

Fatih Sultan Mehmet Vakıf Üniversitesi



Veri Analizi

Veri Madenciliğinde Özel Konular
Dr. Öğr. Üyesi Zeki KUŞ

Final Projesi

Reyhan HOŞAVCI 231221001

Kullanılan Teknolojiler

aws

Bu projede AWS'nin bulut tabanlı hizmetlerinden yararlanıldı:

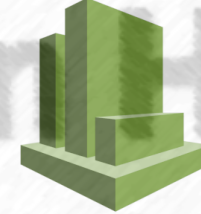
- **AWS Lambda:** Veri işleme ve dönüştürme işlemleri
- **AWS S3:** Yedekleme, arşivleme ve veri paylaşımı
- **AWS Glue:** ETL işlemleri, Veri cataloglama ve sınıflandırma
- **AWS Athena:** SQL tabanlı sorgularla verilerin analizi
- **Amazon QuickSight:** Verilerin görselleştirilmesi ve raporlanması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

Programlama dili : Python

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



AWS Lambda



Amazon CloudWatch



Amazon Glue



Amazon S3



Amazon Athena

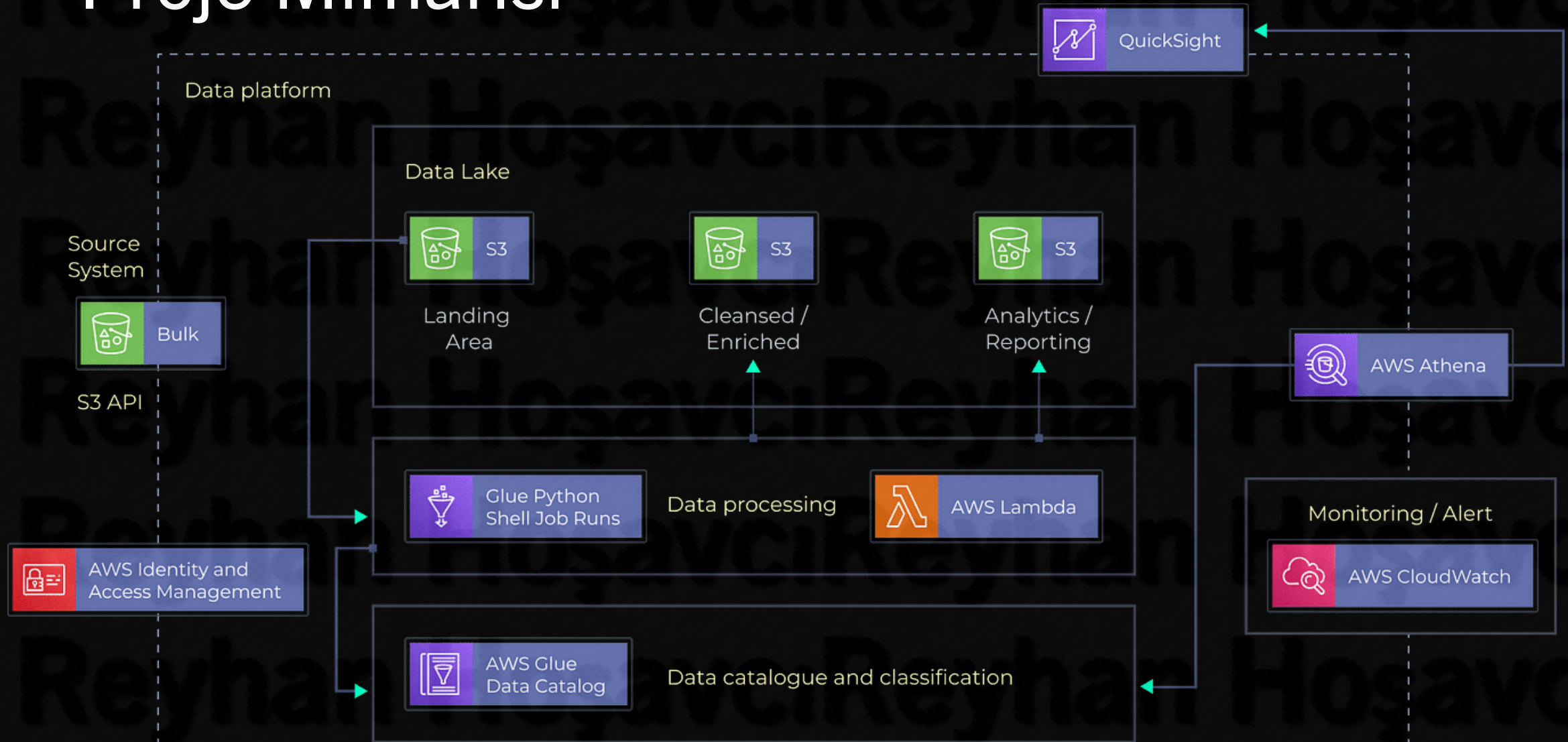


AWS IAM



amazon QuickSight

Proje Mimarisi



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)

```

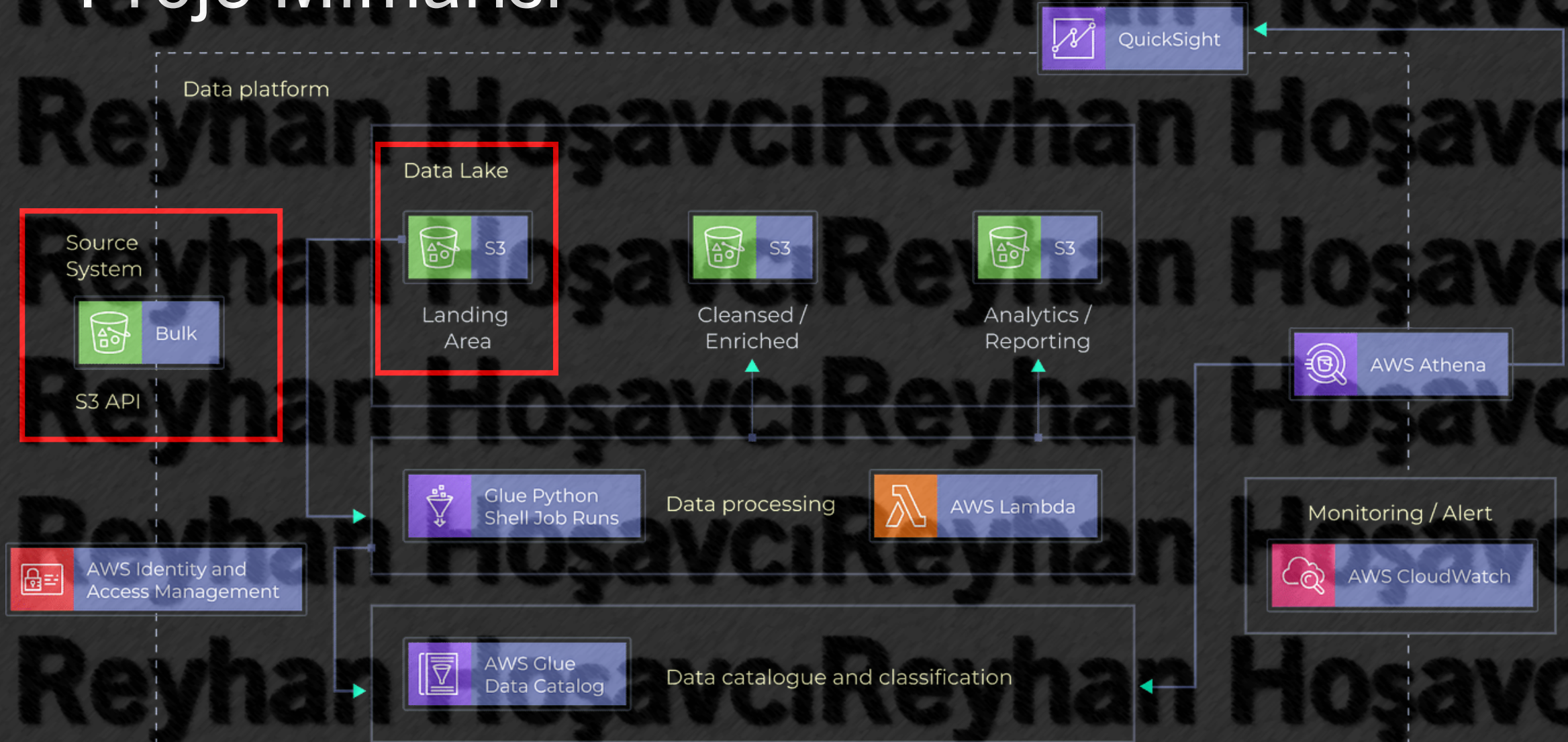
{
  "root": {
    "kind": "youtube#videoCategoryListResponse",
    "etag": "\"1d9biNPKjAjjgV7EZ4EKeEGrhao/1v2mrzYSYG6onNlt2qTj13hkQzk\"",
    "items": [
      {
        "kind": "youtube#videoCategory",
        "etag": "\"1d9biNPKjAjjgV7EZ4EKeEGrhao/Xy1mB4_yLrHy_BmKmpBggy2mZQ\"",
        "id": "1",
        "snippet": {
          "channelId": "UCBR8-60-B28hp2BmDPdntcQ",
          "title": "Film & Animation",
          "assignable": true
        }
      }
    ]
  }
}

```

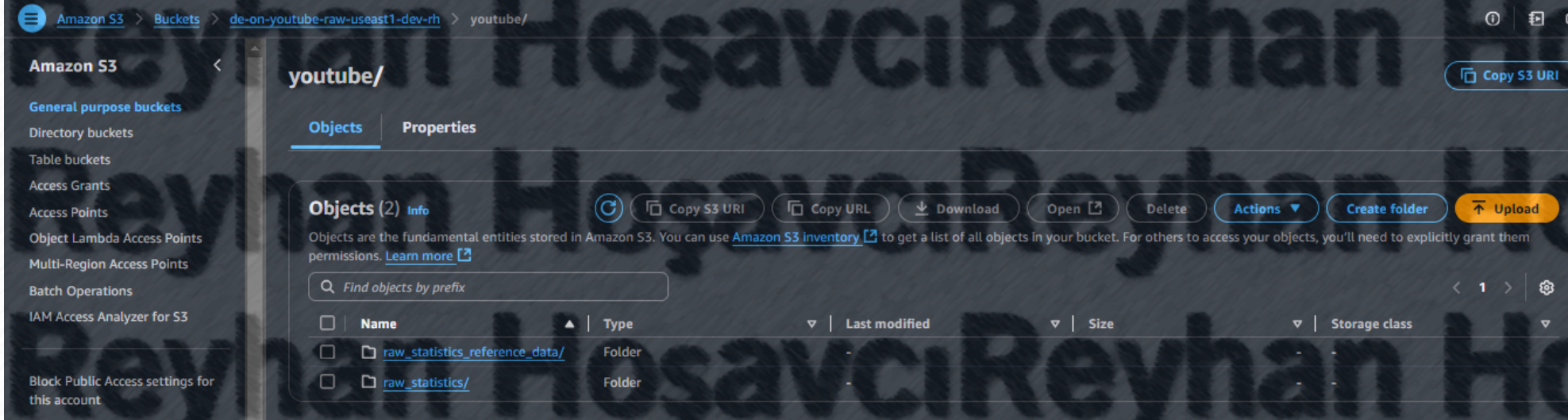
CAvideos.csv (64.07 MB)

Detail	Compact	Column	8 of 16 columns				
video_id	title	channel_title	category_id	publish_time	# views	# likes	
n1Wp7iowLc	Eminem - Walk On Water (Audio) ft. Beyoncé	EminemVEVO	10	2017-11-10T17:00:03.000Z	17158579	787425	
0dBIkQ4Mz1M	PLUSH - Bad Unboxing Fan Mail	iDubbzTV	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	146835	
d380meD8W0M	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000Z	2095828	132239	

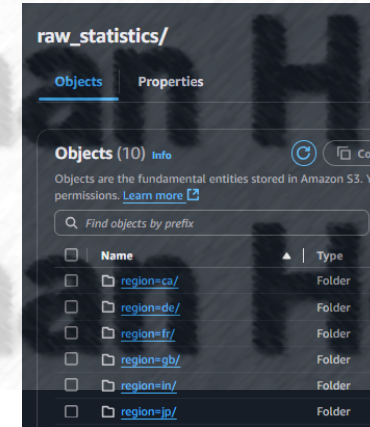
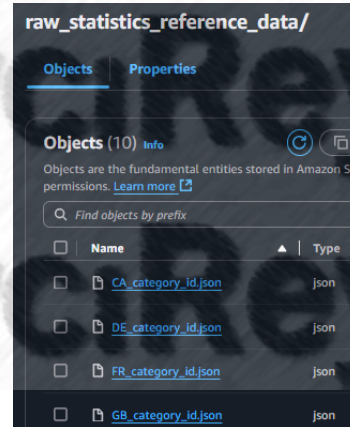
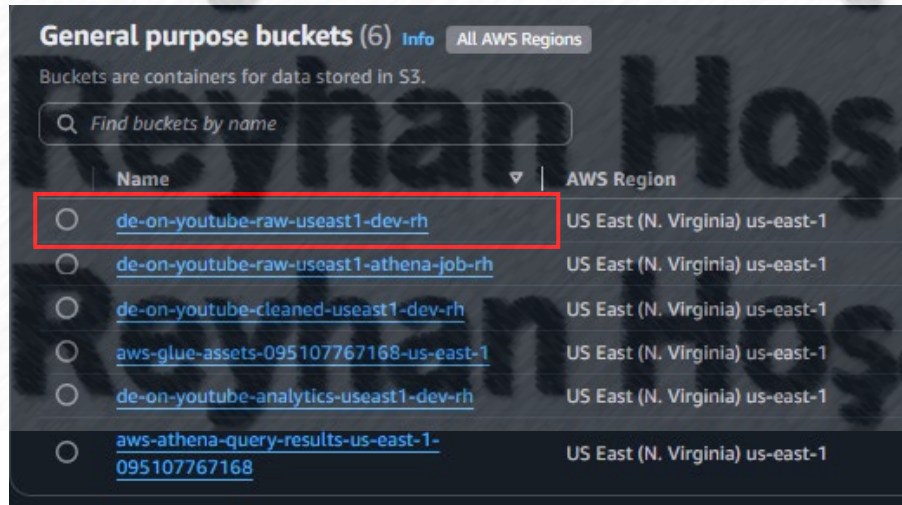
Proje Mimarisi



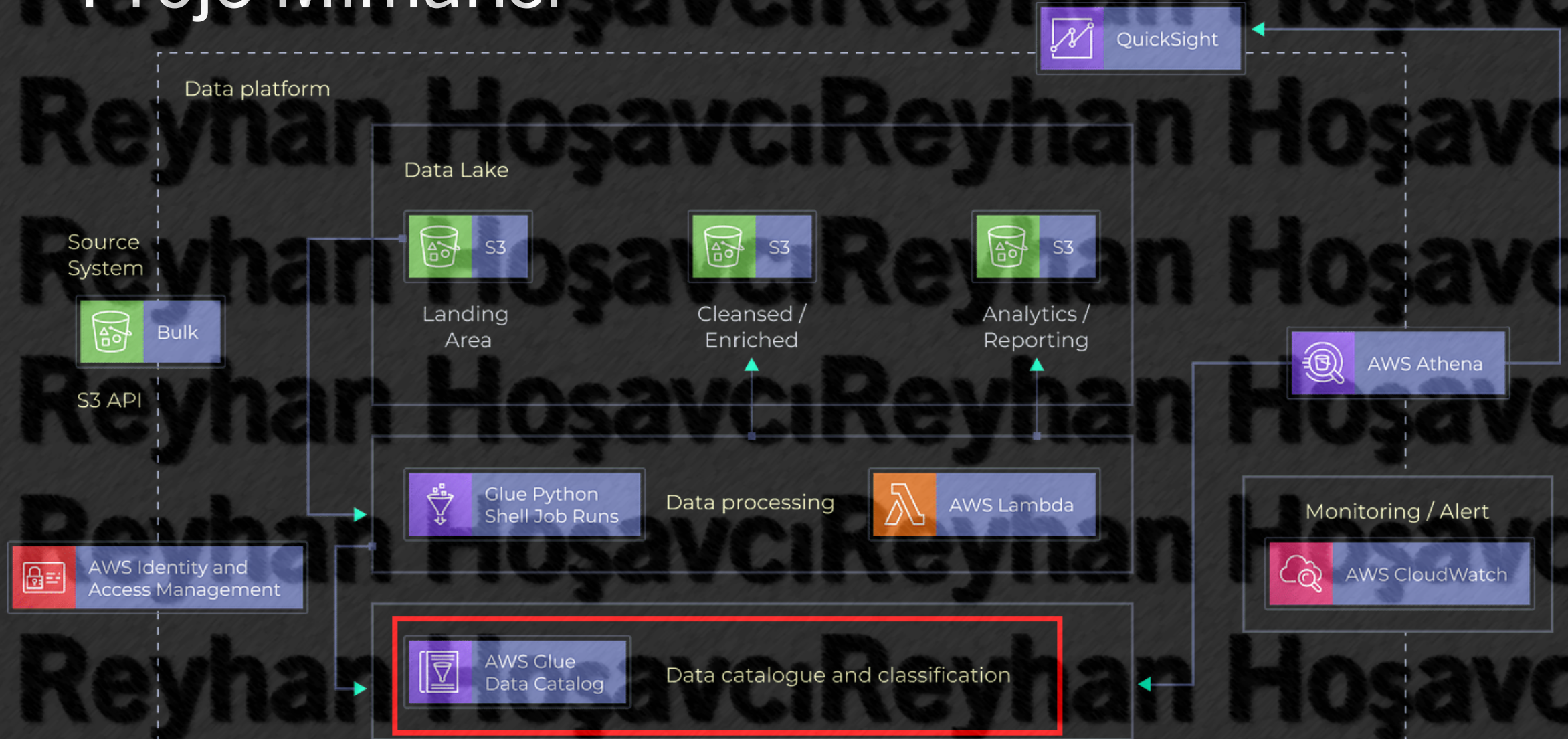
JSON ve CSV formatındaki verilerin S3 (Simple Storage Service) Bucket üzerinde toplanması



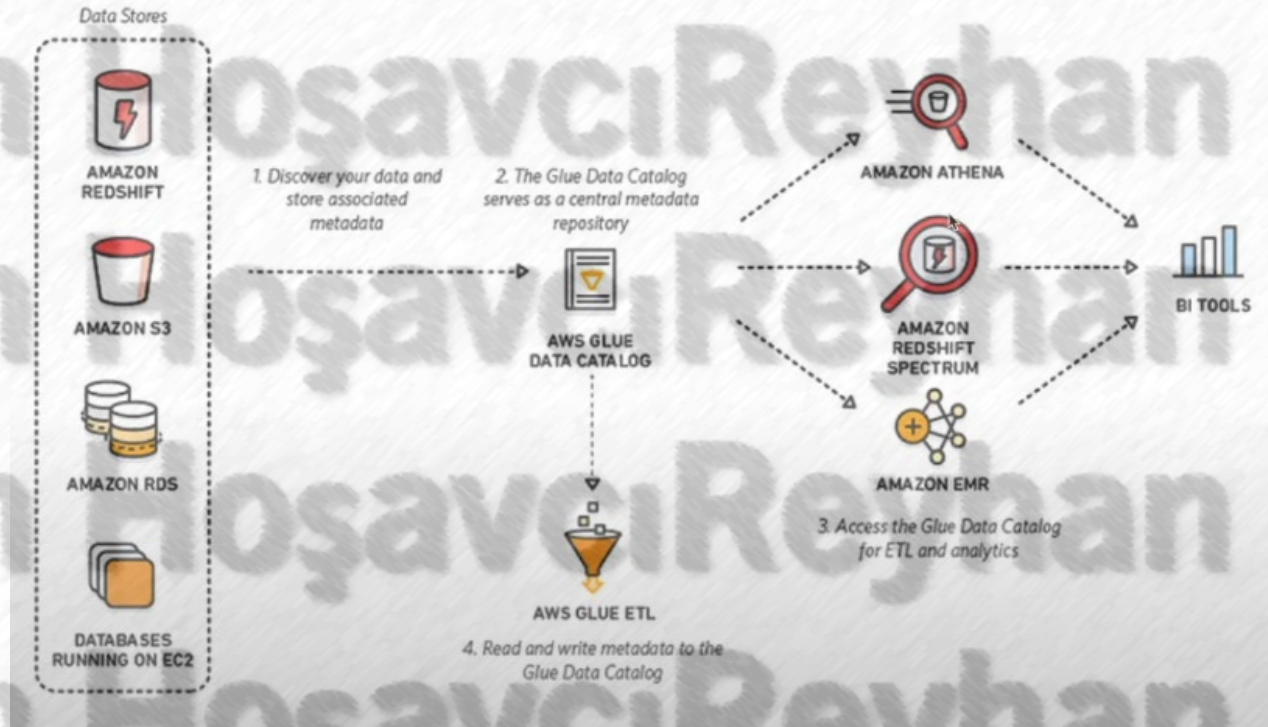
S3, verilerin organize edilmesi ve Data Lake katmanları arasında taşınması için kullanıldı.



Proje Mimarisi



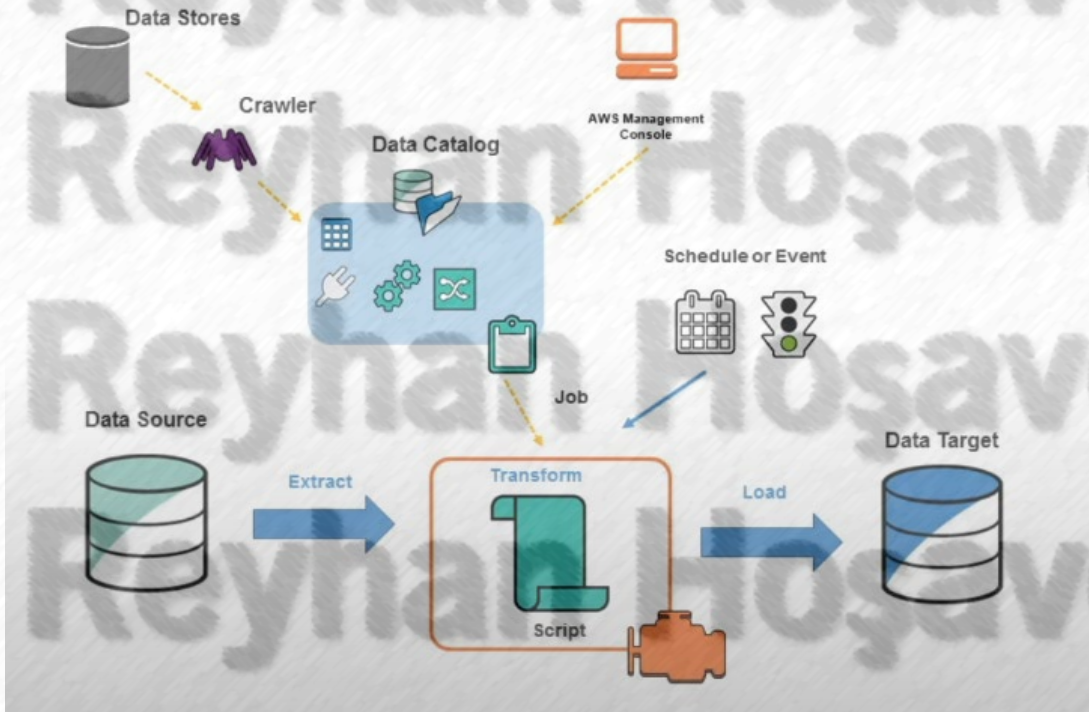
AWS Glue Catalog oluřturulması



- 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.



AWS Glue > Crawlers

Crawlers
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables.

Crawlers (3) Info
View and manage all available crawlers.

Filter crawlers

<input type="checkbox"/>	Name	State	Schedule	Last run
<input type="checkbox"/>	de-on-youtube-cleaned-csv-csv-to-parquet-etl-rh	Ready		Succeeded
<input type="checkbox"/>	de-on-youtube-raw-csv-crawler-01-rh	Ready		Succeeded
<input type="checkbox"/>	de-on-youtube-raw-glue-catalog1-rh	Ready		Succeeded

AWS Glue > Tables

Tables
A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (5)
View and manage all available tables.

Filter tables

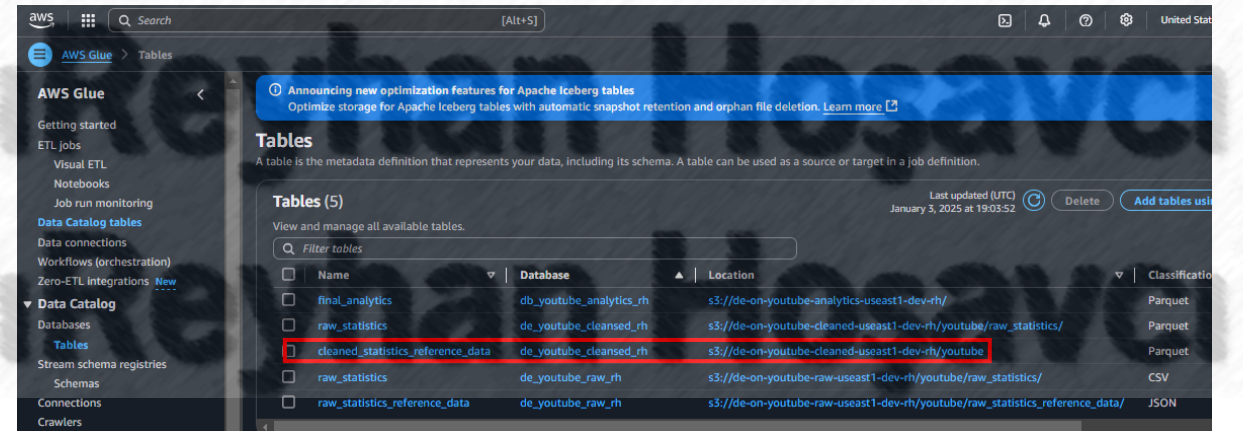
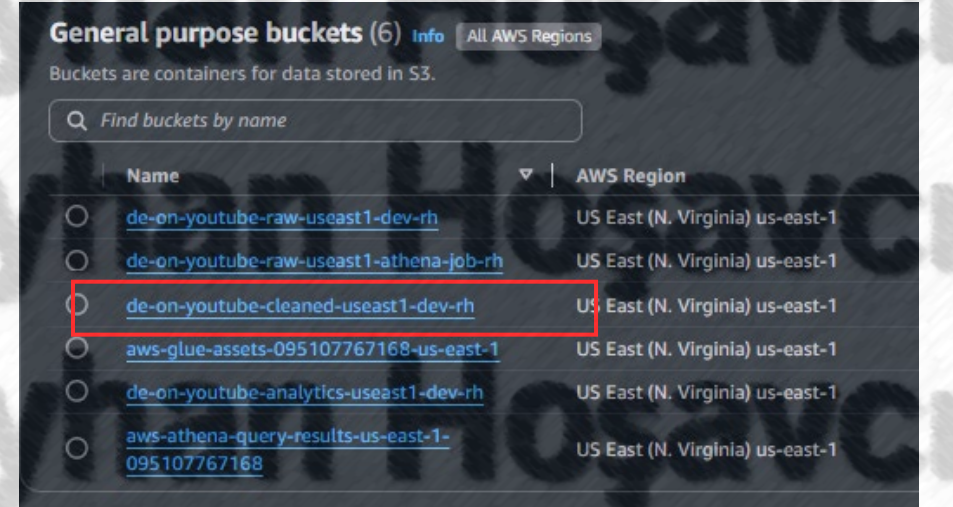
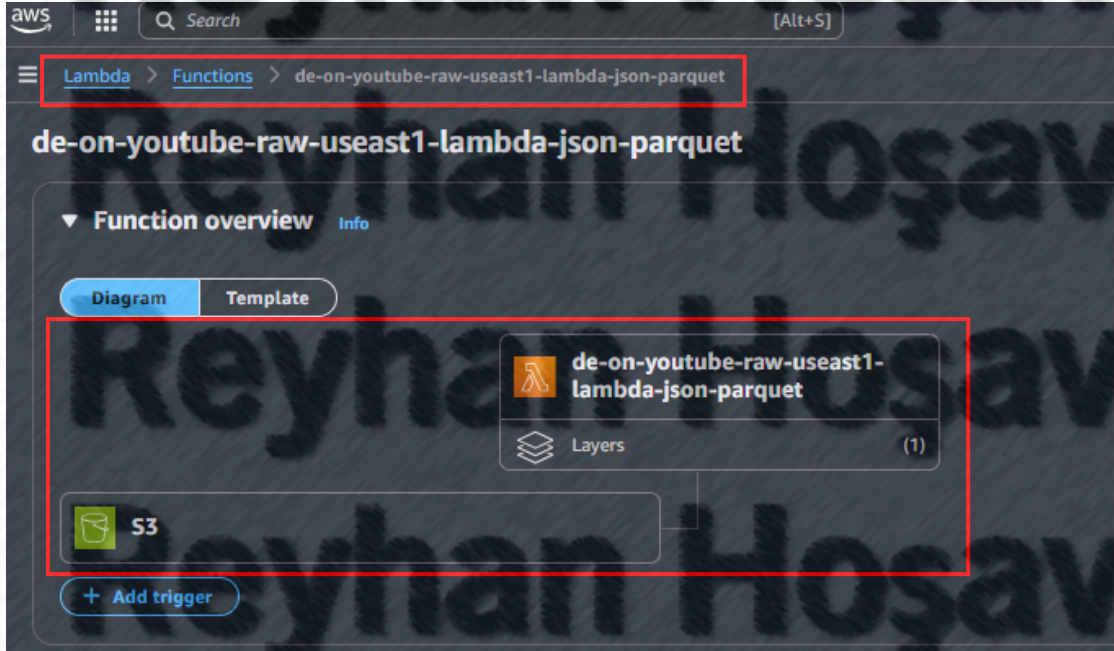
<input type="checkbox"/>	Name	Database	Location	Classification	D
<input type="checkbox"/>	cleaned_statistics_reference_data	de_youtube_cleaned_rh	s3://de-on-youtube-cle	Parquet	-
<input type="checkbox"/>	final_analytics	db_youtube_analytics_rh	s3://de-on-youtube-an	Parquet	-
<input type="checkbox"/>	raw_statistics	de_youtube_cleaned_rh	s3://de-on-youtube-cle	Parquet	-
<input type="checkbox"/>	raw_statistics	de_youtube_raw_rh	s3://de-on-youtube-raw	CSV	-
<input type="checkbox"/>	raw_statistics_reference_data	de_youtube_raw_rh	s3://de-on-youtube-raw	JSON	-

Proje Mimarisi



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.

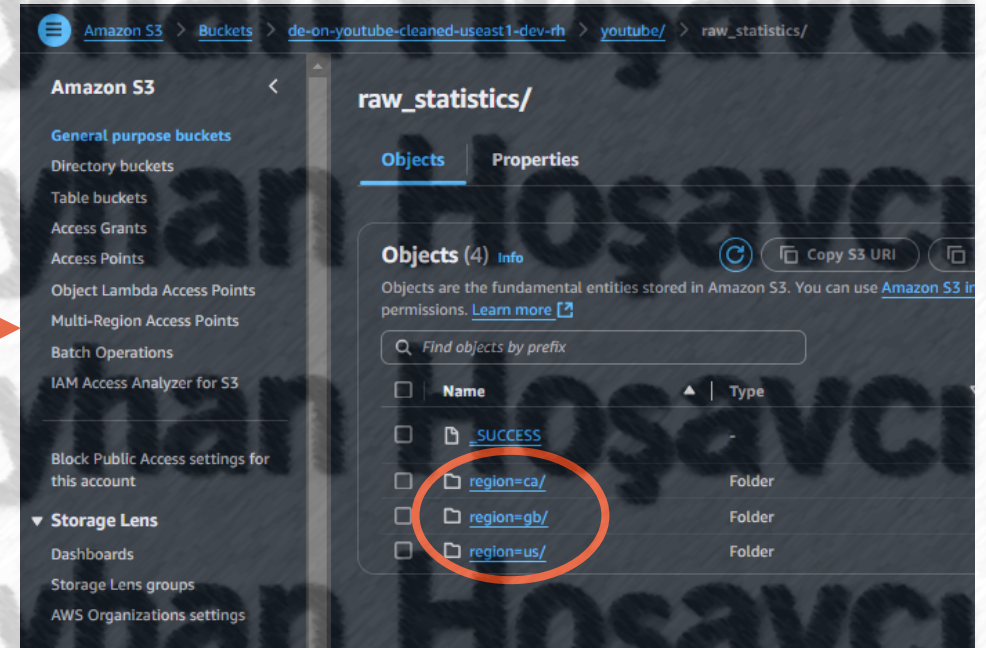
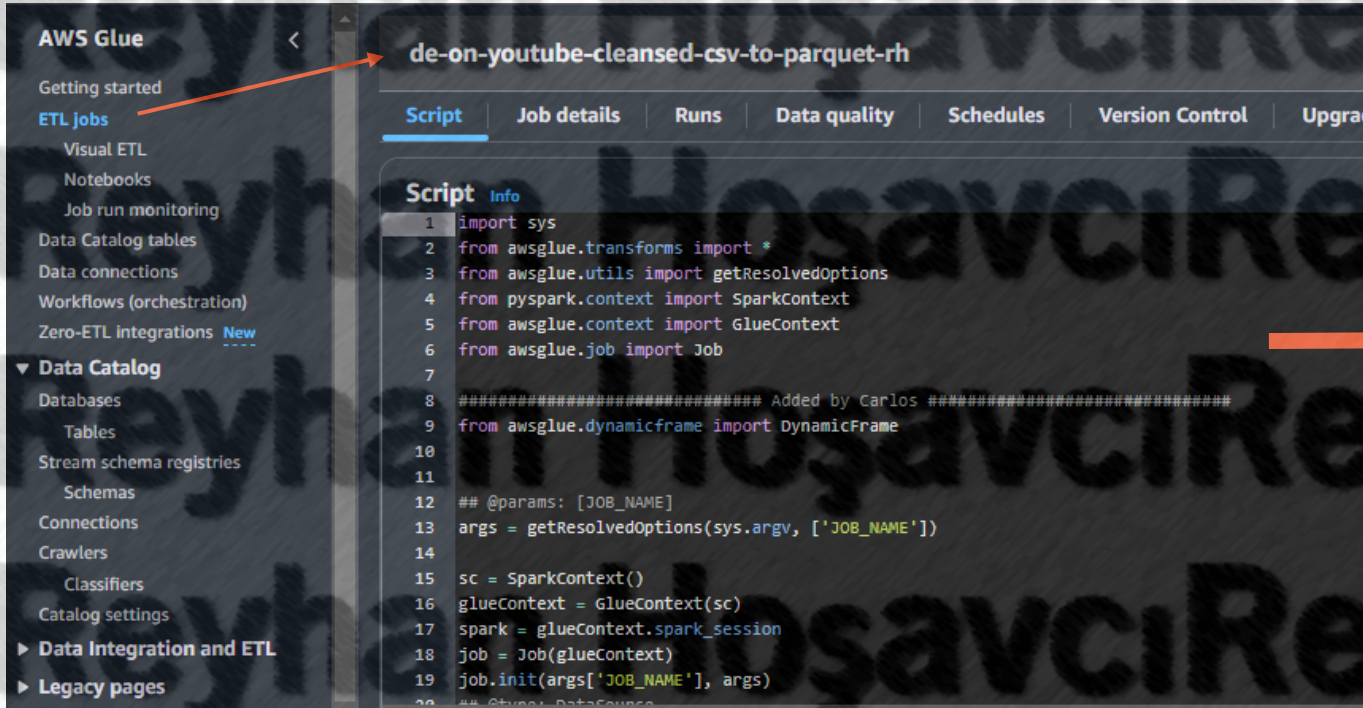


Proje Mimarisi

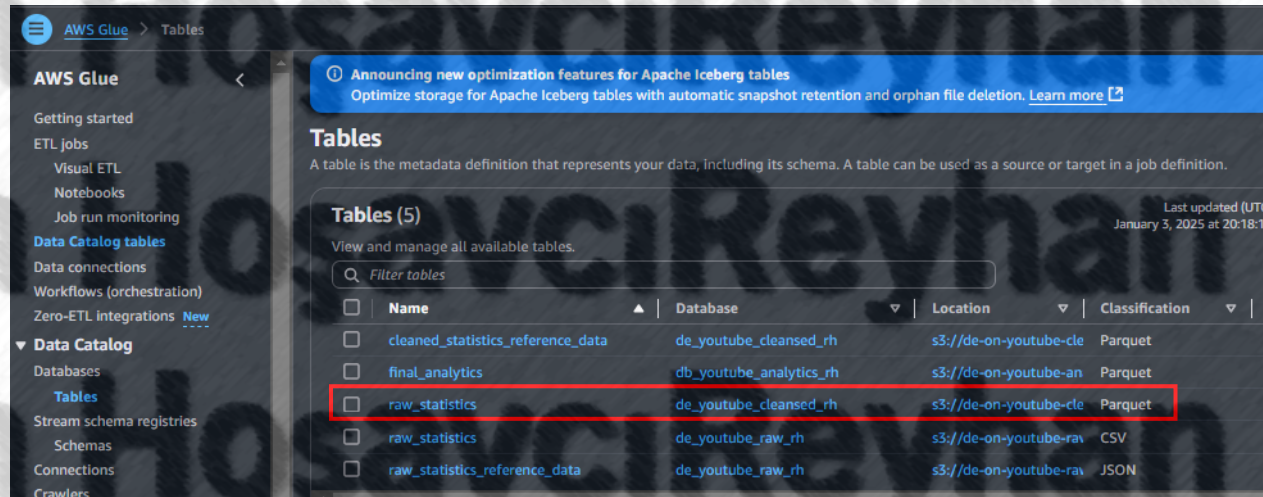
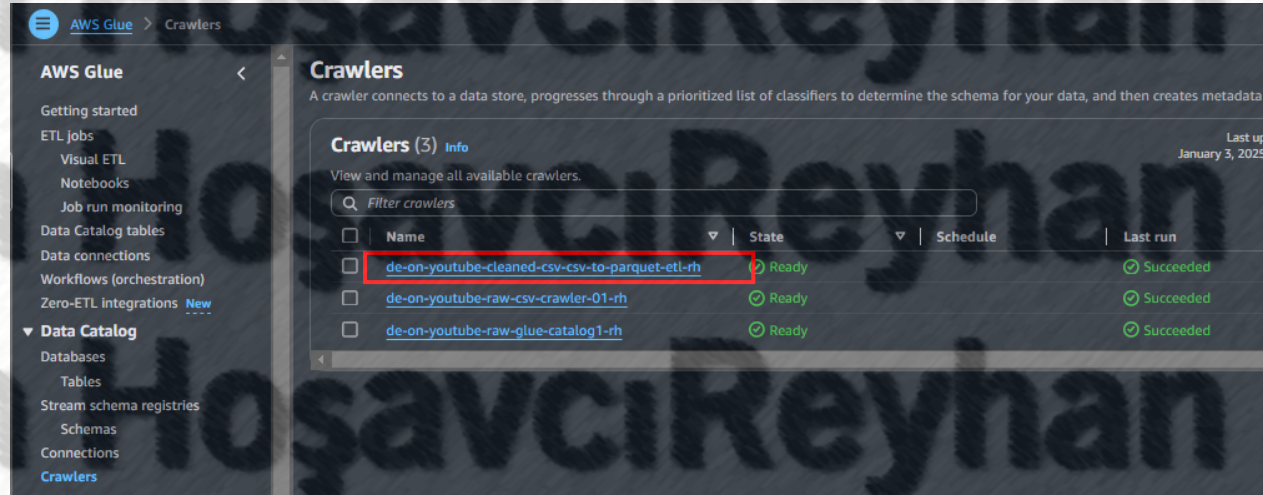


Glue ETL iş akışı (1)

- 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.

AWS Glue > Tables

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Zero-ETL integrations New

▼ **Data Catalog**
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers

Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)

Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (5) Last updated (UTC) January 3, 2025 at 20:18:16

View and manage all available tables.

Filter tables

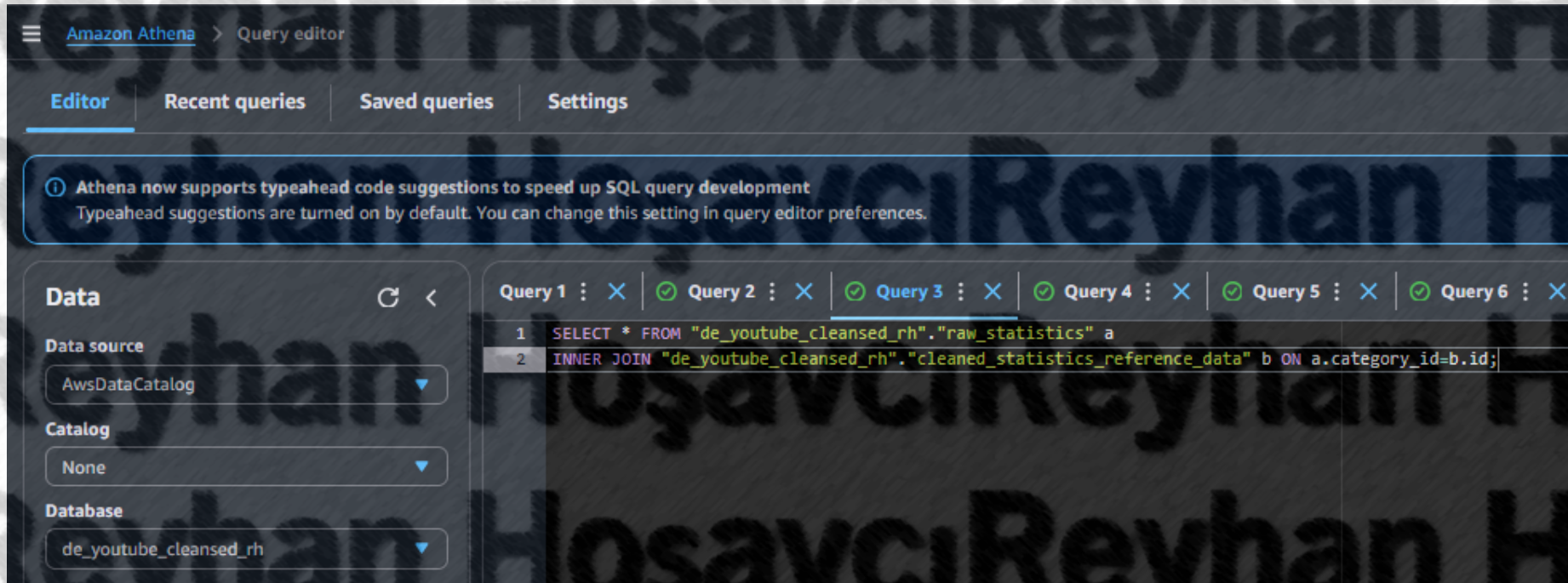
<input type="checkbox"/>	Name	Database	Location	Classification	D
<input type="checkbox"/>	cleaned_statistics_reference_data	de_youtube_cleansed_rh	s3://de-on-youtube-cle	Parquet	-
<input type="checkbox"/>	final_analytics	db_youtube_analytics_rh	s3://de-on-youtube-an	Parquet	-
<input type="checkbox"/>	raw_statistics	de_youtube_cleansed_rh	s3://de-on-youtube-cle	Parquet	-
<input type="checkbox"/>	raw_statistics	de_youtube_raw_rh	s3://de-on-youtube-raw	CSV	-
<input type="checkbox"/>	raw_statistics_reference_data	de_youtube_raw_rh	s3://de-on-youtube-raw	JSON	-

Proje Mimarisi



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.




```
1 SELECT * FROM "de_youtube_cleansed_rh"."raw_statistics" a
2 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 ... > ⚙

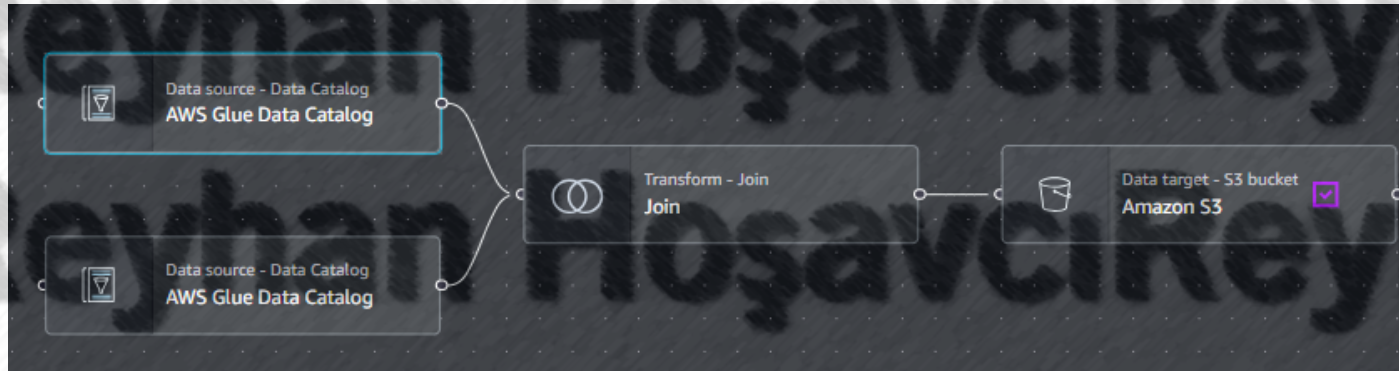
# ▾	video_id ▾	trending_date ▾	title
1	2kyS6vSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
2	2kyS6vSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
3	2kyS6vSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
4	2kyS6vSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
5	2kyS6vSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE
6	2kyS6vSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE

Proje Mimarisi



Glue ETL iş akışı (2) – Tabloların birleştirilmesi

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



Tables (5)
View and manage all available tables.

Filter tables

<input type="checkbox"/>	Name	Database	Location	Classification
<input type="checkbox"/>	final_analytics	db_youtube_analyt	s3://de-on-youtube	Parquet
<input type="checkbox"/>	raw_statistics	de_youtube_cleansi	s3://de-on-youtube	Parquet
<input type="checkbox"/>	cleaned_statistics_r	de_youtube_cleansi	s3://de-on-youtube	Parquet
<input type="checkbox"/>	raw_statistics	de_youtube_raw_rh	s3://de-on-youtube	CSV
<input type="checkbox"/>	raw_statistics_refer	de_youtube_raw_rh	s3://de-on-youtube	JSON

General purpose buckets (6) Info All AV

Buckets are containers for data stored in S3.

Find buckets by name

<input type="radio"/>	Name
<input type="radio"/>	aws-athena-query-results-us-east-1-095107767168
<input type="radio"/>	de-on-youtube-analytics-useast1-dev-rh
<input type="radio"/>	aws-glue-assets-095107767168-us-east-1

region=ca/

Objects (16) Info

Copy S3 URI Copy URL

Objects are the fundamental entities stored in Amazon S3. You need to explicitly grant them permissions. Learn more

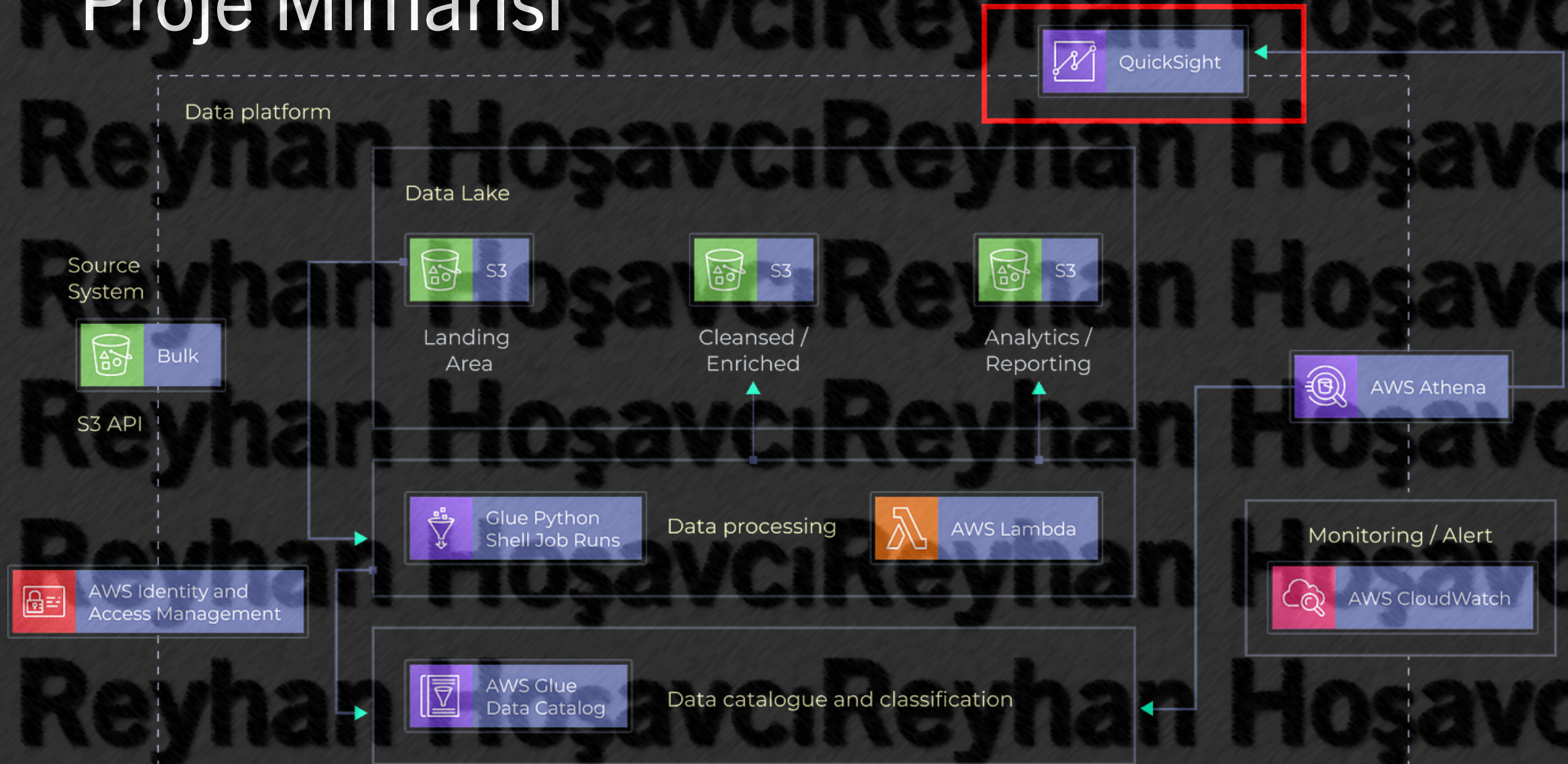
Find objects by prefix

<input type="checkbox"/>	Name	Type
<input type="checkbox"/>	category_id=1/	Folder
<input type="checkbox"/>	category_id=10/	Folder
<input type="checkbox"/>	category_id=15/	Folder
<input type="checkbox"/>	category_id=17/	Folder
<input type="checkbox"/>	category_id=19/	Folder
<input type="checkbox"/>	category_id=2/	Folder
<input type="checkbox"/>	category_id=20/	Folder


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

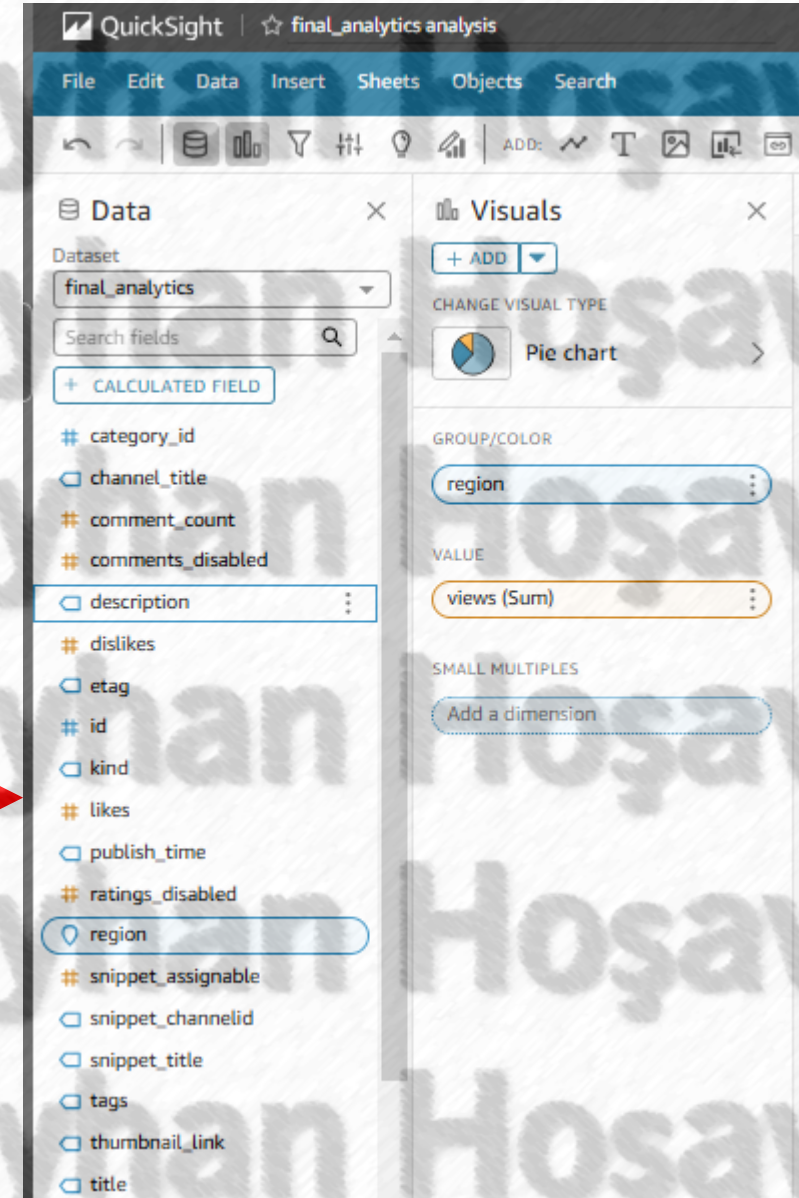
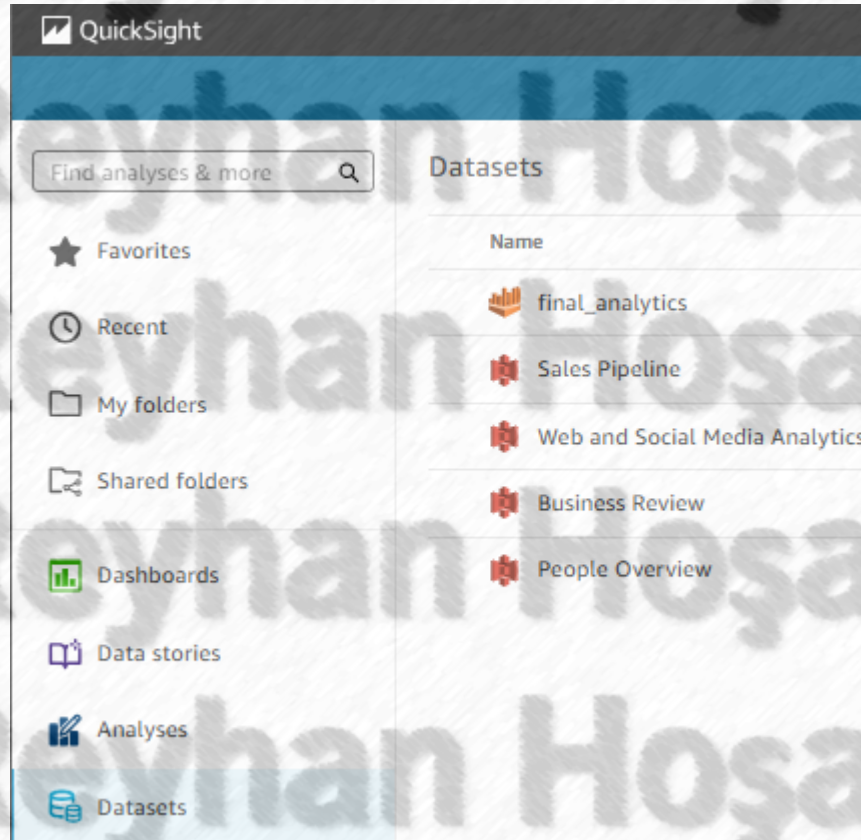
Query results						Query stats	
✔ Completed						Time in queue: 103 ms	Run time: 1.147 sec
Results (10)						Data scanned: 209.18 KB	
<input type="text" value="Search rows"/>						Copy Download results	
<div>< 1 > ⚙</div>							
# ▾	ratings_disabled ▾	comments_disabled ▲	snippet_title ▾	trending_date ▾	etag		
1	false	false	Autos & Vehicles	18.12.01	"XI7nbFXulYBlpL0ayR_gDh3e		
2	false	false	Autos & Vehicles	18.12.01	"XI7nbFXulYBlpL0ayR_gDh3e		
3	false	false	Autos & Vehicles	18.09.01	"XI7nbFXulYBlpL0ayR_gDh3e		
4	false	false	Autos & Vehicles	18.09.01	"XI7nbFXulYBlpL0ayR_gDh3e		
5	false	false	Autos & Vehicles	18.12.01	"XI7nbFXulYBlpL0ayR_gDh3e		

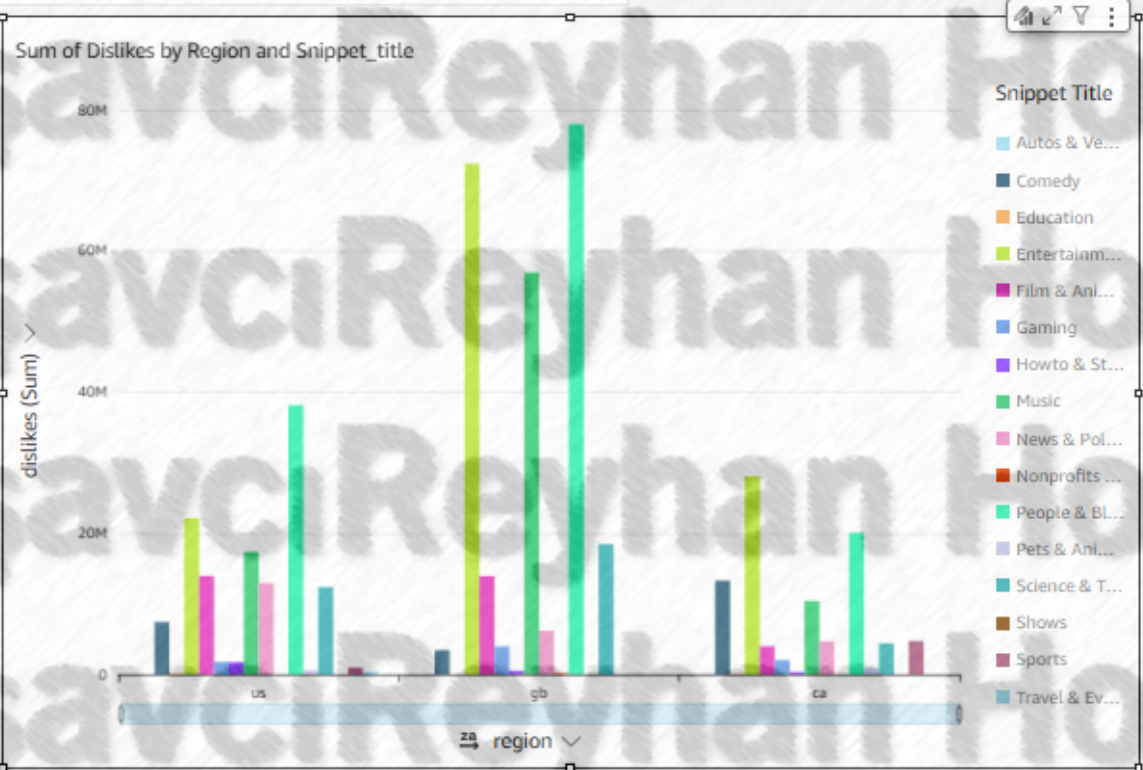
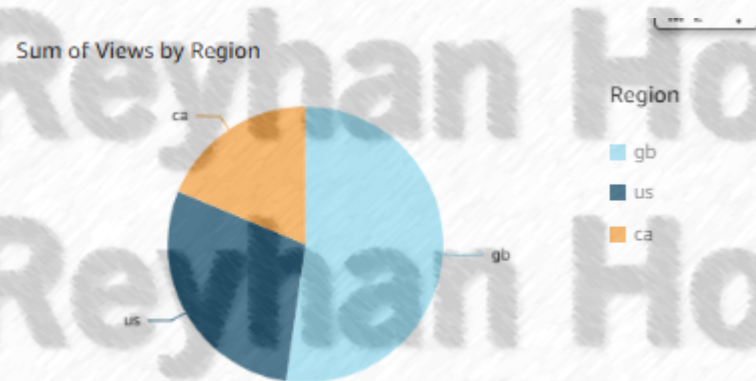
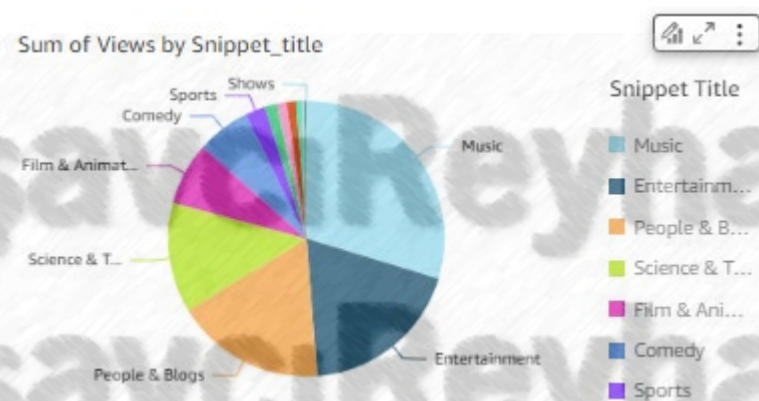
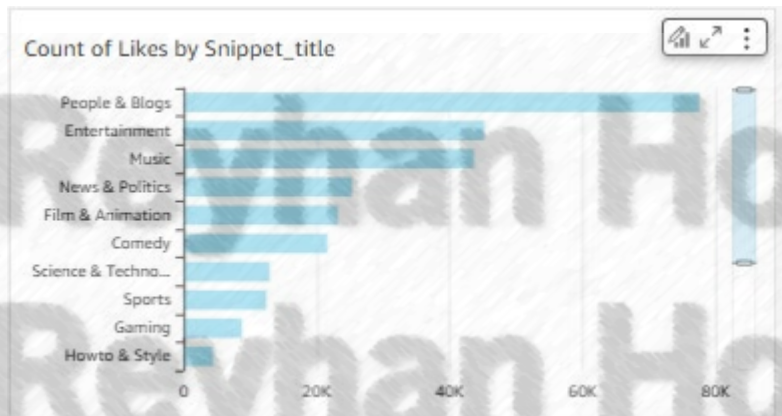
Proje Mimarisi



Quicksight ile analiz sonuçlarının görselleştirilmesi

- Athena kullanarak database aktarıldı





Teşekkürler