# Modern Data Stack (MDS)

Dr. Zhang Zhenjie
Neuron Mobility Pte. Ltd.

# Overview of the talk

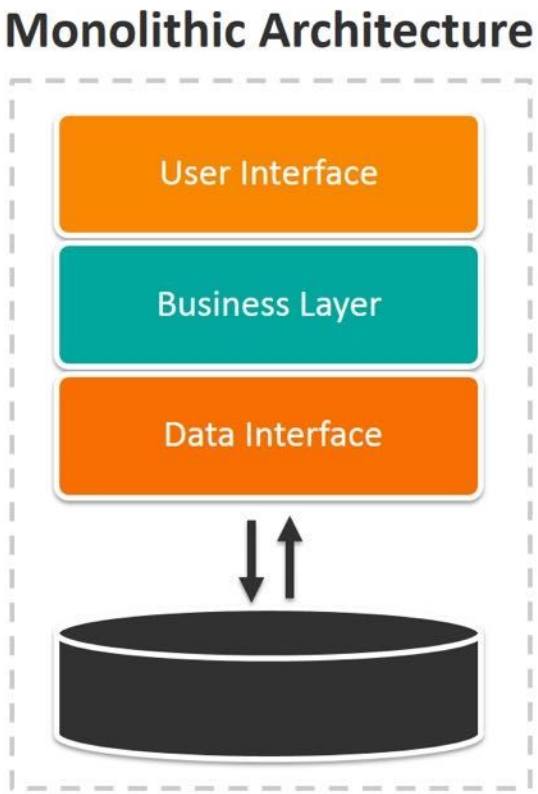
- The history of data stack
- Trends in modern data stacks
- Open source and ecosystem
- Practice at Neuron Mobility

# Ancient data stack

- 20 years ago, banks are the only ***big*** data consumer
  - High cost to own
  - Hard to scale up
  - Poor extensibility

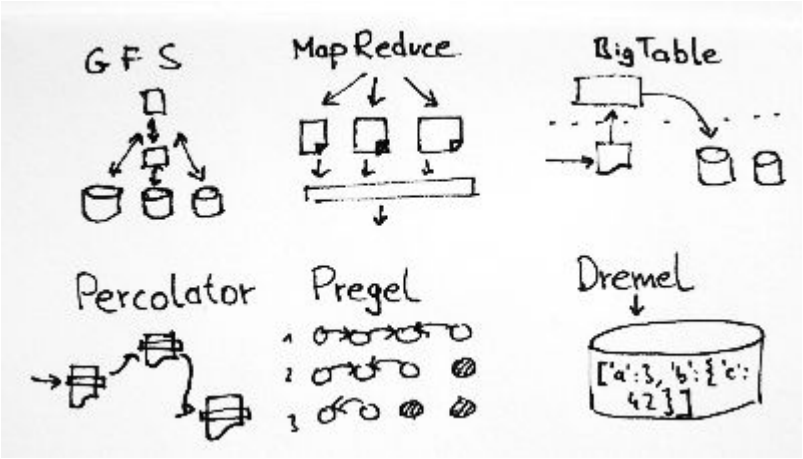
# Monolithic architecture: the challenge to scalability

- Any data application
  - It must be designed and implemented in a unified way
  - All computation and storage are managed as a whole
- Any update in any component
  - May incur unexpected changes on other components
  - We need a lot of tests before delivering a small feature
  - We cannot easily add more computation or storage resource

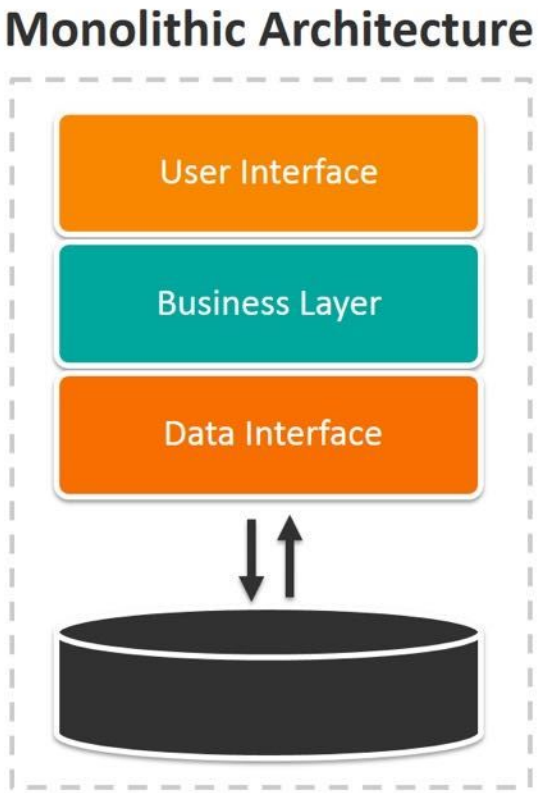


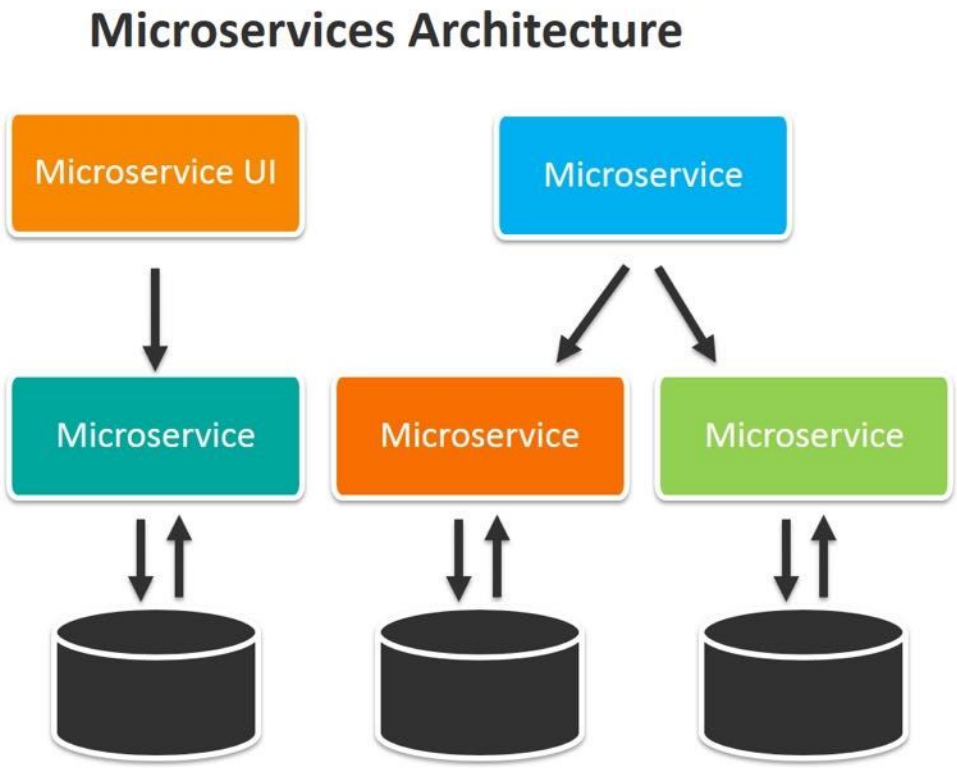
# Google’s initiative

- How can we use commodity PC to solve problems?
  - Data Storage (GFS)
  - Computation of PageRank (MapReduce)
  - Wide table (Big Table)
  - Graph data (Pregel)
- The interesting consequences
  - The beginning of open source era
  - The prevailing of distributed computing
  - The low cost of deployment

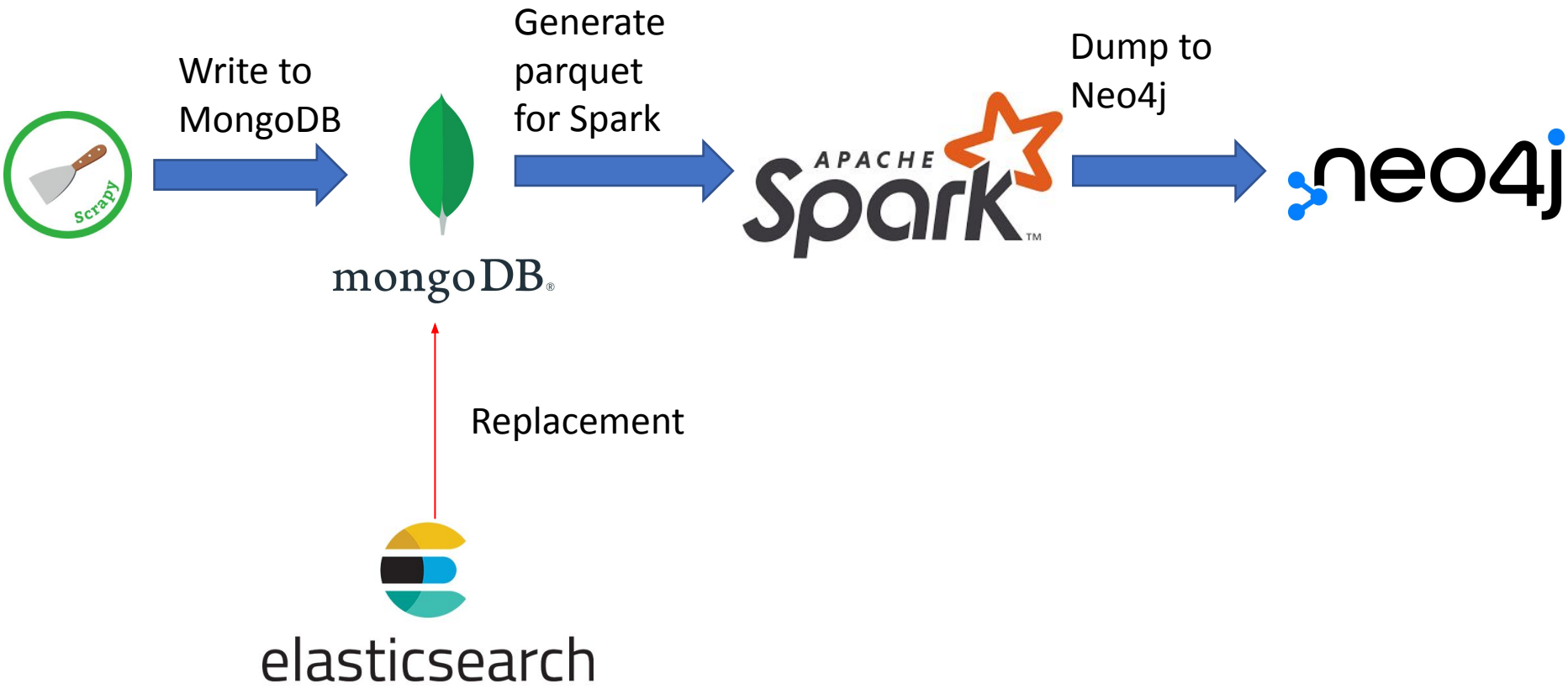


# Microservice: breaking the monolithic application into independent pieces

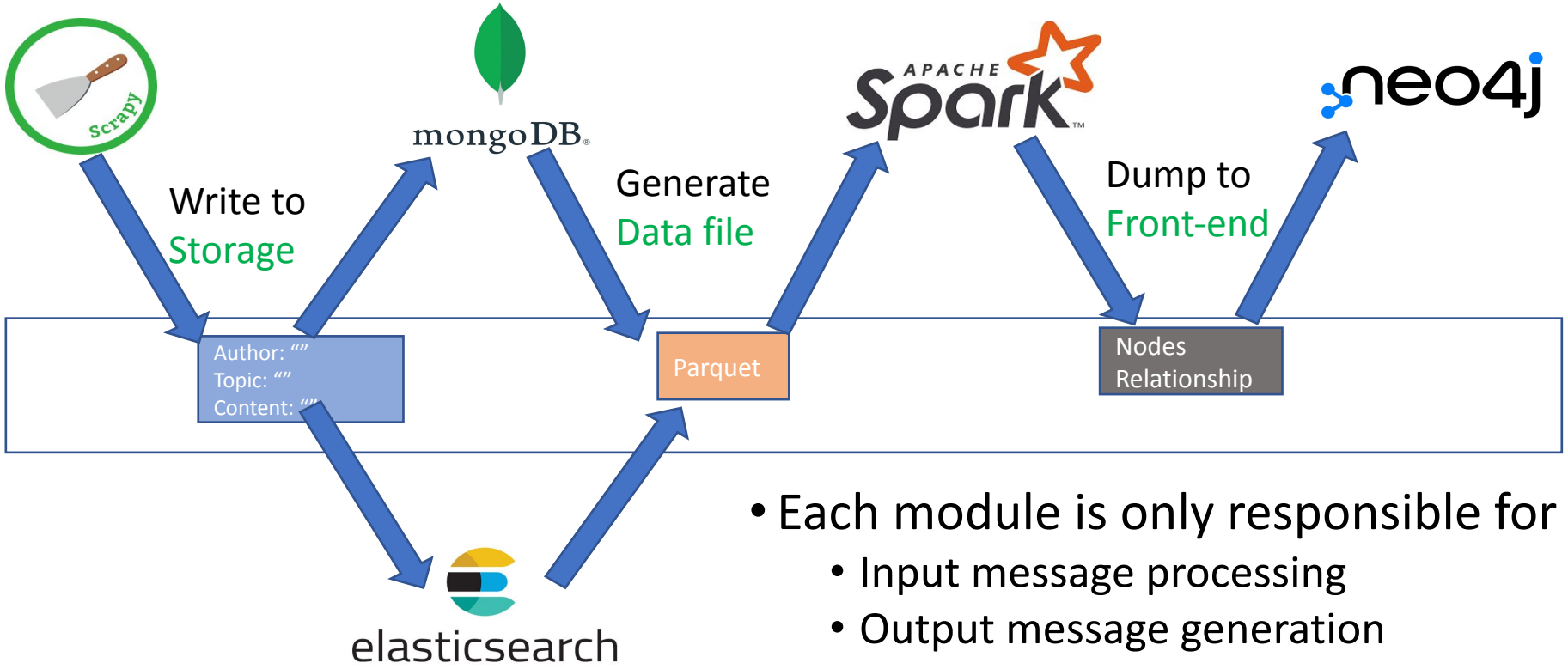


# A simple motivating example of microservice

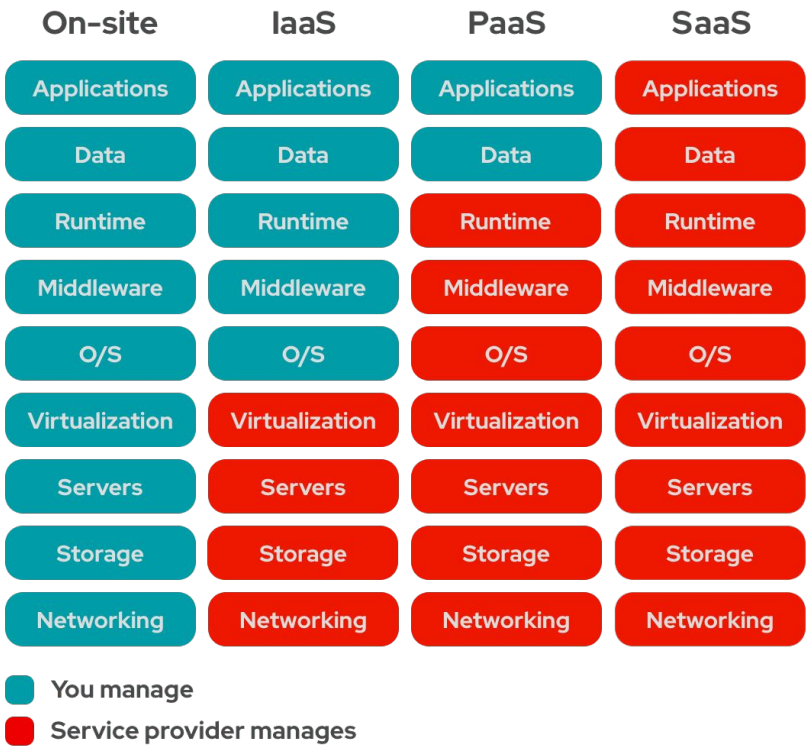


# Decoupling the application into independent services




- Each module is only responsible for
  - Input message processing
  - Output message generation

# The rise of the clouds and SaaS



# Modern data stack is the results of all these technical transformations

- A modern data stacks with the following characteristics
  - Everything is running on the cloud
  - It usually contains a portfolio of different SaaS tools
  - The tools are combined to support specific data tasks
  - The whole architecture is easily extensible
  - Each module can be easily replaced
  - The development of new logics is super easy

# Redshift: the milestone of modern data stack

- Redshift is the first data warehouse product on cloud (AWS)
  - Scalability
  - Low cost
  - Extensibility
- However, Redshift provides functions as a traditional data warehouse
- The modern data stack starts to evolve by answering a series of questions and demands

# Why do I have to rent a big instance when we query the warehouse only a few times a day?

aws Services Resource Groups

New EC2 Experience
Tell us what you think

EC2 Dashboard
Events
Tags
Reports
Limits

INSTANCES
Instances
Instance Types
Launch Templates New
Spot Requests
Dedicated Hosts

EC2 > Instance types

**Instance types** (254)

Filter instance types

| Instance type | vCPUs | Architecture | Memory (MiB) |
|---|---|---|---|
| a1.2xlarge | 8 | arm64 | 16384 |
| a1.4xlarge | 16 | arm64 | 32768 |
| a1.large | 2 | arm64 | 4096 |
| a1.medium | 1 | arm64 | 2048 |

# Snowflake: separation of storage and computation

# How can we load the data from different sources?

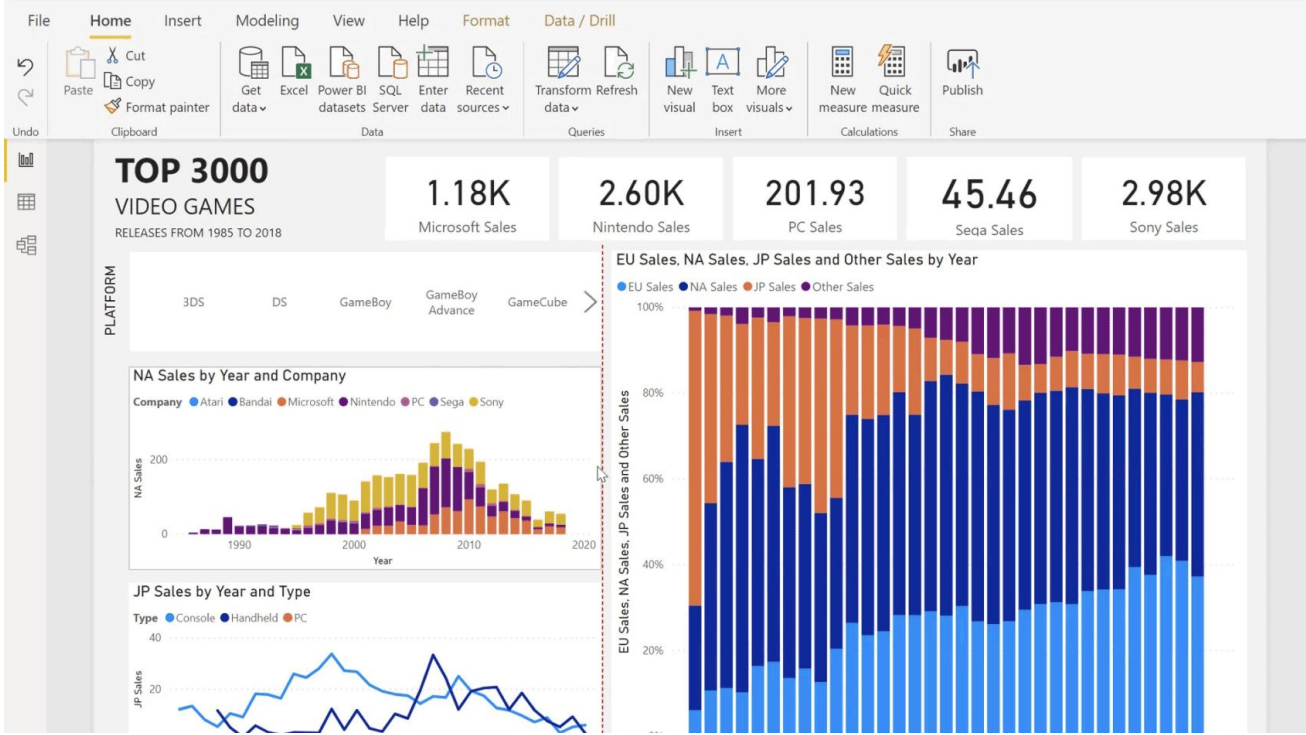
- Tools to support seamless data connection and ETL pipelines

Fivetran

Stitch
A Talend Company

# How can we visualize and present the data?

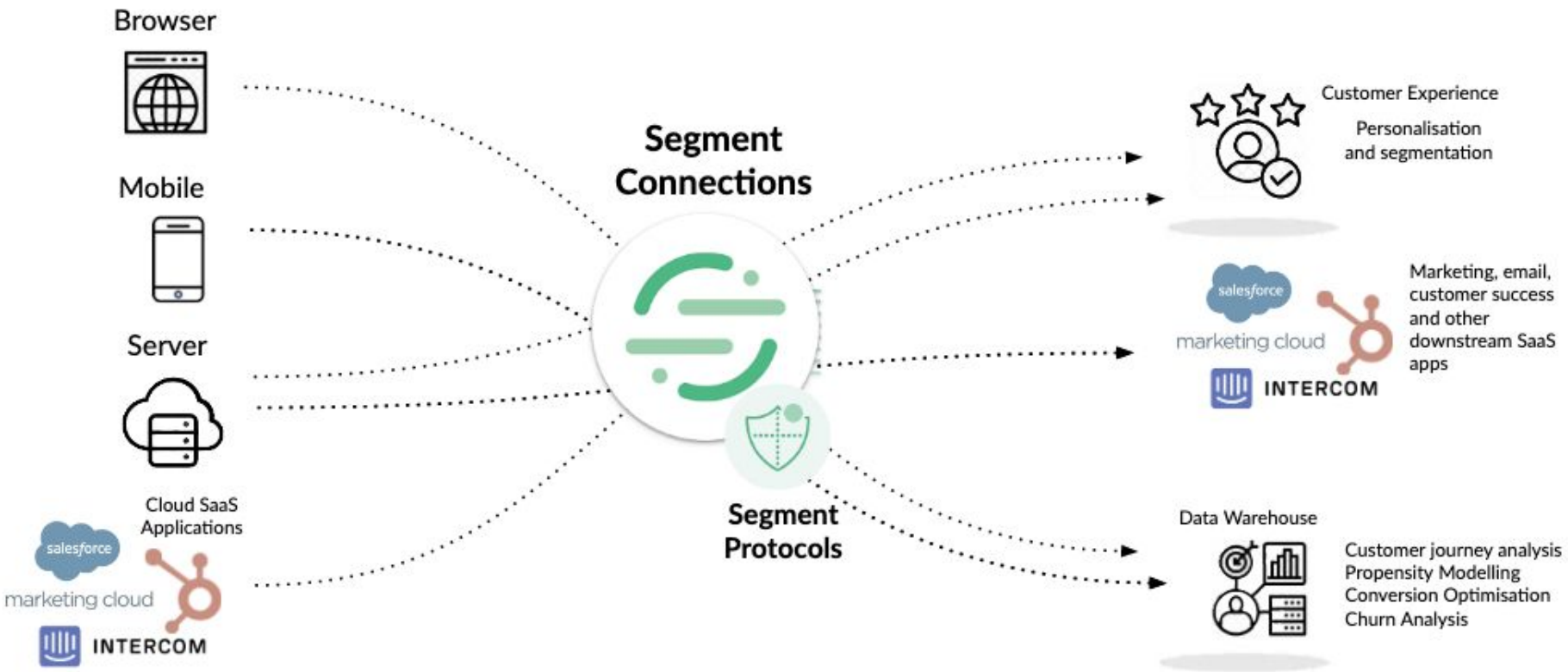
- Visualization of the data for decision makers




Power BI

tableau

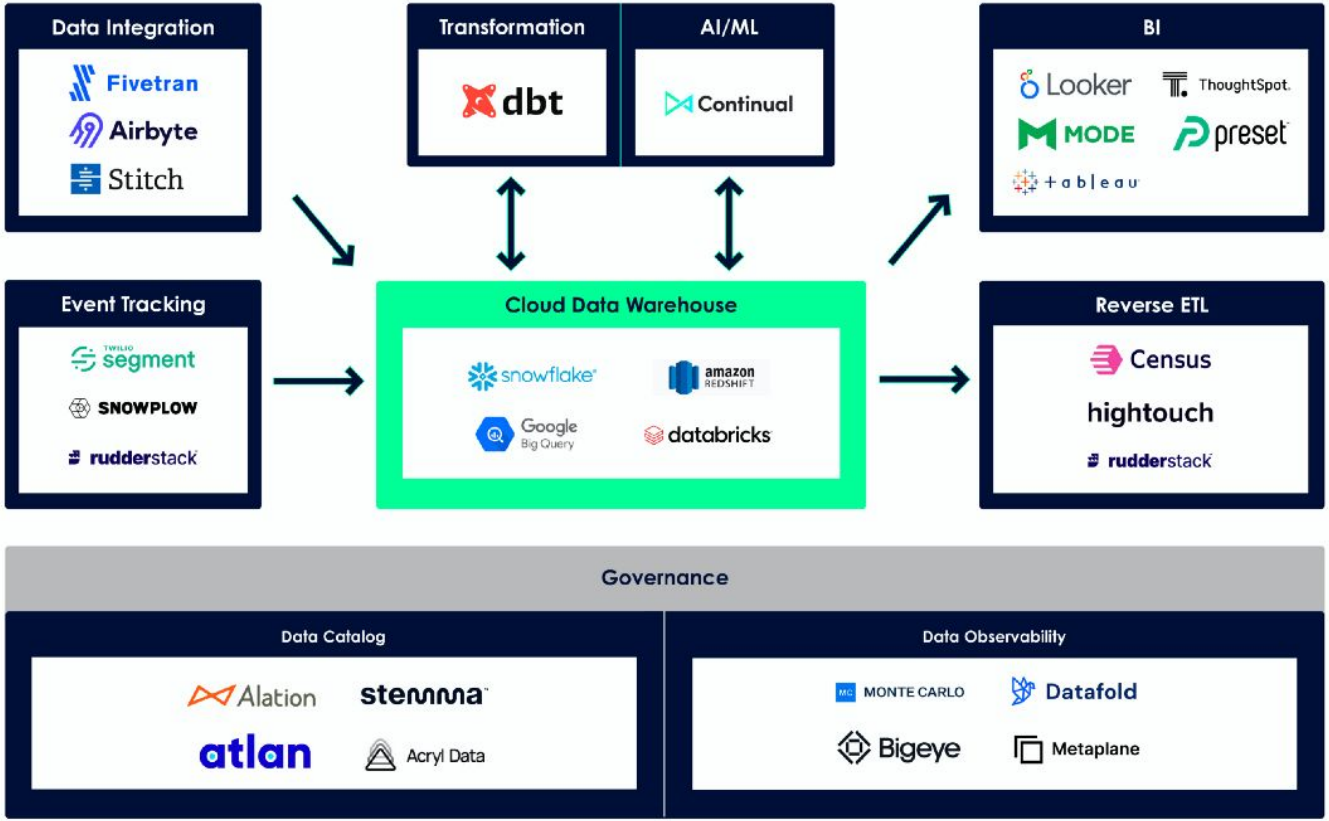
# How can we collect the data from the users?



# How can we better govern the data?

- Maintaining the meta-data
- Data discovery
- Security and privacy management

# A combined modern data stack landscape



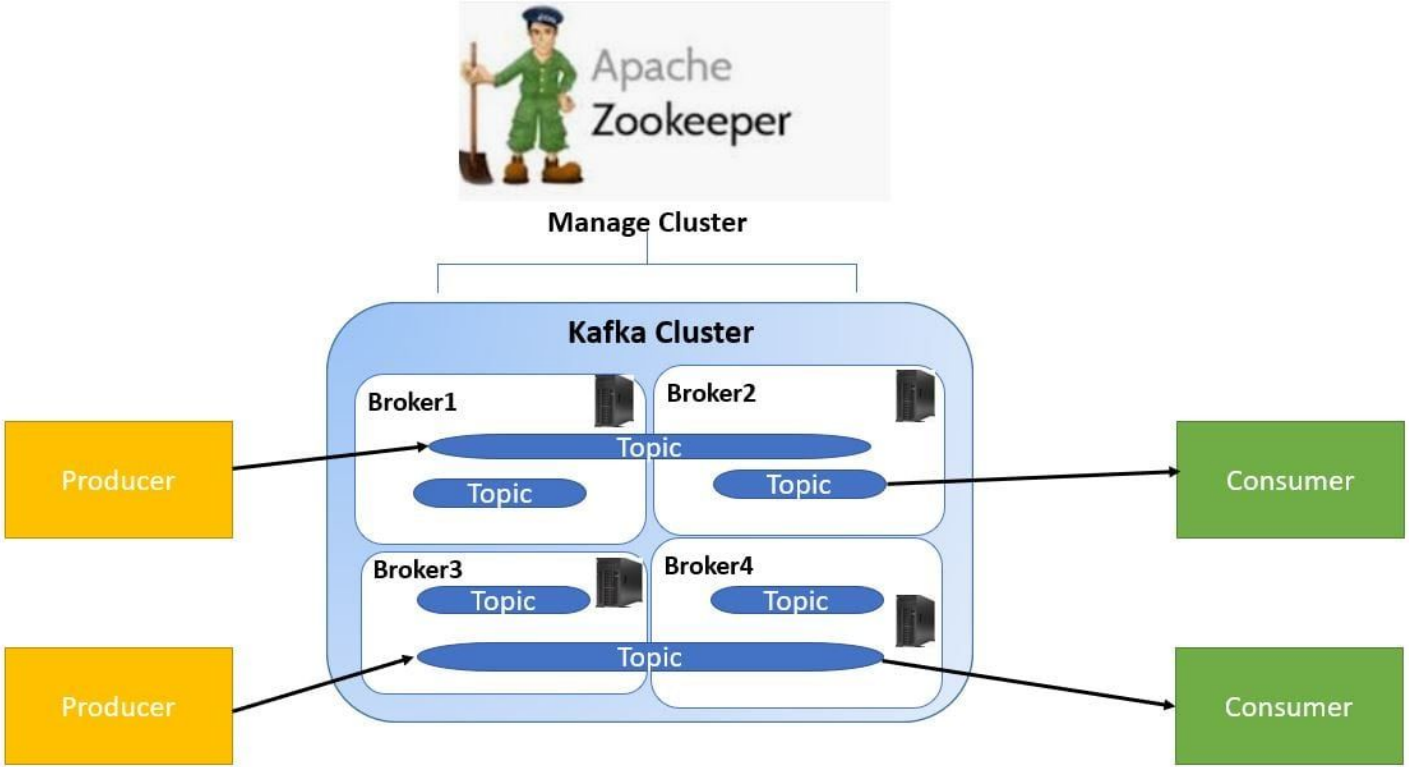
# Overview of the talk

- The history of data stack
- Trends in modern data stacks
- Open source and ecosystem
- Practice at Neuron Mobility

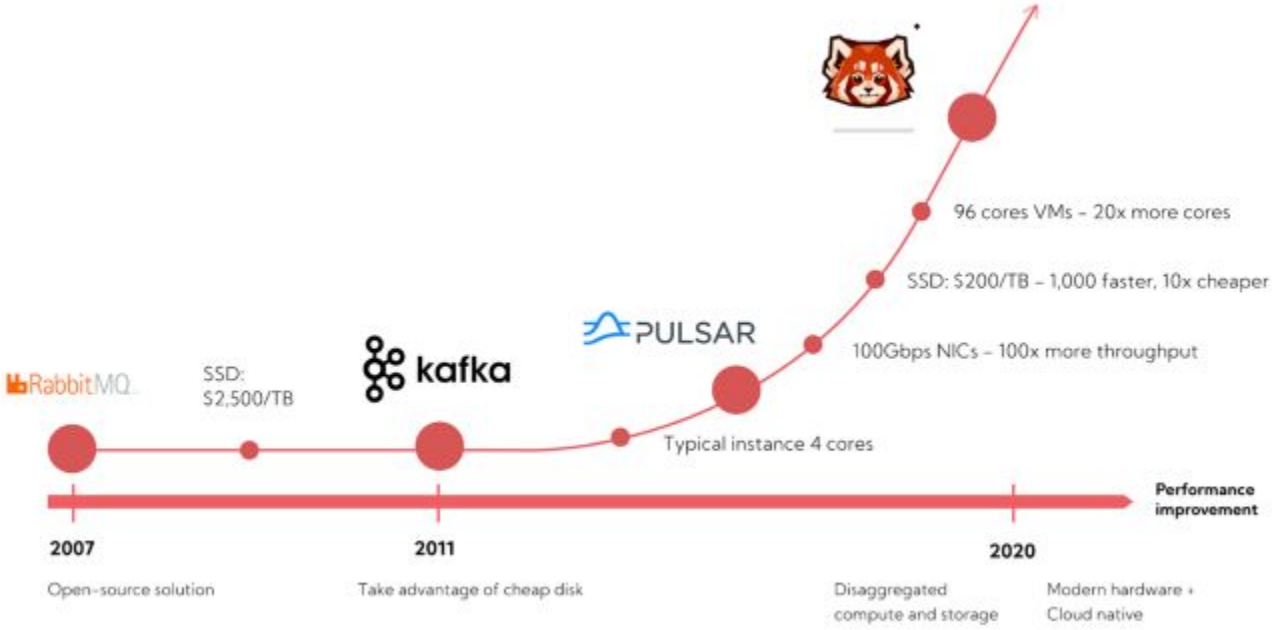
# MDS is evolving on these directions

- Performance optimization
  - Lower latency
  - Better scalability
- Data democratization
- Data tool simplification

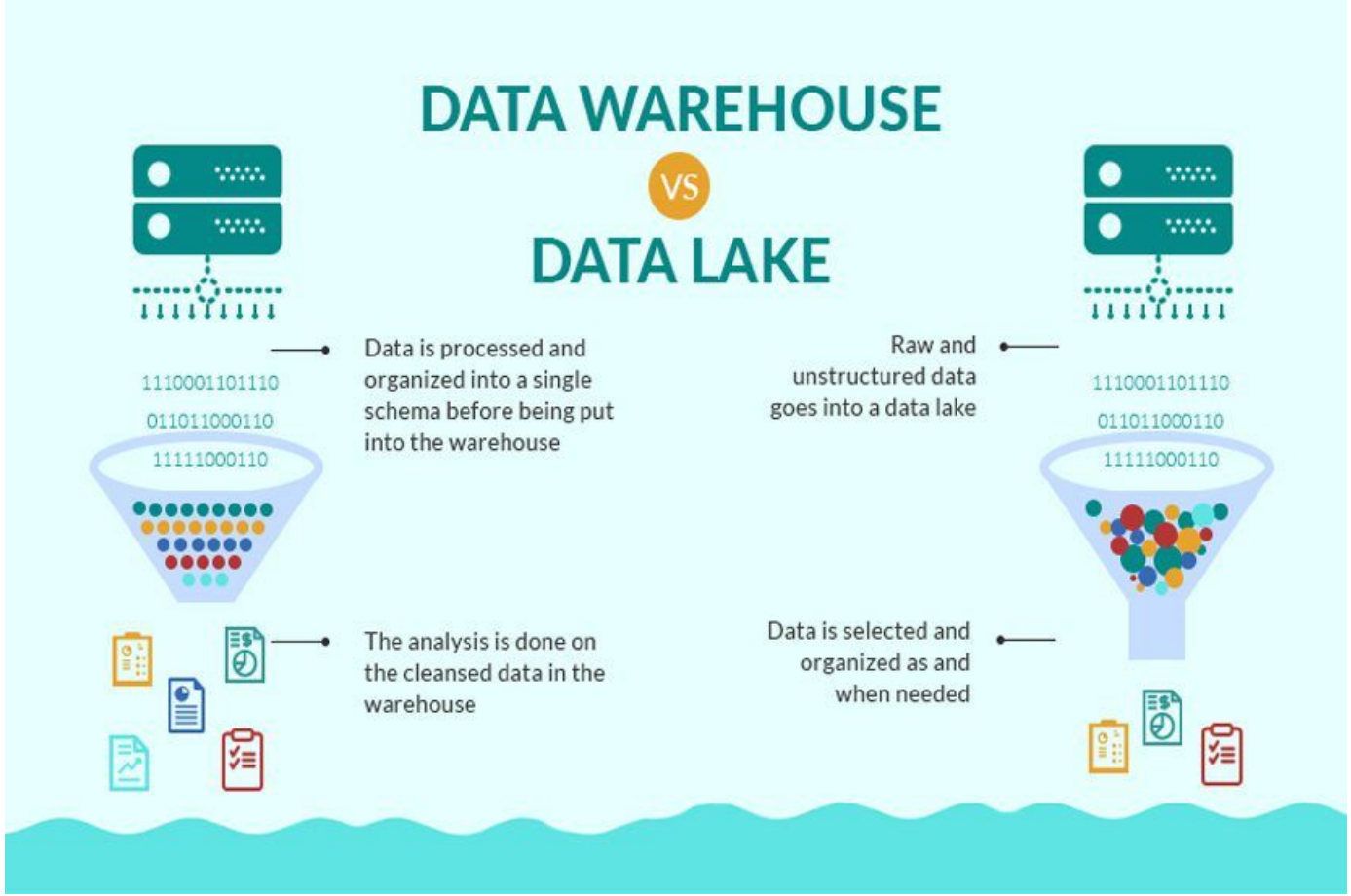
# From Kafka to Redpanda: An example of performance improvement



# From Kafka to Redpanda: An example of performance improvement



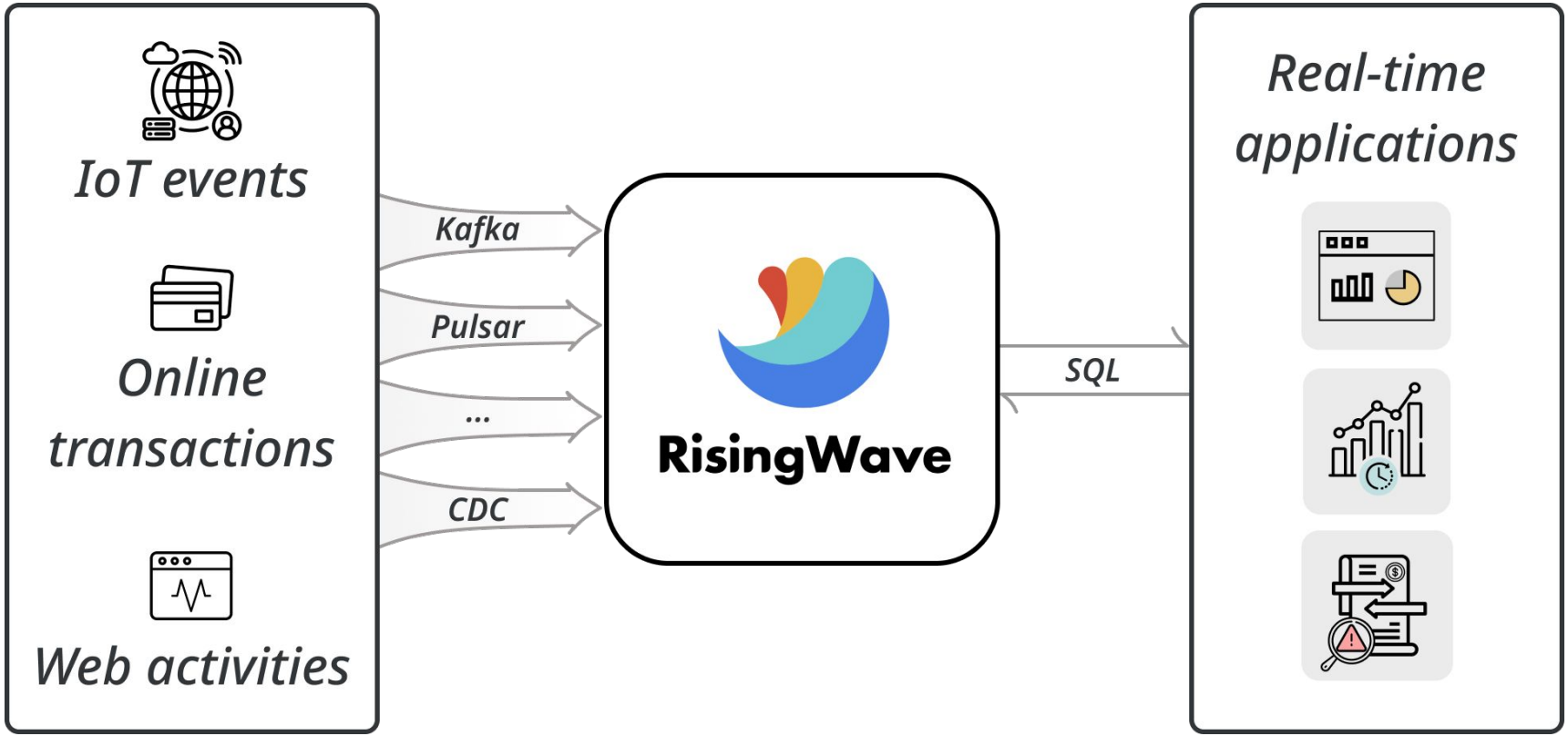
# Data lake


DATA WAREHOUSE
VS
DATA LAKE
Data is processed and organized into a single schema before being put into the warehouse
1110001101110
011011000110
11111000110
Raw and unstructured data goes into a data lake
1110001101110
011011000110
11111000110
The analysis is done on the cleansed data in the warehouse
Data is selected and organized as and when needed

# Database v.s. Data Warehouse v.s. Data Lake

| | Database | Data Lake | Data Warehouse |
|---|---|---|---|
| **Workloads** | Operational and transactional | Analytical | Analytical |
| **Data Type** | Structured or semi-structured | Structured, semi-structured, and/or unstructured | Structured and/or semi-structured |
| **Schema Flexibility** | Rigid or flexible schema depending on database type | No schema definition required for ingest (schema on read) | Pre-defined and fixed schema definition for ingest (schema on write and read) |
| **Data Freshness** | Real time | May not be up-to-date based on frequency of ETL processes | May not be up-to-date based on frequency of ETL processes |
| **Users** | Application developers | Business analysts, application developers, and data scientists | Business analysts and data scientists |

# Real-time database



# Real-time warehouse: ClickHouse

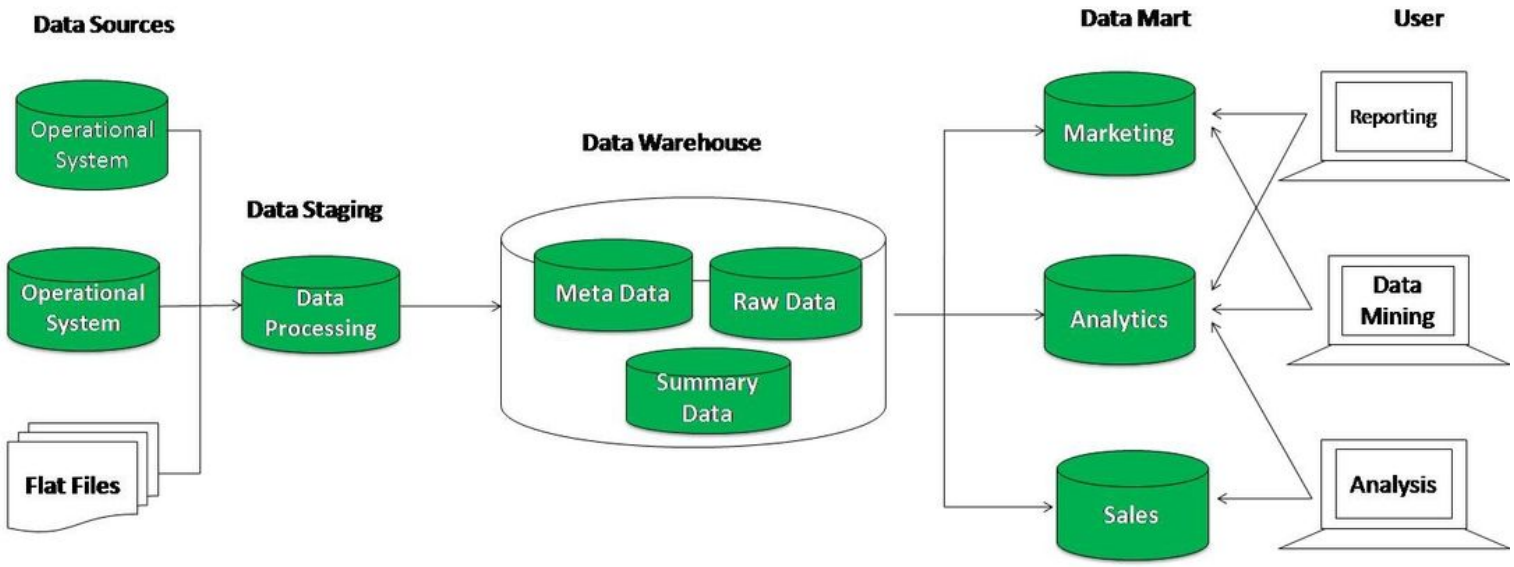
| | Snowflake | ClickHouse |
|---|---|---|
| **Low-latency dashboards** | Dozens of second load times at 100s of GB scale | Sub-second load times at TB scale |
| **Enterprise BI** | Mature and broad Enterprise DW featureset | Limited integrations with Enterprise BI ecosystem tools. |
| **Data Apps (Customer-facing, low latency, high concurrency)** | – Dozens of second load times at 100s of GB scale.<br><br>– Scale-out to more clusters required starting from dozens of concurrent queries. | – Sub-second load times at TB scale.<br><br>– Supports hundreds of concurrent queries on a single cluster. |
| **Ad hoc** | Decoupled storage/compute architecture allows to spin up ad-hoc resources | – Performance is dependent on predefined indexing.<br><br>– Coupled storage/compute means single Ad-Hoc query can easily hog cluster. |

# New concepts are emerging in the domain of MDS

- [Data Democratization] Data Mesh
- [Data Flexibility] Reverse ETL
- [Data Confidence] Metric Layer
- [Data Governance] Smart Data Catalog
- [Data Management] Data Observability

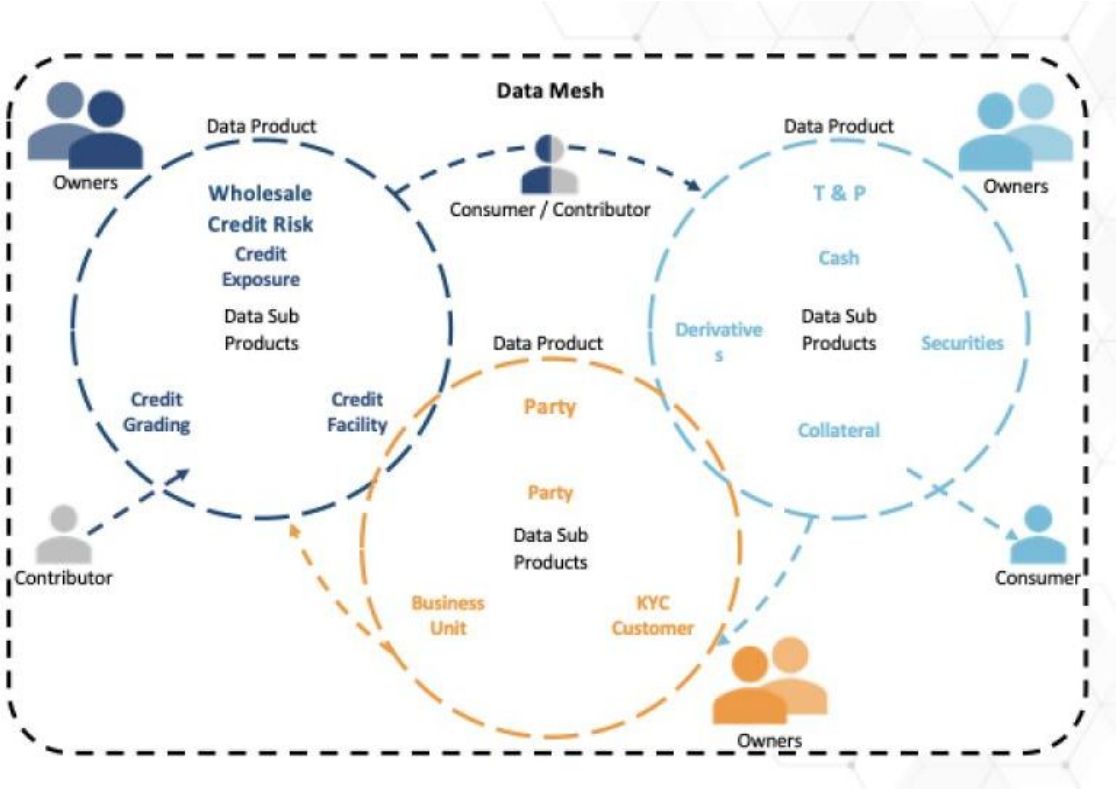
# Data Democratization

- Accountability challenge
  - Data producer is not responsible for data quality
  - Users complaint to data analyst on data quality
  - Data engineer can hardly do anything to enhance data quality



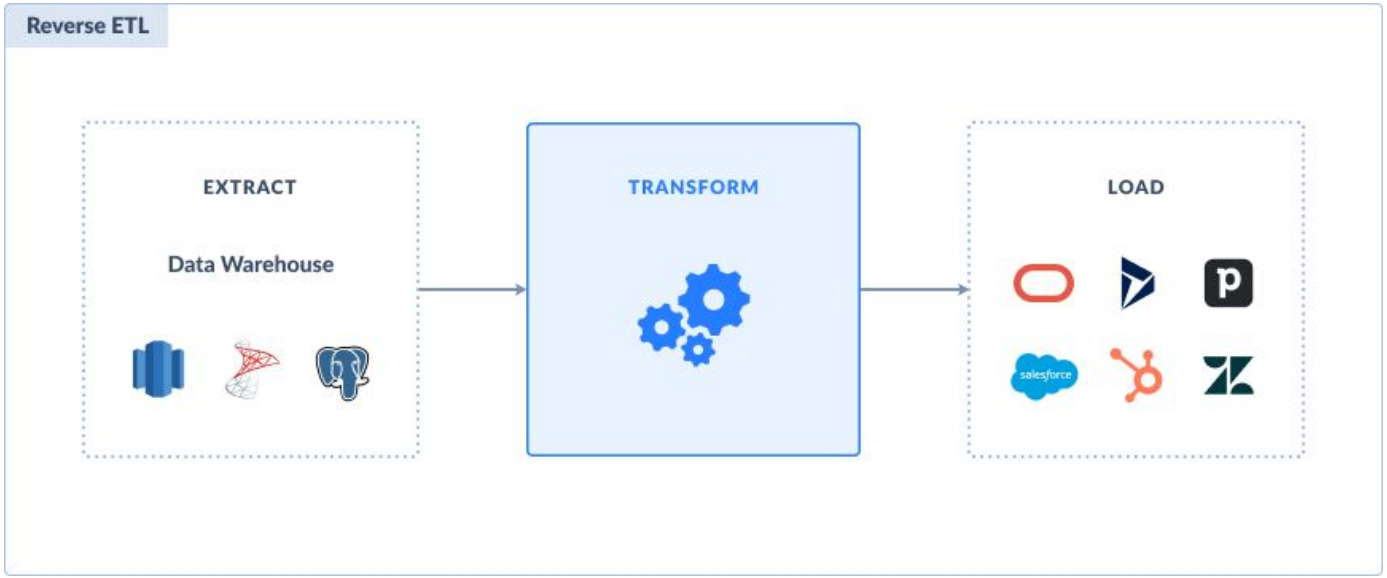
# Data mesh: an innovative way of data decentralization

- Each domain has a domain manager
- Data as a product
- Infrastructure to support data exchange

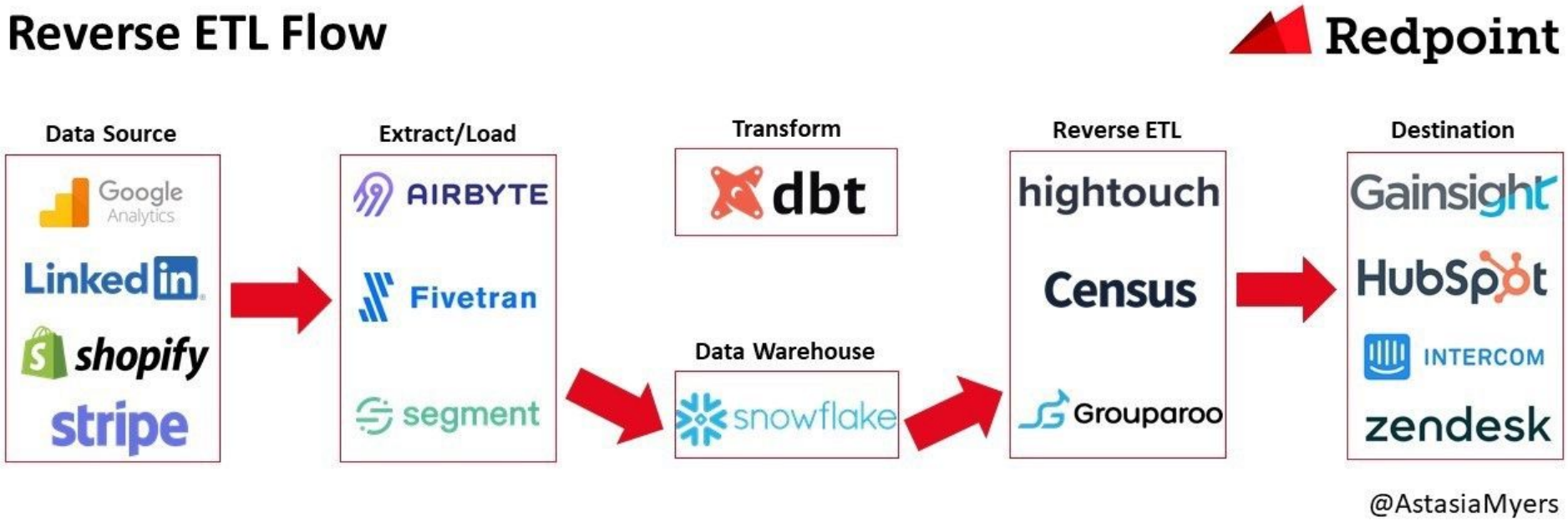


# The diversity of downstream data consumers

- All these consumers need data in very different format
  - SQL
  - JSON
  - Text
- ETL converts dirty data into clean format
- Reverse ETL converts clean data into complex format



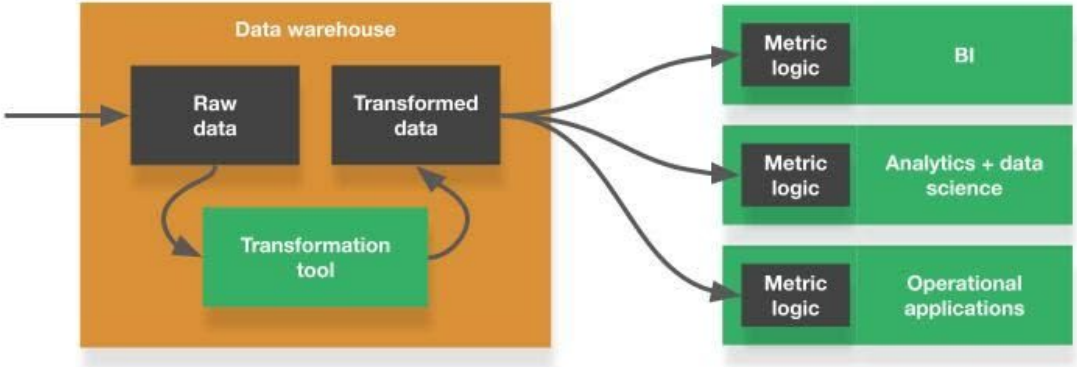
# Reverse ETL closes the loop of data warehouse ecosystem



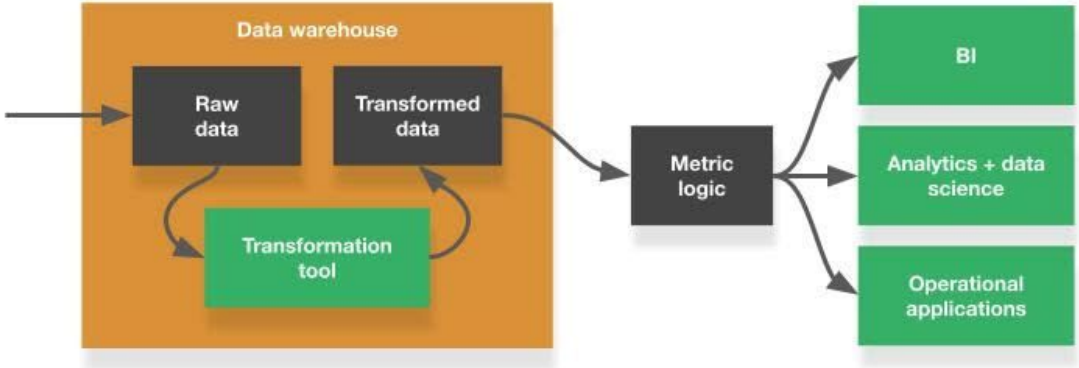
# Metric Layer: unifying all metrics in an organization

- Two dashboards may present different numbers on the same metric (or different names)
- The users (say CFO) may get confused with the contradiction

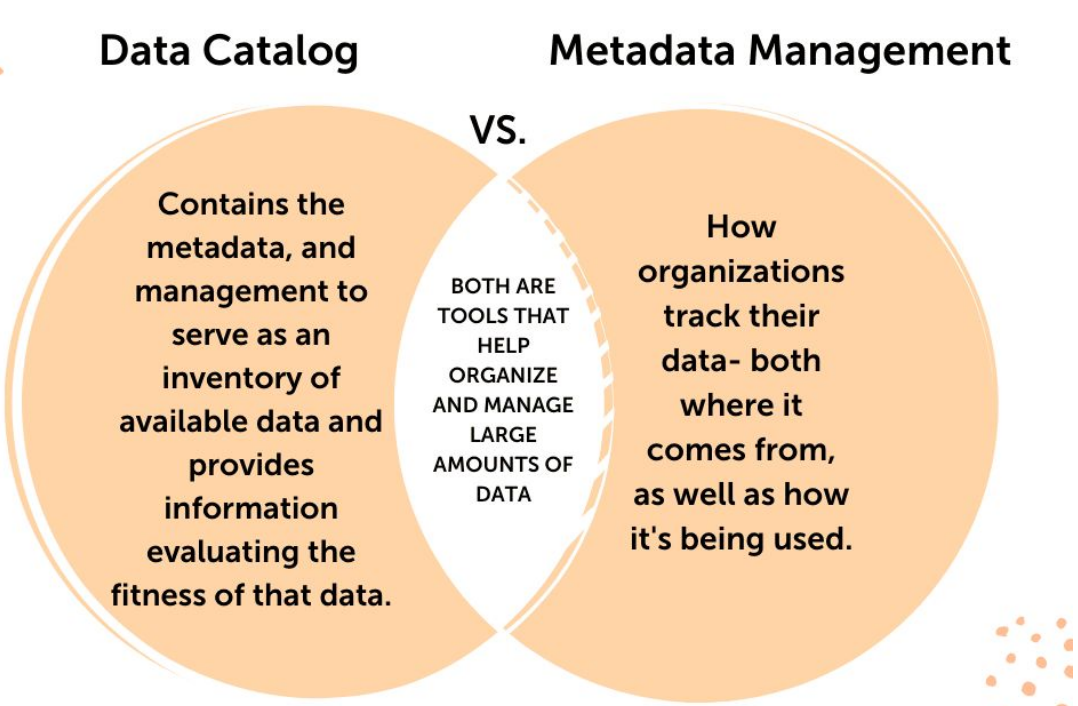
## Today's architecture




## The proposed metrics layer

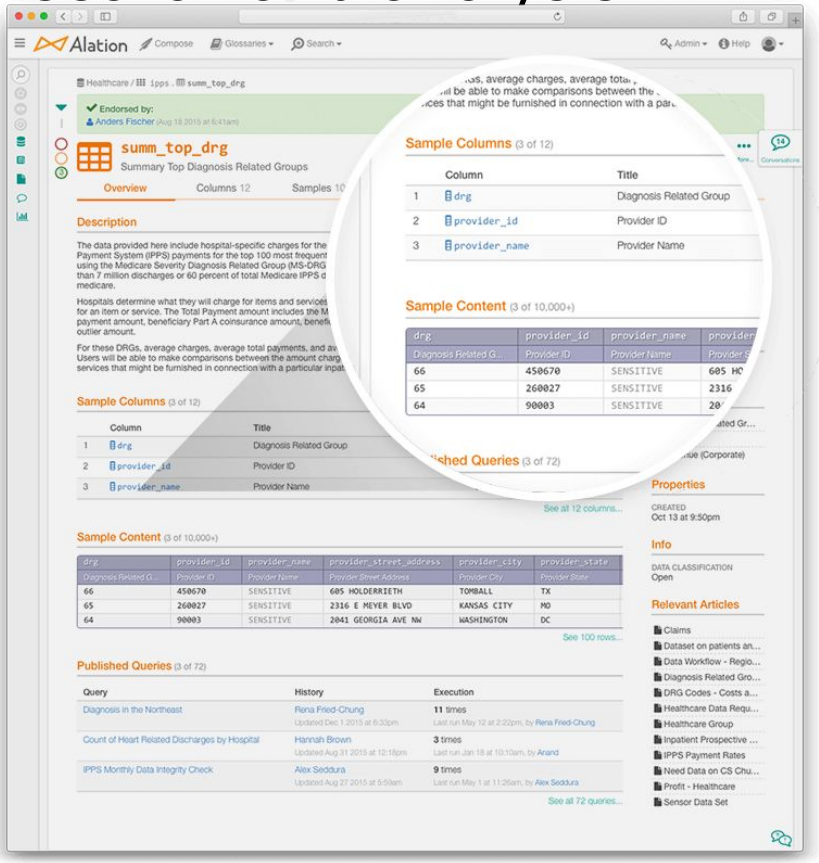


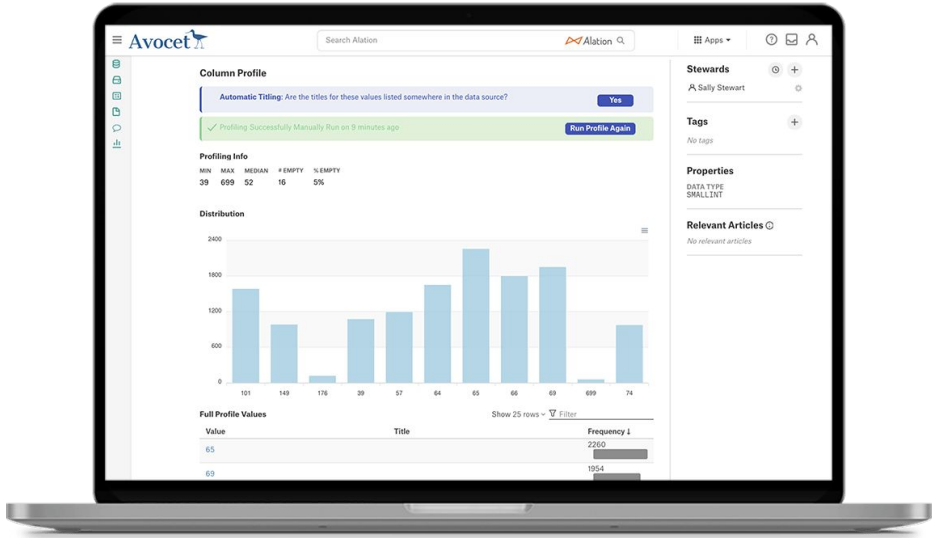
# Smart Catalog: A better management on your data definition




Alation

# Data Catalog: Not only for documentation, but also for search and analysis



# Overview of the talk

- The history of data stack
- Trends in modern data stacks
- Open source and MDS
- Practice at Neuron Mobility

# Open Source licenses are different

| | APACHE | BSD | MIT | GPL v3 Free as in Freedom | LGPL v3 Free as in Freedom | AGPL v3 Free as in Freedom |
|---|---|---|---|---|---|---|
| **Type** | Permissive | Permissive | Permissive | Copyleft | Copyleft | Copyleft |
| **Provides copyright protection** | ✓ TRUE | ✓ TRUE | ✓ TRUE | ✓ TRUE | ✓ TRUE | ✓ TRUE |
| **Can be used in commercial applications** | ✓ TRUE | ✓ TRUE | ✓ TRUE | ✓ TRUE | ✓ TRUE | ✓ TRUE |
| **Provides an explicit patent license** | ✓ TRUE | ✗ FALSE | ✗ FALSE | ✗ FALSE | ✗ FALSE | ✗ FALSE |
| **Can be used in proprietary (closed source) projects** | ✓ TRUE | ✓ TRUE | ✓ TRUE | ✗ FALSE | ✗ FALSE partially | ✗ FALSE for web |
| **Popular open-source and free projects** | Kubernetes<br>Swift<br>Firebase | Django<br>React<br>Flutter | Angular.js<br>JQuery,<br>.NET Core<br>Laravel | Joomla<br>Notepad++<br>MySQL | Qt<br>SharpDevelop | SugarCRM<br>Launchpad |

# An interesting story of source code requests

**Patrycja** @ptrcnull

my favorite corporate interaction so far

Ben commented:

Hi,

You can request the shareable source codes (most of them are not free and owned by MediaTek) at our Shenzhen office (only Chinese speaking) in working hours.
The address is:
405-407 Jinqi Zhigu Building , 4/F , 1 Tangling Road , Nanshan District , Shenzhen City, P.R.C

Kind regards, Ben - UMIDIGI

---

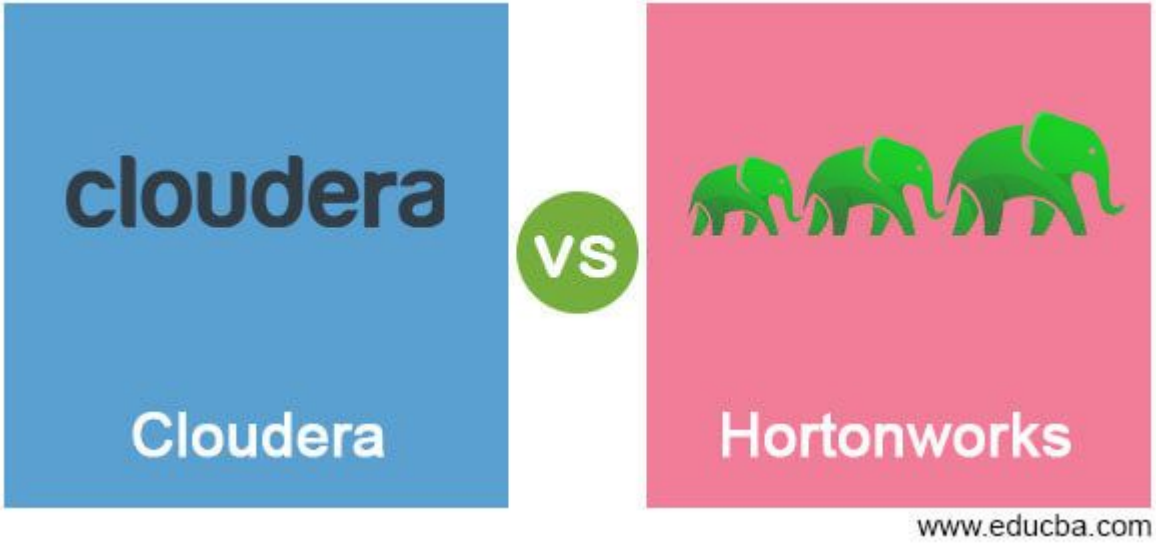
Ben changed the status to Waiting for customer.

9:14 PM · Aug 20, 2021

4.8K Reply Copy link

# Hadoop ecosystem

- Cloudera and Hortonworks are the first generation of open source commercial software vendors
- Hadoop open source community was like US political system 10 years ago.



# Data Brick is the biggest open source company among other famous softwares

- Every new MDS start-up is based on open source software
- The business logic is now quite different
  - Core functionalities are all available in open source version
  - Management and deployment tools are only available in the enterprise version
  - SaaS and pay-as-you-go scheme

databricks

mongoDB

ClickHouse

# Building MDS is like playing with bricks

# Overview of the talk

- The history of data stack
- Trends in modern data stacks
- Open source and MDS
- Practice at Neuron Mobility

# Practice in Neuron Mobility: Business Overview

- Neuron Mobility is a Singapore-based startup company
- We design, build and operate shared scooters in commonwealth countries
- There are more than 150,000 riders per week on Neuron scooters
- We are the dominating operator in Australia

# High diversity in the data flow

- User behaviour flow
- Ground operation flow
- IoT data flow
- Payment flow

# High diversity on data destination: real-time update and analytical workloads

- User app
  - Location update, scooter availability
- Operator app
  - Scooter position and status
- Analytical dashboard
  - Trip distribution
- Google sheets
  - Finance data for manual adjustment

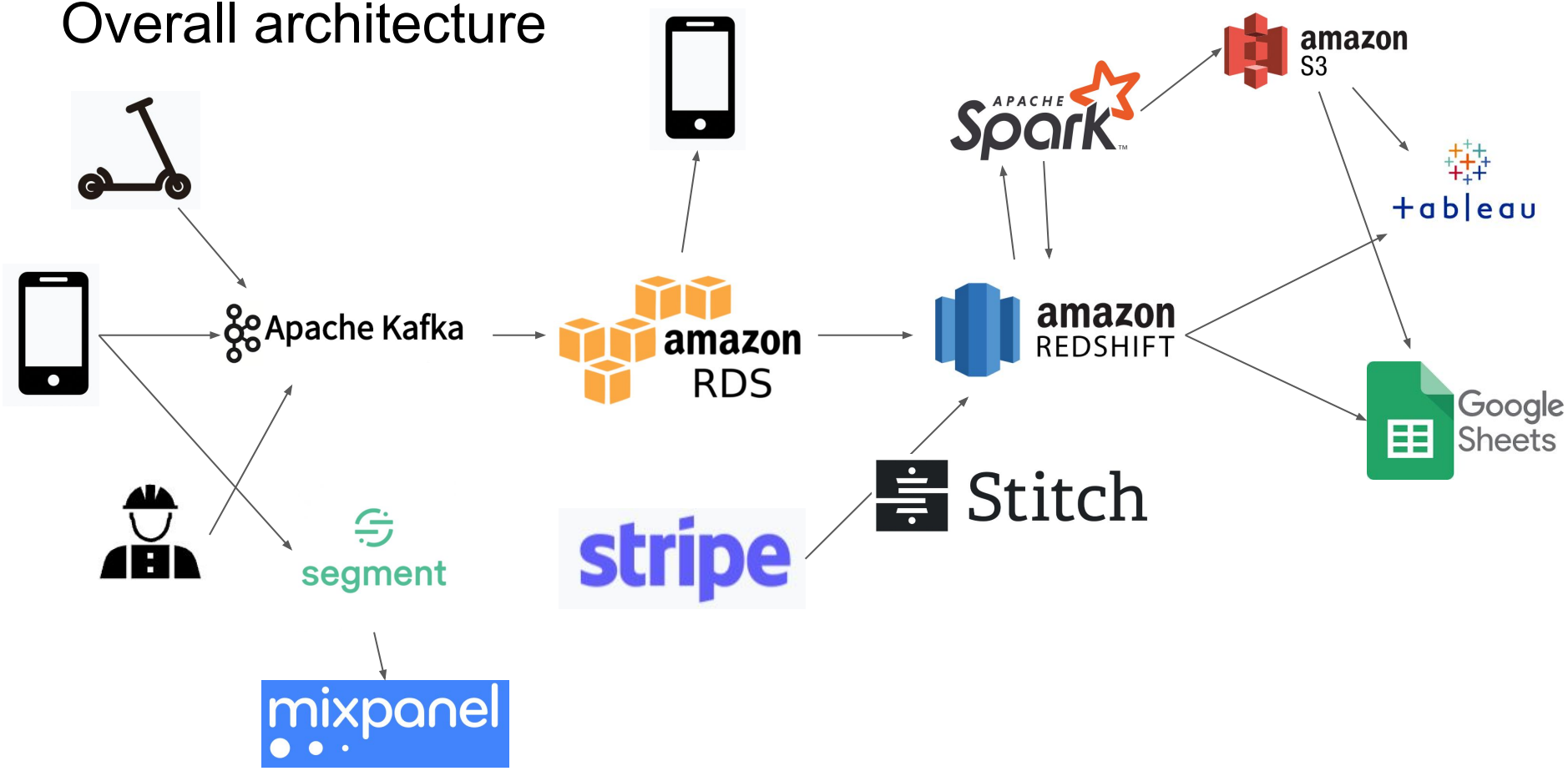
# Geographical and privacy challenges

- Our markets are across four continents
  - Asia, Europe, Oceania and North America
- Each market has its own data privacy requirements
  - GDPR in UK
  - PDPA in Singapore

# Everything is on cloud

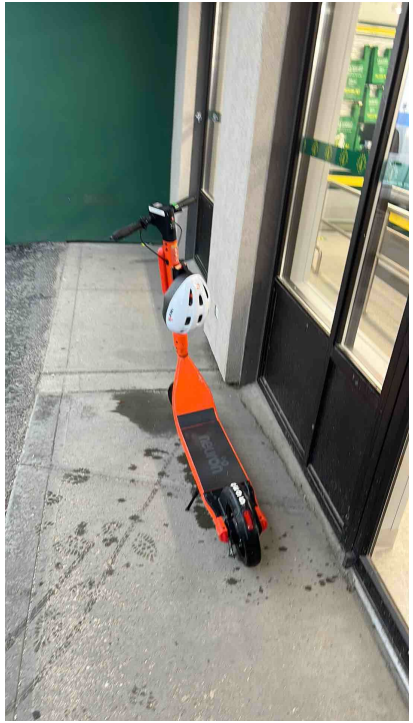
- We don’t have any on-premise servers
- This gives us flexibility to customize data services in each of the market

# Overall architecture



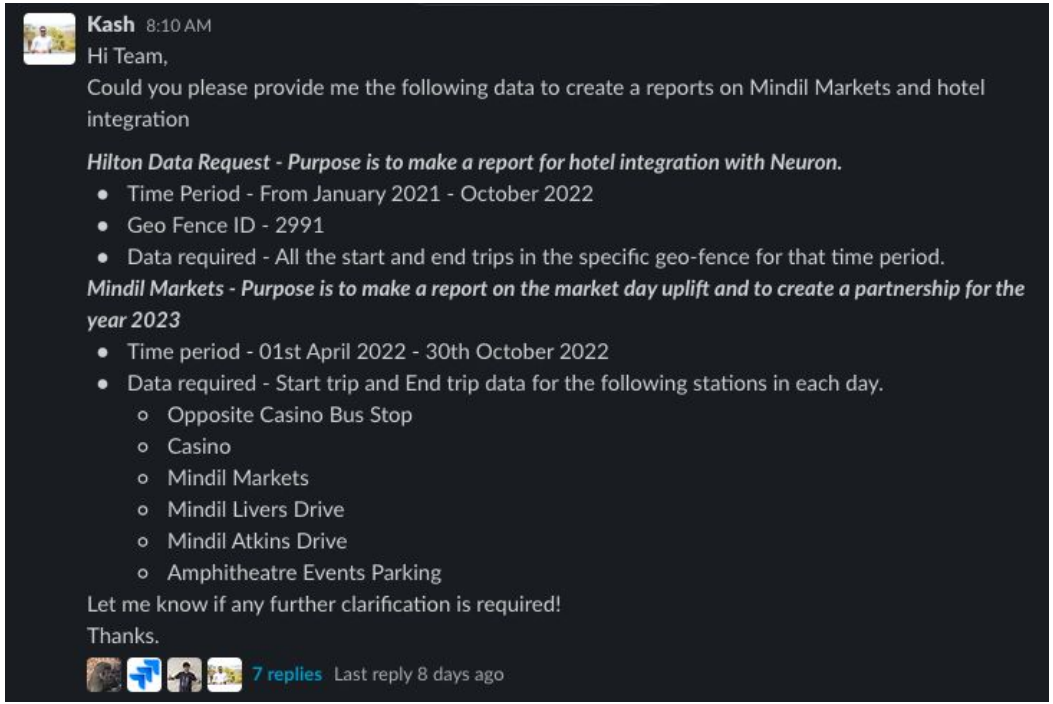
# Data workstreams: Data as a Product

- We build our own data product
  - External reporting to city councils
  - Incident management system
  - User profiling engines
  - Internal tools: data chat bot
  - AI services: parking photo recognition



# Data Workstreams: Data as a Service

- There are always unexpected data questions



**Kash** 8:10 AM

Hi Team,

Could you please provide me the following data to create a reports on Mindil Markets and hotel integration

***Hilton Data Request - Purpose is to make a report for hotel integration with Neuron.***

- Time Period - From January 2021 - October 2022
- Geo Fence ID - 2991
- Data required - All the start and end trips in the specific geo-fence for that time period.

***Mindil Markets - Purpose is to make a report on the market day uplift and to create a partnership for the year 2023***

- Time period - 01st April 2022 - 30th October 2022
- Data required - Start trip and End trip data for the following stations in each day.
  - Opposite Casino Bus Stop
  - Casino
  - Mindil Markets
  - Mindil Livers Drive
  - Mindil Atkins Drive
  - Amphitheatre Events Parking

Let me know if any further clarification is required!

Thanks.

7 replies Last reply 8 days ago

# Data workstreams: Data Consultancy

- Special projects to support business decision making
  - Deployment strategy comparison against competitors
  - And others…

# Takeaway messages

- Modern data stack (MDS) is growing with the development of distributed computing, cloud and SaaS
- MDS is usually a portfolio of SaaS, used to build end-to-end data processing logics
- There are a number of start-up companies in each niche domain of MDS
- Open-source software are becoming increasingly popular and profitable in the industry
- The design of MDS heavily relies on the business model