Alperen CAN 21601740
Dr. Haluk ALTUNEL
CS411 – Software Architecture Design
25 October 2022

Paper Review

As a topic, I chose serverless architecture. The paper I made a review of it is *Serverless Architecture for Big Data Analitics*, written by Mijanur Rahman and Hasibul Hasan at South East University in Bangladesh, published in 2019, and has 9 citations at Google Scholar [1]. The reason why I chose this topic is because serverless architecture model has been becoming more and more widespread and important due to the fact that technological developments have also changed the needs with it. Increased demand, burden and cost accelerated the development of new methods in the field of software. Eventually, with the developing cloud technology and innovations, different important services have entered our lives, such as serverless architecture. Therefore, it seemed more interesting to me, and I wanted to learn about it by starting to analyze a paper.

The paper discusses the importance and the ways of implementing serverless architecture for big data, which refers to the data growing fast and becoming too complex to be dealt with. Therefore, traditional data processing applications is becoming less and less effective by time. In this case, paper emphasizes the importance of having a serverless architecture. However, it also argues that serverless does not come with all capabilities that are feasible to design and implement a data lake. Having explaining the *data lake* concept, which is a vital repository to bring the whole data to a central place, authors talk over the proper way of implementing a data lake by using some serverless architecture services such as *Amazon S3*, *AWS Glue* and more. Ultimately, by comparing the traditional and serverless approaches, paper states that big data analytics will become easier with minimum cost thanks to serverless implementation model.

The paper has valuable points especially for the ones who do not know much about the serverless architecture and big data since it explains the concepts from scratch, so to speak. Having read the paper, it is easily understandable that the idea behind serverless architecture is to provide a structure where applications are hosted by Cloud providers, thus eliminating the need for developers to manage servers, software and hardware. Therefore, it is unambiguous that since managing the servers is under the responsibility of the cloud provider rather than the developer, it allows developers to focus more on the business side. Moreover, the paper does not just explain serverless and data lakes in general, but also clarifying how to implement them by giving the names of specific providers, such as AWS Athena, AWS Glue, Amazon Kinesis and more. In addition, the reasons to use them and where to use them is explained in separate paragraphs.

Subsequently, the relation between these services and tools are ideally visualized thanks to the comprehensible figure representing the layers to implement a proper data lake. Ultimately, comparing the traditional and serverless approaches item by item in a table increased the clarity.

Despite all these positive points and valuable contribution of the paper, some contradictions have appeared when big data was mentioned. Initially, I agree with the paper on the fact that managing the servers in the traditional way results in an additional difficulty for the programmers as the data is becoming more complex. Therefore, from my perspective, authors' following statement has a point: "*By relying on serverless architecture, big data analytics will become easier…"*. However, this thesis statement solely does not sufficient, because serverless also tends to have drawbacks while working on complex projects with big data. For example, if a complex application will be developed in serverless architecture, it is highly likely that there will be a need to design the architecture in accordance with serverless. Therefore, it may also bring the faulty designs and high costs with it. For that reason, the paper should have included that cases in which serverless may not be the correct decision, or instead it should have included that what should be the milestones to move the application to serverless, in order to adapt the project in a more effective manner, i.e.by moving the parts of the application step by step.

Moreover, the paper has some unanswered/left questions. For instance, although it tries to describe the steps to design a data lake in 4 layers, authors do not give details about the ways to do this, thus remains these questions mostly unanswered. To exemplify, while describing the 1st layer, paper describes the main ideas as follows:

> *"How we can bring and save data to our structure or system from various external sources? How can we bring relational or social media data? How can we bring our client's data? Our first step is to bring and store data from all external sources."*

Then paper proceeds to other steps, leaving them unanswered. Having reading this paper and also other sources to have a core knowledge, I can assert that relying and using Amazon services such as AWS Glue, Firehose and Kinesis may be helpful to bring stream data. In addition, relying on DMS may also be helpful to bring relational database data.

References:

[1] Rahman, Mijanur, and Hasibul Hasan. "Serverless Architecture for Big Data Analytics." *IEEE Xplore*, https://ieeexplore.ieee.org/abstract/document/8978443.