Why Does the Cloud Stop Computing? Lessons from Hundreds of Service Outages

Haryadi S et. al.

Overview

The authors of this paper went over the coverage and documentation of outages at 32 popular internet services to understand the causes, severity and fixes associated with these outages. The sources of this information included news coverage and post-mortem reports of these outages. Data mining and analytics were applied to this information to obtain the findings presented in this paper. The findings concerned the number of outages suffered, uptime and downtime statistics, and insights into the single point of failure. The authors also present the causes for these failures and describe each one in some detail. The paper then goes on to discuss impacts and fixes and the pros and cons of their method before concluding that outages are inevitable and expressing their hope that the findings in their paper are of value to cloud developers, operators and users.

Thoughts on the Paper

This paper's methodology of performing data mining on news headlines and post mortem reports on outages is interesting. It was a novel way to leverage all the information that these providers of internet services have made publicly available. The paper's idea to make sense of these vast amounts of information is true to the purpose of data mining. As an enthusiast of text mining in particular, I found this application of text mining compelling. The paper's findings, particularly the ones regarding the root causes are also of particular note.

Unfortunately, I believe that this paper has laid its foundations on some weak assumptions and vague definitions. First of these, I wish the authors of this paper had explained why they picked the 32 services they did, since, in the paper there is no explanation for why these services where selected. I find it hard to believe the criterion was the number of users, since, for example, in the social media services, Google Plus is included, but Reddit is not. Another example is the exclusion of 'gig economy' apps such as Uber and GrubHub. I don't have a problem with the exclusion of Reddit per se, but I would have appreciated some rationale for choosing the services they did, such as the amount of information about outages available, perhaps. As it is, I can only speculate the reasons now.

I also find it discomfiting that on average only two sources were consulted regarding outages. Considering that only five parameters were extracted maybe more than 32 sources could have been considered. It may have been useful to have some more information about the distribution of source data, such as the median, outliers and so on. Regarding the collection of data, I also wish the authors had expanded a bit more on how they collected data, because as it is, they only say that they Googled outages for the service based on month and year. When I tried the same thing ("Google outage march 2012") I received outage information from March of this year. Obviously, the authors would have used some extra measures such as filtering results to appear from a certain time, or using quotes around the time to only obtain results containing that time, but I think it would have been good to have the exact steps the authors took to obtain this information.

I think that a problem of some severity is the definition of full outages and partial outages in this paper, or more specifically, the distinction between ‘essential’ and ‘non-essential’ services. The paper lists as examples of non-essential services changing profile pictures and background images. But while that service may be non-essential for a service like Gmail, when considering social media platforms, I would argue that outages of these services is actually quite severe. Not being able to change your profile picture on Facebook or Instagram is more severe than on Amazon.com. Given that the authors included a diverse range of services, it might have been better for them to consider the different components essential to each kind of service.

The root causes section was very fleshed-out and detailed, discussing the percentage of failures caused by a root cause and details on how that tends to happen. Unfortunately, the suggestions on what action to take seem sparse and not as detailed. For me, this dovetails into the exclusion of any qualitative analysis on fix procedures. The authors say that they could not discuss fixes in depth because of space constraints, but I wish they had worked around them, because similar to the analysis of root causes, I think that a rundown on the fixes and their effectiveness could have been really interesting, and would have enabled the authors to obtain the big picture on how different fixes would work for different outages, and in turn would have made the suggestions in the Root Causes section better tailored to the outages and more detailed.

As it is, the paper doesn’t feel as detailed as it could have. For this assignment, I also read the paper ‘Simple Testing Can Prevent Most Critical Failures’ a paper that has the same premise, i.e. analyzing a large number of failures and drawing conclusions about their causes, and it is difficult to not compare these two papers. The second paper spends a lot of time discussing the causes and potential fixes for the outages and comes across as more informative. This paper feels shallower in comparison, since the majority of the content seems to deal with listing statistics, and in my opinion, doesn’t really deliver on the ‘lessons’ in their title.

A final suggestion that I have for this paper would be that instead of considering as many outages as it did, or considering even more outages as I suggested earlier on, might be to take the opposite approach and only consider a handful of case studies of major outages, for example, the AWS outage that had rippling effects for other services that use it. That way, a lot of data could be mined from news reports, maybe the full text of the news articles instead of only the headlines, the full body of the post mortem report, and the general nature of the study could be more qualitative than quantitative, since I believe that this paper was at its best when it was qualitative in nature.

Conclusion

This paper had an interesting idea regarding mining publicly available outage data for popular services and worked around the constraints of this kind of data fairly well considering the circumstances but unfortunately is undermined by weak assumptions, vague definitions and a lack of detail in crucial areas.