온라인 쇼핑몰 판매 AWS

과목 : 클라우드 가상화

담당교수 : 노 철 우

학과명 : 컴퓨터 공학과

학번/이름 : 201495004 권우주

201495005 김도영

201495019 김홍렬

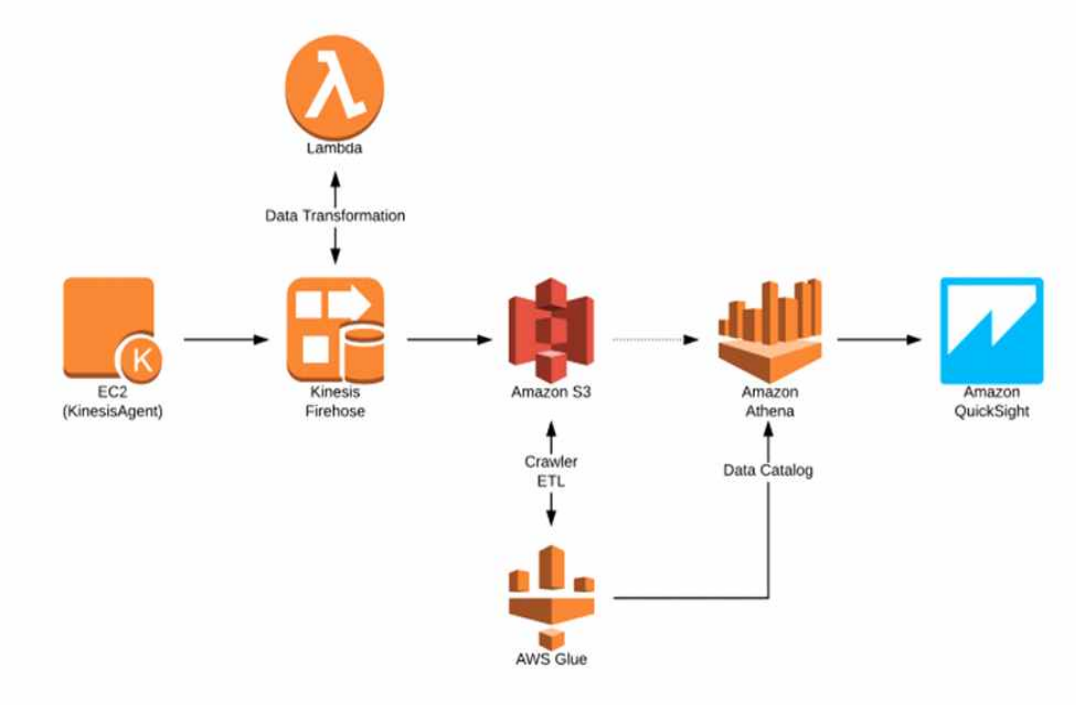
목 차

1. 목적

데이터를 수집 및 분석하는 것은 현대 사회에서 중요한 일이다.
다양한 데이터를 효율적으로 수집하고 분석하고 활용하고 있는 회사는 경쟁 업체에 대비하여 높은 매출을 기록하고 있다. (약 9%)

아마존 상품들을 대상으로 실시간 데이터를 수집 및 분석하고 시각화 해보기

2. 흐름도



Kinesis Agent를 통해 데이터를 발생시키고 Kinesis Firehose 및 S3에 데이터를 수집 및 저장한 후 Glue를 통해 DB를 구축하고 Athena를 통해 분석하고 QuickSight를 통해 데이터를 시각화합니다.

3. 사전준비

(1) key pair 생성

key pair 생성시 주의할 점 : putty를 쓰지않고 cmd에서 openssh를 사용
(ppk이 아닌 pem로 설정)

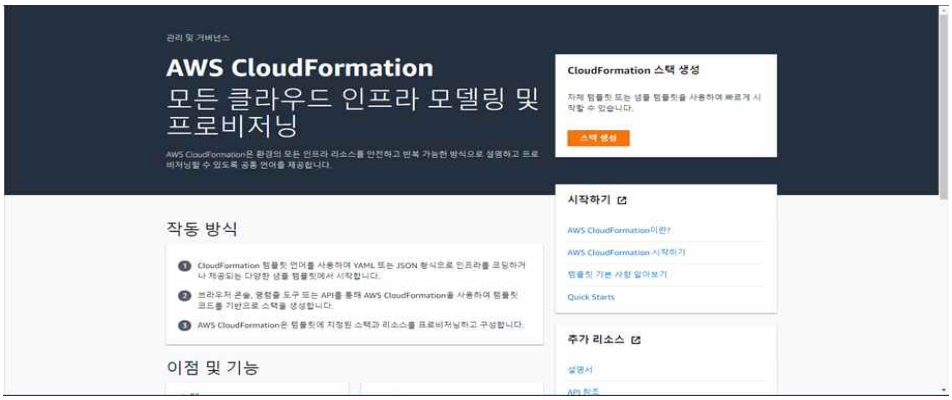
(2) s3 생성

s3는 객체를 저장할 때 사용하여 저장 가능한 객체 수 제한이 없습니다. 또한 99.9%의 내구성 및 암호화(SSE)를 제공합니다.

S3는 윈도우에서 폴더 만들기로 생각하면 됩니다.

우리가 폴더에 여러 가지 문서 및 데이터를 넣듯, S3에 데이터를 저장한다고 생각하면 이해가 편합니다.

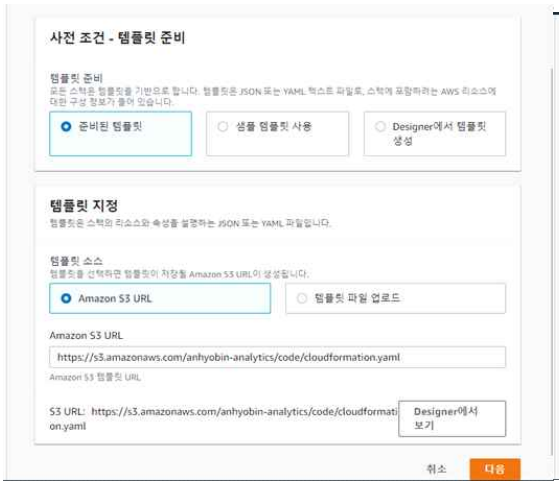
(3) Cloud Formation(Stack) 만들기



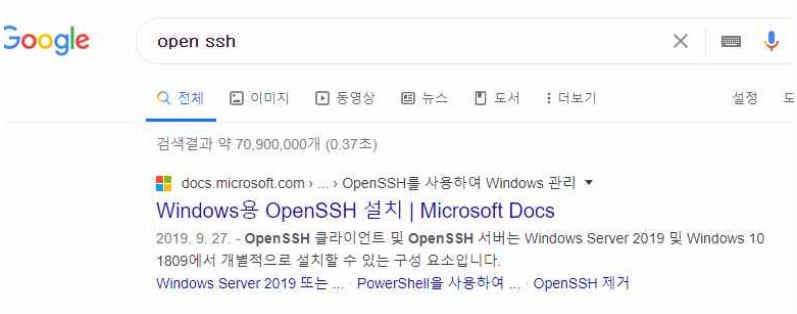
실습에 사용한 IAM 사용자, EC2 인스턴스, Lambda 함수 등은 미리 제공하는 CloudFormation 템플릿을 통해 구성

템플릿이란?
Stack을 구성하는 AWS 리소스를 선언한 것

Stack을 만들어 주면 AWS 리소스를 개별적으로 생성하고 구성할 필요가 없으며 어떤 것이 무엇에 의존하는지 파악할 필요도 없습니다.



(4) SSH 설치

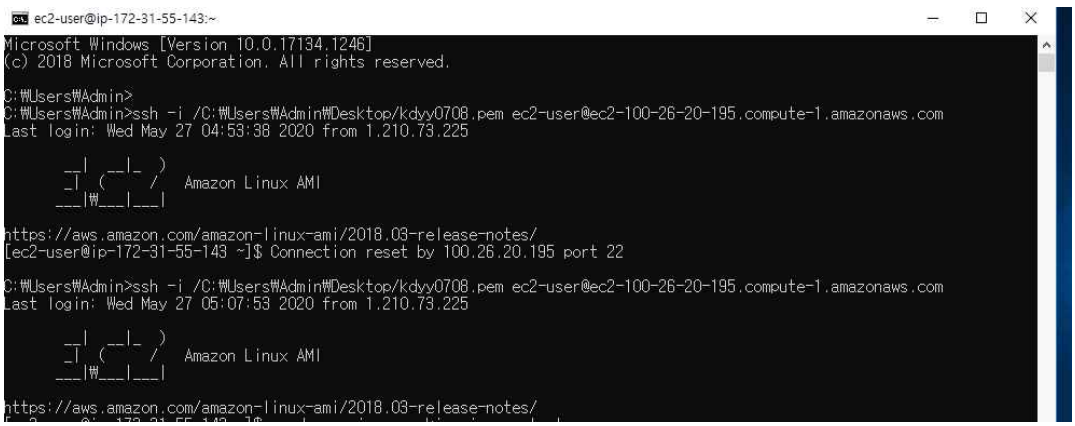


구글에 검색해서 ssh를 설치. **cmd에서 ssh를 사용할 수 있게 됩니다.**

4. 데이터 수집

아마존에서 판매하는 상품을 바탕으로 고객이 구매하는 데이터를 발생시켜 수집 및 저장.

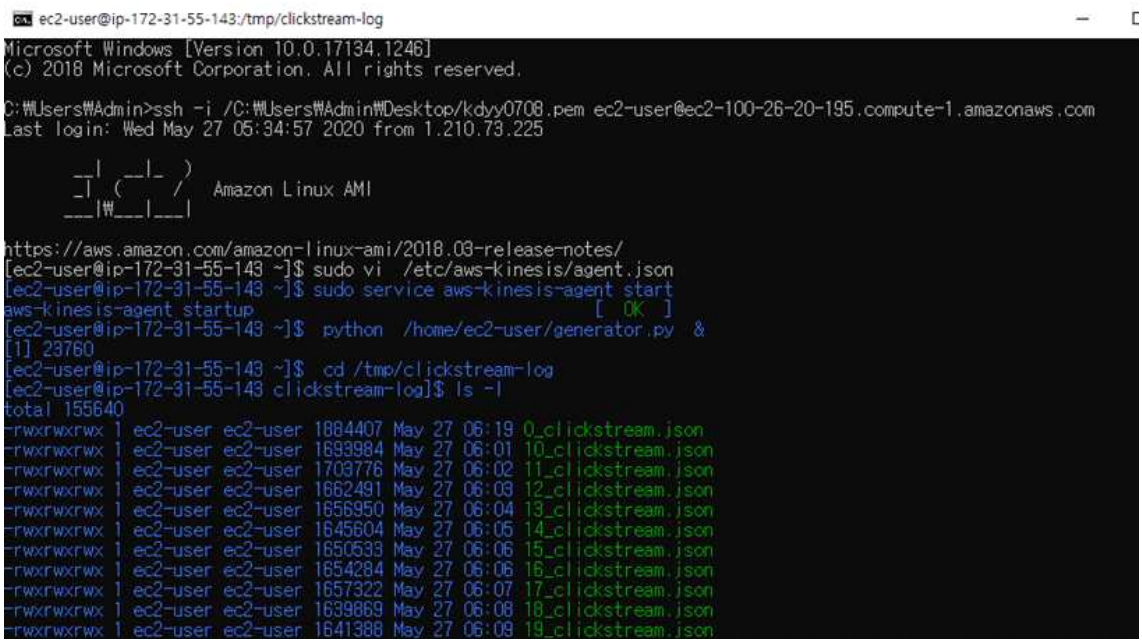
(1) EC2 인스턴스로 접속하기



Stack을 통해 만들어진 EC2가 인스턴스에 들어갑니다.
Ssh –I/key가있는경로/keyname.pem ec2-user@ec2인스턴스의 퍼블릭DNS
현재 바탕화면에 key를 둔 상태(key name=kdyy0708, ec2를 클릭하면 퍼블릭DNS를 얻을 수 있다)

(2) kinesis에 데이터 발생 시키기

kinesis에 임의의 데이터를 발생시킵니다. 그 후 파일을 볼 수 있는 명령어입니다.

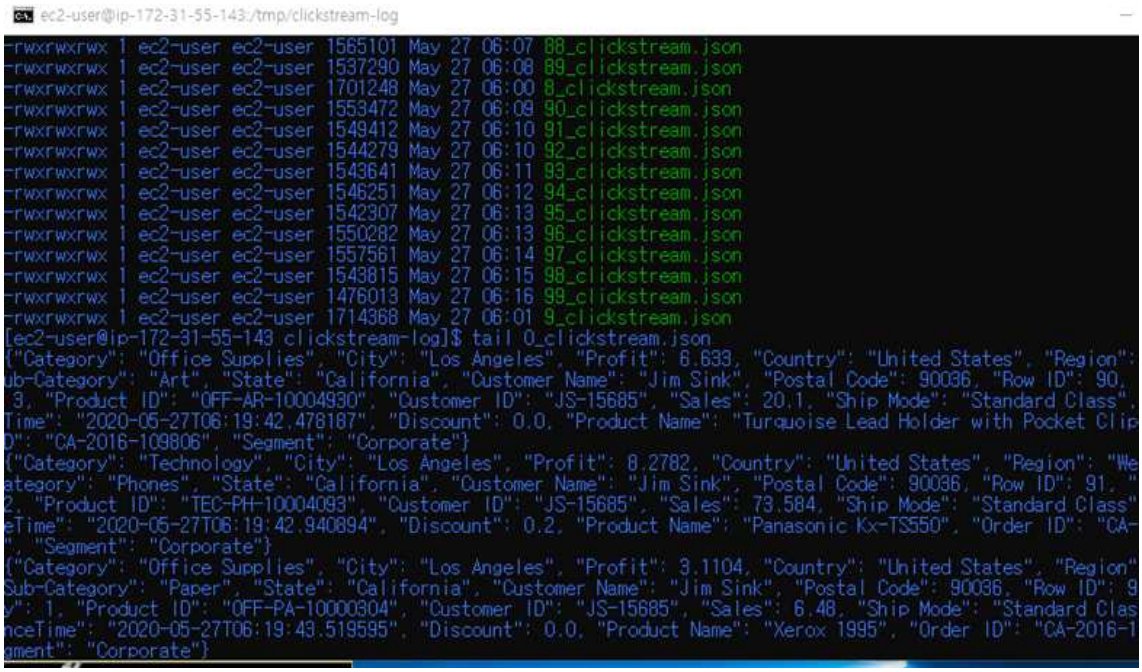


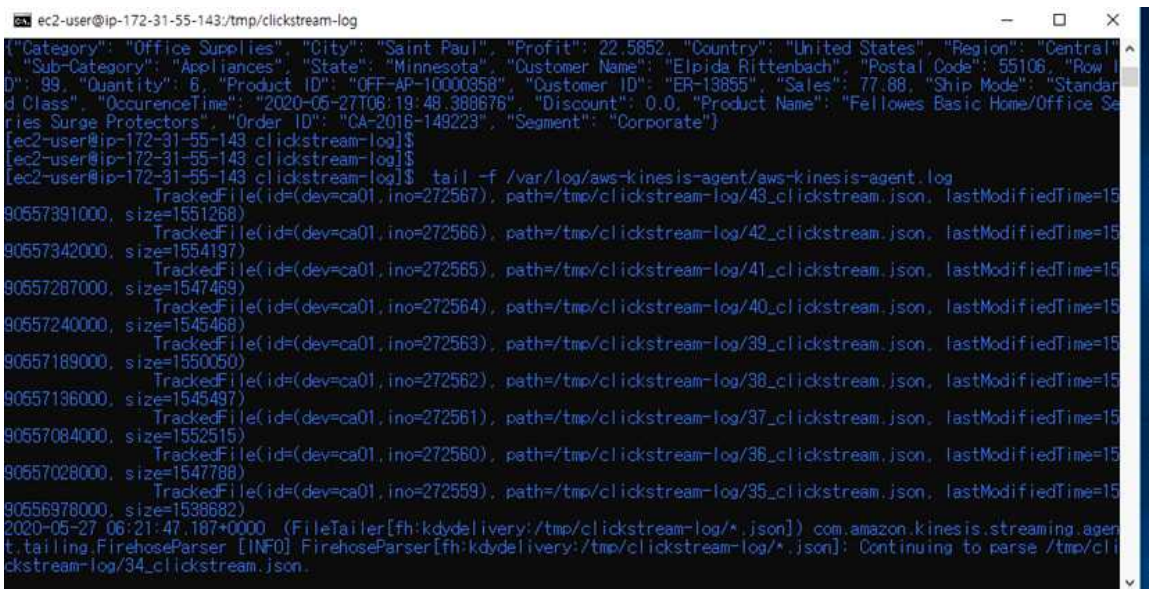
Sudo vi /etc/aws-kinesis-agent start	**데이터 발생 시작**
python /home/ec2-user/generator.py &	**background에서 스크립트를 수집**
cd /tmp/clickstream-log	**파일 들어가기**
ls –ll	**json 파일 내용 보기**
tail 숫자_clickstream.json	**임의의 데이터 보기**

수 많은 clickstream중에서 현재 0_clickstream의 데이터의 내용을 살펴본 사진입니다.

(3) Clickstream 로그 확인 (데이터 전송 확인하기)

Tail –f /var/log/aws-kinesis-agent/aws-kinesis-agent.log
(데이터가 잘 전송되는지 로그를 통해 확인 가능)





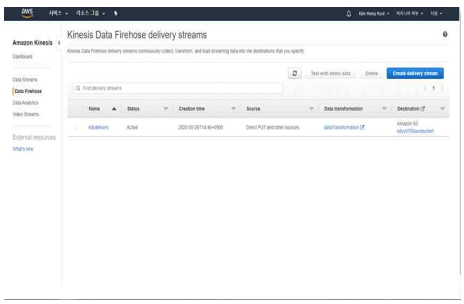
(4) 데이터 수집 및 저장

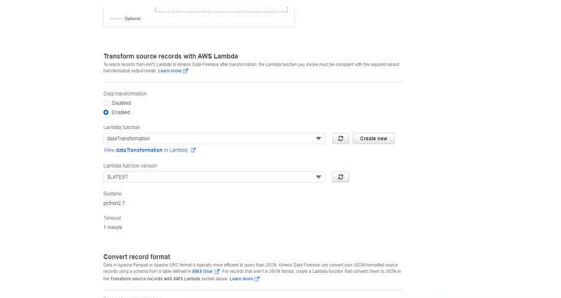
EC2인스턴스에 들어간 뒤 Kinesis Agent를 통해 데이터를 발생시키고 Kinesis Firehose 및 S3에 데이터를 저장합니다.

Kinesis란?
완전 관리형 서비스이며 데이터를 수집할 때 사용
서버리스 방식(S3, ES 등으로 데이터를 로드)

데이터 처리량에 따라 자동으로 확장하는 기능이 있으며 Lambda를 이용한 데이터 전처리가 가능

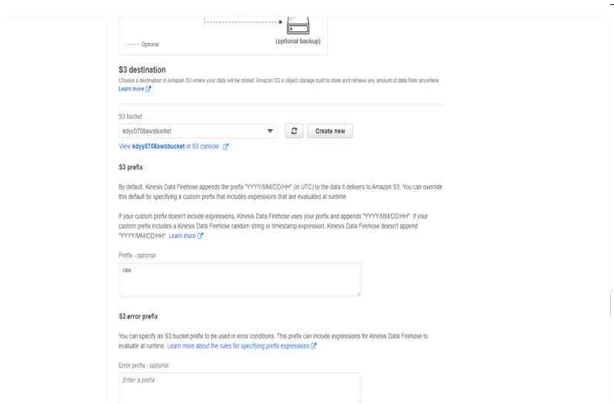
1. Kinesis 만들기

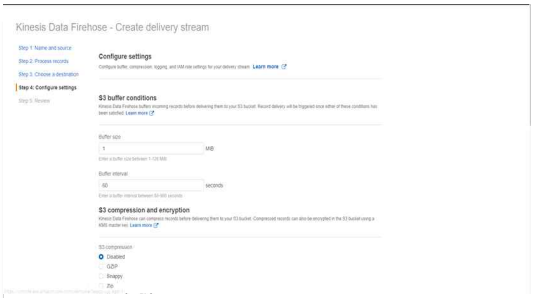




우측 상단에 Create클릭 (파란색) / Kinesis 이름 설정
Lambda데이터 Enabled(활성화) , Data Transformation 활성화

2. 데이터 저장할 버킷 설정 (사전준비에서 만든 버킷으로 설정)



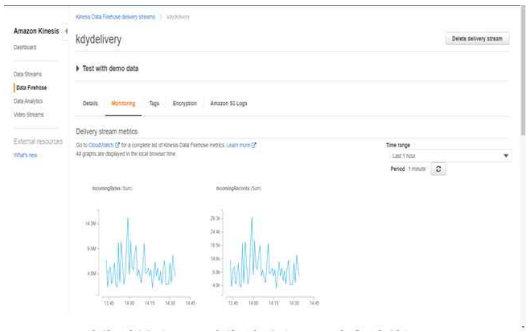


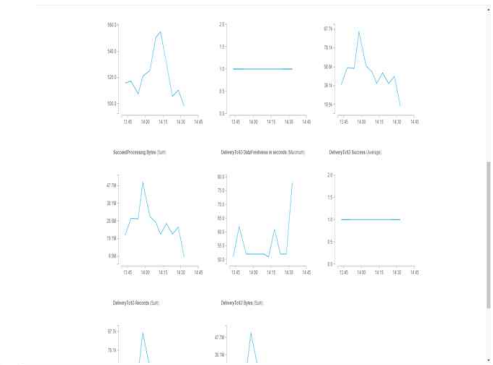
버퍼 사이즈 및 시간 설정

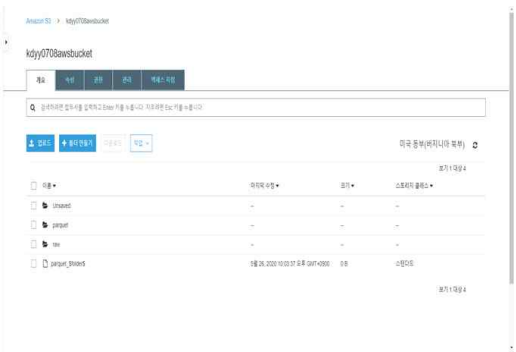
버퍼 사이즈와 시간은 알맞게 해주는 것이 중요 ! (과금주의)

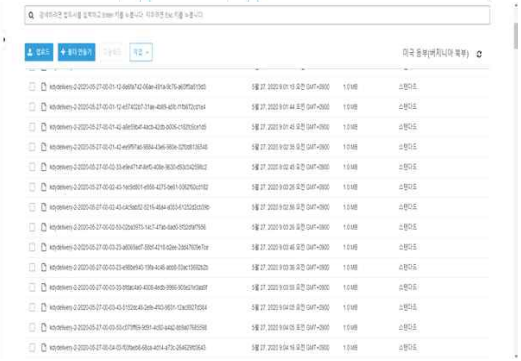
3. 데이터 수집 확인

Kinesis Monitering의 그래프를 통해서 확인하는 방법도 있고 S3 Bucket을 통해 확인









5. 데이터 분석

데이터 ETL(추출, 변환) 작업 (=DB구축)
수집된 데이터를 통해 Glue를 사용하여 DB테이블을 구축

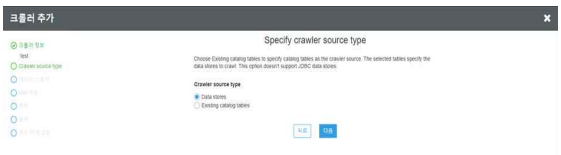
Glue란?
완전 관리형 ETL 서비스가 가능하고 서버리스 형태
AWS에 저장된 데이터를 자동 검색 및 분류하여 빠르게 분석
테이블 및 스키마 등을 Glue 데이터 카탈로그에 저장

Glue 데이터 카탈로그란?
Glue 크롤러를 통해 자동으로 데이터 검색, 스키마를 카탈로그에 저장
Athena에서 즉시 쿼리가 가능해짐과 동시에 데이터 분석이 가능

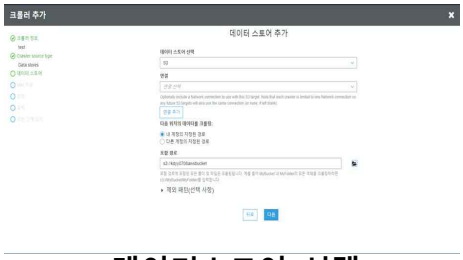
Glue 크롤러란?
분류자 로직을 호출하여 데이터의 스카마, 포맷 및 데이터 유형을 유추
S3에 실시간으로 수집된 데이터를 바탕으로 데이터 유형을 알맞게 유추하여 테이블을 생성

(1) 크롤러 만들기

크롤러 이름 생성

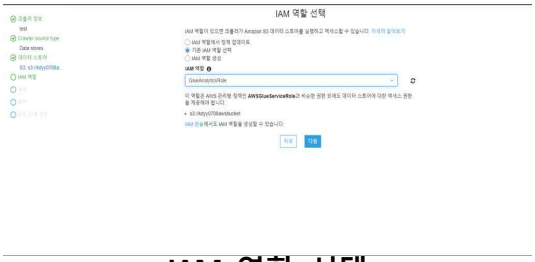


데이터 스토어or테이블 카탈로그 선택



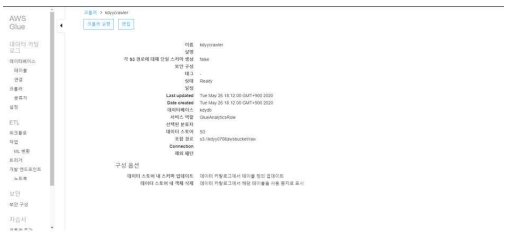
데이터스토어 선택

(데이터를 수집해 둔 S3 bucket으로 설정)

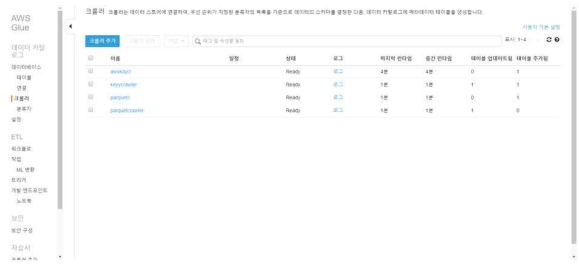


IAM 역할 선택

DB 생성



크롤러 확인 및 크롤러 실행



테이블 추가됨 항목에 1이 되어있음을 확인

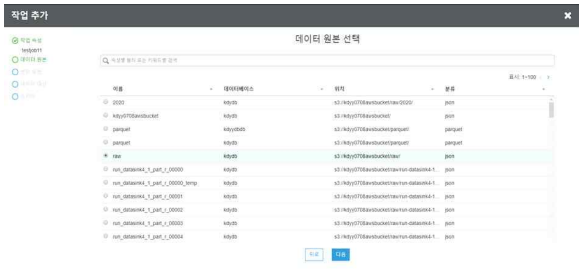
(2) 작업

작업이란?

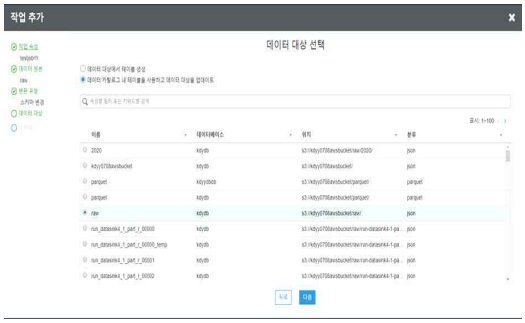
Glue에서는 Python과 Scala 두 가지 언어를 이용해 작성한 ETL 스크립트 수행 가능 (스크립트를 자동 생성하고 테스트하고 실행하는 것도 가능)

Glue Job을 이용하여 S3의 데이터 타입인 Json 데이터를 Parquet 포맷으로 변환가능 , Json에 비해서 Parquet의 비용이 저렴

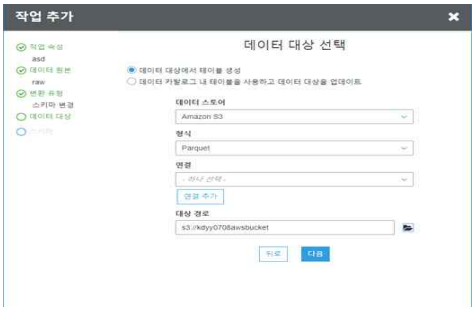
작업 추가하기



데이터 선택1

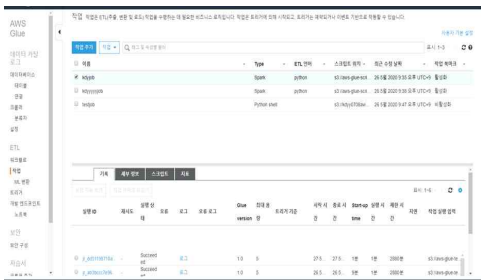


데이터 선택2



테이블로 만들 데이터 대상 및 형식선택

데이터 형식(타입)을 알맞게 매핑



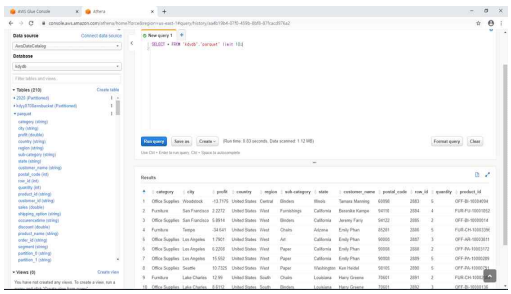
작업 실행 시킨 후
Succeeded가 뜨면 실행완료

(3) Amazon Athena

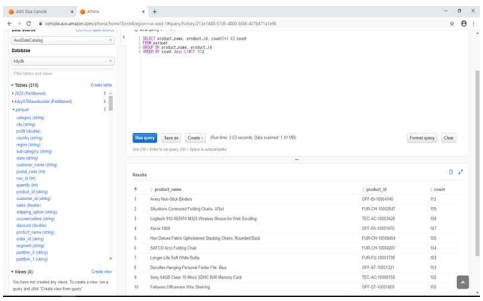
SQL을 사용해 실시간으로 수집한 데이터를 쿼리
S3에 저장된 데이터를 간편하게 분석할 수 있는
대화식 쿼리 서비스 제공

Amazon Athena의 장점?
서버리스 서비스 제공
데이터 분석 가능
데이터를 따로 들고 올 필요없이 S3에 저장된 데이터를 사용

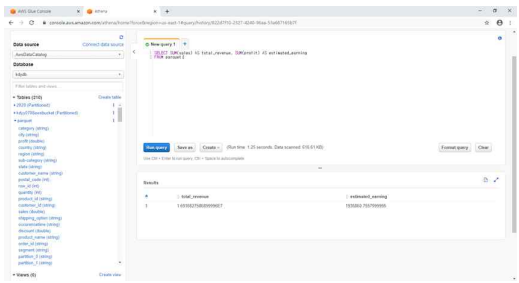
(1) 아테나 사용



10개의 상품보기



상품 이름 및 상품코드와 개수를 나타내는 쿼리
그룹화 및 정렬



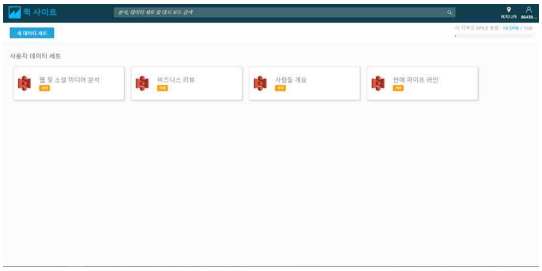
매출 및 이익의 합계

6. 데이터 시각화

Amazon QuickSight

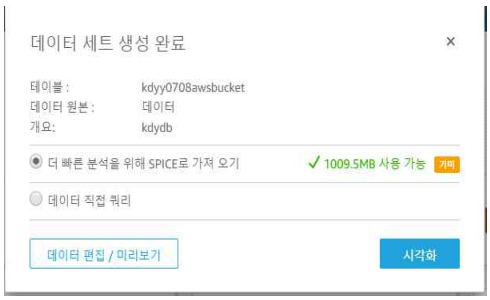
쿼리를 통해 분석한 데이터 시각화
S3, Athena, RDS, Redshift와 같은 AWS 서비스의 데이터를 빠르게 시각화 가능

새 분석 클릭 !



새 데이터 세트 클릭 !

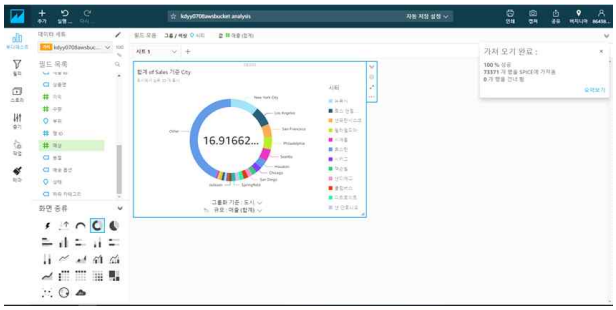
kdydb에 있는 임의의 데이터 파일 선택



데이터 세트 생성완료

도시별 이익의 합계

도시별 할인율의 합계



도시별 매출 합계

7. 개선 및 활용 방안

1. 판매 데이터를 이용한 지욕 물량 발주 예측

2. 지역별 베스트 상품 코너 증설 및 이벤트 생성

3. 지역별 상품 연관성을 이용한 묶어 팔기

4. 이익이 낮은 지역 보완 전략

8. 각자 담당 파트

김홍렬 - key pair 생성 ~ Clickstream 로그 확인 (데이터 전송 확인하기)

권우주 - 데이터 수집 및 저장 ~ 데이터 수집 확인

김도영 - 데이터 분석 ~ 데이터 시각화