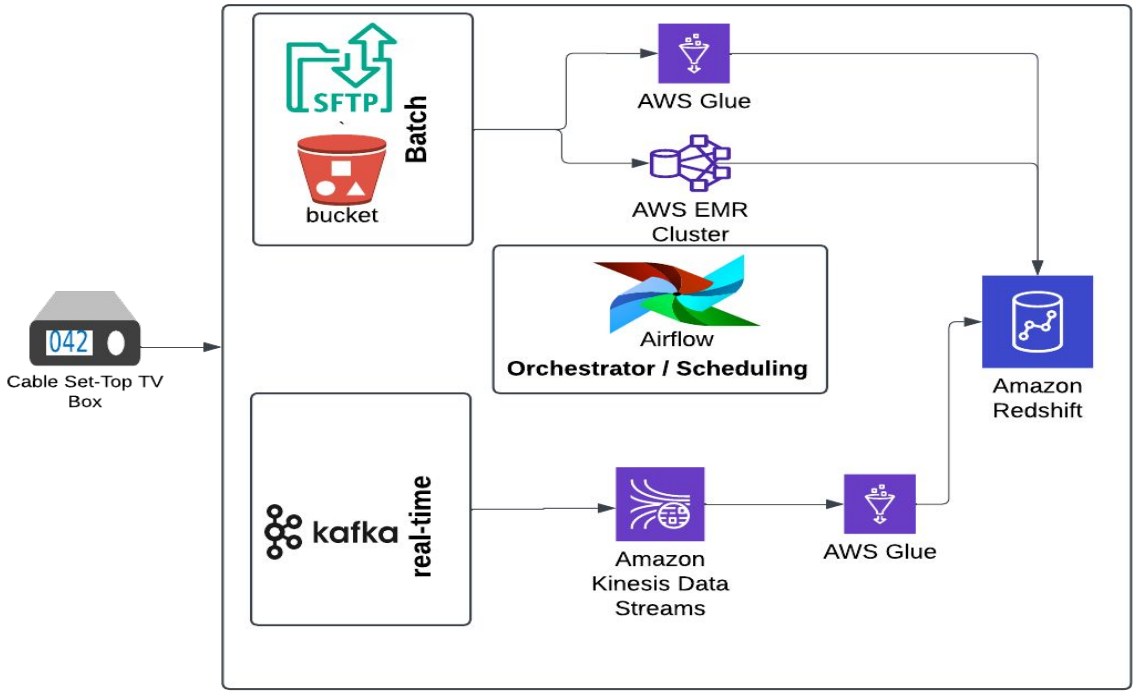
MBC CASE STUDY ASSESSMENT

Submitted by Sahil kukreja

TV Viewership Data Processing Pipeline

Real-Time and Batch ETL Solution in AWS



Architecture Components

Components:

1. **Data Sources**:
   - TV viewership logs, metadata, and demographic data from sources like CSVs, JSON, or APIs.
2. **ETL Layer**:
   - Extraction: PySpark handles data ingestion from source files.
   - Transformation: Cleaning, joining, and enriching data.
   - Loading: Load into a staging area and transform into warehouse tables.
3. **Data Warehouse**:
   - **Staging Layer:** Raw ingested data.
   - **Core Layer:** Transformed and business-ready data.
   - **Analytics Layer:** Aggregations and summaries for business analysis.
4. **Orchestration**:
   - Apache Airflow: Workflow scheduling and monitoring.
5. **Downstream Integration**:
   - Business Intelligence Tools: Dashboards and reports.
   - Machine Learning Pipelines: Predictive analytics.

Design Decisions

1. **Hybrid Data Processing (Batch and Real-Time)**:
    - **Batch Processing**: Data from an SFTP server is ingested into an S3 bucket for periodic processing. This approach supports the processing of large datasets at scheduled intervals.
    - **Real-Time Processing**: Kafka is used for streaming data, which is processed continuously, enabling real-time insights.
2. **AWS Glue**:
    - Acts as an ETL service for data transformation and cataloging.
    - Simplifies schema management and ensures compatibility with downstream systems like Amazon Redshift.
3. **AWS EMR (Elastic MapReduce)**:
    - Used for large-scale distributed data processing (e.g., Spark or Hadoop workloads).
    - Chosen for its cost-effectiveness and ability to handle complex data processing at scale.
4. **Amazon Kinesis Data Streams**:
    - Facilitates real-time data ingestion and processing from Kafka.
    - Integrates seamlessly with AWS Glue for transformation.
5. **Amazon Redshift**:
    - Serves as the data warehouse for storing processed and transformed data.
    - Optimized for analytical queries and reporting.
6. **Apache Airflow**:
    - Acts as the orchestrator, managing workflows and scheduling jobs.
    - Enables seamless coordination between batch and real-time processes.

Tools Used

1. **AWS S3**: Storage for raw data from SFTP, ensuring durability and availability.
2. **Apache Kafka**: Industry-standard streaming platform for real-time data pipelines.
3. **AWS Glue**: Simplifies ETL processes with serverless scaling and integration with other AWS services.
4. **AWS EMR**: Handles batch processing for large datasets.
5. **Amazon Redshift**: Centralized data warehouse for querying and analytics.
6. **Apache Airflow**: Workflow orchestration for monitoring and scheduling.

Future Scalability Considerations

Data Volume:

- S3 and Kinesis can handle growing data volumes. AWS services like Glue and Redshift support horizontal scaling to meet increasing demands.

Integration with Additional Data Sources:

- Kafka supports various connectors, enabling integration with new real-time sources.
- AWS Glue and S3 can accommodate diverse data formats and sources.

Real-Time Analytics:

- Use Amazon Redshift Spectrum or AWS Athena to query data directly in S3 for ad-hoc analytics.
- Add Amazon QuickSight for visualization and reporting.

Fault Tolerance and Availability:

- Leverage multi-AZ setups for critical services like Kafka and Redshift.
- Use AWS Step Functions for better error handling and state management in workflows.

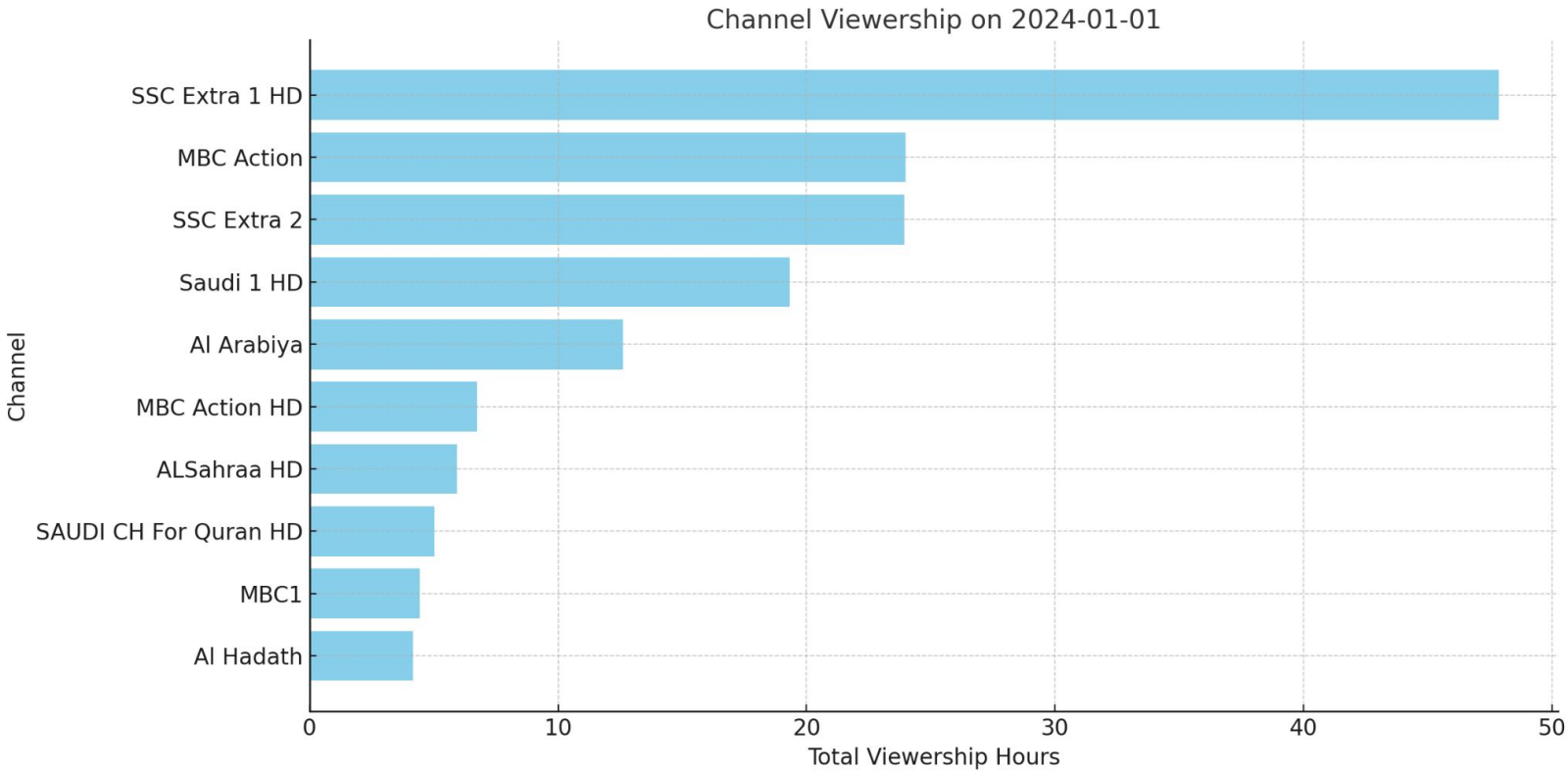
Cost Optimization:

- Utilize spot instances in EMR for batch jobs.
- Implement lifecycle policies for S3 to move less frequently accessed data to cheaper storage classes like Glacier.

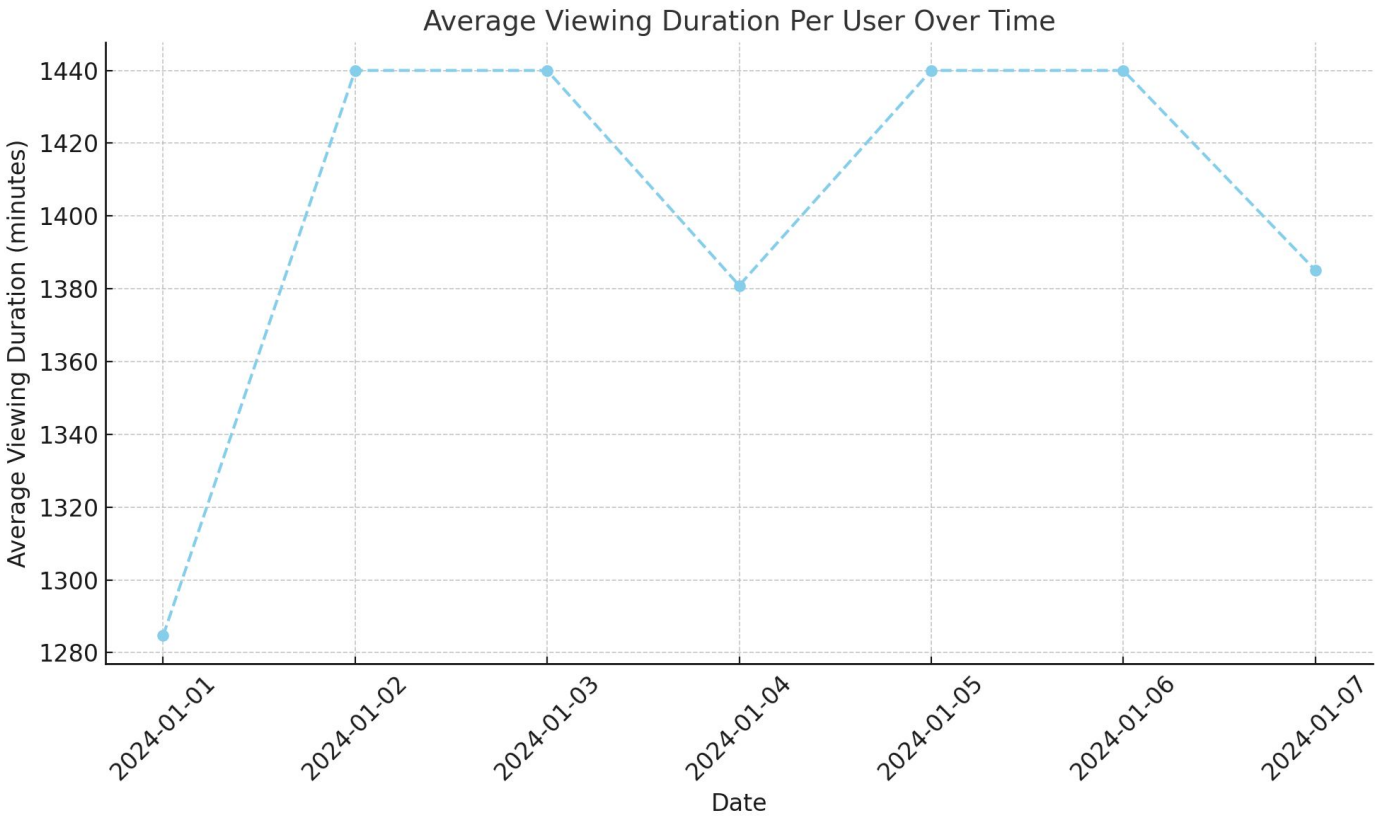
Security Enhancements:

- Integrate AWS IAM roles and policies for secure access.
- Use encryption for data at rest (S3, Redshift) and in transit (SFTP, Kinesis)

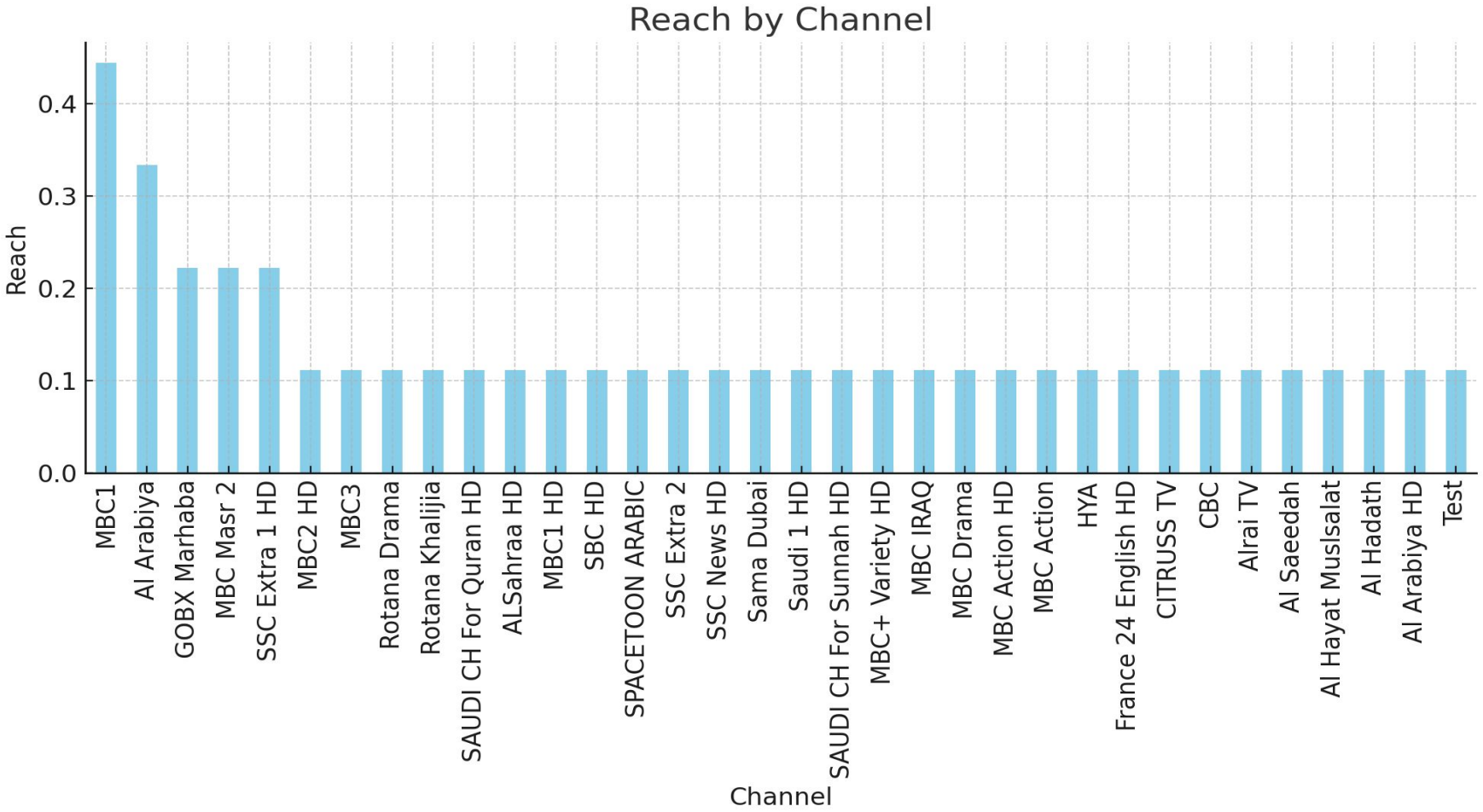
Key Insights: Top 10 viewed channels on 2024-01-01



Key Insights: Average Viewing Duration Per User Over Time



Key Insights: Reach of channels on 2024-01-01



Key Insights: TRP channels on 2024-01-01