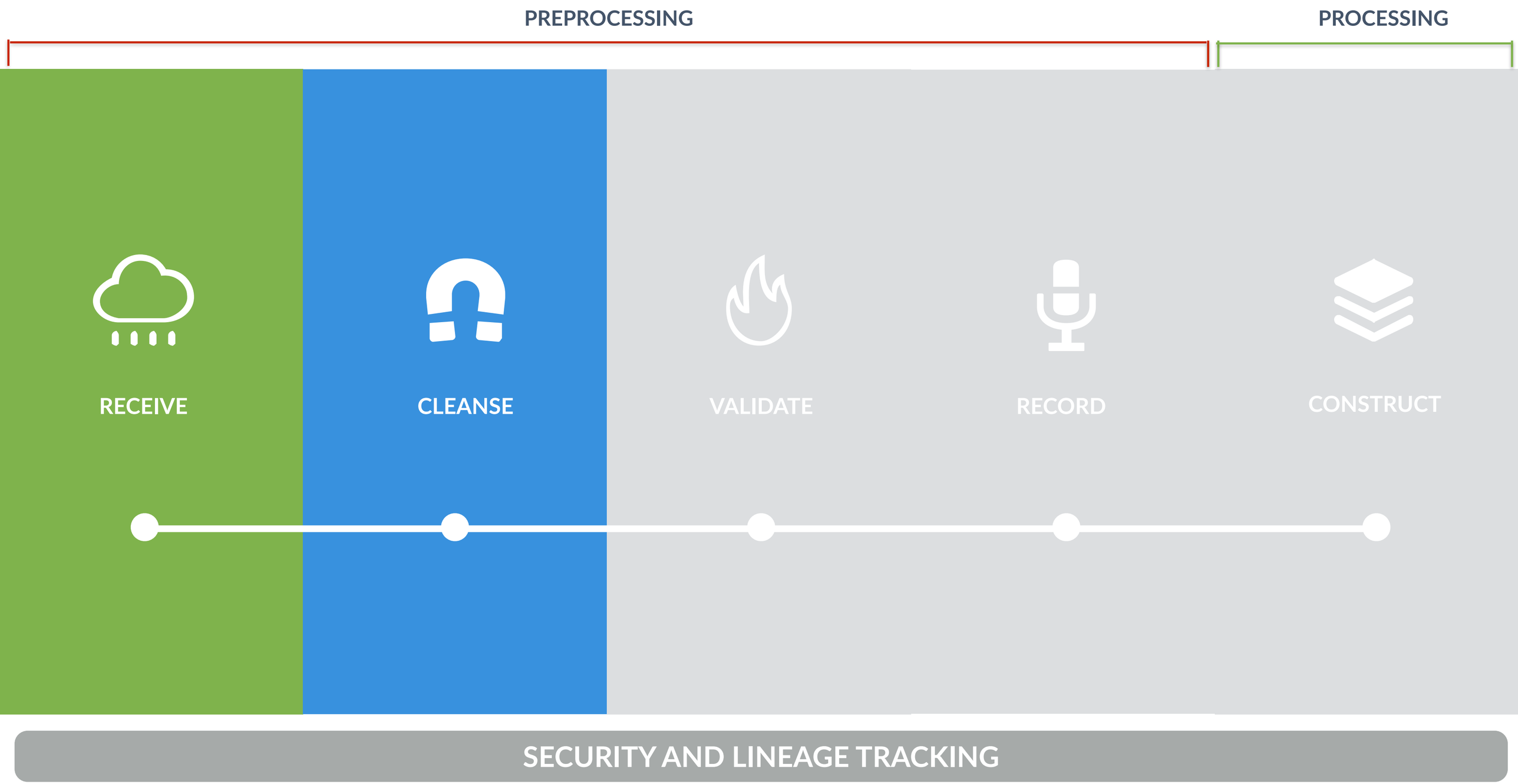
Now

Unified ingestion platform

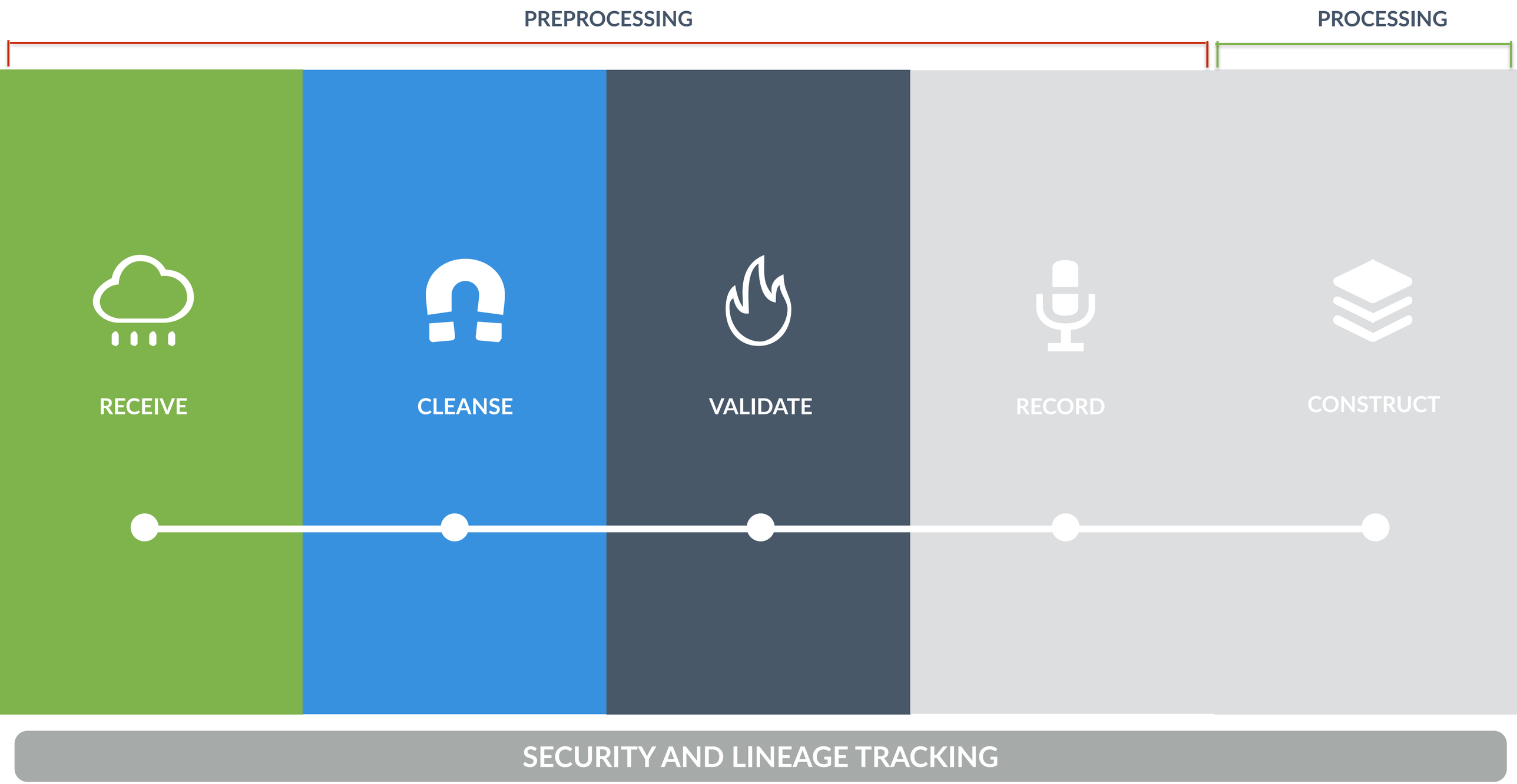
What is our view of Ingestion Framework?



Cleanse


PREPROCESSING
PROCESSING
RECEIVE
CLEANSE
VALIDATE
RECORD
CONSTRUCT
SECURITY AND LINEAGE TRACKING

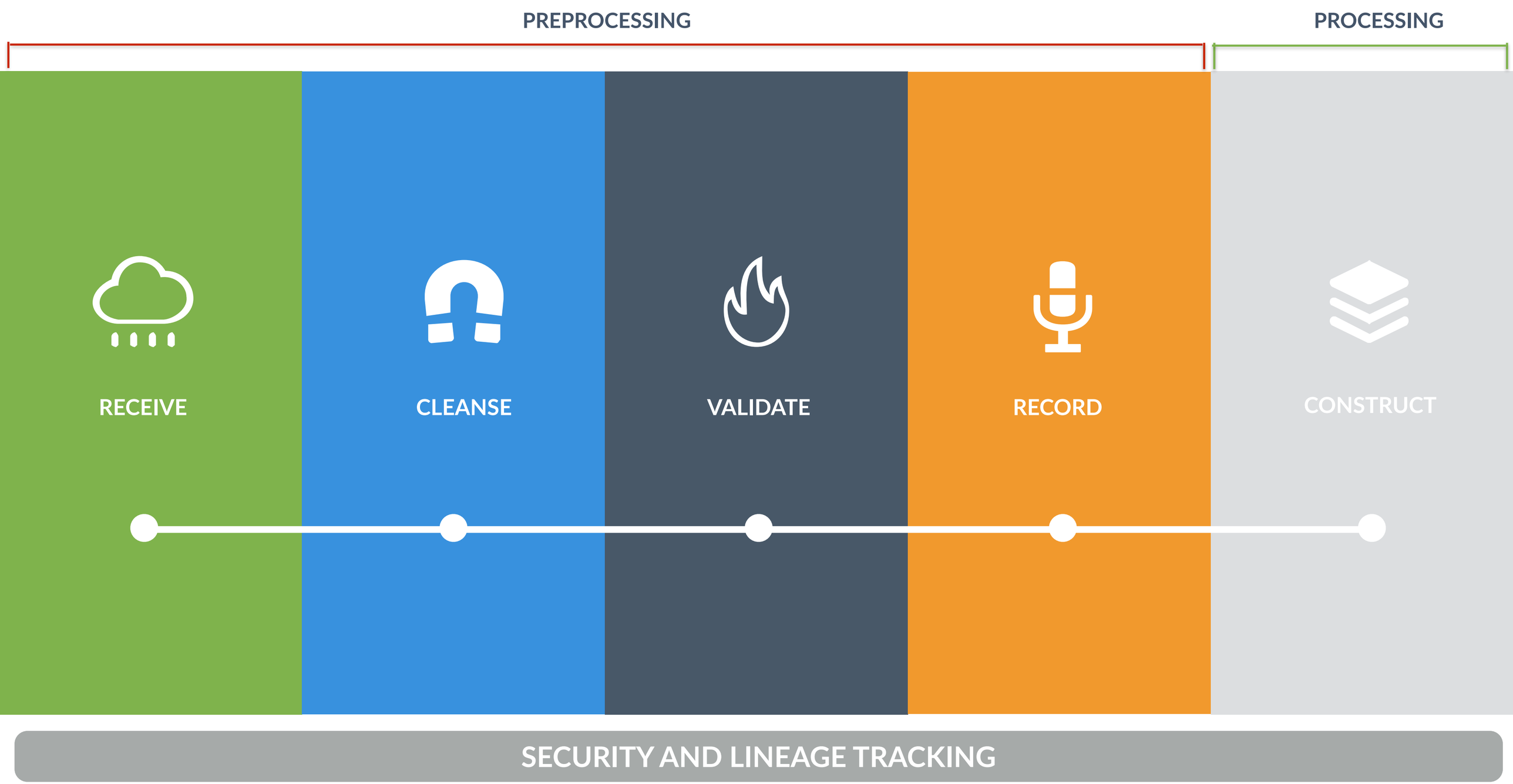
Validate


PREPROCESSING
PROCESSING
RECEIVE
CLEANSE
VALIDATE
RECORD
CONSTRUCT
SECURITY AND LINEAGE TRACKING

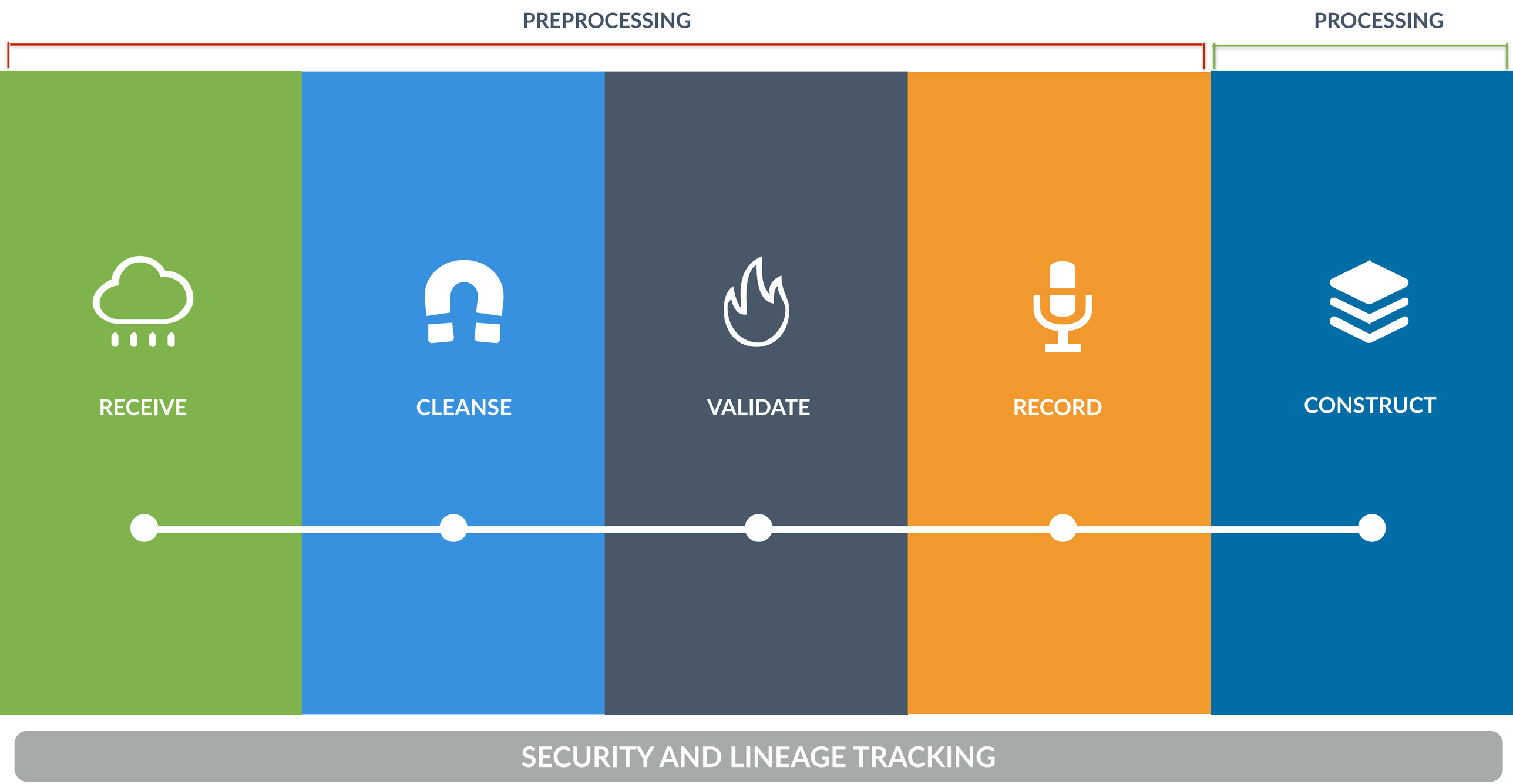
Essential pre-processing

1. Column and Row count validation
2. Embedded new line removal and special character replacement
3. Datatype validation
4. Data transformation
5. Value defaulting

Record


PREPROCESSING
PROCESSING
RECEIVE
CLEANSE
VALIDATE
RECORD
CONSTRUCT
SECURITY AND LINEAGE TRACKING

Construct


PREPROCESSING
PROCESSING
RECEIVE
CLEANSE
VALIDATE
RECORD
CONSTRUCT
SECURITY AND LINEAGE TRACKING

Backing technologies

Generation 2 Awesomeness

kafka

Change Data Capture

APACHE nifi

APACHE Spark

RECEIVE | CLEANSE | VALIDATE | RECORD | CONSTRUCT

1. Consistent tooling for preprocessing, ops data management, error reporting and archival
2. Security and managed concurrency
3. Faster development cycle for new applications. Easier to reason with the flow with NiFi's visual flow representation.
4. Significantly faster processing through Spark
5. ORC performs well for most of our consumption patterns - supporting both predicate and projection push down.

Generation 2 Awesomeness

kafka

Change Data Capture

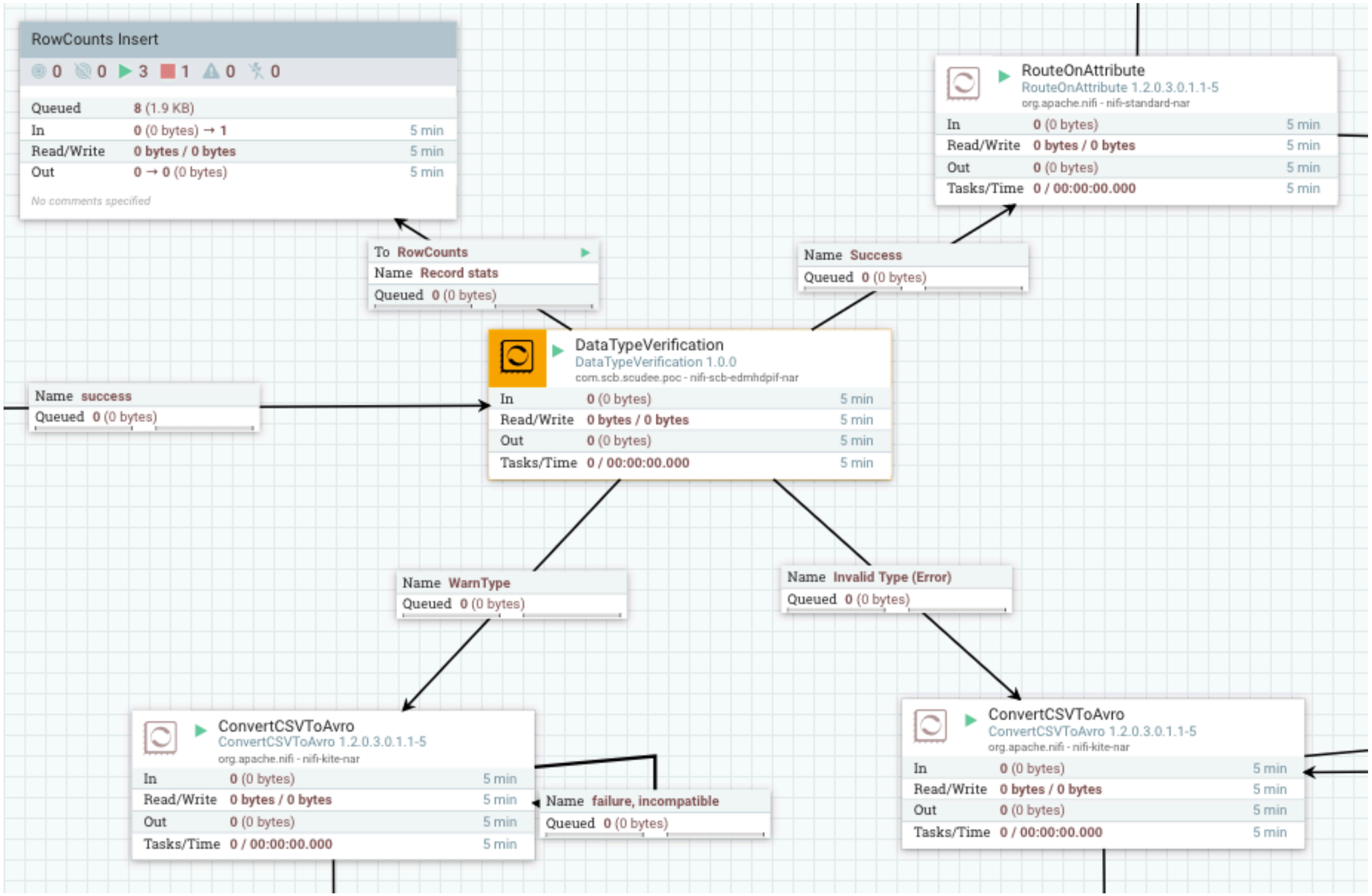
APACHE nifi

APACHE Spark

RECEIVE | CLEANSE | VALIDATE | RECORD | CONSTRUCT

1. Consistent tooling for preprocessing, ops data management, error reporting and archival
2. Security and managed concurrency
3. Faster development cycle for new applications. Easier to reason with the flow with NiFi's visual flow representation.
4. Significantly faster processing through Spark
5. ORC performs well for most of our consumption patterns - supporting both predicate and projection push down.

Extending NiFi via Custom Processors



Good Problems

Don’t bring me anything but trouble.
Good news weakens me.

- Charles Kettering

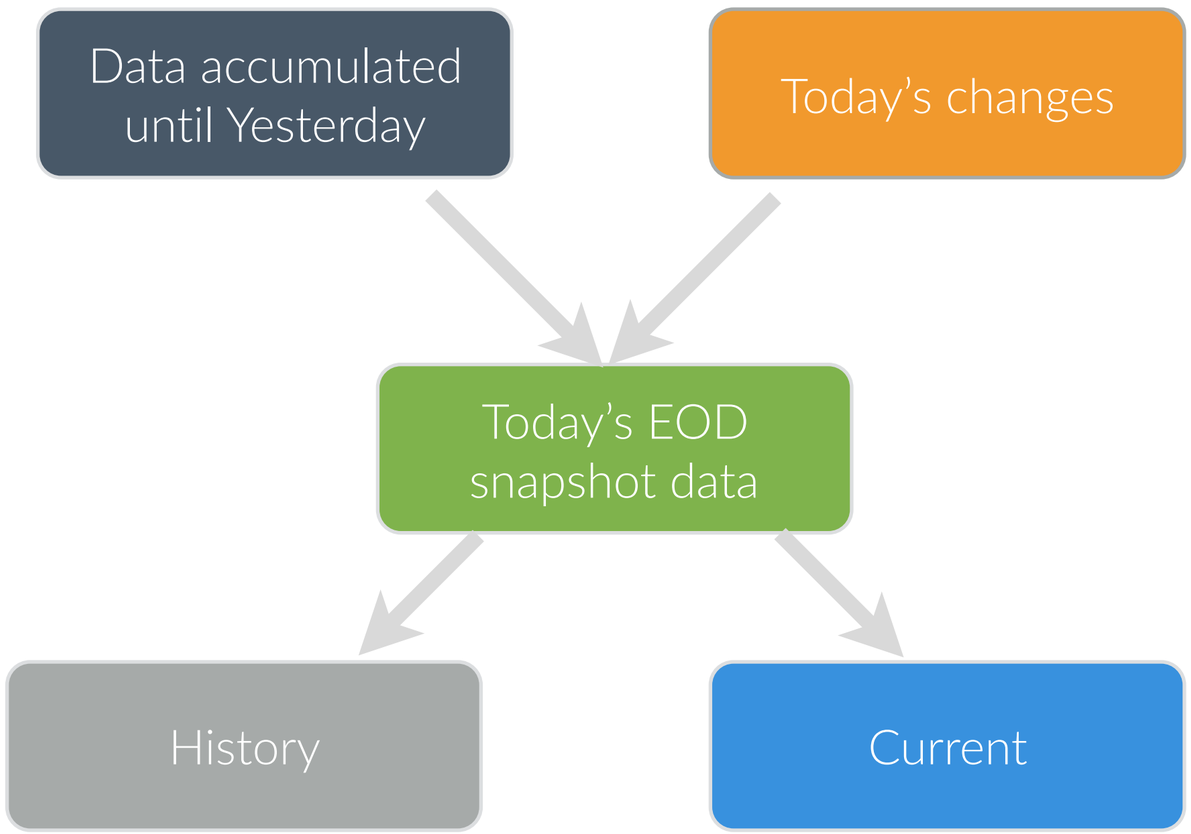
Types of Data

TYPES OF DATA

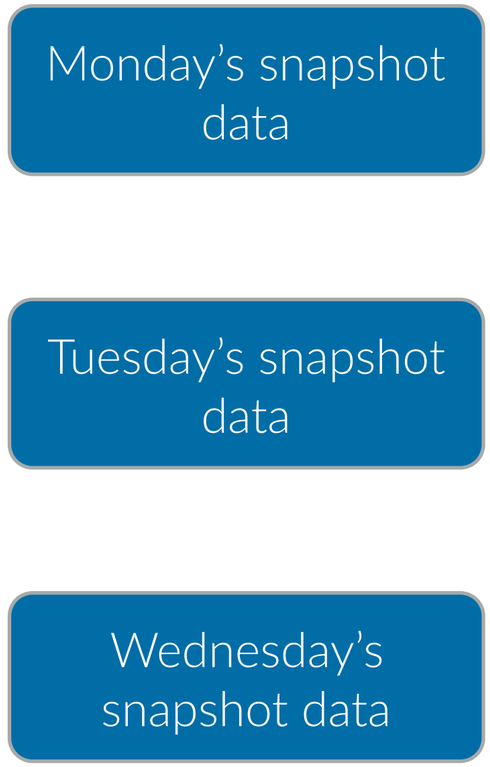
- » **Master** (eg. Customer data)
- » **Transaction** (eg. Banking transactions)
- » **Transaction data as Master** (eg. editable transactions)

Types of Data

Master



Transactional



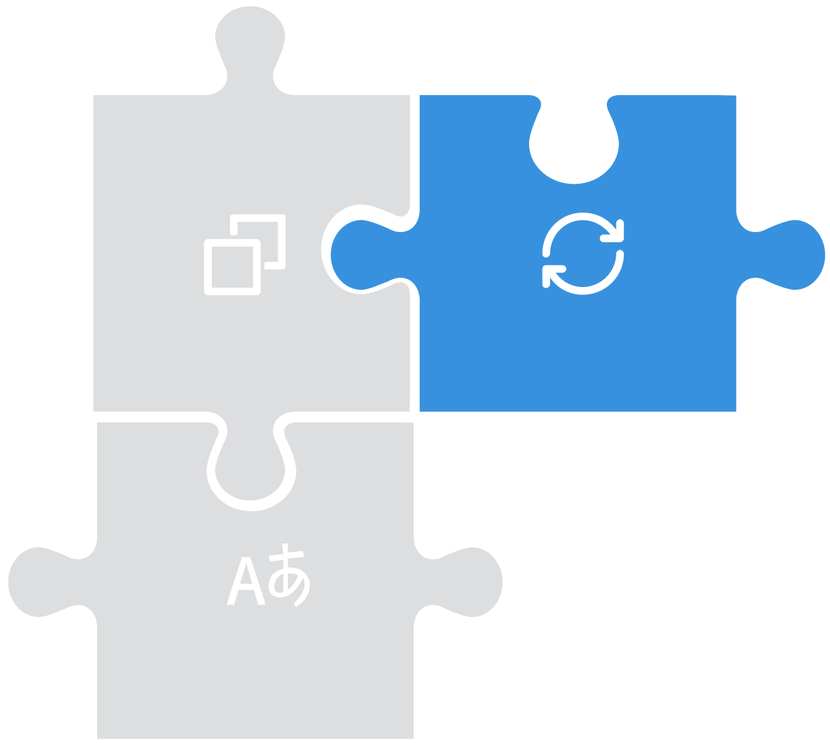
Frequency

TYPES OF DATA

- **Master** (eg. Customer data)
- **Transaction** (eg. Banking transactions)
- **Transaction data as Master** (eg. editable transactions)

DATA FORMATS

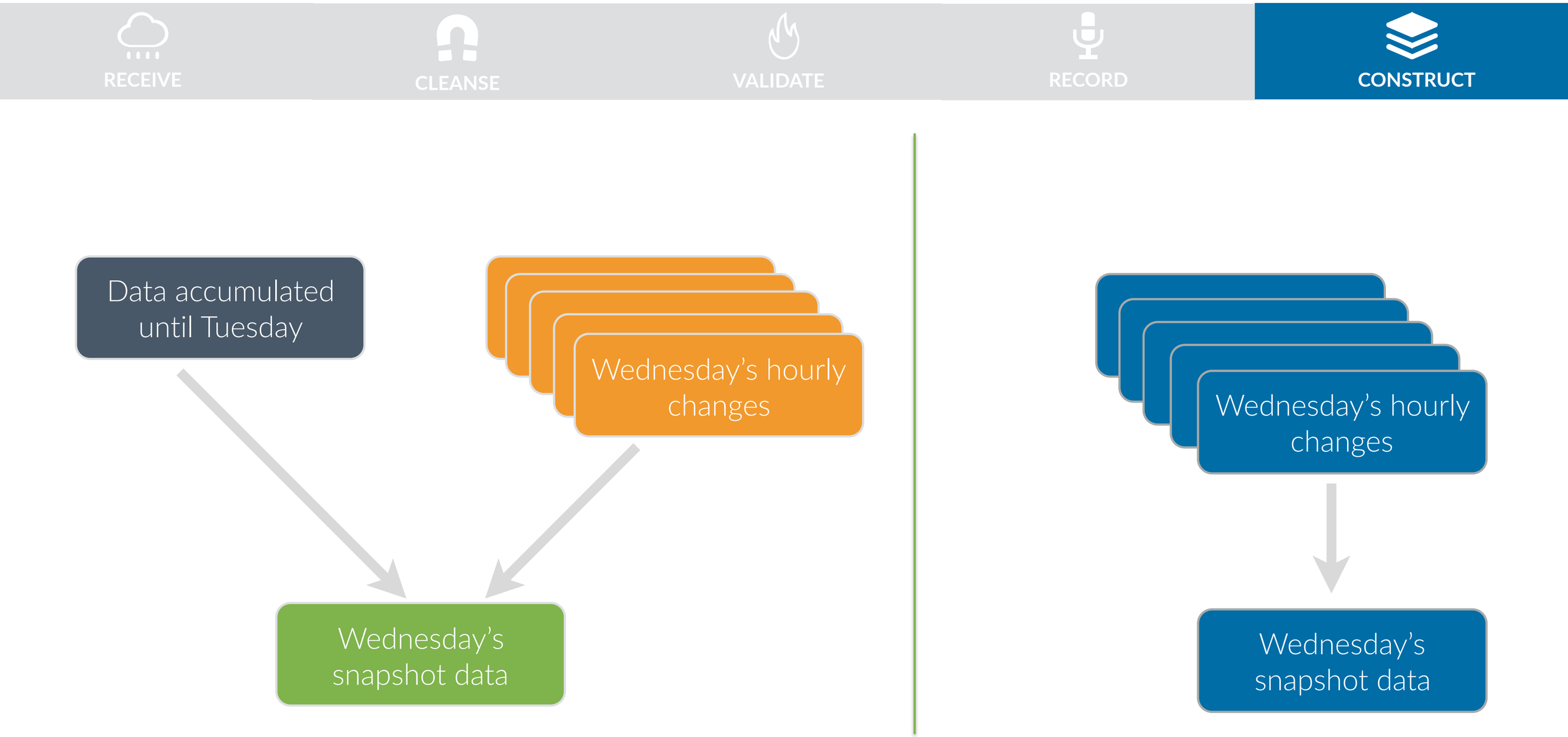
- **Output of Change Data Capture systems (Delimited)**
- **Plain delimited**
- **Fixed width**
- **Avro-JSON messages**
- **Spreadsheets**



FREQUENCY

- Streaming
- **Hourly incremental**
- Daily
- Weekly

Frequency - Hourly Incremental



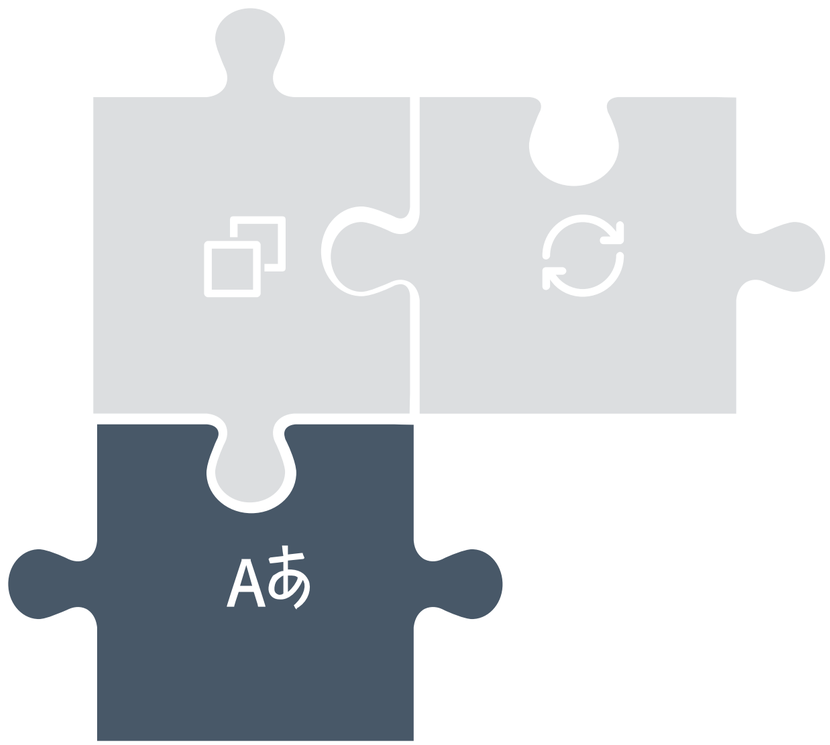
Data formats

TYPES OF DATA

- **Master** (eg. Customer data)
- **Transaction** (eg. Banking transactions)
- **Transaction data as Master** (eg. editable transactions)

DATA FORMATS

- **Output of Change Data Capture systems (Delimited)**
- **Plain delimited**
- Fixed width
- **Avro messages**
- XML
- Spreadsheets



FREQUENCY

- Streaming
- Hourly incremental
- Daily
- Weekly

Data formats - CDC Delimited

TIMESTAMP, TRANSACTION_ID, OPERATION_TYPE, USER_ID,<DATAFIELDVALUE1>,<DATAFIELDVALUE2>....

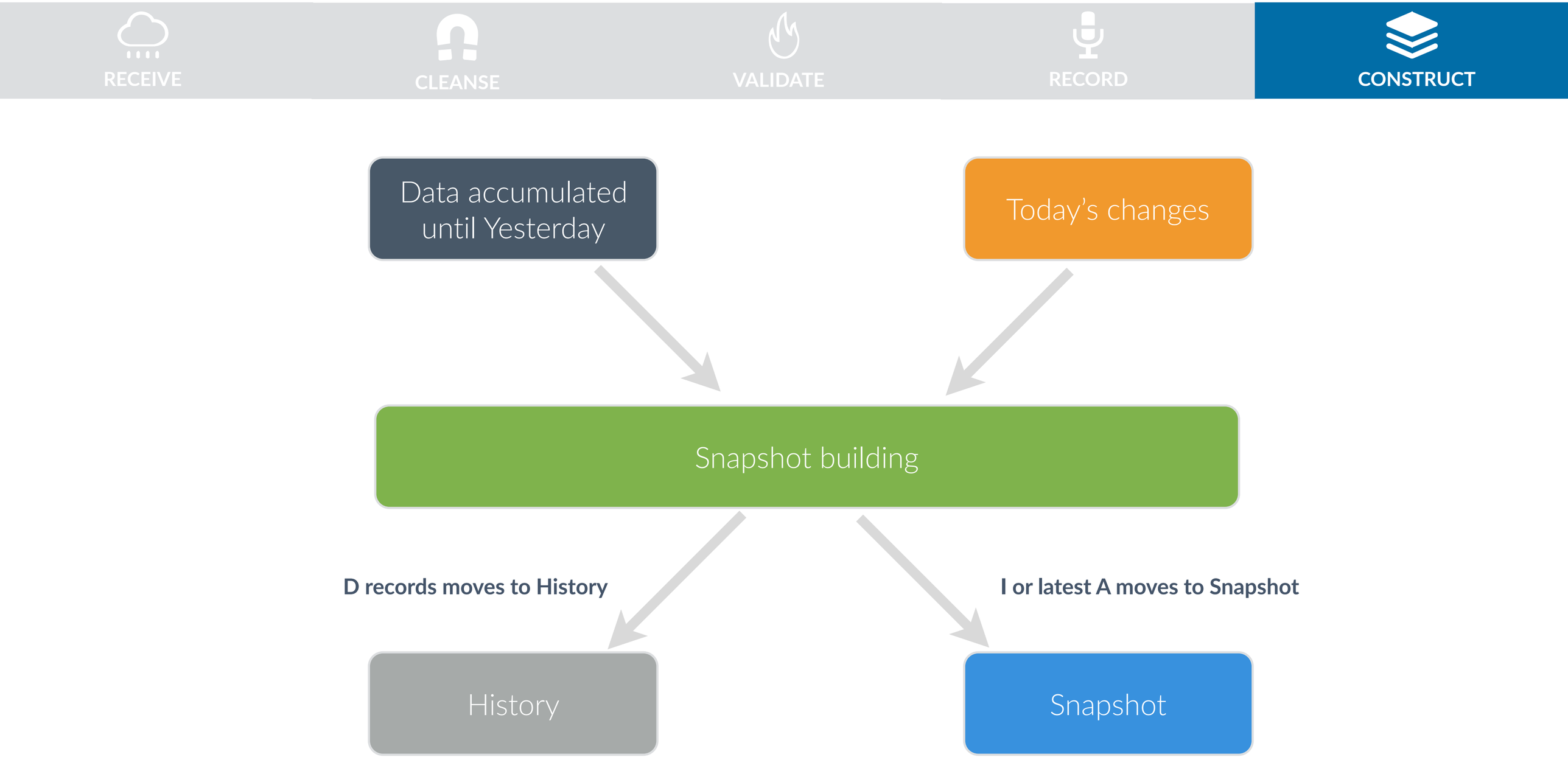
```
2017-07-20 22:38:04.00000|426664065479|B|DWUSER|OFFICE TELEPHONE NO|21234567890
```

CDC Columns

Application Columns

Could be one of I, B, A, D

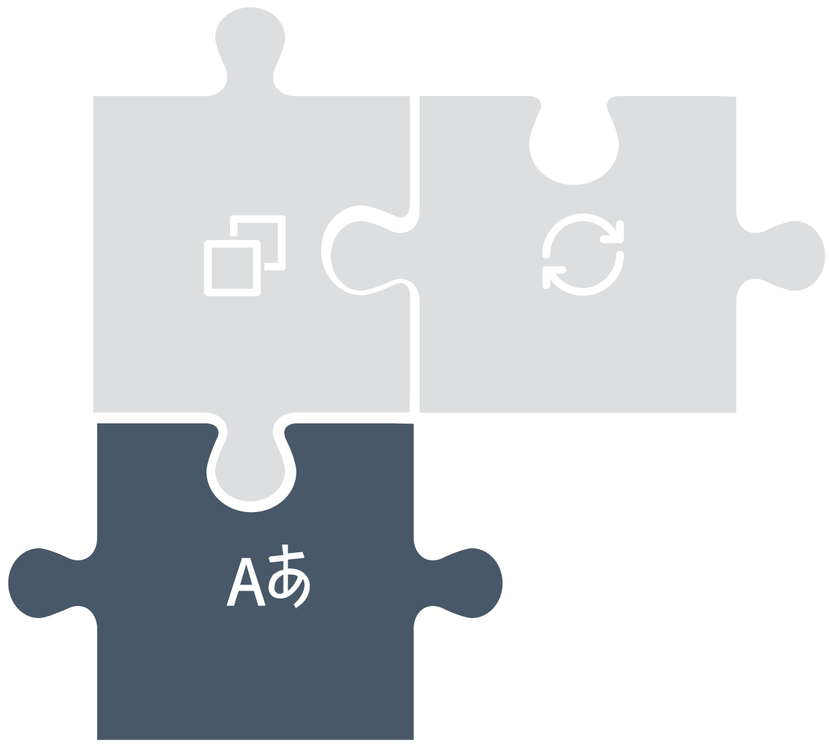
CDC Delimited - Snapshot building



Data formats and Frequency

TYPES OF DATA

- » **Master** (eg. Customer data)
- » **Transaction** (eg. Banking transactions)
- » **Transaction data as Master** (eg. editable transactions)



FREQUENCY

- » **Streaming**
- » **Hourly incremental**
- » **Daily**
- » **Weekly**

DATA FORMATS

- » Output of Change Data Capture systems (Delimited)
- » **Plain delimited**
- » Fixed width
- » Avro messages
- » XML
- » Spreadsheets

Data formats - Delimited with Header/Trailer

» Header and trailer has meta information about the data in the file

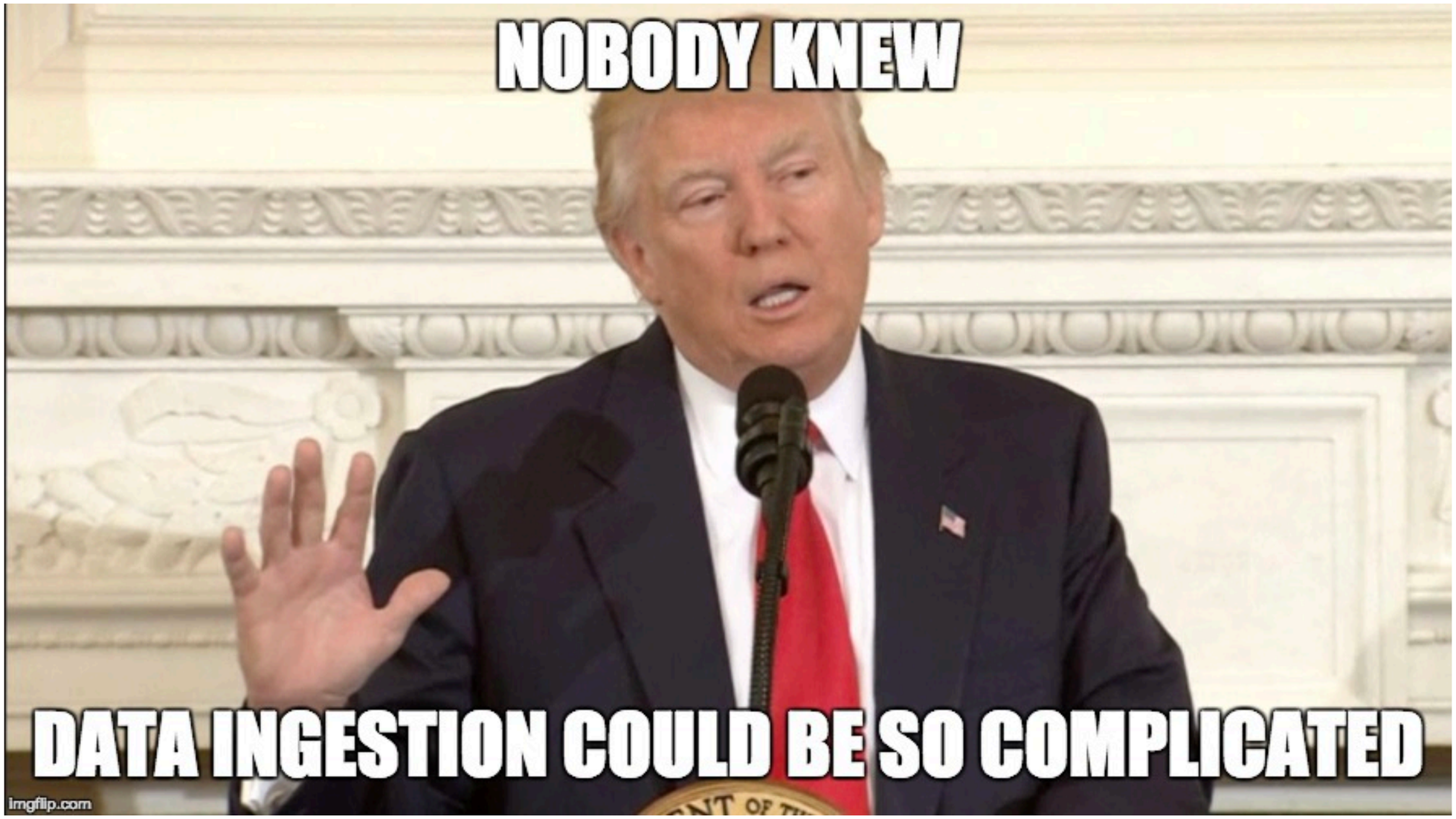
» Varieties of header and trailer formats

```
H3
TRANS_ID_1|USER_ID1|100.00
TRANS_ID_2|USER_ID2|3.00
TRANS_ID_3|USER_ID1|20.00
T123.00
```

Row count of the file

Sum or a rolling checksum of column(s) in that file

* Header and trailer information is used during reconciliation


NOBODY KNEW
DATA INGESTION COULD BE SO COMPLICATED
imgflip.com

The final piece to the puzzle

» TYPES OF DATA

» Master (eg. Customer data)
» Transaction (eg. Banking transactions)
» Transaction data as Master (eg. editable transactions)

DATA FORMATS

» Output of Change Data Capture systems (Delimited)
» Plain delimited
» Fixed width
» Avro messages
» XML
» Spreadsheets

FREQUENCY

» Streaming
» Hourly incremental
» Daily
» Weekly

PROCESSING

Among others already discussed,
» Schema evolution
» Cascading re-runs
» full-dump override

More awesomeness

1. Metadata management UI
2. Schema evolution & retrofitting historic data
3. Reruns, Cascading reruns, full-dump override
4. One-touch deployment

To summarise

1. The volume, the variety and the interesting combinations of data presented us some very interesting problems to solve.
2. With HDF and HDP, we were able to abstract away the cross cutting concerns such as security and concurrency.
3. Our code is (still) manageable and extensible
4. And… we intend to open source the framework for the general audience.

THANK YOU !

Questions?

@arunma