ETL vs ELT Cage Fight: Using RudderStack and ClickHouse to Build Real-Time Data Pipelines

Robert Hodges – Eric Dodds

Altinity

rudderstack

Let's make some introductions

Robert Hodges

Database geek with 30+ years on DBMS systems. Day job: CEO at Altinity

Eric Dodds

Head of Product Marketing at RudderStack, 10 years building data stacks

…And introduce our companies

Altinity

rudderstack

Altinity is the enterprise ClickHouse provider that lets you run anywhere with 100% open source analytic stacks

Real-time analytics in the cloud, on Kubernetes, and on-prem

RudderStack is the Warehouse Native CDP. Collection, unification and activation of customer data.

Real-time event streaming, ETL, rETL, transformations, ID res and more

Explainer: ETL vs. ELT

ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform).

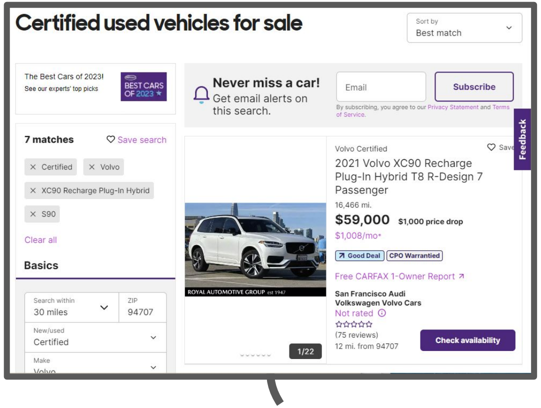
The main difference between ETL and ELT is the order in which the transformation stage is performed.

ETL is useful for structured data that requires transformation before reaching its destination (cleaning, enrichment, integration customizations, privacy). These can happen in batch or streaming formats.

ELT is useful when you want to retain an original copy of the data and are performing various kinds of modeling in the target system (most commonly a database). ELT is also useful when you are working with unstructured or semi-structured data, which can be transformed much more efficiently after being delivered.

The path from data to enlightment

eCommerce Website



Certified used vehicles for sale

Sort by
Best match

The Best Cars of 2023!
See our experts' top picks

BEST CARS OF 2023

Never miss a car!
Get email alerts on this search.

Email

Subscribe

By subscribing, you agree to our Privacy Statement and Terms of Service.

7 matches ♡ Save search

× Certified × Volvo
× XC90 Recharge Plug-In Hybrid
× S90

Clear all

Basics

Search within
30 miles

ZIP
94707

New/used
Certified

Make
Volvo

Feedback

Volvo Certified ♡ Save

2021 Volvo XC90 Recharge Plug-In Hybrid T8 R-Design 7 Passenger

16,466 mi.

$59,000 $1,000 price drop

$1,008/mo*

Good Deal CPO Warrantied

Free CARFAX 1-Owner Report

San Francisco Audi Volkswagen Volvo Cars

Not rated

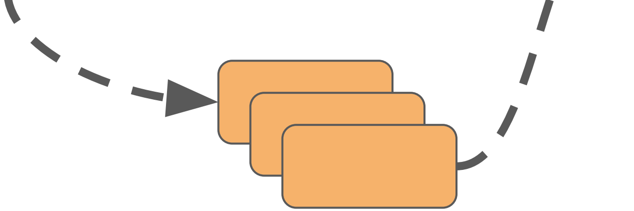
(75 reviews)

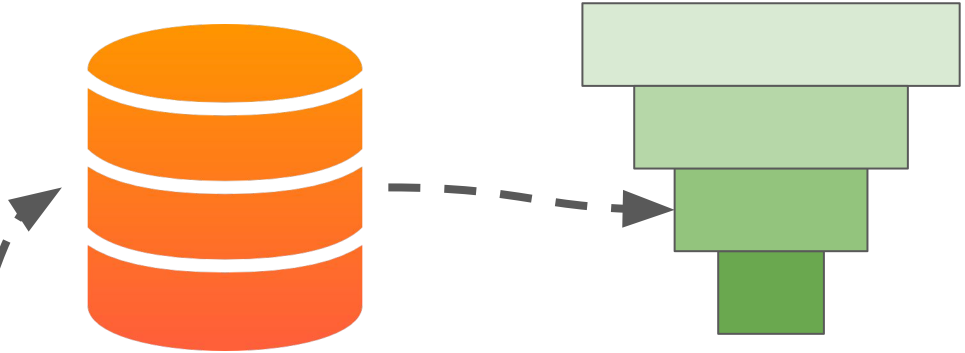
12 mi. from 94707

Check availability

ROYAL AUTOMOTIVE GROUP est 1947

1/22





Analytic Database

Funnel Analysis

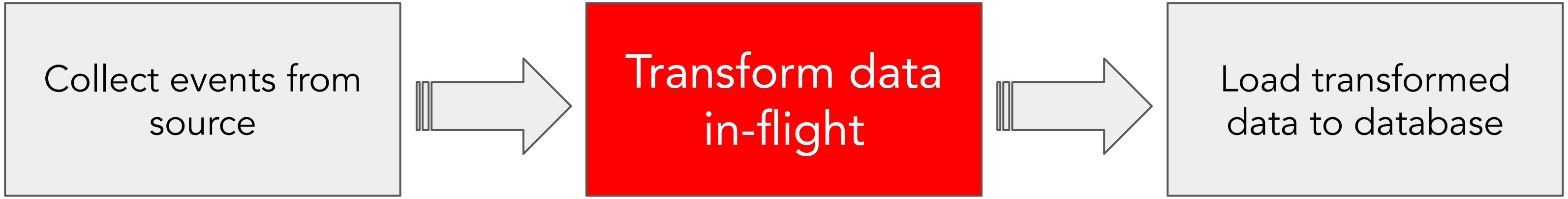
User Visit Events

We transform data in many ways along the way

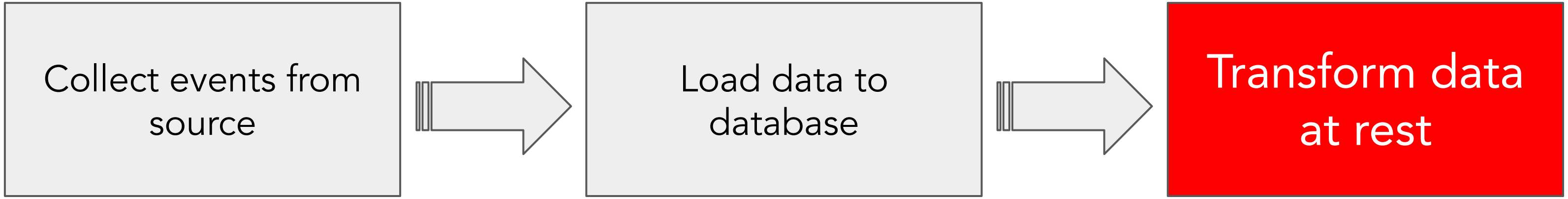
Name	Description	Example
Cleaning	Make data consistent for downstream	Normalizing addresses
Privacy	Remove/anonymize/encrypt sensitive data	Remove SSAN
Security	Allow or block specific data sources	Block invalid IPs
Enrichment	Add additional denormalized information	Add geolocation data
Customization	Specialized changes for applications	Change data to new format
Deduplication	Remove extra copies of data	Drop repeated visit events
Type mapping	Change data for performance/efficiency	Map Int64 to UInt8
Aggregation	Summarize data for quick insight	Website visitors per hour

There are two basic design choices for transformation

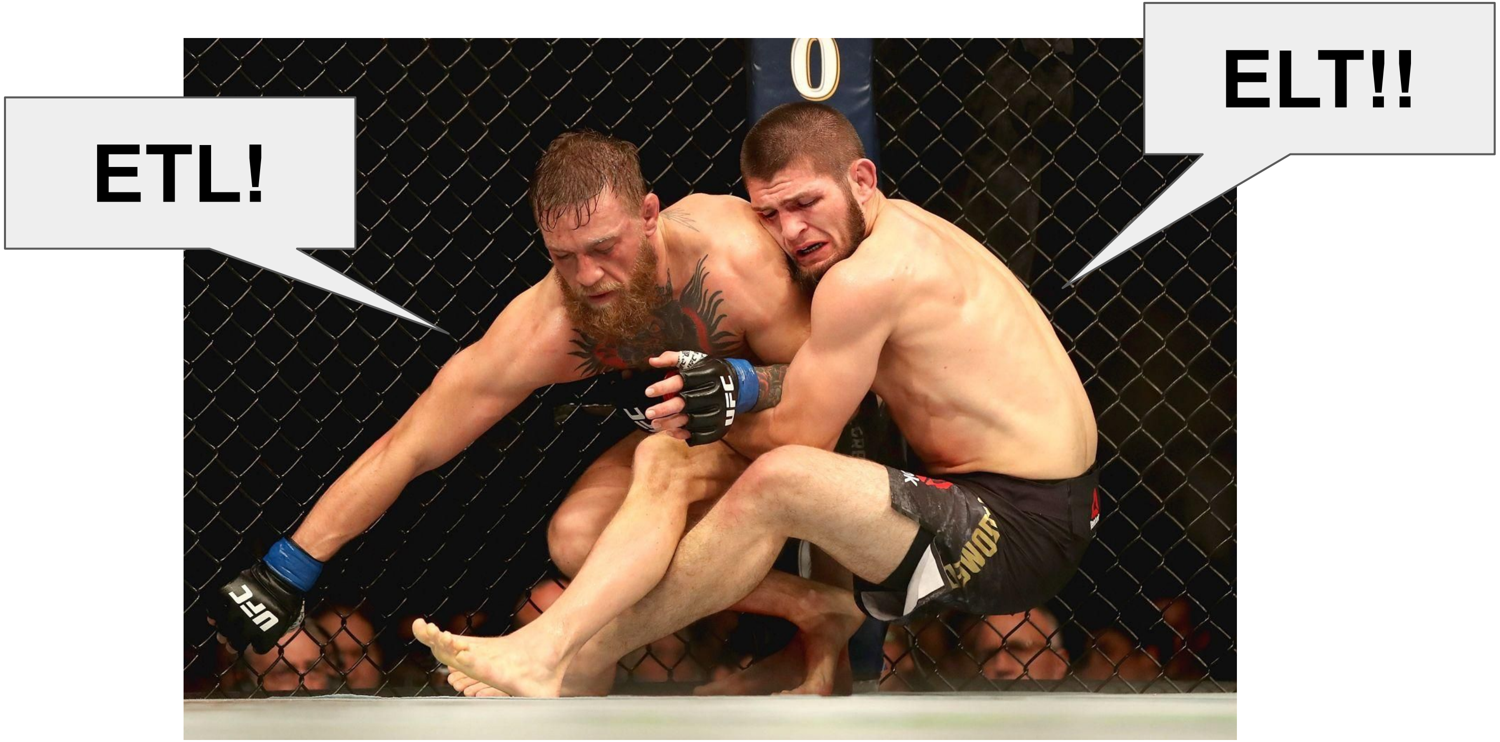
ETL == Extract, Transform, Load



ELT == Extract, Transform, Load



Do we need to fight over the winner?



It does not have to be this way

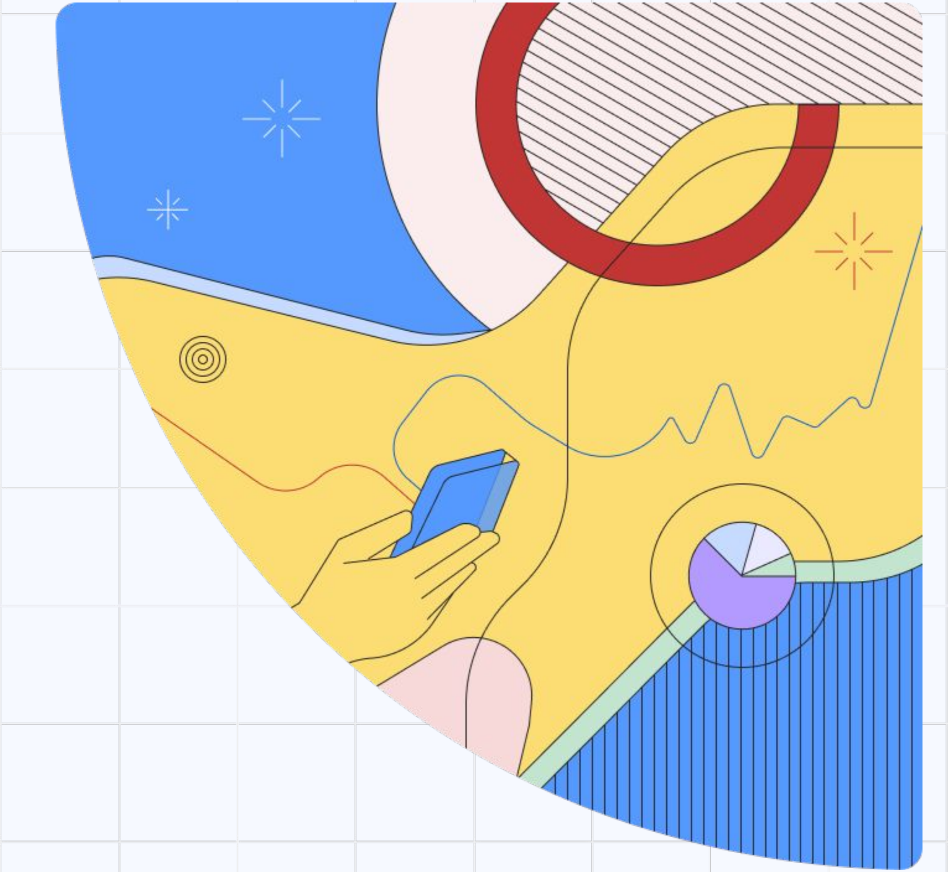
Introduction to RudderStack

RudderStack delivers trustworthy, real-time data to the tools and teams that need it

rudderstack

rudderstack

Make Data Valuable



Altinity

rudderstack

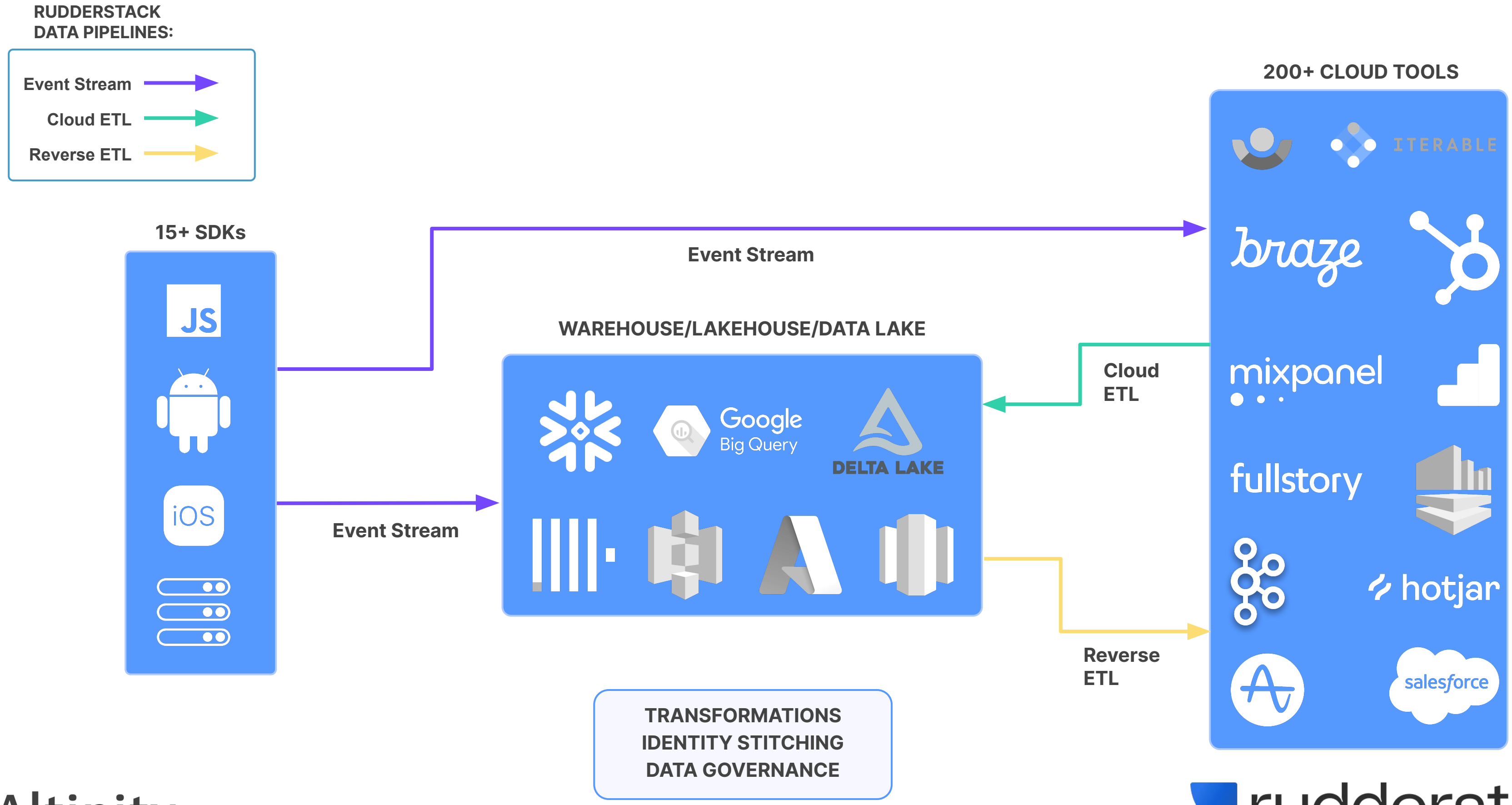
About RudderStack

RudderStack delivers trustworthy, real-time data to the tools and teams that need it.

We provide data pipelines and features that let you:

- Send first-party data across your stack in real-time
- Transform that data in-flight, before reaching your tools
- Activate enriched data back across your tools and teams

RudderStack Architecture Diagram



RUDDERSTACK
DATA PIPELINES:

Event Stream

Cloud ETL

Reverse ETL

15+ SDKs

Event Stream

WAREHOUSE/LAKEHOUSE/DATA LAKE

Event Stream

Cloud ETL

Reverse ETL

200+ CLOUD TOOLS

TRANSFORMATIONS
IDENTITY STITCHING
DATA GOVERNANCE

Altinity

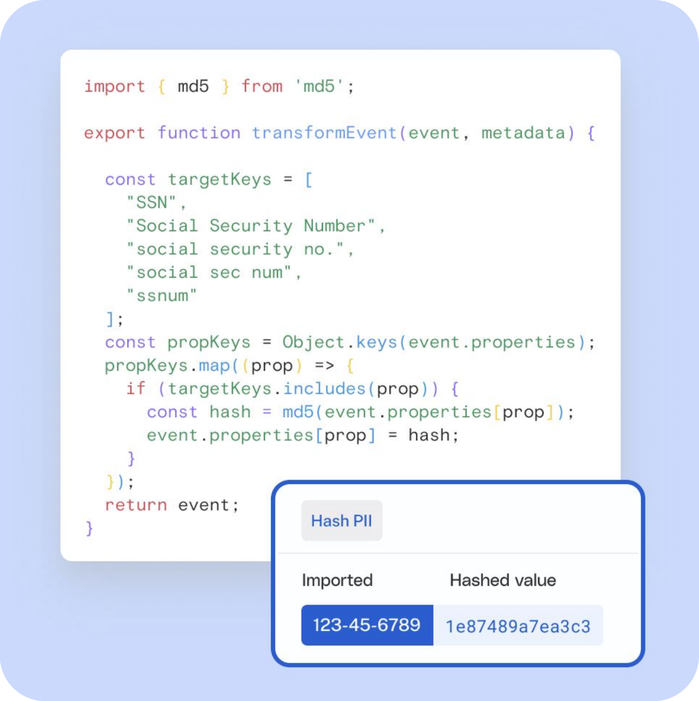
rudderstack

What is RudderStack Transformations?

Transformations lets users customize event data in real-time using JavaScript or Python.

With Transformations, users have the **control** and **flexibility** to:

- Ship data projects faster
- Secure and build data trust
- Quickly adapt to change



```
import { md5 } from 'md5';

export function transformEvent(event, metadata) {

  const targetKeys = [
    "SSN",
    "Social Security Number",
    "social security no.",
    "social sec num",
    "ssnum"
  ];
  const propKeys = Object.keys(event.properties);
  propKeys.map((prop) => {
    if (targetKeys.includes(prop)) {
      const hash = md5(event.properties[prop]);
      event.properties[prop] = hash;
    }
  });
  return event;
}
```

Hash PII

Imported	Hashed value
123-45-6789	1e87489a7ea3c3

RudderStack Transformations Use Cases

RudderStack Transformations allows users to manipulate event data in real-time with custom Javascript or Python code to quickly execute use cases for:

Data Processing & Enrichment

Enrich payloads with user, geo, AI data and more

Flatten schemas to fit downstream tools

Modify events in real-time before they reach your server

Data Security & Governance

Hash/Mask/Replace PII and sensitive data

Block or allow specific events from reaching specific tools

Encrypt or decrypt PII, including those stored in cookies

Custom Integrations & More

Rename event properties to any naming convention

Create custom sources and integrations

Dynamically send events to different paths via webhook

Introduction to ClickHouse

ClickHouse is a real-time analytic database

Understands SQL

Runs on bare metal to cloud

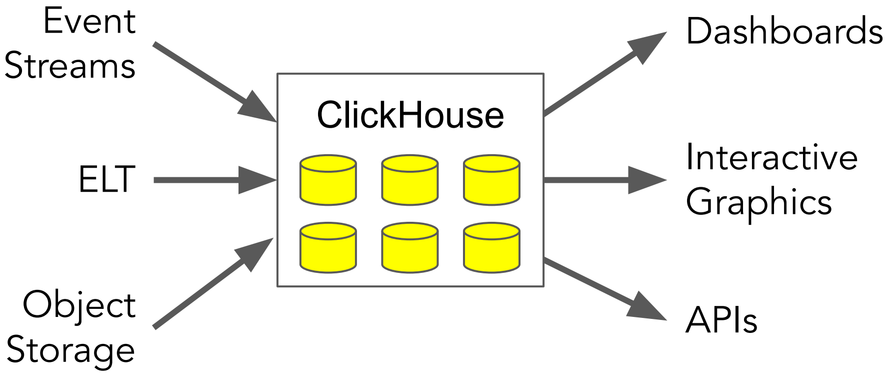
Shared nothing architecture

Stores data in columns

Parallel and vectorized execution

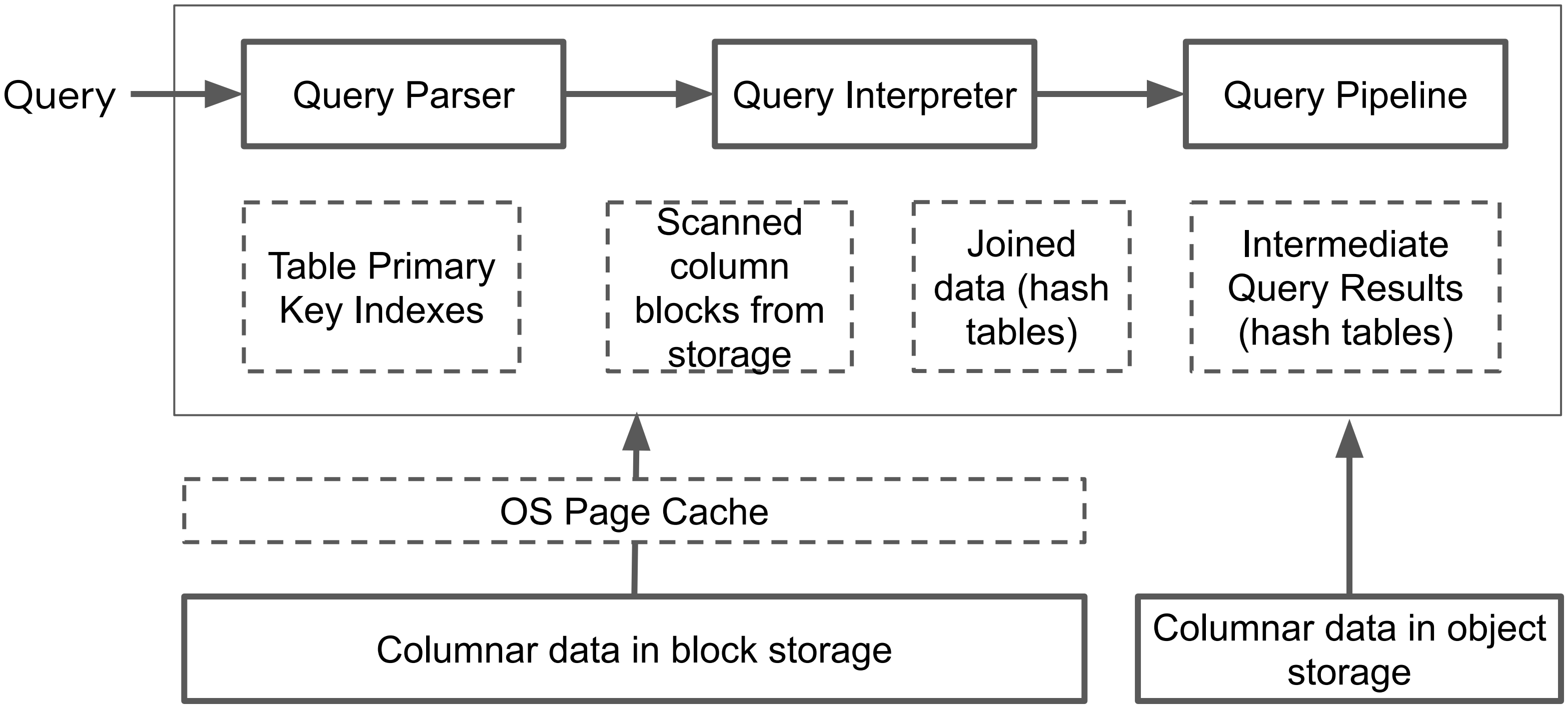
Scales to many petabytes

Is Open source (Apache 2.0)



It's the core engine for low-latency analytics

ClickHouse Server Architecture


Query
Query Parser
Query Interpreter
Query Pipeline
Table Primary Key Indexes
Scanned column blocks from storage
Joined data (hash tables)
Intermediate Query Results (hash tables)
OS Page Cache
Columnar data in block storage
Columnar data in object storage

Altinity

rudderstack

Why is ClickHouse so fast?

Codecs

Data Types

Sharding

Read Replicas

Data Partitioning

Compression

Skip Indexes

Projections

Tiered Storage

Distributed Query

In-RAM dictionaries

Primary key index

Seeing is believing

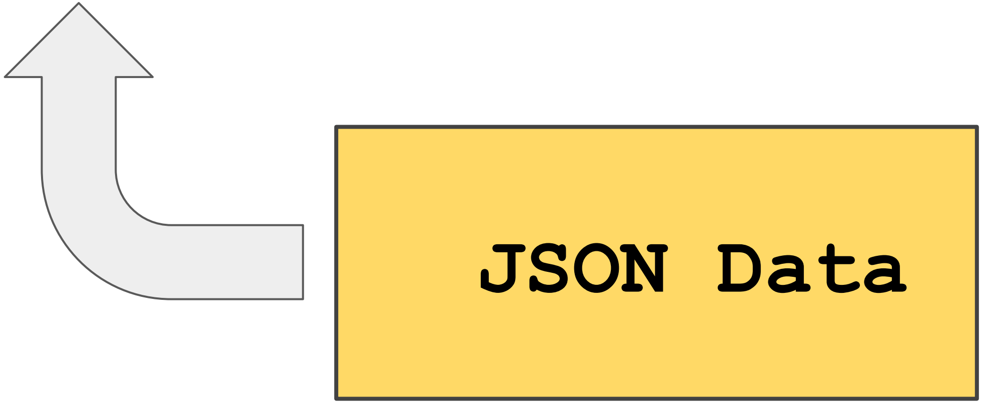
Demo Time!

Sensor Input Data

```
{
  "sensor_id": "0",
  "sensor_type": "1",
  "time": "2019-01-01 00:00:00",
  "msg_type": "reading",
  "temperature": "46.31",
  "message": "",
  "device_type": "0",
}
```

Simplest way to load readings

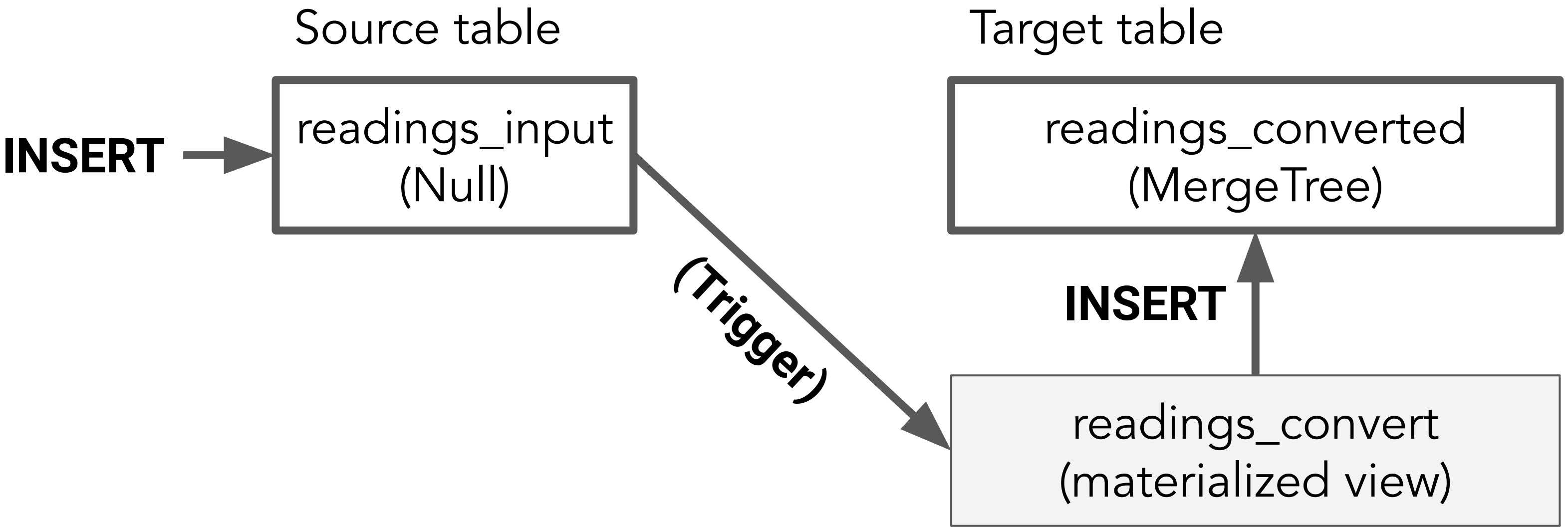
```
INSERT INTO readings(sensor_id,
sensor_type, time, msg_type,
temperature, message)
Format JSONEachRow
```



Pipe input data to clickhouse-client

JSON Data

Materialized views can transform input



Source table definition

```
CREATE TABLE readings_input (
   `event` String
)
ENGINE = Null
```

Target table definition

```
CREATE TABLE readings_converted (
   `sensor_id` Int32 CODEC(DoubleDelta, LZ4),
   `sensor_type` UInt8,
   `time` DateTime CODEC(DoubleDelta, LZ4),
   `date` Date ALIAS toDate(time),
   . . .
   `event` String
) ENGINE = MergeTree
PARTITION BY toYYYYMM(time)
ORDER BY (msg_type, sensor_id, time)
```

Materialized view to convert input to correct datatypes

```
CREATE MATERIALIZED VIEW readings_convert
TO readings_converted
AS
SELECT
  toInt32(JSON_VALUE(event, '$.sensor_id')) AS `sensor_id`,
  toInt8(JSON_VALUE(event, '$.sensor_type')) AS
`sensor_type`,
  toDateTimeOrNull(JSON_VALUE(event, '$.time')) AS `time`,
  . . .
  `event`
FROM readings_input
```

We can do any transformation that SQL can!

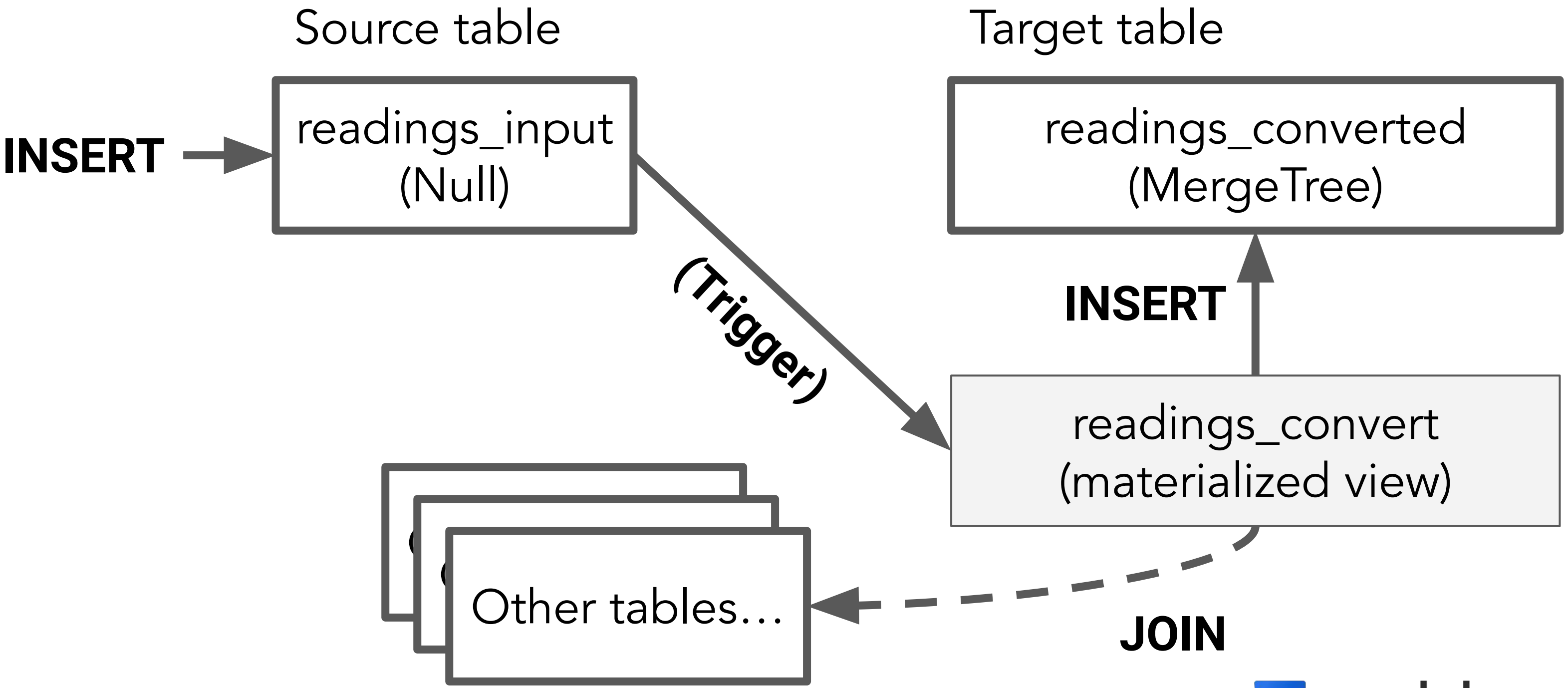
```
CREATE MATERIALIZED VIEW pii_data
TO safe_data
AS
SELECT
  '000-00-00000' as ssan,
  toString(cityHash64(email)) as hashed_email,
  encrypt('aes-256-ofb', name, key) AS encrypted_name,
  . . .
FROM readings_input
```

Zero out SSAN

Hash email

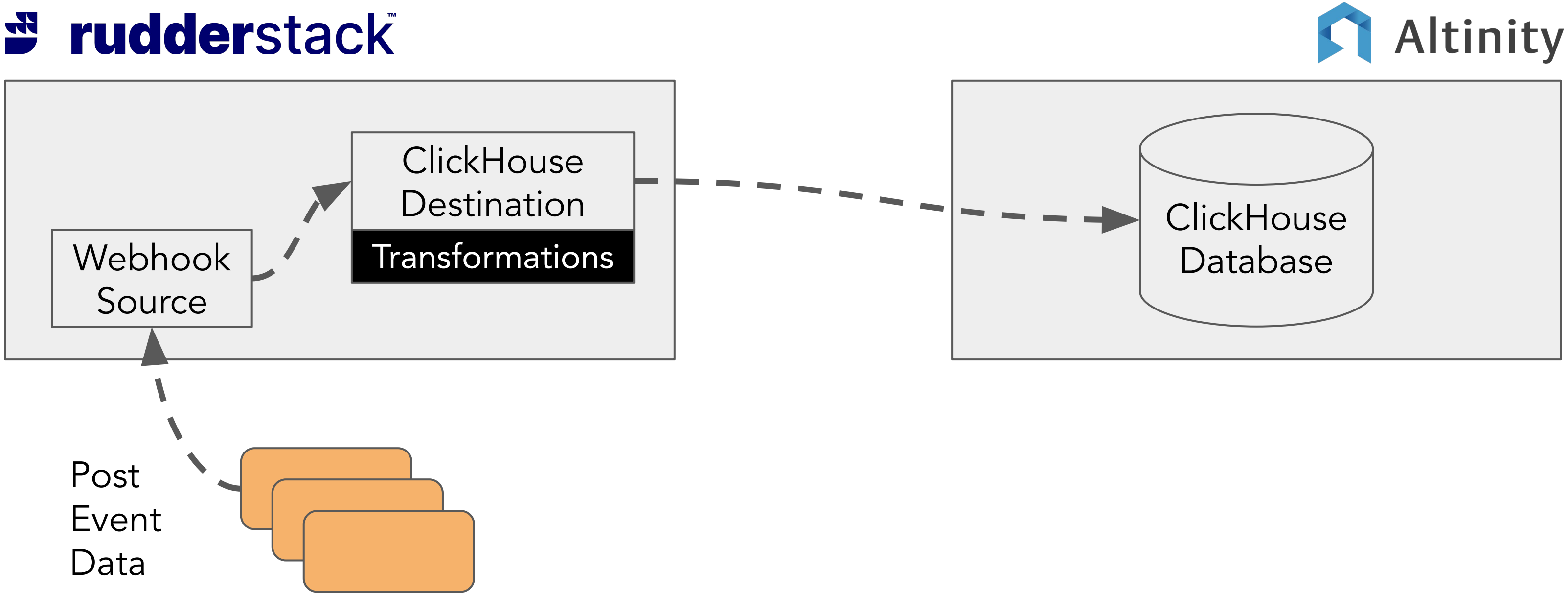
AES-encrypt name

ClickHouse can even join on other ables to add data



RudderStack and ClickHouse Together

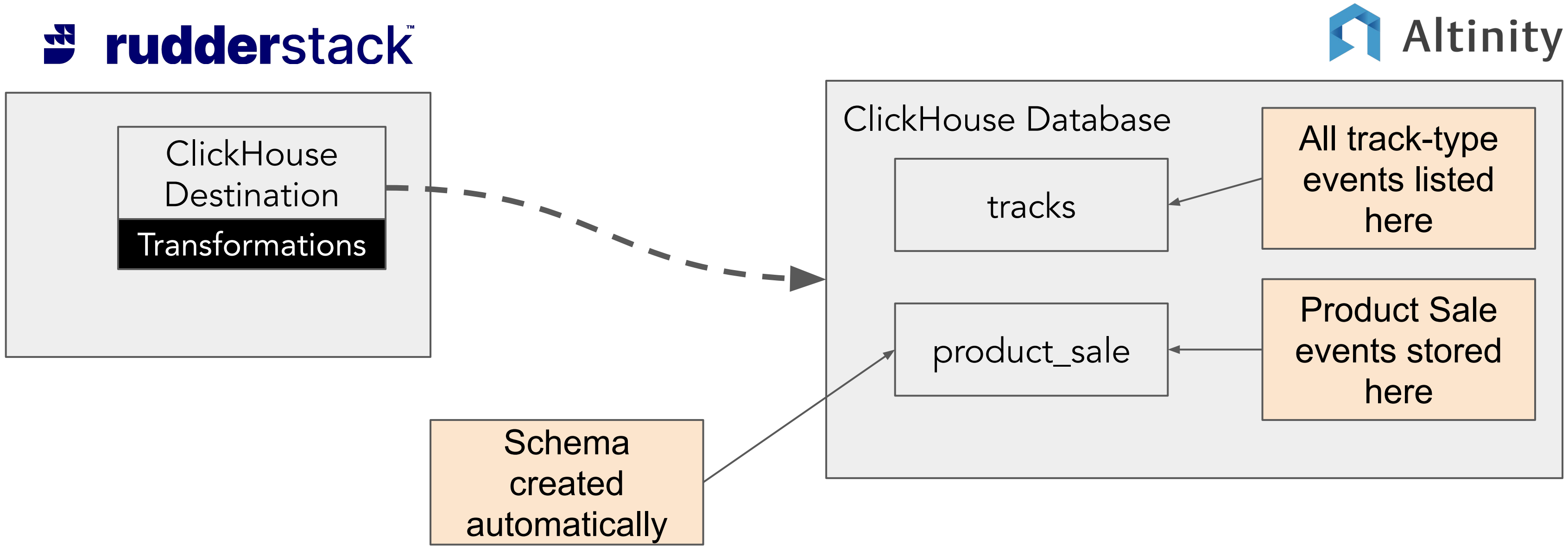
Integrating RudderStack and ClickHouse



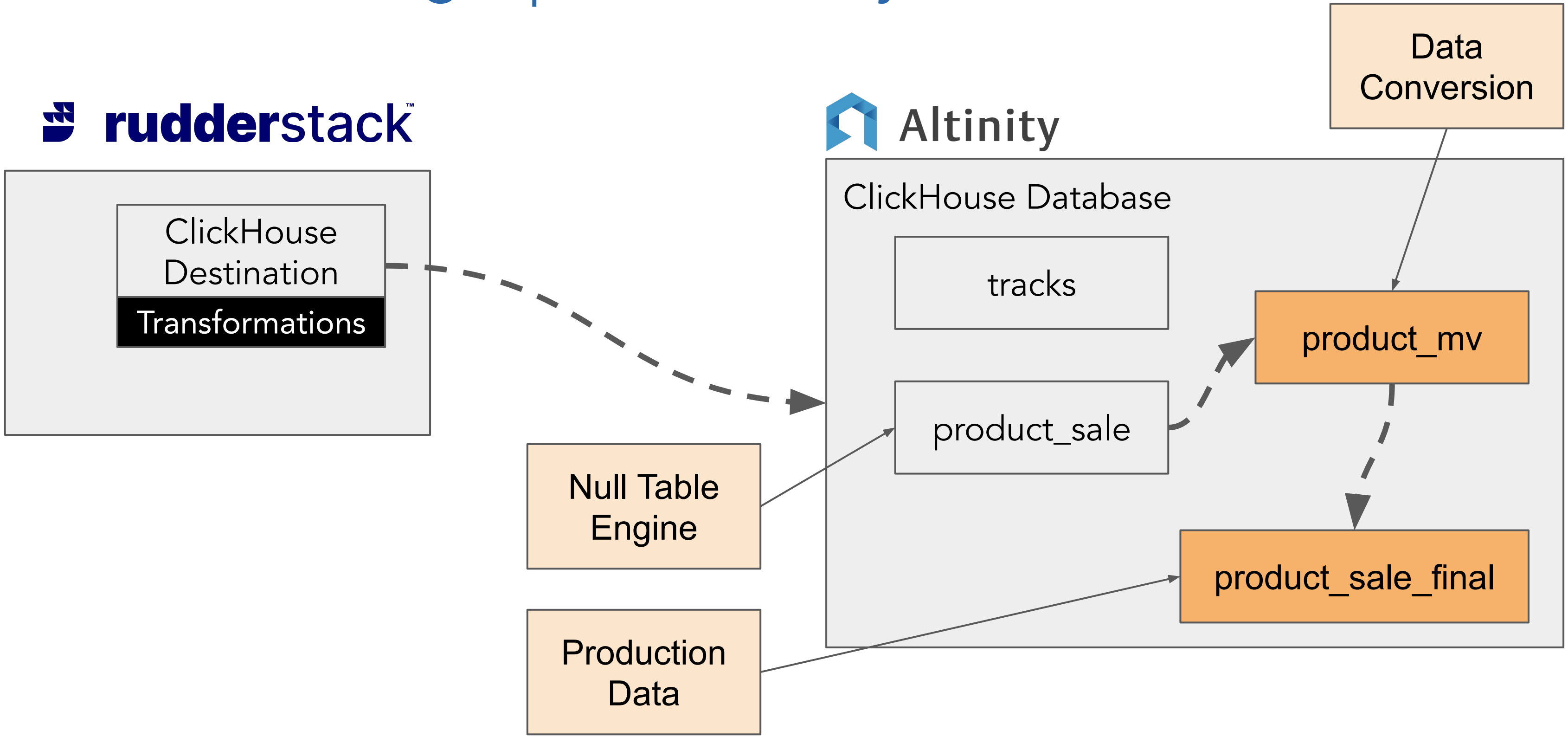
Seeing is believing

Demo Time!

How does schema management work?



Hints for building a production system



Use Rudderstack and ClickHouse for transformations

Name	Description	RudderStack	ClickHouse
Cleaning	Make data consistent for downstream	✅✅	
Privacy	Remove/anonymize/encrypt sensitive data	✅✅	✅
Security	Allow or block specific data sources	✅✅	✅
Enrichment	Add additional denormalized data	✅✅	✅✅
Customization	Specialized changes for applications	✅✅	✅
Deduplication	Remove extra copies of data	✅	✅
Type mapping	Change data for performance/efficiency	✅	✅✅
Aggregation	Summarize data for quick insight		✅✅

A few words about Reverse ETL

Reverse ETL: Send enriched data and audiences from your warehouse to your entire customer data stack

Configure data mapping using a JSON editor: Customize warehouse table sync settings by configuring JSON. Modify keys and add constants to customize payloads for every destination.

Create pipelines by writing SQL: Use our Reverse ETL Models feature to write SQL queries and turn the resulting table into a Reverse ETL job.

- Push warehouse data to all of your business tools
- Support for all major cloud warehouses
- 150+ cloud destinations
- Enable advanced analytics-based use cases like personalization, recommendations, lead scoring and more

Wrap-up

Summary points

- There’s no conflict between ETL and ELT – Use them both together
- RudderStack offers a rich set of tools to move and convert data in-flight
- ClickHouse offers a rich set of tools to convert data at rest
- Get off the ground quickly with RudderStack Cloud and Altinity.Cloud

Thank you!
Questions?

https://altinity.com
Altinity.Cloud
Contact Altinity

https://rudderstack.com
RudderStack Cloud
Contact RudderStack

Altinity

rudderstack