Data Pipelines with AWS Glue (Level 200)

Unni Pillai, Specialist Solution Architect
Thanomsak Ajjanapanya, Data Engineering Manager, Toyota Tsusho – Thailand

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

In this session…

Introduction to AWS Glue

Understand AWS Glue Components

Construct an ETL Flow

Demo – Build a data pipeline using Glue in 4 steps

Learn how  NEXTY Electronics is using AWS Glue

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

What is AWS Glue?

Fully-managed, serverless extract-transform-load (ETL) service **for developers**, built **by developers**

1000s of Developers and jobs

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

There are many tools already in AWS Ecosystem

Amazon Redshift Partner Page for Data Integration


Fivetran

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Still ETL Developers Hand-Code

- Canvas based tools are hard to extend
- Code is flexible, powerful, and easy to share
- Familiar tools and development pipelines
  - IDEs, version control, testing, continuous integration
- Highly customizable tasks can be achieved

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Hand-coding is laborious

schemas change

data formats change

add or change sources

data volume grows

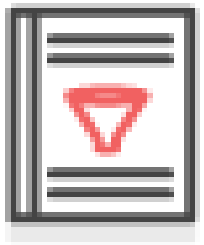
makes hand-coding error-prone & brittle

AWS Glue does the undifferentiated heavy lifting

so developers can easily customize

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AWS Glue Components



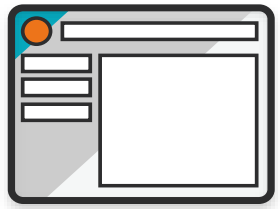
Data Catalog

Discover

Automatic crawling

Apache Hive Metastore compatible

Integrated with AWS analytic services



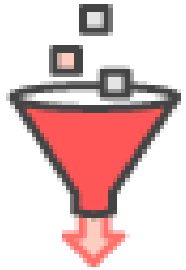
Job Authoring

Develop

Auto-generates ETL code

Python and Apache Spark

Edit, Debug, and Explore



Job Execution

Deploy

Serverless execution

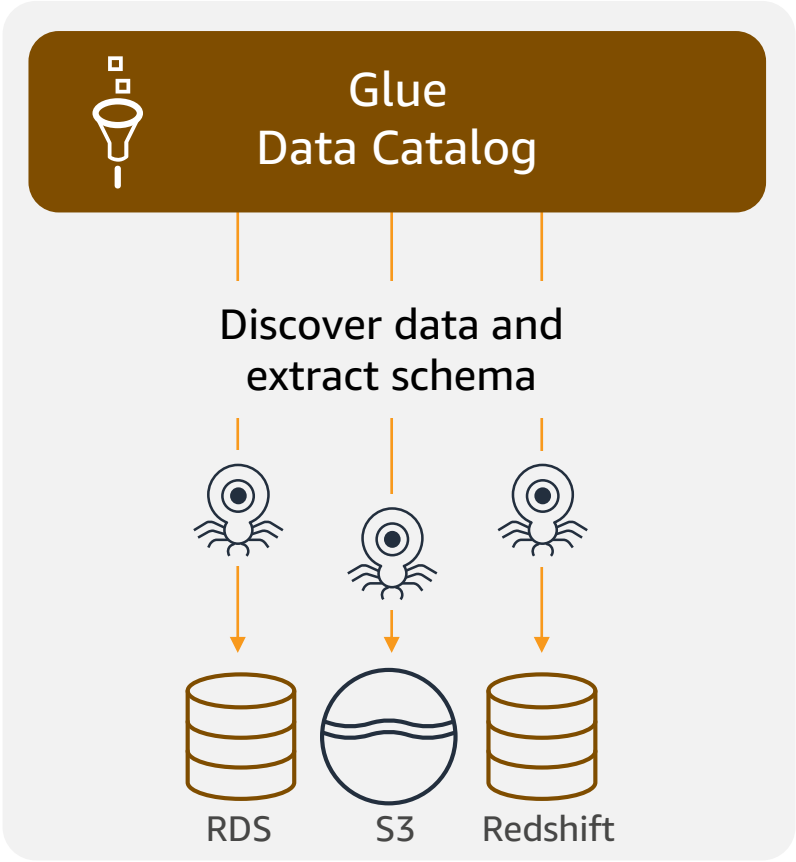
Flexible scheduling

Monitoring and alerting

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AWS Glue - data catalog

Make data discoverable



Automatically **discovers data** and **stores schema**

Catalog makes **data searchable**, and available for ETL

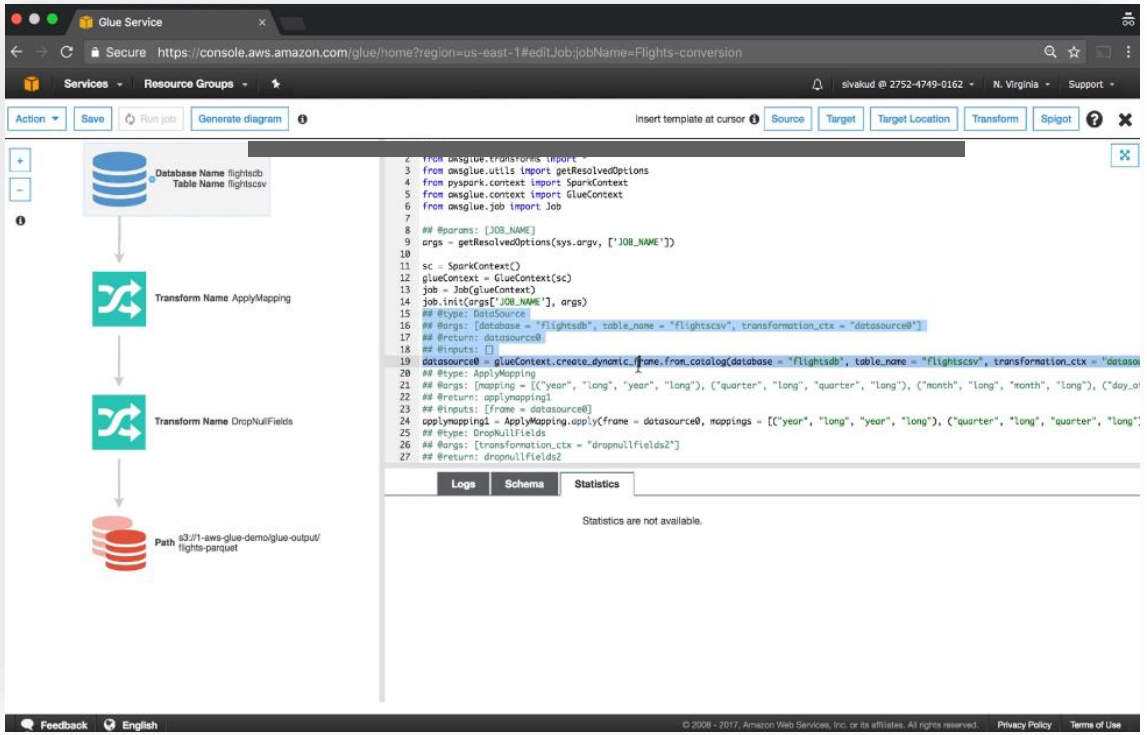
Catalog contains table and job definitions

Computes statistics to make queries efficient

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AWS Glue - ETL service

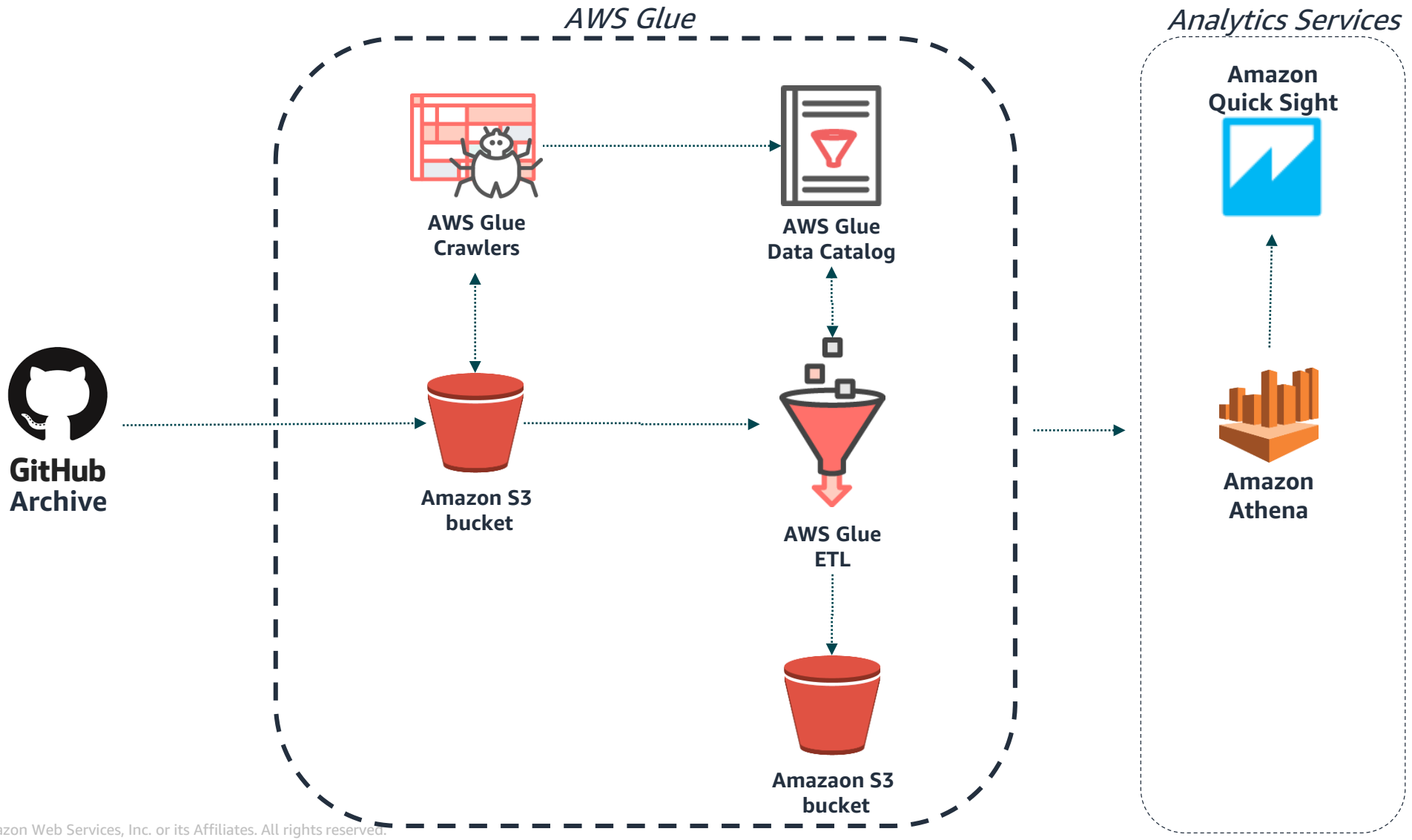
Make ETL scripting and deployment easy



- **Serverless** Transformations
- Based on **Apache Spark**
- Automatically **generates ETL code**
- Code is customizable with **PySpark** and **Scala**
- Endpoints provided to edit, debug, test code
- Jobs are **scheduled** or **event-based**

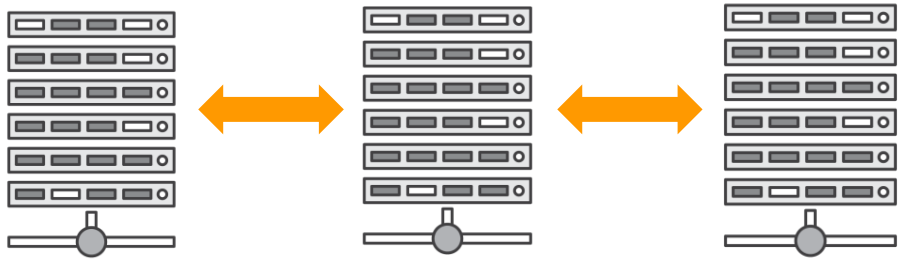
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

ETL example



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Apache Spark and AWS Glue ETL



SparkSQL	AWS Glue ETL
Dataframes	Dynamic Frames
Spark core: RDDs	

What is Apache Spark?

Parallel, scale-out data processing engine

Fault-tolerance built-in

Flexible interface: Python scripting, SQL

Rich eco-system: ML, Graph, analytics, …

AWS Glue ETL libraries

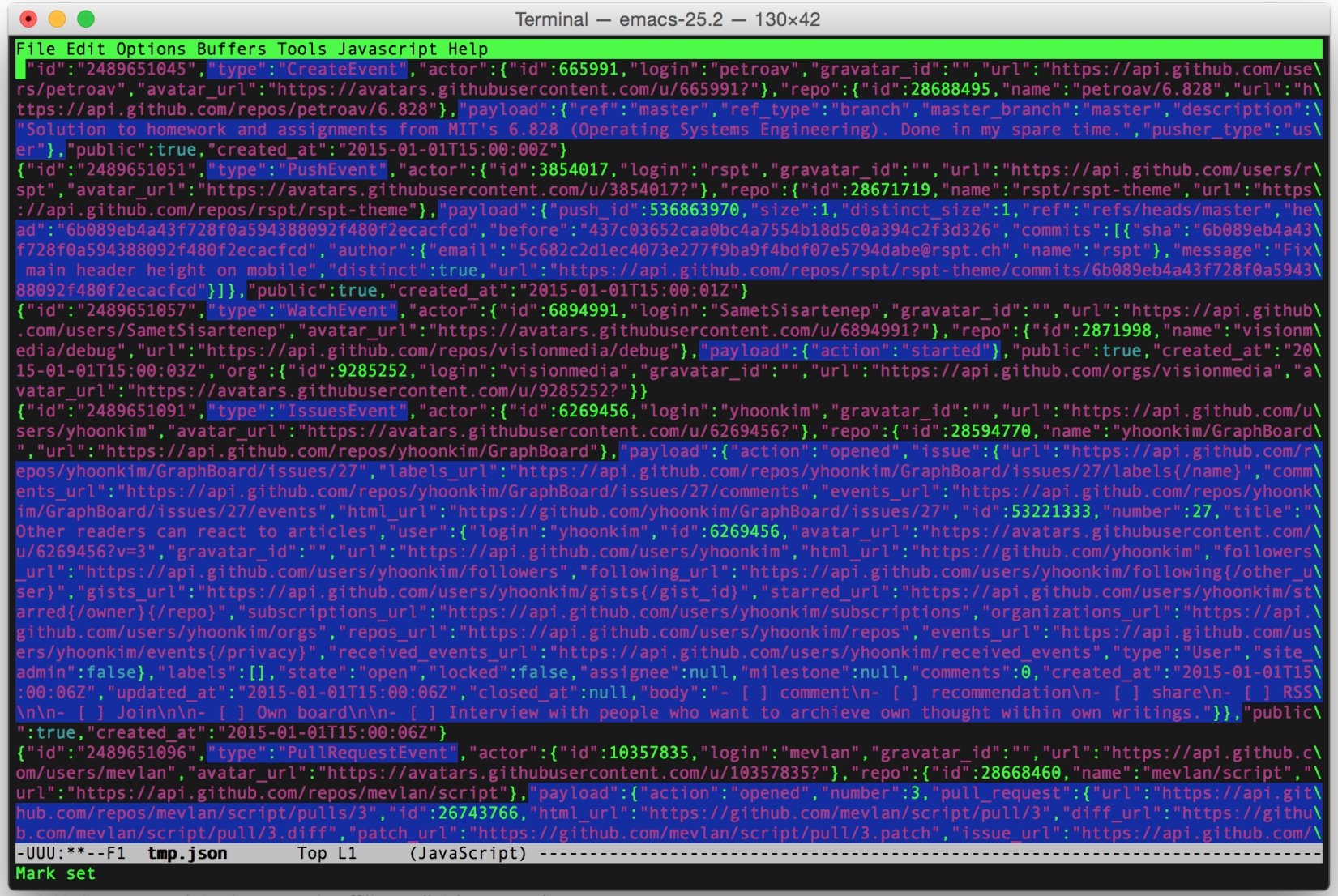
Integration: Data Catalog, job orchestration, code-generation, job bookmarks, S3, RDS

ETL transforms, more connectors & formats

New data structure: Dynamic Frames

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Public GitHub timeline is …



semi-structured

35+ event types

payload structure and size varies by event type

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Dataframes and Dynamic Frames

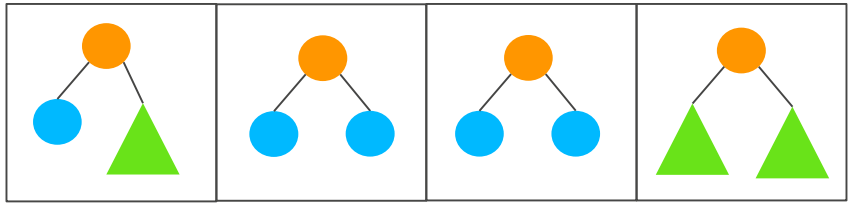
Dataframes

Core data structure for SparkSQL

Like structured tables

- Need schema up-front
- Each row has same structure

Suited for SQL-like analytics



Dynamic Frames

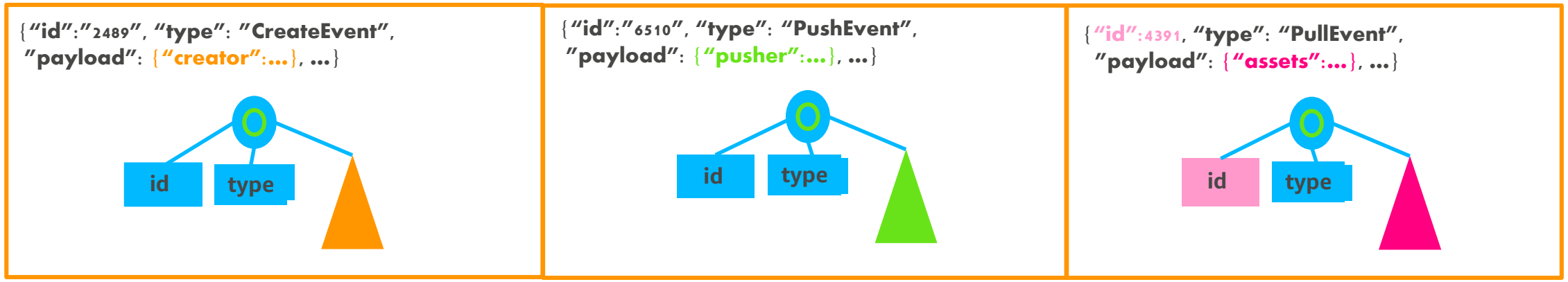
Like dataframes for ETL

Designed for processing **semi-structured** data, e.g. JSON, Avro, Apache logs …

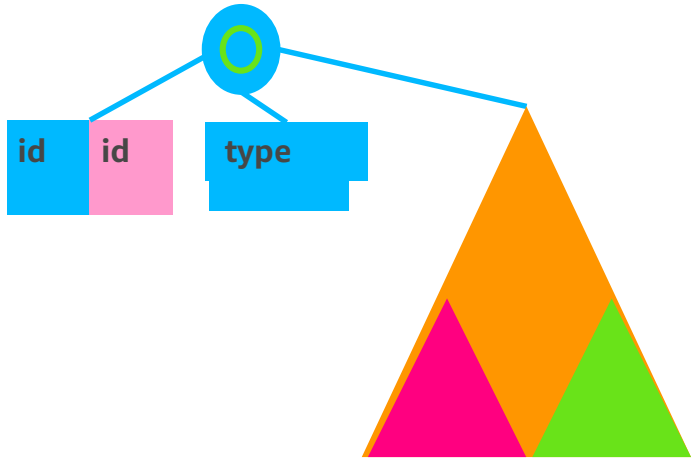
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Dynamic Frame internals

Dynamic Records



Dynamic Frame Schema



Schema per-record, **no up-front schema** needed

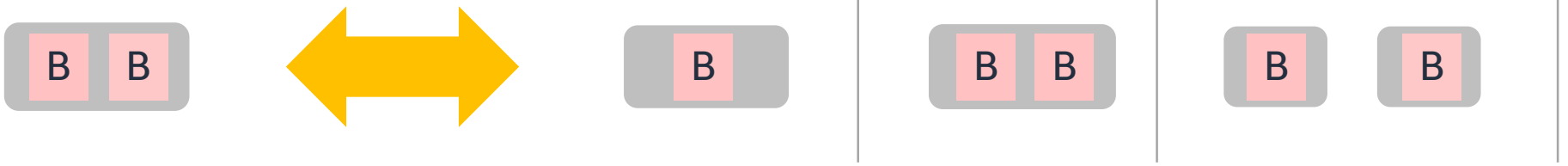
- Easy to restructure, tag, modify
- Can be more compact than dataframe rows
- Many flows can be done in single-pass

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

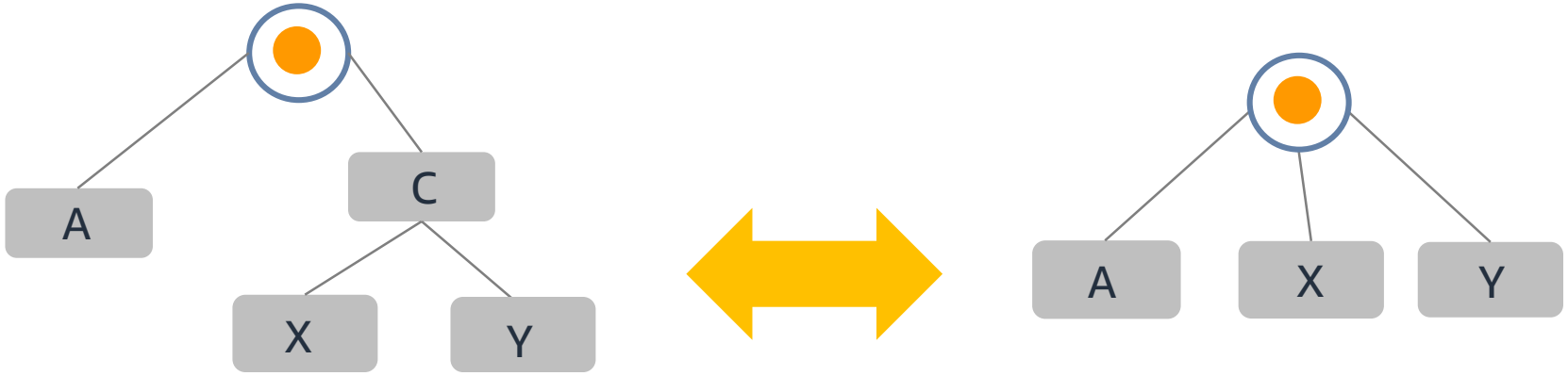
Dynamic Frame transforms

15+ transforms out-of-the box

ResolveChoice()

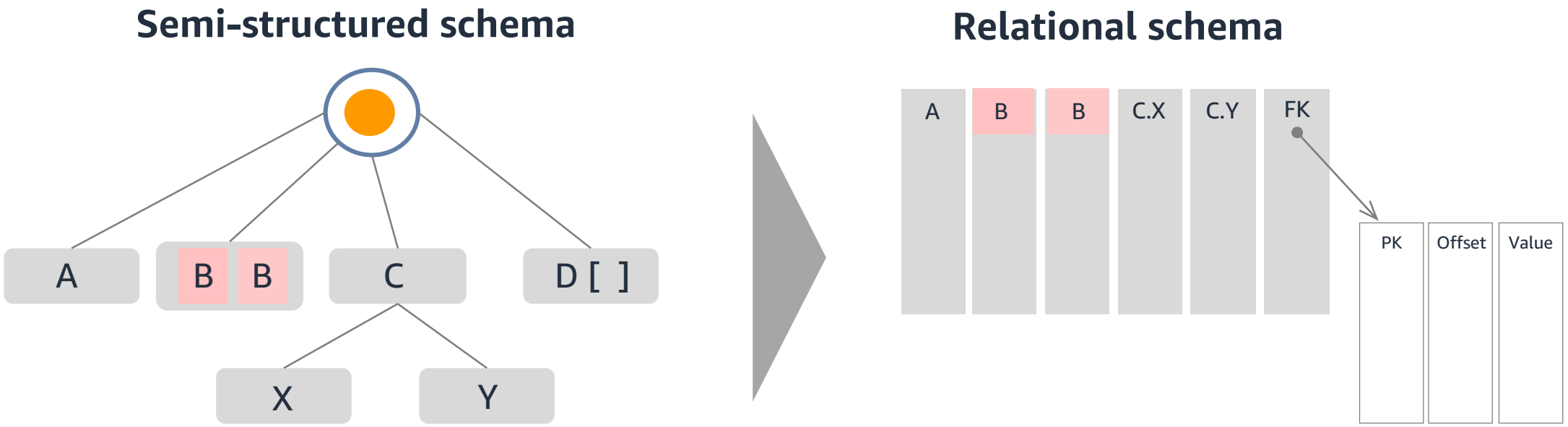


ApplyMapping()



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Relationalize() transform



Transforms and **adds new** columns, types, and tables on-the-fly

Tracks **keys** and **foreign keys** across runs

SQL on the relational schema is orders of **magnitude faster** than JSON processing

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Useful AWS Glue transforms

toDF(): Convert to a Dataframe

Spigot(): Sample data of any Dynamic Frame to S3

Unbox(): Parse string column as given format into Dynamic Frame

Filter(), **Map()**: Apply Python UDFs to Dynamic Frames

Join(): Join two Dynamic Frames

And more ….

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

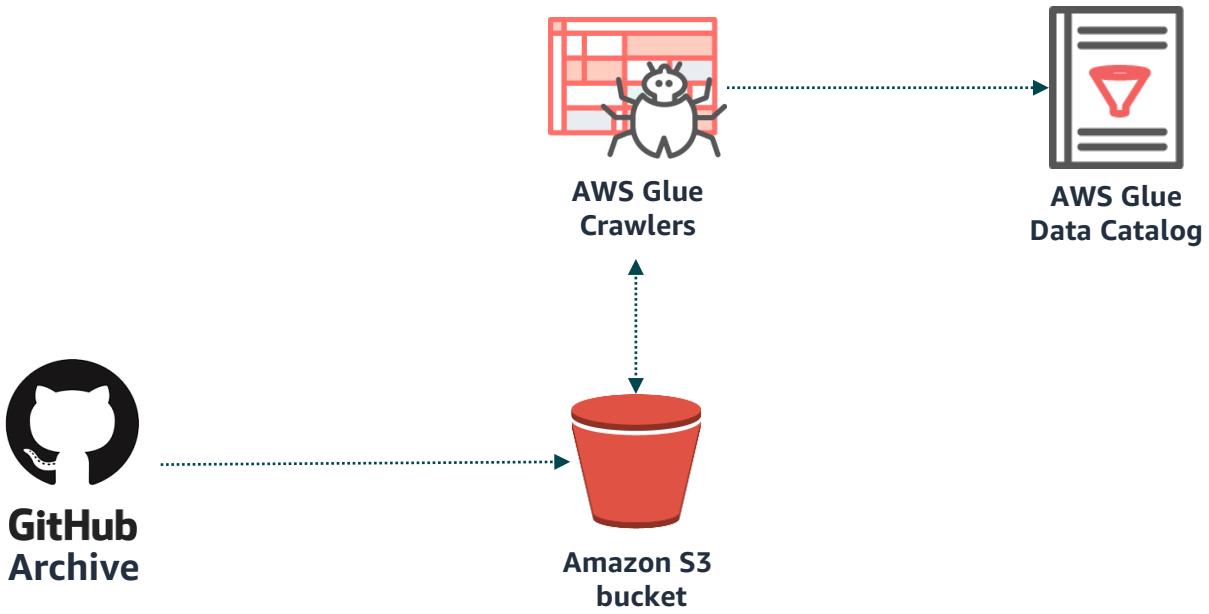
DEMO - Architecture

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

DEMO - Architecture

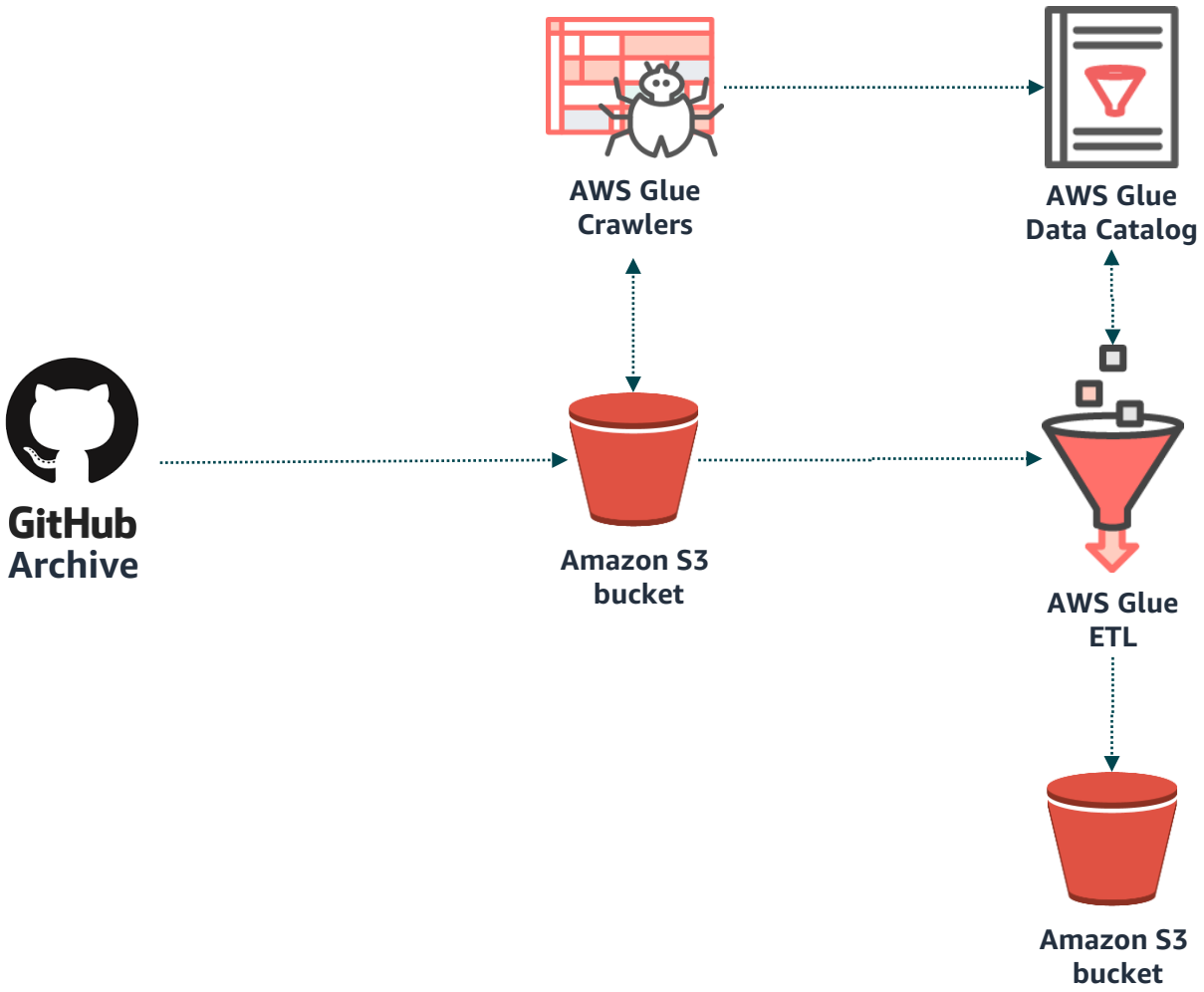
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

DEMO - Architecture



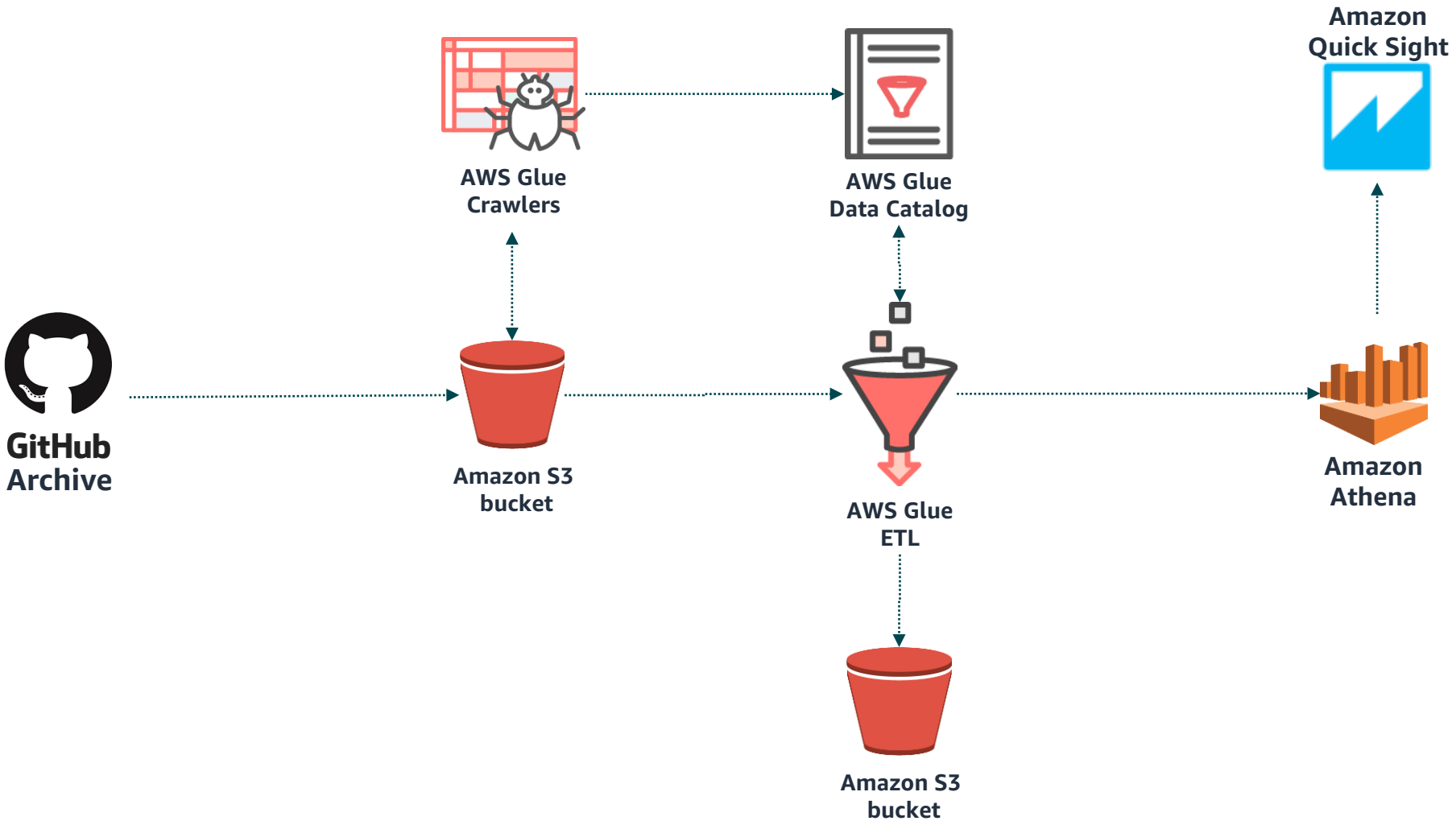
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

DEMO - Architecture



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

DEMO - Architecture



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.


Demo Screen
aws
Services
Resource Groups
admin @ unnikaws
N. Virginia
Support
AWS services
Find a service by name or feature (for example, EC2, S3 or VM, storage).
Recently visited services
All services
Compute
EC2
Lightsail
Elastic Container Service
Lambda
Batch
Elastic Beanstalk
Management Tools
CloudWatch
AWS Auto Scaling
CloudFormation
CloudTrail
Config
OpsWorks
Service Catalog
Systems Manager
Trusted Advisor
Managed Services
Mobile Services
Mobile Hub
AWS AppSync
Device Farm
Mobile Analytics
AR & VR
Amazon Sumerian
Application Integration
Step Functions
Amazon MQ
Simple Notification Service
Storage
S3
EFS
Glacier
Storage Gateway
Media Services
Elastic Transcoder
Helpful tips
Manage your costs
Get real-time billing alerts based on your cost and usage budgets. Start now
Create an organization
Use AWS Organizations for policy-based management of multiple AWS accounts. Start now
Explore AWS
Amazon Relational Database Service (RDS)
RDS manages and scales your database for you. RDS supports Aurora, MySQL, PostgreSQL, MariaDB, Oracle, and SQL Server. Learn more.

Take the demo home…

http://bit.ly/aws-innovate-2018-glue-demo

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

TOYOTA TSUSHO
NEXTY ELECTRONICS
(THAILAND) CO., LTD.

NETH Traffic Information Provisioning with AWS Data Lake

Content Development & Distribution

Thanomsak Ajjanapanya
Group Manager Content Department

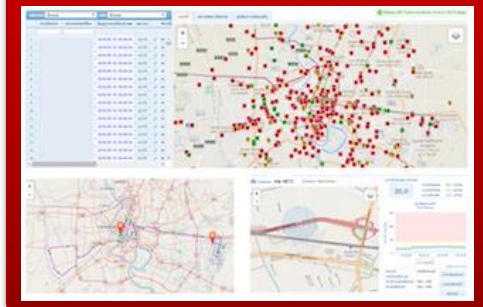
Copyright © NEXTY Electronics Corporation

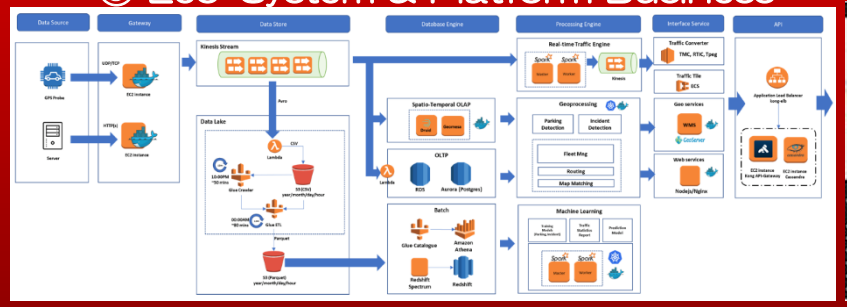
NETH Contents Business Overview

TOYOTA TSUSHO NEXTY ELECTRONICS (THAILAND) CO., LTD.

Automotive (Car-OEM) | Logistic | Parking | ..... | Health Care

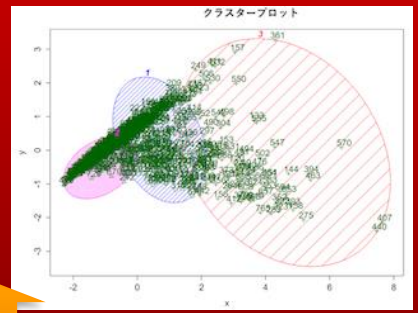
③ Eco-System & Platform Business





① Content Business

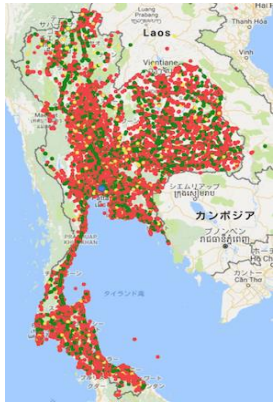
② Data Analysis Business



Copyright © NEXTY Electronics Corporation

NETH Traffic Provisioning Story

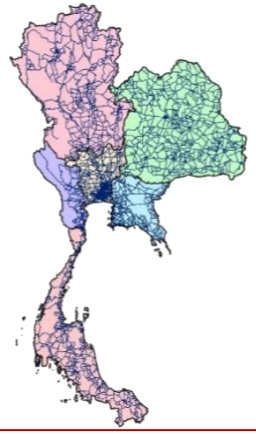
TOYOTA TSUSHO
NEXTY ELECTRONICS
(THAILAND) CO., LTD.



GPS Data

- 140,000 vehicles
- 100 Million data/day
- 3 Billion data/month

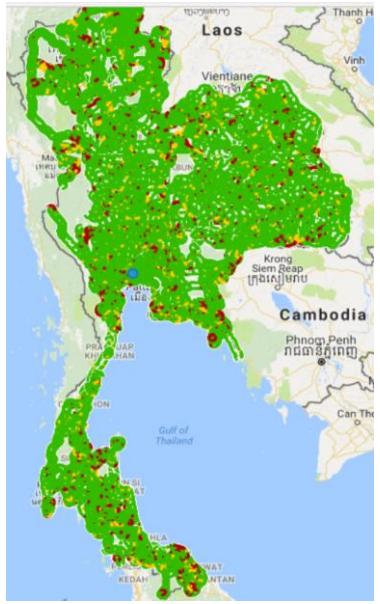
Traffic Info Provisioning

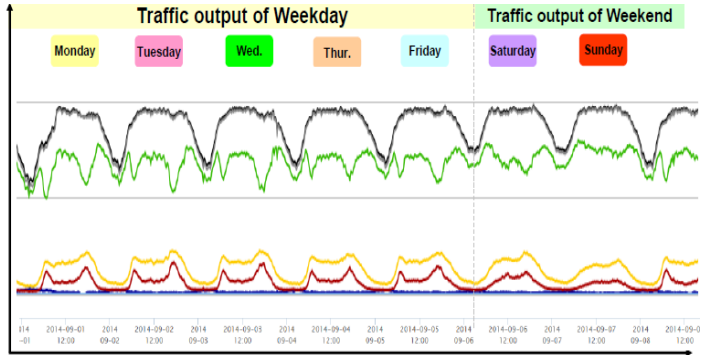


Road Network

110,000 Road Links

Traffic Information

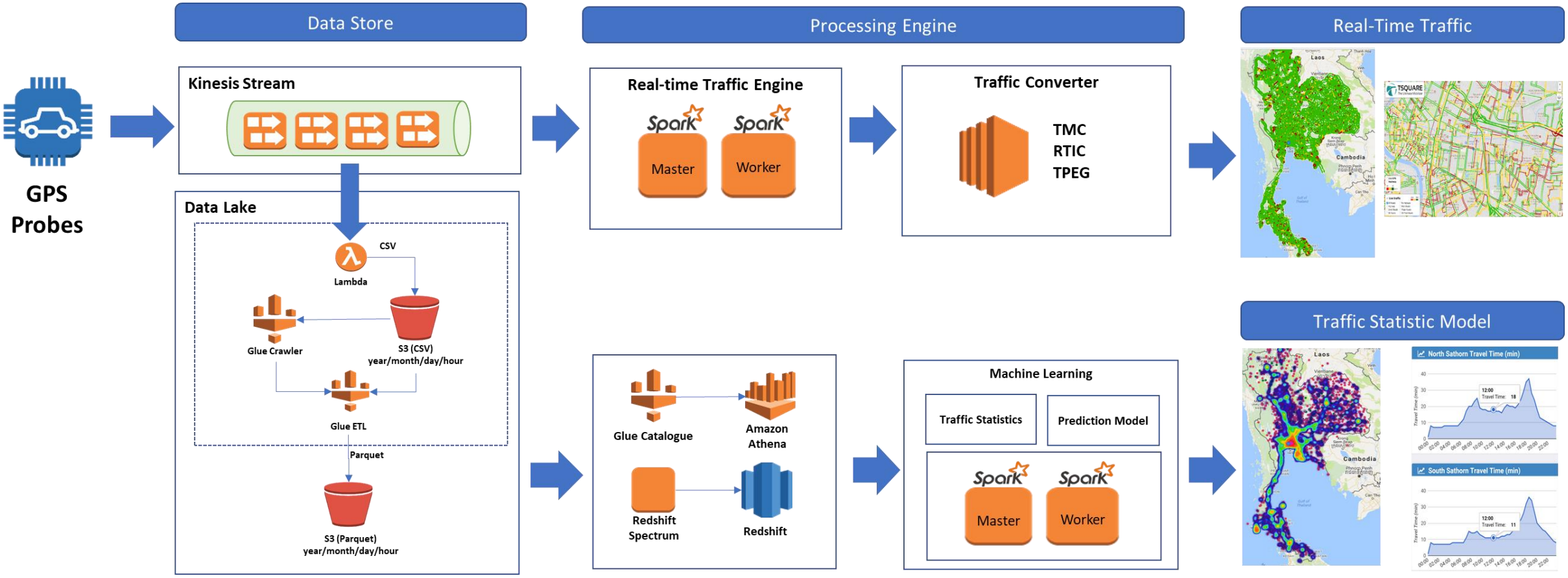




Copyright © NEXTY Electronics Corporation

Data Lake Architecture

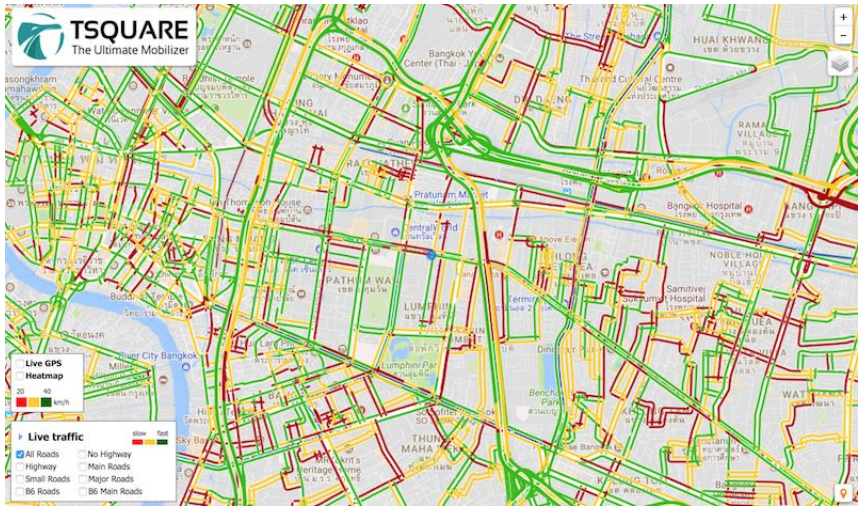
TOYOTA TSUSHO NEXTY ELECTRONICS (THAILAND) CO., LTD.

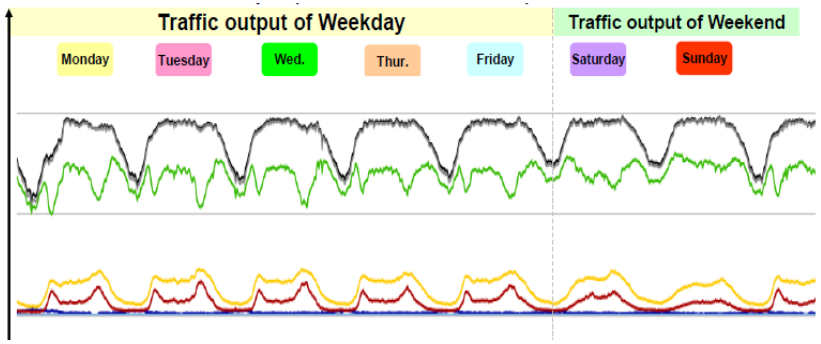


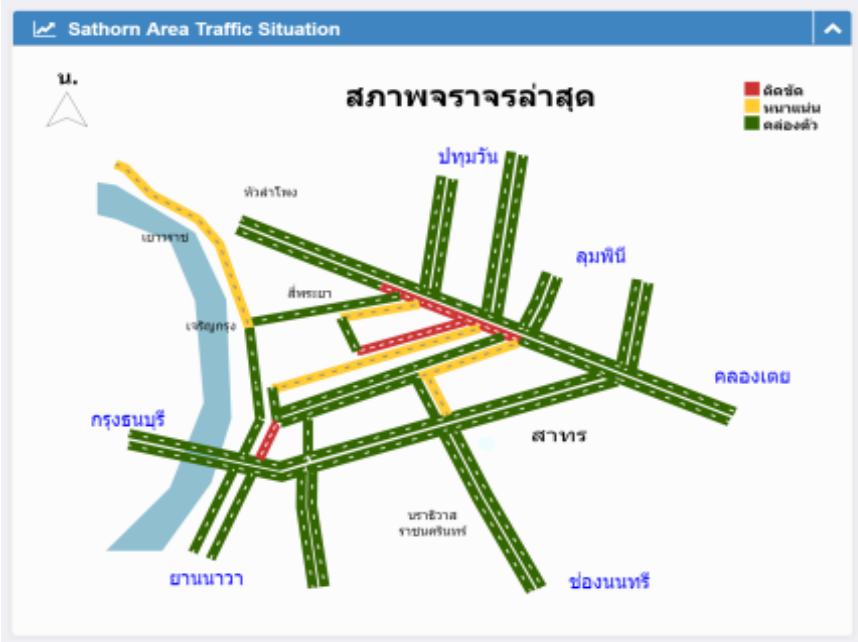
Copyright © NEXTY Electronics Corporation

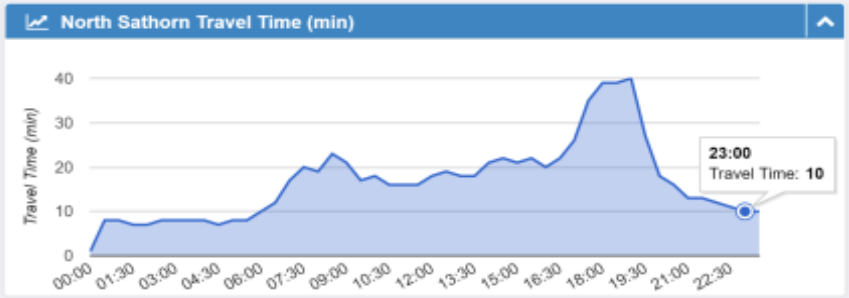
Learning and practices

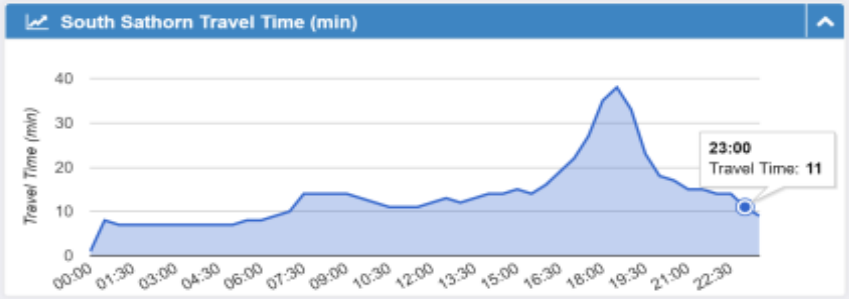
TOYOTA TSUSHO NEXTY ELECTRONICS (THAILAND) CO., LTD.











Copyright © NEXTY Electronics Corporation

Take the demo home…

http://bit.ly/aws-innovate-2018-glue-demo

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Learn from AWS experts. Advance your skills and knowledge. Build your future in the AWS Cloud.

Digital Training

Free, self-paced online courses built by AWS experts

Classroom Training

Classes taught by accredited AWS instructors

AWS Certification

Exams to validate expertise with an industry-recognized credential

Ready to begin building your cloud skills?
Get started at: https://www.aws.training/

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

With deep expertise on AWS, APN Partners can help your organization at any stage of your Cloud Adoption Journey.

AWS Managed Service Providers

APN Consulting Partners who are skilled at cloud infrastructure and application migration, and offer proactive management of their customer's environment.

AWS Competency Partners

APN Partners who have demonstrated technical proficiency and proven customer success in specialized solution areas.

AWS Marketplace

A digital catalog with thousands of software listings from independent software vendors that make it easy to find, test, buy, and deploy software that runs on AWS.

AWS Service Delivery Partners

APN Partners with a track record of delivering specific AWS services to customers.

Ready to get started with an APN Partner?
Find a partner: https://aws.amazon.com/partners/find/
Learn more at the AWS Partner Network Booth

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Thank You for Attending AWS Innovate

We hope you found it interesting! A kind reminder to **complete the survey.**

Let us know what you thought of today's event and how we can improve the event experience for you in the future.


aws-apac-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.