Transdisciplinary Lifecycle Analysis of Systems
R. Curran et al. (Eds.)

doi:10.3233/978-1-61499-544-9-377

Towards Cloud Big Data Services for Intelligent Transport Systems

Gavin KEMP[1], Genoveva VARGAS-SOLAR[2,3], Catarina FERREIRA Da SILVA[1],
Parisa GHODOUS[1], Christine COLLET[2] and Pedropablo LOPEZ AMAYA[1]
[1] Université Lyon 1, LIRIS, CNRS, UMR5202, bd du 11 novembre 1918, Villeurbanne, F69621, France
[2] LIG Grenoble Institute of Technology, 681 rue de la Passerelle, Saint Martin d'Hères, France
[3] LIG-LAFMIA, CNRS 681 rue de la Passerelle, Saint Martin d'Hères, France

Abstract. In later years, the increase in computation power and data storage has opened new perspectives to data analysis. The possibility to analyse big data brings new insights into obscure and useful correlations in data providing undiscovered knowledge. Applying big data analytics to ~~the~~ transport data has brought better understanding to the transports network revealing unexpected choking points in cities. This technology is still largely inaccessible to small companies due to their limited computational resources and complex for large ones due to the time needed to develop a big data analytical system Using the high scalability of Cloud and the use of specialized services in a services oriented architecture, new perspective are to developing efficient and scalable big data infrastructure adapted to transport systems. This paper presents a big data infrastructure using service oriented architecture.

Keywords. ITS, Big Data, Cloud Services, NoSQL

Introduction

During the last five years, the problem of providing intelligent real time data management using cloud computing technologies has attracted attention from both academic researchers, like [1], [2]and industrial practitioners like Google Big Query, IBM, Thales. They mostly concentrate on modelling stream traffic flow, yet they barely combine different data flows with other big data to provide new Intelligent Transport Services (ITS). ITS apply technology for integrating computers, electronics, satellites and sensors for making every transport mode (road, rail, air, water) more efficient, safe, and energy saving. ITS effectiveness relies on the prompt processing of the acquired transport-related information for reacting to congestion, dangerous situations, and, in general, optimizing the circulation of people and goods. Integration, storage and analysis of huge data collections must be adapted to support ITS for providing solutions that can improve citizens' lifestyle and safety.

In order to address these challenges it is important to consider that big data introduce aspects to consider according to its properties described by the 5V's model [3]: Volume, Velocity, Variety, Veracity, Value.

Volume and velocity (i.e., continuous production of new data) have an important impact in the way data is collected, archived and continuously processed. Transport data are generated at high speed by arrays of sensors or multiple events produced by devices and transport media (buses, cars, bikes, trains, etc.). These data need to be processed in real-time, near real-time or in batch, or as streams. Important decisions must be made in order to use distributed storage support that can maintain these data collections in apply on them analysis cycles. Collected data, involved in transport scenarios, can be very heterogeneous in terms of formats and models (unstructured, semi-structured and structured) and content. Data variety imposes new requirements to data storage and database design that should dynamically adapt to the data format, in particular scaling up and down. ITS and associated applications aim at adding value to collected data. Adding value to big data depends on the events they represent and the type of processing operations applied for extracting such value (i.e., stochastic, probabilistic, regular or random). Adding value to data, given the degree of volume and variety, can require important computing, storage and memory resources. Value can be related to quality of big data (veracity) concerning (1) data consistency related to its associated statistical reliability; (2) data provenance and trust defined by data origin, collection and processing methods, including trusted infrastructure and facility.

Processing and managing big data, given the volume and veracity and given the greedy algorithms that are sometimes applied to it, for example, giving value and making it useful for applications, requires enabling infrastructures. Cloud architectures provide unlimited resources that can support big data management and exploitation. The essential characteristics of the cloud lie in on-demand self-service, broad network access, resource pooling, rapid elasticity and measured services [4]. These characteristics make it possible to design and implement services to deal with big data management and exploitation using cloud resources to support applications such as ITS.

The objective of our work is to manage and aggregate cloud services for managing big data and assist decision making for transport systems. Thus this paper presents our approach for developing data storage, data cleaning and data integration services to make an efficient decision support system. Our services will implement algorithms and strategies that consume storage and computing resources of the cloud. For this reason, appropriate consumption models will guide their use.

The remainder of the paper is organized as follows. Section 2 describes work related to ours. Section 3 introduces our approach for managing transport big data on the cloud for supporting intelligent transport systems applications. Section 4 presents a case study of the application that validates our approach. Finally, Section 5 concludes the paper and discusses future work.

1. Related work

This section focuses on big data transport projects, namely to optimize taxi usage, and on big data infrastructures and applications for transport data events.

Transdec [5] is a project to create a big data infrastructure adapted to transport. It is built on three tiers comparable to the MVC (Model, View, Controller) model for transport data. The presentation tier, based on GoogleTM Map, provides an interface to express queries and expose the result, the query interface provides standard queries for the presentation tier and a data tier is spatiotemporal database built with sensor data and traffic data. This work provides an interesting query system taking into account the dynamic nature of town data and providing time relevant results in real-time.

Urban insight [6] is a project studying European town planning. In Dublin they are working event detection through big data, in particular on an accident detection system using video stream for CCTV (Closed Circuit Television) and crowdsourcing. Using data analysis they detect anomalies in the traffic and identify if it is an accident or not. When there is an ambiguity they rely on crowdsourcing to get further information. The project RITA [7] in the United States is trying to identify new sources of data provided by connected infrastructure and connected vehicles. They work to propose more data sources usable for transport analysis. L. Jian et al. [8] propose a service-oriented model to encompass the data heterogeneity of several Chinese towns. Each town maintains its data and a service that allows other towns to understand their data. These services are aggregated to provide a global data sharing service. These papers propose methodologies to acknowledge data veracity and integrate heterogeneous data into one query system. An interesting line to work on would be to produce predictions based on this data to build decision support systems.

H. V. Jagadish et al. [3] propose a big data infrastructure based on five steps: data acquisition, data cleaning and information extraction, data integration and aggregation, big data analysis and data interpretation. X. Chen et al. [9] use Hadoop-gis to get information on demographic composition and health from spatial data. J. lin and D. Ryaboy [10] present their experience on twitter to extract information from log information. They concluded that an efficient big data infrastructure is a balancing speed of development, ease of analysis, flexibility and scalability. Proposing a big data infrastructure on the cloud will make developing big data infrastructures more accessible to small businesses for several reasons: little initial investment, ease of development through Service-Oriented Architecture (SOA) and using services developed by specialist of each service.

N. J. Yuan et al. [11], Y. Ge et al. [12] and D. H. Lee et al. [13] worked a transport project to help taxi companies optimize their taxi usage. They work on optimizing the odds of a client needing a taxi to meet an empty taxi, optimizing travel time from taxi to clients, based on historical data collected from running taxis. Using knowledge from experienced taxi drivers, they built a mapping of the odds of passenger presence at collection points and direct the taxis based on that map. These research works do not use real-time data thus making it complicated to make accurate predictions and react to unexpected events. They also use data limited to GPS and taxi usage, whereas other data sources could be accessed and used.

D. Talia [14] presents the strengths of using the cloud for big data analytics in particular from a scalability stand point. They propose the development of infrastructures, platforms and service dedicated to data analytics. J. Yu et al. [15] propose a service oriented data mining infrastructure for big traffic data. They propose a full infrastructure with services such accident detection. For this purpose they produce a large database with the collected data by individual companies. Individual services would have to duplicate the data to be able to use it. This makes for highly redundant data as the same data I stored by the centralised database, the application and probably the data producers. What is more, companies could be reluctant to giving away their data with no control for its use. H. Demirkan and D. Delen [16] proposes a service oriented decision support system using big data and the cloud. A data service provides a centralised database to which application can query.

The state of the art reveals a limited use of predictions from big data analytics for transport-oriented systems. The heavy storage and processing infrastructures needed for big data and the current available data-oriented cloud services make possible the continuous access and processing of real time events to gain constant awareness, produce big data-based decision support systems, which can help take immediate informed actions. Cloud based big data infrastructure often concentrate around the massive scalability but don't propose a cheap method to simply aggregate big data services.

2. Managing transport data in smart cities

Consider the scenario where a taxi company needs to embed decision support in electric vehicles, to help their global optimal management. The company uses electric vehicles that implement a decision cycle to reach their destination while ensuring optimal recharging, through mobile recharging units. The decision making cycle aims at ensuring vehicles availability both temporally and spatially; and service continuity by avoiding congestion areas, accidents and other exceptional events. The taxis and mobile devices of users are equipped with video camera and location trackers that can emit the location of the taxis and people. For this purpose, we need data on the position of the vehicles and their energies levels, have a mechanism to communicate unexpected events and have usage and location of the mobile recharging station.

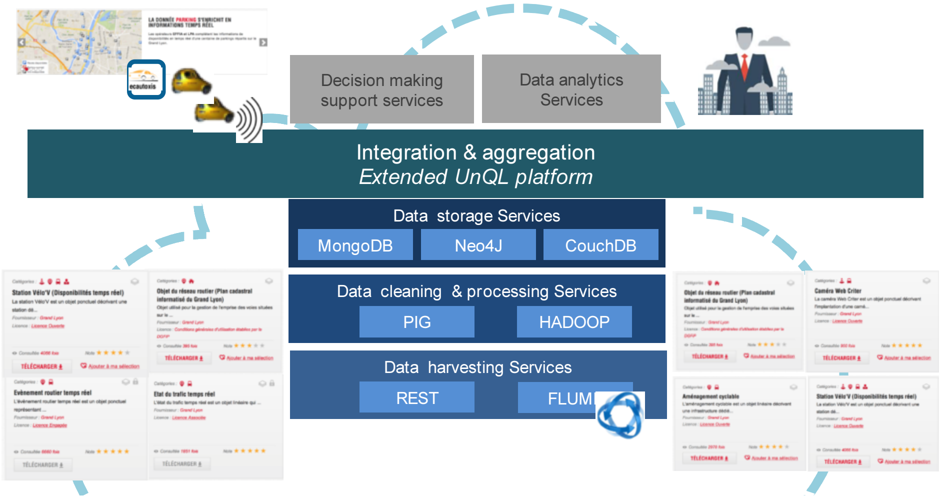

Figure 1. Big Data services.

Figure 1 shows the services that this application relies on. These services concern data acquisition and cleaning and information extraction in one side of the spectrum, and on the other side big data analysis, integration and aggregation services, and decision-making support.

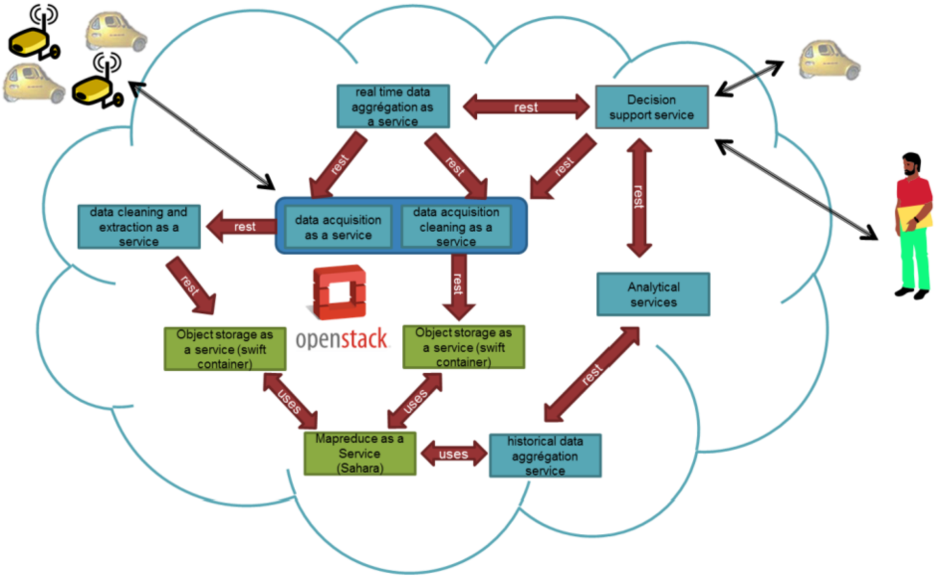
real time data aggrégation as a service
rest
Decision support service
rest
rest
rest
rest
data cleaning and extraction as a service
rest
data acquisition as a service
data acquisition cleaning as a service
rest
rest
Analytical services
openstack
Object storage as a service (swift container)
Object storage as a service (swift container)
uses
uses
rest
Mapreduce as a Service (Sahara)
uses
historical data aggrégation service

Figure 2. Big Data infrastructure.

In cloud computing everything is viewed as a service (XaaS). As a consequence cloud software (SaaS) is built as an aggregate of services exploiting services available on the cloud infrastructure (IaaS). In this spirit, we build a big data infrastructure were individual companies (**Figure 2**), specialized in their level of big data; can propose services on each level of big data. This also means that the companies wanting a big data infrastructure will be able to simply build it from an aggregation of services proposed by specialized companies.

Next section explains how the services are built to insure high scalability of cloud and the controlled cost.

3. Some Cloud Big data services for transport

3.1. Data acquisition service

The first step of a big data infrastructure is well collecting the big data. This is basically hardware and infrastructure services that transfer, to NoSQL data stores adapted to the format of the data, the data acquired by the vehicles, users, and sensors deployed in cities (e.g. roads, streets, public spaces). This is done by companies and entities such as town or companies managing certain public spaces, who have data collecting facilities. These companies propose and sell their data on a cloud infrastructure. Clients using the same cloud infrastructure could then pay to have access the data., in our case the universities Openstack infrastructure [17]. Using object storage such as Swift [18] using MongoDB [19], these companies will have a cheap highly scalable and sharable data store. Also the sharding capability of these data stores offers high horizontal scalability but also faster analysis threw MapReduce and data availability.

In the taxis scenario, data from the town and from the vehicles would be stored on several object data stores known as container with Swift. Since there are several companies involved, they will store there data on separate data stores. When a client wishes to access the data, the company would propose REST (representational state of transfer) [20] services to which the client can query to get access to the data. These companies propose different service to historical data and real-time data. The historical data service provides authentication to the client so they can apply Openstack MapReduce service (Sahara) [21] on their container. The real-time data service provides a Json or XML file of the latest data produced.

We are implementing and testing the data acquisition service. This services uses NodeJS module to acquire the city data from the Grand Lyon [22] but also from Twitter and from Bing search engine using REST requests. Still using REST requests these services will post the data onto a Mongodb database container to store as historical data. The service provides functions to access data via REST either with the key to the data store when wanting to query or analyses the historical data or the latest file acquired when using the real-time data service. The data is stored under XML, Json or the original image file or PDF before data extraction.

3.2. Information extraction and cleaning service

The next step is cleaning and data extraction. This consists of both extracting the information from unstructured data and cleaning the data. This could be done by the company producing the data or an independent company depending on the level of structuration of the data. Highly structured data would likely be cleaned by the company producing the data as they understand best its production and thus know how best to clean it up. For highly unstructured data like sound or video data, highly specialized expert would be needed to extract the information.

This would be used to pot outliers in the data. Using MapReduce, the company acquiring the data or the company contracted to do it would perform statistical analysis to identify for example outliers in the data. This is important as, for example, a malfunction in a sensor loop could either ignore passing traffic or register non-existing traffic. Cleaning these events is important since inaccurate data produced by a dodgy sensor can break a model.

3.3. Integration and aggregation services

The objective of big data analytics is to use the large volume of data to extract new knowledge by searching, for example, for patterns in the data. This often has a consequence of data coming from a wide variety of sources. This means the data has to be aggregated into a usable format for the analytics tools to use. This service proposes services for real-time data aggregation and historical data aggregation.

The real-time data aggregation service gets the data from the individual data stores real-time data services and proposes a formatted file with the data from all the data acquiring service simply by fusing together the data provided by the real-time data acquisition services. Thus we aggregate data from the city, state of recharging stations, having location of people based on the time stamp or the GPS location.

The historical data aggregation will have to find a way to do similar action but with the data stores. The problem is that having data on several separate data stores is not a usable format. Importing all the data into a new huge data store would be redundant on already existing resources making this service potentially excessively expensive and as for temporary stores would be long to build when having to import terabytes of data as well as being expensive on network cost as well as time consuming. To solve this problem, this service will to propose a query interface for simple querying and processing service to process the data masse by converting a form simple programing language into UNQL queries to collect and pre-process the data before being integrated into a model.

3.4. Big data analytical and decision support services

The whole point of big data is to identify and extract information form the mass of data. Predictive tools can be developed to anticipate the future. The role of this service is to provide a computer model of the historical data. It also provides the algorithm applied to the individual pieces of data. Thus using the model provided by the analytical service and the algorithm applied to the real-time data we can approach similar situations and act accordingly.

The decision support service composes several services. On the strategic level and using the model and the algorithm proposed by the big data analytical services, the decision support service provides an interface exposing the data situation in real-time but also predictions on events. For example, regularly observing an increase in the population in one place and traffic jams 30 minutes later we can deduct cause and effect and intervene in future situations so the taxis avoid and evacuate that area.

This service also generates data on the decision taken by the strategists to build more elaborate model including the consequence of this decision to then provide better decision support. On the vehicle level, services will provide advice to the vehicle for optimal economic driving based on the driving conditions. It also provides a database were the information on the dangers of the road is stored.

4. Conclusion

This paper proposes a set of big data services as a starting point of a dedicated and flexible infrastructure for managing and exploiting transport data. Our approach uses NoSQL systems deployed in a multi-cloud setting and makes sharing decisions for ensuring data availability.

Our transport data service architecture is validated in a scalable and adaptable ITS case study of electric vehicles using big data analytics on the cloud. This provides a global view of current status of town transport, helps making accurate strategic decisions, and insures maximum security to the vehicles and their occupants.

For the time being our data transport services concentrate in improving design issues with respect to NoSQL support. We are currently measuring performance with respect to different sizes of data collections. We have noticed that NoSQL provides reasonable response times once an indexing phase has been completed. We are willing to study the use of indexing criteria and provide strategies for dealing with continuous data. These issues concern our future work.

Acknowledgement

We thank the Région Rhône-Alpes who finances the thesis work of Gavin Kemp by means of the ARC 7 programme (http://www.arc7-territoires-mobilites.rhonealpes.fr/), as well as the competitiveness cluster LUTB Transport & Mobility Systems, in particularly Mr. Pascal Nief, Mr. Timothée David and Mr. Philippe Gache for putting us in contact with local companies and projects to gather use case scenarios for our work.

References

[1] V. Gulisano, R. Jiménez-Peris, M. Patiño-Martı́nez, C. Soriente, and P. Valduriez, “StreamCloud: An elastic and scalable data streaming system,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, pp. 2351–2365, 2012.

[2] F. Lecue, S. Tallevi-Diotallevi, J. Hayes, R. Tucker, V. Bicer, M. L. Sbodio, and P. Tommasi, “STAR-CITY,” in *Proceedings of the 19th international conference on Intelligent User Interfaces - IUI ’14*, 2014, pp. 179–188.

[3] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, *Big Data and Its Technical Challenges*, vol. 57, no. 7. 2014.

[4] P. M. and T. Grance, “The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology,” 2008.

[5] U. Demiryurek, F. Banaei-Kashani, and C. Shahabi, “TransDec:A Spatiotemporal Query Processing Framework for Transportation Systems,” *IEEE*, pp. 1197–1200, 2010.

[6] A. Artikis, M. Weidlich, A. Gal, V. Kalogeraki, and D. Gunopulos, “Self-Adaptive Event Recognition for Intelligent Transport Management,” pp. 319–325, 2013.

[7] D. Thompson, G. McHale, and R. Butler, “RITA,” 2014. [Online]. Available: http://www.its.dot.gov/data_capture/data_capture.htm.

[8] L. Jian, J. Yuanhua, S. Zhiqiang, and Z. Xiaodong, “Improved Design of Communication Platform of Distributed Traffic Information Systems Based on SOA,” in *2008 International Symposium on Information Science and Engineering*, 2008, vol. 2, pp. 124–128.

[9] X. Chen, H. Vo, A. Aji, and F. Wang, “High performance integrated spatial big data analytics,” in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data - BigSpatial ’14*, 2014, pp. 11–14.

[10] J. Lin and D. Ryaboy, “Scaling big data mining infrastructure : The twitter Experience,” *ACM SIGKDD Explor. Newsl.*, vol. 14, no. 2, p. 6, Apr. 2013.

[11] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, “T-finder: A recommender system for finding passengers and vacant taxis,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, pp. 2390–2403, 2013.

[12] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, “An energy-efficient mobile recommender system,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’10*, 2010, p. 899.

[13] D.-H. Lee, H. Wang, R. Cheu, and S. Teo, “Taxi Dispatch System Based on Current Demands and Real-Time Traffic Conditions,” *Transp. Res. Rec.*, vol. 1882, pp. 193–200, 2004.

[14] D. Talia, “Clouds for scalable big data analytics,” *Computer (Long. Beach. Calif).*, vol. 46, no. 5, pp. 98–101, 2013.

[15] J. Yu, F. Jiang, and T. Zhu, “RTIC-C: A Big Data System for Massive Traffic Information Mining,” in *2013 International Conference on Cloud Computing and Big Data*, 2013, pp. 395–402.

[16] H. Demirkan and D. Delen, “Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud,” *Decis. Support Syst.*, vol. 55, no. 1, pp. 412–421, 2013.

[17] Open, “Openstack,” 2015. [Online]. Available: http://www.openstack.org/.

[18] openstack, “swift,” 2015. [Online]. Available: http://docs.openstack.org/developer/swift/.

[19] P. J. Sadalage and M. Fowler, *NoSQL Distilled*. 2012.

[20] F. Valverde and O. Pastor, “Dealing with REST Services in Model-driven Web Engineering Methods,” Jan. 2009.

[21] openstack, “sahara,” 2015. [Online]. Available: http://docs.openstack.org/developer/sahara/.

[22] GrandLyon, “Smart Data,” 2015. [Online]. Available: http://data.grandlyon.com/.