Development of Alert Software for Real-time Geospatial Tracking System
(Technical Topic)

Impacts and Obstacles Behind Real-time Data Processing in Ridesharing Services
(STS Topic)

A Thesis Project 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

Kayla Lewis

Fall, 2021

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:

STS Advisor: Kathryn A. Neeley PhD Department of Engineering and Society

Technical Writing Advisor: Rosanne Vrugtman PhD Department of Computer Science

Technical Advisor: Daniel G. Graham PhD Department of Computer Science

Applications of Real-time Data Streaming in Geospatial Systems

It is estimated that by 2025, 150 billion devices across the globe will be connected and generating real-time data (Condon, 2018). Real-time data streaming is the process by which large amounts of data are presented to and processed by a system in a nearly instantaneous manner. This type of process is often used by organizations to analyze big data quickly enough to adapt to changing sociotechnical systems. Safaei (2017) explains that "Monitoring (e.g., network traffic, sensor networks, healthcare, etc.), surveillance, web-clicks stream, financial transactions, fraud and intrusion detection are some applications of streaming big data" (p.2). These applications are continuing to gain critical roles in society and because of that, there are significant concerns that must be addressed.

Real-time data streaming systems often have little to no capacity for long-term data storage, so the data that passes through these systems has a small range of time to be processed before it is removed from the system or loses its value. Systems utilizing real-time data streaming need to have low storage latency costs and high processing efficiency (Stonebraker, M et al., 2005). If these systems are not properly designed and maintained to ensure they are at peak efficiency, these systems could fail to operate as intended. For example, real-time data streaming is often used in critical systems such as network anomaly detection. If this security system were to have significant slow down or failure from inefficient data processing, this could lead to cyber attacks and information theft (Ariyaluran Habeeb et al, 2019).

Real-time data streaming can provide many great benefits to certain technical systems, especially those that center on complex and ever-changing data like geospatial

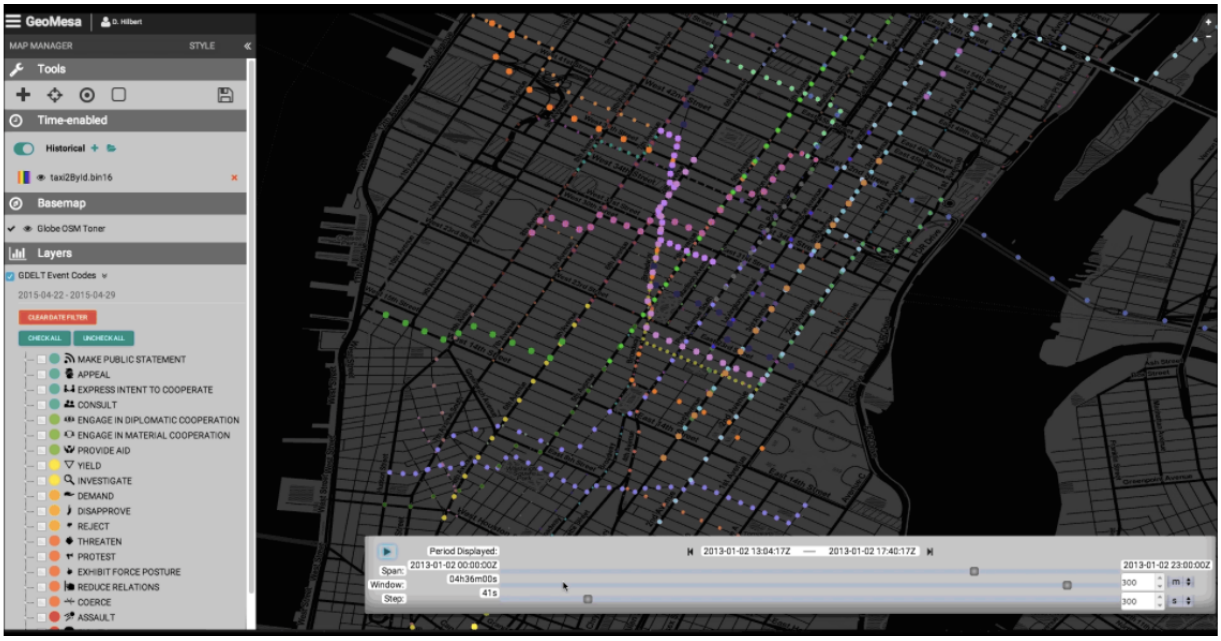
systems. For my technical project, I addressed the complication of expiring location data within a real-time geospatial system by creating an application that is able to analyze location data in real-time and alert users when an event that they have specified occurs. For my STS topic, I will investigate the impacts and obstacles behind real-time data streaming while paying special attention to the use of these systems in ridesharing services. The goal of the STS portion of my project is to draw attention to the criticality and importance of real-time data streaming systems within ridesharing services, and how the efficiency of this system could have a huge impact on urban areas.

Development of Alert Software for Real-time Geospatial Tracking System

Geospatial systems use location data that is highly complex and operates on a large scale (Gu, K., Yang, L., & Yin, B. , 2018). Because of this, geospatial systems that analyze big location data must be efficient in every way possible. In the case of a geospatial tracking system, any real-time location data must either be stored somewhere to be processed later, which can be very costly in terms of storage space and time, or it must be acted upon immediately. The geospatial tracking system I worked with for my technical project used real-time data from numerous sources to track aerial and maritime traffic, and this data was not placed into any form of long term storage. Real-time geospatial systems like this require large amounts of location data, which often needs to move though the system quickly and securely, or else there is a risk of some form of system failure (Li, S., Dragicevic, S. et al., 2016). In the case of my technical project, system failure would be a noticeable processing slowdown which would affect the accuracy of the location data.

A user viewing the graphic interface of the geospatial tracking system I worked with could see aerial or maritime traffic events on a global map that were currently taking place similarly to the system shown in Figure 1 which happens to be displaying some form of location data within a city.

Figure 1: Geomesa Geospatial Interface



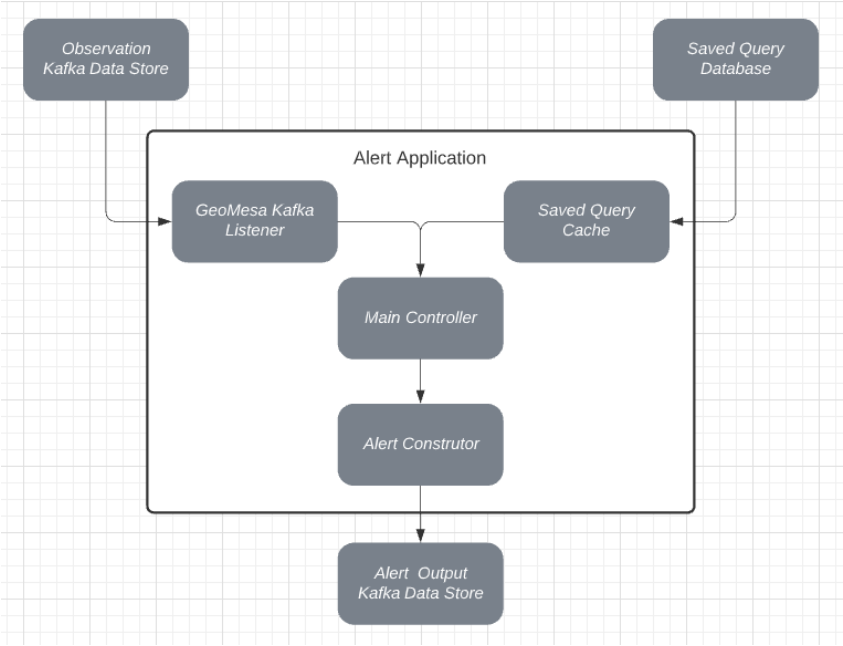
This image shows the graphical interfaces of a geospatial spatial system that uses real-time data in a similar manner to that of my technical project. The panel on the left allows users to select different layers of live data (Geomesa, n.d., n.p.).

In a tracking system like this, it is often useful to track potential future events as well. It was my task to develop an application that contained this crucial missing functionality and could be run in parallel to this existing geospatial tracking system. Within this system, a geospatial event could often be defined as an entity or group of entities entering or exiting a certain geographic region. The tracking system already contained a way for users to click on a side panel, similar to the panel on the left of Figure 1, and create saved queries for potential geospatial events, but those saved queries were not yet utilized. If this problem

was not properly addressed by the development of an alert application, there would not be a way for users to track future events because the data in this system is displayed and analyzed in real-time and then discarded. This need for an alert system is a great example of an imposed requirement that real-time data streaming places on a system. Despite this, the benefit of real-time data streaming in this system, such as little to no storage space and query latency, outweighs the extra procedures needed to utilize the location data effectively.

My technical project builds onto the functionality of an existing geospatial tracking system by reading data from the same input data streams and saved query databases that are used in the tracking system. This design is seen in Figure 2 where the two inputs to the alert application are the location observations and saved user queries.

FIgure 2: Alert Application Components



This diagram shows the general flow of data between components with the components inside the main square being the components that are a part of the alert application (Created by Author).

A lot of the challenge of my project revolved around ensuring the data used by this massive system ran smoothly through the alert application, so that the state of the application was in synchronization with that of the existing geospatial tracking system. This means that the alert application needs to process location data as fast as or faster than the tracking system . The overall method and design of my technical project relates to the general research topics of big data optimization and analytics techniques. My technical project also relates to and utilizes research centered on data streaming frameworks such as that in "A Study of Apache Kafka in Big Data Stream Processing" (Hiraman, B. R. et al., 2018).

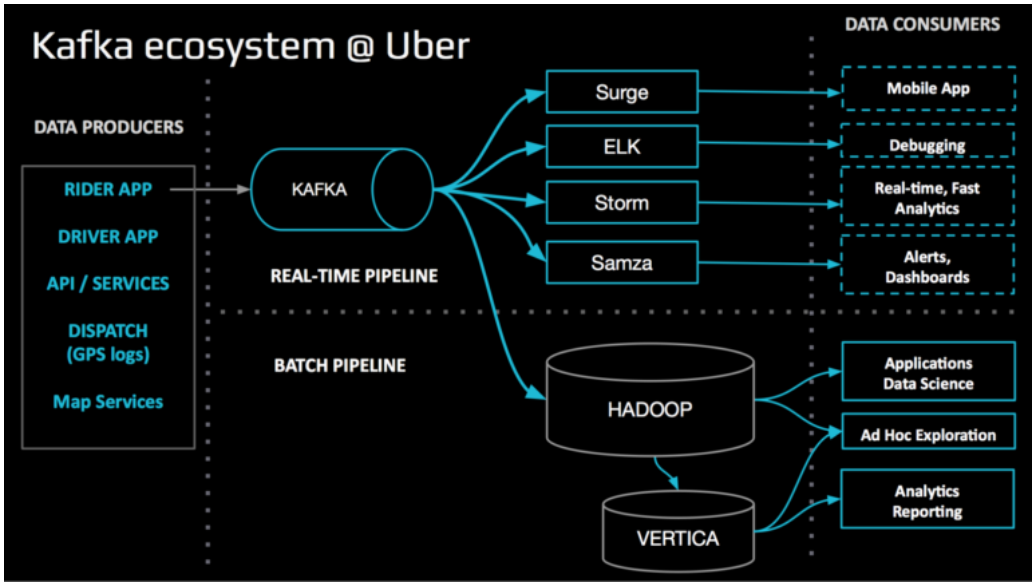
Impacts and Obstacles Behind Real-time Data Processing in Ridesharing Services

The wide use of smartphones and social media has made the adoption of ridesharing services in everyday life trivial. As the proportion of people living in cities and urban areas increases, the usage of on-demand ridesharing services is likely to increase as well (Simonetto, A. et al, 2019). Gambella (2020) explains that "The aim of the ridesharing service is to find an assignment of available vehicles to customers so as to optimize a performance indicator, such as vehicle travel times" (p. 2). This is done by processing real-time data from driver locations and user vehicle requests, and this real-time data processing requires a certain level of speed and processing power in order to remain effective (Stonebraker, M et al., 2005). This system needs to be maintained even as the use of ridesharing services and data needed to run them increases.

There is some uncertainty about the ability of ridesharinging services to scale and combat the natural challenges of real-time data processing. Uber is one of a few popular

ridesharing service providers that utilizes real-time data streaming in order to process and synchronize data within their platform's ecosystem. As seen in Figure 3, Uber primarily uses Apache Kafka as its core data streaming framework in order to have a high-performance, real-time data pipeline between data producers and consumers within this ecosystem. This style of data streaming and analytics is typical amongst ridesharing services and works efficiently with current demands, but there is uncertainty in whether or not these systems could scale if current demands were to dramatically increase. Kafka is extremely scalable and capable of providing data from producers to consumers in near real-time, but ridesharing services require complex and sometimes time consuming algorithms to carry out ride assignments (Simonetto, A. et al, 2019). This assignment algorithm acts as the bottleneck of the system which means that this algorithm should be the focus when it comes to future improvements.

Figure 3: Kafka Pipeline in Uber Ecosystem



This diagram shows how data produced by different services at Uber flows through a Kafka pipeline towards various data consumers (Soman, C, 2016, n.p.).

Ridesharing is a service that has increasing importance in urban environments. Many people use ridesharing as their main method of transportation within busy cities. As Gambella (2020) explains, “Ridesharing emerged in the past few decades amongst the shared-mobility services as a means to limit traffic congestion and achieve environmental benefits” (p.1). Failing to implement the real-time data processing efficiently in a system like this could create huge problems in a city that heavily relies on it. This would be an especially critical issue in an Internet of Things (IoT) based city which is a theoretical design that society is quickly approaching. The transportation system in an IoT based city would have a large amount of transportation actors that are constantly sending data and communicating with each other (Gambella, C. et al, 2020). There is also the risk that failure in these ridesharing systems causes users to lose trust and stop patronizing these services.

My research will take into account the multiple alternatives for and perspectives on the current and future state of high-performance data streaming and processing in ridesharing services in order to establish strong problem definition. I will also build on and respond to research that addresses the increasing use and development of ridesharing platforms that utilize real-time data streaming and processing to provide users a quick easy way to use the service. The real challenge for this research will be finding and evaluating valuable sources that discuss the future of data streaming in ridesharing because the uncertainty of this topic stems from future possibilities instead of present issues. The analytical system behind the current generation of ridesharing platforms has a high chance of being the basis for an IoT connected smart city, so it is critical these systems are

developed efficiently and responsibly (Gambella, C. et al, 2020). My STS research should provide a deep understanding of the impacts and requirements that data streaming imposes on ridesharing services and how this could affect urban societies and the people who live within them.

Project Deliverables and Significance

The deliverable for both my technical project and STS research shows the uses and compilations behind real-time data streaming. The deliverable of my technical work was a backend application that ran in parallel with a geospatial tracking system and provided real-time data processing to create alert notifications for specific geospatial events. The anticipated deliverable of my STS research is a deeper understanding of the benefits and drawbacks of the real-time data processing architecture in ride sharing services and how increased real-time data processing efficiency can affect the future of transportation in urban areas. My technical project resolved the issues of a missing alert functionality by allowing user-created queries to be applied to the live stream of location data. This was effectively the development of an efficient real-time data processing system. My STS research should provide a deeper understanding of the benefits and drawbacks that real-time data processing in ridesharing services and show why it is necessary to determine what effect these services could have on cities as well as the problems they could create as the scale of the system increases. Both deliverables attempt to provide solutions for the problems data streaming can cause so that systems utilizing real-time data streaming can operate at their full potential.

Word Count: 1971

References

Ariyaluran Habeeb, R. A., Nasaruddin, F., Gani, A., Targio Hashem, I. A., Ahmed, E., & Imran, M. (2019). Real-time big data processing for anomaly detection: A Survey. *International Journal of Information Management*, *45*, 289–307. https://doi.org/10.1016/j.ijinfomgt.2018.08.006

Baturina, O. (2019). Big Data Statistics: 40 Use Cases and Real-life Examples. https://www.scnsoft.com/blog/big-data-use-cases-stats-and-examples

Condon, S. (2018). By 2025, nearly 30 percent of data generated will be real-time, IDC says. ZDNet. Retrieved November 1, 2021, from https://www.zdnet.com/article/by-2025-nearly-30-percent-of-data-generated-will-be-real-time-idc-says/

Emrouznejad, A., & Marra, M. (2016). Big Data: Who, What and Where? Social, Cognitive and Journals Map of Big Data Publications with Focus on Optimization. In A. Emrouznejad (Ed.), Big Data Optimization: Recent Developments and Challenges (pp. 1–16). Springer International Publishing. https://doi.org/10.1007/978-3-319-30265-2_1

Gambella, C., Monteil, J., Dekusar, A., Barros, S. C., Simonetto, A., & Lassoued, Y. (2020). A city-scale IoT-enabled ridesharing platform. *Transportation Letters-the International Journal of Transportation Research*, *12*(10), 706–712. https://doi.org/10.1080/19427867.2019.1694206

Geomesa (n.d.). *Store, index, query, and transform spatio-temporal data at scale in HBase, Accumulo, Cassandra, Redis, Kafka and Spark*. Retrieved October 20, 2021, from https://www.geomesa.org/

Gu, K., Yang, L., & Yin, B. (2018). Location Data Record Privacy Protection Based on Differential Privacy Mechanism. *Information Technology and Control*, *47*(4), 639–654. https://doi.org/10.5755/j01.itc.47.4.19320

Hiraman, B. R., Viresh M., C., & Abhijeet C., K. (2018). A Study of Apache Kafka in Big Data Stream Processing. *2018 International Conference on Information , Communication, Engineering and Technology (ICICET)*, 1–3. https://doi.org/10.1109/ICICET.2018.8533771

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., & Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Pascalev, M. (2017). Privacy exchanges: Restoring consent in privacy self-management. *Ethics and Information Technology*, *19*(1), 39–48. https://doi.org/10.1007/s10676-016-9410-4

Safaei, A. A. (2017). Real-time processing of streaming big data. *Real-Time Systems*, *53*(1), 1–44. https://doi.org/10.1007/s11241-016-9257-0

Simonetto, A., Monteil, J., & Gambella, C. (2019). Real-time city-scale ridesharing via linear assignment problems. *Transportation Research Part C: Emerging Technologies*, *101*, 208–232. https://doi.org/10.1016/j.trc.2019.01.019

Soman, C. (2016, August 4). uReplicator: Uber Engineering's Robust Apache Kafka Replicator. Uber Engineering. https://eng.uber.com/ureplicator-apache-kafka-replicator/

Stonebraker, M., Çetintemel, U., & Zdonik, S. (2005). The 8 requirements of real-time stream processing. *ACM SIGMOD Record*, *34*(4), 42–47. https://doi.org/10.1145/1107499.1107504

Wang, S., Sinnott, R., & Nepal, S. (2017). Privacy-protected Place of Activity Mining on Big Location Data. In J. Y. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. BaezaYates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, & M. Toyoda (Eds.), *2017 Ieee International Conference on Big Data (big Data)* (pp. 1101–1108). Ieee. https://www.webofscience.com/wos/woscc/full-record/WOS:000428073701017