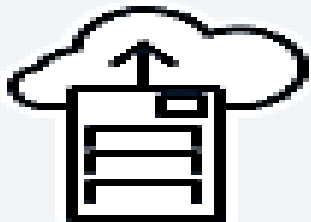


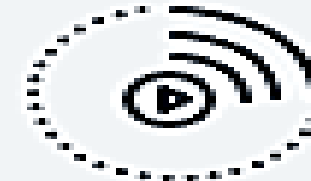
Machine Learning



Analytics



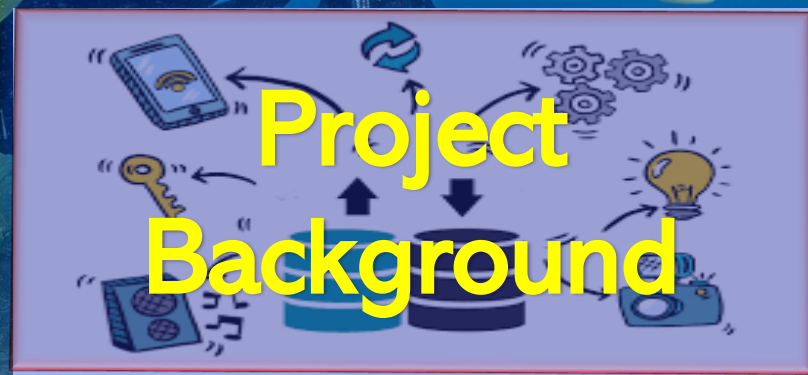
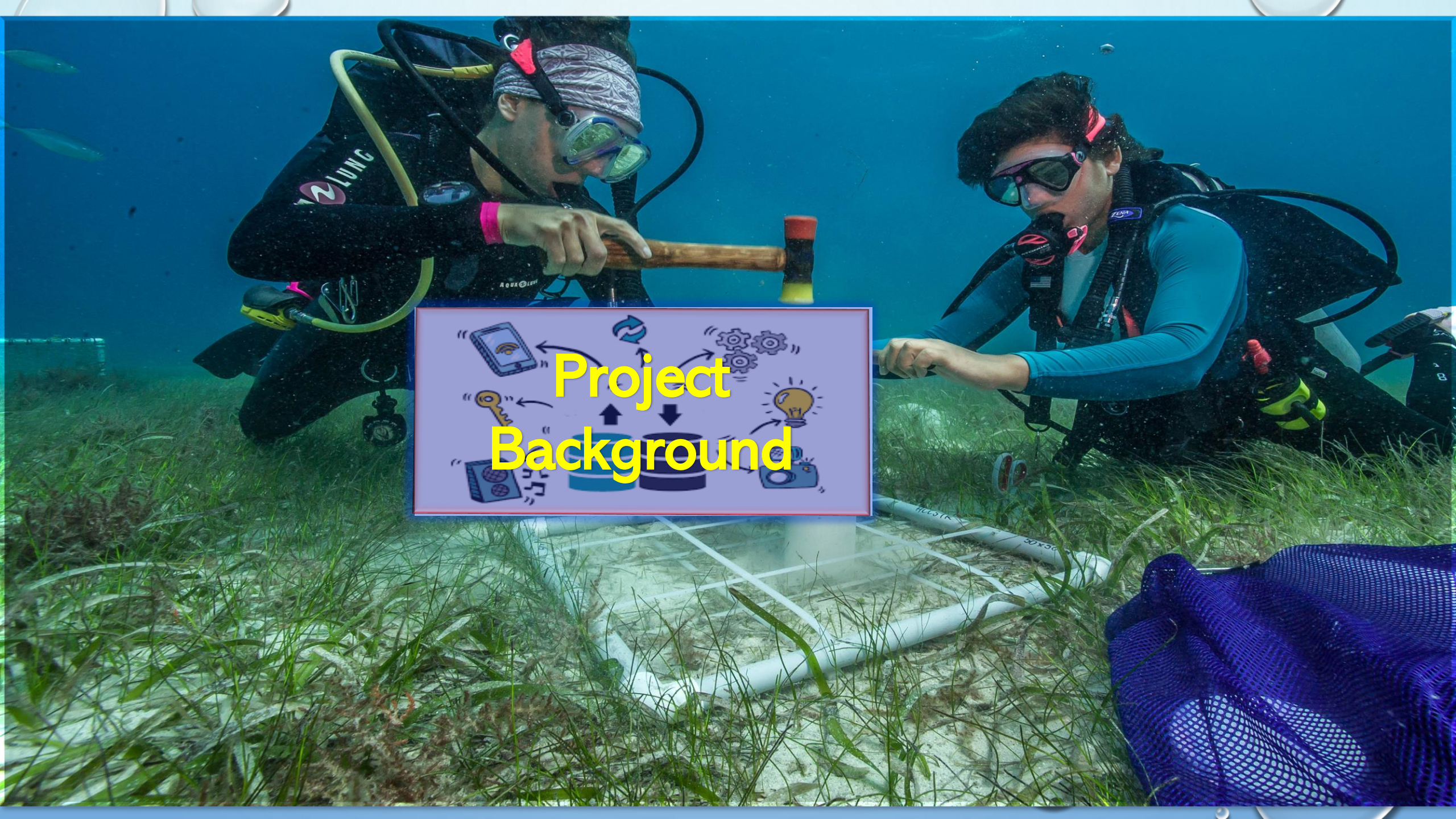
On-premises
Movement



Real-time Data

Content

- ✓ Project Background
- ✓ Asking Business Questions
- ✓ Data lake & Analytics Solutions CSA On AWS
- ✓ Achievements and Recommendations
- ✓ Reference
- ✓ Acknowledgment




➤ Problem Statement

- ❑ I'm working at a company that is planning to use Amazon Simple Storage Service (Amazon S3) as the storage layer for their data lake solution. Initially, the data that will be ingested into the data lake will come from three locations:
 - ✓ Internet of Things (IoT) sensors that send real-time data
 - ✓ A database with historical records
 - ✓ Supplemental data from third-party entities for enriching internally generated data
- ❑ These components are currently running in the data center on physical servers. Currently, if a power outage occurred in the data center, all systems would be brought offline. Because of this issue (in addition to other benefits of the cloud), my customer wants to migrate all components to the cloud and, when possible, use AWS services to replace on-premises components.

➤ What will I do as a Cloud Solutions Architect?

- ❑ The company has tasked me with designing solutions for ingesting this data into their data lake, and each location (IoT sensors, database, and third party) will need its own ingestion solution.
- ❑ From there, I will need to also design a solution for how to clean or transform the data so that it can be analyzed. The company currently uses Apache Hadoop-based software. When possible, they prefer to use similar technologies in the cloud so that they don't need to retrain their analytics team on too many new technologies at one time.
- ❑ The company also has a requirement to create dashboards that show visual representations of the insights they derive from the data.

A man with dark hair, wearing a blue and white patterned button-down shirt, is seated in a classroom or meeting. He is raising his right hand, holding a yellow pencil, as if to ask a question or make a point. He has a name tag on his chest. In the background, other people are blurred, including a woman in a black blazer and a man in a white shirt. The scene is brightly lit, suggesting an indoor setting with large windows.

**What,
Why and
How?**

**Asking Business
Questions?**

➤ Some Crucial Questions to be answered

- ☐ How will I ingest data from IoT sensors into Amazon S3? Which AWS services will you use to stream, process, and store the real-time data?
- ☐ How will I ingest data from the database into Amazon S3? Which AWS services will I use to extract, transform, and load (ETL) the historical records?
- ☐ How will I ingest data from third-party entities into Amazon S3? Which AWS services will I use to acquire, validate, and integrate the supplemental data?
- ☐ How will I clean and transform the data in Amazon S3 so that it can be analyzed? Which services will I use to perform data quality checks, partitioning, compression, encryption?
- ☐ How will I leverage Apache Hadoop-based software in the cloud for analyzing the data in Amazon S3? Which services will I use to run Hadoop clusters and access data from S3?
- ☐ How will I create dashboards that show visual representations of the insights derived from the data in Amazon S3? What tools or services will I use to connect to Amazon S3 and visualize the data?

Source Data

(examples)



OnPremise Data

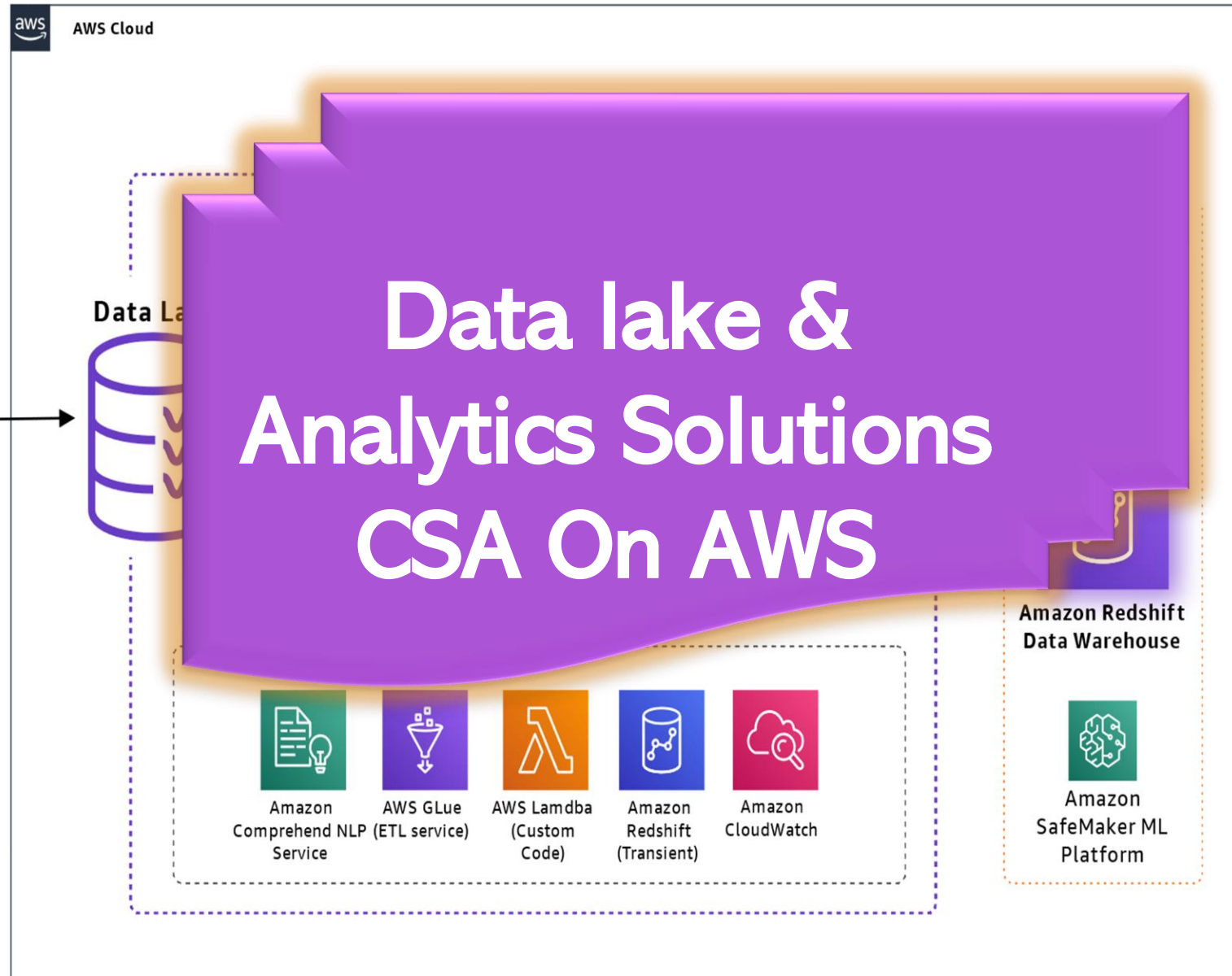


Social Media data



Streaming data

Store, Ingest and Backup



Visualize



End Users



Insights, Reports

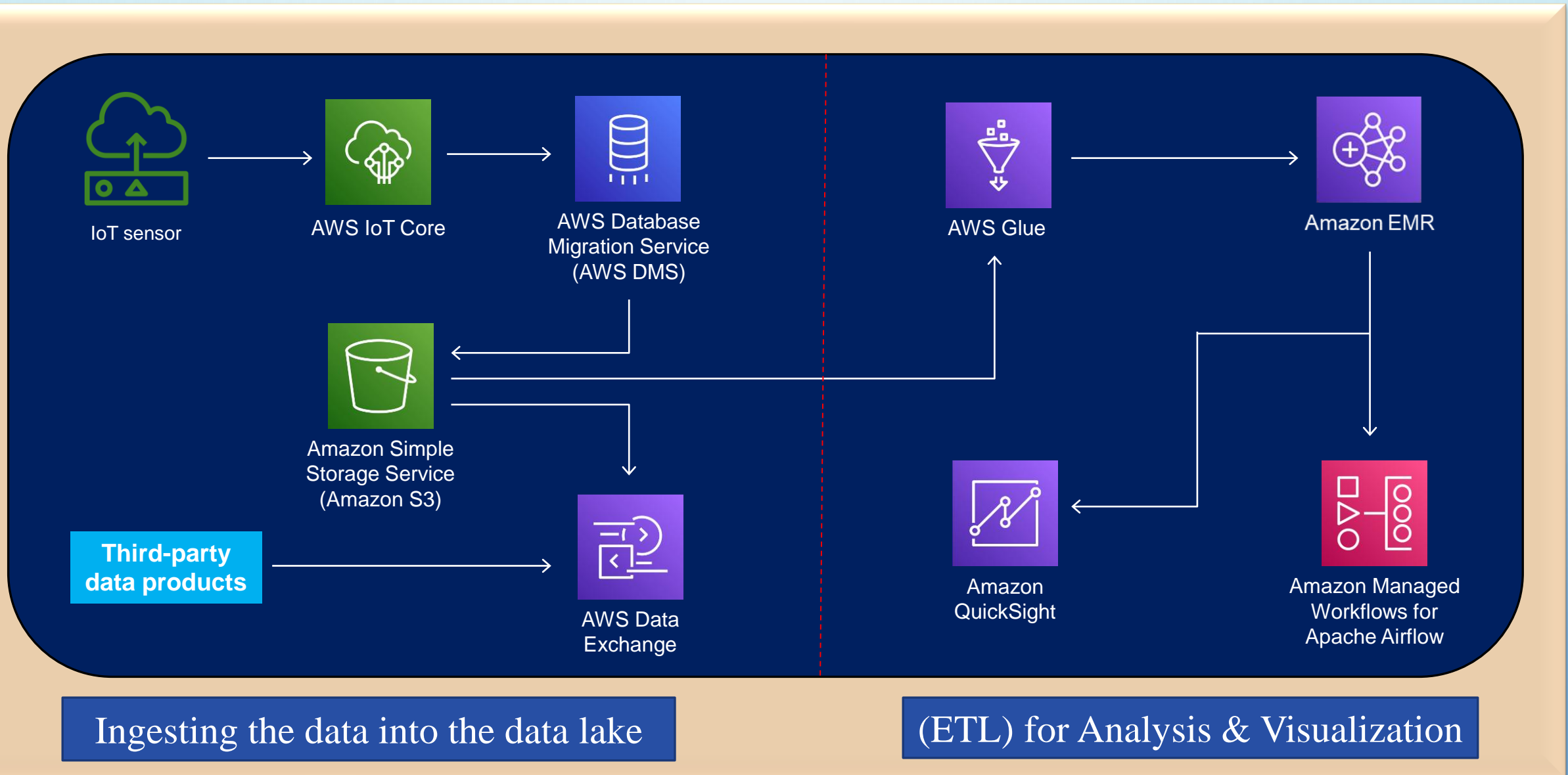


Operational
Systems and
Processes

➤ Solution Architecting

- ❑ For ingesting data from IoT sensors into Amazon S3, I can use AWS IoT Core to collect and process data from connected devices. For the database with historical records, I can use AWS Database Migration Service (DMS) to migrate data to Amazon S3. For supplemental data from third-party entities, I can use AWS Data Exchange to find and subscribe to third-party data products.
- ❑ Once the data is ingested into Amazon S3, I can use AWS Glue for cleaning and transforming the data. Since my company is already familiar with Apache Hadoop-based software, I can use Amazon EMR (Elastic MapReduce) for processing and analyzing the data.
- ❑ For creating dashboards that show visual representations of insights derived from the data, I can use Amazon QuickSight.

➤ Data lake and Analytics Solutions CSA on AWS



➤ Workflow of Data lake and Analytics Solutions

- ❑ The data-lake-and-analytics solutions for the company would involve using several AWS services to ingest, process and analyze data from various sources. The architecture workflow stated as follow
- ❑ The previous page cloud solution architecture (CSA) for data lake and analytics solutions diagram shows how data flows from IoT sensors to AWS IoT Core, then to AWS DMS, which migrates historical records from the company's database and third-party data products from AWS Data Exchange to Amazon S3. Amazon S3 serves as the storage layer for the data lake, where data can be cleaned and transformed by AWS Glue and processed and analyzed by Amazon EMR using Apache Hadoop-based software. Finally, Amazon QuickSight creates dashboards that show visual representations of insights derived from the data.

➤ Why I Choose Those Specific AWS Services for CSA

- ❑ AWS IoT Core was chosen for collecting and processing data from IoT sensors in real-time because it is a managed cloud service that lets connected devices easily and securely interact with cloud applications and other devices.
- ❑ AWS Database Migration Service (DMS) was chosen for migrating historical records from the company's database to Amazon S3 because it is a cloud service that makes it easy to migrate relational databases, data warehouses, NoSQL databases, and other types of data stores.
- ❑ AWS Data Exchange was chosen for finding and subscribing to third-party data products that can be used to enrich internally generated data because it makes it easy to find, subscribe to, and use third-party data in the cloud.
- ❑ Amazon S3 was chosen as the storage layer for the company's data lake because it is an object storage service that offers industry-leading scalability, data availability, security, and performance.

➤ Cont...

- ❑ AWS Glue was chosen for cleaning and transforming ingested data so that it can be analyzed because it is a fully managed extract, transform, and load (ETL) service that makes it easy to move data between data stores.
- ❑ Amazon EMR (Elastic MapReduce) was chosen for processing and analyzing data using Apache Hadoop-based software that the company is already familiar with because it is a managed cluster platform that simplifies running big data frameworks such as Apache Hadoop and Apache Spark on AWS.
- ❑ Amazon QuickSight was chosen for creating dashboards that show visual representations of insights derived from the data because it is a fast business intelligence service that makes it easy to deliver insights to everyone in my organization.



Achievements & Recommendations

➤ Final Results and Achievements

- ❑ I have successfully ingested data from different sources into Amazon S3 using AWS IoT Core, AWS DMS, and AWS Data Exchange.
- ❑ I have cleaned and transformed the data using AWS Glue, which is a serverless data integration service that simplifies ETL (extract, transform, and load) tasks.
- ❑ I have processed and analyzed the data using Amazon EMR, which is a managed cluster platform that simplifies running big data frameworks such as Apache Hadoop and Apache Spark on AWS.
- ❑ I have created dashboards that show visual representations of insights derived from the data using Quicksight, which is a cloud-based business intelligence service that makes it easy to create and share interactive dashboards.

➤ Recommendations

- ❑ I suggest to optimize the company's data ingestion pipeline by using AWS IoT Analytics to filter, transform, enrich, and store IoT data before loading it into Amazon S3.
- ❑ I advice to improve the company's data quality by using AWS Lake Formation to define policies and govern access to your data lake.
- ❑ I propose to enhance company's data analysis by using Amazon Athena to query your data in Amazon S3 using standard SQL without setting up any servers or clusters.
- ❑ I advocate to leverage machine learning capabilities by using Amazon SageMaker to build, train, and deploy ML models on your data.

Reference

❑ AWS- Cloud Solutions Architect Specializations Professional Certificate

- ✓ Visit My GitHub Portfolio: <https://github.com/kedibeki>
- ✓ Visit the Specializations Page: <https://www.coursera.org/professional-certificates/aws-cloud-solutions-architect>
- ✓ AWS Training & Certification: <https://www.aws.training>

Acknowledgment

☐ AWS (Amazon Web Services) and Coursera- AWS Cloud Solutions Architect Specializations Instructors

- ✓ **Morgan Willis:** Senior Cloud Technologist @AWS Training & Certification
- ✓ **Rafael Lopes:** Principal Cloud Technologist @AWS Training & Certification
- ✓ **Seph Robinson:** Cloud Technologist @AWS Training and Certification
- ✓ **Alana Layton:** Technical Curriculum Developer

A data lake is more than just a storage repository. It is a powerful platform for analytics and machine learning that can transform your raw data into valuable insights and ML Predictions.



THANK YOU!

