International Journal of Software Innovation
Volume 12 • Issue 1

A Novel Spatial Data Pipeline for Orchestrating Apache NiFi/MiNiFi

Chase D. Carthen, University of Nevada, Reno, USA*
https://orcid.org/0009-0006-7027-5212

Araam Zaremehrjardi, University of Nevada, Reno, USA

Vinh Le, University of Nevada, Reno, USA

Carlos Cardillo, University of Nevada, Reno, USA

Scotty Strachan, Nevada System of Higher Education, USA

Alireza Tavakkoli and Maketitle, University of Nevada, Reno, USA

Frederick C. Harris Jr., University of Nevada, Reno, USA
https://orcid.org/0000-0002-0857-6931

Sergiu M. Dascalu, University of Nevada, Reno, USA

ABSTRACT

In many smart city projects, a common choice to capture spatial information is the inclusion of lidar data, but this decision will often invoke severe growing pains within the existing infrastructure. In this article, the authors introduce a data pipeline that orchestrates Apache NiFi (NiFi), Apache MiNiFi (MiNiFi), and several other tools as an automated solution to relay and archive lidar data captured by deployed edge devices. The lidar sensors utilized within this workflow are Velodyne Ultra Puck sensors that produce 6-7 GB packet capture (PCAP) files per hour. By both compressing the file after capturing it and compressing the file in real-time; it was discovered that GZIP and XZ both saved considerable file size being from 2-5 GB, 5 minutes in transmission time, and considerable CPU time. To evaluate the capabilities of the system design, the features of this data pipeline were compared against existing third-party services, Globus and RSync.

KEYWORDS

Big Data, Data Pipeline, Data Transfer, Edge Computing, IOT, Lidar, MiNiFi, NiFi, PCAP, Smart City, Spatial Data

INTRODUCTION

As cities begin employing more and more complex sensing devices to either conduct traffic analysis or provide a measure of infrastructure, creating a system for data transferal becomes a crucial challenge. For smart city projects, spatial information such as Light Detection and Ranging (lidar) is especially a concern. Due to the massive amount of data generated by lidar point clouds, data

DOI: 10.4018/IJSI.333164

*Corresponding Author

collection and transferal from edge device to central repository tends to suffer from bottle-necking issues, such as low throughput networking, high latency, and packet-loss. These constraints must be considered as most cities in the United States may have difficulty placing fiber optic infrastructure in their cities (Cooper, 2022).

As part of ongoing smart city developments in the city of Reno, Nevada, the work presented within this paper involves a 100 mbps fiber network provided by the city of Reno. While this network was deployed to specifically address the cyber-infrastructure needs within the city of Reno, this called for the development of a Software Data Pipeline (SDP) that could enable reliable data transformation, transferal, and logging between edge computers and the fog computing network.

In this paper, the authors developed an SDP that leverages NiFi/MiNiFi to facilitate the movement of lidar data generated at various edge computing locations placed around the city of Reno, specifically the Virginia Street corridor. This data is relayed to the fog computing network located at the University of Nevada, Reno (UNR), which is then finally piped towards its destination, UNR's Pronghorn High Performance Computing Cluster, for archival storage. The software on the edge environments use Docker Compose with MiNiFi to hook into the NiFi-based data pipeline in which the lidar point-clouds are compressed and then transmitted off. The software within the UNR Data Center uses Kubernetes to scale up NiFi hosts and receive the lidar point clouds, which are then processed for storage. To ease any confusion, the name "UNR-Virginia SDP" was chosen as the colloquial name to refer to the SDP approach presented in this paper.

The UNR-Virginia SDP does offer some insights for those interested in establishing a scalable pipeline for spatial data collection within smart city infrastructure (Duygan et al., 2022). With the increasing interest in smart city development, the UNR-Virginia SDP provides a template so that other cities with similar network infrastructure may easily incorporate lidar data collection as part of their normal workflow. Due to the versatility of lidar data, lidar collection presents more opportunities for cities to better utilize big data methodologies for effective planning or the establishment of new data-driven solutions (McCrae & Zakhor, 2020; Zhao et al., 2019).

As a form of evaluation for the UNR-Virginia SDP, the authors conducted an analysis of different compression algorithms, compared the discussed approach with the present network bandwidth, and finally performed a feature comparison with major established third-party services, RSync (Davison, 2023) and Globus (Foster et al., 2012). Furthermore, additional metrics gathered from the UNR-Virginia SDP were recorded, such as the bandwidth usage, resource usage on edge devices, and recording time for message transfer. To elaborate, this involved testing different compression methods in terms of resource usage, average CPU usage, average memory usage, total duration time, and size of messages. As part of the feature comparison, RSync and Globus were compared against the UNR-Virginia SDP for basic functionality of data transmission and receiving, load balancing, parallel streaming support, the customization of data flow, and file verification.

The remainder of this paper is structured as follows: the first section presents background information of the technologies explored and used by the UNR-Virginia SDP, the second section describes the design of the UNR-Virginia SDP with considerations and expected requirements of the data pipeline, the third section details the resulting implementation of the planned design and data flow, the fourth section presents the overall performance evaluation of the software data pipeline with benchmarks and comparisons of other methods, and the final section discusses possible uses of the data pipeline and outlines future work to extend its functionality.

BACKGROUND AND RELATED WORKS

Data Pipeline Approaches

With the increasing interest in both cloud and fog computing, different approaches have been explored to facilitate streams of data that require high throughput, intense bandwidth usage, and consistent access across different network scenarios. In general, the very nature of these data streams presents