Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

TOWARDS AN ONLINE DATA ANALYSIS ARCHITECTURE FOR LARGE-SCALE DISTRIBUTED SIMULATIONS

Xiaorui Du

Department of Informatics — Intelligent Cloud Technologies Laboratory
Technical University of Munich — Huawei Munich Research Center
Boltzmannstr. 3 — Riesstrasse 25
Garching, 85748, GERMANY — Munich, 80992, GERMANY

ABSTRACT

Online data analysis plays an important role in simulations. However, with the rise of distributed and large-scale simulations, designing an efficient online data analysis architecture is particularly challenging, since it requires efficiently retrieving and processing the massive data produced by the distributed simulation. Many of existing solutions use high performance computing resource or big data platforms to build online data analysis systems. However, none of them can be applied to distributed simulations. In our work, we propose a novel online data analysis architecture based on the concept of Modelling and Simulation as a Service (MSaaS) with the goal of supporting efficient data analysis in large scale distributed simulations.

1 INTRODUCTION AND MOTIVATION

Online data analysis plays an important role in simulations: traffic authorities use it to test hypothesis in real-time and support decision making, developers use it for monitoring the current status of simulation nodes or for troubleshooting, and traffic engineers use it for external control of the simulation. However, with the rise of distributed and large-scale simulations, we are facing great challenges: 1) How to efficiently retrieve data from the distributed simulation? 2) How to efficiently process massive data online? 3) How to hide low-level details to allow users to use data analysis systems transparently in distributed simulations?

There are many research works describing solutions for efficient online data analysis in large scale simulations. For example, Zehe et al. (2016) present a cloud-based online data analysis system that aims to improve the efficiency of data processing by employing high performance computing resources. Amini et al. (2017) aim to solve the challenges of fault tolerance and low latency of data analysis by applying Kafka and HDFS in their online data analysis architecture. However, they are all built on shared memory simulations and lack the applicability to distributed simulations. Therefore, we propose an novel online data analysis architecture based on the concept of Modeling and Simulation as a Service (MSaaS) with the goal of supporting efficient data analysis in large scale distributed simulations.

2 ORIGINAL CONTRIBUTION

As illustrated in Figure 1, the proposed architecture consists of two components: the **Front-End** and the **Back-End**,where the **Back-End** includes the **Proxy** and the **Simulation** sub-components. The **Front-End** is used by the users to interact with the simulation, including submitting queries, controlling simulation, and subscribing to the interested topics, and it's also the place to receive the results. The **Proxy**, as a bridge between the users and the distributed simulation, is the core component in our online data analysis architecture. It consists of four parts including **Optimizer**, **Query Executor**, **Data Collector** and **Publisher**.

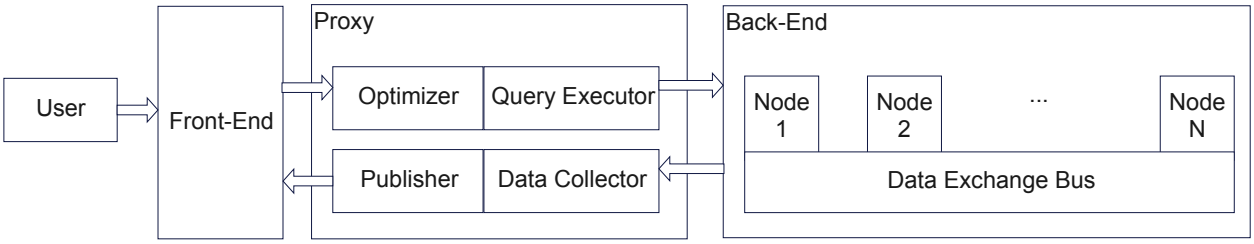

Figure 1: Overall architecture.

- The **Optimizer** optimizes queries, for instance, by exploiting common sub-queries. Common sub-queries only have to be evaluated once and their results can be re-used, saving bandwidth and computational resources, thus improving the performance of data analysis.
- The **Query Executor** is used to rewrite queries and issues them to simulation nodes accordingly. Generally, this is done in the following steps: the **Query Executor** (1) parses a query and obtains a list of simulation nodes involved in that query based on a predefined partition map, (2) rewrites the query to multiple sub-queries (one for each node involved) and issues them to the corresponding simulation nodes.
- The **Simulation** executes a single simulation on multiple nodes. Each node generates the required data according to the issued queries, and all data is sent to the **Data Collector**.
- The **Data Collector** is used to gather and aggregate the data from the distributed nodes. The state machine is applied in this module for data processing synchronization. Every time the **Query Executor** issue a new query to the **Simulation**, a corresponding table and state machine are registered in the **Data Collector**. The table is used to collect data continuously and the state machine is used to trigger collection completion. Once all data required for a query is received, the corresponding computation is executed. Additionally, incremental computation is also allowed in the **Data Collector**.
- The **Publisher** rewrites the result from the **Data Collector** into Google Protocol Buffers (Protobuf) and publishes it. Protobuf provides a rich set of serialization and deserialization methods which can facilitate the data exchange between components in our architecture.

3 ONGOING AND FUTURE WORK

Currently, the proposed system is under development. We already implemented the **Data Collector** which collects the data from nodes and allow the retrieval of the traffic metrics both in a streaming and batch mode online (e.g. computing the average speed of a corridor.

Future work will be focused on refining the functionality of each component including: 1) defining a uniform API for the Front-End, 2) defining output formats for different query types, and 3) providing more optimization strategies for the optimizer module. We believe that this architecture will allow users to obtain and analyse data from large scale distributed simulations more efficiently and easily.

REFERENCES

Amini, S., I. Gerostathopoulos, and C. Prehofer. 2017. "Big Data Analytics Architecture for Real-time Traffic Control". In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems*. June 26th-28th, Napoli, Italy, 710–715.

Zehe, D., V. Viswanathan, W. Cai, and A. Knoll. 2016. "Online Data Extraction for Large-scale Agent-based Simulations". In *Proceedings of the 2016 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. May 15th-18th, Banff, AB, Canada, 69–78.