IAC Publishing Labs Turns Data in to a Business Accelerator with Snowflake on AWS

IAC PUBLISHING LABS

About IAC Publishing Labs: One Company – Global Websites

Headquartered in Oakland, California's vibrant tech district, IAC Publishing Labs is a digital media and incubator group within the IAC Publishing Labs organization. IAC Publishing Labs is one of the largest collections of premium publishers on the web. Web properties supported and operated by IAC Publishing Labs includes Ask.com, shopping.net, shop411.com, reference.com, informationvine.com, and consumersearch.com.

In all, IAC Publishing Labs creates the digital experiences that entertain the millions of people throughout the world visits those sites.

The business intelligence (BI) team of IAC Publishing Labs provides centralized analytics for data across several world wide web publishing businesses. Through the main web pipeline alone, the BI team manages more than 300 million events, 50-100 million keywords and marketing terms for bidding and monetization, and imports more than 1.5 terabytes of raw data every day. Previously IAC Publishing Labs stored six months of completely processed data in the main data warehouse (roughly 50-60 terabytes of analyzed data).

Along with ingesting and analyzing the huge volume of data, the BI team must adopt and incorporate different data sources on a frequent basis as IAC Publishing Labs expands through the acquisition of new websites.

"The executive level wants the ability to see initiatives and detailed progress across all these businesses," said Erika Bakse, Head of BI, IAC Publishing Labs.

Our Challenge: A Rigid System that Could Not Keep Pace with the Growing Business

As web properties and acquisitions bloomed and expanded, IAC Publishing Labs had a data environment that was inflexible, difficult to modify, unstable, and unable to keep paces with the growing business. Their legacy environment didn't allow them to monetize the data fast enough for advertising technology and retargeting purposes, thus not allowing IAC Publishing Labs to meet the needs of their internal customers. As a result, the reputation of the BI team suffered as they attempted to service many global companies.

As the business grew, the challenge faced by the team was scaling the data warehouse to keep pace. This legacy data warehouse system was unstable. There was a high propensity for the cluster to shut down when running a query longer than 30 minutes. "When we ran a query longer than 30 minutes, there was a 70% chance the (on-premises) database would shut down," said Bakse.

Breaking Away from Inefficiencies and Complexities to Process Diverse Data

Originally, the architecture to manage the large volume of data comprised of two separate on-premises systems: a large MPP warehouse and an even larger Hadoop cluster. In addition, the processing of JSON data required complex pre-processing, where scripting technologies such as C, Python, and Perl were required simply to coalesce and manage a dataset. Additionally, considering the size of the environment, it became hard to create and manage a development environment that would

Previously IAC Publishing Labs stored six months of completely processed data in the data warehouse (roughly 50-60 terabytes of analyzed data).

allow IAC Publishing Labs to effectively test changes.

"We had a business intelligence environment that was not looked well upon by the business units," said Bakse, the leader of a six-person team.

Due to this complex environment, IAC Publishing Labs faced the following challenges:

- Rigid, time-consuming processes
- An inability to scale the size of their environment
- Poor quality of data as changes moved into production haphazardly
- A legacy system that often froze while servicing requests
- Lack of proliferation in the user community

Creating a Path for Better BI Customer Service

During the process of re-evaluating its complete data environment needs and requirements, the BI team took a long, hard look at how data was being consumed by internal users. The team quickly realized they needed to think of data and its processing in different terms. "We determined that scaling and maintaining a legacy environment that could meet the wide range of consumption patterns of our analysts was cost prohibitive and would have taken years to roll out," said Erika Bakse. It was concluded and quite clear that the existing approach had to change.

The BI team proceeded to identify 120+ metrics that would help evaluate a next-generation environment to replace the legacy environment. They ran an internal proof-of-concept (POC) and benchmark testing that included Snowflake on Amazon Web Services (AWS) and other cloud data warehouse alternatives. Considering the large legacy investment, the BI team also included the incumbent vendor, to determine if a managed service approach on the existing BI technology stack would work. In the end, the BI team realized that a move to an 'as-a-service' architecture provided the most value to the business and would meet its scalability needs.

> With Snowflake and AWS, the BI team retired the 36 node MPP data warehouse and a Hadoop cluster, providing one vision of the truth and insights for sites managed by IAC Publishing Labs.

"After completing our evaluations and choosing Snowflake Elastic Data Warehouse on AWS, we were in full production in just three months," stated Erika Bakse.

Snowflake on AWS Offers a New Way to Look at Data

IAC Publishing Labs now has a single environment for processing data and producing results. This enables the team to directly query JSON data from the web logs, allowing the team to pinpoint logging, instrumentation errors, and errors from specific pages in near-real time. Functionality to query native JSON has simplified, and the team uses SQL in the database and processes data by spinning up different-sized warehouses as required.

Scalability with Greater Control on the Cloud

"Snowflake gives us more comfort and clarity around data quality," said Bakse. "With one system processing data and producing results, the team is able to pinpoint logging, instrumentation errors, and errors from specific pages. This added visibility enables the team to make adjustments for a superior data warehouse."

Snowflake's innovative data warehouse architecture built on AWS enables the team to spin up as many new warehouses as needed (for production or development) using production data. Having a non-production environment consistent with production data and using the cloning feature of Snowflake enabled by Amazon Simple Storage Service (Amazon S3), provides the IAC Publishing Labs

BI team greater control in building processes around their code and creates a higher-quality final product with fewer bugs and better-quality data. Moreover, because Snowflake enables new dev-and-test environments to use production data without the overhead of copying and managing another physical data warehouse, additional or more thorough testing can be completed.

The current platform efficiently deals with concurrency, through the use of Snowflake virtual warehouses enabled by unlimited processing capabilities of the Amazon Elastic Compute Cloud (Amazon EC2) on AWS. It also increases system stability by providing separate but controlled access to new or inexperienced users. The legacy system was brittle; any user could log on and write a job that could render the cluster useless for other jobs and engineers. Today, the BI team can effectively extend the cloud data warehouse environment to new users and scale the environment while aligning its user base growth.

"Migrating to and working with Snowflake has been so remarkably smooth and effortless that my CTO is in disbelief." said Erika.

Enhanced Service Levels through Better Consistency and Responsiveness

"Consistency of performance and concurrency has been a huge win for us," said Bakse. The BI team can now have 30 or more analysts concurrently query the environment with no issues. The team can also load the data 24 hours a day, 365 days a year, without causing contention across time zones or impeding the work of groups across the globe.

These capabilities limit the amount of change that is visible to users, which has increased stability and has built trust in the environment. Overall system adoption and use are on a steady upward trajectory.

A greater volume of data is available with the new system. The BI team has gone from storing approximately six months of data to keeping two years of data. They have lifted former retention policies—policies that were put in place because of performance and instability concerns.

In addition, queries process far more quickly, enabling the BI team to provide a higher level of service with greater efficiency.

Instead of loading data every five hours, the BI team can load data every 15 seconds. Instead of processing data once a day for three to four hours, they now process data every hour, in as little as five to 10 minutes. As a result of this acceleration, the BI team can bring new businesses on-board very quickly and start to monetize data. "The ability of the BI team to pull data into the data warehouse from various sources and combine the data very easily with the overall business picture has significantly increased the BI team's credibility and reputation that we can keep up with the rest of the business," said Bakse.

A Transition from a Cost Center to a Value Center

The team now has one elastic data warehouse, a single source of truth, with a centralized data warehouse model that gives real-time visibility across businesses. Snowflake on AWS holds both internal and external data together, serving both data scientists and business analysts. Because they now have a fast and flexible data warehouse, business intelligence is moving from a being cost center to being a value center. "We are acting in an advisory role regarding how to best get the data to work for the organization," said Bakse.

The increased flexibility has allowed the team to quickly integrate new data sources and roll them into new metrics, at the speed of business. With the scale that the new open platform provides, it's easy for many different businesses to join. The new platform allows the BI team to provide value to the overall business without incurring major capital expenses and infrastructure rollouts.

"Because the data warehouse platform is managed in the cloud as a service by a team of expert resources, our internal BI team can be wholly committed to providing value to the business," said Bakse. "No longer do we burn cycles on hardware disk replacements, backup strategies and software upgrades. Instead, we find ourselves more

embedded with website general managers' day-to-day lives, working on metric definitions, market opportunities and product roadmaps." The perception of the BI team has changed to a fast-moving, fluid organization. They are now considered a consultative partner integral to our business growth rather than a necessary technical team.

By choosing Snowflake on AWS, IAC Publishing Labs achieves impressive achievements that include:

- Establishing one source of truth in a centralized data warehouse that serves both data scientists and business analysts across all the company's businesses
- Consolidating technologies and eliminating legacy platforms
- Providing enhanced BI service levels through superior consistency and responsiveness of the data environment
- Changing the BI team from a cost center to a value center and making it a respected, often-consulted resource within the company

About Snowflake:

Snowflake Computing, the cloud data warehousing company, has reinvented the data warehouse for the cloud and today's data. The Snowflake Elastic Data Warehouse is built from the cloud up with a new architecture that delivers the power of data warehousing, the flexibility of big data platforms and the elasticity of the cloud – at a fraction of the cost of traditional solutions. Snowflake is headquartered in Silicon Valley and can be found online at snowflake.net.

About AWS:

For 10 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud platform. AWS offers over 70 fully featured services for compute, storage, databases, analytics, mobile, Internet of Things (IoT) and enterprise applications from 35 Availability Zones (AZs) across 13 geographic regions in the U.S., Australia, Brazil, China, Germany, Ireland, Japan, Korea, Singapore, and India. AWS services are trusted by more than a million active customers around the world – including the fastest growing startups, largest enterprises, and leading government agencies – to power their infrastructure, make them more agile, and lower costs. To learn more about AWS, visit http://aws.amazon.com.