Amazon Web Services in Support of Big Data and Analytics

Peter Russell
Indiana University
petrusse@iu.edu

ABSTRACT

Executives are constantly looking for ways to find the pulse of their competitive landscape along with ways to gauge the sentiment among their customers. The emergence of the Big Data movement has given businesses the unique opportunity to gain perspective on these fronts, in addition to many others. Amazon Web Services has placed itself at the epicenter of this data movement and now offers tools that allows decision makers to quantify their businesses in ways that were previously computationally impossible or were prohibitively expensive. As a result, with Amazon Web Services, companies now have the ability to gain deep insights into customer activity, which can be used as real-time feedback or guidance to make future experiences more personalized.

KEYWORDS

i523,HID334, Cloud Computing, AWS, Big Data Analytics

1 INTRODUCTION

Amazon Web Services (AWS), the cloud service arm of Amazon, is currently the most dominant company in the cloud computing marketplace. With a market share of 31%, AWS holds a larger share than the next three closest competitors (Google, Microsoft and IBM) and contributes $10 billion a year to Amazon[16]. Aside from its financial importance to Amazon though, AWS has become critical for businesses that are looking to gain insights from the data they have at their disposal, especially as this data becomes more abundant [15].

With this business need in mind, AWS offers several products under their "Analytics" platform of services. This is just one of their 18 categories or platforms used to classify their 108 different products. This platform is a particularly interesting area because it is allowing companies to perceive their competitive landscape through an analytical lenses on a scale and frequency not previously seen. Namely, vast data sets in real-time if desired [24].

Our particular focus will be on a high-level description of the products offered in this "Analytics" category, their current utilization by businesses, recent developments in this platform and how it impacts Big Data.

2 ANALYTICAL PRODUCTS

To discuss the impact AWS is having on modern businesses, it's necessary to give a concise description of each analytical service offered. Subsequent sections will then be able to mention these services by name with a basic understanding of that service's function.

2.1 Amazon Athena

Amazon Athena that allows users to analyze data in Amazon Simple Storage Service (S3) as an SQL query. S3 is Amazon's web interfaced data storage and retrieval service, which can be accessed from anywhere, and can be more broadly be described as an Infrastructure as a Service (IaaS). This was designed for queries that may be unique and one-off. Athena remains one of the newest products introduced on the Analytics platform as it was released in late 2016 [1].

2.2 Amazon Elasticsearch Service

Amazon Elasticsearch Service (ES) is a managed service that implements Elasticsearch, which is an open source engine that allows for the indexing of large data sets. This indexing allows for analysis to better understand the events generating the data, such as with a user of an application [3].

2.3 Amazon Elastic MapReduce

Amazon Elastic MapReduce (EMR) is aimed at analysis of large data sets as it allows users to take advantage of a managed Hadoop framework without the traditional setup costs. Hadoop is advantageous over traditional database models because it parses the large data sets over several nodes, allowing for parallel computing and greatly increased efficiency. EMR allows for the iteration over a massive amount of text files while ES is concerned with indexing these files[4].

2.4 Amazon Quicksight

Amazon Quicksight is the data visualization tool that allows for seamless charting and integrating with AWS databases. It also recognizes data types and suggests the best type of visualization for a given analysis [6].

2.5 AWS CloudSearch

AWS CloudSearch is a managed search engine service that can be integrated into an application for a company's users. This allows an easier experience for the user without the company having to dedicate the resource costs that historically came with developing and maintaining the search feature [2]. In fact, AWS CloudSearch uses the same logic and intelligence for search queries that is used on Amazon.com. As one might suspect, AWS CloudSearch is similar to Amazon Elasticsearch Service. However, AWS CloudSearch is fully managed while Amazon Elasticsearch remains the more flexible and popular of the two.

2.6 AWS Data Pipeline

AWS Data Pipeline is designed to ease the maintenance of regular data sets by allowing users to schedule or automate changes to files along with the movement of that data set to other AWS services [8].

2.7 AWS Glue

Broadly speaking, AWS Glue is similar to AWS Data Pipeline in terms of automated transfer and modification of data. However, AWS Glue automates much of this data transformation whereas AWS Data Pipeline offers more flexibility for those who desire it [9].

2.8 AWS Kinesis

The work of AWS Kinesis is likely the most known product of AWS to the common consumer as it is responsible for the processing of real-time data for analysis or alert triggering. A dashboard that displays trending topics on social media or fraud detection at a bank is likely fed by an AWS Kinesis setup [5].

2.9 AWS Redshift

AWS Redshift was created to meet the database storage and maintenance needs of businesses. With Redshift, companies are able to reduce their capital expenditure and time to implementation, both of which could especially critical for nascent companies [7]. This line of business should prove to be increasingly important as data collection by businesses continues to grow. In 2012, it was already estimated that the cost of storage on AWS Redshift was just 10% the cost of traditional database costs [19].

3 RELEVANCE TO BIG DATA: USE CASES

Amazon has stated that they currently have one million active users, which is defined as using their services at least once a month [13]. In exploring current uses it becomes clear that the users are rarely consumers of just one product, opting instead to take advantage of the AWS ecosystem through multiple services. This section will touch upon the most popular AWS products and their interesting uses in the business environment.

3.1 Yelp

Yelp is a search based website that allows users to find different types of businesses while also showing user contributed reviews for these businesses. Started in 2004, Yelp's website now averages 28 million unique mobile users and 83 million unique desktop users per month. These users have contributed 135 million reviews in aggregate [26].

The impact of AWS on Yelp's business planning came when the company was trying to decide how to optimize its advertising revenue [11]. Specifically, Yelp stores log data daily on attributes, such as user location, user query, user clicks and displayed ads. This is all in an effort to better formulate search results given the available data and display ads that are most relevant to users [23].

Of the services discussed earlier, Yelp adopted AWS EMR and AWS Redshift to meet its analytical needs. EMR was implemented to allow multiple teams to analyze the data simultaneously and Redshift was used for easy retrieval. EMR is also used to enhance the user's search experience by returning useful results in the case of misspellings, auto-completion or features such as "People Who Viewed This Also Viewed."[21] As stated earlier, EMR allows this retrieval of information from the stored in nearly real-time. In all, the utilization of these services allows Yelp to be more dynamic as its data analysis time is dramatically reduced while also improving the customer experience and ultimately, retention [14].

3.2 Zillow

Zillow is an online real estate listing marketplace where users can find homes for sale, recently sold homes or foreclosures. One of the largest draws to the site though, is the modeling of a specific property value through a feature they refer to as a "Zestimate." Through the use of AWS Kinesis for data collection and AWS EMR for data processing, Zillow is able to generate home value estimations in virtual real-time for 100 million properties across the United States, which is said to be a function of over 100 input variables[12]. Some of these inputs need to be as real-time as possible, such as recent sales data, for the most accurate estimate, which made Kinesis so impactful [18]. This integration of technologies has dramatically improved their calculation time for these estimates from hours to seconds[12]. Once again, this enhanced user experience through the utilization of Big Data analytics keeps the website relevant and best suited to meet customer needs.

3.3 Netflix

Netflix is a worldwide media provider, offering on-demand movies and shows along with a DVD rental service. Currently, the company has nearly $9 billion in annual revenue with 104 million subscribers [20]. Incredibly, users in aggregate are watching one billion hours of content *a week* and during peak times, Netflix can be servicing over ten thousand streams a second [22]. Perhaps as impressive as the company's success with its user base is the foresight the company had in early as 2008 to begin moving operations to AWS as it began rolling out its internet streaming services. By 2016, they moved their entire infrastructure to the cloud and can have up to a hundred thousand AWS instances running during peak hours [17].

As likely one of the largest AWS users by market capitalization, Netflix casts a wide net across the use of AWS services. By their own admission, insights gleaned from the data they collect play a pivotal role on business and product decisions. Through the AWS Elasticsearch Service, Netflix is able to properly classify its 1.3 PB of data per day (24 GB per second) across different indices, such as viewing activities, error logs and diagnostics [25]. Similarly, Netflix uses AWS Kinesis as the pipeline used to stream this log data and the real-time functionality allows them to identify potential issues immediately [10]. Whether for business or troubleshooting purposes, this data on AWS can be easily visualized through AWS Quicksight for inferences.

Netflix is perhaps the best example of how a company can leverage AWS to outsource the burdens of data management as the volume of data grows. This allows them to focus on the core competencies and customer experience, which like the other examples, maintains or advances their position in the marketplace.

4 RECENT ADVANCEMENTS IN AWS

The most recent advancements in AWS as it relates to Analytics platform have come directly from the introduction of Athena in 2016 and Glue in 2017. Indirectly, AWS has been developing a new product line that is complementary to the Analytics category. In 2016, AWS launched its "Artificial Intelligence" platform, which

is now comprised of seven new services and is clearly an area of growth and focus for Amazon.

Of these new services, Amazon Machine Learning will likely be the most attractive new offering for businesses. This service will allow business users to discover underlying trends in their data and formulate more accurate forecasts.

5 CONCLUSIONS

In these use cases, we've seen that AWS has had a positive impact on Big Data for two reasons. First, businesses are better able to embrace the Big Data movement by making data collection and analysis a priority without the major cost that has historically been associated with such an initiative. Second, we would expect that the successful implementation of cloud analytics will help businesses be more successful, in turn incentivizing them to collect more data and therefore, further expanding the Big Data universe.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the Assistant Instructors for their feedback and help.

REFERENCES

[1] Amazon. 2017. Amazon Athena. Website. (2017). https://aws.amazon.com/athena/
[2] Amazon. 2017. Amazon CloudSearch. (2017). https://aws.amazon.com/cloudsearch/
[3] Amazon. 2017. Amazon Elasticsearch Service. Website. (2017). https://aws.amazon.com/elasticsearch-service/
[4] Amazon. 2017. Amazon EMR. Website. (2017). https://aws.amazon.com/emr/
[5] Amazon. 2017. Amazon Kinesis. Website. (2017). https://aws.amazon.com/kinesis/
[6] Amazon. 2017. Amazon QuickSight. Website. (2017). https://quicksight.aws/
[7] Amazon. 2017. Amazon Redshift. Website. (2017). https://aws.amazon.com/redshift/
[8] Amazon. 2017. AWS Data Pipeline. (2017). https://aws.amazon.com/datapipeline/
[9] Amazon. 2017. AWS Glue. Website. (2017). https://aws.amazon.com/glue/
[10] Amazon. 2017. Netflix and Amazon Kinesis Streams Case Study. Website. (2017). https://aws.amazon.com/solutions/case-studies/netflix-kinesis-streams/
[11] Amazon. 2017. Yelp Data Analytics Case Study. Website. (2017). https://aws.amazon.com/solutions/case-studies/yelp-data-analytics/
[12] Amazon. 2017. Zillow Provides Near-Real-Time Home-Value Estimates Using Amazon Kinesis. Website. (2017). https://aws.amazon.com/solutions/case-studies/zillow-zestimate/
[13] Jeffrey P. Bezos. 2015. Annual Letter to Shareholders. Press Release. (April 2015).
[14] Niraj Dawar. 2016. Use Big Data to Create Value for Customers, Not Just Target Them. Website. (Aug. 2016). https://hbr.org/2016/08/use-big-data-to-create-value-for-customers-not-just-target-them
[15] The Economist. 2017. Data is giving rise to a new economy. Website. (May 2017). https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy
[16] Synergy Research Group. 2016. AWS Remains Dominant Despite Microsoft and Google Growth Surges. Website. (Feb. 2016).
[17] Neil Hunt. 2016. Website. (2016). https://aws.amazon.com/solutions/case-studies/netflix/ Conference Presentation at AWS re:Invent 2016.
[18] Eric Knorr. 2016. Hot property: How Zillow became the real estate data hub. Website. (April 2016). https://www.infoworld.com/article/3060773/big-data/hot-property-how-zillow-became-the-real-estate-data-hub.html
[19] Ingrid Lunden. 2013. Amazon Takes Redshift, Its Cloud-Based Data Warehouse Killer, Global. Website. (Feb. 2013). https://techcrunch.com/2013/02/15/amazon-takes-redshift-its-cloud-based-data-warehouse-killer-global/
[20] Netflix. 2017. Investor Relations - Financial Statements. Website. (Sept. 2017). https://ir.netflix.com/
[21] David M. Search. 2010. mrjob: Distributed Computing for Everybody. Website. (Oct. 2010). https://engineeringblog.yelp.com/2010/10/mrjob-distributed-computing-for-everybody.html
[22] Softpedia. 2017. Netflix Users Spend 1 Billion Hours per Week Watching Movies. Website. (April 2017). http://news.softpedia.com/news/netflix-users-spend-1-billion-hours-per-week-watching-movies-514989.shtml
[23] Jeremy Stoppelman. 2013. Fast Company Innovation Uncensored. Panel Discussion. (Nov. 2013). http://blog.fastcompany.com/post/66283564254/yelp-ceo-jeremy-stoppelman-talks-big-data-in-this
[24] Laura Winig. 2016. GE's Big Bet on Data and Analytics. Website. (Feb. 2016). https://sloanreview.mit.edu/case-study/ge-big-bet-on-data-and-analytics/ Case Study.
[25] Steven Wu, Allen Wang, Monal Daxini, Manas Alekar, Zhenzhong Xu, Jigish Patel, Nagarjun Guraja, Jonathan Bond, Matt Zimmer, and Peter Bakas. 2016. Evolution of the Netflix Data Pipeline. Website. (Feb. 2016). https://medium.com/netflix-techblog/evolution-of-the-netflix-data-pipeline-da246ca36905
[26] Yelp. 2017. Fact Sheet. Website. (June 2017). https://www.yelp.com/factsheet

6 BIBTEX ISSUES

7 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

7.1 Uncaught Bibliography Errors

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

7.2 Formatting

DONE:

Incorrect number of keywords or HID and i523 not included in the keywords

DONE:

Other formatting issues - missing references section, probably due to bibtex issues

7.3 Writing Errors

7.4 Character Errors

DONE:

Erroneous use of quotation marks, i.e. use “quotes” , instead of ” ”

7.5 Structural Issues

DONE:

Acknowledgement section missing