SCAPE: An Application Platform for Environmental Big Data Analysis in Smart Cities

Rundong Fan and Gang Wu
Shanghai Jiaotong University, 800 Dongchuan Rd, Shanghai 200240, China

Abstract—**Environment protection is a crucial part of a smart city. Environmental big data analysis is an important way to improve environment modeling and city planning, which plays an important role in environment protection. Applications for environmental big data analysis involves collecting, storing, processing, modeling and visualizing a huge amount of highly heterogeneous data, so a powerful application platform is needed to support them. In this paper, we propose an application platform for environmental big data analysis in smart cities, named SCAPE. Its design combines dataflow management and stream processing, big data platform and visualization engine. We evaluated the SCAPE platform by introducing a practical application about air pollution modeling. The case study of the practical application shows the SCAPE platform's capabilities to support environmental big data analysis applications.**

Keywords-big data platform; smart city; environment protection; environment modeling

I. INTRODUCTION

Nowadays urbanization is progressing very fast all around the world [1], with more and more modern cities facing great challenges including dense population, management of complex city assets and environment protection. As a result, people find it increasingly hard to manage modern cities in traditional governance models. On the other hand, the revolution of information technology enables the concept of "smart cities" [2]. A smart city is a rapidly evolving vision of urban development to integrate various information technology and infrastructure to manage city resources and provide public services to citizens.

Smart cities promise sustainability and improved quality of life, therefore, environment protection is an important part of a smart city. In the field of environment protection, in order to store dynamic environment status for future analysis and allow for real time responses to environmental challenges, a smart city usually employs a various of Internet of Things (IoT) solutions and information and communication technology (ICT) to create a network of environmental sensors to monitor multiple environment-related indicators such as pollution index, weather status, wind status, road traffic and the running status of counter-pollution equipment. It also needs to get data from other information services, for example, weather forecast and hospital admission rate. A smart city is then supposed to do several calculations and modelings based on the huge collected dataset to get enough knowledge about the city's environment status and trend, which is often powered by big data platforms. Finally, the knowledge is fed back to the city's environment by various ways including visualizing to the policy-makers, alerting related administrative agencies and broadcasting suggestions to citizens.

In the scenario above, to cope with challenges of environment protection, a smart city needs to have the capability of receiving, storing, processing and modeling a huge amount of heterogeneous data, which might be real-time or historical, from different IoT solutions and information systems. In addition, a smart city should be able to present the analyzed result to related policy-makers, government agencies, companies and citizens in order to guide policy-making, city-planning, business activities and daily life.

Building up a reliable, efficient and flexible application platform for environmental big data analysis is the key to the success of environment protection in smart cities. This paper introduces a smart city application platform for environment, namely SCAPE, which fully meets the demand of environmental big data analysis in smart cities. The SCAPE platform is specifically designed to cope with the challenges of environmental issues in smart cities, other related work and similar systems are discussed in Section II. The architecture of the SCAPE platform consists of dataflow management and stream processing, big data platform and visualization engine, which will be described in Section III. A practical application case about environment modeling is discussed in Section IV to prove and demonstrate the capabilities of the SCAPE platform. Then we discusses and analyses the features and capabilities of the platform in Section V. Finally, Section VI presents the conclusions and possible future work.

II. RELATED WORK

Smart city is a promising concept, so there're a lot of efforts about construction of mobile sensors network [3], deploying testbed simulators [4], building smart city ecosystems [5] and developing specific services in smart cities [6]. However, in this paper, the SCAPE platform focuses more on building a platform to support applications of environmental big data analysis in smart cities.

Still, there're several studies about architectures of big data platforms for smart cities. Some of them are commercial products and the design details are not open, for example, IBM's Smarter City Solutions on Cloud [7] and Microsoft's CityNext [8]. There're also some ambitious gigantic framework and testbed, one of them is SmartSantander [9], which is a framework implemented on a real-world city testbed, and is continually evolving in recent years [10]. The

SmartSantander project developed a similar data platform called CiDAP [11] which is a general data platform that does not focus on dataflow management. There’re also approaches that focuses on dataflow [12], however its architecture is highly abstract and cannot be used to support actual applications.

As a comparison, the SCAPE platform is built for specific need of applications for environmental big data analysis in smart cities. It covers not only the topic of general data collection and processing, it aims at solving the common issues in smart cities that applications of environmental big data analysis share, which include heterogeneous data compatibility, multiple data source support and flexible real-time visualization of heterogeneous data. To summarize, the SCAPE platform could be a valuable prototype for real-world application platform for environmental big data analysis in smart cities.

III. Platform Design

A. Overview

The overall architecture of the SCAPE platform is shown as Figure 1. The main body of the SCAPE platform is a combination of open source big data products and interface applications.

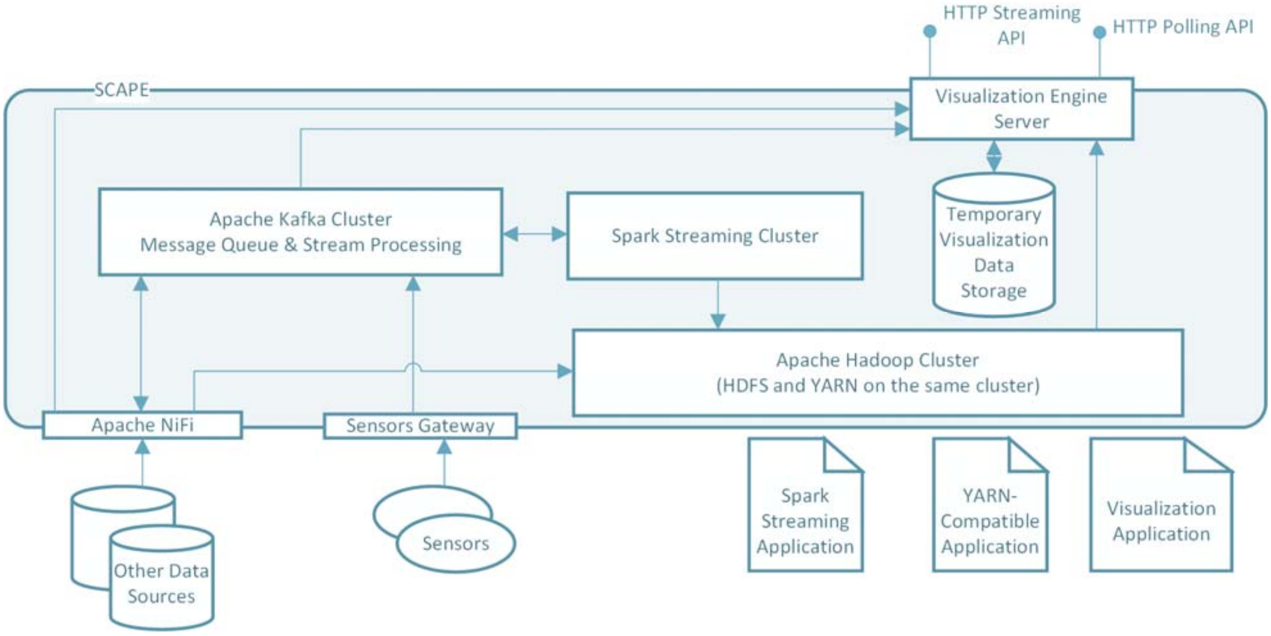
HTTP Streaming API
HTTP Polling API
SCAPE
Visualization Engine Server
Apache Kafka Cluster
Message Queue & Stream Processing
Spark Streaming Cluster
Temporary Visualization Data Storage
Apache Hadoop Cluster
(HDFS and YARN on the same cluster)
Apache NiFi
Sensors Gateway
Other Data Sources
Sensors
Spark Streaming Application
YARN-Compatible Application
Visualization Application

FIGURE I. SYSTEM ARCHITECTURE OF THE SCAPE PLATFORM

There’re 3 major parts in this architecture: dataflow management and stream processing, big data platform and visualization engine.

The first part is the dataflow management and stream processing, which contains the Apache NiFi, the sensors gateway application, the Apache Kafka cluster and the Spark streaming cluster. These components manage and collect the dataflow from various data sources including external databases, local plaintext files, third-party HTTP APIs and real-time data from sensors. The dataflow is then classified, processed and sent to pre-configured destination including Kafka message queues, Spark streaming applications, HDFS storages and the visualization engine server.

The second part is the big data platform. It consists of Apache Hadoop clusters, including co-locating HDFS and YARN clusters. YARN-compatible applications such as MapReduce and Spark applications run here, loading from and saving back to the HDFS cluster during the process.

The third part is the visualization engine, its main components are the visualization engine server and the temporary visualization data storage, which is a single-node MongoDB database currently. The visualization engine server receives certain real-time updates from the Kafka cluster through subscribing topics. And it also receives certain HTTP requests with data from NiFi, loads data from HDFS eventually and uses the temporary storage to load and save status data that are related to real-time visualization. The server also provides HTTP streaming API and polling API for visualization applications, which are usually web and mobile applications.

There’re also some supporting infrastructure components not mentioned in Figure 1, including an Apache Zookeeper cluster that provides distributed services to several big data products, and a server with Apache Ambari for rapid deployment of Hadoop environment.

B. Dataflow Management and Stream Processing

The dataflow management and stream processing part is specifically designed for environment related big data applications. Environment related sensors speak different protocols. And then they are translated by the sensors gateway application. JavaScript object notation, namely JSON is used as the intermediate format as shown in Figure 2. The translation code needs to be manually coded in sensors gateway application, and different sensors might connect to different sensors gateway applications, which leaves room for online extension.

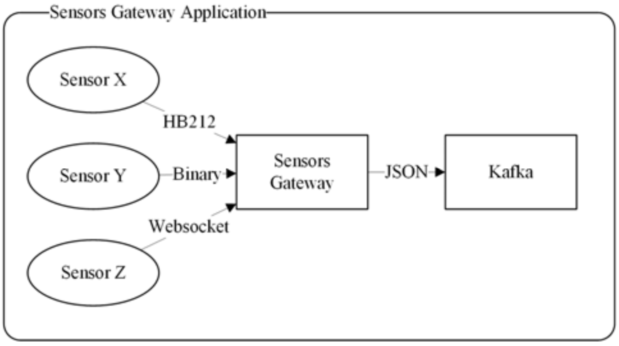

FIGURE II. SENSORS GATEWAY

Other data sources including external relational databases like MySQL, third-party HTTP APIs and files are managed by NiFi. Through its web GUI, multiple data sources can be added, and the direction of dataflow can be pre-configured, while it also supports online reconfiguration. For example, we can configure the NiFi on the SCAPE platform to load hospital admission rate from an external MySQL database and request third-party HTTP APIs to retrieve weather forecast and then save them to HDFS for archiving and further processing.

Kafka serves as a de-coupler and data buffer on the SCAPE platform. Data from different sources are published to different Kafka topics to get processed. While Kafka has several stream processing features, the SCAPE platform mainly utilizes the Spark streaming cluster to do the stream processing. The Spark streaming cluster pulls data from Kafka and then processes data and saves the result to HDFS, and most sensors data got processed in this way. In some cases the result is sent back to Kafka, in order to enter next stream processing phase, it can also be used to notify the visualization engine server when some important visualization related states changed.

C. *Big Data Platform*

Big data platform is mainly a co-locating HDFS and YARN cluster, which serves as the heavyweight processing unit of the SCAPE platform. Almost all of the data that needs to be persisted is stored to the HDFS cluster for archiving and future analyzing. The YARN cluster is capable of running MapReduce and Spark applications, which is a solid solution to mainstream big data analysis requirements.

D. *Visualization Engine*

Visualization engine is an important part on the SCAPE platform, and it is designed to be flexible and real-time. The visualization engine server itself is an application server which is subscribing notification topics from Kafka for real-time data updating and loads data from the HDFS cluster eventually. It's also a web server, and it receives HTTP requests from NiFi for real-time data updates. It has an HTTP based streaming API and normal polling HTTP API to provide visualization related data to external visualization applications.

Specifically, visualization related data are pre-configured and mapped to HDFS files, Kafka topics, and NiFi requests as shown in Figure 3. The server uses the streaming API to push data that comes from certain Kafka topics or NiFi requests, and serves HDFS data using the polling API.

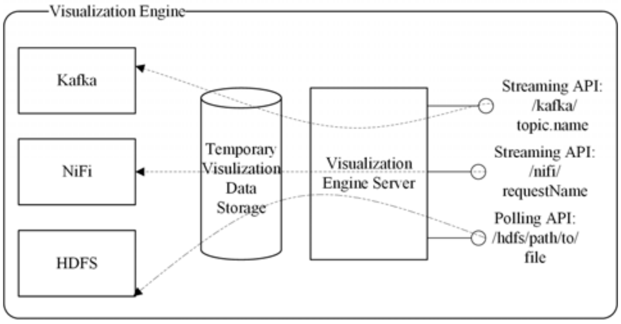
Visualization Engine
Kafka
NiFi
HDFS
Temporary Visulization Data Storage
Visualization Engine Server
Streaming API: /kafka/ topic.name
Streaming API: /nifi/ requestName
Polling API: /hdfs/path/to/ file

FIGURE III. VISUALIZATION ENGINE

IV. APPLICATION CASE STUDY

The SCAPE platform is aimed at providing a solid solution to support applications for environmental big data analysis in smart cities. We'll describe a practical application example about environment modeling and its interaction with the SCAPE platform to show the platform's power and capabilities.

Air pollution can be a serious problem, especially for citizens who is suffering from allergic reactions. Under smart city conditions, we could collect air pollution data and use the spatial-temporal approach to model air pollution data and predict spikes of heavy air pollution [13], so concerned citizens can be alerted prior to the pollution and take measures to lower the risk of allergy. Application like this could be a practical example for the SCAPE platform, and we'll use the name Allergy Guard to indicate this application in the following part of this paper.

In order for Allergy Guard to function, the history data of a city's air pollution in different areas must be preserved. On the SCAPE platform, we could create a Spark streaming application to do this job. In this application, we could get sensors' data from Kafka and then save sensors' data into HDFS and organize them by date, time and area. Another thing is to get other related data for better model training, for example, weather data could be useful because it has a huge impact on air pollution status. On the SCAPE platform, we could configure NiFi to get weather information through third-party HTTP APIs and then save it to HDFS. After preparing the data, we could start a scheduled Spark job on the YARN cluster to train with the history data and produce next hour's prediction. The result will be stored on HDFS, and mapped to the visualization engine server's API endpoint. Finally, with proper web or mobile application that can request the visualization engine server and get the prediction data, citizens could get the recent forecast and take corresponding measures.

Case study shows that practical applications like Allergy Guard can be implemented in an organized way on the SCAPE platform. Although features like actual visualization need to be implemented at the application side.

V. Discussion and Analysis

Environment related big data applications in smart cities poses a lot of challenges. Its heterogeneous sensor data, diverse external data sources, flexible real-time visualization and the huge amount of data generated everyday are major difficulties that needs to be overcome.

The SCAPE platform uses sensors gateway applications to translate the sensors' data into a universal intermediate format and then sends them to Kafka to fully de-couple the sensor layer. This method is very flexible, however it is hard for maintenance when the number of sensor protocols rises. A configuration-based solution can be a better one, but the flexibility would drop drastically.

The SCAPE platform also uses NiFi to manage various external data sources, NiFi is a powerful and well-supported project, and it showed the strength to manage the dataflow between various data sources and data processors. However, its cluster support is not mature enough.

To support real-time data, the SCAPE platform uses Spark streaming to do real-time processing, and provides streaming APIs to support real-time visualization. It's powerful but very heavyweight. As for visualization, the SCAPE platform again uses a very primitive way, which is only providing visualization related data, but not doing the actual visualization. The flexibility is very high, but it leaves too much work for visualization applications to do, and this part needs further improvement.

The SCAPE platform relies on mature big data products such as Hadoop and Spark to ensure efficiency and reliability while handling big data.

Application case study shows that the SCAPE platform is capable of supporting applications for environmental big data analysis in smart cities. However, further practice is needed to fully determine the platform's capabilities and current solution is lack of visualization features.

To summarize, the SCAPE platform is a viable solution to supporting environment related applications in smart cities, although it needs improvement in simplifying its infrastructure and adding more features.

VI. Conclusion and Future Work

We discussed the architecture design of the SCAPE platform in this paper, and then demonstrated and analyzed its capability of supporting applications for environmental big data analysis. The design and the demonstration in the laboratory environment of the SCAPE platform showed its value on supporting environment related big data applications in smart cities.

Our future study will be focused on the real-world deployment of such an application platform, the real-world applications can be different, and there're far more issues including high availability, ecosystems, deployment, operations and security to be solved in real-world scenarios. We will seek to further improve the SCAPE platform by gathering more production experience.

References

[1] Montgomery, M. (2007). United Nations Population Fund: State of World Population 2007: Unleashing the Potential of Urban Growth. Population and Development Review, 33(3), 639-641.

[2] Hollands, R. G. (2008). Will the real smart city please stand up? Intelligent, progressive or entrepreneurial?. City, 12(3), 303-320.

[3] Lee, U., Zhou, B., Gerla, M., Magistretti, E., Bellavista, P., & Corradi, A. (2006). Mobeyes: smart mobs for urban monitoring with a vehicular sensor network. IEEE Wireless Communications, 13(5), 52-57.

[4] Lugaric, L., Krajcar, S., & Simic, Z. (2010, October). Smart city—Platform for emergent phenomena power system testbed simulator. In 2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe) (pp. 1-7). IEEE.

[5] Schaffers, H., Komninos, N., Pallot, M., Trousse, B., Nilsson, M., & Oliveira, A. (2011, May). Smart cities and the future internet: Towards cooperation frameworks for open innovation. In The Future Internet Assembly (pp. 431-446). Springer Berlin Heidelberg.

[6] Chinrungrueng, J., Sunantachaikul, U., & Triamlumlerd, S. (2007, January). Smart parking: An application of optical wireless sensor network. In Applications and the Internet Workshops, 2007. SAINT Workshops 2007. International Symposium on (pp. 66-66). IEEE.

[7] Naccarati, F., & Hobson, S. (2014). IBM Smarter City Solutions on Cloud.

[8] Martin, K. (2014). CityNext and Sitecore getting onto Australian councils' agenda. Government News, 34(5), 35.

[9] Sanchez, L., Galache, J. A., Gutierrez, V., Hernandez, J. M., Bernat, J., Gluhak, A., & Garcia, T. (2011, June). Smartsantander: The meeting point between future internet research and experimentation and the smart cities. In Future Network & Mobile Summit (FutureNetw), 2011 (pp. 1-8). IEEE.

[10] Sanchez, L., Muñoz, L., Galache, J. A., Sotres, P., Santana, J. R., Gutierrez, V., ... & Pfisterer, D. (2014). SmartSantander: IoT experimentation over a smart city testbed. Computer Networks, 61, 217-238.

[11] Cheng, B., Longo, S., Cirillo, F., Bauer, M., & Kovacs, E. (2015, June). Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander. In 2015 IEEE International Congress on Big Data (pp. 592-599). IEEE.

[12] Vilajosana, I., Llosa, J., Martinez, B., Domingo-Prieto, M., Angles, A., & Vilajosana, X. (2013). Bootstrapping smart cities through a self-sustainable model based on big data flows. IEEE Communications Magazine, 51(6), 128-134.

[13] Russo, A., & Soares, A. O. (2014). Hybrid model for urban air pollution forecasting: A stochastic spatio-temporal approach. Mathematical Geosciences, 46(1), 75-93.