International Journal for Applied Information Management
Vol. 1, No. 1, April 2021, pp. 01-05

ISSN -
1

Data Analytics Architectures for E-Commerce Platforms in Cloud

John Yeung [1,*], Simon Wong [b,2], Alvin Tam [b,3]

[1] Data Science Academy, Hong Kong
[2] Hong Kong Community College, The Hong Kong Polytechnic University, Hong Kong
[3] Data Science Academy, Hong Kong
[1] john.yeung@ymail.com*; [2] ccswong@hkcc-polyu.edu.hk; [3] alvin_tyl@hotmail.com
* corresponding author

(Received April 24, 2021, Accepted April 24, 2021, Available online April 25, 2021)

Abstract

Today, organizations not only need to manage larger volumes of data, but also generate insights from existing data. These insights help them understand better about their customers and predict market trends. With this initiative, they can take advantage of the cloud platform to achieve this goal because it manages higher data volume, speed and variation. This cloud platform enables them to provide elasticity and efficient computing and storage resources. They also provide many ready-to-use tools for building data analytics in various stages. Additionally, an on-demand pricing model allows organizations to pay for what they consume. It changes the organizational consumption model from capital expenditure to operational expenditure. It greatly minimizes initial capital investment to build data analytics solutions and implement other innovative ideas. This paper highlights the main reasons for encouraging organizations to build data analytics in the cloud. It also shows how to articulate data analytics frameworks for ecommerce platforms in the cloud and how to integrate machine learning models into data analytics processes, to create more sophisticated analyzes. AWS Amazon Web Services' premier public cloud platform is adopted to demonstrate these concepts and practices with real-life business cases.

Keywords: Cloud Computing; Data Analytics; Machine Learning; E-Commerce;

1. Introduction

Over the past two decades, consumer shopping habits have shifted from in-store to online. This creates massive business changes or opportunities for retailers to be present online and businesses to drive sales and growth, especially as they expand their business globally [1]. However, as their online businesses start to grow with more concurrent transactions, they will face some distinctive technical challenges in terms of enabling platform scalability, changing cost models, and addressing data analytics challenges [2-3].

In this paper, we would firstly analyze the common challenges of E-Commerce platforms from the technical perspective. Secondly, we explain how cloud computing technologies can be leveraged to address some of these challenges and come up with new business values [4]. Furthermore, to understand customer behaviors by analyzing data in the E-Commerce platform is one of practical approaches to forecast the market trend [5]. It is considered as a key success factor for market players to have competitive advantages against other competitors. Some cloud platforms provide advanced technologies like Artificial Intelligence AI and Machine Learning ML for customers to integrate with their existing business systems or processes We would cover the methodology of how to leverage these AI and ML services to integrate with the data analytics process to formulate a more comprehensive analytics platform in the cloud [6-8]. A practical real-life case would be explained to strengthen this model.

2. Common Challenges of E-Commerce Platform

As online businesses start to grow with more simultaneous transactions, these e-commerce players will face the following typical technical challenges:

Challenge 1: Scalability limit due to on-site infrastructure

Challenge 2: Large upfront investment in on-site infrastructure

Challenge 3: Dealing with the rapidly growing volume and variety of data for data analytics

Challenge 4: Make use of more complex analytical models to get accurate results

3. E-Commerce Platform on Cloud

3.1. Platform Scalability and Agility

Cloud computing enables E-Commerce platforms to handle the dynamic demands and scenarios of the market. It enables these platforms having the elasticity to upscale or downscale the services, i.e. compute and storage, in order to meet the actual demands, and seasonal spikes [9-11]. In the traditional environment, i.e. on-premise data centers, people always make oversize provisions on hardware and software capacities in order to prevent from un-sufficient resources to meet these seasonal spikes. This results in wasting a lot of resources in the normal operation most of the time.

3.2. Changing the Cost Model

E-Commerce platform owners need to make an upfront investment in hardware and software infrastructure before they can earn positive income in such a business. With cloud computing, this changes the cost model from Capital Expenditure (CAPEX) to Operating Expenditure (OPEX) [12,16]. Thus, the cost of ownership [1] can be very flexible according to actual business needs. Even though the business grows, operating costs will not change dramatically. This protects business discovery on the platform infrastructure, and shifts business focus to more valuable areas such as understanding customer behavior and market trends [13].

3.3. Challenges of Handling Big Data

Ecommerce players need better, faster, and more relevant insights from their data to stay ahead of the competition. As data volume, variety, and volume increase, they need to have more sophisticated tools to collect, categorize, and turn data into valuable insights [14]. Advanced analytics related products have been the key to generating value. The data lake concept was introduced to the market. It serves as a solid basis for storing large amounts of data in a storage center within an organization [15]. Data lakes need to be very cost-effective, scalable and secure. Done right, organizations can open the door to creating advanced analytics, facilitating data science and machine learning.

3.4. Demanding for more Complicated Analysis Models

The adoption of data science is creating significant business value to E-Commerce players [16]. Business decisions can be more data-driven, and new product lines can be created using data insights unveiled by Business Intelligence BIand Machine LearningML [17].

4. Data Analytics Architecture on Cloud

As shown in figure 1, the entire data analysis architecture basically consists of four stages: collecting, storing, processing and consuming. Each stage will have a different cloud service specifically handling specific tasks. These services can work together to streamline the entire process, namely orchestration.

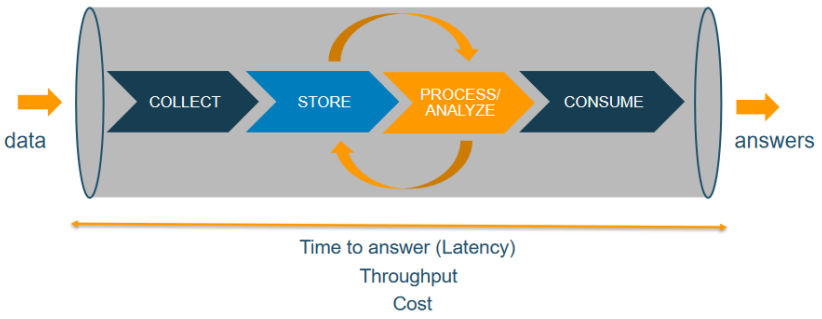

Fig. 1. Data Processing Pipeline

4.1. Data Collecting Stage

We can classify data as structured, semi-structured and unstructured types. Structured data is highly normalized by general schema and stored in relational databases, which support transactional lines of business applications [17]. This data is easily accessible via SQL or a data extraction tool. Semi-Structured Data contains identifiers without following a predefined schema, often stored in NoSQL databases such as JSON and XML. This data is easily accessible but requires some preparation to be ready for data analysis. Unstructured data does not fit into the data model and is usually stored as individual files. Some examples are text, image, audio and video documents.

After identifying the nature of the data, we then decide on the data collection method as Batch Load or Streaming. Batch Load periodically extracts data from various data sources and moves it to the Data Lake. This process usually involves querying the database and includes several transformation processes including Extracting, Transforming, and LoadingETL.

4.2. Data Storing Stage

Data Lake is a centralized response for storing all data types. In contrast to the Data Warehouse which is a database that is optimized for analyzing relational data sourced from transactional applications. The data structure and schema in the Data Warehouse are well defined in advance to optimize SQL query performance. The results are typically used for operational reporting and analysis via several business intelligence tools. Cleaned, modified and stored database table data.

On the other hand, aData Lake stores relational data from transactional applications and non-relational data from mobile applications, Internet of Things IoT devices, or social media platforms. The data structure or schema is not determined when the data is collected. Different types of analysis such as SQL queries, big data analytics, full-text search, and machine learning can be used to find insights from data lakes.

4.3. Data Processing Stage

Data processing includes cleansing, transforming, sorting and aggregating data. A typical example is the ETL process. Sometimes, the whole data processing may have several iterations between the storing data stage and the data processing stage.

In addition, more enterprises are exploring the approaches of deploying machine learning algorithms to spot patterns and catch more insights based on perceived data. One of common applications in the E-Commerce industry is recommendation engines. Customer behaviors like purchasing history and browsing preferences are recorded in the platforms. These data would then be handled by some analysis processes with ML models deployed. This approach makes the whole mechanism more interactive and sophisticated enough to extract more in-depth analysis.

4.4. Data Consuming Stage

This is the final stage of the whole data pipeline. It is about how to have insights from data with visualization tools like Business Intelligence software or provide data to other entities or applications by API calls.

5. Integrating ML to Data Analytics on Cloud

Machine learning, as a branch of Artificial Intelligence, was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks [2]. The iteractive aspect of machine learning is important because models are exposed and adapt to new training data . It is supposed that models would be more “mature” to produce more reliable and predictive result when keep learnings from previous computations.

In our research, we aim to integrate ML technologies to the data analytics processes and recommend a more comprehensive pattern for the E-Commerce industry. In our example, we are using Amazon SageMaker as the platform to build, train and deploy ML models on Cloud. The whole ML process typically involves the steps in Figure 2 and 3:

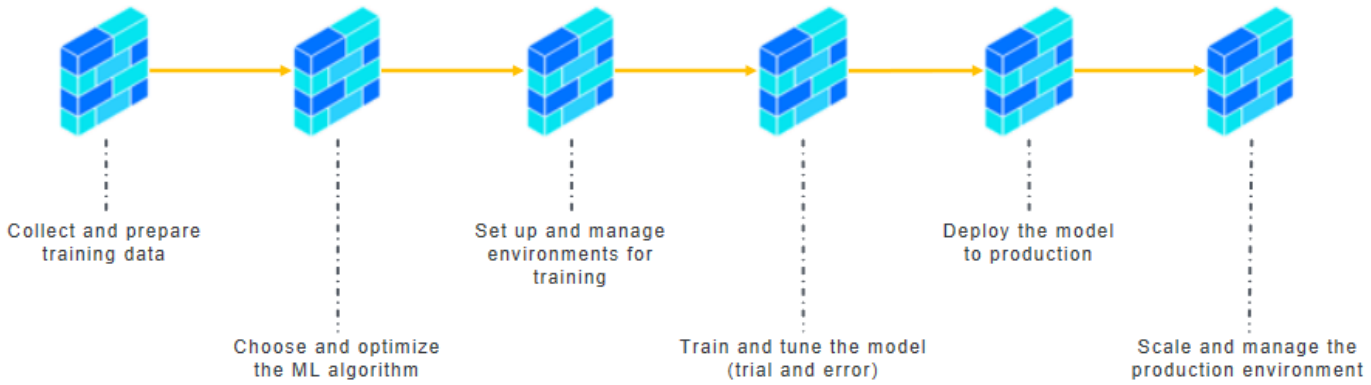

Fig. 2. Machine Learning Processes

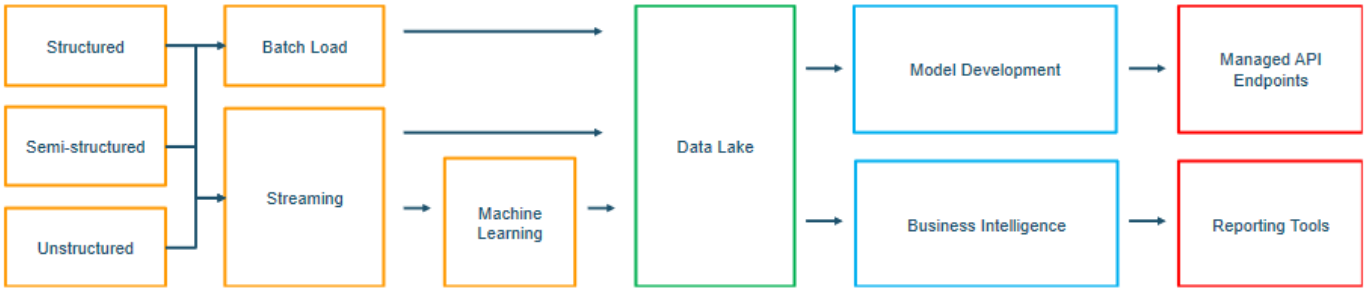

Fig. 3. Integrating ML and Data Analytics

6. Conclusion Remarks and Future Works

In this study, the researchers highlighted the common challenges of E-Commerce platforms and reasons why Cloud addressed them and brought additional values for future developments in areas of data analytics and machine learning. With respect to future work of this study, the researchers propose:

- gathering reference cases to prove the models mentioned in this paper
- setting up some demonstrations to illustrate the whole architecture.

References

[1] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi, "Cloud computing - The business perspective," *Decis. Support Syst.*, vol. 51, no. 1, pp. 176–189, 2011, doi: 10.1016/j.dss.2010.12.006.

[2] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics," *J. Parallel Distrib. Comput.*, vol. 74, no. 7, pp. 2561–2573, 2014, doi: 10.1016/j.jpdc.2014.01.003.

[3] A. A. Cárdenas *et al.*, "SYSTEMS SECURITY Big Data Analytics for Security," no. December, pp. 74–76, 2013.

[4] M. I. T. Sloan, M. Review, M. I. T. Sloan, and M. Review, "Big Data, Analytics and the Path from Insights to Value," *MIT Sloan Manag. Rev.*, no. 52205, pp. 1–18, 2010, [Online]. Available: https://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/.

[5] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *J. Big Data*, vol. 2, no. 1, pp. 1–32, 2015, doi: 10.1186/s40537-015-0030-3.

[6] K. Eric, "What cloud computing really means | InfoWorld," *Infor World*, no. April, 2008, [Online]. Available: http://www.infoworld.com/article/2683784/cloud-computing/what-cloud-computing-really-means.html.

[7] T. Dillon, C. Wu, and E. Chang, "Cloud computing: Issues and challenges," *Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA*, pp. 27–33, 2010, doi: 10.1109/AINA.2010.187.

[8] R. L. Grossman, "The case for cloud computing," *IT Prof.*, vol. 11, no. 2, pp. 23–27, 2009, doi: 10.1109/MITP.2009.40.

[9] L. Wang *et al.*, "Cloud computing: A perspective study," *New Gener. Comput.*, vol. 28, no. 2, pp. 137–146, 2010, doi: 10.1007/s00354-008-0081-5.

[10] Y. Ding, M. Korotkiy, and B. Omelayenko, "Golden Bullet: Automated Classification of Product Data in E-Commerce," *Bus. Inf. Syst. Proc. BIS 2002*, pp. 1–9, 2002, [Online]. Available: http://www.cs.jyu.fi/ai/vagan/course_papers/Paper_30_SW.pdf.

[11] Q. Guo *et al.*, "Securing the Deep Fraud Detector in Large-Scale E-Commerce Platform via Adversarial Machine Learning Approach. BT - The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019," pp. 616–626, 2019, [Online]. Available: https://doi.org/10.1145/3308558.3313533.

[12] H. K. Rao, Z. Zeng, and A. P. Liu, "Research on personalized referral service and big data mining for e-commerce with machine learning," *2018 4th Int. Conf. Comput. Technol. Appl. ICCTA 2018*, pp. 35–38, 2018, doi: 10.1109/CATA.2018.8398652.

[13] D. Fensel *et al.*, "Integration in B2B," 2001.

[14] N. H. Trang, "Limitations of Big Data Partitions Technology," *J. Appl. Data Sci.*, vol. 1, no. 1, pp. 11–19, 2020.

[15] T. Wahyuningsih, "Problems , Challenges and Opportunities Visualization on Big Data," *J. Appl. Data Sci.*, vol. 1, no. 1, pp. 20–28, 2020.

[16] A. W. Services, "Cost Management in the AWS Cloud," *Amaz. Web Serv.*, no. March, 2018.

[17] A. S. M. Al-rawahnaa, A. Yahya, and B. Al, "Data mining for Education Sector , a proposed concept," *J. Appl. Data Sci.*, vol. 1, no. 1, pp. 1–10, 2020.