

**SeUng Jung**

**Data Life Cycle on AWS**

*Lab 3-1. Glue Data Catalog for Analytics*

**Table of Contents**

Lab 설명 3

Lab Architecture 4

Glue Data Catalog for Analytics 5

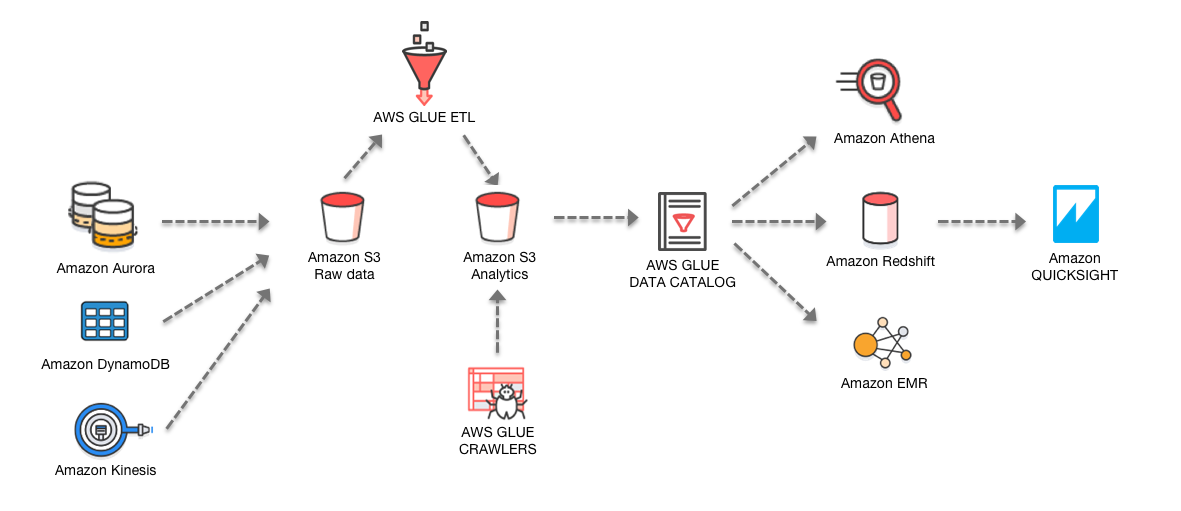
# Lab 설명

이번 Lab은 다양한 데이터 정제 작업을 통해 준비된 Analytics 용 S3에 저장된 데이터를 통해 모든 AWS의 데이터 분석 도구를 통해 사용할 수 있는 단일된 뷰인 Data Catalog를 생성합니다.

Glue를 통해 생성하는 Data Catalog는 이후 분석을 위해서 Athena, Redshift, EMR 에서 손쉽게 Query 할 수 있습니다.

# Lab Architecture

데이터 분석을 위한 Data Catalog를 생성합니다. S3 분석용 버킷에 저장된 데이터는 Athena, Redshift, EMR을 통해 쉽게 조회 / 분석이 가능합니다.



단계별 상세 Lab은 다음과 같습니다.

Lab 2-1. RDS to S3 with Glue

Lab 2-2. DDB to S3

Lab 2-3. Data Transforming with Glue

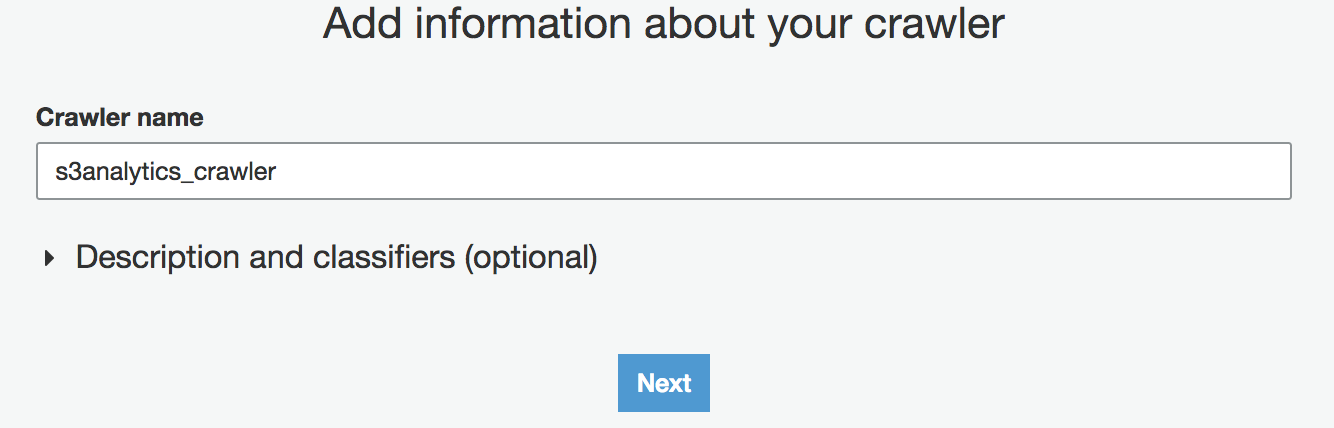
Lab 2-4. ETL to Redshift with Glue

**Lab 2-5. Glue Data Catalog for Analytics**

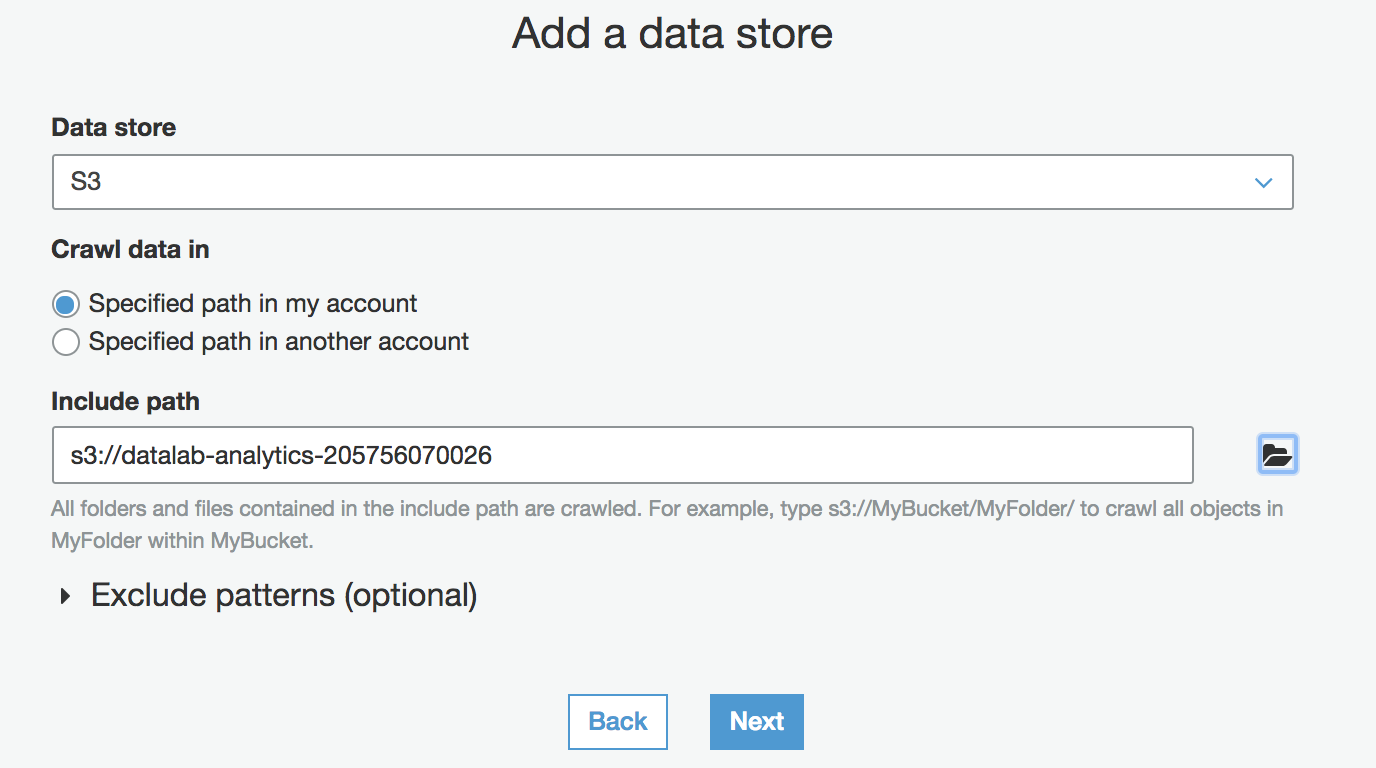
Lab 2-6. Query with Athena and Redshift spectrum

# Glue Data Catalog for Analytics

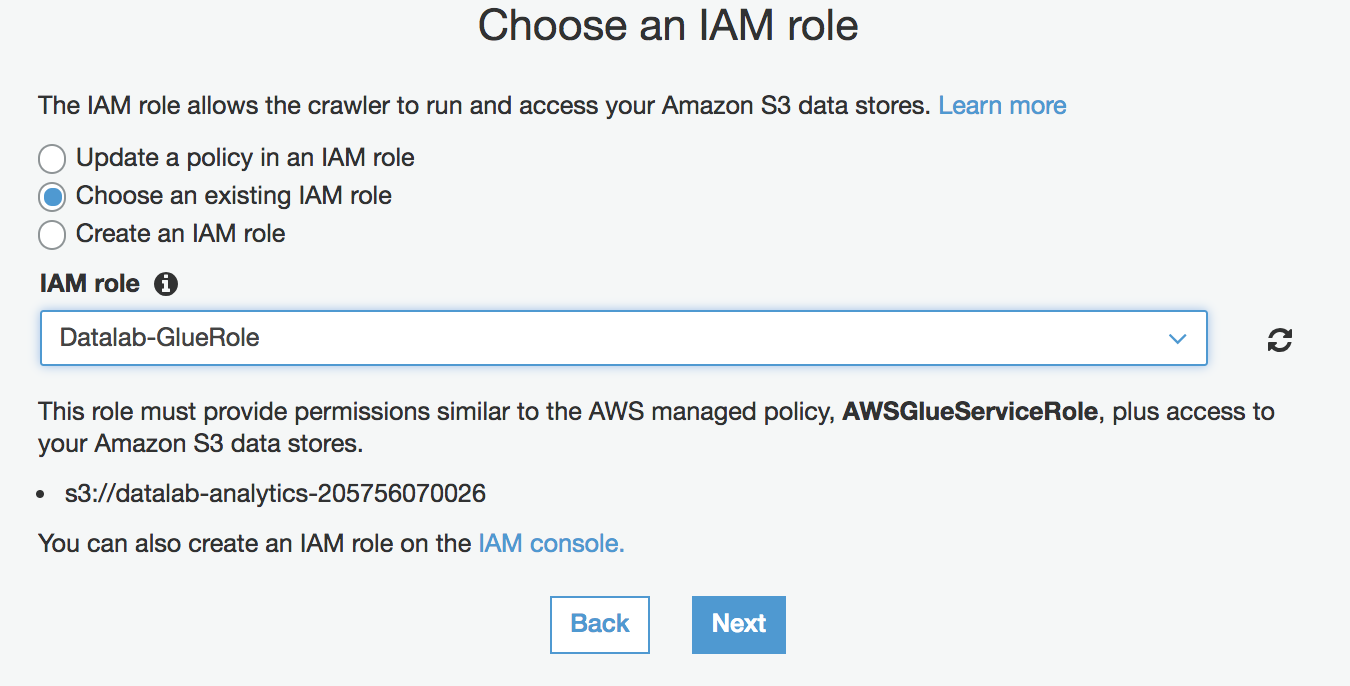
1. 우측 상단의 리전을 **Oregon (US-WEST)**으로 선택합니다.
2. AWS Management Console에서 AWS Glue 서비스 화면으로 이동, Crawler를 생성합니다.
3. [Add Crawler]를 선택하고 Crawler name은 **s3analytics\_crawler**를 입력합니다.



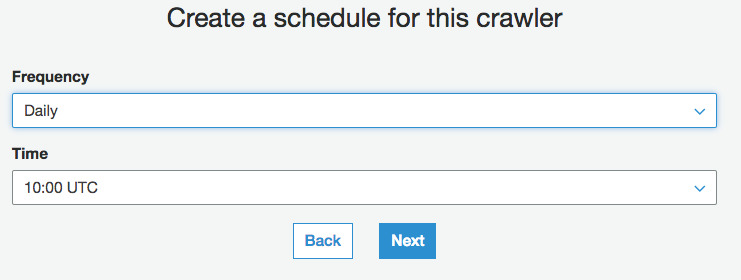
1. 타겟 Data store로 S3를 지정합니다.



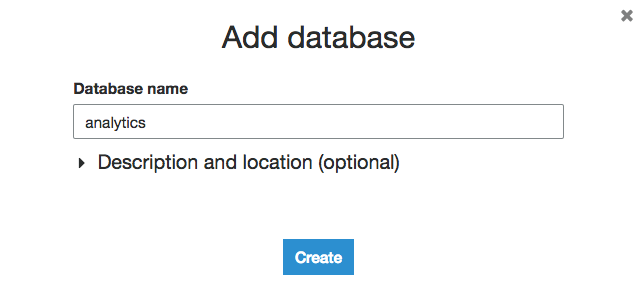
1. IAM Role에서 **[Datalab-GlueRole]**을 선택합니다.



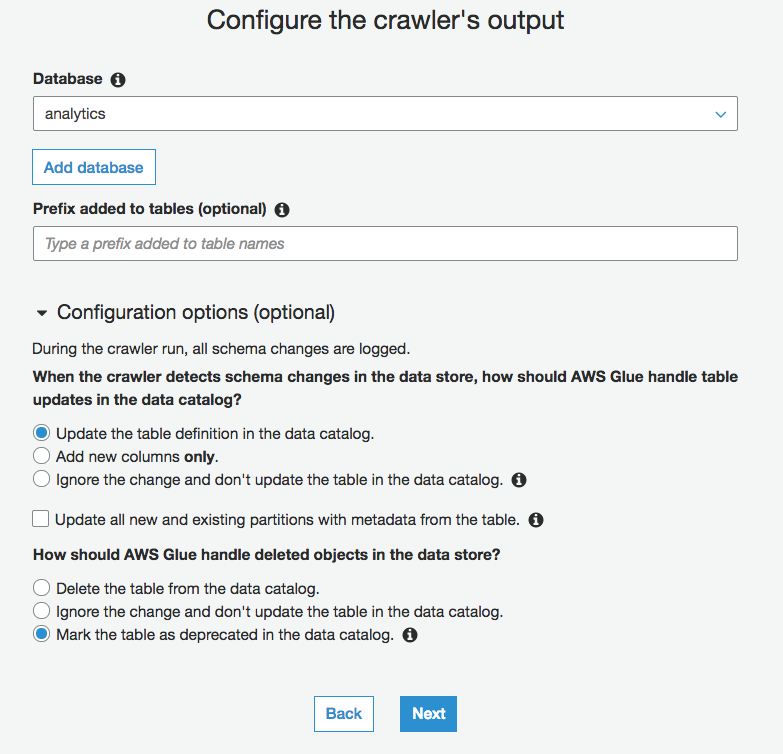
1. 일단위로 스케줄을 설정합니다. Frequency는 **[Daily]**를 선택합니다.



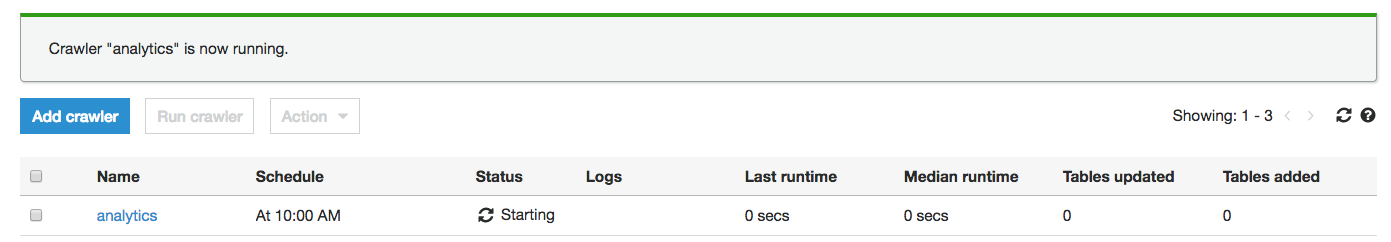
1. **[Add database]**을 누르고Database명은 이후 코드 실행을 고려하여 반드시 analytics로 설정하시기 바랍니다.



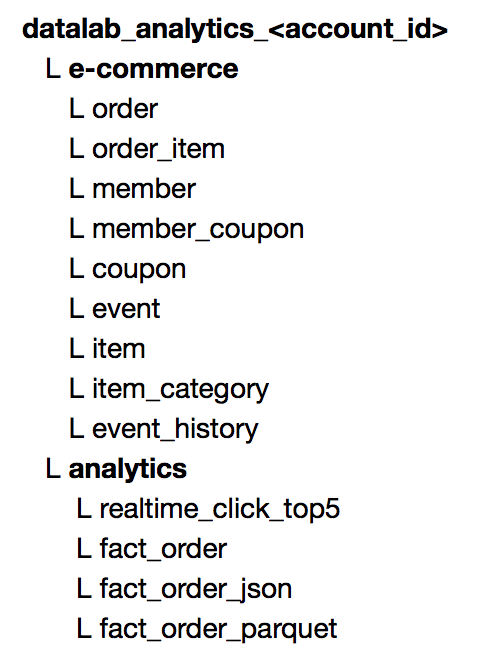
1. 스키마 업데이트 설정은 기본 값을 사용합니다.



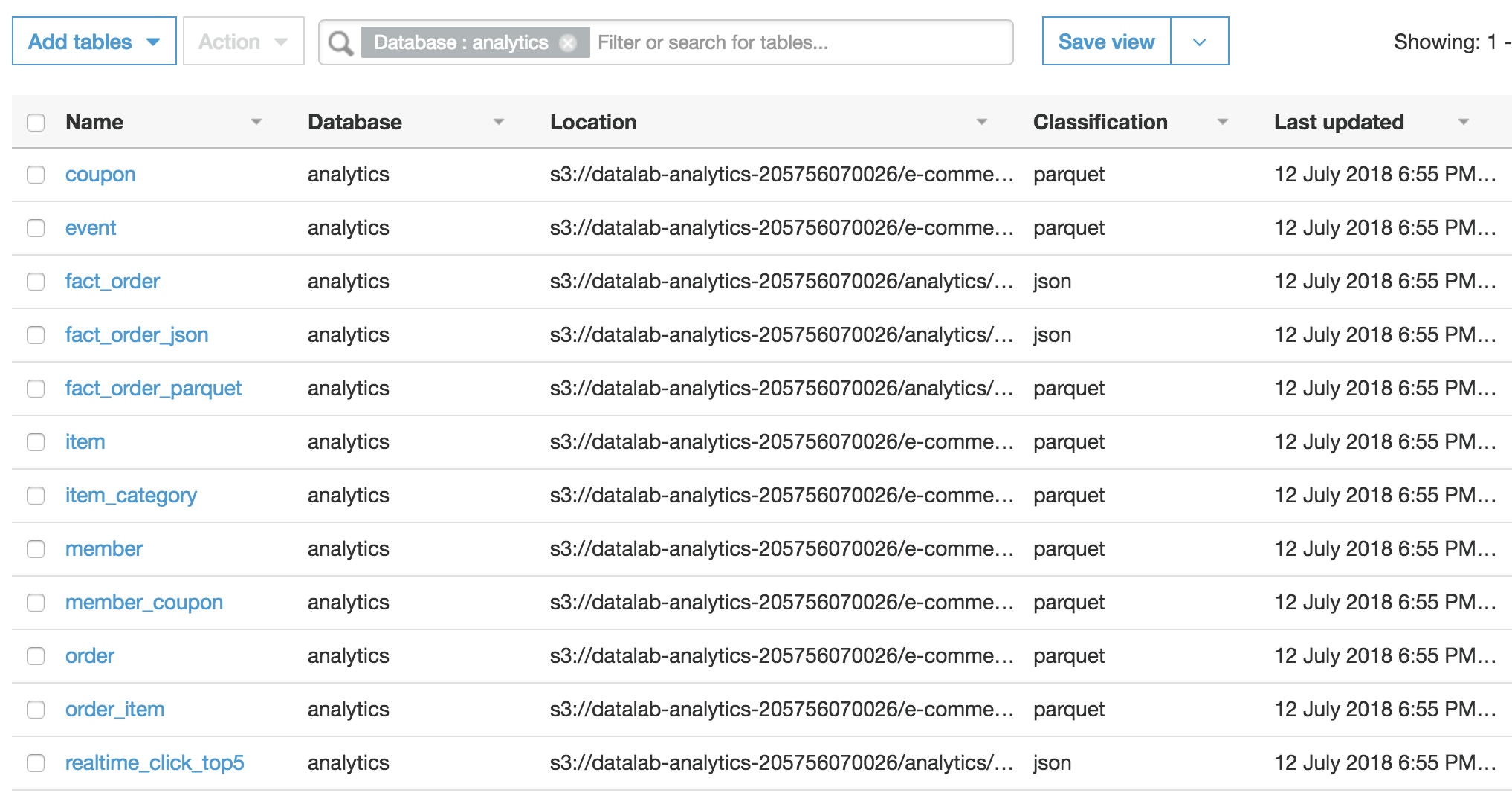
1. 생성된 크롤러를 실행합니다.



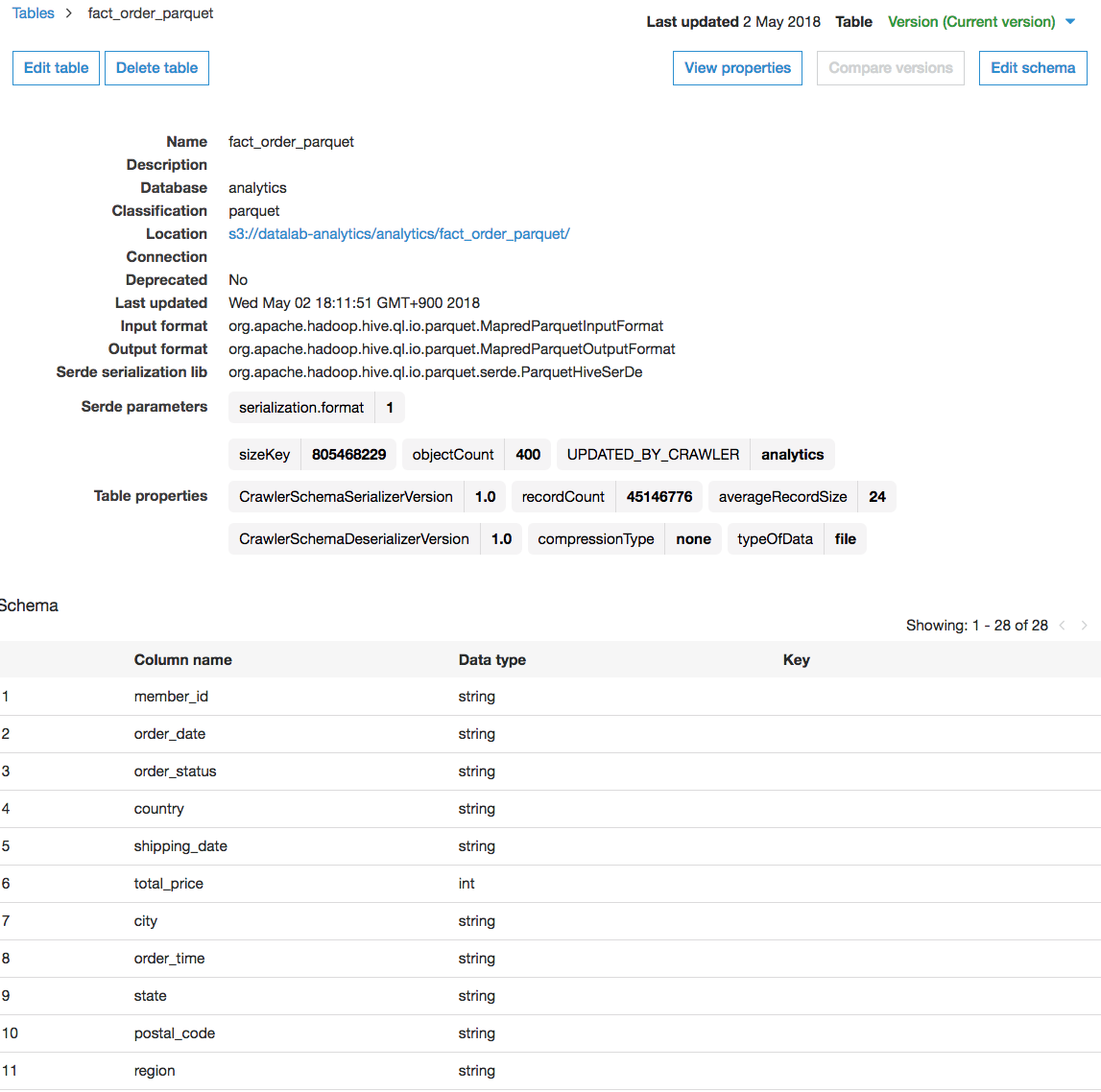
1. 해당 S3 버킷의 개략적인 구조는 다음과 같습니다.



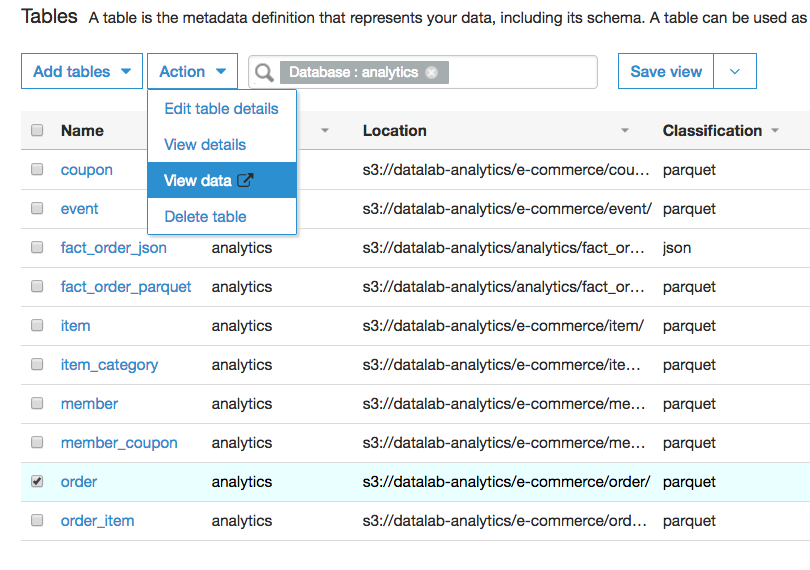
1. Glue Database에서 **[analytics]**를 선택하면 다음과 같이 테이블 리스트를 볼 수 있습니다.

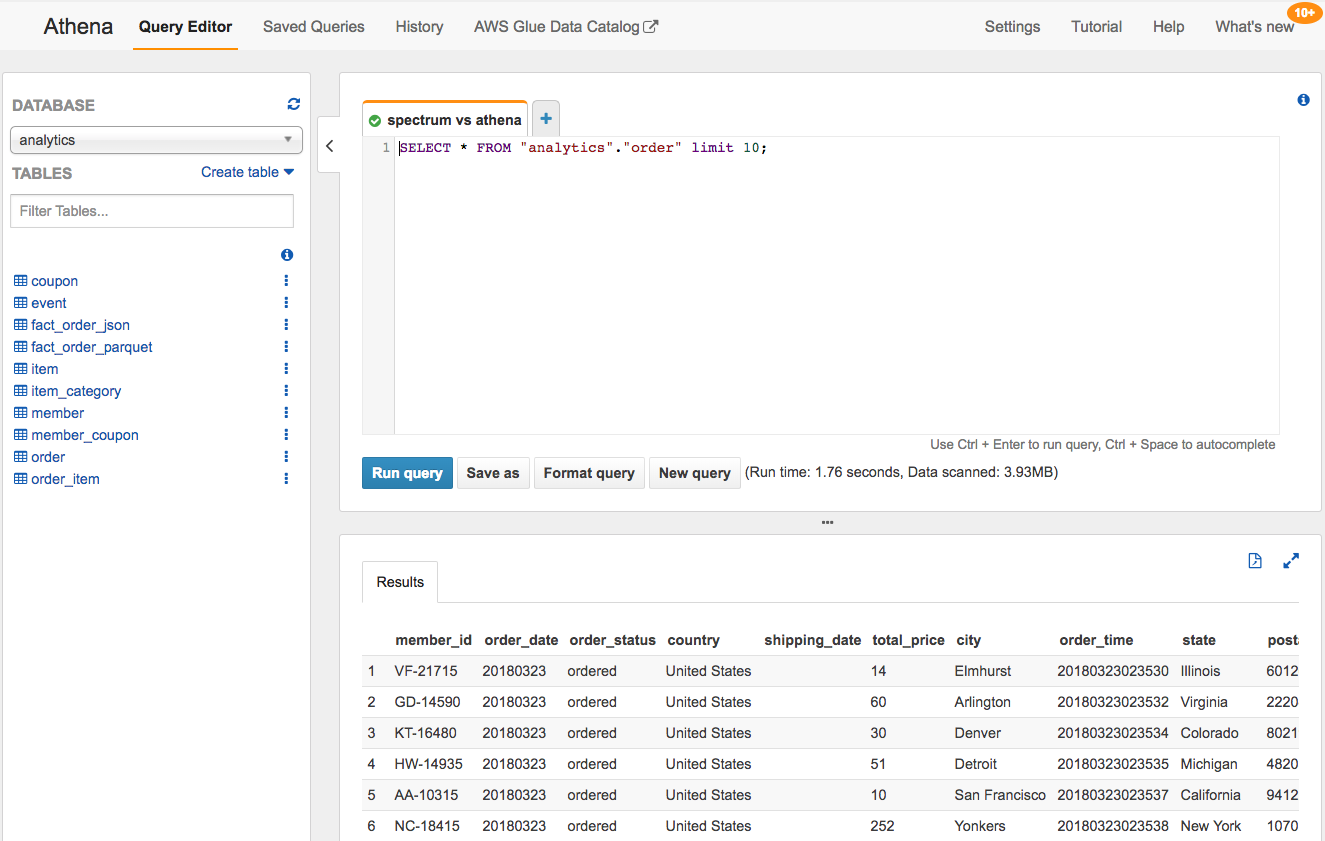


1. 테이블을 선택하면 해당 테이블의 일반적인 특성과 컬럼명, 데이터 타입등을 확인 가능합니다.



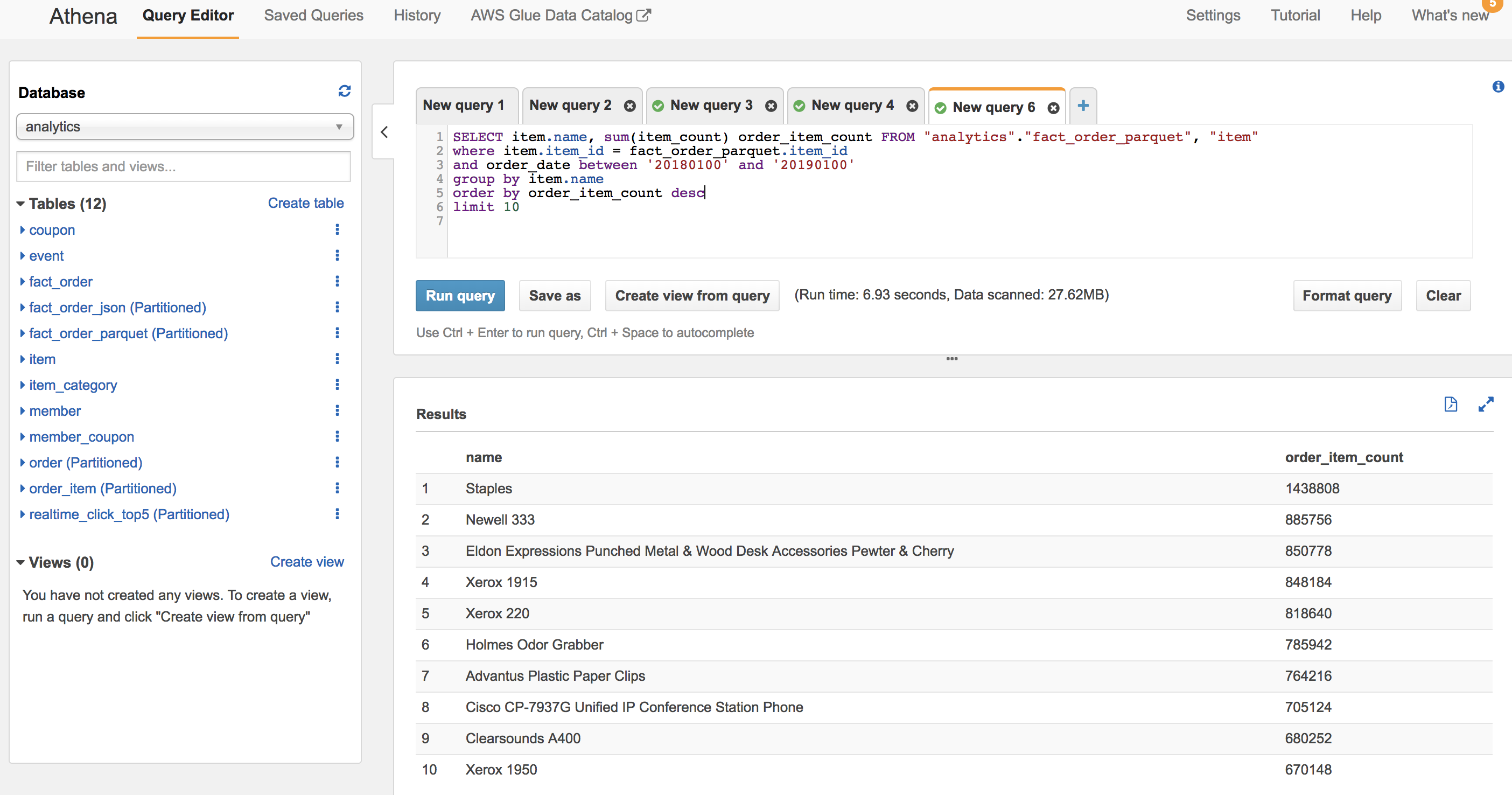
1. 테이블 리스트에서 하나의 테이블을 선택하고, **[View Data]**를 누르면 Athena 화면에서 데이터를 바로 조회 할 수 있습니다.





1. 모든 주문 데이터를 통합한 **fact\_order\_parqeut** 테이블을 통해 먼저 올해 한해 동안 가장 많이 판매된 상품이 무엇인지 조회해 봅니다.

|  |
| --- |
| SELECT item.name, sum(item\_count) order\_item\_count FROM "analytics"."fact\_order\_parquet", "item"  where item.item\_id = fact\_order\_parquet.item\_id  and order\_date between '20180100' and '20190100'  group by item.name  order by order\_item\_count desc  limit 10; |



1. Kinesis Analytics를 통해서 수집 / 처리된 Top5 클릭 상품의 리스트도 바로 조회가 가능합니다. **realtime\_click\_top5**테이블을 통해 올해 가장 많이 클릭 Top5에 선택된 상품의 클릭 수를 쿼리해봅니다.

|  |
| --- |
| SELECT item, sum(item\_count) FROM "analytics"."realtime\_click\_top5"  where partition\_0 = '2019'  group by item  order by 2 desc  limit 10; |

