

Ways Data Collection and Analysis Pipelines Can Be Built Through Cloud Services

(Technical Paper)

Problems of Relying on Data as Basis for Important Decision Making

(STS Paper)

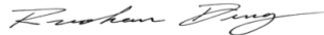
A Thesis Prospectus Submitted to the
Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Ruohan Ding

Fall, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments



Signature _____ Date 10/31/2022
Ruohan Ding

Approved _____ Date _____

Briana Morrison, PhD, UVA Computer Science Department in School of Engineering

Approved _____ Date _____

STS Advisor: Richard D. Jacques, Ph.D., Department of Engineering & Society

Introduction

Data is the defining feature of the 21st century. There is an incredible amount of data in the world but most of it is unprocessed and unused. A common problem many companies face is how best to conduct useful data collection and analysis while also investing a minimal amount of capital and time. Companies like Palantir provide data analysis services using Machine Learning and Artificial Intelligence. Customers then purchase these services and give up their data in return for valuable insight generated from the models.

However relying on third-party companies does present potential problems. For one there are security concerns with and possibly regulations against sharing sensitive data with an outside company. Some companies prefer to develop their own data-processing pipelines, as this allows for complete control over the entire process. Unfortunately, the services and tools necessary to successfully develop these pipelines require significant time and capital investments that most companies cannot afford. Therefore, using existing cloud services designed specifically to build data-processing pipelines is a good midpoint between relying on a third-party company and creating proprietary software.

Any company that can successfully leverage the wealth of data available to them will be able to generate greater revenue, make better informed decisions, and improve their overall efficiency (Gavin et al., 2019). There are many existing cloud architectures for the design of data processing pipelines. One popular technique is to combine together many different cloud

services. Even some existing companies that specialize in data analysis find it easier to pay for existing cloud services instead of creating their own.

Technical Discussion

Data processing is a huge problem that almost every industry experiences. Small startups and big corporations alike need a system to properly collect, analyze, and visualize their data to provide them with valuable insight. Therefore many companies use existing cloud services like Amazon Web Services (AWS) that are specifically designed to resolve this problem (Mesbahi, Rahmani, Hosseinzadeh, et al., 2018). There are many existing cloud architectures for the design of data processing pipelines. One popular technique is to combine together many different AWS services. Even some existing companies that specialize in data analysis find it easier to use AWS services instead of creating their own.

Palantir, a leader in machine learning data analysis, chooses to develop their own data processing systems in conjunction with existing AWS Cloud services. Their complex Artificial-Intelligence powered analysis is hosted on AWS servers and much of their data deployment and service hosting is also handled by AWS. This shows that these cloud services can be highly flexible. Companies can use as little or as much of these services as fits their needs.

AWS also provides many official guides on how best to leverage their services for a variety of projects. These projects range from large scale applications deployed by big corporations to small services developed by university students. The low-cost of AWS helps to

lower the entrance bar for many users to get started. Also there is very extensive documentation as well as a very active community that can help new users to familiarize themselves with the services and to more easily find solutions to their wide-ranging problems.

My project can be split up into two main parts. First, a cloud system that can process, store, and analyze large amounts of data in a reasonable amount of time. The system will be receiving thousands of data logs every day and should be able to parse that data without crashing or taking too long. That data then must be formatted and inserted into a long-term storage database. This data will then be used by the next phase of this project.

The second phase consists of building out a web-app dashboard that is accessible and easy-to-use. Users should be able to access the dashboard and create an account for themselves. This account must be secure and have multi-factor verification enabled. Users then should be able to add data through this dashboard as well as generate graphs and tables from existing data.

Basic requirements such as the security of the data, scalability of the database in respect to the amount of data, and the ability to handle spikes of users are inherently addressed by the cloud services. AWS services have advanced built in load balancing, data scaling, and data protection. This makes implementing any new project a lot easier as these simple but highly important aspects of the project are essentially abstracted away.

The architecture I chose to build in order to meet the requirements detailed above was inspired by many related works found through online research and by official guides provided by

Amazon. Many times tradeoffs had to be made when choosing between two different implementation routes.

The frontend web-app dashboard was mainly written in React. The ChartJS framework was used extensively in order to generate detailed charts and graphs of the data. The dashboard was hosted on AWS Amplify which worked with Cognito in order to provide two-factor authentication to the web-app. Amplify was also configured with IAM roles that gives it permissions to communicate with the backend through Lambda functions. The backend consists of the AWS Aurora Serverless database. The database was connected to API Gateway which allowed outside cloud functions to access and parse it. Cloud Lambda functions were written and connected to API Gateway in order to gain access to the database. These cloud functions are called by the frontend in order to insert, delete, modify, and retrieve data from the database.

AWS CodeBuild was also set up with a repository in CodeCommit that contains the code for running nightly-tests, collecting the results, and storing the data. A pipeline was then built out in CodePipeline. This pipeline would trigger the CodeBuild functionality which in turn ran the code contained in the CodeCommit repository in it's own virtual environment. CodeBuild and CodePipeline both were connected to Aurora Serverless's Data API which allowed them to directly conduct Create, Read, Update and Delete (CRUD) operations on the data without the need to use Lambda and API Gateway. After CodeCommit finished, the pipeline would then use EventBridge and SES to send out email notifications of success or failure to stakeholders. The pipeline is triggered at a set time every night by a Jenkins job.

STS Discussion

Data can be a wonderful tool that provides previously hidden and valuable insight, but it can also be misleading and discriminative. This is an issue that impacts almost everyone in the world. Flaws in data collection, mistakes in the data analysis algorithms, or even misleading visualization of data can misdirect and give wrong information. Companies that rely on this data to make decisions will ultimately make flawed decisions that can have lasting impacts. There can never be a 100% guarantee that the data being used isn't flawed and thus that presents the question of how reliable it is to use data for decision making?

We need to first identify where our blind spots are before we can find ways to fix them. However, it seems that these blind spots exist in almost every aspect of our data pipelines and in every type of data we collect. For example, Black Americans are about four times as likely to develop kidney failure when compared to White Americans, but the algorithm that controls transplant placement order often places Black Americans lower on lists than White Americans (Christensen, Manley, Resendez et al., 2021). The flaw in these algorithms can be hard to resolve and they most aptly can be attributed to flawed data. These types of bias can be found everywhere and in all types of industries. Data often gives a comprehensive snapshot of the population, and therefore it is good at representing the average person under average circumstances. That means disabled people, racial minorities, and people in poverty all suffer from biases in these data algorithms, and the effects of these biases can have devastating impacts on these people's lives.

These algorithms and their underlying data control many aspects of people's lives that many don't even realize. They heavily influence everything from terms for bank loans to secure a mortgage or buy a car to the effectiveness of facial recognition technology (Brancaccio,

Conlon et al., 2021). The problem has become so prevalent and that it can no longer be attributed to simple errors in the algorithm but rather points to consistent flaws in the entire data processing lifecycle.

Now that the problem has been identified the natural next step is to find ways to resolve it. This part has proven to be the hardest part due to the fact that companies have spent a lot of time and resources into building these existing pipelines and they work well for the majority of the population. Actually solving the core issue would require spending even more resources to dismantle these networks and rebuild with mitigation of bias in mind. For example, the algorithms used by autonomous vehicles requires them to be trained on massive amounts of data for long periods of time (Christianson et al., 2020). If an inherent flaw were to be identified in them then the solution would not be as simple as tweaking a few lines of code but rather the entire process would need to be restarted from scratch with the proper data. This would cost companies years and render their past resource investments essentially worthless. Most companies would rather choose to ignore the problem unless they have the proper incentives to take action.

The goal of this research is to find out exactly how prevalent is the problem of flawed data, and what if anything can be done to streamline future fixes to the flaws. We also seek to explore what companies with existing flawed data pipelines should do to mitigate the damage that they cause. Many people tend to believe that data does not lie and therefore they place a great amount of trust on it. Top executives and world leaders rely on data to make important decisions, and when that data is wrong it triggers a cascading effect that may result in severely negative consequences. many people's lives.

Conclusion

Cloud technology is widely available and provide companies with powerful tools to develop services that can process massive amounts of data. Building these services is key to successfully using the data available to us on a mass scale. However it is necessary to keep in mind possible biases in the data and try to build pipelines that can help to eliminate these flaws.

Data is an inescapable part of our lives. We need to learn to leverage it to achieve our own goals but also need to be aware of its possible flaws.

Word count: 1774

References

- Gavin, M. (2019, July 16). *Business analytics: What it is & why it's important: HBS Online*. Business Insights Blog. Retrieved October 31, 2022, from <https://online.hbs.edu/blog/post/importance-of-business-analytics>
- Mesbahi, M. R., Rahmani, A. M., & Hosseinzadeh, M. (2018). Reliability and high availability in cloud computing environments: A reference roadmap. *Human-Centric Computing and Information Sciences*, 8(1). <https://doi.org/10.1186/s13673-018-0143-8>
- Raji, D. (2021, January 21). *How our data encodes systematic racism*. MIT Technology Review. Retrieved October 31, 2022, from <https://www.technologyreview.com/2020/12/10/1013617/racism-data-science-artificial-intelligence-ai-opinion/>
- Christensen, D. M., Manley, J., & Resendez, J. (2021, September 9). *Medical algorithms are failing communities of color*. Health Affairs Forefront. Retrieved October 31, 2022, from <https://www.healthaffairs.org/doi/10.1377/forefront.20210903.976632/full/>
- Brancaccio, D., & Conlon, R. (2021, August 25). *How mortgage algorithms perpetuate racial disparity in home lending*. Marketplace. Retrieved October 31, 2022, from <https://www.marketplace.org/2021/08/25/housing-mortgage-algorithms-racial-disparities-bias-home-lending/>
- Christianson, P. (2020, September 30). *Billions of miles of data: The Autonomous Vehicle Training Conundrum*. CloudFactory Blog. Retrieved October 31, 2022, from <https://blog.cloudfactory.com/autonomous-vehicle-training-conundrum>