Fatih Sultan Mehmet Vakıf Üniversitesi

YouTube Veri Analizi

Veri Madenciliğinde Özel Konular

Dr. Öğr. Üyesi Zeki KUŞ

Final Projesi

Reyhan HOŞAVCI 231221001

06.01.2025

Kullanılan Teknolojiler

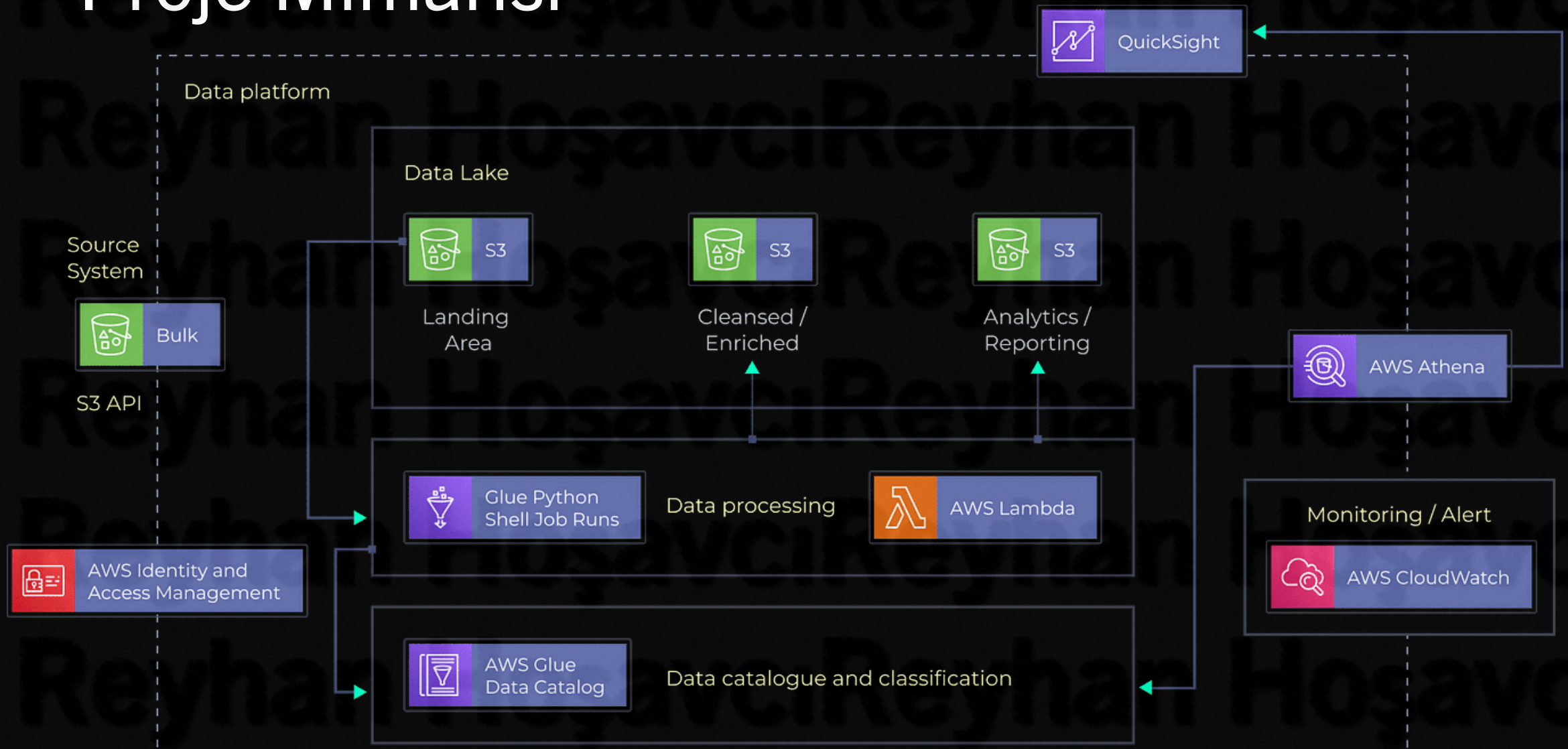
Bu projede AWS'nin bulut tabanlı hizmetlerinden yararlanıldı:

- **AWS Lambda:** Veri işleme ve dönüştürme işlemleri
- **AWS S3:** Yedekleme, arşivleme ve veri paylaşımı
- **AWS Glue:** ETL işlemleri, Veri cataloglama ve sınıflandırma
- **AWS Athena:** SQL tabanlı sorgularla verilerin analizi
- **Amazon QuickSight:** Verilerin görselleştirilmesi ve raporlanması
- **AWS IAM (Identity and Access Management):** Erişim kontrolü ve güvenliği
- **AWS CloudWatch**: İzleme ve uyarı mekanizmaları

Programlama dili : Python

Kaynak video:https://www.youtube.com/watch?v=yZKJFKu49Dk

Proje Mimarisi



Mimari diagram: https://www.projectpro.io/project-use-case/aws-etl-data-pipeline-python-example

Kullanılan Veri Seti

- Kaggle
- YouTube'da günlük trend olan videolara ait birkaç aylık veriyi içerir
- Veri bölgeleri:
  - **US, GB**, DE, **CA**, FR (ABD, İngiltere, Almanya, Kanada, Fransa)
  - RU, MX, KR, JP, IN (Rusya, Meksika, Güney Kore, Japonya, Hindistan)
- video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count
- Her bölge için farklılık gösteren **category_id** alanı bulunuyor. Bu alana JSON dosyalarından erişiliyor.

CA_category_id.json (7.91 kB)


```
"root" : { 3 items
  "kind" : string "youtube#videoCategoryListResponse"
  "etag" : string ""ld9biNPKjAjgjV7EZ4EKeEGrhao/1v2mrzYSYG6onNLt2qTj13hkQZk""
  "items" : [ 31 items
    0 : { 4 items
      "kind" : string "youtube#videoCategory"
      "etag" : string ""ld9biNPKjAjgjV7EZ4EKeEGrhao/Xy1mB4_yLrHy_BmKmPBggty2mZQ""
      "id" : string "1"
      "snippet" : { 3 items
        "channelId" : string "UCBR8-60-B28hp2BmDPdntcQ"
        "title" : string "Film & Animation"
        "assignable" : bool true
      }
    }
```

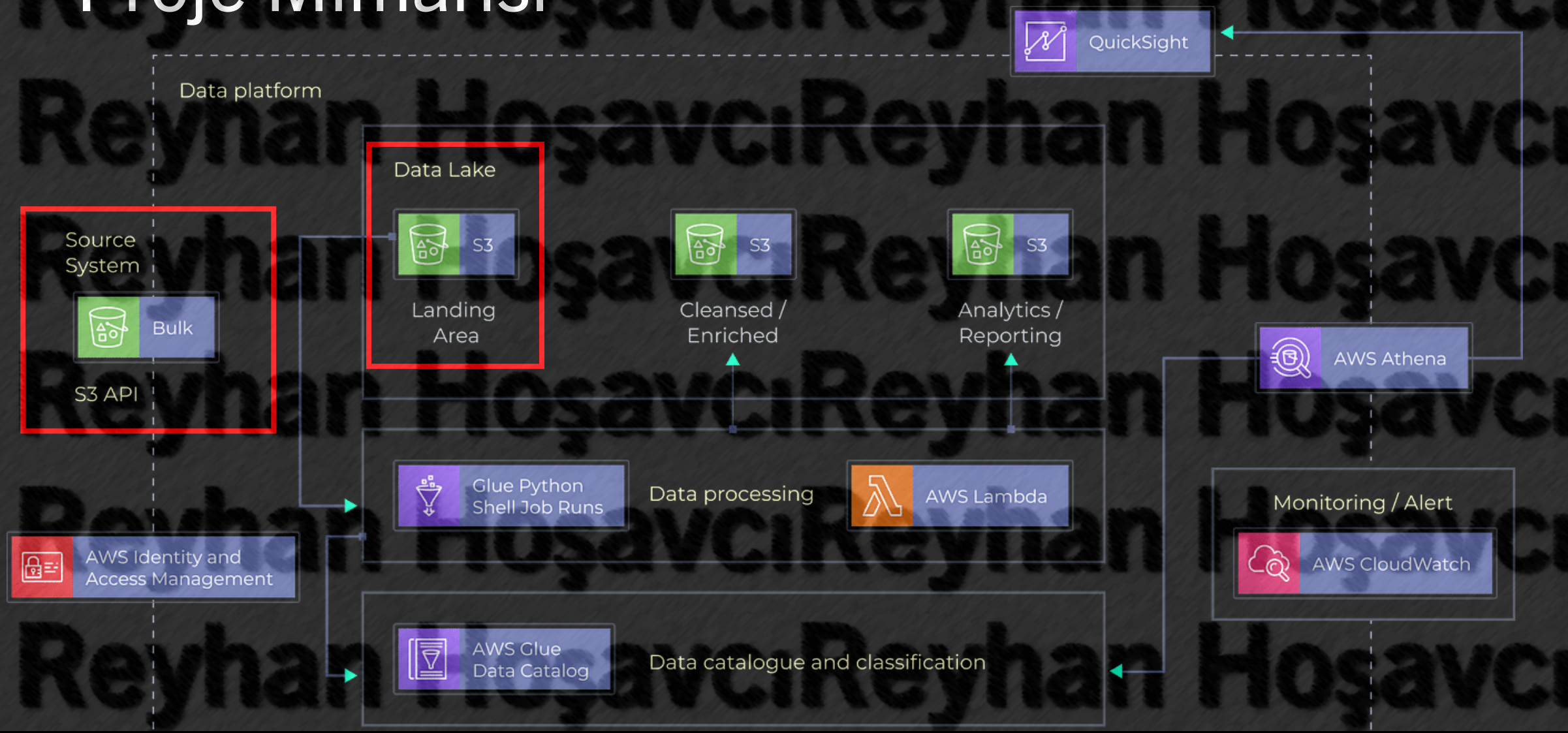

CAvideos.csv (64.07 MB)

Detail | Compact | Column — 8 of 16 columns

video_id	title	channel_title	category_id	publish_time	views	likes
n1WpP7iowLc	Eminem - Walk On Water (Audio) ft. Beyoncé	EminemVEVO	10	2017-11-10T17:00:03.000Z	17158579	787425
0dBIkQ4Mz1M	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	2017-11-13T17:00:00.000Z	1014651	127794
5qpjK5DgCt4	Racist Superman | Rudy Mancuso, King Bach & Lele Pons	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	3191434	146035
d380meD0W0M	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000	2095828	132239

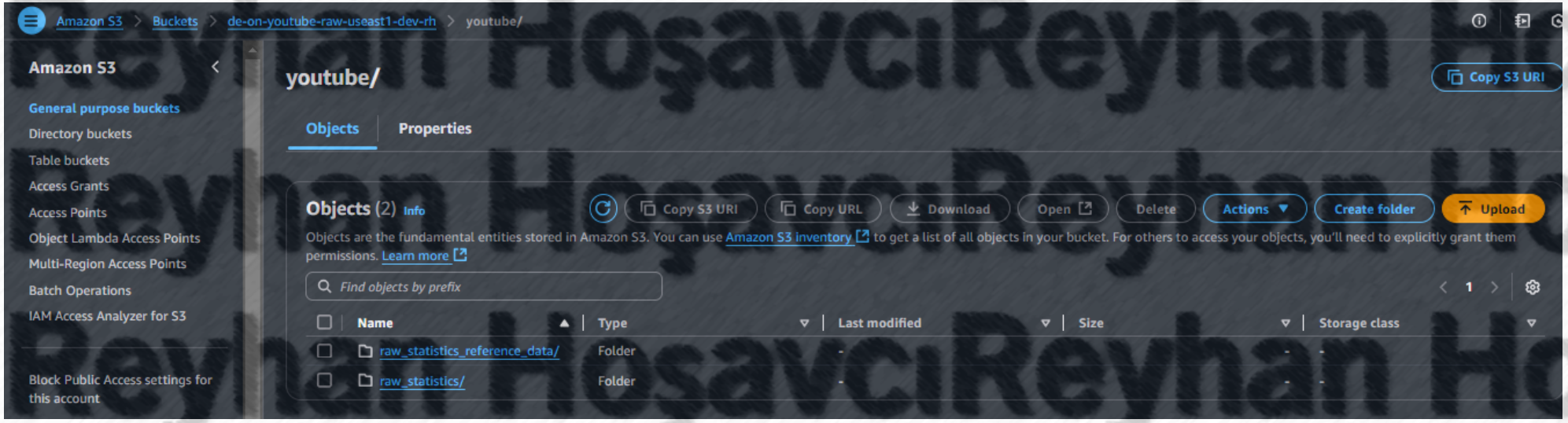
JSON ve CSV formatındaki verilerin S3 Bucket üzerinde toplanması

Proje Mimarisi

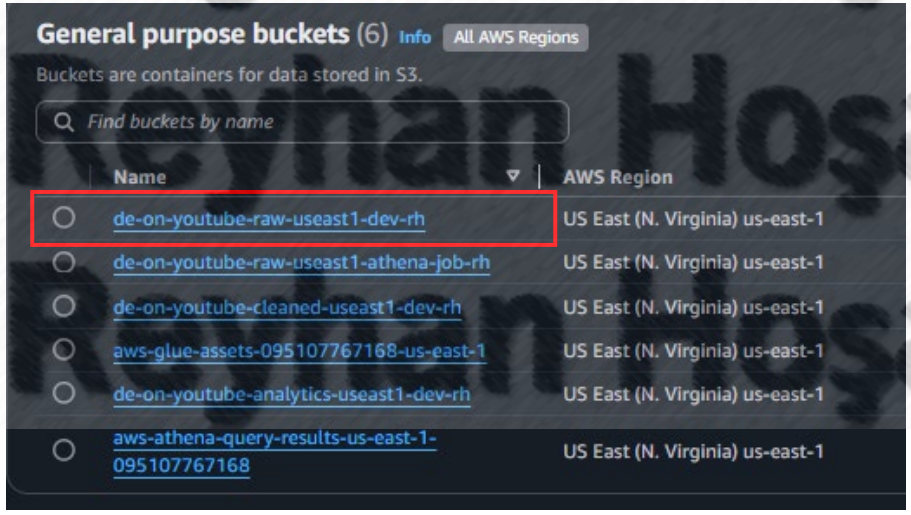


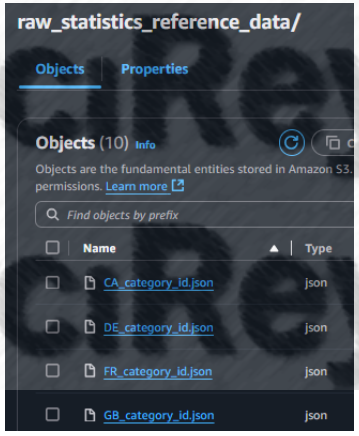
Mimari diagram: https://www.projectpro.io/project-use-case/aws-etl-data-pipeline-python-example

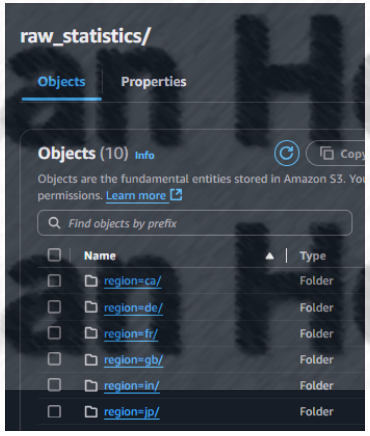
JSON ve CSV formatındaki verilerin S3 (Simple Storage Service) Bucket üzerinde toplanması



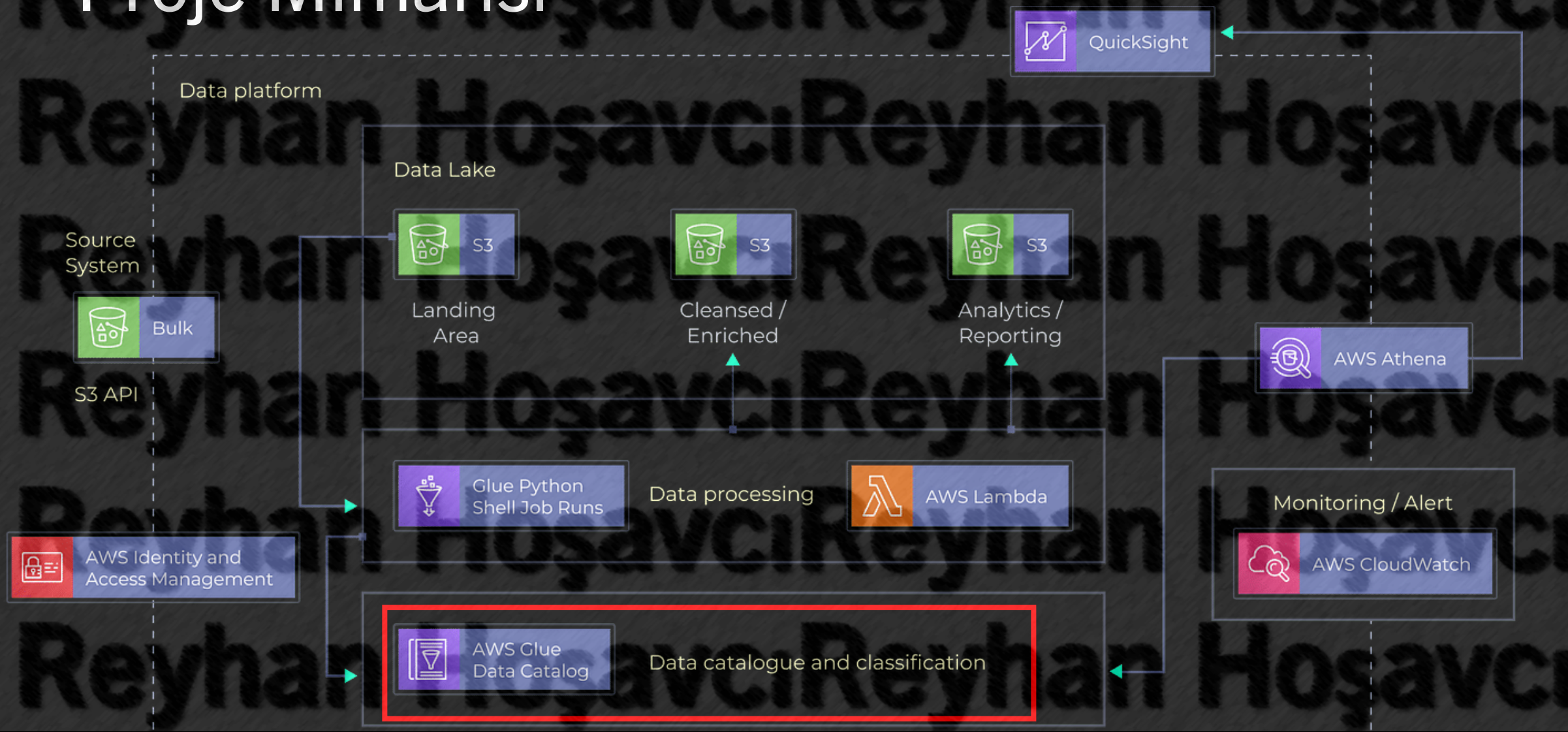
S3, verilerin organize edilmesi ve Data Lake katmanları arasında taşınması için kullanıldı.





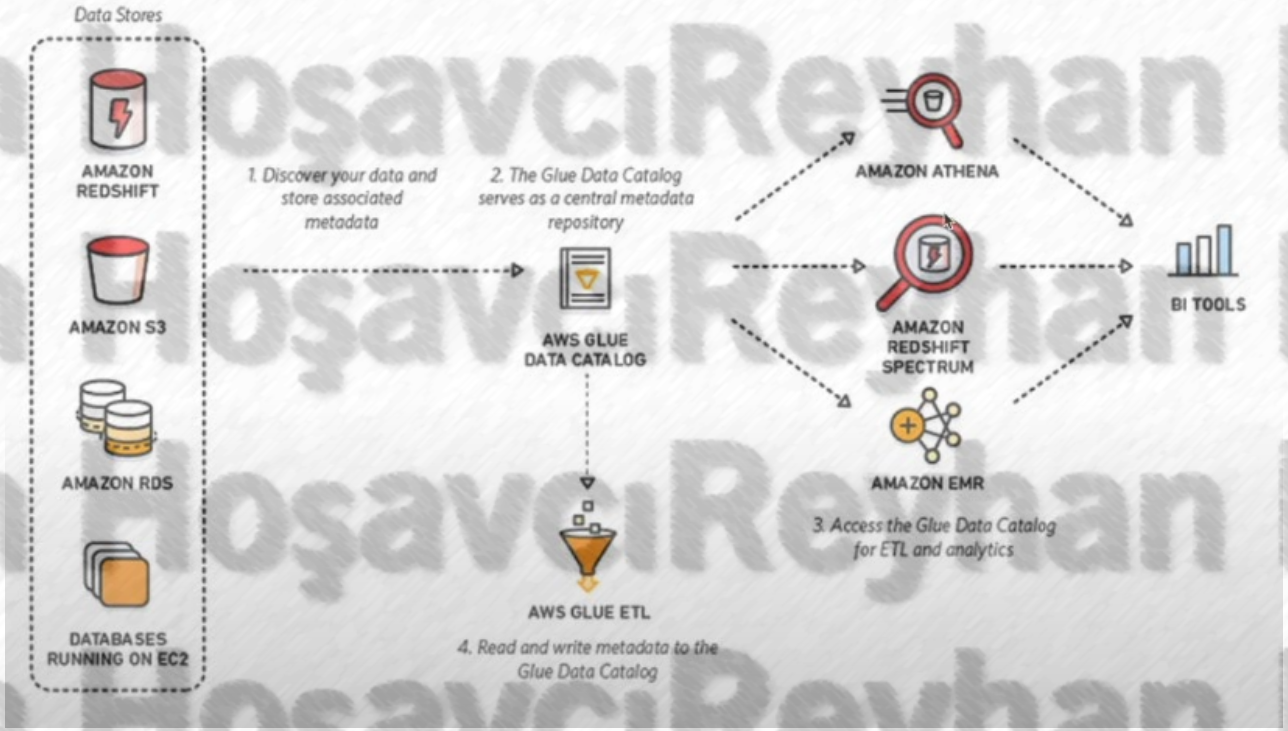


Proje Mimarisi



Mimari diagram: https://www.projectpro.io/project-use-case/aws-etl-data-pipeline-python-example

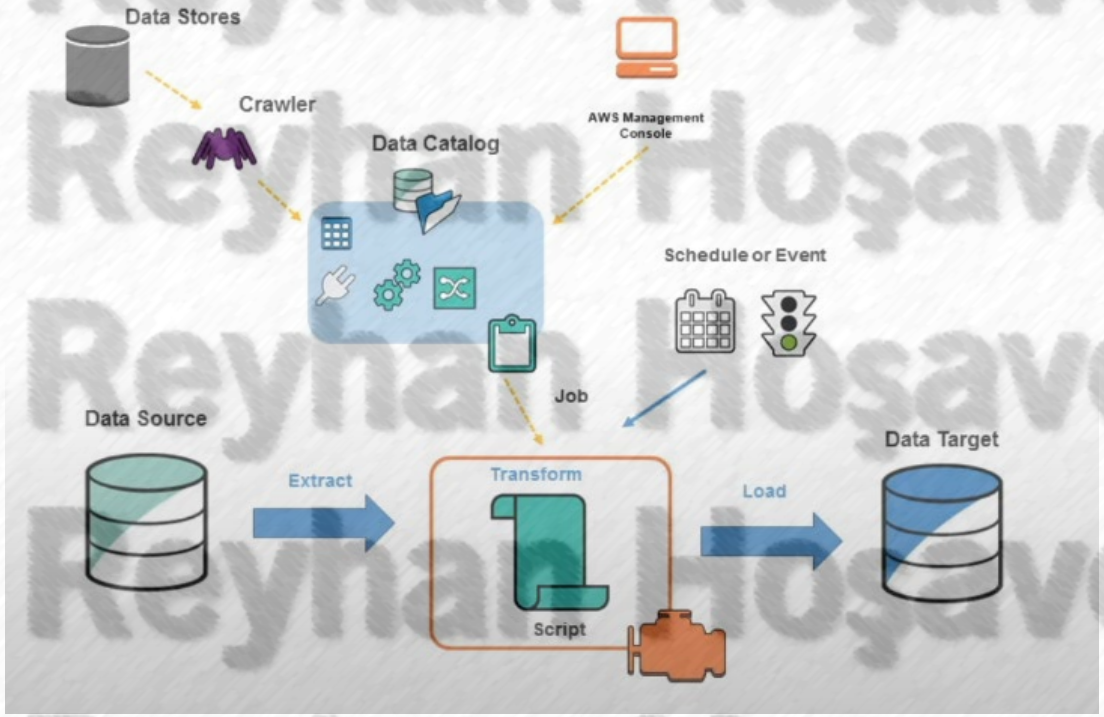
AWS Glue Catalog oluşturulması

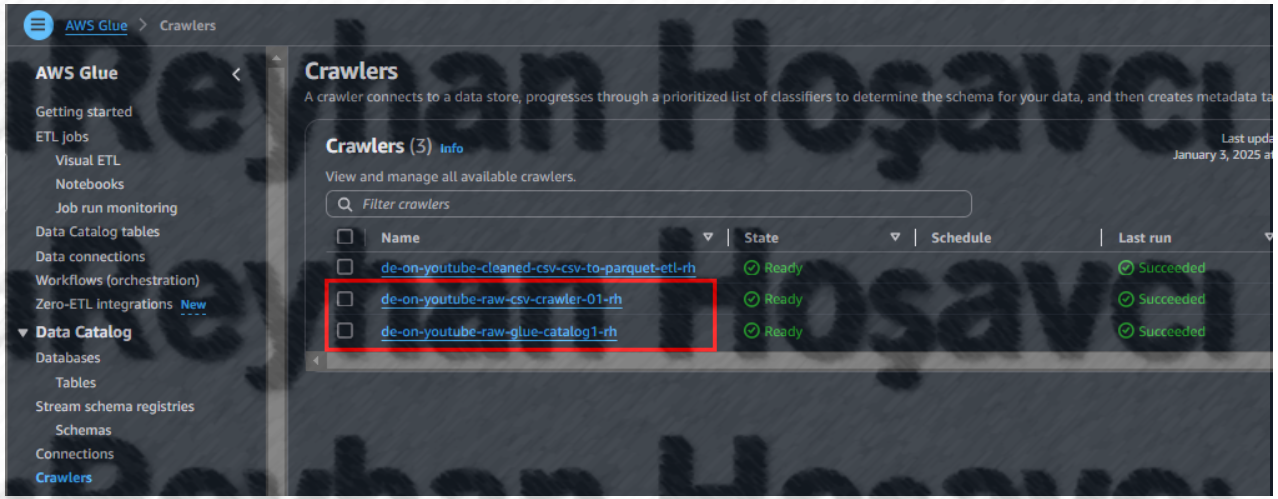


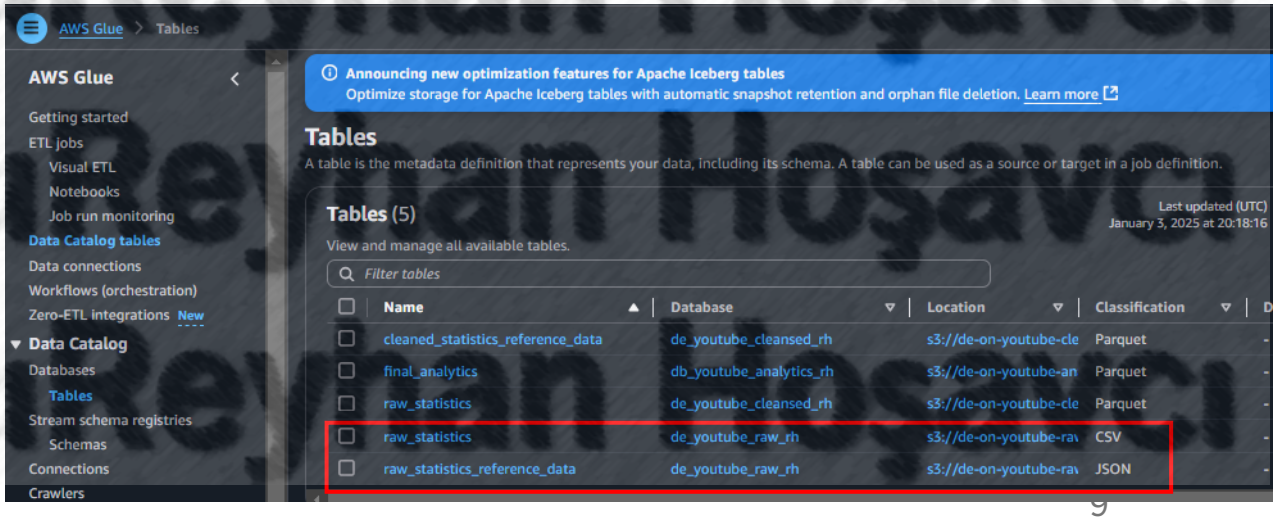
- AWS Glue Data catalog, verilerle ilgili meta verileri saklayan ve yöneten bir hizmettir.
- Glue ETL işleri, veri şemalarını tanımlamak için Data Catalog u kullanır.
- Glue Catalog, S3 veri kaynağının şemalarını, sütunlarını tanımak ve depolamak üzere kullanıldı.

Glue Crawler

- Glue Crawler ile S3 bucket taranıp JSON ve CSV dosyalarına ait şemalar çıkarıldı, ve Data Catalog üzerinde bir veritabanı ve tablolar oluşturuldu.





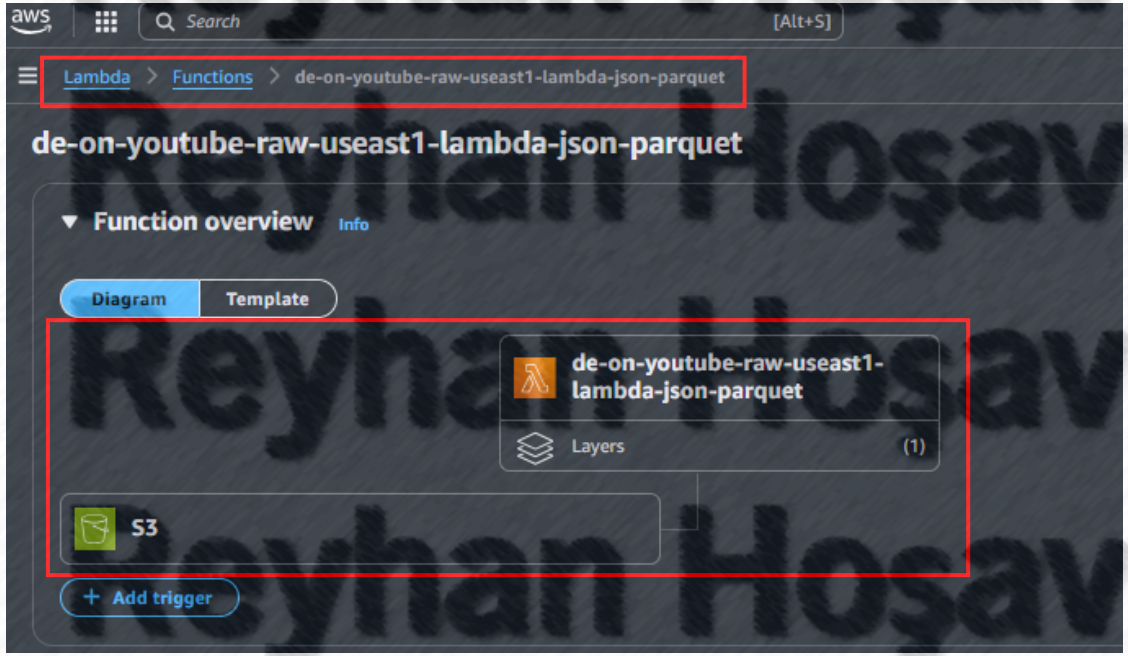


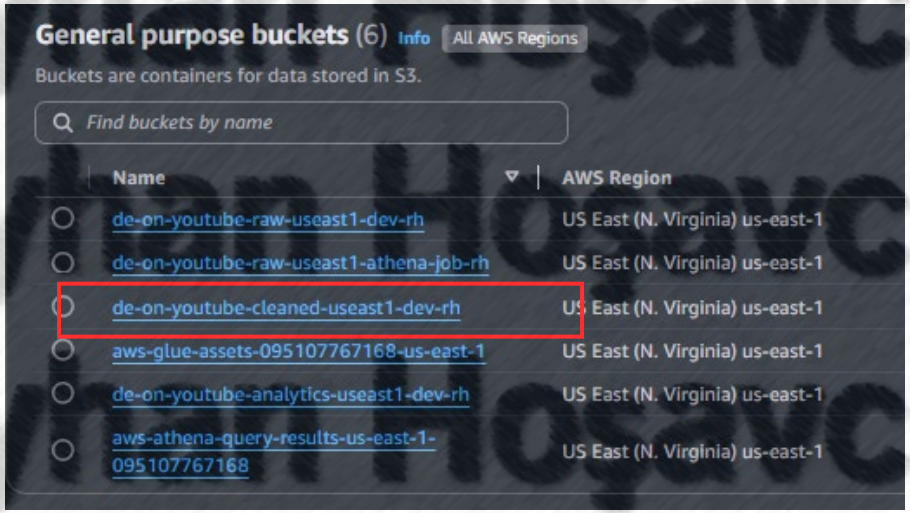
Proje Mimarisi

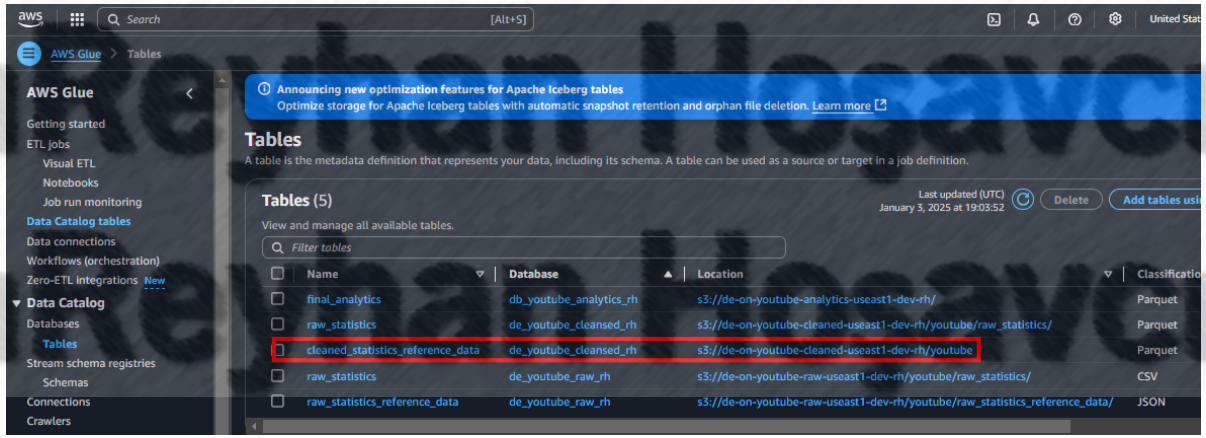
Mimari diagram: https://www.projectpro.io/project-use-case/aws-etl-data-pipeline-python-example

AWS Lambda ile verilerin standartlaştırılması

- AWS Lambda içerisinde bir fonksiyon oluşturularak ETL iş akışı gerçekleştirildi.
- JSON formatındaki veriler, Parquet formatına dönüştürülerek AWS Glue iş akışları için hazır hale getirildi.
- S3 tetikleyicisi eklenerek, Lambda fonksiyonunun S3 bucket'a yüklenen JSON verilerini otomatik olarak Parquet formatına dönüştürmesi sağlandı.
- Veriler tekrar S3 bucket üzerinde depolandı.
- Aynı zamanda bu fonksiyon ile AWS Glue üzerinde Tablo oluşturuldu.







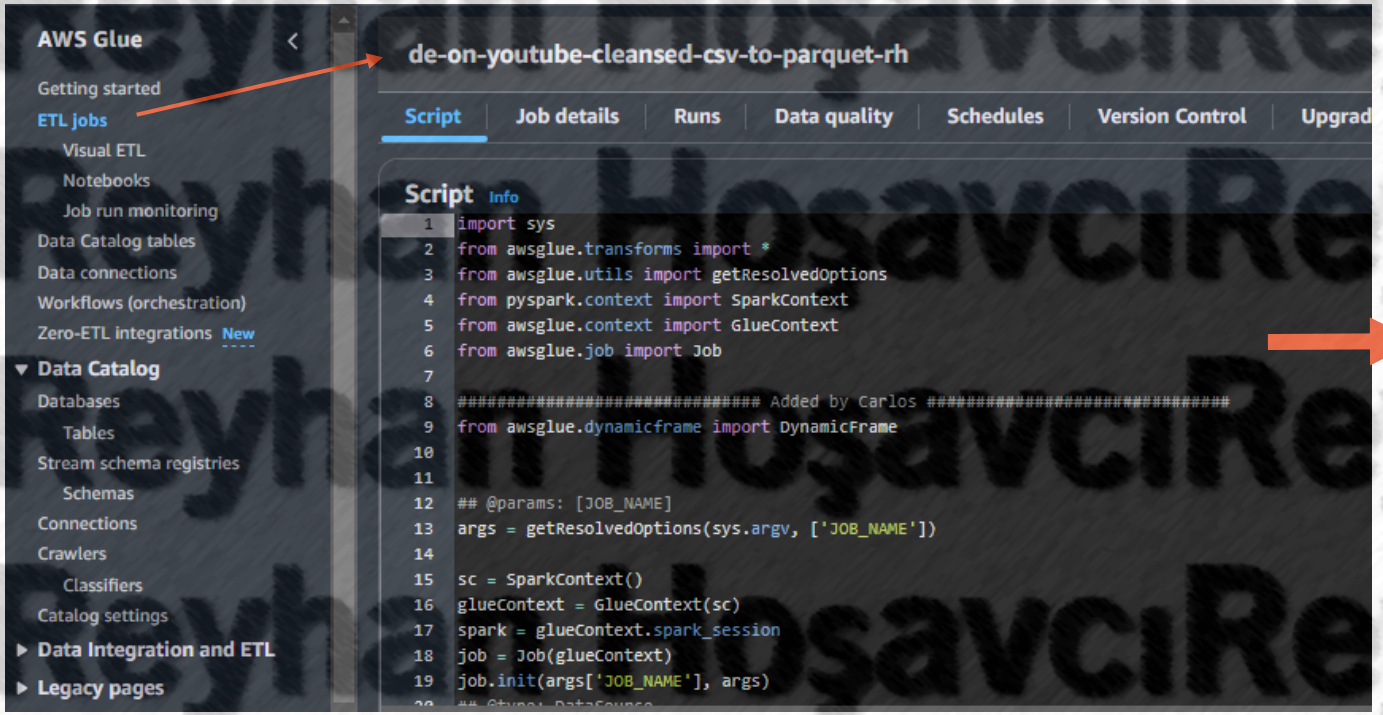
Glue ETL iş akışı

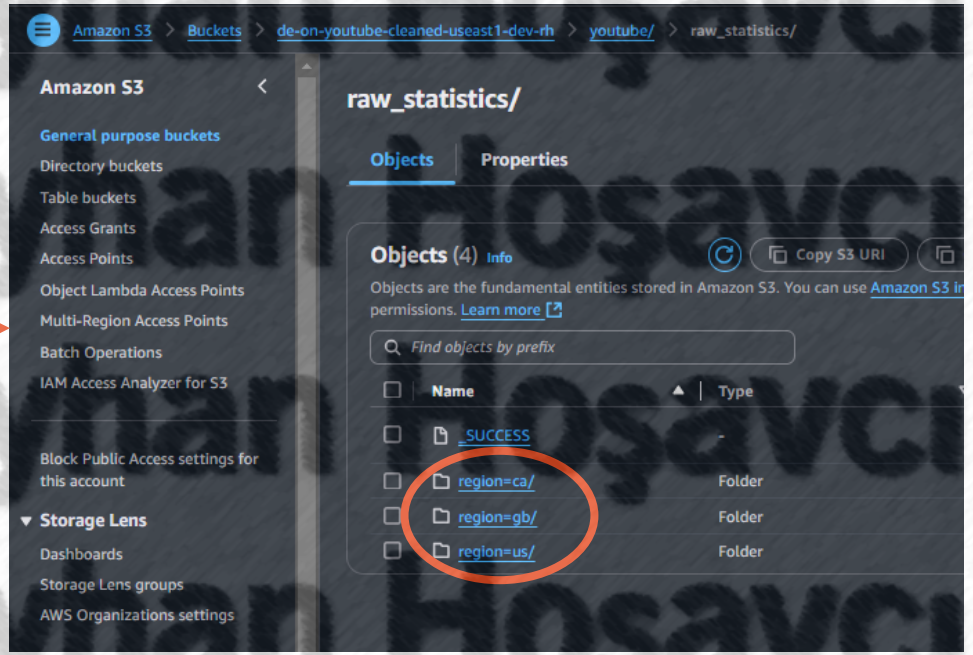
Proje Mimarisi

Mimari diagram: https://www.projectpro.io/project-use-case/aws-etl-data-pipeline-python-example

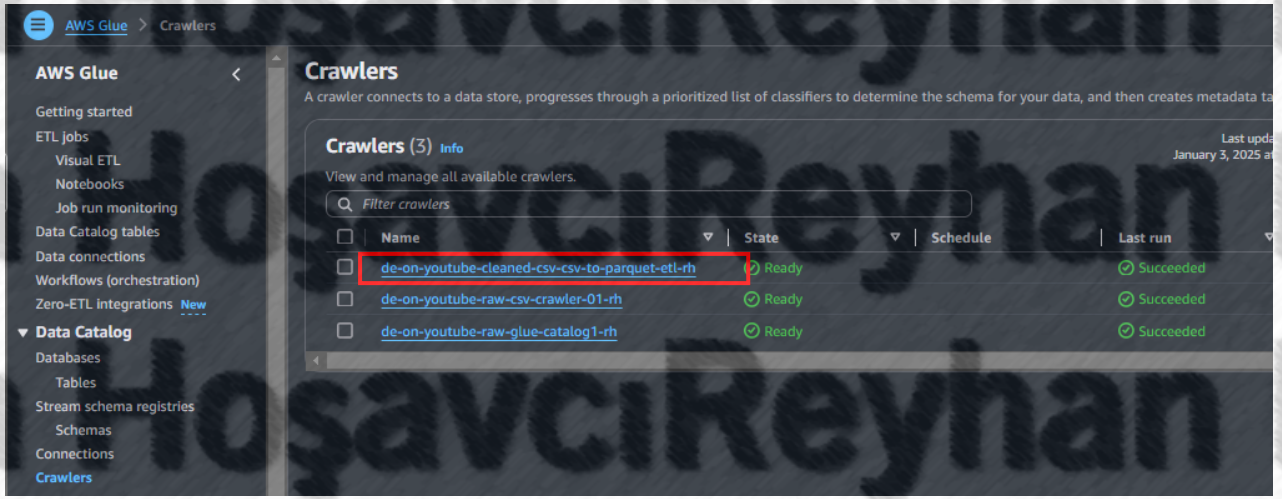
Glue ETL iş akışı (1)

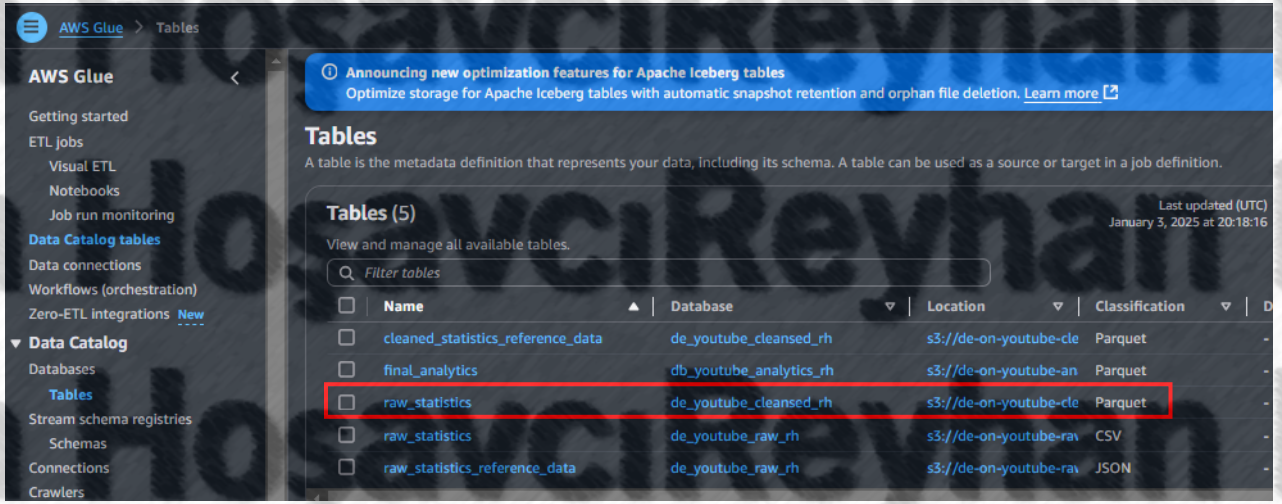
- AWS Glue üzerinde ETL iş akışı oluşturuldu (de-on-youtube-cleansed-csv-to-parquet-rh)
- Bu ETL işi ile beraber veriler parquet formatındaki veriler üzerinde temizleme, filtreleme ve sınıflandırma işlemleri yapıldı. (Örneğin belirli bölgeler filtrelendi: ca, gb, us)
- Temizlenmiş veri Parquet formatında, region (bölge) bazlı olarak S3e kaydedildi.



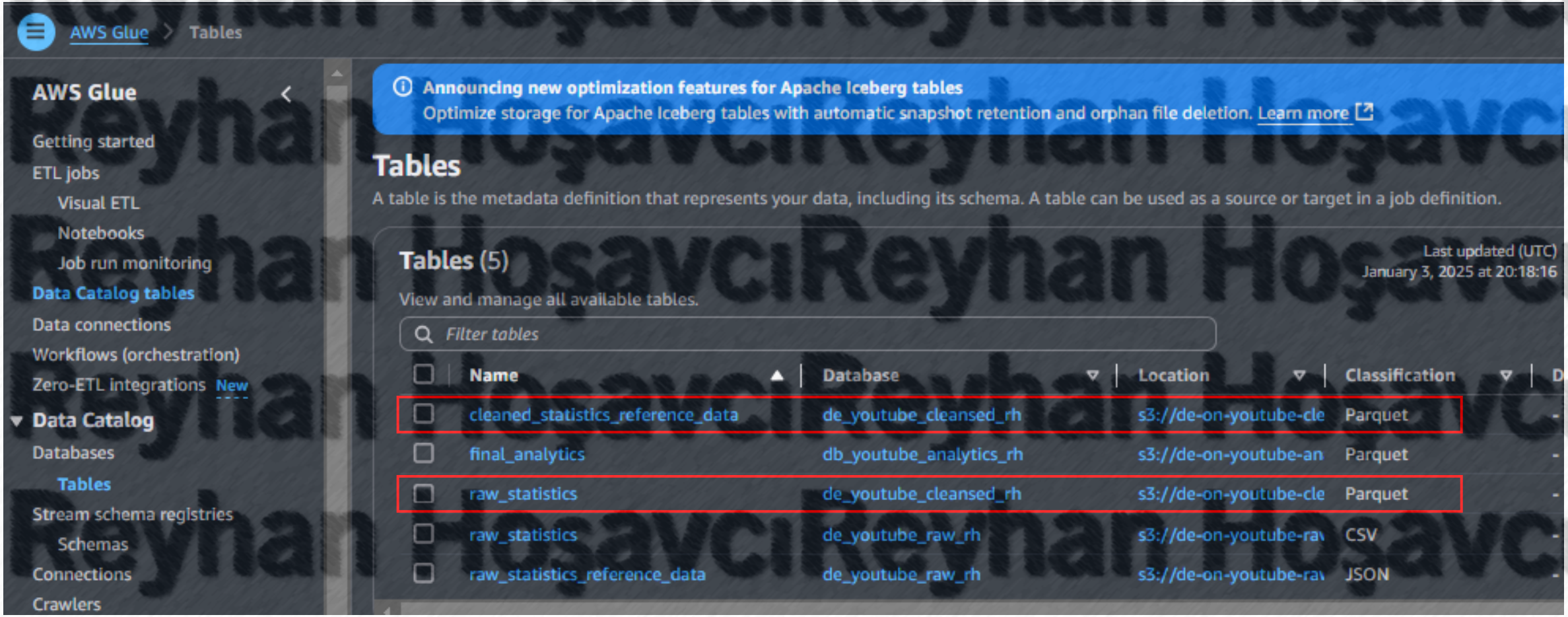


- Glue Crawler ile temizlenmiş JSON ve CSV dosyalarına ait şemalar çıkarıldı, ve Data Catalog üzerinde bir veritabanı ve tablolar oluşturuldu.



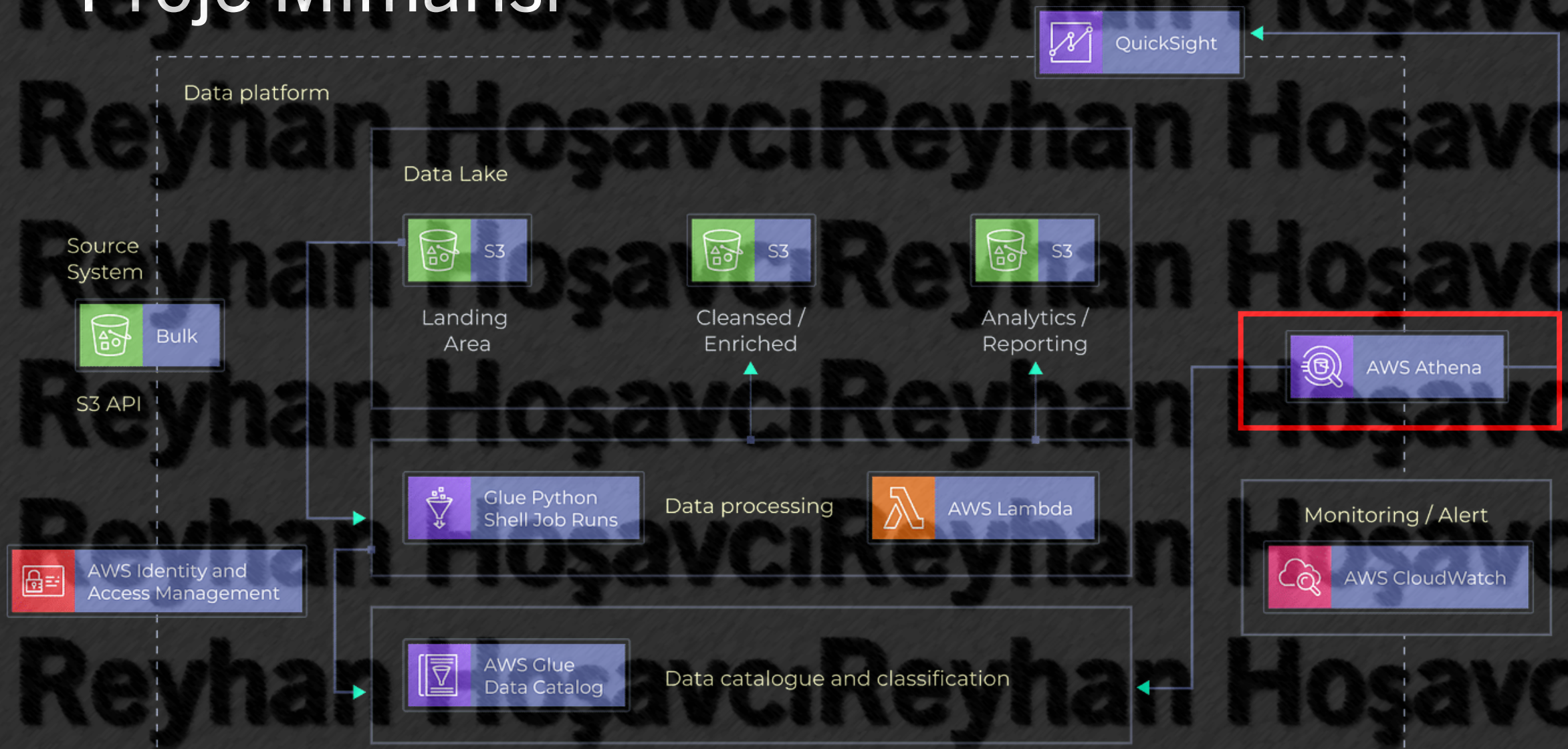


- Parquet formatında filtrelenmiş/temizlenmiş iki farklı tablo var. Bu tablolar category_id alanı üzerinden birbiri ile ilişkilendirilebilir.



Athena ile verilerin analiz edilmesi

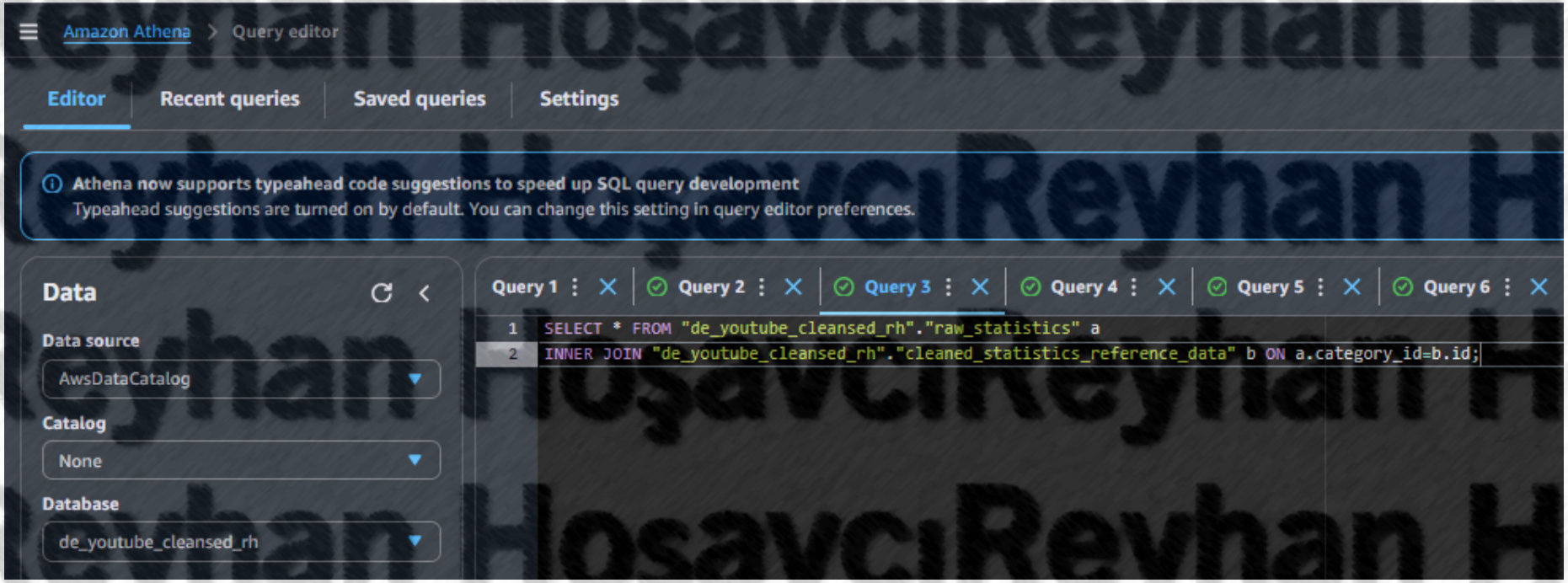
Proje Mimarisi



Mimari diagram: https://www.projectpro.io/project-use-case/aws-etl-data-pipeline-python-example

Athena ile verilerin analiz edilmesi

- AWS Athena ile veriler üzerinde SQL tabanlı sorgular çalıştırıldı ve analiz sonuçları verimli bir şekilde elde edildi.
- Category_id ile ilintili olan iki tablo SQL komutları ile analiz edildi.



```sql
SELECT * FROM "de_youtube_cleansed_rh"."raw_statistics" a
INNER JOIN "de_youtube_cleansed_rh"."cleaned_statistics_reference_data" b ON a.category_id=b.id;
```

Query results | Query stats

Completed | Time in queue: 100 ms | Run time: 4.911 sec | Data scanned: 3.49 MB

Results (283.288)

Copy | Download results

Search rows

1 2 3 4 5 ... >

#	video_id	trending_date	title
1	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
2	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
3	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
4	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
5	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
6	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE

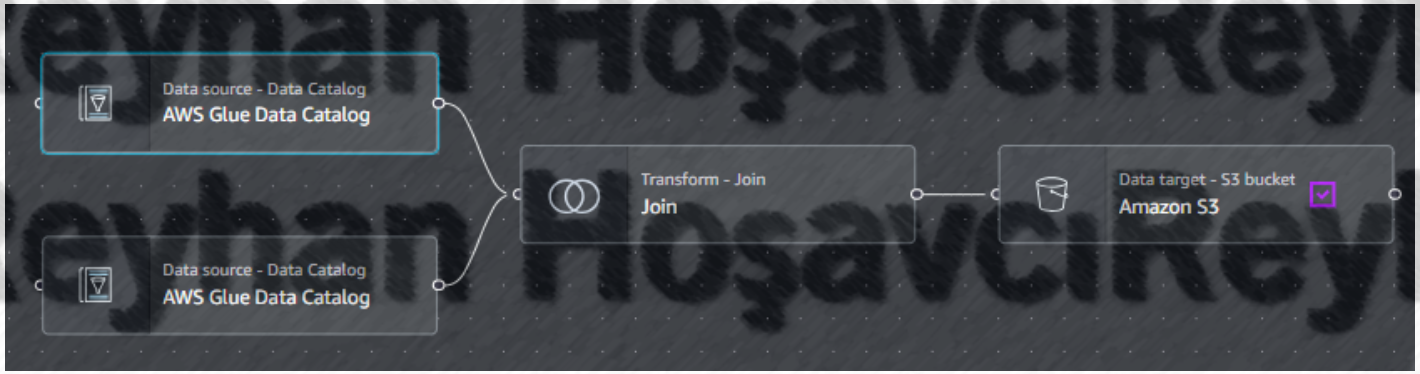
Glue ETL iş akışı – Tabloların birleştirilmesi

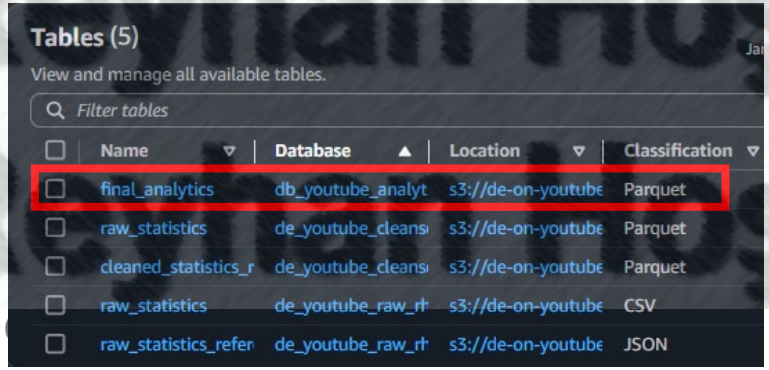
Proje Mimarisi

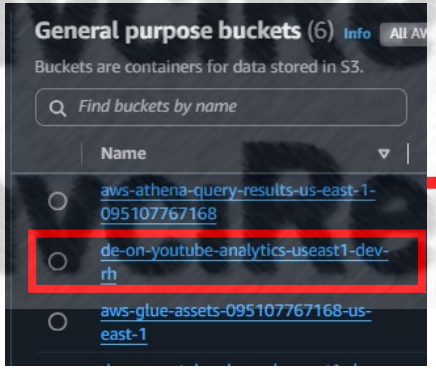
Mimari diagram: https://www.projectpro.io/project-use-case/aws-etl-data-pipeline-python-example

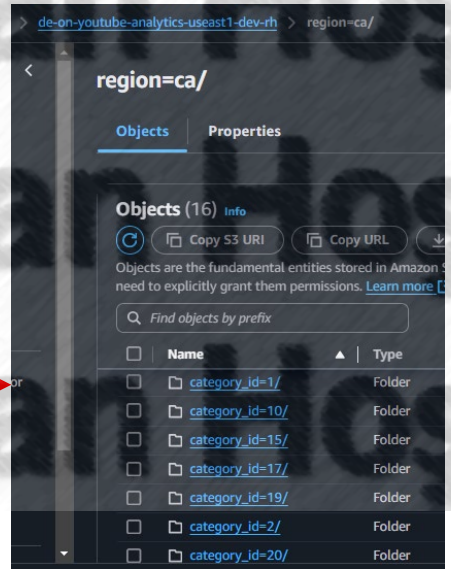
Glue ETL iş akışı (2) – Tabloların birleştirilmesi

- İki tablo category_id üzerinden join edildi.
- Yeni oluşturulan tablo yeni database/bucket/tablo ‘e aktarıldı.









```sql
SELECT * FROM "db_youtube_analytics_rh"."final_analytics" limit 10;
```

Query results | Query stats

Completed | Time in queue: 103 ms | Run time: 1.147 sec | Data scanned: 209.18 KB

Results (10) | Copy | Download results

Search rows | 1

#	ratings_disabled	comments_disabled	snippet_title	trending_date	etag
1	false	false	Autos & Vehicles	18.12.01	"XI7nbFXulYBIpL0ayR_gDh3e
2	false	false	Autos & Vehicles	18.12.01	"XI7nbFXulYBIpL0ayR_gDh3e
3	false	false	Autos & Vehicles	18.09.01	"XI7nbFXulYBIpL0ayR_gDh3e
4	false	false	Autos & Vehicles	18.09.01	"XI7nbFXulYBIpL0ayR_gDh3e
5	false	false	Autos & Vehicles	18.12.01	"XI7nbFXulYBIpL0ayR_gDh3e

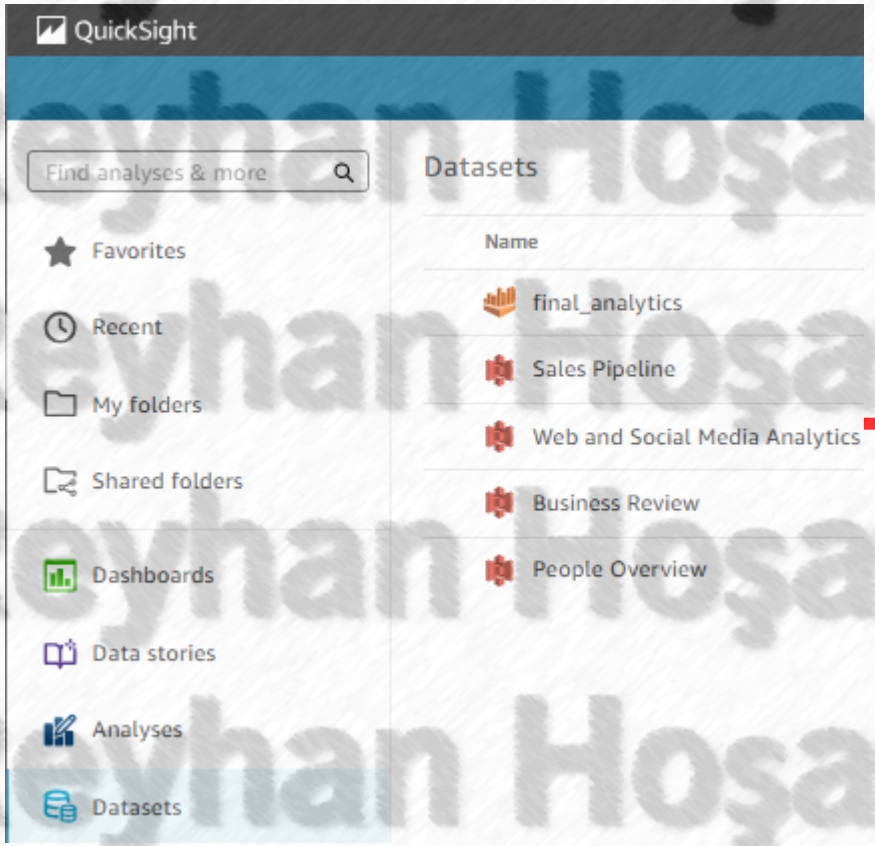
Quicksight ile analiz sonuçlarının görselleştirilmesi

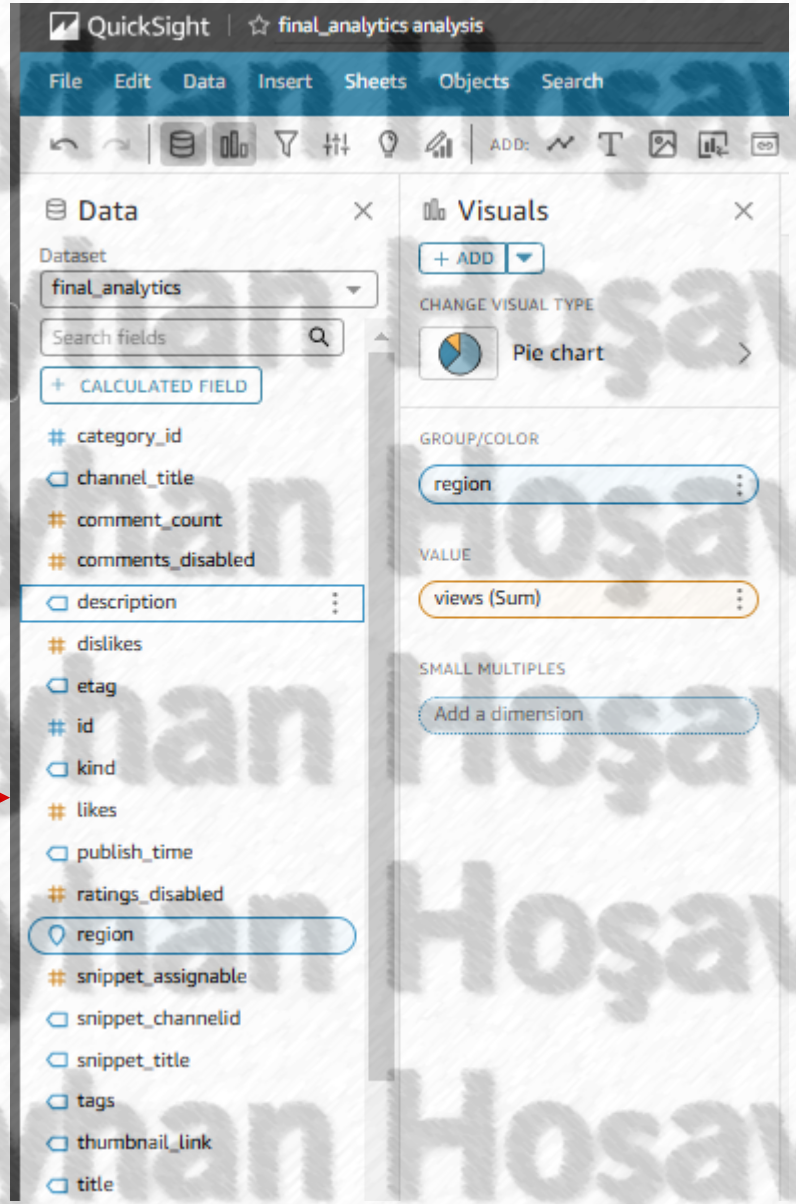
Proje Mimarisi

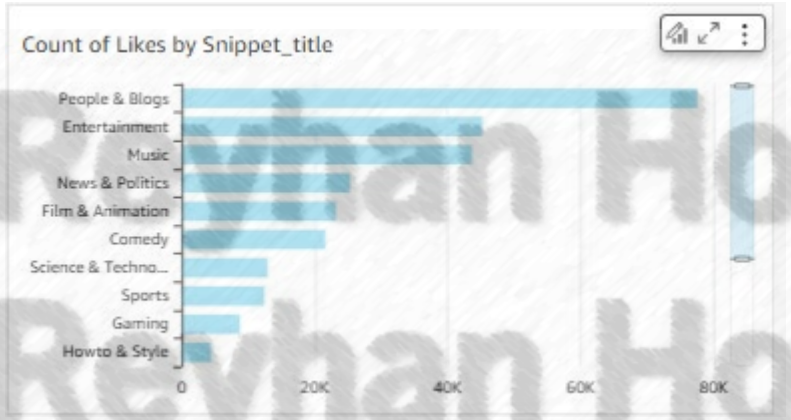
Mimari diagram: https://www.projectpro.io/project-use-case/aws-etl-data-pipeline-python-example

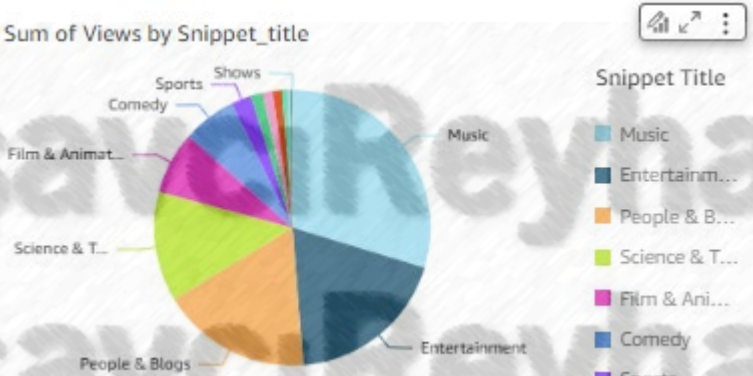
Quicksight ile analiz sonuçlarının görselleştirilmesi

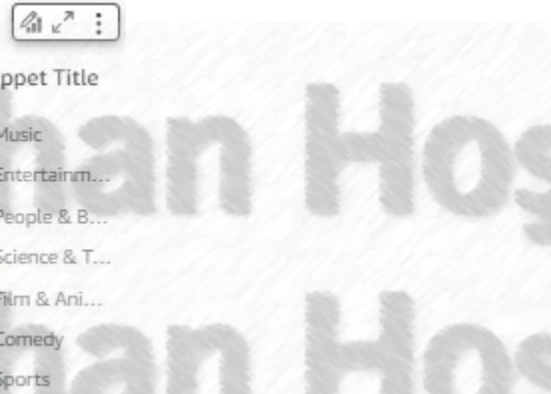
- Athena kullanılarak database aktarıldı

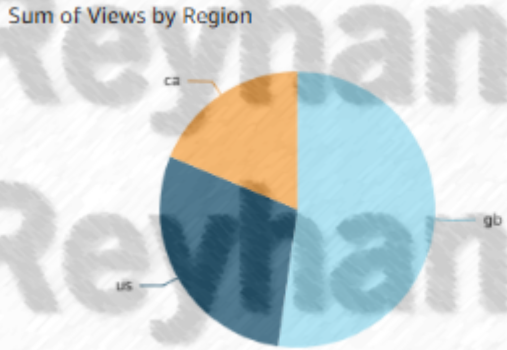


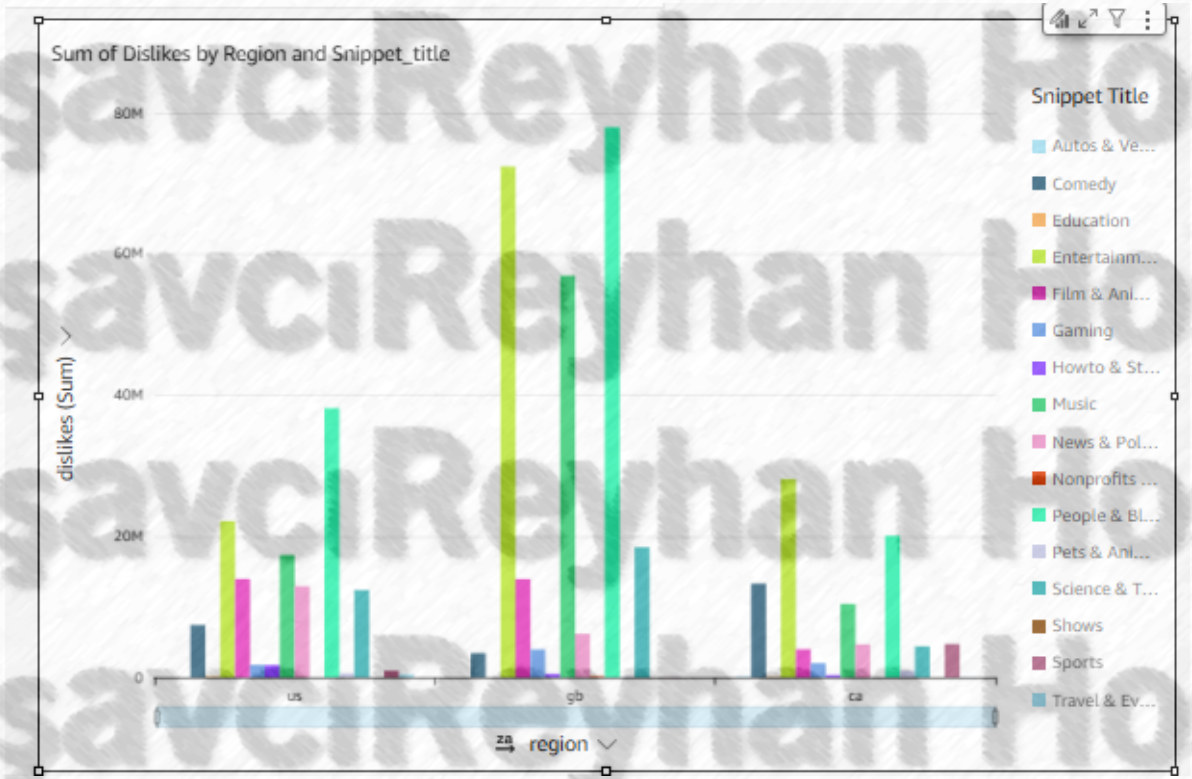












Teşekkürler