Cloud Data Architecture

ผู้ช่วยศาสตราจารย์ ดร.สมเกียรติ โกศลสมบัติ

สาขาวิชาวิทยาศาสตร์และนวัตกรรมข้อมูล

วิทยาลัยสหวิทยาการ มหาวิทยาลัยธรรมศาสตร์

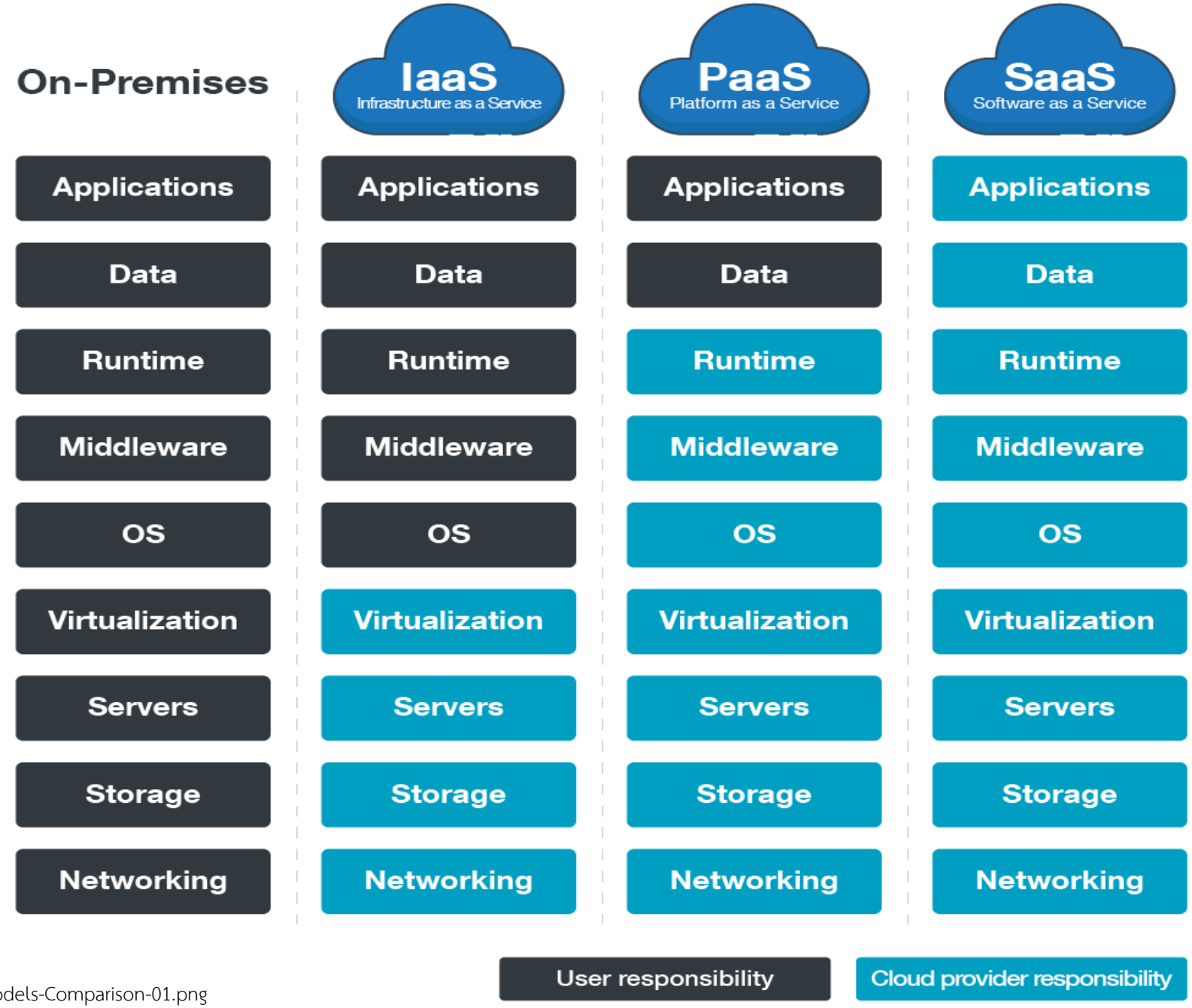
Contents

Introduction to Cloud Data Architecture

- Cloud Data Architecture คือ โครงสร้างการจัดการข้อมูลที่ออกแบบให้ทำงานบนแพลตฟอร์ม Cloud โดยรวมถึงวิธีการเก็บข้อมูล การประมวลผล การวิเคราะห์ และการเข้าถึงข้อมูลในสภาพแวดล้อมแบบกระจาย (Distributed Environment)
- Flexibility, Scalability, Real-time หรือ Near Real-time

ลักษณะ	Traditional Architecture	Cloud Data Architecture
โครงสร้างพื้นฐาน	On-premise (ตั้งอยู่ในองค์กร)	บริการบน Cloud
Scalability	จำกัด โดยต้องจัดซื้อ H/W ใหม่	ปรับโดยอัตโนมัติ
Cost	ค่าใช้จ่ายสูง	Pay-as-you-go (Pay per use)
การจัดการข้อมูล	Manual	Automation and Serverless
การเข้าถึงข้อมูล	จำกัดอยู่ในองค์กร	ผ่าน API or Web ทุกที่

Cloud Service Models



https://documents.trendmicro.com/images/TEx/articles/Cloud-Service-Models-Comparison-01.png

Cloud Service Models

- Function as a Service (FaaS)
  - Serverless, Run code without provisioning or managing servers
  - AWS Lambda, Google Cloud Functions, Azure Functions
- Database as a Service (DBaaS)
  - Relational DB (Amazon RDS, Azure SQL)
  - NoSQL (Google Firestore / Big Table, Azure Cosmo DB, MongoDB Atlas, Amazon DynamoDB)

Cloud Service Models

- Machine Learning as a Service (MLaaS)
  - Use ML tools and models via API, AutoML (Azure ML Studio, IBM Watson)
- Anything as a Service (XaaS)
  - Desktop as a Service (DaaS) Amazon WorkSpaces, Azure Virtual Desktop
  - Network as a Service (NaaS), Analytics as a Service (AaaS), Storage as a Service (STaaS)

Core Components of Cloud Data Architecture

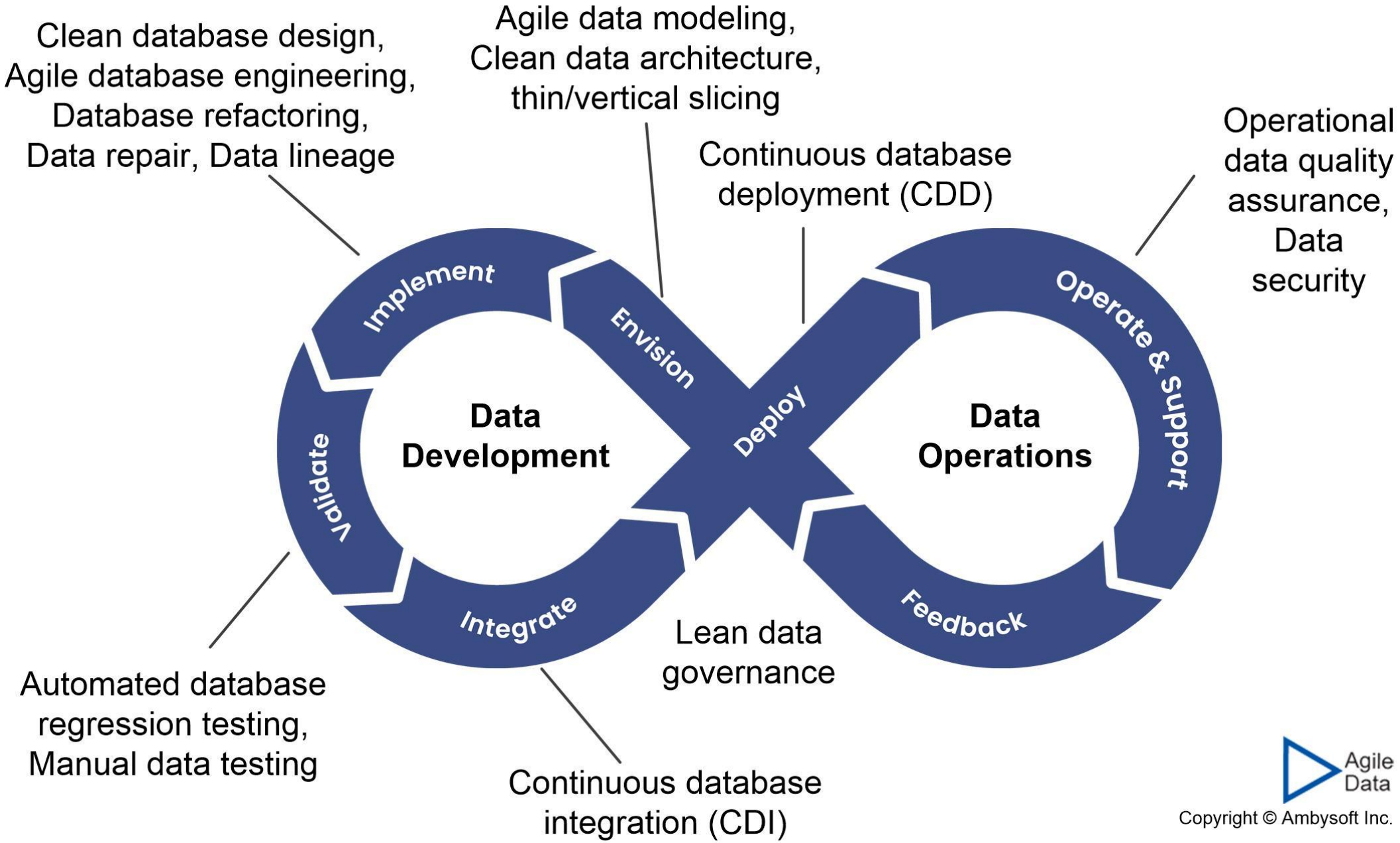
- Data Sources
  - IoT, Web Apps, Databases, APIs
- Data Ingestion
  - Batch (AWS Glue, Google Cloud Dataflow)
  - Real-time (Apache Kafka, AWS Kinesis, Azure Event Hubs)
- Data Storage
  - Data Lake (Amazon S3, Azure Data Lake)
  - Data Warehouse (Snowflake, Google BigQuery, AWS Redshift)
  - SQL / NoSQL

Core Components of Cloud Data Architecture

- Data Processing
  - ETL / ELT Pipelines
  - Stream Processing (Apache Spark, Apache Flink)
  - Serverless (AWS Lambda, Google Cloud Functions)
- Data Orchestration (Workflow Orchestration)
  - Airflow, Prefect
- Data Governance & Security
  - Data Catalog (Apache Atlas, DataHub)
  - Access Control, Encryption
  - Compliance
  - Identity and Access Management (IAM)

Data Operations (DataOps)

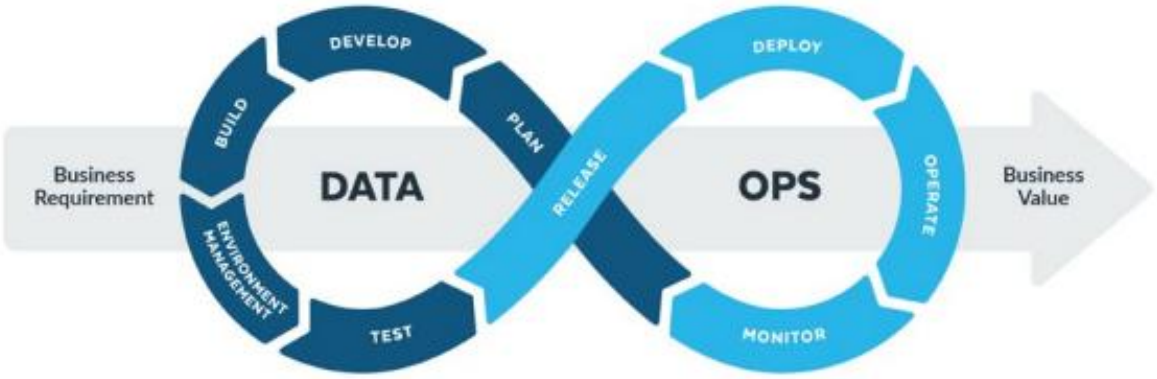
- DataOps is a set of collaborative data management practices intended to speed delivery, maintain quality, foster collaboration and provide maximum value from data.
- Modeled after DevOps practices, DataOps' goal is to ensure that previously siloed development functions are automated and agile.
- While DevOps is concerned with streamlining software development tasks.
- DataOps focuses on automating the data management and data analytics process.



https://agiledata.org/wp-content/uploads/2023/09/DataOps.jpg

Continuous Integration / Continuous Deployment (CI/CD)

- Continuous Integration (CI) involves regularly integrating data models, scripts, and other data assets into a shared repository.
- This process is accompanied by automated tests that quickly identify and fix errors, thereby improving data quality and consistency.
- Continuous Deployment (CD) automates the deployment of data models and pipelines to production environments, reducing manual intervention and speeding up data delivery.



https://miro.medium.com/v2/resize:fit:640/format:webp/1*5OzigyopmT0OkeGN1no_KQ.png

References

- https://www.ibm.com/think/topics/dataops
- https://medium.com/@AmirKheirollah/embracing-dataops-ci-cd-for-enhanced-data-management-and-deployment-e4aa0648c647