o·o·y·a·l·a™

Designing a Scalable Database for Online Video Analytics

White Paper

Online Video Analytics

A deep understanding of how video content is consumed is critical to maximizing revenue and viewership for online video content. Publishers are increasingly looking for sophisticated ways to monetize and syndicate their video to ensure they are reaching the largest number of viewers and maximizing the value of their online video content. Video analytics is the key to unlocking the full potential of online video content, and leveraging the proper technology and infrastructure to manage the vast amounts of viewership data is an important component for an online video analytics platform.

There are a vast number of possible databases to choose from to store viewership data, ranging from the more traditional relational databases (such as MySQL) to distributed databases (like Cassandra and HBase). Traditional relational databases fall short when confronted with the terabytes of data generated from online video. Forcing a relational database model on these large-scale datastore problems often results in sharding architectures that are difficult to maintain and scale. Fortunately, there are alternate solutions designed to more effectively handle large amounts of loosely structured data like data generated from online video viewership.

Designing a good online video analytics system

Before exploring some of the alternatives to traditional relational databases, it's useful to take a closer look at some of the key requirements for an online video analytics database. They include:

- **Scalability**: The database should scale to handle terabytes and even petabytes of data.
- **High Availability**: The database should strive to exhibit 100% up-time and be robust to node failure.
- **Flexibility**: The database should allow a user to easily add new types of analytics information.
- **Real-Time**: The database should allow a user to quickly write and read data.

Scalability

Every viewer watching content generates a wealth of data which can be captured and ultimately leveraged to drive higher engagement with content or maximize monetization opportunities. Consider a typical viewing experience and some of the relevant session information generated:

- The geographic location of the viewer (City, State, Country, DMA, etc)
- How much of the video was watched
- Which other videos were watched during the session
- The page the video was watched on
- When they watched the video

Two primary factors are driving the growth of online video analytics data. First, consumers continue to view more online video content which is resulting in a commensurate increase in the volume of viewership analytics data. Second, publishers require more granular data in order to drive more sophisticated monetization campaigns and drive higher viewership of their content. Increasing the granularity and detail of analytics data also results in increased storage requirements.

High Availability

Publishers rely on their data being available 100% of the time. Losing viewership information or the inability to access viewership data often has a direct impact on revenue. A highly available database should be robust to and gracefully handle the failure of any particular database node.

Flexibility

The online video space continues to evolve at a rapid pace and publishers will require increasingly sophisticated metrics and analyses to ensure they are optimizing every aspect of the online video experience. The database must be flexible enough to tackle the known challenges of today, but easily scale to handle unknown challenges facing publishers tomorrow.

Real-Time

Viral videos can generate hundreds of thousands of hits in a single day. In addition, live events have a short life cycle, typically several hours, in which they can generate a large number of viewers. Understanding the performance of your video in real-time allows publishers to maximize revenue and viewership to ensure content is reaching the largest possible audience.

Building a scalable, flexible, real-time analytics engine using Cassandra

Some of the problems confronting online video analytics systems are very similar to those confronting other web properties dealing with large amounts of loosely structured data. Many of these sites are choosing to migrate away from a more traditional relational model and instead leverage a distributed datastore.

Cassandra

Cassandra is a distributed database based on both Google's BigTable and Amazon's Dynamo architecture. Facebook released the first version of Cassandra to the open-source community in 2008. Cassandra leverages a peer-to-peer architecture to distribute and duplicate data across numerous nodes resulting in both a highly available and robust system. Several of the largest web properties today leverage Cassandra to store user data and help drive their sites, including Facebook, Digg, and Twitter.

Scalability

Cassandra was designed for horizontal scalability, meaning that as storage requirements increase, new nodes can be added to both increase the amount of available storage and to increase the throughput the system in a linear fashion. This differs from more traditional relational databases which often require complex master-slave or sharding architectures at scale and tend to be difficult to implement and maintain.

High-Availability

The distributed nature of Cassandra means that the failure of any particular node will not necessarily result in the loss of viewership data or the ability to write new data analytics data. Node failure is inevitable, and maintaining multiple copies of information across different nodes reduces the risk of permanently losing viewership data.

Flexibility

Cassandra is a loosely structured database meaning that it has some of the advantages of a structured representation, but also that it remains flexible and allows users to easily add additional columns of data.

Real-Time

The distributed nature of Cassandra increases the speed of both reading information from the database and writing information to the database because different nodes are responsible for different pieces of data.

> *"We spent a lot of time evaluating different databases for our viewership data which would extend our current analytics solution and allow us to scale as publishers deliver more video content. Cassandra ended up being the best choice and is allowing us to continue to build innovative products around analytics and monetization."*
>
> *Edmond Lau, Technical Lead of Analytics and Monetization Team, Ooyala*

Analytics are enabling the future of online video

Choosing the right database to store analytics information is a critical decision for all video companies looking to collect extensive viewership information capable of scaling with increasing online video consumption. Traditional relational approaches are well suited for smaller-scale storage and retrieval but become cumbersome as terabytes and petabytes of viewer information is stored. Analytics will become an increasingly critical part of a publisher's online video strategy and choosing the right system today will enable the next generation of online video products and features. As a result of our migration to Cassandra, Ooyala will soon be introducing a number of exciting products including advanced content recommendations, hyper-targeted advertising campaigns, and tools for rapidly maximizing your overall monetization strategy.