KDT 2기 멘토링 프로젝트 소개

- Engineering Part

AI센터 개발운영팀 박동우

- **주제: AWS 환경 기반 데이터 수집(웹 크롤링) 자동화 및 분석 환경 구축**

- **미션**

- 분석에 필요한 데이터 테이블 스키마 설계
- 자동화 구현 범위 수립 및 AWS 아키텍처 설계
- 웹 크롤링 봇 구현 및 테스트
- 코드 형상관리 및 변경 이력에 따른 코드 배포 자동화
- 데이터 수집 및 분석 환경 위한 AWS 서비스 구현
- 데이터 변경에 따른 AWS 서비스 최신화 자동화 구현

- [Optional] 분석 결과 시각화 웹서버 구현
- [Optional] 데이터 변경에 따른 실시간 웹서버 자동 배포

- **필요/기대 역량**

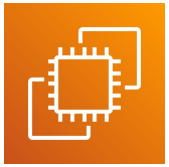
- 웹 크롤링 기초 역량(HTML Tag, Selenium)
- Docker 기초 지식, CI/CD 개념 입문
- AWS 클라우드 서비스 기초 경험

Sample

- AWS 기술 스택 예시


AWS Glue


Amazon Athena


Amazon Elastic Compute Cloud (Amazon EC2)


AWS Lambda

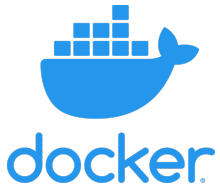

Amazon Elastic Container Registry (Amazon ECR)


Amazon Simple Storage Service (Amazon S3)


Amazon CloudWatch

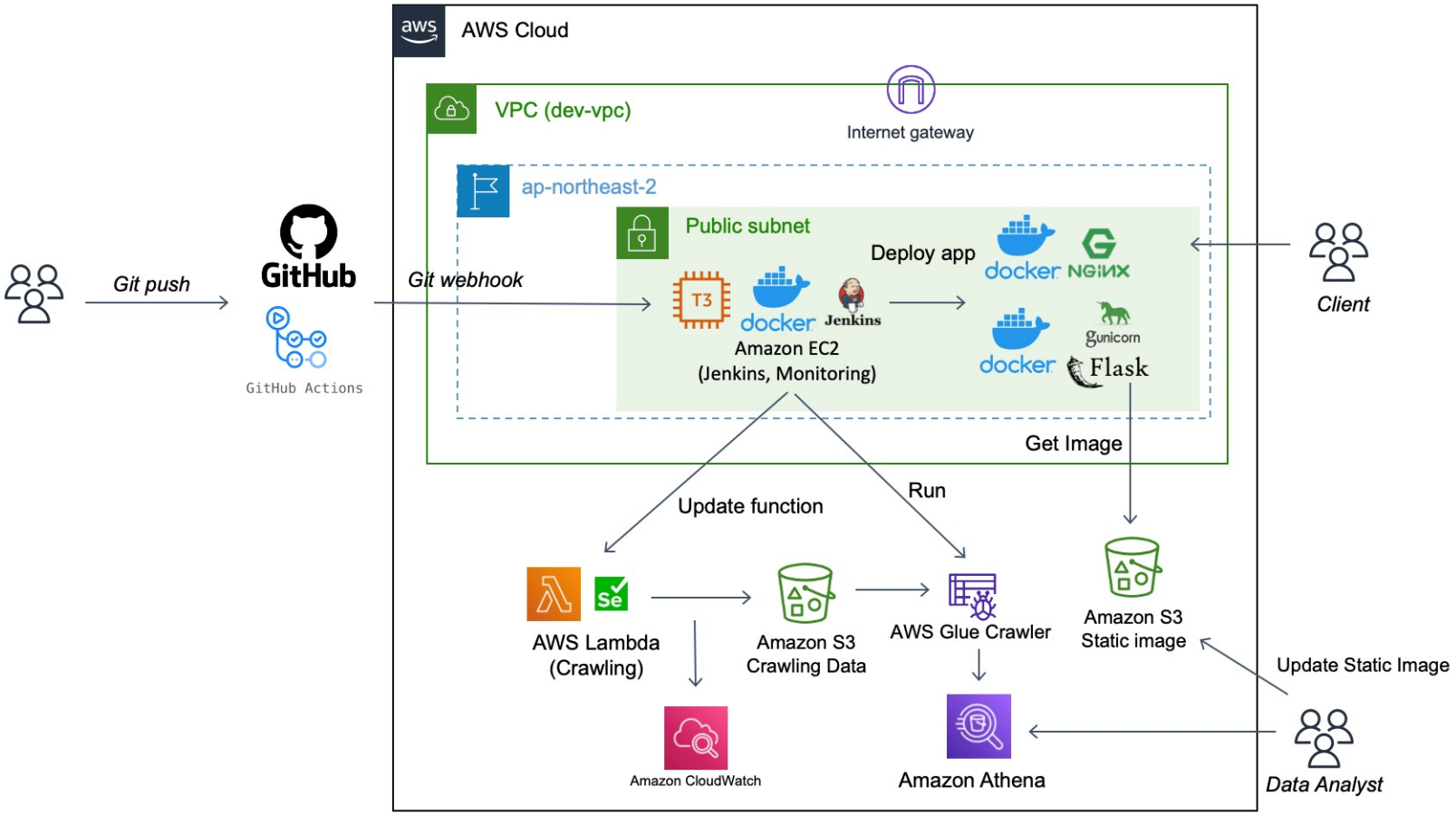
Sample

- Library/Tools/SDK



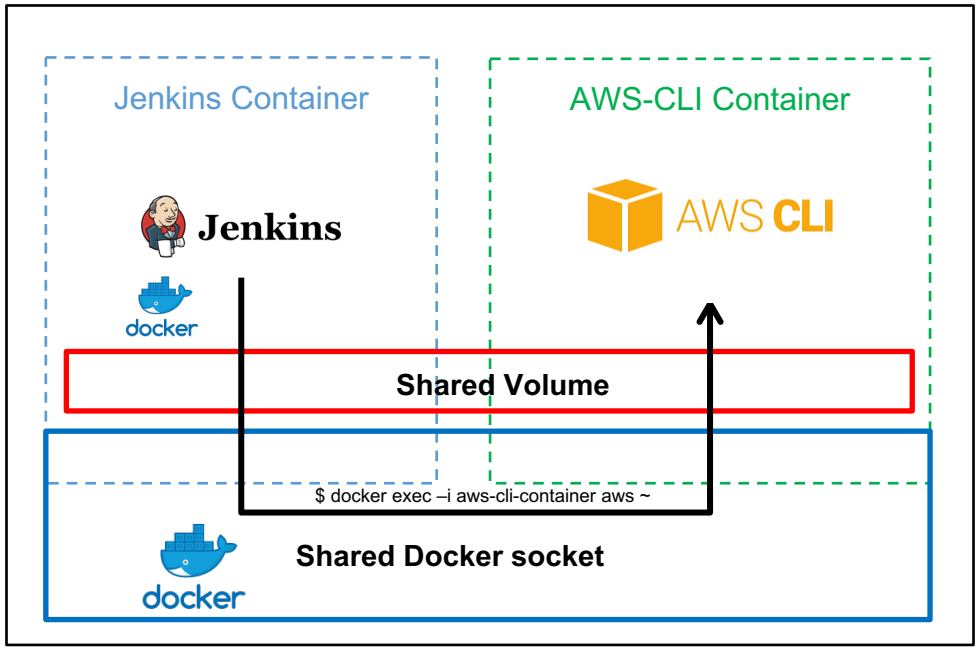
Sample

- AWS Architecture Concept Example



Sample

- Docker Container Batch Example



Sample

- Table Schema / AWS Glue Crawler Example

Schema | Partitions | Indexes

Schema (7)
View and manage the table schema.

Filter schemas

#	Column name	Data type	Partition key
1	userid	string	-
2	lesson	string	-
3	gubun	string	-
4	code	string	-
5	predicted	double	-
6	yyyy	string	Partition (0)
7	mm	string	Partition (1)

Crawler properties

Name	IAM role	Database	State
birt_athena_crawler_HBEdu_App	AWSGlueServiceRole-birt-athena-crawler	birt_athena	READY

Description	Security configuration	Lake Formation configuration	Table prefix
-	-	-	HBEdu_App_

Maximum table threshold
-

▶ Advanced settings

Crawler runs | Schedule | Data sources | Classifiers | Tags

Crawler runs (32)
The list of crawler runs for this crawler.

Stop run | View CloudWatch logs | View run details

Filter data | Filter by a date and time range | 1 2

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
May 3, 2023 at 08:00:51	May 3, 2023 at 08:04:23	03 min 32 s	Completed	0.173	126 table changes, 115 partition changes
April 3, 2023 at 08:00:44	April 3, 2023 at 08:04:26	03 min 42 s	Completed	0.102	108 table changes, 135 partition changes
March 3, 2023 at 08:00:55	March 3, 2023 at 08:03:56	03 min	Completed	0.182	115 table changes, 513 partition changes
February 3, 2023 at 08:00:55	February 3, 2023 at 08:05:23	04 min 28 s	Completed	0.148	111 table changes, 111 partition changes
January 27, 2023 at 06:20:18	January 27, 2023 at 06:21:50	01 min 31 s	Completed	0.151	116 table changes, 221 partition changes
January 3, 2023 at 08:00:55	January 3, 2023 at 08:04:36	03 min 40 s	Completed	0.154	66 table changes, 303 partition changes
December 3, 2022 at 08:00:54	December 3, 2022 at 08:03:36	02 min 41 s	Completed	0.203	116 table changes, 984 partition changes
November 3, 2022 at 08:00:52	November 3, 2022 at 08:03:32	02 min 39 s	Completed	0.115	47 table changes, 315 partition changes

Reference

- 참고 링크
  - aws-cli docs: https://docs.aws.amazon.com/cli/index.html
  - Jenkins docs: https://www.jenkins.io/doc/
  - Docker docs: https://docs.docker.com/

- 추천 도서
  - 업무에 바로쓰는 AWS 입문: http://www.yes24.com/Product/Goods/116626210
  - 그림과 실습으로 배우는 도커 & 쿠버네티스: http://www.yes24.com/Product/Goods/108431011