



Zillow Increases Accuracy of 'Zestimates' Using Amazon Kinesis



"

We can compute Zestimates in seconds, as opposed to hours, by using Amazon Kinesis Data Streams and Spark on Amazon EMR. As a result, the Zestimates are more up-to-date and accurate, because they're built with the absolute latest data. That's a huge benefit for our users, who depend on this information to influence their buying or selling decisions.

Jasjeet Thind

Vice President of Data Science and Engineering, Zillow Group

"

Supporting a Fast-Growing Home-Valuation Tool

The most popular online real-estate website and mobile app in the United States, [Zillow](#) provides updated home information to tens of millions of buyers and sellers every day. A primary feature of the Zillow website is the Zestimate—a home-valuation tool that provides buyers and sellers with the estimated market value for a specific home. Zillow currently offers Zestimates for more than 100 million homes in the U.S., with hundreds of attributes for each property.

Zillow uses a wide variety of public-record data—including tax assessments, sales transactions, images of homes, MLS listing data, and other information provided by homeowners—as inputs to its Zestimate algorithm. Because Hadoop and other distributed big-data technologies didn't exist more than 10 years ago, Zillow built an in-house machine-learning framework that ran on-premise to process vertically scaling workloads. However, as the amount of data has grown and the complexity of machine-learning models for accurate Zestimates has increased, the company struggled to provide timely Zestimates for all new

homes. “With more than 100 million homes and millions of different machine-learning models, we got to the point where we couldn’t scale fast enough to meet that growth using our existing technology,” says Jasjeet Thind, vice president of data science and engineering at Zillow Group.

The company specifically sought a distributed platform, which would enable the fast creation and execution of massively parallel machine-learning jobs. “We wanted to be able to do machine learning on multiple nodes, because it was taking way too long to do it previously,” says Thind. “It sometimes took more than a day to compute Zestimates, which meant our customers weren’t getting updated information fast enough.”

Using Amazon Kinesis, Apache Spark, and Amazon EMR to Power Zestimates

Zillow initially moved its website to Amazon Web Services (AWS) to address image-system scalability, performance, and disaster-recovery challenges. Then, it decided to expand its use of AWS to solve the scalability and performance problems it faced with the Zestimate tool. Zillow chose to run Apache Spark on [Amazon Elastic MapReduce](#) (Amazon EMR).

By running Zillow’s machine-learning algorithms using Spark on Amazon EMR, Zillow can quickly create scalable Spark clusters and use Spark’s distributed-processing capabilities to process large data sets in near real time, create features, and train and score millions of machine-learning models. “Spark on Amazon EMR appealed to us because it means we don’t need to manage Spark clusters ourselves,” Thind says.

Zillow uses [Amazon Kinesis Data Streams](#) to ingest a variety of data, including public-property records, home tax assessments, sales transactions, images and video, MLS-listing data, and user-provided information. All this data is ingested and pushed into Spark on Amazon EMR, which runs machine-learning models and gives users near-real-time Zestimates. The same dataset is sent in parallel from Amazon Kinesis Data Streams to [Amazon Kinesis Firehose](#), which batches the data in 15-minute intervals and delivers it to Zillow’s centralized data lake on [Amazon Simple Storage Service](#) (Amazon S3). From the data lake, a number of applications such as personalization, advertising optimization, and recommendations use the data without storage scalability concerns.

Overall, Zillow manages petabytes of data in its Amazon S3 data lake.

Increasing the Speed and Accuracy of Zestimates

Relying on Spark on Amazon EMR, Zillow can execute massively parallel machine-learning jobs on a distributed platform. As a result, the organization can run distributed machine learning across multiple nodes to calculate Zestimates. “Previously, given our scale limits using existing proprietary technology, it could take us an entire day or longer to compute a Zestimate,” says Thind. “Now, we can do it in hours nationwide using Spark on Amazon EMR, which enables us to quickly perform calculations in parallel on multiple machines.”

Zillow can compute Zestimates faster and more frequently, because Amazon Kinesis Data Streams and Spark on Amazon EMR enable near-real-time data processing. “We can compute Zestimates in seconds, as opposed to hours, by using Amazon Kinesis Data Streams and Spark on Amazon EMR,” Thind says. “As a result, the Zestimates are more up-to-date and accurate, because they’re built with the absolute latest data. That’s a huge benefit for our users, who depend on this information to influence their buying or selling decisions.” Also, using Amazon Kinesis Data Streams, Zillow does not have to be concerned with managing and scaling a fleet of servers for ingesting real-time streaming data.

Scaling to Support Growth

In addition, Zillow can easily scale its platform as it continues to grow. “We can put raw and historical data in one place using Amazon S3, so we now have a central place to enable infinite storage scalability at a low cost,” says Thind. “We can also get easy compute scale using Spark on Amazon EMR.”

Based on its success with Zestimate, Zillow is now using Amazon Kinesis Data Streams and Spark on Amazon EMR to power machine learning for all Zillow Group brands. “We are using Amazon Kinesis Data Streams and Spark on Amazon EMR to do shopping personalization, ad targeting and optimization, and advanced analytics and machine learning across the company,” says Thind. “We are confident this is the technology to move us forward.”

Learn More

Learn more about [Amazon EMR](#).

[Sign In to the Console](#)

[Learn About AWS](#)

[Resources for AWS](#)

[Getting Started](#)

[Developers on AWS](#)

[Developer Center](#)

What Is AWS?	Training and Certification	SDKs & Tools
What Is Cloud Computing?	AWS Solutions Portfolio	.NET on AWS
AWS Inclusion, Diversity & Equity	Architecture Center	Python on AWS
What Is DevOps?	Product and Technical FAQs	Java on AWS
What Is a Container?	Analyst Reports	PHP on AWS
What Is a Data Lake?	AWS Partner Network	Javascript on AWS
AWS Cloud Security		
What's New		
Blogs		
Press Releases		

Help

[Contact Us](#)[AWS Careers](#)[File a Support Ticket](#)[Knowledge Center](#)[AWS Support Overview](#)[Legal](#)[Complete Sign Up](#)

Amazon is an Equal Opportunity Employer: *Minority / Women / Disability / Veteran / Gender Identity / Sexual Orientation / Age.*

Language

[عربي](#) |
[Bahasa Indonesia](#) |
[Deutsch](#) |
[English](#) |
[Español](#) |
[Français](#) |
[Italiano](#) |
[Português](#) |
[Tiếng Việt](#) |
[Türkçe](#) |
[Русский](#) |

ไทย |

日本語 |

한국어 |

中文 (简体) |

中文 (繁體)

Privacy

|

Site Terms

|

Cookie Preferences

|

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.