Media Stream Analytics

1. Business Challenge/Requirement

Objective

MediaStream Analytics is a big data solution designed to deliver actionable insights from large-scale media consumption data. Using scalable AWS services, it enables real-time and batch processing of viewership logs, ad revenue, and user demographics to optimize programming and monetization.

Key Data Sources

- ***Viewership Logs (Kinesis)***
- ***Ad Revenue Reports (S3)***
- ***Channel Metadata***
- ***Demographic Information***
- ***Cooked PII Data (Hashed via Lambda)***

Below is an abstract of end to end process:

- ***As the viewers are watching channels,Viewership logs are generated and received from an online process which are sent to kinesis for further processing.***
- ***A bigdata processing system captures the log files and send them to snowflake and s3 for further querying***
- ***Demographic , ad revenue and channel metadata are stored in S3 which needs to be cleaned before they are used for further processing***
- ***The data has to be extracted ,transformed and cleaned before it can be used for reporting***

Core Business Questions

1. ***Which channels generate the highest ad revenue per day?***
2. ***What demographic segments generate the most ad revenue per minute?***

3. What are the peak viewing hours across India?

4. How does genre popularity vary by gender?

5. Which channels are most popular among premium subscribers?

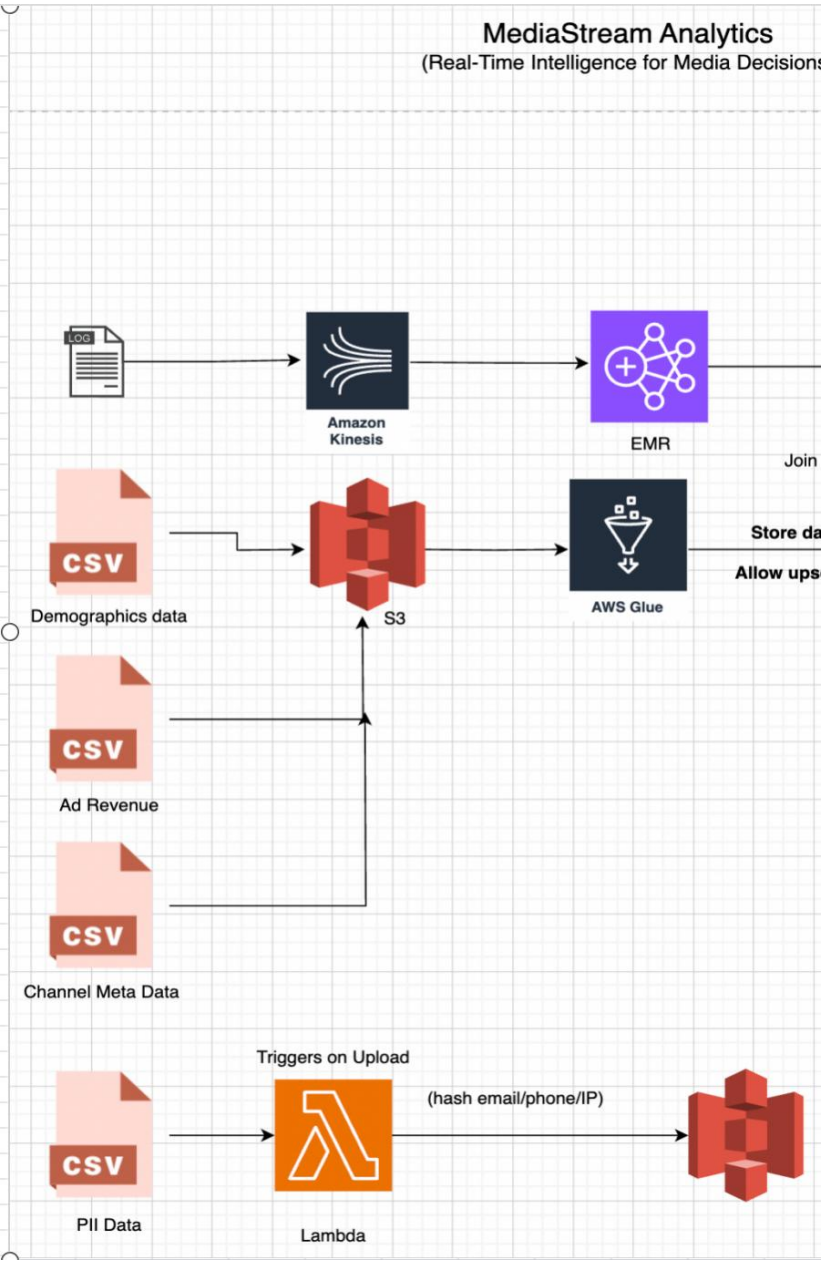
6. Which shows drive higher engagement in female metro audiences?

2. The Goal of the Project

Below are some of the high-level technical and non-technical goals for this project:

- ***Get an overall understanding of the Media domain***
- ***Learn the fundamentals & standards of ETL and data warehousing using spark, snowflake, kinesis***
- ***Real-time and batch ingestion of data from various sources to Big Data and processing them using EMR/Spark and storing them in snowflake and S3***
- ***Using reporting tools like Quick sight***
- ***Lambda architecture where data can be processed in both batch and real-time***
- ***Reporting KPIs(Key Performance Indicators)***

3. Data Flow Architecture/Process Flow



#	Business Question	Required Fields	Data Source(s)
1	Which **age group** watches the most across all channels?	user_id, age_group, channel_id	Demographics, Viewership Logs
2	What is the **average viewing time** per user segmented by region?	user_id, region, duration_minutes	Demographics, Viewership Logs
3	Do **mobile users** watch more content than smart TV users?	user_id, device_type, duration_minutes	Viewership Logs, Demographics
4	What is the **total ad revenue per channelper** day?	channel_id, ad_revenue, timestamp	Ad Revenue Data
5	Which **demographic segment** generates the **highest ad revenue per minute**?	user_id, age_group, gender, region, channel_id, ad_revenue, duration_minutes	Demographics, Viewership Logs, Ad Revenue
6	What are the **peak hours** of viewership across India?	timestamp, region, duration_minutes, channel_id	Viewership Logs, Demographics
7	How does **content genre popularity** vary by gender?	user_id, channel_id, genre, gender	Channel Metadata, Demographics, Viewership Logs
8	Which channels are most popular among **premium subscribers**?	user_id, channel_id, subscription_type	Demographics (or Subscription Table), Viewership Logs
9	Are certain shows driving higher engagement among **female users in metro cities**?	user_id, gender, region, channel_id, show_name, duration_minutes	Demographics, Viewership Logs, Channel Metadata

4. Dataset Explanation and Schema

We have 2types of data sources:

- ***Data stored in S3***
- ***Real-time data for the current date for viewership logs***

4.1 Data stored in S3

We have the below datasets -

Ad-revenue

column Name	*Data type*	*Column description*	*sample value*
Channel_id	*Int*	*Unique Channel id*	
channel_name	*varchar(50)*	*Name of the channel*	*Starsports*
date	*varchar(50)*		
ad_revenue	*double*	*Revenue generated from Ads*	

Channel meta data

column Name	*Data type*	*Column description*	*sample value*
channel_id	*int*		
channel_name	*varchar(50)*		
genre	*varchar(50)*		
language	*varchar(50)*		
launch_year	*int*		

Demographics

column Name	*Data type*	*Column description*	*sample value*
user_id	*int*		
gender	*varchar(10)*		
age_group	*varchar(50)*		
region	*varchar(50)*		
subscription_type	*varchar(10)*		

4.2 Real-time data for the current date for viewship logs

Viewership_logs

column Name	*Data type*	*Column description*	*sample value*
user_id	*int*		
channel_name	*varchar(75)*		
channel_id	*varchar(2)*		
timestamp	*varchar(2)*		
duration	*int*		
region	*varchar*		
subscription	*varchar*		
device	*varchar*		
platform	*varchar*		
is_live	*boolean*		
genre	*varchar*		
ads_watched	*int*		

ad_revenue	*double*		
engagement	*int*		
buffer_count	*int*		
completion_pct	*int*		
session_id	*varchar*		
show_name	*varchar*		

Copy the file - generate_media_data.py to your VM and generate real-time data. This real-time simulator script which will generate json data and writes it to a kinesis data stream

5. ***Problem Statements/Tasks***

1. Generate the realtime data (Viewership logs) and send them to a kinesis stream

2. Intercept this data in an EMR cluster and write them to snowflake and send the stream data in parallel to S3

3.Load the ad revenue, Channel_metadata, and demographics to S3 and use glue to transform this data in to S3.
4.The solution should also allow data to be updated in S3 as soon as the source data is updated for any of the files in ad revenue, channel_metadata or demographics using iceberg tables
5.New files(Updates or new records) pushed to S3 for the above files(step 3), apache airflow(AMAA) should run a glue job to merge the existing records or insert the new records for those tables
6.Use athena to run the required queries

7. Create dashboard for the queries in quicksight with snowflake as the source

Queries to run and visualize

1. Total viewership duration per channel
2. Average engagement by device
3. Daily ad revenue per channel
4. Gender-wise average completion percentage
5. Most watched genres in each region
6. Top 5 channels with highest ad revenue in the past 7 days
7. Peak viewership hours by region
8. Which age group watches the most live content?
9. Subscription type driving most revenue per channel
10. High engagement sessions (more than 90% completion and >1 min)
11. Channels with above-average ad revenue per day
12. Most engaging show for female users in metro regions
13. Genre-wise ad revenue and completion comparison
14. Channels with highest buffer counts but good engagement
15. which age group watches the most across all the channels
16. what is the average viewing time per user segmented by region
17. Do mobile users watch more content than Smart tv users
18. What is the total ad revenue per channel per day
19. what are the peak hours of viewership accross india
20. how does content genre popularity vary by gender
21. which channels are more popular among premium subscribers
22. are certain shows driving higher engagement among female users in metro cities