BIG DATA PROCESSING FOR ANALYTICS USING ADVANCED AWS CLOUD PLATFORM TOOLS

Kaviyarasi.,
Data Engineer,
India.

Meenakshi.,
Product Engineer-Team Leader,
India.

ABSTRACT

This paper details the challenges involved Big Data for analytics purposes. How constantly evolving AWS cloud platform tools and landscape makes it easy for business to jump in Big Data analytics space.

Keywords— Big Data on Cloud, Big Data Analytics, Big Data Analytics on AWS Cloud, AWS Big Data tools, AWS.

1. INTRODUCTION

As we become a more digital society, the amount of data being created and collected is growing and accelerating significantly. Analysis of this ever-growing data becomes a challenge with traditional analytical tools. We require innovation to bridge the gap between data being generated and data that can be analyzed effectively. Big data tools and technologies offer opportunities and challenges in being able to analyze data efficiently to better understand customer preferences, gain a competitive advantage in the marketplace, and grow your business. Data management architectures have evolved from the traditional data warehousing model to more complex architectures that address more requirements, such as real-time and batch processing; structured and unstructured data; high-velocity transactions; and so on. Amazon Web Services (AWS) provides a broad platform of managed services to help you build, secure, and seamlessly scale end-to-end big data applications quickly and with ease. Whether your applications require real-time streaming or batch data processing, AWS provides the infrastructure and tools to tackle your next big data project. No hardware to procure, no infrastructure to maintain and scale—only what you need to collect, store, process, and analyze big data. AWS has an ecosystem of analytical solutions specifically designed to handle this growing amount of data and provide insight into your business.[1][2][3]

2. BIG DATA DIMENSION

Big data has four dimensions as described below [19]:

• Volume – Current data existing is in petabytes, which is already problematic; it is predicted that in the next few years it is to increase to zettabytes (ZB) . This explosion of data is mainly due to social media and mobile devices.

• Velocity – Refers to both the rate at which data is captured and the rate of data flow. As Live data is too large and continuously in motion, it causes challenges for traditional analytics.

• Variety – As data collected is not of a specific category or from a single source, Data exists in numerous raw data formats obtained from the web, texts, sensors, e-mails, etc. which are structured or unstructured. It is not from a specified source or from a single category. Traditional analytical methods cannot manage this kind of data known as big data.

• Veracity – Ambiguity within data typically from noise and abnormalities within the data is the primary focus in all four V's. Big Data helps enterprise to develop big data driven e-commerce architecture which aids in gaining extensive "insight into customer behavior, industry trends, more accurate decisions to improve just about every aspect of the business, from marketing and advertising, to merchandising, operations, and even customer retention [4][9][19].

3. ANALYTICS

Computational analysis of data which is done in systematic way is known as Analytics.[1] Analytics helps to discover, interpret, and communicate meaningful patterns in data. These data patterns help towards effective decision making. Analytics is valuable when there is abundant recorded information. Computer programming, operation research and simultaneous application of statistics is required quantify performance of Analytics. To improve business performance, predict and describe; Organizations applies analytics to business data. Specifically, areas within analytics include Big Data Analytics, retail analytics, supply chain analytics, predictive analytics, web analytics, call analytics, speech analytics, prescriptive analytics, enterprise decision management, descriptive analytics, cognitive analytics, predictive science, graph analytics, credit risk analysis, and fraud analytics, store assortment and stock-keeping unit optimization, marketing optimization and marketing mix modelling, sales force sizing and optimization, price and promotion modelling. Analytics requires extensive computation. All latest technologies in computer science, mathematics and statistics are used in algorithms and software for analytics.[5][6][7]

4. CLOUD COMPUTING

Cloud computing basically provides databases, storage, servers, networking, software, intelligence, and analytics through the Internet known as cloud. Cloud computing gives the power to have faster innovation, scale economically and use resources as per your need [19]. Cloud computing helps you to lower your operating costs by giving flexibility to pay only for cloud services you use. Thus, running your infrastructure more efficiently and scale as your business according to the need.[8]

5. ADVANTAGES IN BIG DATA ANALYTICS USING AWS

Analyzing large data sets requires significant compute capacity that can vary in size based on the amount of input data and the type of analysis.[17] This characteristic of big data workloads is ideally suited to the pay-as-you-go cloud computing model, where applications can easily scale up and down based on demand. As requirements change, you can easily resize your environment (horizontally or vertically) on AWS to meet your needs, without having to wait for additional hardware or being required to over invest to provision enough capacity. For mission-critical applications on a more traditional infrastructure, system designers have no choice but to over-provision, because a surge in additional data due to an increase in business need must be something the system can handle.[10] By contrast, on AWS you can provision more capacity and compute in a matter of minutes, meaning that your big data applications grow and shrink as demand dictates, and your system runs as close to optimal efficiency as possible. In addition, you get flexible computing on a global infrastructure with access to the many different geographic regions that AWS offers, along with the ability to use other scalable services that augment to build sophisticated big data applications.[18] These other services include Amazon Simple Storage Service (Amazon S3) to store data and AWS Glue to orchestrate jobs to move and transform that data easily.[11] AWS IoT, which lets connected devices interact with cloud applications and other connected devices. As the amount of data being generated continues to grow, AWS has many options to get that data to the cloud, including secure devices like AWS Snowball to accelerate petabyte-scale data transfers, delivery streams with Amazon Kinesis Data Firehose to load streaming data continuously, migrating databases using AWS Database Migration Service, and scalable private connections through AWS Direct Connect. AWS recently added AWS Snowball Edge, which is a 100 TB data transfer device with on-board storage and compute capabilities. You can use Snowball Edge to move large amounts of data into and out of AWS, as a temporary storage tier for large local datasets, or to support local workloads in remote or offline locations. Additionally, you can deploy AWS Lambda code on Snowball Edge to perform tasks such as analyzing data streams or processing data locally. As mobile continues to rapidly grow in usage you can use the suite of services within the AWS Mobile Hub to collect and measure app usage and data or export that data to another service for further custom analysis. These capabilities of the AWS platform make it an ideal fit for solving big data problems, and many customers have implemented successful big data analytics workloads on AWS. For more information about case studies, see Big Data Customer Success Stories.[12][13]

The following services for collecting, processing, storing, and analyzing big data are described in order [20]:

- Amazon Kinesis
- AWS Lambda
- Amazon Elastic MapReduce
- Amazon Glue
- Amazon Machine Learning
- Amazon DynamoDB
- Amazon Redshift
- Amazon Athena
- Amazon Elasticsearch Service
- Amazon QuickSight

In addition to these services, Amazon EC2 instances are available for self-managed big data applications.[14][20]

6. VARIETY OF REAL-TIME DATA PROCESSING SYESTEM IN CLOUD

- Real-time File Processing – You can trigger Lambda to invoke a process where a file has been uploaded to Amazon S3 or modified. For example, to change an image from color to gray scale after it has been uploaded to Amazon S3.

• Real-time Stream Processing – You can use Kinesis Data Streams and Lambda to process streaming data for click stream analysis, log filtering, and social media analysis. • Extract, Transform, Load – You can use Lambda to run code that transforms data and loads that data into one data repository to another.

• Replace Cron – Use schedule expressions to run a Lambda function at regular intervals as a cheaper and more available solution than running cron on an EC2 instance.

• Process AWS Events – Many other services, such as AWS CloudTrail, can act as event sources simply by logging to Amazon S3 and using S3 bucket notifications to trigger lambda functions.[15][16]

7. CONCLUSION

Big Data is not a new term but has gained its spotlight due to the huge amounts of data that are produced daily from different sources. From our analysis we saw that big data is increasing in a fast pace, leading to benefits but also challenges. Cloud Computing is the best solution for storing, processing, and analyzing Big Data. Companies like Amazon, Google and Microsoft offer their public services to facilitate the process of dealing with Big Data. From the analysis we saw that there are multiple benefits that Big Data analytics provides for many different fields and sectors such as healthcare, education, and business. We also saw that because of the interaction of Big Data with Cloud Computing there is a shift in the way data is processed and analyzed. Data analytics requires flexible, scalable, and highperformance tools so that it can provide insights more quickly. As more and more data are generated and collected so new tools emerge every now and then, but it is difficult to choose the right tool and to keep pace as most of them "die" very soon. AWS platform makes it easier to scale, deploy and build big data applications. It provides various managed services to collect, process, and analyze big data. AWS provides various solutions to help in your big data analytic requirements so that you can focus on business problems instead of updating and managing these tools. To achieve a flexible and scalable big data architecture most business use Multiple AWS tools to build a complete solution. This approach helps meet stringent business requirements in the most cost-optimized, performance, and resilient way possible. In future scope helps data analysts and data scientists of all technical levels understand, combine, clean, and transform data, it can automate filtering anomalies, correcting invalid values, converting data to standard formats and other tasks.

REFERENCES

[1] M. Hillbert and P. Lopez, ”The World's Technological Capacity to Store, Communicate and Compute Information,” Compute Information.Science, vol. III, pp. 62-65, 2011.

[2] J. Hellerstein, ”Gigaom Blog,” 8 November 2019. [Online]. Available: https://gigaom.com/2008/11/09/mapreduce-leads-the-way-forparallelprogramming/. [Accessed 20 January 2021].

[3] Statista, ”Statista,” 2020. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/. [Accessed 21 January 2021].

[4] D. Reinsel, J. Gantz and J. Rydning, ”Data Age 2025: The Evolution of Data To-Life Critical,” International Data Corporation, Framingham, 2017.

[5] S. Kaisler, F. Armour and J. Espinosa, ”Big Data: Issues and Challenges Moving Forward,” Wailea, Maui, HI, s.n, pp. 995 - 1004., 2013.

[6] Wikipedia, ”Wikipedia,” 2018. [Online]. Available: https://www.en.wikipedia.org/wiki/Big data/. [Accessed 4 January 2021].

[7] J. Weathington, ”Big Data Defined.,” Tech Republic, 2012.

[8] PCMagazine, ”PC Magazine,” 2018. [Online]. Available: http://www.pcmag.com/encyclopedia/term/62849/big-data.. [Accessed 9 January 2021]

[9] D. Gewirtz, ”ZDNet,” 2018. [Online]. Available: https://www.zdnet.com/article/volume-velocity-andvarietyunderstanding-the-three-vs-of-big-data/. [Accessed 1 January 2021].

[10] S. M. F. Akhtar, Big Data Architect’s Handbook, Packt, 2018.

[11] WhishWorks, ”WhishWorks, ”, 2019. [Online]. Available: https://www.whishworks.com/blog/data-analytics/understanding-the3- vs-of-big-data-volume-velocity-and-variety/. [Accessed 23 January 2021].

[12] S. Yadav and A. Sohal, ”Review Paper on Big Data Analytics in Cloud Computing,” International Journal of Computer Trends and Technology (IJCTT), vol. IX, 2017.

[13] R. Kimball and M. Ross, The data warehouse toolkit: The definitive guide to dimensional modeling, 3rd ed. John Wiley & Sons, 2013.

[14] LaprinthX, ”LaprinthX, ”, 2018. [Online]. Available: https://laptrinhx.com/better-faster-smarter-elt-vs-etl-2084402419/. [Accessed 22 January 2021].

[15] Xplenty, ”XPlenty, ”, 2019. [Online]. Available: https://www.xplenty.com/blog/etl-vs-elt/#. [Accessed 20 January 2021]

[16] Forbes, ”Forbes, ”, 2018. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2019/11/06/fivebenefit s-of-big-data-analytics-and-how-companies-cangetstarted/?sh=7e1b901417e4. [Accessed 13 January 2021]

[17] EDHEC, ”EDHEC, ”, 2019. [Online]. Available: https://master.edhec.edu/news/three-ways-educators-are-using-bigdataanalytics-improve-learning-process#. [Accessed 6 January 2021]

[18] Google Cloud, ” BigQuery, ”, 2020. [Online]. Available: https://cloud.google.com/bigquery. [Accessed 5 January 2021] [19] Forbes, ”Forbes ”, 2020. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-muchdatado-we-create-every-day-the-mind-blowing-stats-everyone-shouldread/?sh=5936b00460ba

[19] Building Big Data and Analytics Solutions in Cloud (ibm.com) By Wei-Dong Zhu, Manav Gupta, Ven Kumar, Sujatha Perepa, Arvind Sathi and Craig Statchuk Available: http://www.redbooks.ibm.com/redpapers/pdfs/redp5085.pdf [Online]

[20] Big Data Analytics Options on AWS (awsstatic.com) [Online] Available: https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on _AWS.pdf.