Knowledge from Nothing: Creating Data Pipelines for User Insights from Zero

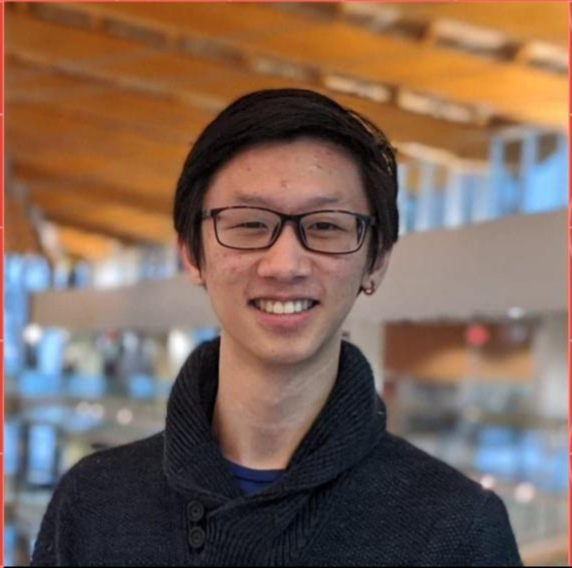
Jeffrey Leung

Software Developer

NEW YORK INSTITUTE OF TECHNOLOGY
Office of Development & Alumni Relations

Google Developer Groups
Vancouver

Women Techmakers
Vancouver

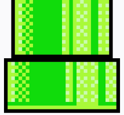


Google Developer Groups

devfest

Vancouver

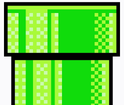
```
child: Col
  crossAxi
  children
    /*2*/
    Contai
      padd
      chil
        'D
        st
          
        ),
      ),
    ),
    Text(
      'Mul
      styl
        co
```

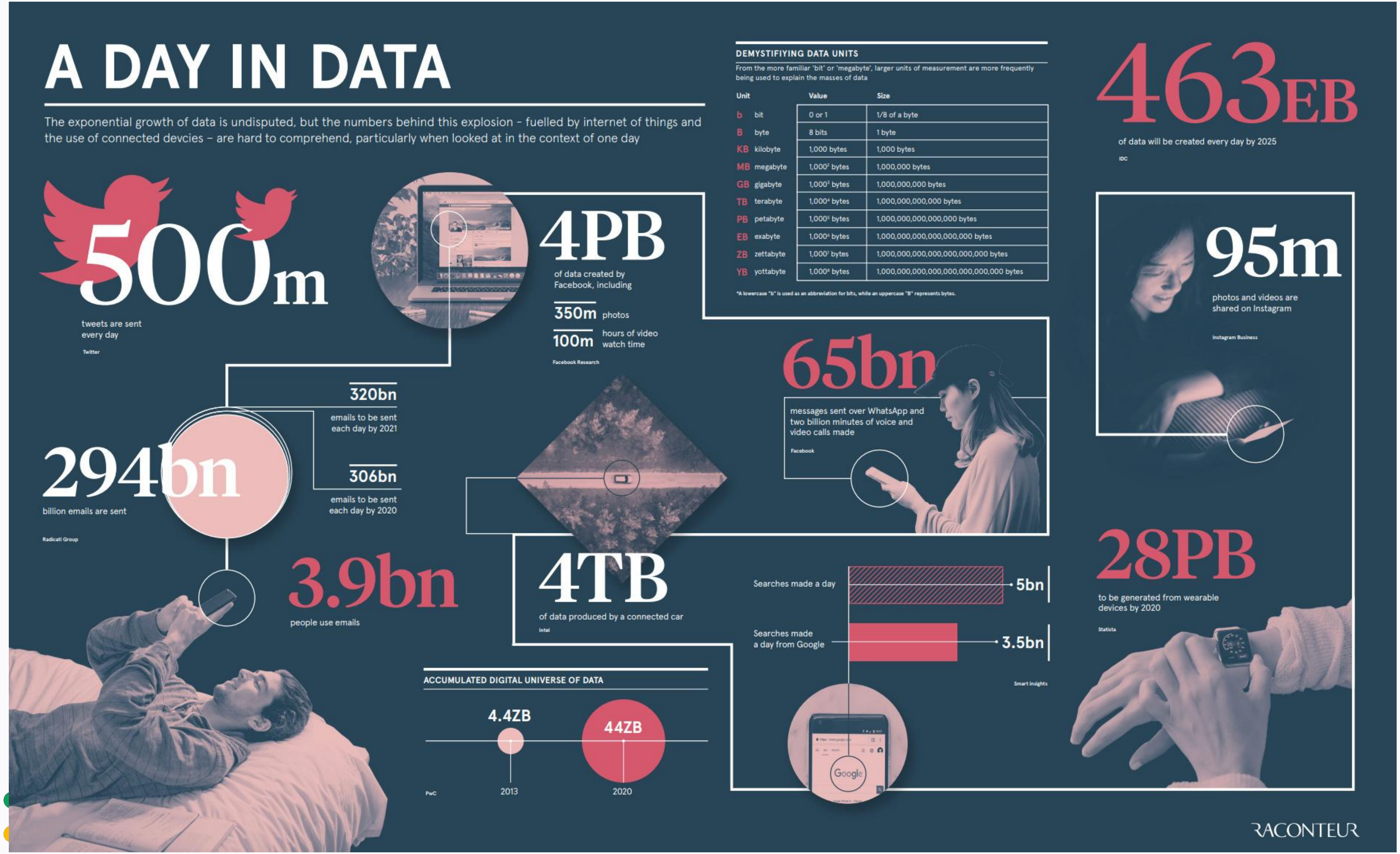


Knowledge from Nothing

Creating Data Pipelines for User Insights from Zero

JEFFREY LEUNG





A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m
tweets are sent every day
Twitter

4PB
of data created by Facebook, including
350m photos
100m hours of video watch time
Facebook Research

294bn
billion emails are sent
Radicati Group

320bn
emails to be sent each day by 2021

306bn
emails to be sent each day by 2020

3.9bn
people use emails

4TB
of data produced by a connected car
Intel

ACCUMULATED DIGITAL UNIVERSE OF DATA

PwC

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB
of data will be created every day by 2025
IDC

95m
photos and videos are shared on Instagram
Instagram Business

65bn
messages sent over WhatsApp and two billion minutes of voice and video calls made
Facebook

Smart insights

28PB
to be generated from wearable devices by 2020
Statista

RACONTEUR

Why do you(r leaders) want data?

Analyze past performance.

Understand current performance.

Experiment on new features.

Make informed business decisions.

Machine Learning for new insights.

Outline

0. data engineering fundamentals

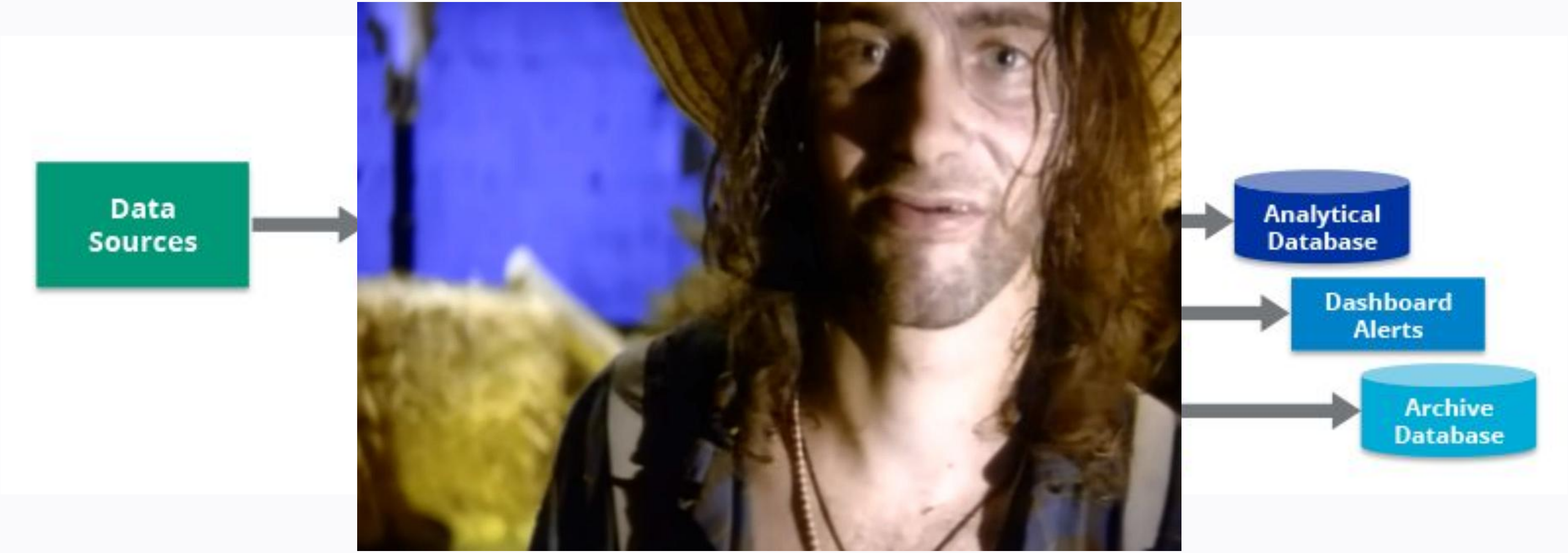
1. starter tools

2. meaning from a mess

3. organizational maturity

0. the fundamentals

Where does it come from, where does it go?



Data Sources

Analytical Database

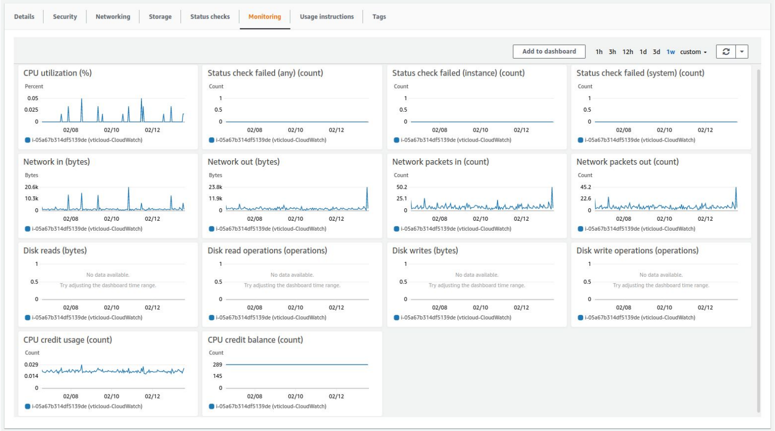
Dashboard Alerts

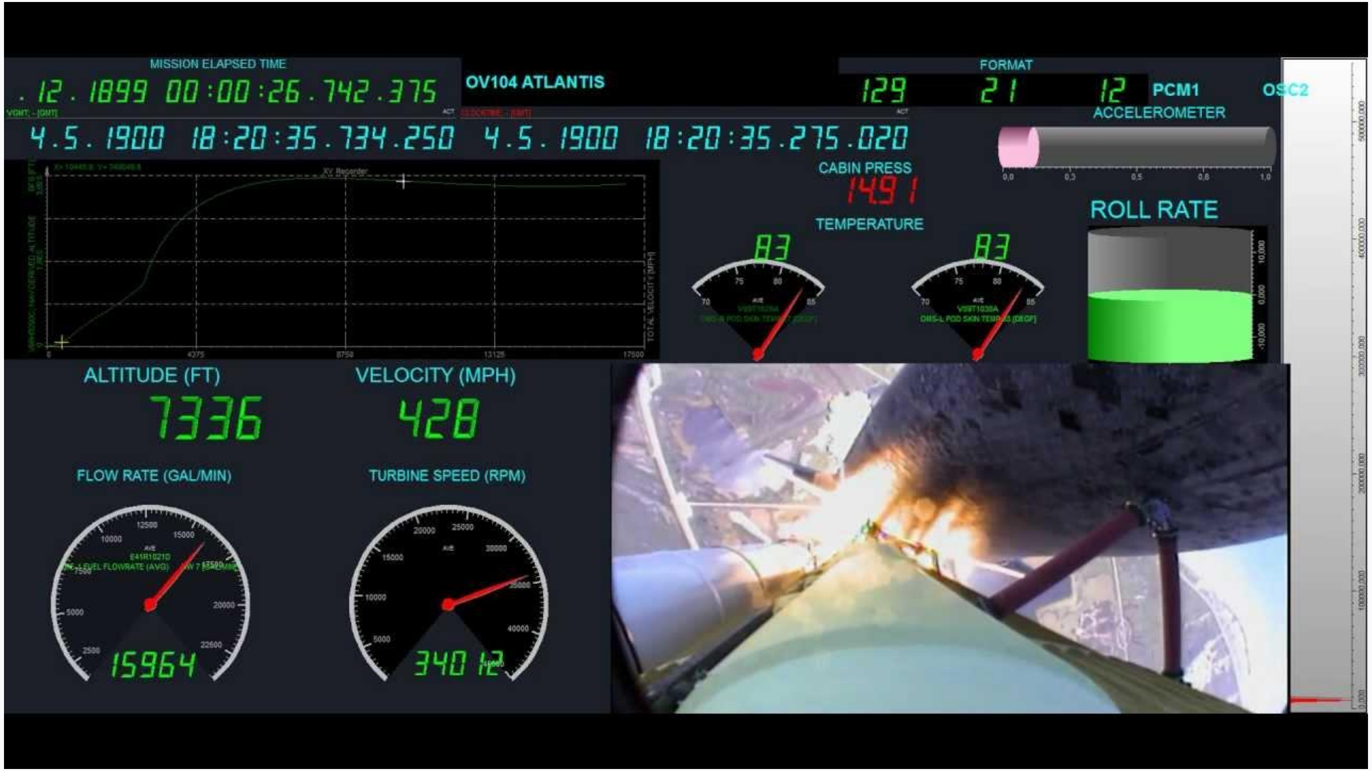
Archive Database

What kind of data?

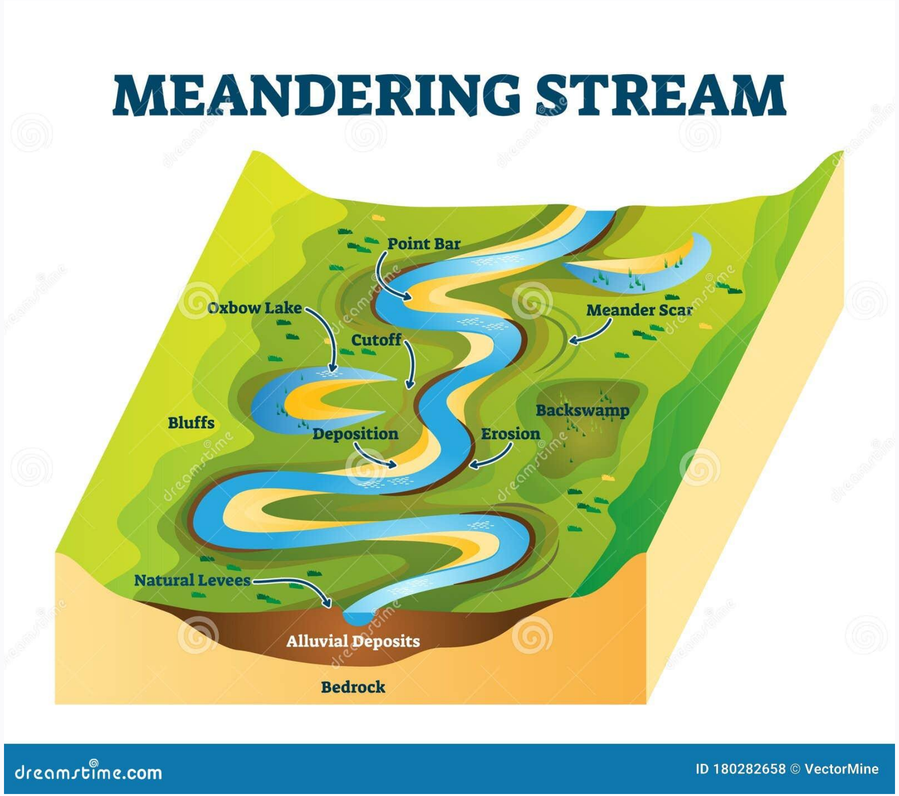
User behaviour
(e.g. signups, interactions)

Operational metrics
(e.g. cpu, memory, latency, network)

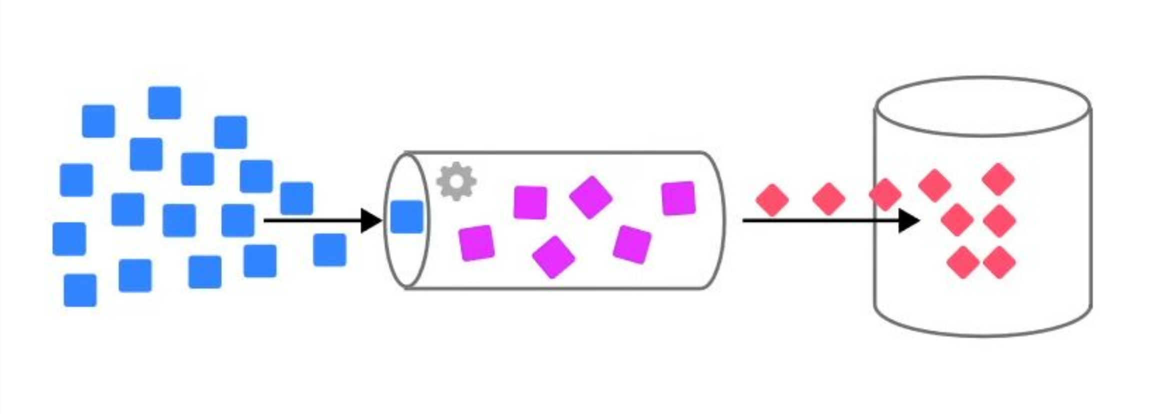

Details
Security
Networking
Storage
Status checks
Monitoring
Usage instructions
Tags
Add to dashboard
1h 3h 12h 1d 3d 1w custom
CPU utilization (%)
Percent
Status check failed (any) (count)
Status check failed (instance) (count)
Status check failed (system) (count)
Count
Network in (bytes)
Bytes
Network out (bytes)
Network packets in (count)
Network packets out (count)
Disk reads (bytes)
Disk read operations (operations)
Disk writes (bytes)
Disk write operations (operations)
No data available.
Try adjusting the dashboard time range.
CPU credit usage (count)
CPU credit balance (count)
02/08
02/10
02/12
i-05a67b314df5139de (vticloud-CloudWatch)

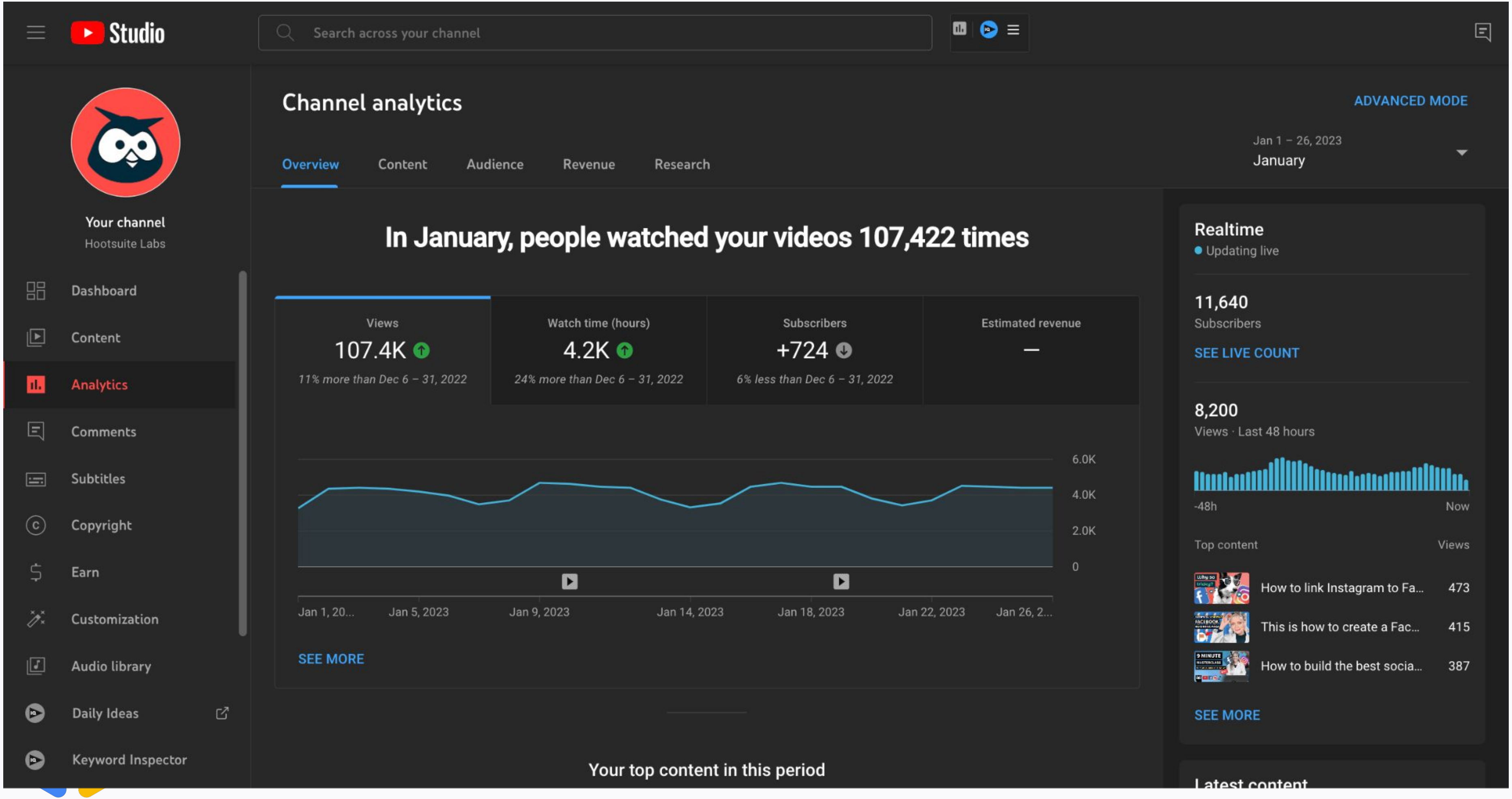

MISSION ELAPSED TIME
. 12 . 1899 00 : 00 : 26 . 742 . 375
OV104 ATLANTIS
FORMAT
129
2 1
12
PCM1
OSC2
4 . 5 . 1900 18 : 20 : 35 . 734 . 250
4 . 5 . 1900 18 : 20 : 35 . 275 . 020
ACCELEROMETER
CABIN PRESS
1491
TEMPERATURE
83
83
ROLL RATE
ALTITUDE (FT)
7336
VELOCITY (MPH)
428
FLOW RATE (GAL/MIN)
15964
TURBINE SPEED (RPM)
34012

Streaming Data Pipeline



Streaming Data Pipeline





Channel analytics

In January, people watched your videos 107,422 times

Views	Watch time (hours)	Subscribers	Estimated revenue
107.4K	4.2K	+724	–
11% more than Dec 6 – 31, 2022	*24% more than Dec 6 – 31, 2022*	*6% less than Dec 6 – 31, 2022*	

Realtime

Updating live

11,640

Subscribers

8,200

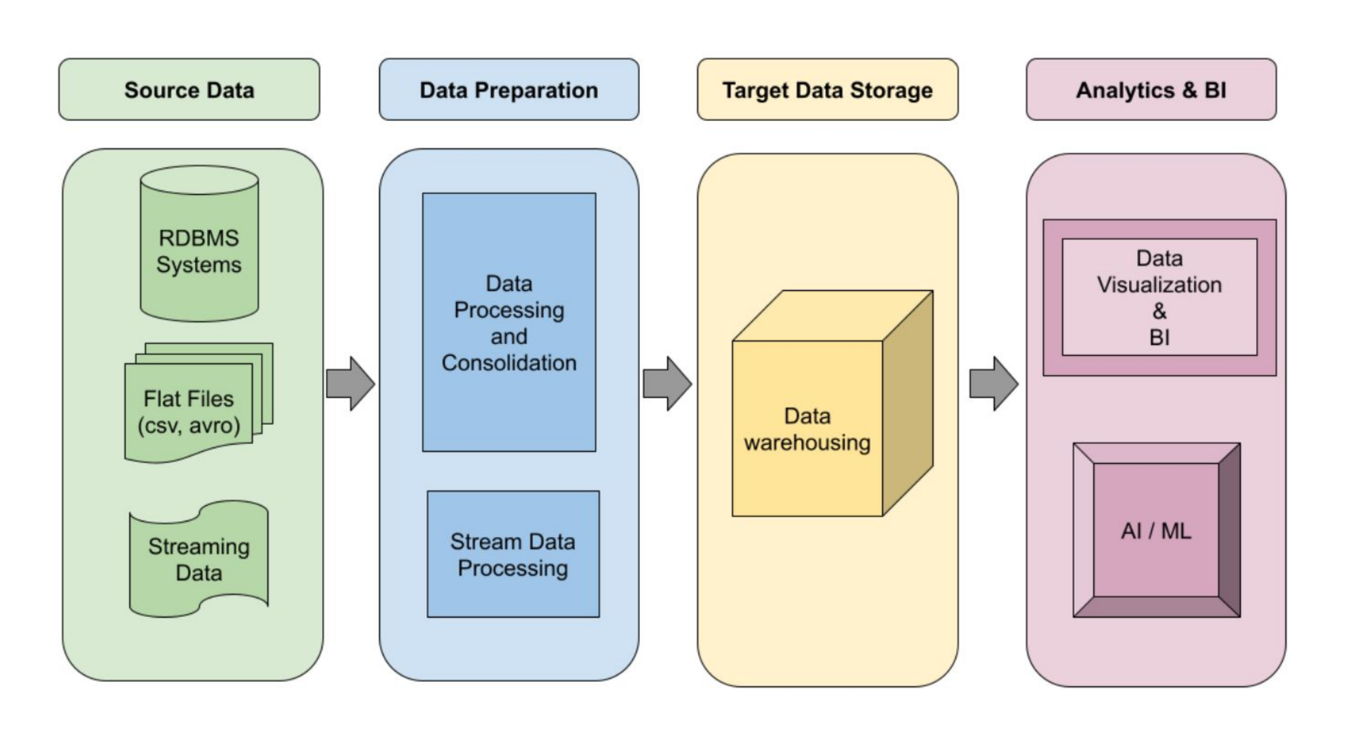
Views · Last 48 hours

Top content	Views
How to link Instagram to Fa...	473
This is how to create a Fac...	415
How to build the best socia...	387

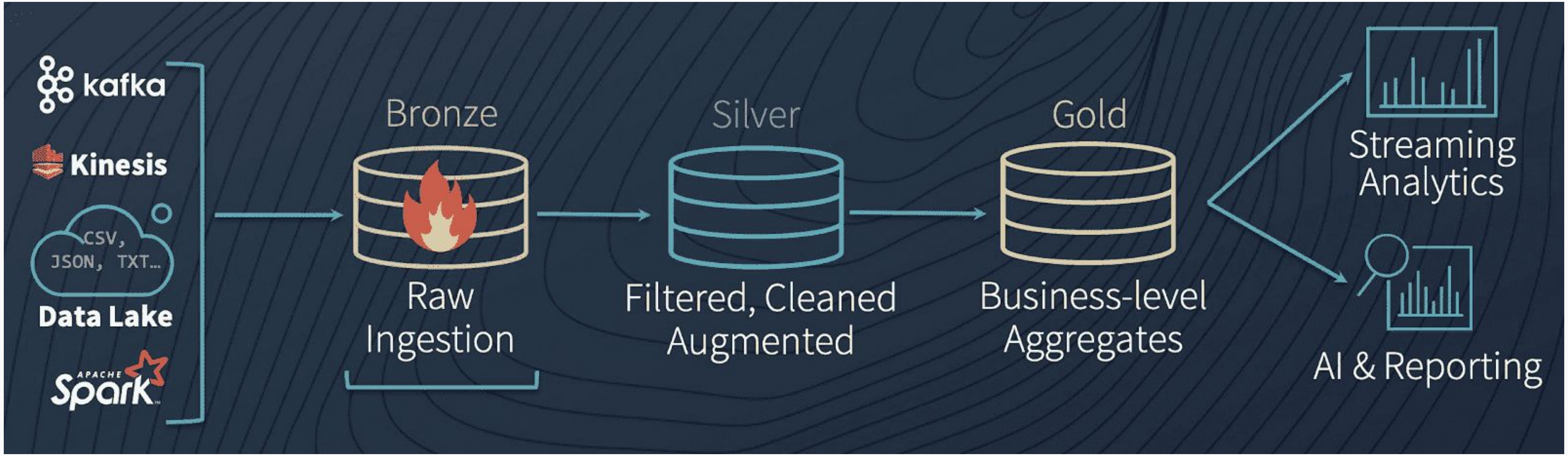
Your top content in this period

Latest content

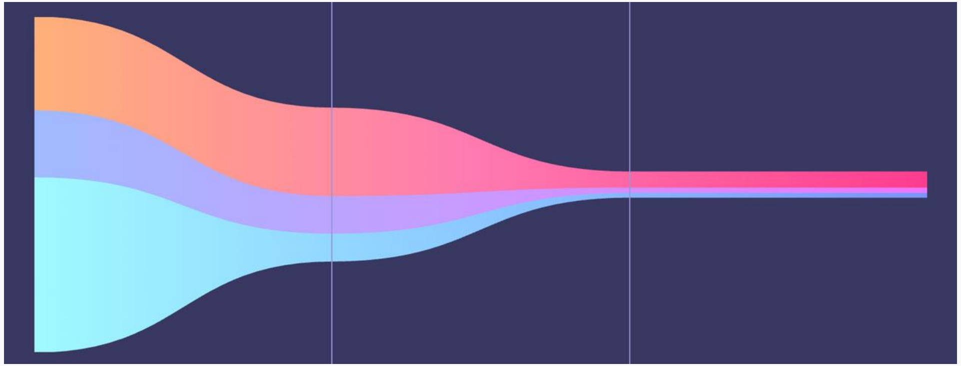
Batch Data Pipeline



Data Pipeline Architecture



The Dataflow



01. ingestion and storage

pseudonymization
anonymization
replayability
verification from source

02. cleaning and filtering

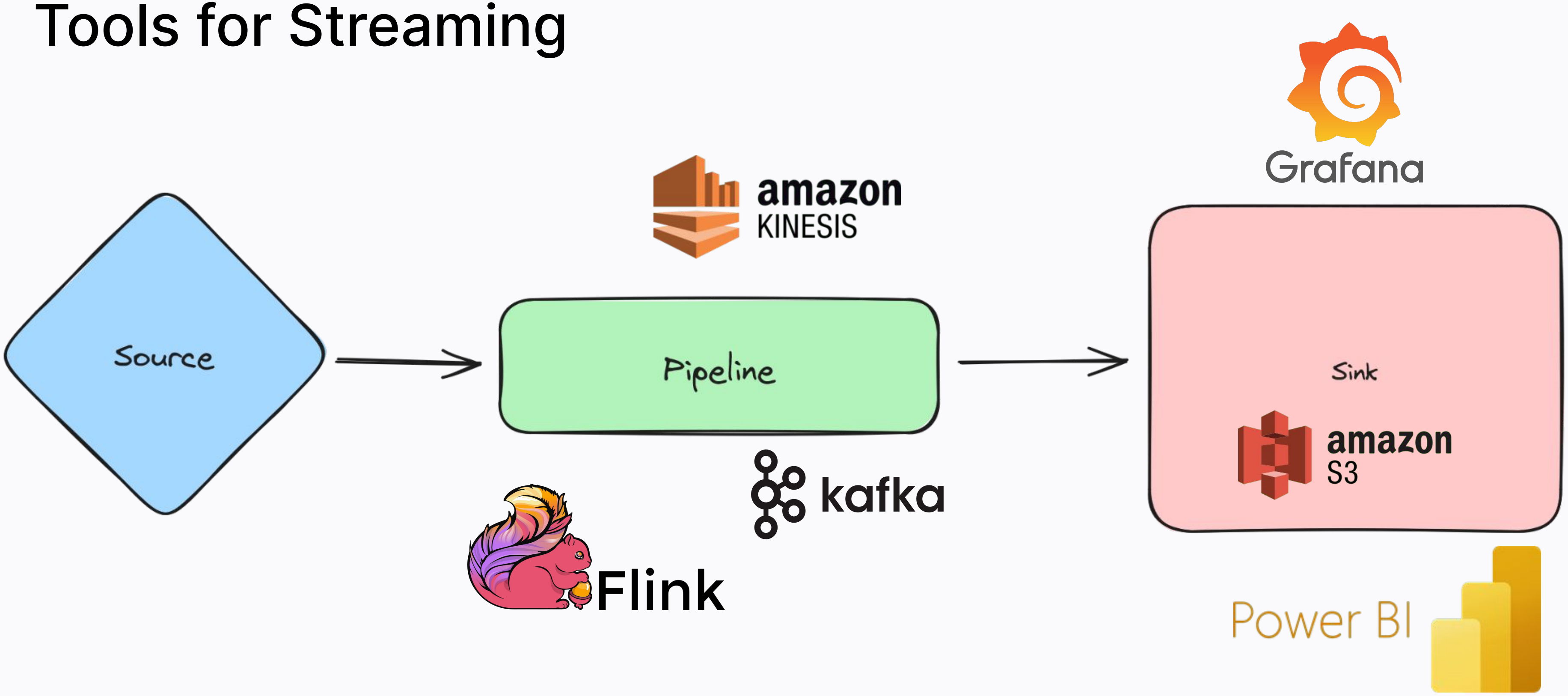
sanity checks
standardization
elimination of invalid data

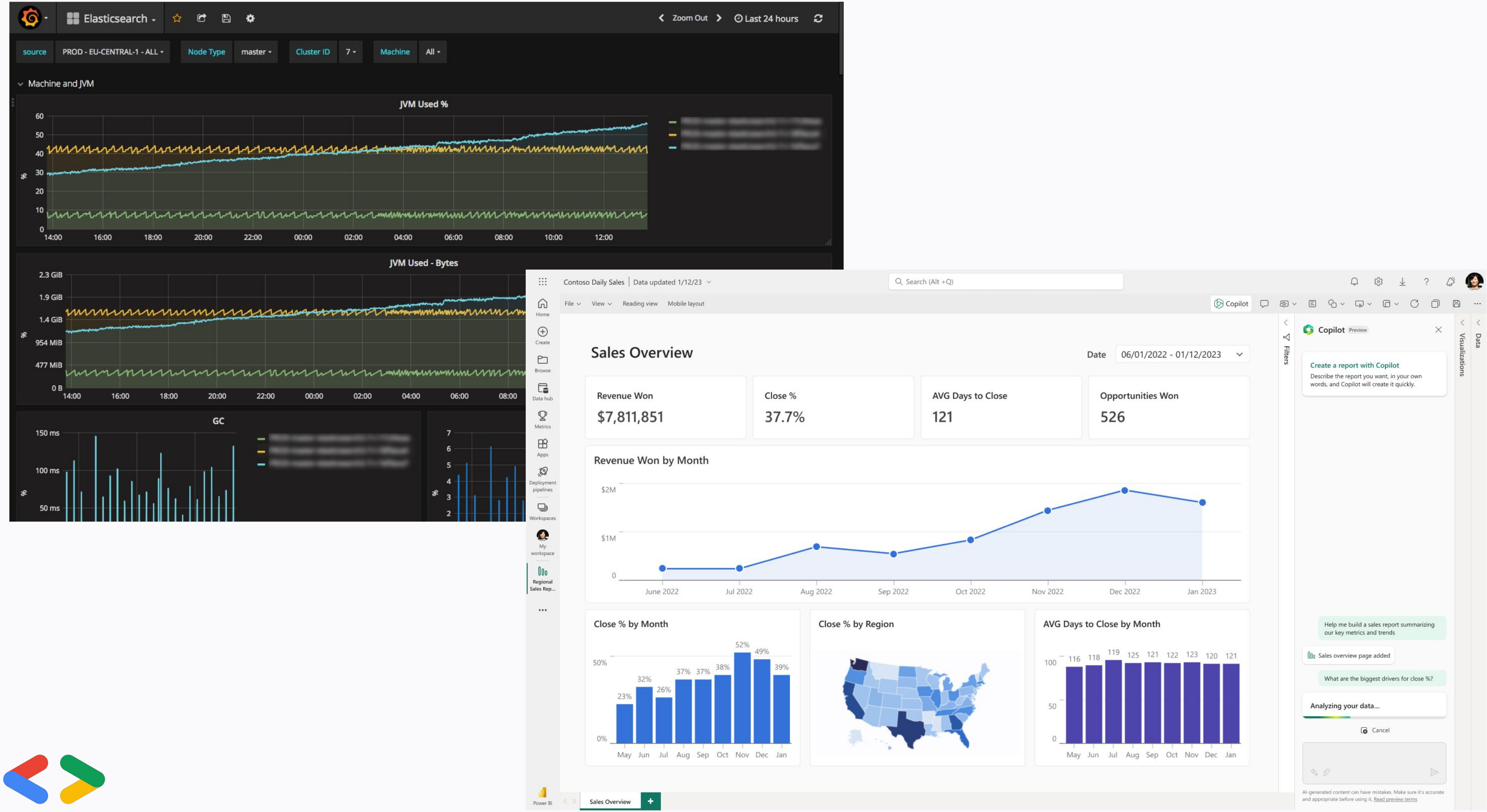
03. aggregating and analyzing

averaging
determining patterns
alerting
p50/p99
diffs across days, weeks, months, or years
graphing time series data
predictive analysis

1. the tools

Tools for Streaming





Contoso Daily Sales | Data updated 1/12/23

Sales Overview

Date 06/01/2022 - 01/12/2023

Revenue Won $7,811,851

Close % 37.7%

AVG Days to Close 121

Opportunities Won 526

Copilot Preview

Create a report with Copilot

Describe the report you want, in your own words, and Copilot will create it quickly.

Help me build a sales report summarizing our key metrics and trends

Sales overview page added

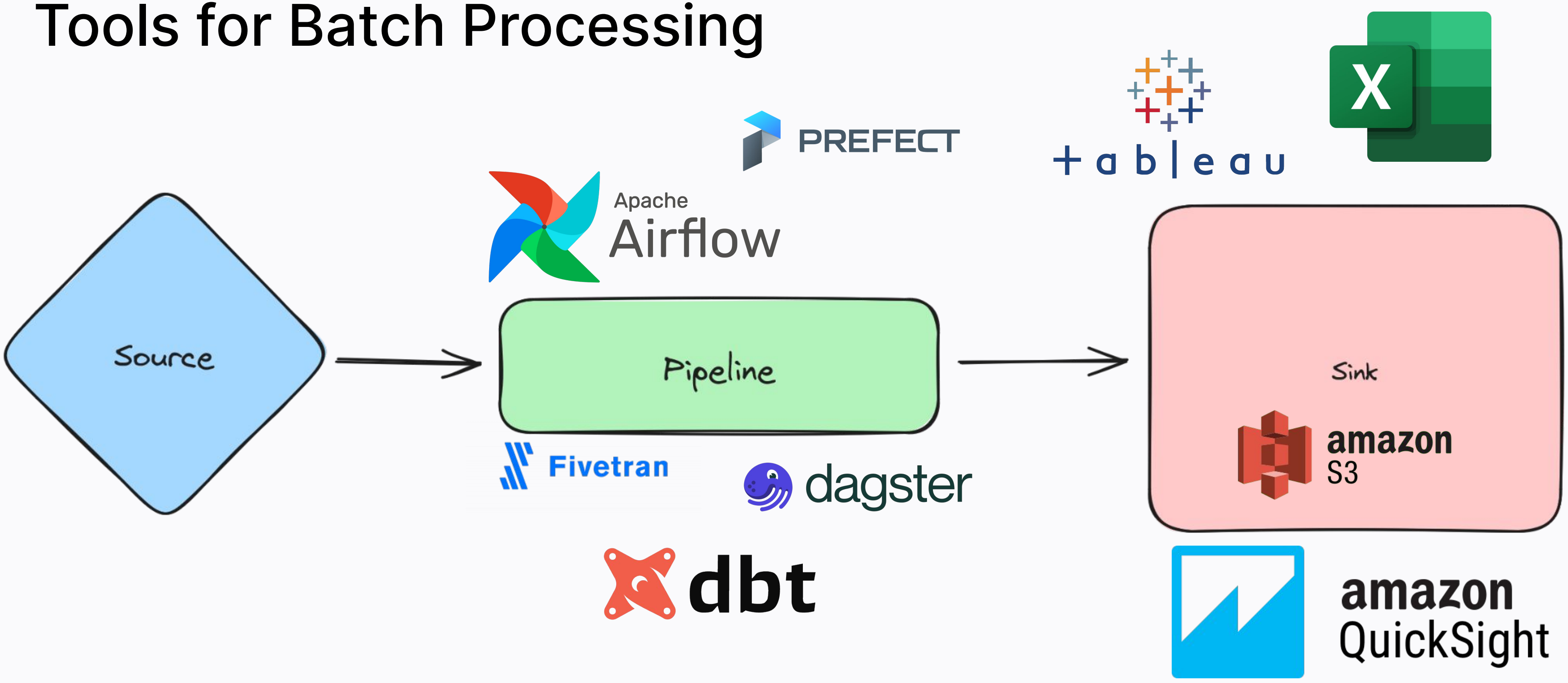
What are the biggest drivers for close %?

Analyzing your data...

Cancel

AI-generated content can have mistakes. Make sure it's accurate and appropriate before using it. Read preview terms

Tools for Batch Processing



Data Warehouses

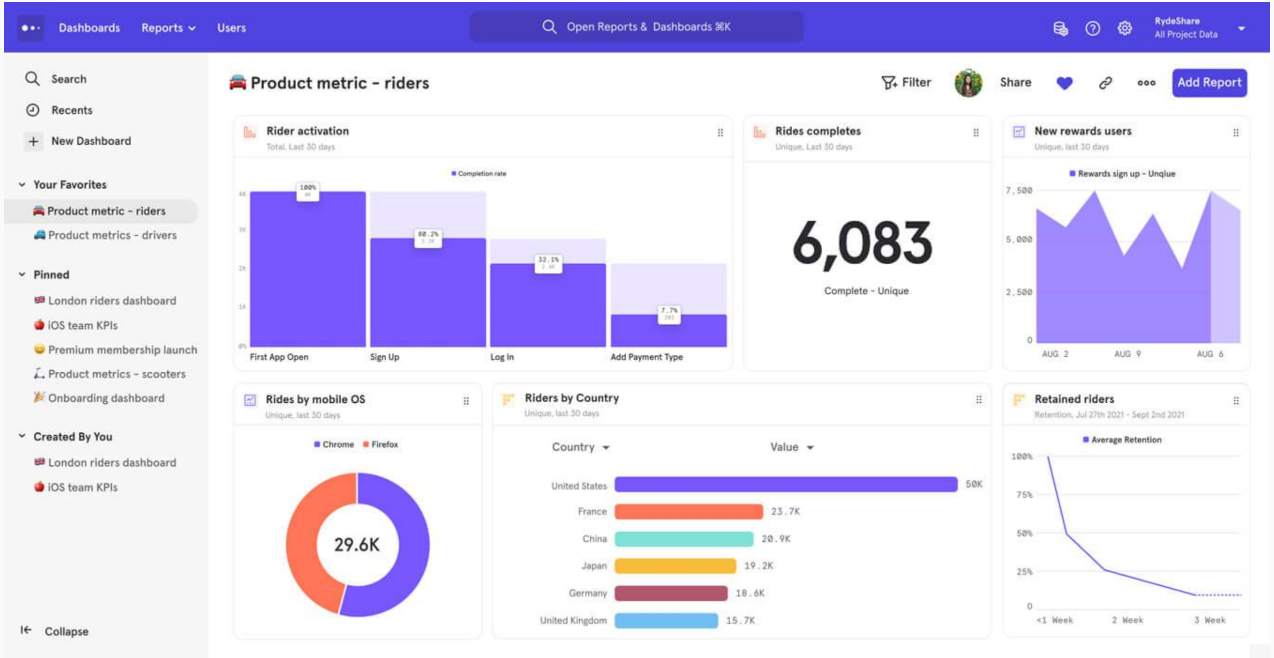
Google BigQuery

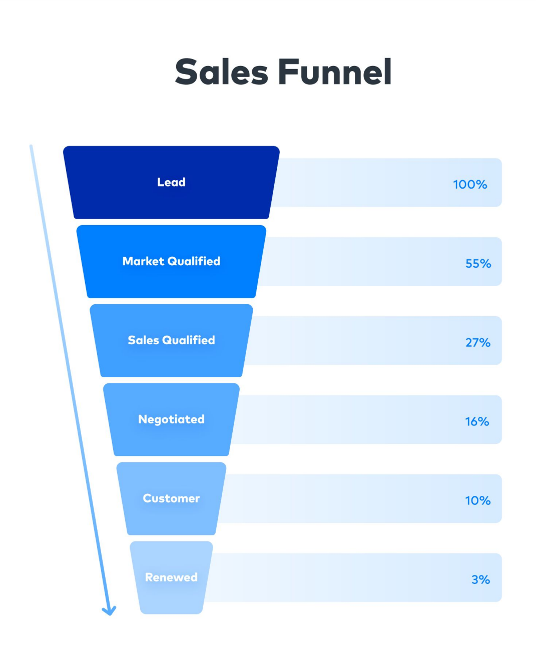
databricks

snowflake

amazon REDSHIFT

Something Else: Mixpanel





2. the **why**

Who wants this data?

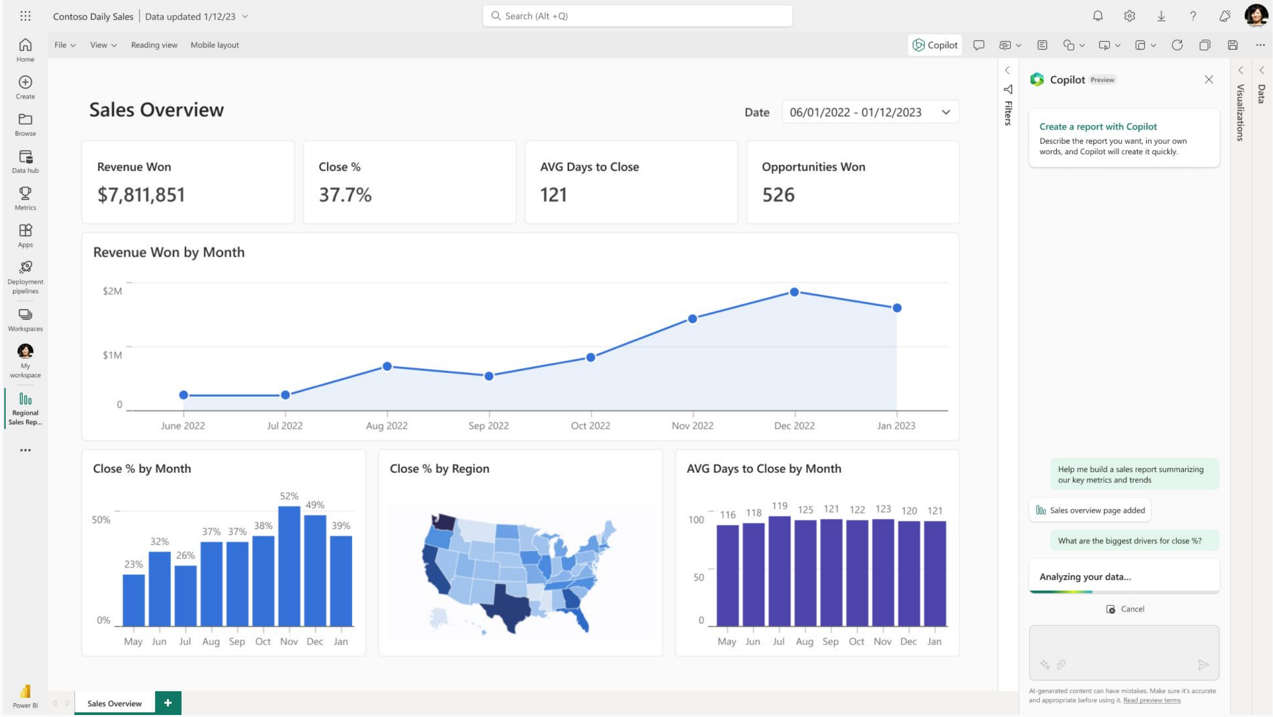
How do they want to consume it?

What decisions do they want to make with it?

What is this data intended to measure?

How do we measure what we actually want data on?

Align on expected output.



Contoso Daily Sales | Data updated 1/12/23

Sales Overview

Date 06/01/2022 - 01/12/2023

Revenue Won	Close %	AVG Days to Close	Opportunities Won
$7,811,851	37.7%	121	526

Copilot Preview

Create a report with Copilot

Describe the report you want, in your own words, and Copilot will create it quickly.

Help me build a sales report summarizing our key metrics and trends

Sales overview page added

What are the biggest drivers for close %?

Analyzing your data...

Cancel

AI-generated content can have mistakes. Make sure it's accurate and appropriate before using it. Read preview terms

What **business value** does this data provide?

Is this the best way to provide it?

What I
does t ?

Is this the best w

Data is your **Product.**

Understand the requirements.

Validate the user need.

Design an MVP.

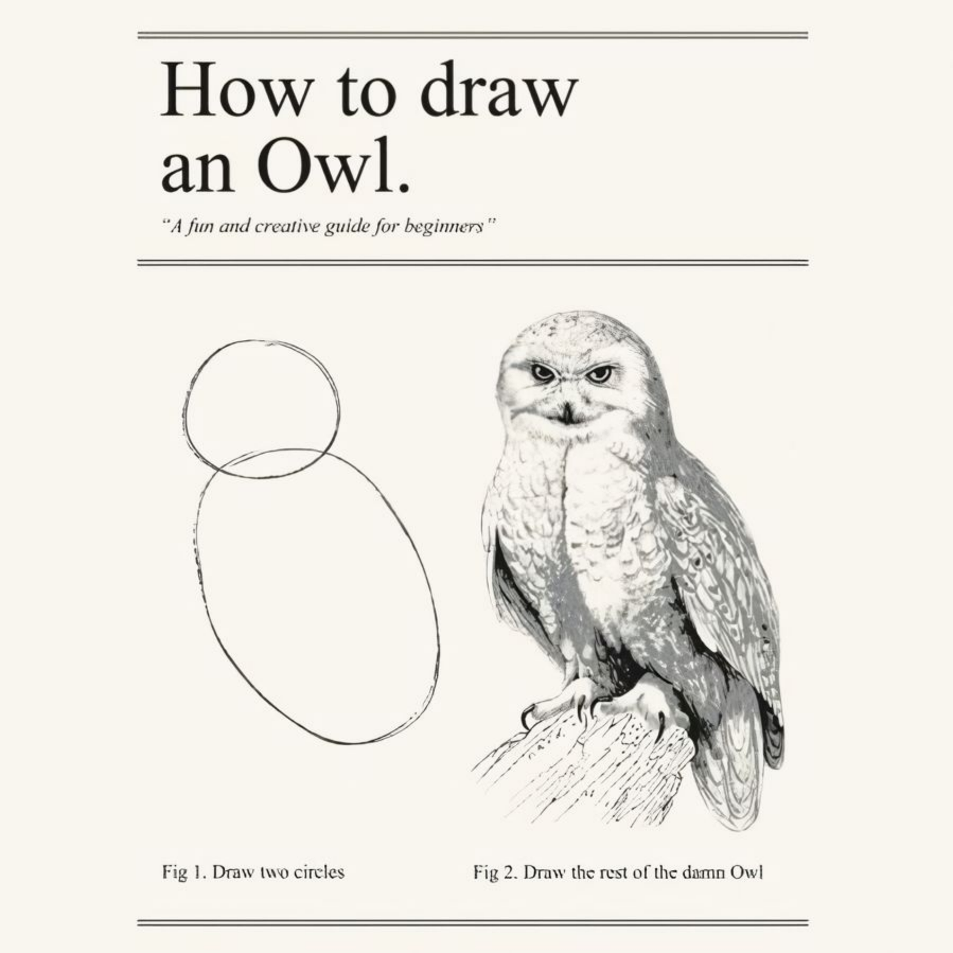
Present **clearly**.

Use graphs and charts.

Compare for context.

p50, p90, p99.

3. data org maturity



How to draw an Owl.

"A fun and creative guide for beginners"

Fig 1. Draw two circles

Fig 2. Draw the rest of the damn Owl

Follow these people

Zach Wilson

Staff Data Engineer
Prev. Facebook, Netflix, Airbnb

Benjamin Rogojan

Seattle Data Guy
Prev. Facebook

Get that data integrity

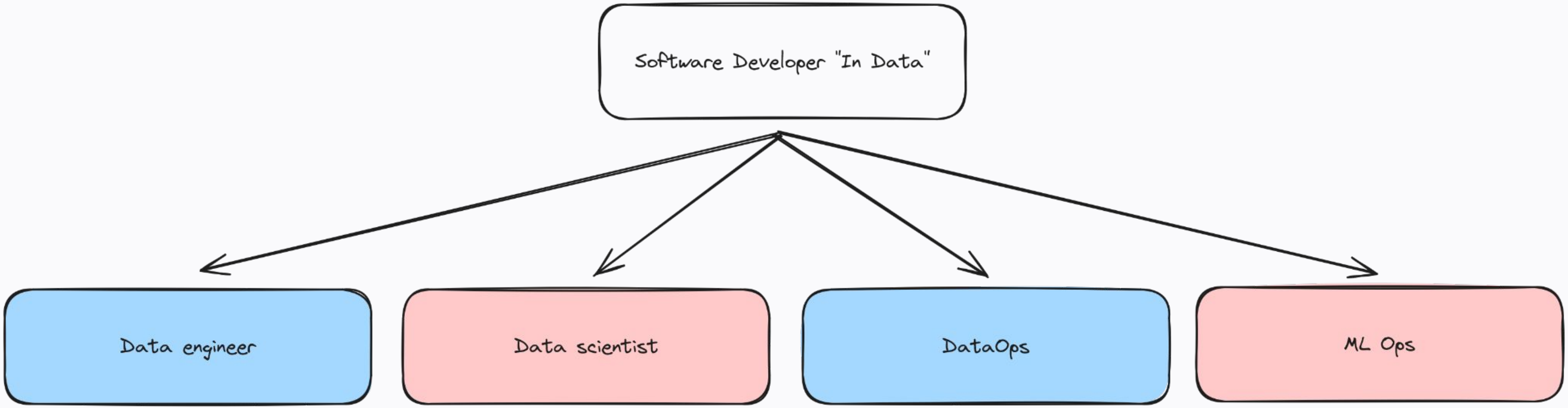
Compliance (GDPR, DSAR, COPPA, DMA) - anonymization, retention, deletion

Quality verification (e.g. Great Expectations, deequ)

Documentation (e.g. Amundsen)

Monitoring and alerting

Evolution of a data-driven organization



Software Developer "In Data"

Data engineer

Data scientist

DataOps

ML Ops

Evolution of a data-driven organization

Experimentation.

Long-range performance analytics.

Deep learning for understanding/prediction.

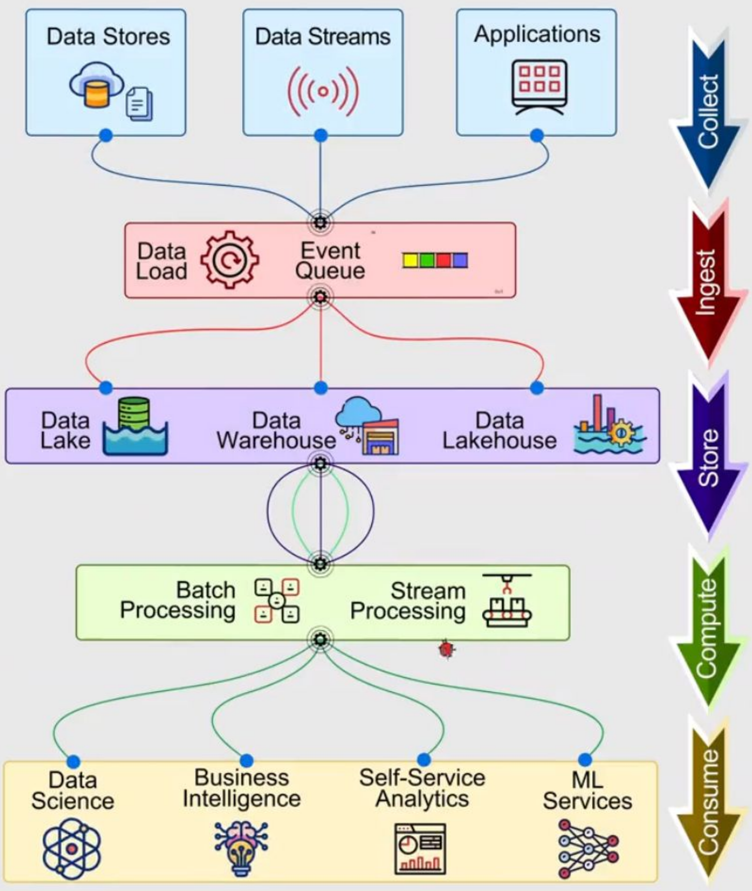
Questions?

Jeffrey Leung

GitHub: **jleung51**

Data Pipeline Overview

ByteByteGo



The Dataflow

01. ingestion and storage

retention: **short**

02. cleaning and filtering

retention: **medium**

03. aggregating and analyzing

retention: **long**