PUBLIC SECTOR SYMPOSIUM

BRUSSELS | MARCH 28, 2023

BTT203

Choose the right service for your analytics

Luka Riester (he/him)
Sr. Solutions Architect
WWPS @ AWS

Tom Zouaoui-Cusell
Analytics Engineer
@ Studocu

Agenda

Dissambiguation

Analytics the AWS way

Use Case Runthrough

Customer case: Studocu – Self-service analytics

Dissambiguation

Serverless / Analytics

Data Analytics

What

raw data -> insights

How

tools, technologies, and processes to find trends and

solve problems by using data.

Why

shape business processes, improve decision-making, and

foster organizational growth.

According to estimates, by 2030, humans will generate an ocean of data up to 572 zettabytes, which is equal to 572 million petabytes.*

How do we prepare our IT environments for the challenges of continually growing demand for infrastructure without inflating infrastructure costs and spending countless hours on maintenance?

*AWS estimate from re: Invent 2022

Serverless Computing

Do

Run code

Gather data

Integrate applications

Analytics

Don’t

Manage servers

Serverless is more than compute

COMPUTE


AWS Lambda


AWS Fargate

DATA STORES & ANALYTICS


Amazon S3


Amazon Aurora Serverless


Amazon DynamoDB


Amazon Redshift Serverless


Amazon EMR Serverless

INTEGRATION


Amazon EventBridge


AWS Glue

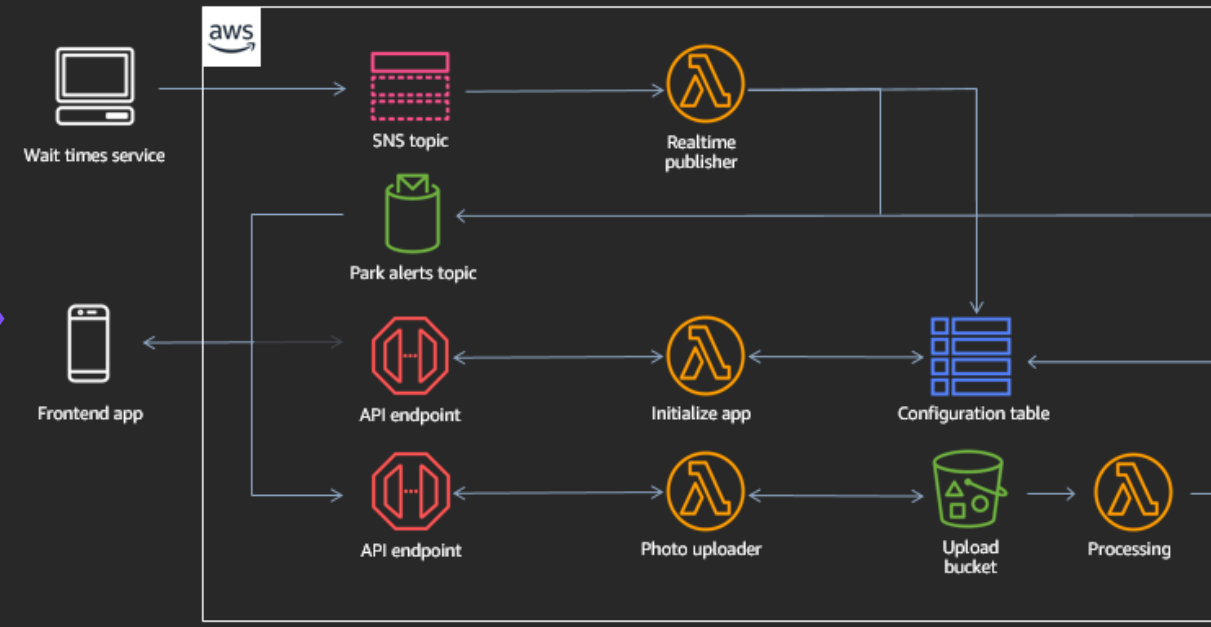

Amazon SQS


Amazon SNS


AWS Step Functions


AWS AppSync

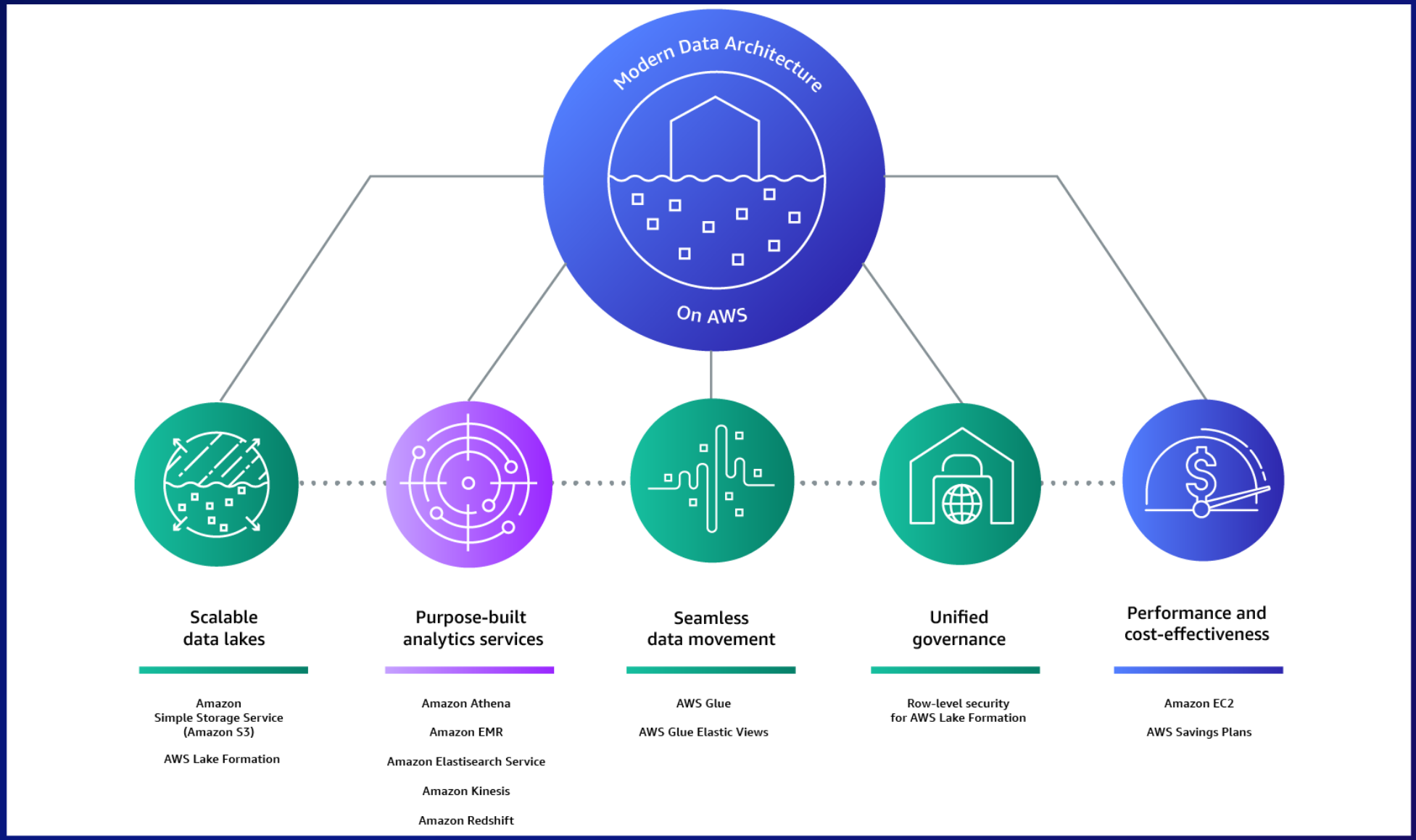
Small pieces, loosely joined



Analytics the AWS way

Analytics - How it started / how it's going

Modern Data Architecture - Tenets



Analytics Flywheel

Analytics portfolio

Data, visualization, engagement, & machine learning


Data Exchange


QuickSight


Pinpoint


SageMaker


Comprehend


Lex


Polly


Rekognition


Translate

+ many more

Analytics


Redshift


EMR (Spark & Hadoop)


AWS Glue (Spark & Python)


Athena


Elasticsearch Service


Kinesis Data Analytics

Data lake infrastructure & management


S3/Glacier


Lake Formation


AWS Glue

Data movement

Database Migration Service | Snowball | Snowmobile | Kinesis Data Streams | Kinesis Data Firehose | Managed Streaming for Apache Kafka

AWS brings ML closer to data

Databases

+

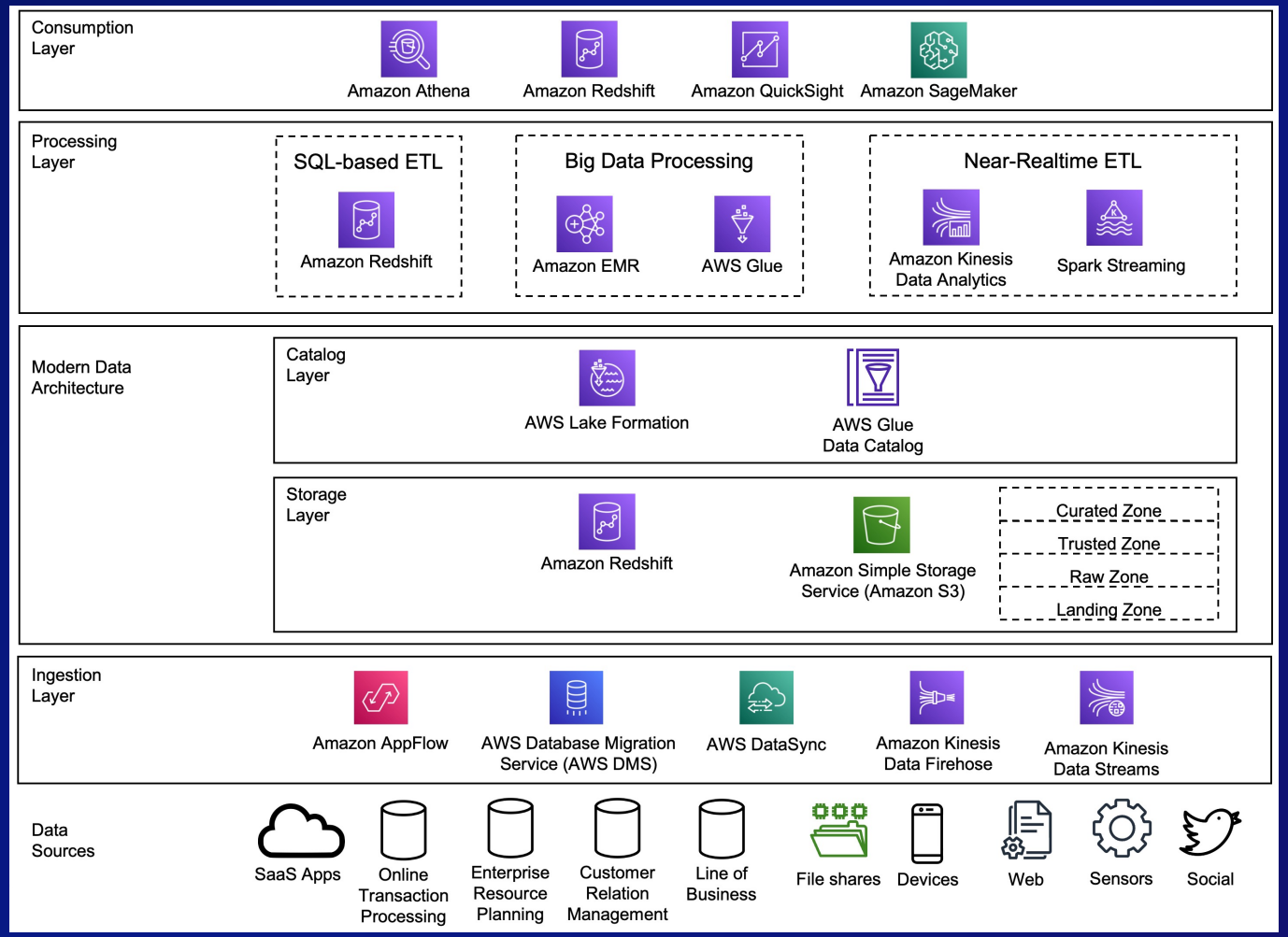
Data warehouses
+ data lakes

+

Business
intelligence tools

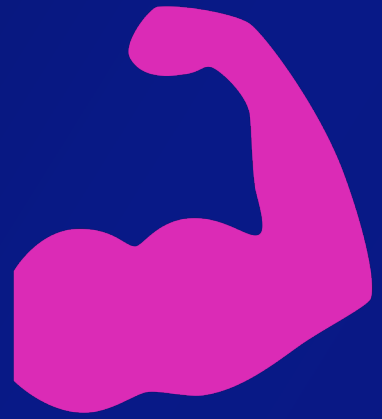
AMAZON AURORA ML	AMAZON NEPTUNE ML	AMAZON REDSHIFT ML	AMAZON ATHENA ML	AMAZON QUICKSIGHT ML
				

AWS Reference Architecture (Analytics Lens)



https://docs.aws.amazon.com/wellarchitected/latest/analytics-lens/reference-architecture.html

Focus on analytics and results, not infra



Flexible scaling, built-in availability and fault tolerance

AWS has the most serverless data analytics services in the cloud → remove your operational overhead

Analytics Usecase – built in AWS

Top-down approach

Concrete Analytics Case

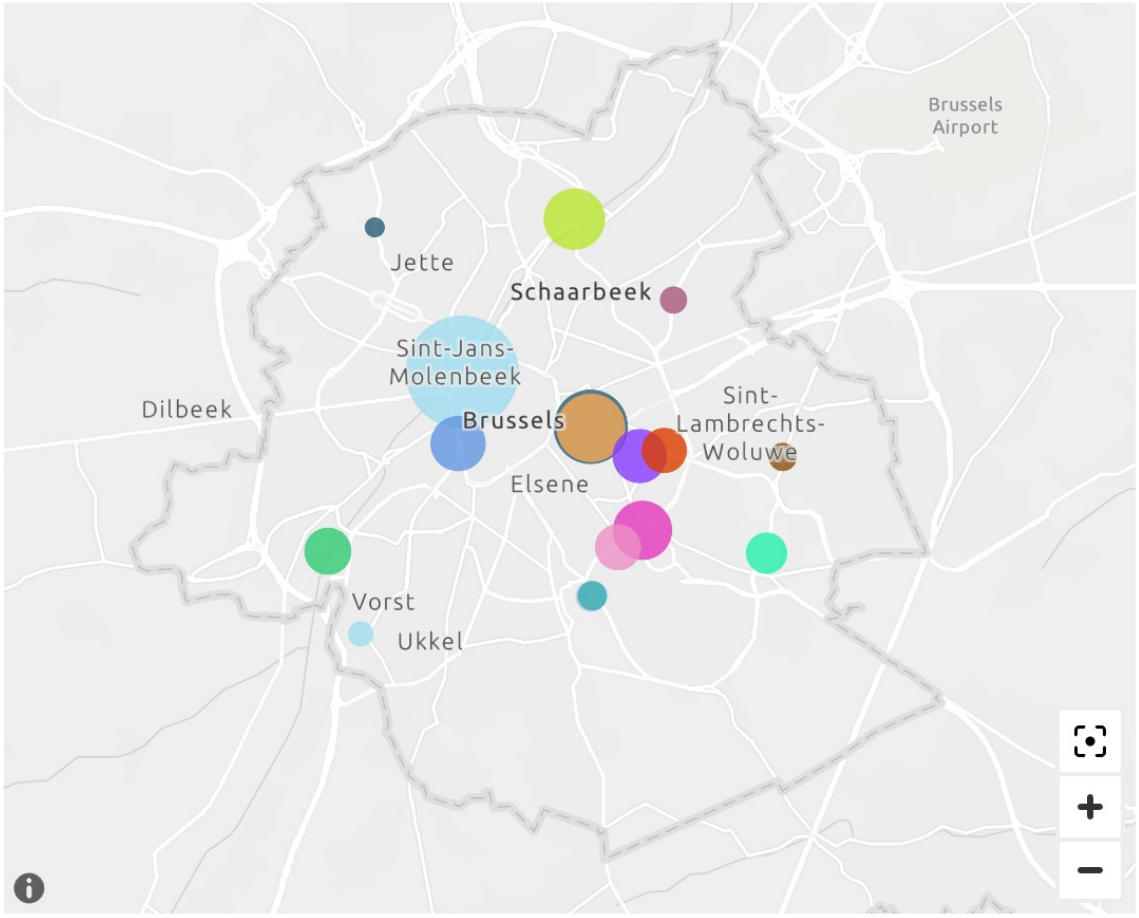
“Le nombre de cyclistes a augmenté de 20% l’an dernier à Bruxelles“

From RTBF.be

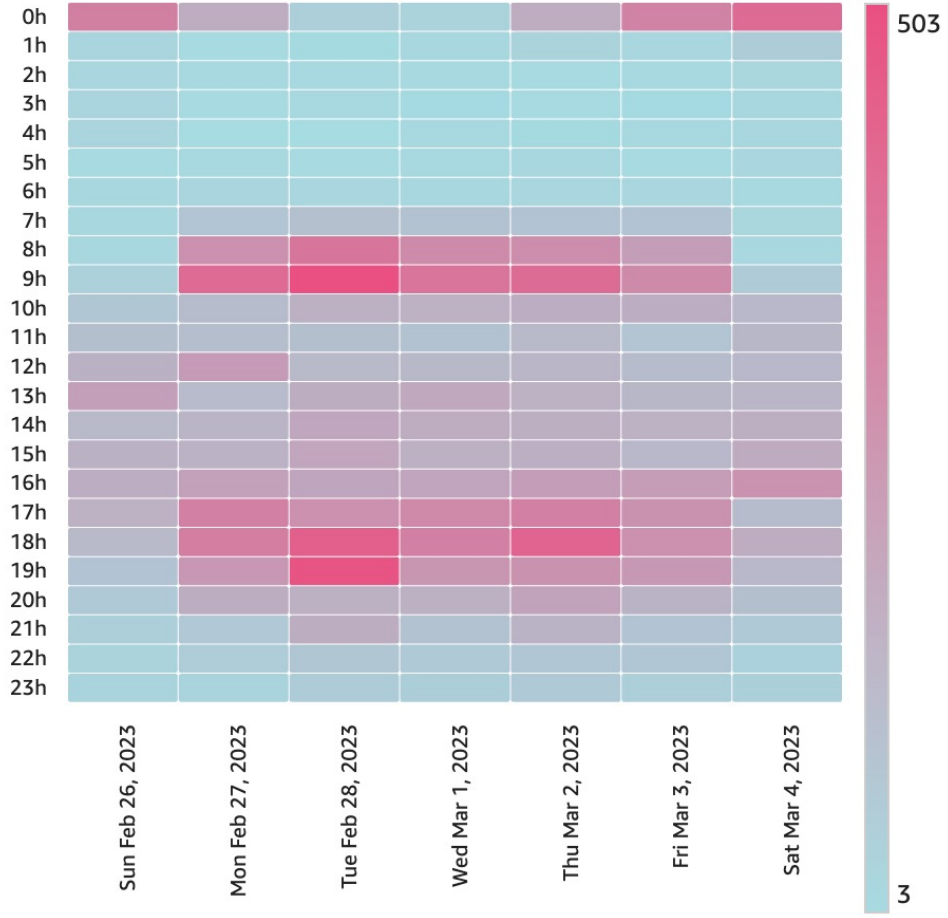
When is the good time to go and leave to/from work?

Bike usage in my city – where & when

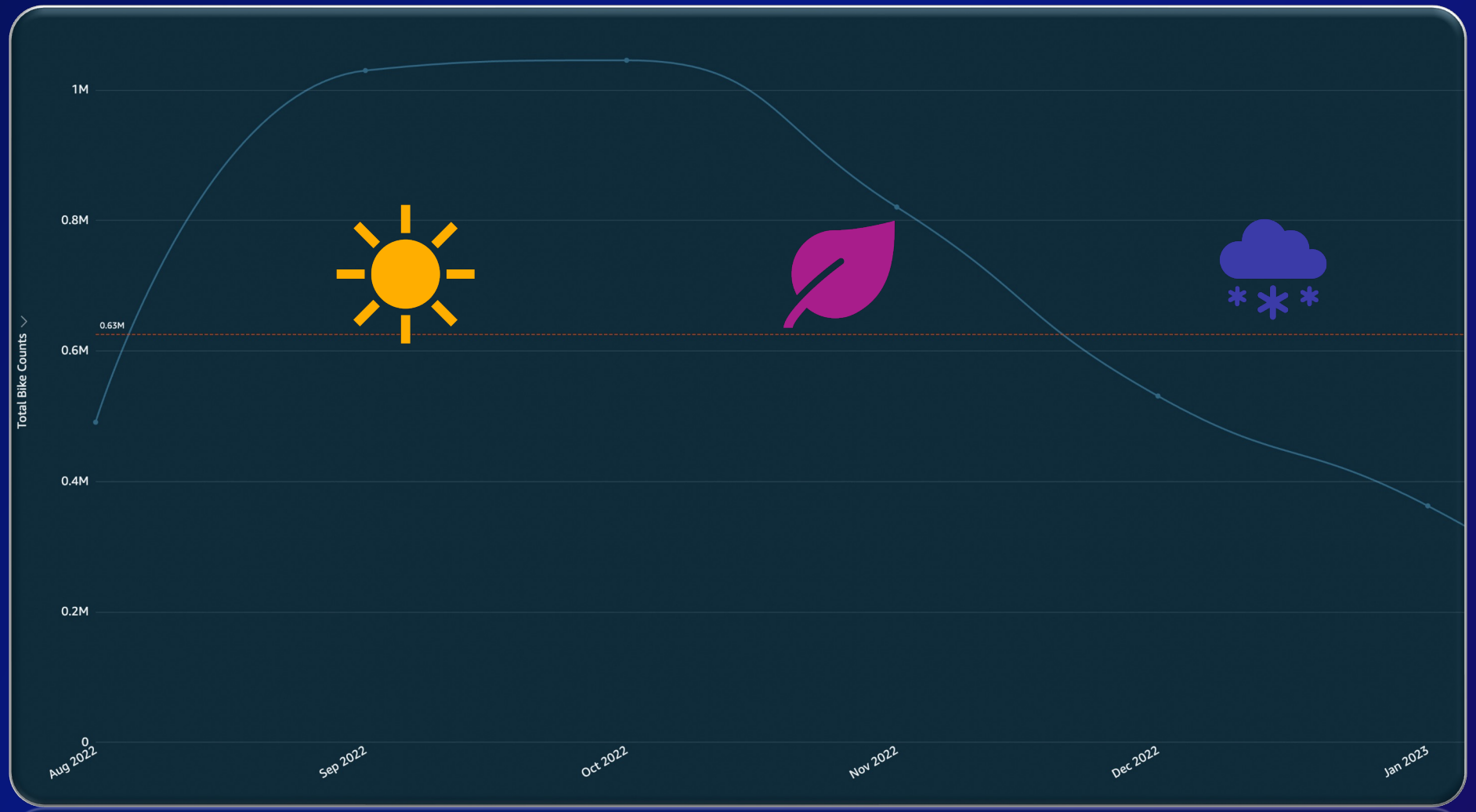
Map of Counting Stations with Bike Counts



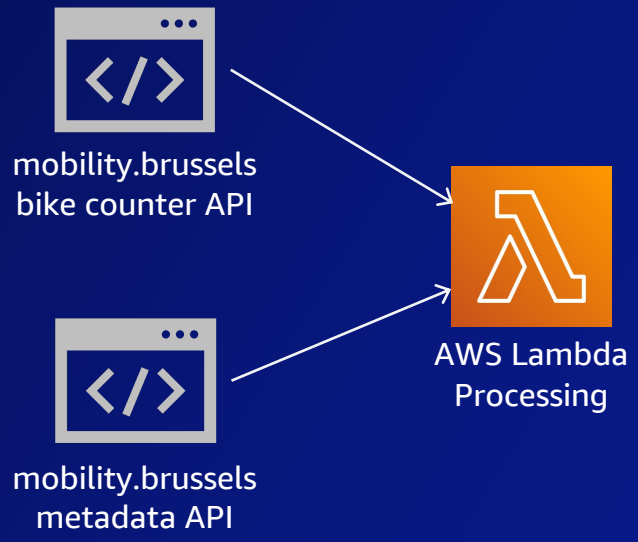
Weekly heatmap of bikes per hour of day



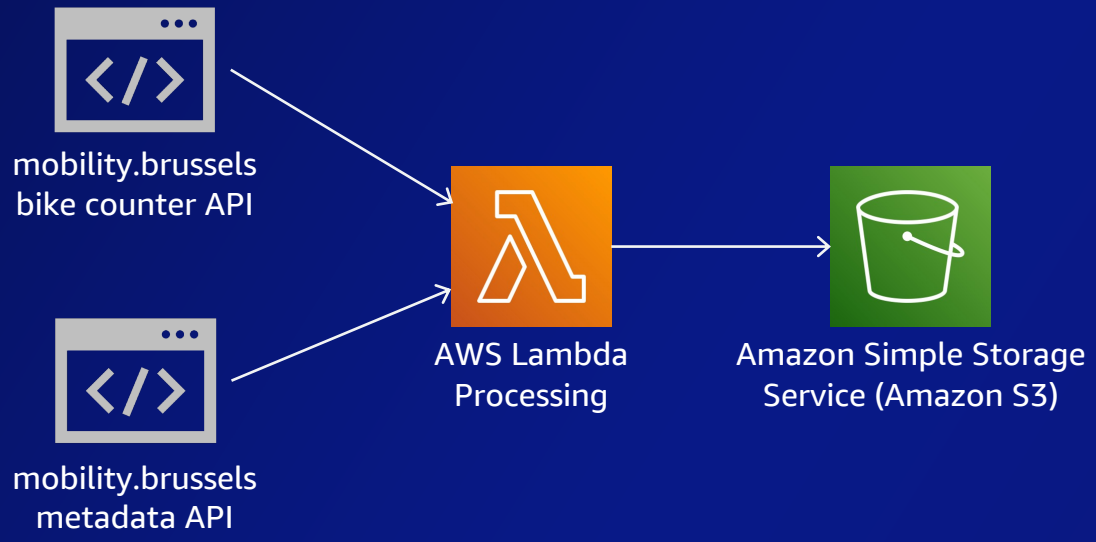
Monthly bike traffic evolution



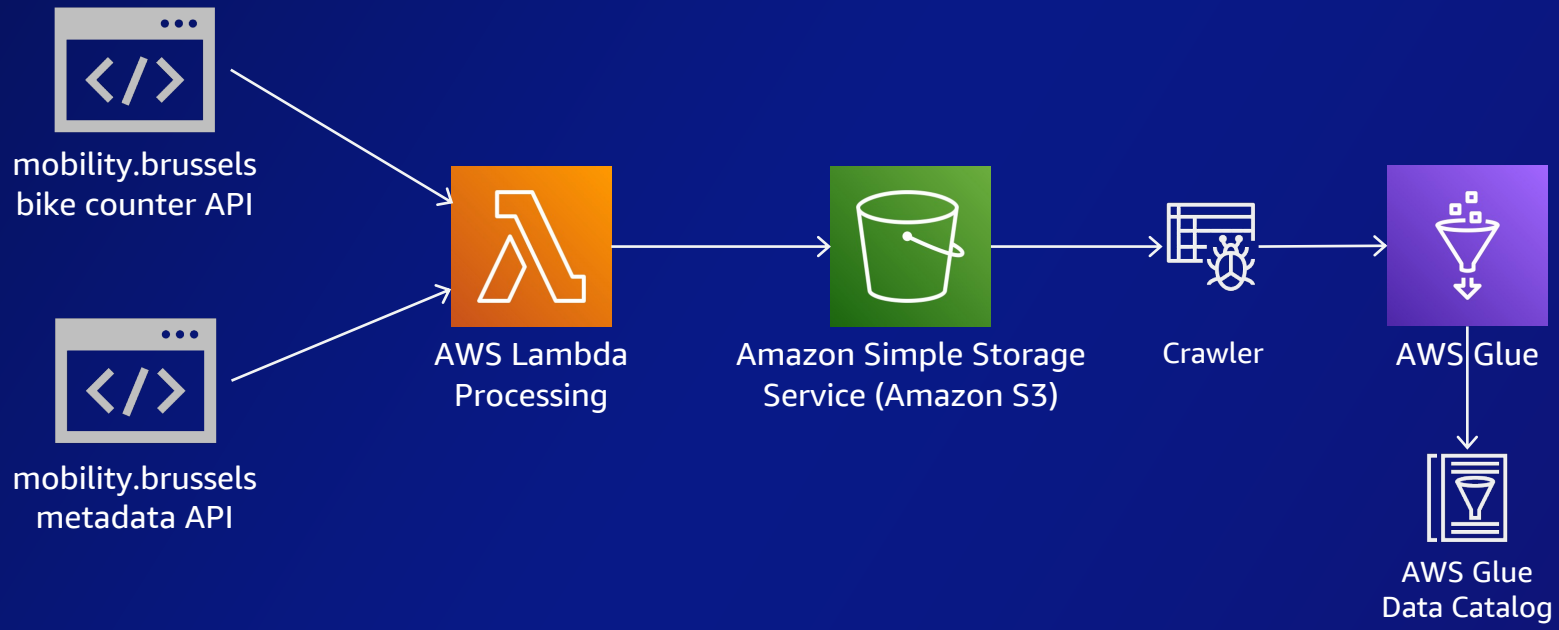
Serverless Analytics – Get



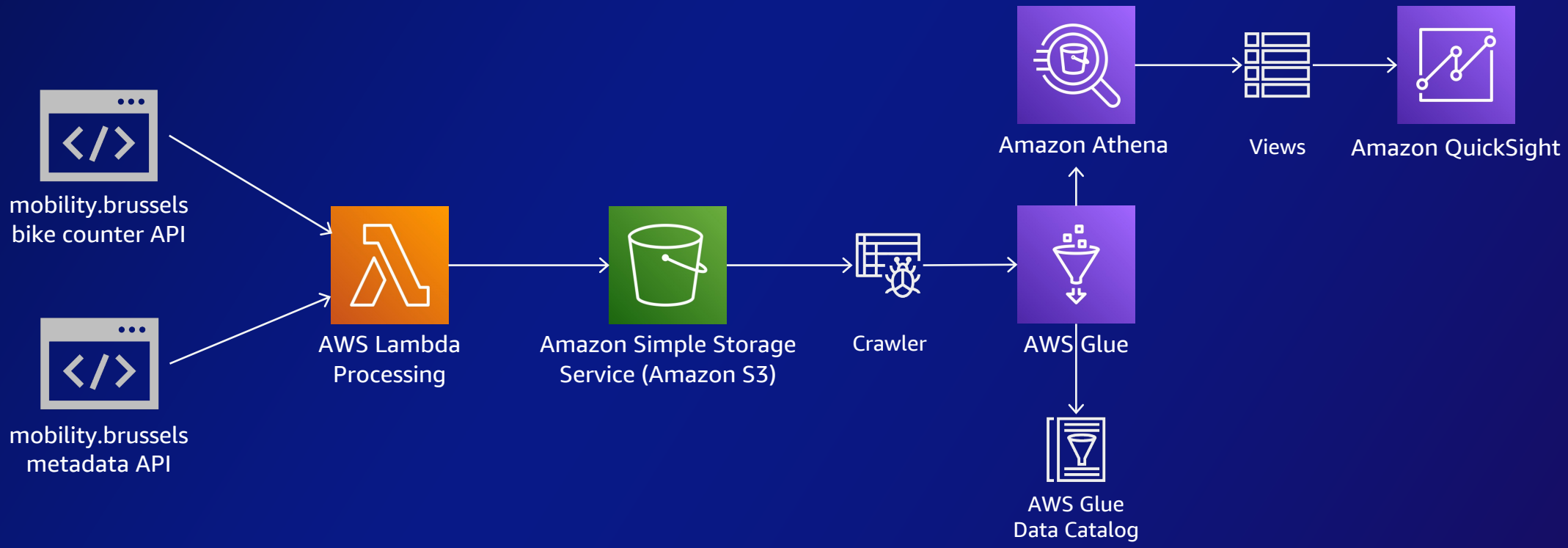
Serverless Analytics – Store



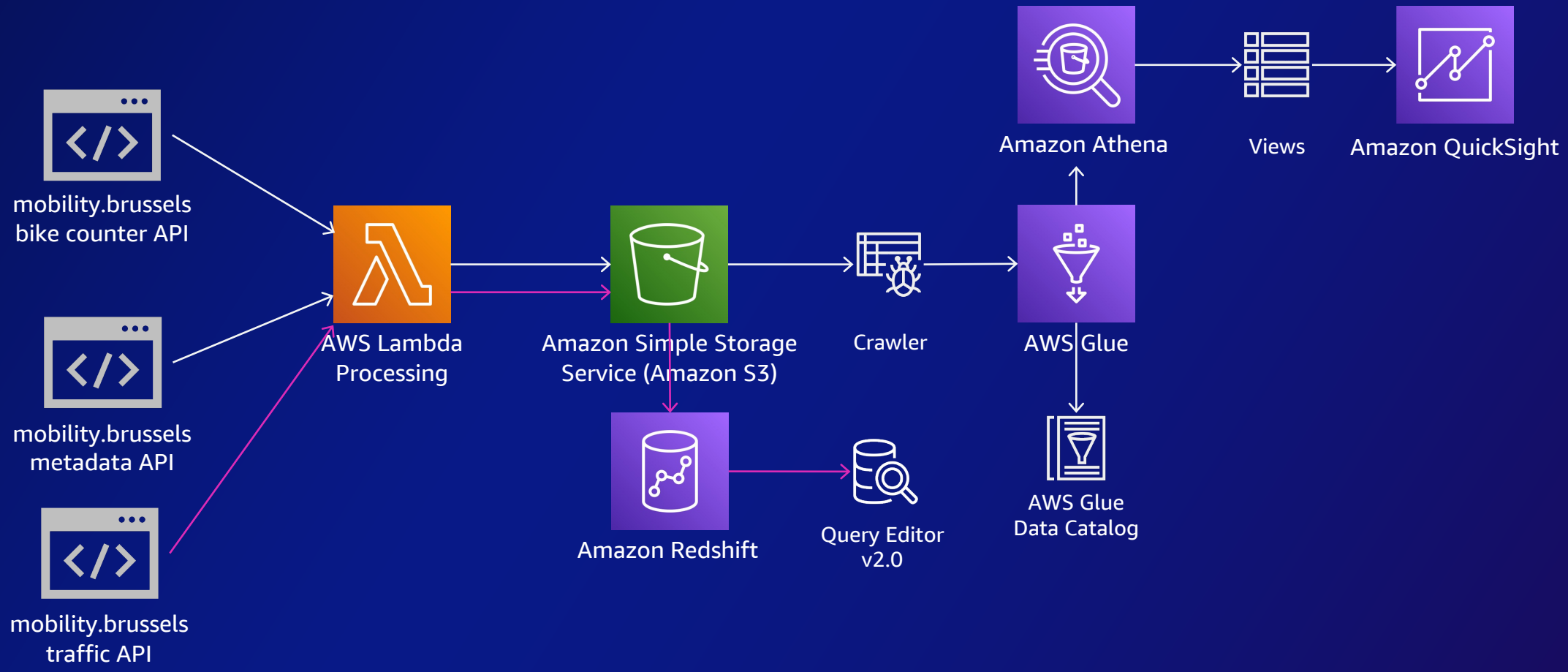
Serverless Analytics – Catalog



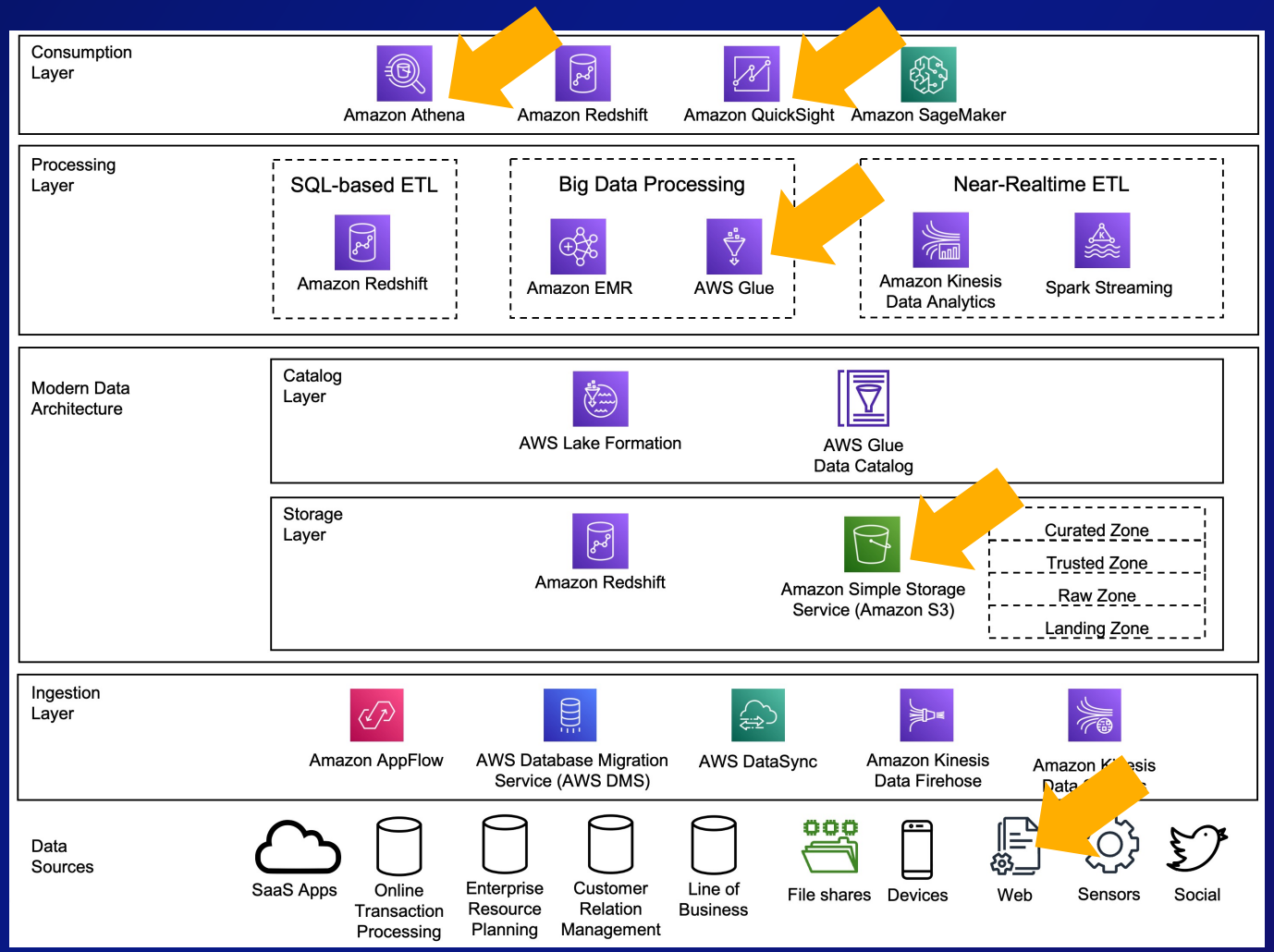
Serverless Analytics – Analyse



Serverless Analytics – Extend

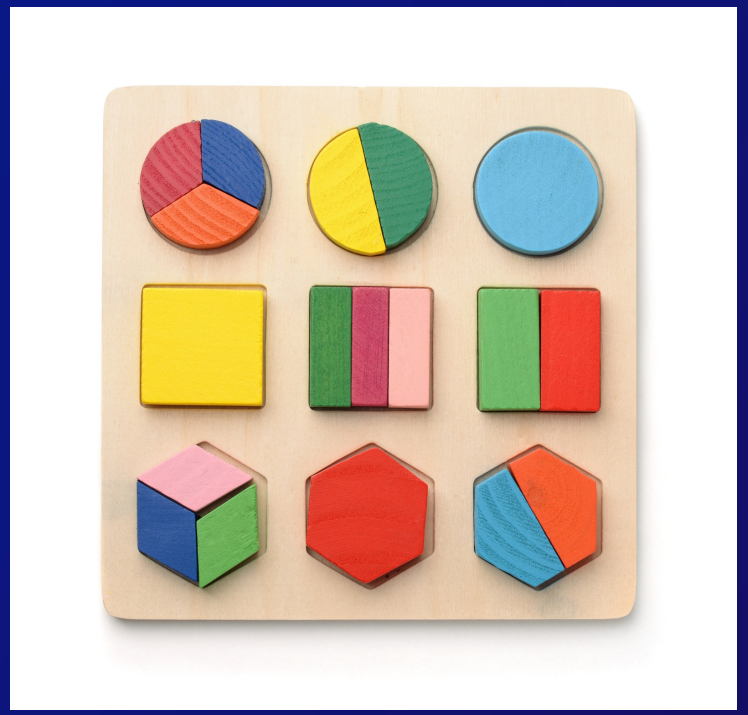


AWS Reference Architecture (Analytics Lens)



https://docs.aws.amazon.com/wellarchitected/latest/analytics-lens/reference-architecture.html

The Modern Analytics Approach



Self-service analytics

Tom Zouaoui-Cusell
Analytics Engineer
@ Studocu

Our values and mission

Empower everyone to excel at their studies by providing the best tools to study more efficiently.

Our formula

EFFICIENCY + SCALABILITY + DATA = **SUCCESS**

Our numbers

28 000
universities

24M
documents

30M
students

Time-to-insights

Autonomous data-driven teams

Faster iterations (A/B tests)

Proactive instead of reactive

The goal of our data team

Enable everyone at the company to answer their own questions without bottlenecks

Our data stack

Almost serverless infrastructure

studocu data sources

1. Get data from different sources

1

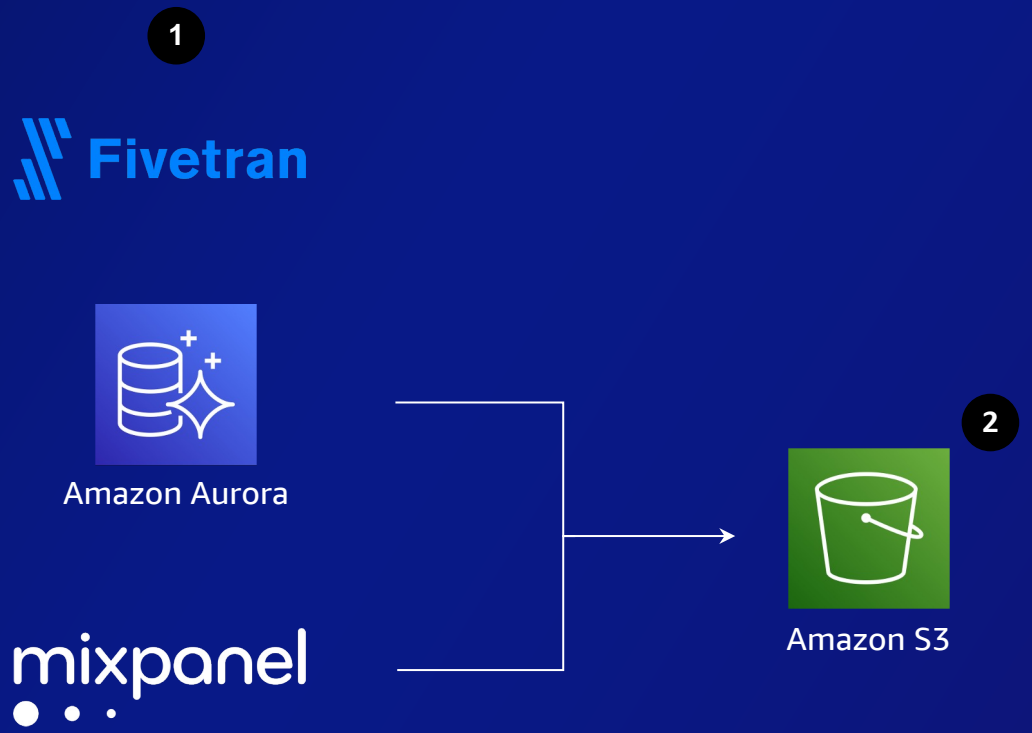
Fivetran

Amazon Aurora

mixpanel

studocu data lake

1. Get data from different sources
2. Store raw data in our data lake



studocu data catalog

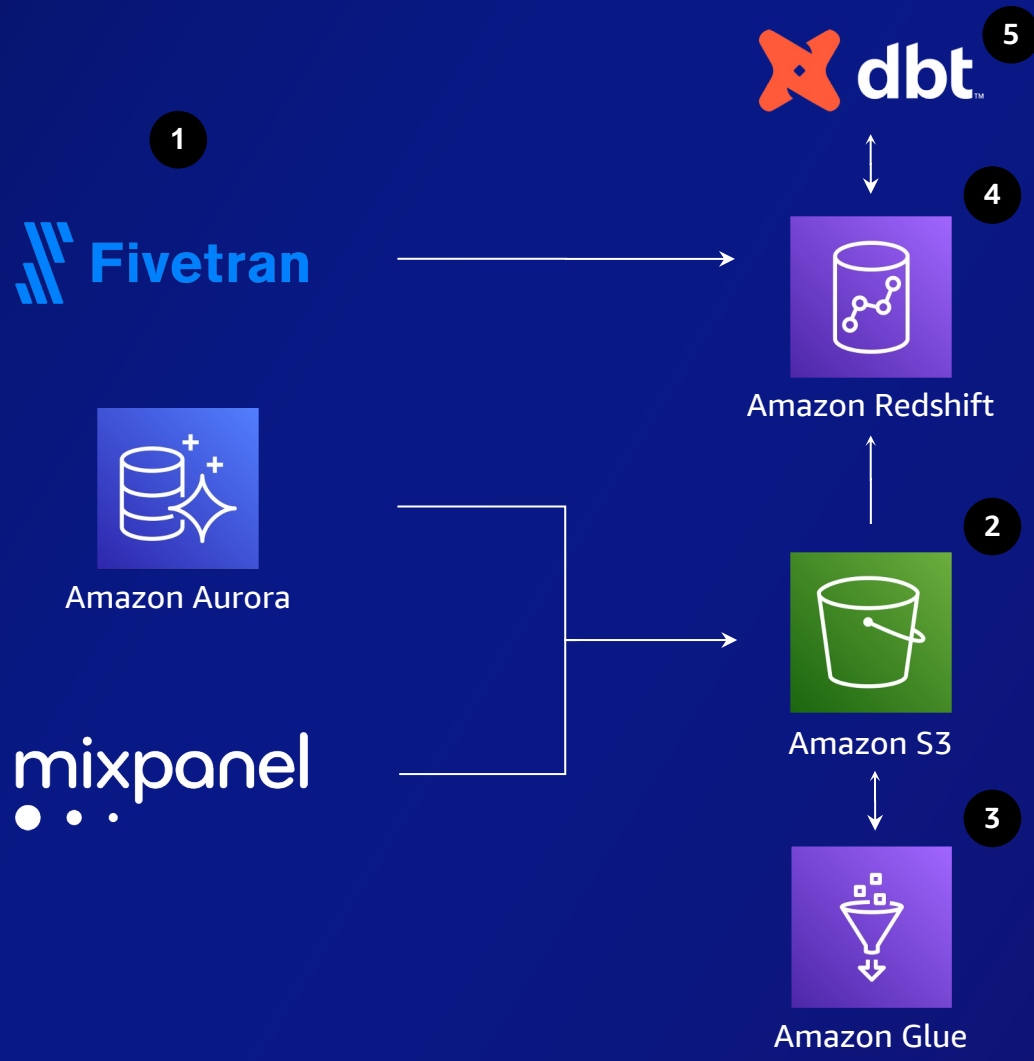
1. Get data from different sources
2. Store raw data in our data lake
3. Extract the schema of our raw data

studocu data warehouse

1. Get data from different sources
2. Store raw data in our data lake
3. Extract the schema of our raw data
4. Push our structured data to our data warehouse

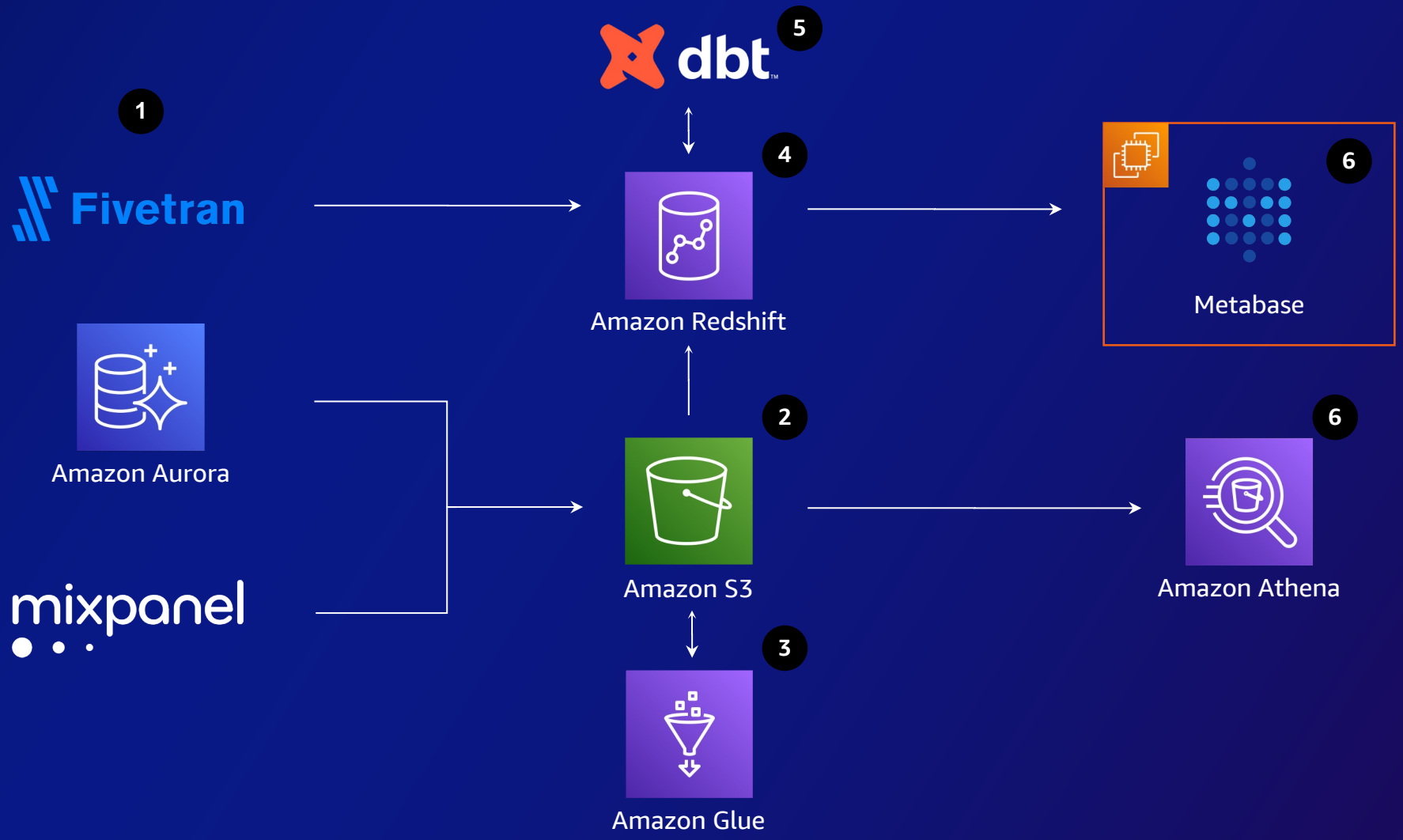
studocu transformation layer

1. Get data from different sources
2. Store raw data in our data lake
3. Extract the schema of our raw data
4. Push our structured data to our data warehouse
5. Apply transformations and business logic with **dbt**



studocu self-service analytics

1. Get data from different sources
2. Store raw data in our data lake
3. Extract the schema of our raw data
4. Push our structured data to our data warehouse
5. Apply transformations and business logic with **dbt**
6. Query data with or without **SQL knowledge**



Next steps

AI-generated documents

Better recommendations

Forecasts

Thank you!

Please complete the session survey in the mobile app

Luka Riester

Sr. Solutions Architect

WWPS @ AWS

Tom Zouaoui-Cusell

Analytics Engineer

@ Studocu