Data Mesh: Practices from a Single Domain Perspective

(Technical Paper)

The Impact of Cloud Computing as a Socio-Technical Phenomenon

(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
By
Brian Mbogo

Fall 2022

Technical Team Members:

Brian Mbogo

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.

ADVISORS

Joshua Earle, Department of Engineering and Society

Briana Morrison, Computer Science

Introduction

It is often, and correctly stated, that "there is no cloud, it's just someone else's computer," which has profound implications for our relationship to modern technology, as my STS project will address. My technical report is a proposal motivated by my experience as a summer intern at Pluralsight. It aims to answer the question of what specific team and technical practices are important for successfully implementing data mesh data architecture. I will also address how these practices are impacted by company and team culture. The most important feature of data mesh is that it divides the handling of data in a company by the domain to which it is related rather than the operations performed on it. For example, at Spotify, data would hypothetically be split into the domains of music, podcasts, and accounts rather than data acquisition, data processing, and data display. Domain-based division allows a single team with thorough understanding of a domain to handle the data's entire lifespan. My STS project will focus more closely on cloud computing, the capability for people to request services like storage and computation from powerful, physically remote computers, which is a major part of modern data analytics. In my STS project, I aim to describe how the existence of cloud computing changes existing relations between companies, technologists, and everyday people.

The questions I answer in my technical report are important to address because in the rapidly changing technology field, the people whom new methodological developments are supposed to help often lack the bandwidth to appropriately implement these new methods. By providing specific guidelines with reference to an actual company, my technical report gives confidence to readers that cannot be found in the (rightly) very general initial descriptions of data mesh architecture, thus enabling them to adopt and adapt it to their specific work. The questions addressed by my STS project are important because they focus on a specific development in the

computing world which has wide-ranging applications, and thus, implications. Cloud computing has been wholeheartedly embraced by individuals and institutions, yet not much attention is paid to the potential negatives of its widespread adoption. By exploring the overall change in relations caused by cloud computing, my STS project will be able to pick apart both the positives and negatives.

In this prospectus I will include the abstract of the technical report, a discussion of the scope of the STS project, and a review of the important texts that I will use for the STS project. The project scope includes the research question, relevant social groups, methods, and a timeline.

Technical Project

As modern businesses' data usage grows, so does their need for methods and technologies that preserve their agility for meeting business objectives in the face of the challenges posed by data. Pluralsight, a technology up-skilling and education platform uses a data mesh architecture to meet these modern data challenges. It ensures that each data endpoint team has the necessary data for its work serving independently defined internal and external customers and reduces duplication of effort when multiple teams are working with the same data.

This report will use the perspective of the analytics team in which I was embedded to offer practical proposals for the implementation of a data mesh. The analytics team works with two internal teams and a single idea of a "large" external customer. The team's preferred technology stack includes Apache Spark, cloud analytics infrastructure, a continuous integration/deployment platform and the necessary interface for company-wide data transfer. Within this team, the result of the chosen technology stack is an efficient, largely formulaic development process for any given data product. Lastly, this report will present possible directions for future work in data architecture that build on existing team practices.

STS Project

As an outgrowth from my technical report, my STS project will focus on the social implications of cloud computing, which is a major part of modern data analytics (Khan et al., 2022). Cloud computing is a term that encompasses a variety of practices, but they all center around the existence of large scale, remote computing hubs, server farms, or the term which I will use throughout this work, data centers. As businesses and technologists seek to reduce their overhead for information technology (IT), these groups have increased their reliance on computation by people with specialized resources (Ranger, 2022). Additionally, with people creating and storing more digital files than ever, cloud storage is gaining more individual use for personal and business purposes because of the space and flexibility it offers (Sebastien, 2021).

I aim to explore how the existence of cloud computing changes existing relations between companies, technologists, and everyday people. This appears to be an important question to me not only because of the increasing proliferation of cloud computing, but also because of the curiosity that a colleague from my internship sparked in me. He said that his hometown is in the middle of a period of rapid development, that one of the big five tech giants (formerly FAANG) is building a new data center there, and that the construction site of that data center is invisible on some major map applications. I have an improved awareness of the complexity of the social systems through which technologies operate thanks to STS, so when I revisited this anecdote, I could not help but think of all the other effects that cloud computing can have on people's lives outside of its "simple" function. Visible social effects such as in my colleague's anecdote and the speed of the proliferation of cloud computing compound into an urgency to address how cloud computing changes existing relations.

It is important, for the sake of a consistent analysis, to properly define the "companies, technologists, and everyday people" the change in whose relationships my STS project will explore. I will define a company as any enterprise with specific products or services, and a target customer base, regardless of size. The distinction of company size is not of great importance when talking about cloud computing because of its availability at multiple tiers (Sebastien, 2021). The social group of technologists is defined as people who are knowledgeable and participate in technological innovation through some sort of tinkering, regardless of the success or results of that tinkering. This group consists of institutional researchers and people who independently explore and modify technology for personal fulfillment or amusement. They either directly explore cloud computing or they use services that depend on it and are familiar with the nature of that dependence. Everyday people are those who simply rely on cloud-supported products without familiarity with the nature of the dependence.

Because all of these groups are defined in direct relation to the pure function of cloud computing, I am leaving out social groups that don't use the functionality of cloud computing but may be related to it in some way. These groups include the people involved in the resource extraction necessary to create data center equipment and people isolated from the cloud computing phenomenon for a multitude of other reasons. It is important to consider which groups are not included in my primary consideration in order to avoid, as much as possible, giving an incomplete account of the social relations. I will need to explicitly mention the reasons for their exclusion.

I will primarily use actor-network theory (ANT) to help answer my research question. ANT is the best framework to address my research question because the research question deals with changes in relationships between groups. The groups can be easily understood as the actors

in the network of interest and the relationships as the connections between those actors. To aid the development of the actor network, I will unpack the stories that are told about cloud computing through media. For these stories, I will primarily look at scientific literature and popular technology magazines. Identifying protagonists and antagonists in stories will help expand the network because each constituent of a story, when explored, can be part of another story. Some examples include discovering competing stories from antagonists or groups left out of the original story.

The timeline for the STS project will be in multiple stages ranging from the research to the final draft. The research phase will begin with an iterative process to build up the resources. The basic steps of this iteration will be looking at sources, developing the actor network and then analyzing (especially human) actors that are left out. I will then use these to find sources that explicitly include those actors. Because actor networks get very big very fast, I will limit third-order connections as much as possible without detracting from the quality of the research. In the second phase, after I have decided on a robust set of sources, I will create brief explanations of their contributions to the project. In the third phase, I will do the writing. I will likely need to add or subtract some sources during this phase as the writing develops more clearly. Finally, I will use my notes from the second phase to complete the citations.

Key Texts

The first text I am considering for my STS project is “The limits of computation: A philosophical critique of contemporary Big Data research” by Peter Törnberg and Anton Törnberg. In this article, Törnberg and Törnberg discuss the ways in which modern Big Data research relates to the methodology of the social sciences, especially with a postmodernist approach. This is an important article for my project because its discussion of methodological

changes also includes discussion of system-like, fluid social structure in general, which is of interest in my exploration of the changes in social relations brought about by cloud computing.

The second text I am considering is “The politics of buzzwords at the interface of technoscience, market and society: The case of ‘public engagement in science’” by Bernadette Bensaude Vincent. In this article, Vincent explores the phenomenon of buzzwords, specifically focusing on “public engagement in science.” Vincent describes specific functions Buzzwords perform. By describing how the social phenomenon of this specific buzzword affects the development of science from the lab and governmental level, this article offers a way for me to explore cloud computing in the same way.

The third text I am considering is “The Value of Accountability in the Cloud: Individual Willingness to Pay for Transparency” by Wouter M.P. Steijn and Maartje G.H. Niezen. In this article, Steijn and Nizen explore the people’s relationship to accountability in cloud computing. They discuss the importance of accountability and people’s willingness to pay for it in cloud computing services. This is a valuable paper for my project because it mentions the power difference at play in cloud computing accountability and a method of addressing it.

The fourth text I am considering is “Systematic analysis of software development in cloud computing perceptions” by Habib Ullah Khan *et al*. In this article, Khan *et al* combine a participating entities discovery with a systematic literature review of cloud computing with specific attention to software development practices. The discovery of participating entities is especially helpful for me to begin formulating my actor network, but of course, the literature review will also provide direction to my research on technologists. Another strength of this article is its definitions of technical terms.

Works Cited

Bensaude Vincent, B. (2014). The politics of buzzwords at the interface of technoscience, market and society: The case of ‘public engagement in science.’ *Public Understanding of Science*, *23*(3), 238–253. https://doi.org/10.1177/0963662513515371

Khan, H. U., Ali, F., & Nazir, S. (2022). Systematic analysis of software development in cloud computing perceptions. *Journal of Software: Evolution and Process*, *n/a*(n/a), e2485. https://doi.org/10.1002/smr.2485

Ranger, J. (2022). *What is cloud computing? Everything you need to know about the cloud explained*. ZDNET. https://www.zdnet.com/article/what-is-cloud-computing-everything-you-need-to-know-about-the-cloud/

Sebastien, N. (n.d.). *Usage & Trends of Personal Cloud Storage: GoodFirms Research*. Retrieved October 25, 2022, from https://www.goodfirms.co/resources/personal-cloud-storage-trends

Steijn, W. M. P., & Niezen, M. G. H. (2015). The Value of Accountability in the Cloud: Individual Willingness to Pay for Transparency. *IEEE Technology and Society Magazine*, *34*(4), 74–82. https://doi.org/10.1109/MTS.2015.2494373

Törnberg, P., & Törnberg, A. (2018). The limits of computation: A philosophical critique of contemporary Big Data research. *Big Data & Society*, *5*(2), 2053951718811843. https://doi.org/10.1177/2053951718811843