An evolutionary service solution for spatio-temporal data analysis in highway domain

Jie Zhou and Weilong Ding*

School of Information Science and Technology,
North China University of Technology,
No. 5 Jinyuanzhuang Road, Beijing, 100144, China
and
Beijing Key Laboratory on Integration and
Analysis of Large-scale Stream Data,
No. 5 Jinyuanzhuang Road, Beijing, 100144, China
Email: 13656429823@163.com
Email: dingweilong@ncut.edu.cn
*Corresponding author

Abstract: In highway domain, many routine analyses are required on business spatio-temporal data to monitor and control the traffic situation in time. Such analytics as big data applications through traditional ways remain inherent challenges due to the inflexibility of the holistic procedure, the variety of the computational jobs and the diversity of the customised visualisation. In this paper, we propose a service solution in a specific highway domain for business technicians to build their own data analytical applications conveniently and rapidly. On massive toll data through respective services, our solution provides evolutionary capacity to load, process and reveal spatio-temporal data for building comprehensive data analysis. In a practical project, our work proves the feasibility and advantages by exhaustive case studies.

Keywords: highway; data analysis; evolutionary solution; spatio-temporal data.

Reference to this paper should be made as follows: Zhou, J. and Ding, W. (2019) 'An evolutionary service solution for spatio-temporal data analysis in highway domain', *Int. J. Intelligent Internet of Things Computing*, Vol. 1, No. 1, pp.43–52.

Biographical notes: Jie Zhou is a postgraduate student at the School of Information Science and Technology, North China University of Technology, Beijing, China. Her research interests are intelligent transportation system and service computing.

Weilong Ding is an Associate Professor at the North China University of Technology, Beijing, China. He obtained his PhD from the Institute of Computing Technology, Chinese Academy of Sciences. He focuses on the real-time data processing, distributed system and service computing. He has published more than 50 academic articles in journals and conferences. He owns five invention patents in related fields.

1 Introduction

With the growth of inter-city traffic, congestion in highway has become one of most serious problems worldwide, and the highway transportation systems are significant to govern urban situations (Holguín-Veras and Preziosi, 2011). With the popularity of electronic toll collection (ETC) in recent decades, the passing efficiency spurs the burst of data generated from sensors at toll stations (Kim et al., 2008). For example, in Henan Province of China, the records of toll data would be around one million a day and more than 0.2 billion in the year 2016. Such toll data is typical spatio-temporal, because a record would imply when a vehicle passing through a specific location (entry or exit station in the highway network). Compared with other sensory data in highway domain, it has the advantages of finite coverage, exact locality and higher quality. Traditionally, after necessary pre-processing (Ding and Cao, 2016), the toll data would be analysed as analytical applications through statistic models on small samples (Ghesmoune et al., 2016) or through individual big data jobs on massive data (Ding et al., 2017). Many analytics are required progressively with the development of domain business and technology, such as the multi-perspective traffic volume analyses of various types of vehicles.

However, it faces inherent limitations for continual development and operation when business requirements are mutable in practical condition. Firstly, the analytical procedure is holistic and too complex to update even if only small parts change. It is normal to delay the release for the regular statistical reports in metropolises (Hochreiner et al., 2016). Secondly, each data analysis is developed separately, which is tedious and error-prone. In fact, in highway domain, analytical jobs have the universality in the processing fashion due to the spatio-temporal characteristics of toll data. Thirdly, the customised visualisation is required for the business management to get high-level result presentation. Interactive and synthesised perspectives are not always easy to achieve (Luo et al., 2016).

Accordingly, the government officials want to find effective ways for the data analysis. In this paper, a service solution is proposed in highway domain to achieve the analyses on massive toll data. Through the services, it provides evolutionary capacity in full procedure and demonstrates the advantages in practice.

2 Preliminary

2.1 Motivation and related work

Our work was initiated by *highway big data analysis system in Henan Province*. It is expected to improve the routine business analysis through big data technologies for public travelling and government management. Operated by Henan Transport Department since October 2017, the toll data from toll stations is kept. The record of such data has the structure in Table 1, which is typical spatio-temporal and contains 12 attributes including six entity attributes, two temporal attributes and four spatial attributes.

The traditional holistic procedure of data analysis in highway domain makes it inconvenient to update analytical jobs or the executive environment. Generally, those jobs are executed in a certain period (e.g., once a month) and on the same input (e.g., last

monthly data). We found that the procedure can be divided into loosely coupled common steps due to the jobs' universality. It can be depicted as general services with configurable settings in respective steps, which is just our motivation.

Table 1 The structure of toll data

Attribute	*Notation*	*Type*
collector_id	Toll collector identity	Entity
vehicle_license	Vehicle identity	
vehicle_type	Vehicle type	
card_id	Vehicle passing card identity	
etc_id	Vehicle ETC card identity	
etc_cpu_id	ETC card chip identity	
entry_time	Vehicle entry timestamp	Time
exit_time	Vehicle exit timestamp	
entry_station	Identity of entry station	Space
entry_lane	Lane number of entry station	
exit_station	Identity of exit station	
exit_lane	Lane number of exit station	

For the data analytics in transportation fields, there has been many works to handle massive data with big data technologies (Zheng et al., 2014). It is a hot topic nowadays to build high-level applications by reusing services (Du et al., 2012), but it still faces challenges in domain-specific solution. As uniform abstractions, the blueprint template (Nguyen et al., 2011) and (OASIS topology and orchestration specification for cloud applications) TOSCA standard are proposed formally for developers as an easy guide to design, configure and deploy cloud services. But they are all conceptual models for service level agreement (SLA) restrictions without the design details of services. Moreover, some frameworks are defined to reduce the development complexity for service solution in cloud environment (Houssaini et al., 2015), manufacturing fields (Giret et al., 2016), and visualisation (Zhou et al., 2016). However, neither of them concerns the spatio-temporal feature and specific analytical requirements in highway domain.

2.2 Methodology

We designed the system architecture with multiple layers as Table 1. The data layer maintains the *basic data* in relational database and the *toll data* in No-SQL database. The former includes the profile of station, section and highway line; the latter is the business data generated continuously. As a dedicated platform as a service (PaaS), the processing layer is the big data computation environment, in which jobs like *daily traffic flow of ETC vehicle*, and *monthly vehicle type proportion* are submitted for routine business analyses. The application layer shows job results as plots in multiple perspectives (e.g., time, space and vehicle). As a dedicated infrastructure as a service (IaaS), the infrastructure layer supplies the system with the virtual resources (i.e., computation, storage and network) from a private cloud.

We reconstruct the domain analytical procedure as Figure 2. In processing layer, the executive jobs contain the parameters of different granularity in similar dimensions like time (e.g., five minutes, 15 minutes, 1 hourly, …), space (e.g., station, section, line, …) or statistic object (e.g., traffic flow, predictive trends, …). The procedure works as typical 'IPO' fashion, and each step exploits a dedicated service to complete its duty. In the input step, an ETL service loads toll data from production database to the data layer (Zhang et al., 2018). In the processing step, the jobs of the processing layer can be developed rapidly through service template, and their results would be write into relational or No-SQL storage. In the output step, the visualisation service presents the jobs' results as interactive plots in the application layer.

Figure 1 System architecture (see online version for colours)

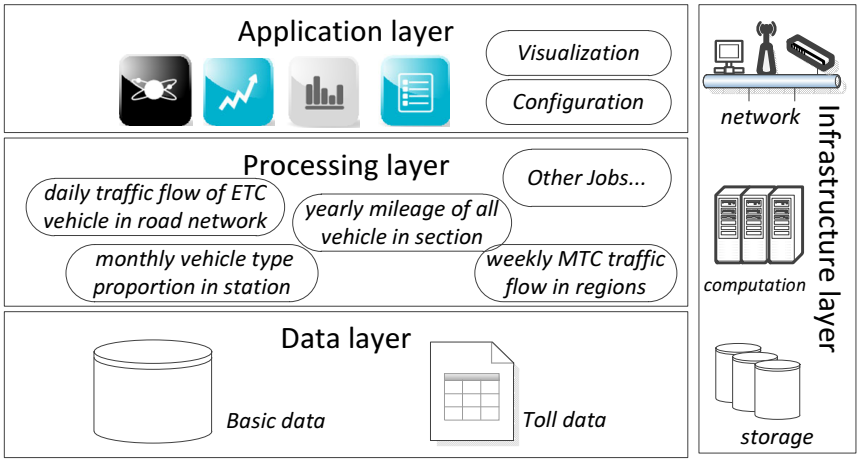
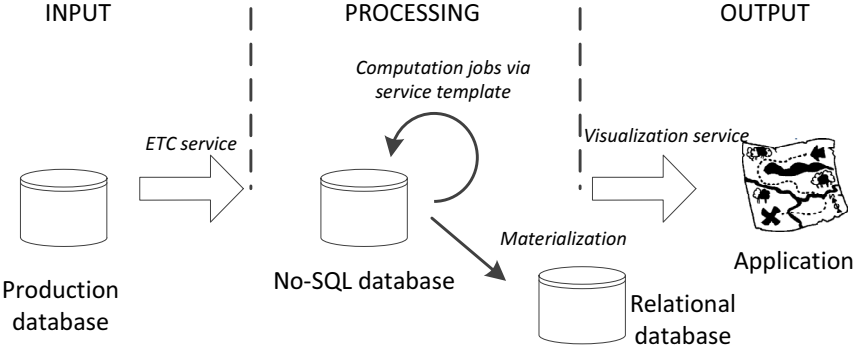


Figure 2 Data analysis procedure



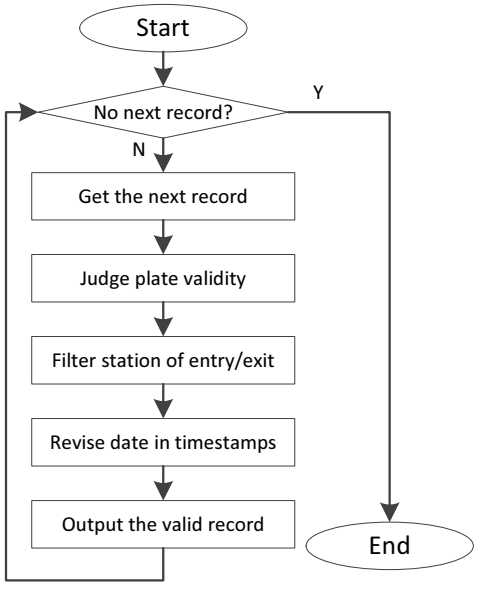
The services in each step would be discussed in detail next.

3 Evolutionary data analysis service

3.1 ETL service with data cleaning

In the input step, an ETL service is designed as a Hadoop MapReduce job to extract original data from the outer storage to a distributed No-SQL database, after necessary operations on data. Before loading data to the target, the service transforms the records to the valid ones as the process in Figure 3. When getting a record, the service would examine the entity, temporal and spatial attributes. Like the operations in Xia et al. (2019), for the any vehicle plate, it is checked by predefined regular expressions; for either entry or exit station, it would be examined its exsistence in the basic data. The ones with invalid plate or unknown stations are abandoned directly without further processing. For the timestamp of either entry or exit, the inconsistent date would be revised by referring to the other one in the same record. For example, the date 1990-01-01 in either timestamp of a record is firmly invalid if the other is 2017-12-28, and could be corrected accordingly. If both timestamps are illegal, the record would be discarded then.

Figure 3 Data cleaning in ETL service



The ETL service is evolutionary: when executive condition or constrains change, the business logic here could be updated independently with other jobs and other steps in the same job. As the first step of the analysis procedure, it is executed only when new original data is ready, and opaque to other steps.

3.2 Service template

In the processing step, the domain jobs are executed on the cleaned data from the No-SQL database. The service template (Lehrig et al., 2018) is proposed to avoid the repetitive development for the jobs with the common business logics. According to our extensive study, the statistical or predictive jobs in highway domain can be described as the composition of five dimensions: time granularity, space granularity, vehicle type, statistic object and ordering style.

Therefore, the service template is modelled as an abstract job of Hadoop MapReduce, and each dimension above becomes an individual parameter in it. The assignment of any parameter is listed as Table 2 including the default values. The common algorithms of the specific statistic object in domain have been implemented in the service template. The concrete domain job is the parameterisation of this template and can be developed rapidly just by inheritance and extension from the template. The details can be found in our work (Ding et al., in press). It is feasible for most of domain jobs of routine data analysis. For example, a job to count daily traffic flow of ETC vehicle in every station, and can be implemented from service template, just by assigning the parameters *time*, *space* and *vehicle* as *one day*, *station* and *ETC*, respectively.

Table 2 The five dimensions in the service template

Time	*Space*	*Vehicle*	*Object*	*Ordering*
5 minutes				
15 minutes				
1 hour*	Station	Local	Traffic flow*	Ascend
1 day	Section	Ecdemic	Traffic trend	Descend
1 week	Line	ETC	Proportion	No order*
1 month	Region	MTC	Mileage	
1 year	Network*	ALL*		
Until now				

Note: *The default value in each dimension.

After the jobs' completion, the results would be materialised. The volume estimation is supported by the service template to choose the appropriate persistence: the relational database is the default output; No-SQL storage is used when results volume is more than 100 thousands. It is sound because relation-based database has better read efficiency via indexes and the column-based No-SQL storage owns better write performance via parallelism. Generally, the job with the parameters of fine-grained time (e.g., five minutes) or coarse-grained space (e.g., network) would produce large size results.

Due to the general service template, the jobs are evolutionary, because their universality can be abstracted as multiple dimensional parameters. Additionally, when requirements change, new parameters and their values can be supplemented to the template once instead of updating various concrete jobs tediously.

3.3 Visualisation service

In the output step, the visualisation service provides multiple dedicated components to draw the interactive plots on the jobs' results in the storage (Zhang et al., 2017). The common statistics or predictive plots of the domain web portal can be composed by the service's JavaScript application programming interface (API) toolkits.

As Table 3, there are main three components in the visualisation service: base map, plot and overlay. The plots in the highway domain can be drawn on an essential base map, and it can be switched to another type. For example, some line plots about regional traffic flow can instantly re-plotted as histogram in order. Moreover, for interactive user-defined search, the composite box is supported by the overlay component to integrate multiple plots in a synthesised view. All those only require a few lines of codes.

The visualisation service is evolutionary, because customised visual effects to present results can be configured by the service components without much hard-coding.

Table 3 The components of the visualisation service

Base map	*Plot*	*Overlay*
Region	Point	Composite
Road	Line	Read-only
Region + road	Pie	No
	Histogram	

4 Case study

Our work is evaluated by case studies in the practical project mentioned in Section 2. The infrastructure layer is supported by *WoCloud* of *China Unicom*. Four virtual machines are supplied for the system, each of which owns 8 cores CPU, 32 GB RAM and 750 GB storage with RHEL 6.3 ×86_64 operating system. To support the data layer and the processing layer, three of those machines make up a Hadoop 2.6.0 cluster, where the data layer consists of MySQL 5.1 and HBase 1.6.0; the processing layer is the encapsulation of Hadoop MapReduce. As the application layer, the fourth machine runs the web portal to present interactive data analyses.

The system has been adopted by Henan Transport Department since October 2017. It has done the analyses on the toll data of the years 2016 and 2017, which involves 175 million vehicles, 32 highway lines, and 384 toll stations. The data analyses would be carried out in every first day of a month as follows. The toll data of the last month would be imported into the system through the ETL service. When data cleaning is finished, 13 domain jobs are triggered to execute in parallel. After the jobs' completion, their results are available instantly in the web portal by the virtualisation service. The execution duration in such a procedure is no more than ten minutes.

The portal contains more than 50 functions in seven categories, where three typical cases are discussed here. It shows the current situation in provincial road network in

Figure 4, where five plots (as *point*, *histogram*, *line* and *pie*) on two base maps (*region* and *road*) are presented. It involves four domain jobs for *traffic flow*, *traffic trend* and vehicle *proportion*, extended from the service template with the parameterisation of *one day* for *time* and *network/region* for *space*. In a typical temporal perspective, Figure 5 illustrates parts of the monthly situation for March 2017. The eight plots (as *histogram*, *line* and *pie*) come from five domain jobs, whose *time* parameter is *one month* (i.e., March 2017). In those jobs, parameter *space* is *station*, *region* and *network*, respectively; parameter *object* is *traffic flow* and *proportion*. The *overlay* component of visualisation service is used in three *line*-plots for results comparison. In another spatial perspective, Figure 6 reveals parts of the highway line situation from three domain jobs, whose *space* parameter is *line*. Besides the composite *overlay* for trends comparison, the plot here is interactive and linked with others when specific line is selected in any plot.

Figure 4 Current situation overview (see online version for colours)

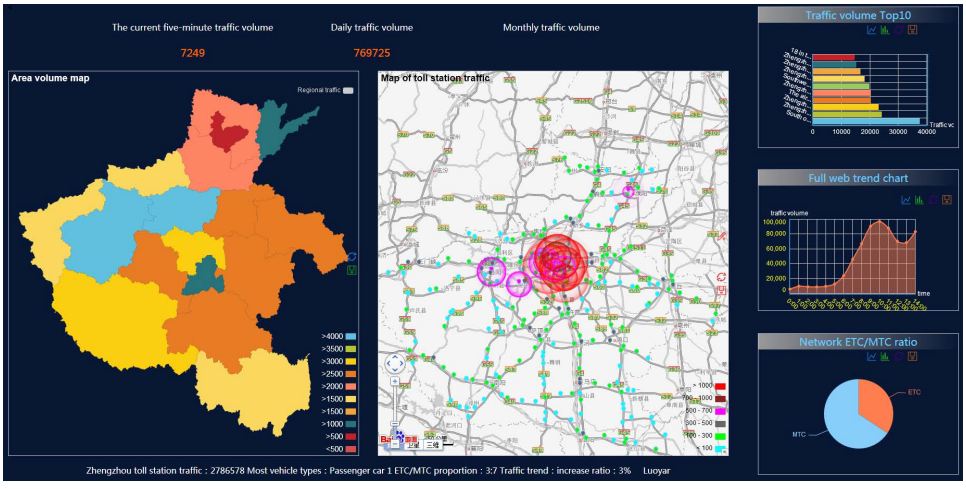
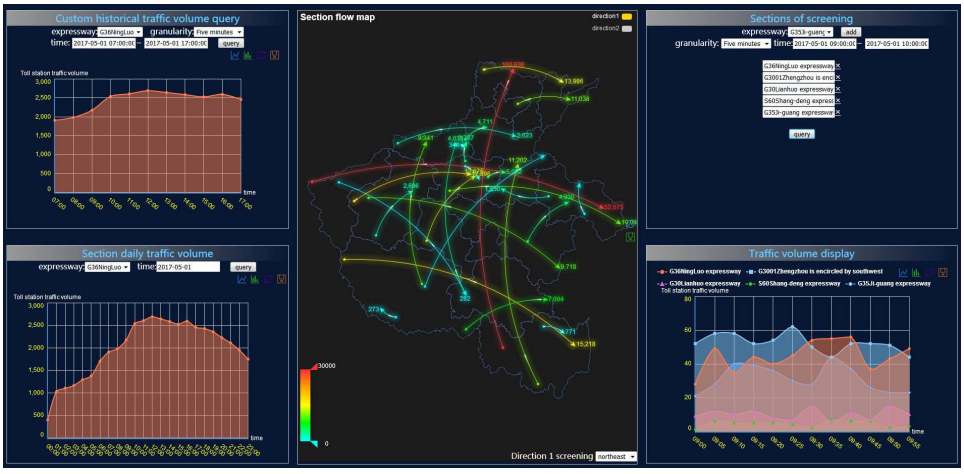


Figure 5 Monthly situation (see online version for colours)

Figure 6 Highway line situation (see online version for colours)



As a result, through three types of services, the toll data analyses are evolutionary in procedural step of loading, processing and presentation on spatio-temporal data.

5 Conclusions and future work

On massive toll data, a service solution is proposed in highway domain for routine statistical and predictive analyses in an evolutionary way. Through the ETL service, analytical service template, and visualisation service, the business data analysis can be implemented rapidly and conveniently by the domain technicians without much development knowledge. In the future, more analytical service templates are expected to promote the efficiency for widely adoption in highway domain.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61702014), Beijing Municipal Natural Science Foundation (No. 4192020), and the Top Young Innovative Talents of North China University of Technology (No. XN018022).

References

Ding, W. and Cao, Y. (2016) 'A data cleaning method on massive spatio-temporal data', in Wang, G., Han, Y. and Martínez Pérez, G. (Eds.): *Advances in Services Computing: 10th Asia-Pacific Services Computing Conference, APSCC 2016, Proceedings*, Springer International Publishing, Cham, Zhangjiajie, China, 16–18 November, pp.173–182.

Ding, W., Zhang, S. and Zhao, Z. (2017) 'A collaborative calculation on real-time stream in smart cities', *Simulation Modelling Practice and Theory*, Vol. 73, No. 4, pp.72–82.

Ding, W., Zou, J. and Zhao, Z. (in press) 'A multidimensional service template for data analysis in highway domain', *International Journal of Internet Manufacturing and Services*.

Du, Y., Li, X. and Xiong, P. (2012) 'A Petri net approach to mediation-aided composition of web services', *IEEE Transactions on Automation Science and Engineering*, Vol. 9, No. 2, pp.429–435.

Ghesmoune, M., Lebbah, M. and Azzag, H. (2016) 'State-of-the-art on clustering data streams', *Big Data Analytics*, Vol. 1, No. 13, pp.1–27.

Giret, A., Garcia, E. and Botti, V. (2016) 'An engineering framework for service-oriented intelligent manufacturing systems', *Computers in Industry*, Vol. 81, pp.116–127.

Hochreiner, C., Vogler, M., Waibel, P. and Dustdar, S. (2016) 'VISP: an ecosystem for elastic data stream processing for the internet of things', in *IEEE 20th International Enterprise Distributed Object Computing Conference (EDOC2016)*, Vienna, Austria, pp.1–11.

Holguín-Veras, J. and Preziosi, M. (2011) 'Behavioral investigation on the factors that determine adoption of an electronic toll collection system: passenger car users', *Transportation Research Part C: Emerging Technologies*, Vol. 19, No. 3, pp.498–509.

Houssaini, C.E., Nassar, M. and Kriouile, A. (2015) 'A cloud service template for enabling accurate cloud adoption and migration', *2015 International Conference on Cloud Technologies and Applications (CloudTech)*, IEEE, Marrakech, Morocco, pp.1–6.

Kim, M., Park, J., Oh, J., Chong, H. and Kim, Y. (2008) 'Study on network architecture for traffic information collection systems based on RFID technology', in *IEEE Asia-Pacific Services Computing Conference (APSCC 2008)*, Yilan, Taiwan, pp.63–68.

Lehrig, S., Hilbrich, M. and Becker, S. (2018) 'The architectural template method: templating architectural knowledge to efficiently conduct quality-of-service analyses', *Software: Practice and Experience*, Vol. 48, No. 2, pp.268–299.

Luo, H., Ding, W. and Gui, S. (2016) 'A service composition method through multiple user-centric views', in Wang, G., Han, Y. and Martínez Pérez, G. (Eds.): *Advances in Services Computing: 10th Asia-Pacific Services Computing Conference, APSCC 2016, Proceedings*, Springer International Publishing, Cham, Zhangjiajie, China, 16–18 November, pp.228–238.

Nguyen, D.K., Lelli, F., Taher, Y., Parkin, M., Papazoglou, M.P. and van den Heuvel, W-J. (2011) 'Blueprint template support for engineering cloud-based services', *European Conference on a Service-based Internet (ServiceWave 2011)*, Springer, Berlin, Heidelberg, pp.26–37.

Xia, Y., Wang, X. and Ding, W. (2019) 'A data cleaning service on massive spatio-temporal data in highway domain', in *Service-Oriented Computing – ICSOC 2018 Workshops*, Hangzhou, China, pp.229–240.

Zhang, C., Zhang, K., Yuan, Q., Peng, H., Zheng, Y., Hanratty, T., Wang, S. and Han, J. (2017) 'Regions, periods, activities: uncovering urban dynamics via cross-modal representation learning', *Proceedings of the 26th International Conference on World Wide Web (WWW2017)*, International World Wide Web Conferences Steering Committee, Perth, Australia, pp.361–370.

Zhang, Z., Liu, C., Li, X., Han, Y., Lv, C. and Ding, W. (2018) 'A declarative service-based method for adaptive aggregation of sensor streams', *IEEE Access*, Vol. 7, No. 1, pp.89–98.

Zheng, Y., Capra, L., Wolfson, O. and Yang, H. (2014) 'Urban computing: concepts, methodologies, and applications', *ACM Transactions on Intelligent Systems and Technology*, Vol. 5, No. 3, pp.1–55.

Zhou, X., Li, R., Chen, T. and Zhang, H. (2016) 'Network slicing as a service: enabling enterprises' own software-defined cellular networks', *IEEE Communications Magazine*, Vol. 54, No. 7, pp.146–153.