clearscale

aws partner network

Premier Consulting Partner

SmugMug Gains Robust Cloud Data Infrastructure and Data Pipeline

SmugMug

Executive Summary

SmugMug is an all-in-one photography sharing and hosting service that enables users to store, protect, share, and sell their photos on a single platform. The site also comes with website-building capabilities, giving creators flexibility over how they share their art.

After acquiring the photo production and management application, Flickr, SmugMug needed specialized cloud support around setting up scalable data repositories and pipelines for advanced analytics down the road. Fortunately, ClearScale had the Amazon Web Services (AWS) expertise to design the ideal data ecosystem for SmugMug.

> "Our engagement with ClearScale has been essential to achieving our goal of accelerating our data strategy. ClearScale helped us to build a new end-to-end data pipeline that will ingest all our data in real-time, giving us confidence that all of our data is in the right place, precisely when needed."
>
> **Scott Kinzie,** VP of Marketing and Business Development, SmugMug

The Challenge

Back in 2018, SmugMug acquired another media industry giant - Flickr. Through the acquisition, Flickr's transactional workloads were migrated from a sharded MySQL cluster to Amazon Aurora as an OLTP. Analytical workloads required a temporary workaround - SmugMug leveraged Aurora read replicas, which couldn't substitute fully-featured OLAP, nor handle the hundreds of terabytes that Flickr gathered. The company had to figure out a better long-term solution.

SmugMug also didn't have the infrastructure and tools to mine its data in a timely manner. To enable advanced analytics, SmugMug needed a data pipeline that could transform data appropriately for further processing without increasing costs or administrative overhead.

The company decided to reach out to ClearScale, an AWS Premier Consulting Partner with extensive experience in designing and implementing cloud-native data infrastructure.

The ClearScale Solution

SmugMug had three specific objectives in mind for ClearScale:

1. Build an AWS-native data warehouse to aggregate data for advanced analytics
2. Replace legacy third-party and home-grown data analytics and business intelligence (BI) tools, and seamlessly integrate them with the new warehouse and lake
3. Ingest existing transactional data into the data warehouse and establish ongoing replication for newly created, updated, or deleted entries with minimal delay

With the goals set, ClearScale got to work.

Implementing the Data Warehouse and Data Lake

ClearScale implemented Amazon Redshift as the foundation of SmugMug's new data platform. Redshift provides cost-effective data warehousing through a fully managed offering, which means SmugMug's engineers don't have to worry about provisioning servers, deploying patches, or ongoing maintenance. AWS takes care of these back-end administrative tasks. In addition, Redshift is entirely compatible with PostgreSQL, enabling the SmugMug team to reuse tools, queries, and code snippets they already knew well.

Next, ClearScale used AWS Lake Formation to add a data lake on top of Amazon S3 for object storage purposes. AWS Lake Formation was valuable because the service makes it easy to implement permissions-based security through AWS Identity and Access Management (IAM). with constructs not found in other solutions like columnar-level access controls. ClearScale configured SmugMug's data lake to store information in columnar formats like ORC and Parquet so that only the data needed for specific queries gets read, significantly decreasing related costs.

Adding the Analytics and Visualization Layer

For the analytic engine, ClearScale used Amazon Athena, a serverless, interactive query service that allows users to analyze their data stored in Amazon S3 using standard SQL. ClearScale added Amazon QuickSight alongside Athena to give SmugMug a BI tool capable of scaling quickly and publishing interactive dashboards. Like Athena, QuickSight doesn't require users to deploy or manage any new infrastructure.

Both of these tools integrated seamlessly with SmugMug's new data warehouse and data lake. The company now had the infrastructure necessary to aggregate all of its data, run complex queries, and publish findings directly to easy-to-understand visuals. Next, ClearScale turned to the ingestion side of SmugMug's data pipeline.

Setting Up Data Pipeline

Before SmugMug could mine any data, it had to first be ingested into the data warehouse from somewhere. In SmugMug's case, this somewhere was a sharded Amazon Aurora cluster. In this service Amazon implemented the same programmatic interfaces available in MySQL, allowing straightforward change data capture on top of the binary log (binlog). The initial data ingestion was executed via AWS Glue, the only serverless Spark offering available on the market.

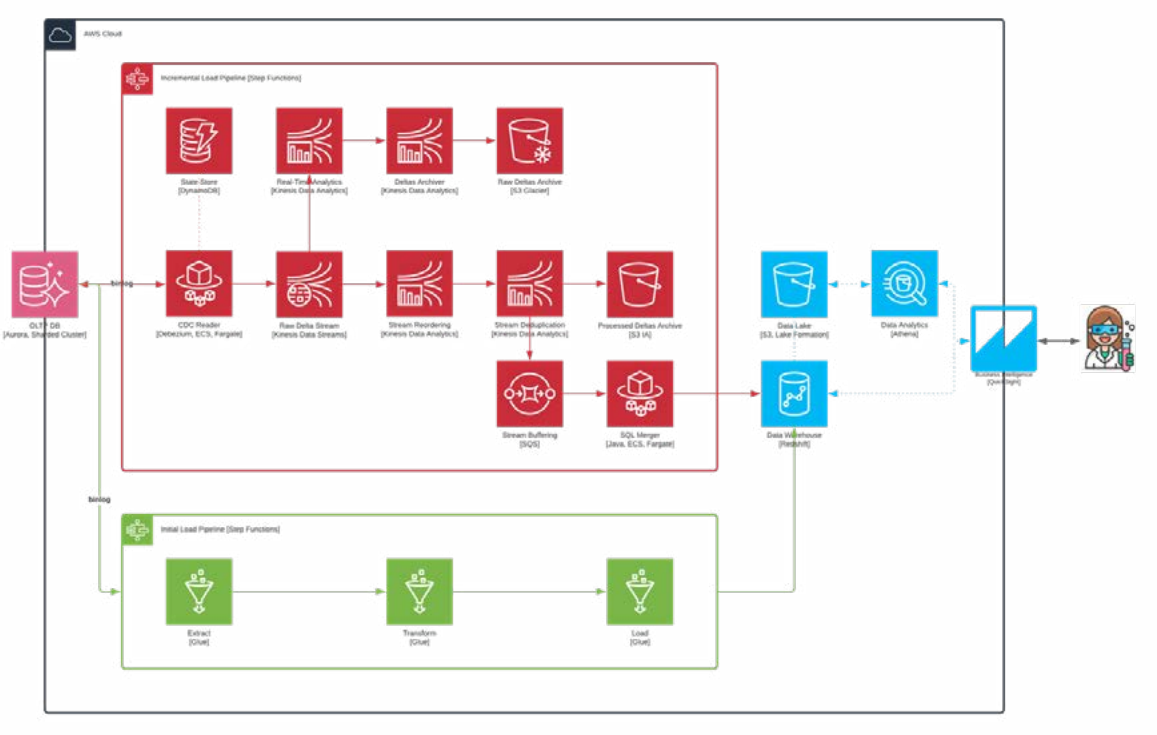
For ongoing ingestion, the ClearScale team used a relatively new open-source tool, Debezium, as well as AWS Fargate, a serverless compute engine designed to support Docker containers. Together, these services provided SmugMug with an elastic rebalancing capability and increased resiliency.

ClearScale then augmented Debezium to write passed deltas into Amazon Kinesis Data Streams, which served as a buffer layer to ensure every delta is saved before getting ingested into the data warehouse. At this point, ClearScale designed the data pipeline to fork into two branches.

The first supports real-time aggregates and queries against the raw data stream, as well as archiving into Amazon S3 Glacier to meet compliance and regulatory requirements. The second is more complex - it uses Amazon Kinesis Data Analytics as a platform for Apache Flink to organize data as it comes in and deduplicate identical entries. The second branch also performs a series of custom transformations to merge related changes and alleviate some of the downstream pressure on Redshift.

SmugMug's data pipeline is orchestrated by AWS Step Functions, inspired by Apache Airflow, which Flickr used extensively in the self-managed OLAP. The serverless tool provides stateflow visualization, automatic exceptions retry, and native integration with most AWS services and many third-party tools. With the pipeline in place, ClearScale's job was done.

Architecture Diagram



The Benefits

Thanks to ClearScale's help, SmugMug now has an end-to-end data pipeline that reliably and cost-effectively ingests various types of data in real time. All of this data is stored in a secure data warehouse and data lake upon which analysts can run ad-hoc queries and leverage BI tools from anywhere in the world. Furthermore, the company's in-house development team no longer has to worry about whether or not data gets where it needs to go.

Overall, this data platform is not just a solution for old problems, but a sound base for future growth - from revenue forecasting Machine Learning models to ingestion of clickstream data right from the app.