Proceedings of the 2018 Pre-ICIS SIGDSA Symposium

Special Interest Group on Decision Support and Analytics (SIGDSA)

12-12-2018

Data Analytics with Personal Cloud: Cloud and Edge Data Replication using Ensemble Methods

Cipson Jose Chiriyankandath
Dakota State University, cjchiriyankandath@pluto.dsu.edu

Data Analytics with Personal Cloud: Cloud and Edge Data Replication using Ensemble Methods

Research-in-Progress

Cipson Jose Chiriyankandath
Dakota State University
cjchiriyankandath@pluto.dsu.edu

Abstract

Out of various personal cloud flavors, the IoT based Private Personal Cloud (PrPC) is an interesting concept because of their enhanced privacy options and easy to setup configurations. Earlier studies found that the content access performance in home-network using IoT based Edge computing outperforms content access performance using Cloud computing. Based on resource distribution concepts, this research proposes a novel hybrid approach to dynamically distribute content between Edge and Cloud using data analytics. Data analytics helps selecting the right content at right time to move from cloud to edge for better user experience. This research brings a unique value proposition by connecting user and user's data using data analytics. Contribution from this research is three folded. First, it characterizes the private personal cloud system. Second, it proposes a measurement model to empirically evaluate performance of private personal cloud. Third, it proposes a data analytics model for content replication.

Keywords

Cloud Computing, Fog Computing, Edge Computing, Personal Cloud, Cloud Storage, Application Performance, IoT, Data Analytics, Ensemble methods.

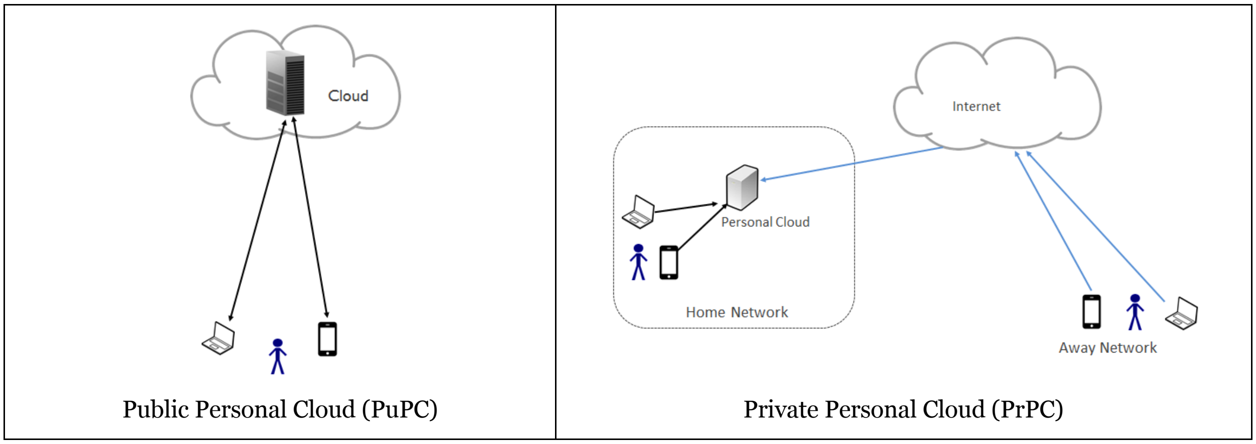
Introduction

Cloud computing is generally divided in to three types, public, private and mix of both (known as hybrid). Cloud computing for personal use is called Personal Cloud and it typically utilize the public cloud as backbone to achieve this. Recent popularity of personal cloud made individuals to increasingly use cloud storage systems for files access across multiple devices. This trend become a norm in recent years because of its tight coupling with smartphone platforms, iCloud for iOS users and GoogleDrive for Android users. Pure-play players like Dropbox is also attracting huge number of users to the personal storage space. Dropbox has half a billion users with 400 billion files in the cloud (Dropbox; Smith, 2018). Dropbox is just one of the leading players in this crowded personal cloud market. As Zhang (2016) commented, online backup is the most developed user focused application based on cloud storage.

Individual has little control over the personal cloud sitting in public cloud infrastructure (PuPC). This infrastructure is shared by many individuals across many locations. Data on PuPC are separated typically by authorization. There is an alternative approach in recent years that provides the convenience of the public cloud and the security of private cloud by setting up mini cloud hardware in user's premises, namely private personal cloud (PrPC). Notable products in this market are Western Digital My Cloud, Seagate Personal Cloud, Synology DiskStation and NETGEAR ReadyNAS. Personal cloud devices act as a central secure location for user to store and access files. This method allows user to retain full control over data while making the content accessible from anywhere. These products offer great deal of privacy over PuPC as the content is always staying at users' premises.

Private personal cloud (PrPC) functions work like public personal cloud (PuPC), except the fact that user has to run the personal cloud and connect it to internet all the time. Once user has set it up, any device can connect to it from anywhere (both from home-network and from away-network). However, this personal content storage has been challenged by some constraints that are unique to the PrPC environment. For example, even with their broad adoption, very little is known about the quality of service of the locally hosted personal clouds. Security and privacy studies on this area of personal computing is not exist. Current PrPC solutions in the market are painfully slow when accessing remotely compared to the PuPC solutions like Dropbox or Google Drive. The main bottleneck in the PrPC system is the limited network bandwidth available for the storage system. When user is away from home network, the fastest access speed user can expect from a private personal cloud is the user's internet connection's upload speed. Data asymmetry between downlink and uplink on average can be 4:1. Average US household internet upload speed is less than 20 Mbps while the average download speeds is about 60 Mbps (statista.com 2017).

Edge computing is a recently proposed computing paradigm that extends the cloud computing and services to the edge of the network. The motivation of edge computing is to place the contents and application services as close as possible to their consumers. Edge can augment the capabilities of the Cloud by extending the cloud to user's premises. By moving resource near the edge allows latency intolerant application performs well in-home network. Edge computing is referred as Fog computing in industry. Cisco Systems introduced the term Fog to indicate that the closer vicinity of cloud to user. The whole concept of privately hosted personal cloud (PrPC) can consider as edge computing system. PrPC enables low latency service delivery using Edge computing. Figure 1 depicts the public and private personal cloud systems.



Public Personal Cloud (PuPC)

Private Personal Cloud (PrPC)

Figure 1. PuPC and PrPC Systems

As Tolia, Andersen & Satyanarayanan (2006) mentioned, when network quality degrades, interactive performance suffers. Even with the high adoption rate, PuPC solutions (Dropbox, Google Drive, One Drive etc.) have some performance constraints due to the high network latency and the slow network bandwidth. This problem is more relevant when user access his/her own content in in-home network. For example, a high definition (4K Video) sitting in the public cloud may not play well when user tries to play it on his/her own TV by streaming from public cloud. Accessing latency intolerant contents like Video or Games or VR can experience performance issues when accessing from public cloud.

On other hand, locally hosted PrPC solutions provide greater performance when accessing the content in home network. When user tries to play the high definition video on his/her own TV by streaming from a locally hosted PrPC, it provides greater streaming performance compared to public cloud, as the streaming is happening within the local LAN (home network). However, accessing the same content in non-home network is painfully slow.

This research investigates how data analytics helps in enhancing the content replication policies between cloud and edge. The key decision criteria for model selection is rapid exploration of models. Ensemble methods built on many independent models can best suited to predict unique user content access patterns.

It is clear that moving contents' proximity to the user is important when connectivity degrades and latency intolerance increases. Allowing data to be closer to users by reducing the distance of content to users yield better user experience, but this require careful movement of resource between locations. User's content access pattern allows machine learning techniques to predict the potential use of content and allow the resource to be duplicated in Cloud or Edge. The main challenge in this process is to predict the content that needs to be replicated or moved based on access patterns and spatio-temporal variables. End-user computing satisfaction (EUCS) is an appropriate theory foundation to verify the predictive resource replication performance.

Prior Research

There is a lot of interest within research community around personal cloud topic, but the term personal cloud is blurry and used differently by researchers in different environments. Researches on Personal cloud are mostly revolving around the public clouds that are tuned towards personal use (e.g. Dropbox, GoogleDrive, and OneDrive). There has been little theoretical development and research on personal clouds within the cloud computing and other steams. There are some research works that are related to public personal clouds (PrPC) in recent years. Drago et al. (2013) is talking about benchmarking the personal cloud storages, however this research is more towards public clouds for individuals. Another research in this line is from Gracia et al. (2013), it presents a measurement study of three major personal clouds: DropBox, Box and SugarSync. Natu et al. (2016) address the performance monitoring issues of hybrid clouds.

There are some studies in the early years of cloud computing that talks on different approaches of cloud computing. Especially, the work of Tian, Song, & Huh (2001) propose the development of methodologies tailored for personal cloud computing. However, this study also focusses towards public clouds with emphasis on security architecture of the personal cloud. Riva et al. (2001) proposes a scalable policy-based replication system that addresses the resource distribution within cloud; however, the proposed system is based out public cloud (Amazon) to demonstrate the personal cloud concept. Another research from Ardissono et al. (2009) proposes a unified environment for handling personal cloud activities. This one resembles this paper's research problem; however, it distributes content across different cloud provider to benchmark the performance.

Even though Edge or Fog (these two terms are interchangeably used by research community) computing is a new paradigm, there is an increased interest among researchers on this form of cloud computing. Bonomi et. al (2012) defines the characteristics of Edge computing as Low latency, location awareness, distributed, strong presence of streaming and real time applications. An interesting research by Luan et al. (2015) on how the fog can extend the cloud to user's premises is well addressed some of the challenges and approaches. The same paper also talks about the dedicated and localized service applications for better performance. In similar manner, Zhu et al. (2013) propose fog computing approach for improving web sites performance. A recent study by Aazam & Huh (2015) proposes dynamic resource estimation and pricing model for internet of things application. Hong et al. (2013) proposes a programming model for large-scale applications called Mobile fog. Zhang et al. (2016) propose a distributed edge environment that process data in IoT environment. Arkian et al. (2017) propose a fog-based resource provisioning using data analytics. Bashir & Gill (2016) propose a framework to store and analyze large amount of IoT data using big data analytics.

A work by Satyanarayanan et al. (2009) shows that latencies over internet can be high and it can interfere with interactive applications. Lillethun et al. (2013) proposed an integrated architecture for pervasive and high-performance computing called MediaBroker to address the latencies. Another work by Clinch et al. (2012) proposes 'Cloudlets' to address latencies for interactive applications.

The work by Hong et al. (2013) provides a high–level programming model that simplifies development of fog devices distributed over a wide area network. The framework also allows applications to dynamically scale based on their workload using on–demand resources in the fog and in the cloud. This research will utilize the mobile fog programming model suggested by Hong et al. (2013) in personal cloud context.

End-user computing satisfaction (EUCS) is an instrument developed by Doll and Torkzadeh (1988). It represents five underlying dimensions of end-user satisfaction, which are Content, Accuracy, Format, Ease of use, and Timeliness. Timeline is the best dimension that can be used in this research.

Research Gap

There has been little research on user hosted private personal clouds. Most literature available on personal cloud area are related the public clouds that are tuned towards personal usage (Dropbox, GoogleDrive, SkyDrive etc.). Studies on application performance for low latency requirements (e.g. video streaming, video gaming, Virtual Reality applications etc.) in personal cloud context are rare. User experience of private personal cloud is rarely studied.

Research Question and Hypotheses

This study addresses three research questions related to the personal cloud user experience. First, how well the personal cloud performs under different network conditions like in-network (home network) and out-network (away network)? Second, can we increase the content access performance by dynamically replicate resources/contents across in-network and out-network using 'Edge' computing concepts? Finally, will the improved content access performance enhance the overall user experience of personal cloud?

Effective information presentation can enhance user's perception and can lead to more effective content utilization (Dull & Tegarden, 1999). Therefore, this research believes that by dynamically replicating contents across in-network and out-network will improve content access performance and hence user experience. Based on this belief, following hypotheses were derived.

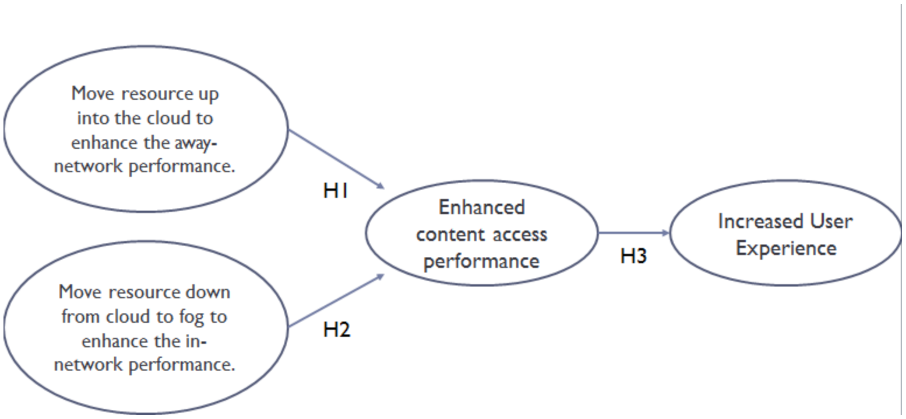

Move resource up into the cloud to enhance the away-network performance.
Move resource down from cloud to fog to enhance the in-network performance.
H1
H2
Enhanced content access performance
H3
Increased User Experience

Figure 2. Conceptual Model

Hypothesis-1

H1: Content Access performance will increase when resource is in cloud while accessing from out-network.

Hypothesis-2

H2: Content Access performance will increase when resource is in edge while accessing from in-network.

Hypothesis-3

H2: End User experience will increase when content access performance is increased.

Research Methodology

This research follows design science research approach guidelines by Hevner et al. (2010). The objective of this research is to develop a conceptual model that increases the user experience by improving content access performance of personal cloud using data analytics.

Proposed System

To test the hypotheses, this research will implement a prototype system that comprises three parts, a 'Cloud' component, a 'Edge component, and a 'Client' component. Public personal cloud (PuPC) criteria will meet when client access content from cloud component. Private personal cloud (PrPC) criteria will meet when client access content from locally hosted device (Edge device). Figure 3 depicts the component diagram for the proposed system.

The 'Cloud' component will reside in cloud and will help user to access content from out-network (away network). The Cloud component in this prototype system will be a server software running on Amazon webserver (a Java server program). 'Edge' component will sit in the in-network (local LAN) and accelerate content access in in-network (home network). The edge component will be developed as desktop software intended to run on desktop machines (a Python application running on Windows OS). The 'Client' component will be smart phone application developed for Android platform.

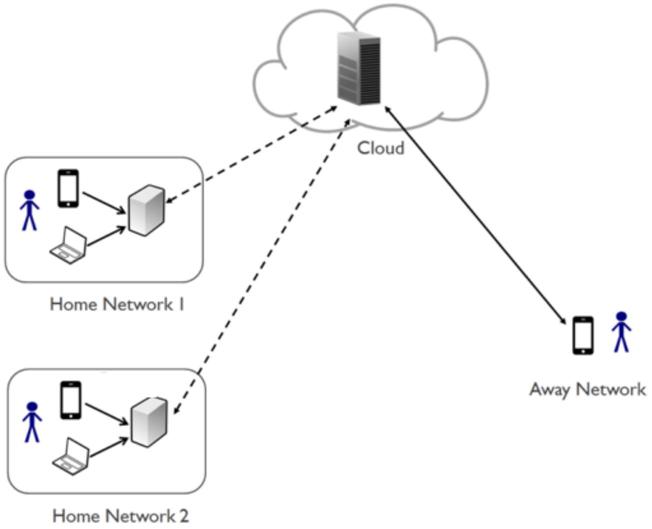


Figure 3. Proposed System

Measurement Model

Focus of this measurement model is to benchmark both PrPC and PuPC services to help users and business to choose the right personal cloud service under real life conditions. Ensuring the accuracy of content access performance is the key in cloud benchmarking. The test results of these PuPC and PrPC may vary based on user's conditions. Client applications communicate with cloud using different protocols, but they fall under 3 categories. Protocol based on HTTP, Protocols based on FTP, and private protocols. (Zhang 2013). The proposed measurement model described in Table-1 is agnostic of any protocols. Detailed description of measurement criteria is explained in Table-2.

The proposed measurement model is designed to handle various conditions that meet the various research validity criteria. Number of tests in each category is designed to have statistically significant iterations. The measurement model is designed in such a way that new constraints can be easily added to the measurement metrics. The model is generic enough to supports single and multiple clients.

Time is the measurement criteria for each iteration. Time is measured as the difference between the first and the last packet of the payload communicated between client and personal cloud. All experiments are under the same network condition.

File Management				
	Cloud type 1	**Step #1**		
		Home Network	**Step #2**	
			Upload	**Step #3**
				No of files
				File type
				File size
				Time of the day
				Day of a week
				Parallel and serial
			Download	Repeat **#3**
		Corporate network	Repeat **#2**	
		Public network	Repeat **#2**	
	Cloud type 2	**Repeat #1**		

Table 1. Measurement Iterations

Content replication using Ensemble model

The proposed concept of moving or replicating content on cloud or edge is guided by analytics. Ensemble method has several advantages over other methods particular to personal cloud data replication. Combining the resource allocation models from several classifiers not only enhance the performance of the model, but also reduce the risk of choosing non-performing model. Ensemble methods are better positioned to address too-little or too-much data problem (Polikar, 2006). This is particularly import in this case, as users may have varied resource access behavior than others.

The scale of user content access and content pattern gives us ample opportunity for predicting user's desired content access using data analytics methods. Out of various data analytics methods, this paper tries to explain why ensemble-based methods more suitable than single classifier methods. This research utilizes popular ensemble methods like bagging, boosting, AdaBoost, stacked generalization, and bayesian parameter averaging.

Data Collection

The hypotheses will be tested through two type of validations. A controlled laboratory experiment will validate H1 and H2 by benchmarking the content access performance using the measurement model explained Table-1. A survey method will validate H3 by evaluating the end user experience.

In controlled experiment benchmarking, 3 different scenarios will be tested against default public personal cloud (PuPC) and edge assisted private personal cloud (PrPC).

- File browsing
- Audio streaming
- Video streaming:

To evaluate end user experience, this research will follow the EUCS guidelines given by Doll and Torkzadeh (1988). Participants will be asked to perform in their own environment the same three scenarios defined in pervious experiment using both PuPC and PrPC systems. Participants will be asked to answer a survey to validate the end user experience.

Variable	Description
Home Network	Typical home network where a router connected to ISP network. Typical download/upload speed is 60/15 mbps (based on FCC 2017 internet speed report)
Corporate Network	This measurement vary organization to organization, but the idea is to get a network where download upload asymmetry is less.
Public Network	Public library, Coffee shop internet, City provided internet etc. comes under this network.
Number of files	This test validates the performance against number of files in the test. The test numbers are chosen in curvilinear way 1, 2, 4, 8, 16, 32, 64.
File type	This test validates the performance against different types of files. Test file types includes text (.txt), image (.jpeg), binary (.exe), compressed (.zip)
File size	This test validates the performance against the size of the files in the test. The file is chosen in nonlinear way 1kb, 10kb, 100kb, 1Mb, 10Mb, and 100 Mb
Time of the day	This test validates the performance against the time of the day. To capture peak and non-peak time, the test validates; 12.00 AM, 3.00 AM, 6.00 AM, 9.00 AM, 12.00 PM, 4.00 PM, 7.00 PM, 10.00 PM
Day of a week	This test validates the performance against the day of the week to understand weekday and weekend difference, the test validates 7 days a week.
Parallel and serial	This test validates the performance against concurrency.

Table 2. Measurement variables

Contributions

The immediate contribution of this research is to provide a framework for enhancing private personal cloud (PrPC) experience. Privacy and Security of personal content make it compel for individuals to adopt private personal cloud. Furthermore, the proposed content replication framework based on data analytics could serve as a model for the development of similar systems.

Theoretical contributions are around the validation of End-user computing satisfaction (EUCS) theory in personal cloud context.

Conclusion

This study represents one of the first efforts on quantifying content access performance on IoT based private personal cloud. The emergence of edge computing allows personal clouds to distribute content/resource to different storage locations based on user's presence. As Luan et. al (2015) describes, the motivation of the edge computing is to place the contents and application services as close as possible to their consumers. When compared to cloud, edge computing can provide enhanced user experience by reducing service latency and response time. Unique user content access behavior and content access patterns make ensemble data analytics method as a compelling choice for content replication problem. This research described some of the key challenges of private personal cloud and propose a measurement model for evaluating its performance. This research is one of its kind in this area. In particular, it proposes a framework to increase end-user satisfaction of private personal cloud systems. Early results show the proposed dynamic content replication significantly enhance in-home and away-network content access performance.

References

Aazam, M., & Huh, E. N. (2015, March). Dynamic resource provisioning through Fog micro datacenter. In Pervasive Computing and Communication Workshops, 2015 IEEE International Conference on (pp. 105-110). IEEE.

Ardissono, L., Goy, A., Petrone, G., & Segnan, M. (2009, September). From service clouds to user-centric personal clouds. In Cloud Computing, 2009. CLOUD'09. IEEE International Conference on (pp. 1-8). IEEE.

Arkian, H. R., Diyanat, A., & Pourkhalili, A. (2017). MIST: Fog-based data analytics scheme with cost-efficient resource provisioning for IoT crowdsensing applications. Journal of Network and Computer Applications, 82, 152-165.

Bashir, M. R., & Gill, A. Q. (2016, December). Towards an iot big data analytics framework: Smart buildings systems. In High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016 IEEE 18th International Conference on (pp. 1325-1332). IEEE.

Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012, August). Fog computing and its role in the internet of things. In Proceedings of the first edition of the MCC workshop on Mobile cloud computing (pp. 13-16). ACM.

Cao, Y., Chen, S., Hou, P., & Brown, D. (2015, August). FAST: A fog computing assisted distributed analytics system to monitor fall for stroke mitigation. In Networking, Architecture and Storage (NAS), 2015 IEEE International Conference on (pp. 2-11). IEEE.

Clinch, S., Harkes, J., Friday, A., Davies, N., & Satyanarayanan, M. (2012, March). How close is close enough? Understanding the role of cloudlets in supporting display appropriation by mobile users. In Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on (pp. 122-127). IEEE.

Do, C. T., Tran, N. H., Pham, C., Alam, M., Rabiul, G., Son, J. H., & Hong, C. S. (2015, January). A proximal algorithm for joint resource allocation and minimizing carbon footprint in geo-distributed fog computing. In Information Networking (ICOIN), 2015 International Conference on (pp. 324-329). IEEE.

Drago, I., Bocchi, E., Mellia, M., Slatman, H., & Pras, A. (2013, October). Benchmarking personal cloud storage. In Proceedings of the 2013 conference on Internet measurement conference (pp. 205-212). ACM.

Dropbox | Company Info. (n.d.). Retrieved from https://www.dropbox.com/news/company-info

Dull, R. B., & Tegarden, D. P. (1999). A comparison of three visual representations of complex multidimensional accounting information. Journal of Information Systems, 13(2), 117-131.

Giang, N. K., Blackstock, M., Lea, R., & Leung, V. (2015, October). Developing iot applications in the fog: a distributed dataflow approach. In Internet of Things (IOT), 2015 5th International Conference on the (pp. 155-162). IEEE.

Gracia-Tinedo, R., Artigas, M. S., Moreno-Martinez, A., Cotes, C., & Lopez, P. G. (2013, June). Actively measuring personal cloud storage. In 2013 IEEE Sixth International Conference on Cloud Computing (pp. 301-308). IEEE.

Hevner, A., & Chatterjee, S. (2010). Design science research in information systems. In Design research in information systems (pp. 9-22). Springer, Boston, MA.

Hobfeld, T., Schatz, R., Varela, M., & Timmerer, C. (2012). Challenges of QoE management for cloud applications. Communications Magazine, IEEE, 50(4), 28-36.

Hong, K., Lillethun, D., Ramachandran, U., Ottenwälder, B., & Koldehofe, B. (2013, August). Mobile fog: A programming model for large-scale applications on the internet of things. In Proceedings of the second ACM SIGCOMM workshop on Mobile cloud computing (pp. 15-20). ACM.

Luan, T. H., Gao, L., Li, Z., Xiang, Y., & Sun, L. (2015). Fog computing: Focusing on mobile users at the edge. arXiv preprint arXiv:1502.01815.

Meurisch, C., Seeliger, A., Schmidt, B., Schweizer, I., Kaup, F., & Mühlhäuser, M. (2015). Upgrading wireless home routers for enabling large-scale deployment of cloudlets. In Mobile Computing, Applications, and Services (pp. 12-29). Springer International Publishing.

Na, S. H., Park, J. Y., & Huh, E. N. (2010, December). Personal cloud computing security framework. In Services Computing Conference (APSCC), 2010 IEEE Asia-Pacific (pp. 671-675). IEEE.

Natu, M., Ghosh, R. K., Shyamsundar, R. K., & Ranjan, R. (2016). Holistic Performance Monitoring of Hybrid Clouds: Complexities and Future Directions. IEEE Cloud Computing, 3(1), 72-81.

Nishio, T., Shinkuma, R., Takahashi, T., & Mandayam, N. B. (2013, July). Service-oriented heterogeneous resource sharing for optimizing service latency in mobile cloud. In Proceedings of the first international workshop on Mobile cloud computing & networking (pp. 19-26). ACM.

Nunamaker Jr, J. F., Chen, M., & Purdin, T. D. (1990). Systems development in information systems research. Journal of management information systems, 7(3), 89-106.

Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and systems magazine, 6(3), 21-45.

Riva, O., Yin, Q., Juric, D., Ucan, E., & Roscoe, T. (2011, October). Policy expressivity in the Anzere personal cloud. In Proceedings of the 2nd ACM Symposium on Cloud Computing (p. 14). ACM.

Satyanarayanan, M., Bahl, P., Caceres, R., & Davies, N. (2009). The case for vm-based cloudlets in mobile computing. Pervasive Computing, IEEE, 8(4), 14-23.

Smith, C. (2018, October 10). 33 Staggering Dropbox Statistics and Facts. Retrieved from https://expandedramblings.com/index.php/dropbox-statistics

Stojmenovic, I., Wen, S., Huang, X., & Luan, H. (2016). An overview of fog computing and its security issues. Concurrency and Computation: Practice and Experience, 28(10), 2991-3005.

Tian, Y., Song, B., & Huh, E. N. (2011, April). Towards the development of personal cloud computing for mobile thin-clients. In Information Science and Applications (ICISA), 2011 International Conference on (pp. 1-5). IEEE.

Tolia, N., Andersen, D. G., & Satyanarayanan, M. (2006). Quantifying interactive user experience on thin clients. Computer, (3), 46-52.

Zhang, Q., Zhang, X., Zhang, Q., Shi, W., & Zhong, H. (2016, October). Firework: Big data sharing and processing in collaborative edge environment. In 2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb) (pp. 20-25). IEEE.

Zhu, J., Chan, D. S., Prabhu, M. S., Natarajan, P., Hu, H., & Bonomi, F. (2013, March). Improving web sites performance using edge servers in fog computing architecture. In Service Oriented System Engineering (SOSE), 2013 IEEE 7th International Symposium on (pp. 320-323). IEEE.