Big Data Collections and Services For Building Intelligent Transport Applications

Gavin Kemp, Pedropablo López Amaya, Catarina Ferreira da Silva, Genoveva Vargas-Solar, Parisa Ghodous, Christine Collet

BIG DATA COLLECTIONS AND SERVICES FOR BUILDING INTELLIGENT TRANSPORT APPLICATIONS

Gavin Kemp[1], Pedropablo López Amaya[1], Catarina Ferreira Da Silva[1], Genoveva Vargas-Solar[2], Parisa Ghodous[1*] and Christine Collet[3]
[1]*Univ Lyon, Université Claude Bernard Lyon 1, LIRIS UMR 5205 CNRS, F-69621, Villeurbanne, France*
[2]*CNRS, LIG-LAFMIA, 681 rue de la Passerelle, Saint Martin d'Hères (38400), France*
[3]*Grenoble Institute of Technology, LIG, Saint Martin d'Hères (38400), France*

ABSTRACT

This paper presents an approach for building data collections and cloud services required for building intelligent transport applications. Services implement Big Data analytics functions that can bring new insights and useful correlations of large data collections and provide knowledge for managing transport issues. Applying data analytics to transport systems brings better understanding to the transport networks revealing unexpected choking points in cities. This facility is still largely inaccessible to small companies and citizens due to their limited access to computational resources. A cloud service oriented architecture opens new perspectives for democratizing the use of efficient and personalized big data management and analytics.

Keywords: Intelligent Transport System, Big Data, Cloud Services, NoSQL

1. INTRODUCTION

Given the accumulation over the years of data and to the continuous production of streams by different kinds of providers like sensors, smart devices and smart cities infrastructure, there is an explosion in available data [2]. In contrast, even if computing power has also increased, it has not increased at the same rate as data. Even if Moore's law seems to be showing its limit [27] to process large data collections, cloud architectures deliver unlimited resources necessary for storing data and to process it using greedy algorithms requiring important computing and memory resources. This has made it possible to manage and analyze large data collections that often require resources that go beyond the capacities of classic systems (i.e., Big Data). In order to address Big Data storage and analytics and decide which cloud services are required it is important to consider Big Data properties described by the 5V's model [17]: Volume, Velocity, Variety, Veracity, Value.

Volume and *velocity* (i.e., continuous production of new data) determine the way data is collected, archived and continuously processed. Transport data are generated at high speed by sensors farms, devices and transport media (buses, cars, bikes, trains, etc.). These data need to be processed in real-time, recurrently, continuously or in batch. Important decisions must be made in order to use distributed storage support that can maintain these data collections, and prepare them to 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 that should dynamically adapt to the data format, in particular scaling up and down. Data transformation, processing and managing can be challenging, given volume and veracity and the greedy algorithms that are sometimes applied to it. Intelligent Transport Systems (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.

* **Corresponding author:**
parisa.ghodous@univ-lyon1.fr

Giving value to data and making it useful for applications, requires enabling infrastructures. The cloud provides a ready to use execution environments with the required physical resources and platforms to be used for Big Data management, and elasticity management mechanisms to adapt the provision of resources at runtime [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.

Thus, this paper presents an approach for building data collections and cloud services required for building intelligent transport applications. It introduces the implementation of data collection and storage strategies provided as services to manage Big Data collections with high degree of variety and velocity. These strategies are implemented by a multi-holder service oriented Big Data infrastructure used by Intelligent Transport Systems.

Accordingly, the remainder of the paper is organized as follows. Section 2 describes work related to transport big data, big data analytics and service oriented big data. Section 3 presents the architecture and the individual service for this service oriented architecture and primarily the services involved. Section 4 presents the data collection and storage services needed in the proposed architecture. Finally, Section 5 concludes the paper and discusses future work.

2. RELATED WORK

2.1 Intelligent Transport Solutions

Existing work and projects show that using Big Data for transport can provide very interesting applications. We describe examples of projects and applications that explicitly deal with Big Data for providing solutions for intelligent transport systems. Transdec [11] is a project to create a big data infrastructure adapted to transport. This work provides an interesting query system taking into account the dynamic nature of town data and providing time relevant results in real-time. It is based on GoogleTM Map to provide an interface to express queries and present results. Queries are evaluated on top of a spatio-temporal database built with data harvested from sensors and traffic.

Urban insight [3] is a project studying European town planning in Dublin. They detect events through big data analysis, in particular for detecting accidents using video stream for CCTV (Closed Circuit Television) and crowdsourcing. They detect anomalies in the traffic and identify if it is an accident or not using data analytics algorithms. When there is an ambiguity they rely on crowdsourcing to get further information. The project RITA [31] in the United States identifies new sources of data provided by connected infrastructure and connected vehicles. They work to propose more data sources usable for transport analysis. L. Jian and co [18] 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. N. J. Yuan and co [35], Y. Ge and co [12] and D. H. Lee and co [21] 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 [28] 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 and co [33] propose a service oriented data mining infrastructure for big traffic data. They propose a full infrastructure with services such accident detection. Therefore, 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. Thus, data is duplicated into multiple databases that must be integrated to provide a global view useful for target applications. What is more, companies could be reluctant to giving away their data with no control for its use.

These approaches 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.

2.2 Collecting Data from the City

Collecting data in the city can be daunting task but the availability of highly efficient mobile network as well as social networks have opened the possibilities to enable collective data harvesting.

Crowd sensing is a method of data collection using human as a sensor through the use in particular of smart phones sensors and GPS. Rajib Rana and co. [25] collected data to create a noise mapping using smart phones' application called MobSLM. This application measures noise level from the environment through the smartphone microphone when the latter is not used.

Active crowdsourcing is the act of collection data from the population through the use of pools and application encouraging users to give information. The OpenStreetMap [14] is a map generated by users using a map definition application. It works very much in the same way than Wikipedia where a community of ordinary users collaboratively produces maps and geographic information. Urban insight [3] uses crowdsourcing to confirm accidents using a mobile application that pools the population in the vicinity of a supposed accident.

Data is also collected form the population in a more passive way. On the transport front several people have collected anonymous data using public transport. Neal Lathia and Licia Capra [20] collected data using London Oystercard to minimise travellers the spending when going through London. Pierre Borgnat and co. [6] analyzed data from the bike renting program in the city of Lyon called Velo'v to understand the dynamic movement of population in that city. Julian Candi and co. [7] collected phone logs for mobiles phone operators to analyses human behavior. This paper observed measurable tendencies from anomalous events. Jie Bao and co [4] use the data from location based social networks like Foursquare, Loopt, and GeoLife to produce recommendation tools that compare the users history with the history of location experts to provide location the user would be interested in visiting. Jing Yuan and co [34] used what is known as floating vehicles. These vehicles, taxi cabs this case, act as mobile GPS sensors mirroring the fluidity of the roads.

2.3 Transport Big Data Analytics on the Cloud

The state of the art reveals a limited use of predictions from Big Data analytics for transport systems. Big data analytics is a domain combining both existing methods and new ones that the data expert does not necessarily master. The use of the cloud for Big Data analytics has shown good results because of the elastic provision of resources. The heavy storage and processing infrastructures needed for Big Data and the current available data-oriented cloud services enable 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 architectures often concentrate around the massive scalability but do not propose a method to simply aggregate Big Data services.

H. V. Jagadish et al. [17] 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. [8] use Hadoop-gis to get information on demographic composition and health from spatial data. J. lin and D. Ryaboy [24] present their experience on Twitter to extract information from logs. They concluded that an efficient Big Data infrastructure is a balancing speed of development, ease of analysis, flexibility and scalability.

Siamak Tavakoli and Ali Mousavi [30] demonstrated their cloud infrastructure for scientific analysis. Using Hadoop map-reduce, they classified the scientific algorithms according to how easy they could be adapted to map-reduce. Thus class 1 is when an algorithm can be executed with one map-reduce; class 2 is when the algorithm need sequential map-reduce; class 3 is when each iteration of an algorithm executes one map reduce cycle; and class 4 is when each iteration needs multiple map-reduce cycles.

Kurt Thearling [5] introduces the main families and technics for data mining. He classified technics into two main families: classical and next generation. The classical technics include statistical models very good for making predictions, nearest neighbor, clustering and generally technics visualizing data as space with as many dimensions as variable. The next generation of techniques includes decision trees, neural networks and rules induction; they view data analysis as a series of tests. There are also more advanced methods proposed in [32].

Finally, the system Ricardo [9] is a tool that integrates the R scripting language with Hadoop. The objective is to provide data analyst tools to easily use map-reduce. It provides an Application Programing Interface to R that connects it to a Hadoop cluster. It can convert R objects into JaQL [23] queries to analyze the data. It has been proven successful with analytical technics like Latent-Factor Model or principal component analysis.

Domenico Talia [28] proposes a three level of Big Data analytical services. The SaaS provides data mining algorithms and knowledge discovery tools. The PaaS provides a platform for the development of new data analytical services. The IaaS provides the low level tools to data mining. In the same way Zibin Zheng et al. [36] have proposed a similar vision applied to analyzing logs.

H. Demirkan and D. Delen [10] propose a service oriented decision support system using Big Data and the cloud. They do this by combining data from multiple databases into a single database then duplicate de database to the services using the data.

Eric E. Schadt et al. [26] demonstrate the efficiency that cloud computing could have for Big Data analytics, showing that analysis of 1 Petabyte of data in 350 minutes for 2040 dollars.

Zhenlong Li et al. [22] proposed a service oriented architecture for geoscience data where they separate the modelling service for geoscience, the

data services, processing services and the cloud infrastructure.

Several articles have demonstrated the strength of cloud and Big Data in particular for instancing large quantities of computing power [1, 15, 29].

3. MANAGING TRANSPORT BIG DATA IN SMART CITIES

Consider a scenario where a taxi company needs to embed decision support in vehicles, to help their global optimal management. The company uses vehicles that implement a decision making 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. 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. More details about this scenario can be found in [19].

Figure 1 shows the services that these applications rely on. They concern data acquisition, cleaning and information extraction; and big data analysis, integration and aggregation, and decision making support services.

In cloud computing, everything is viewed as a service (XaaS). In this spirit, we are building a Big Data infrastructure (Figure 2), specialized in three types of complex operations: collection, storage and analytics. Services implementing different algorithms and strategies for implementing these operations can be used for the personalization of an Intelligent Transport System (ITS) application. Thus, companies using our Big Data services will be able to simply build their applications by aggregating services according to their requirements.

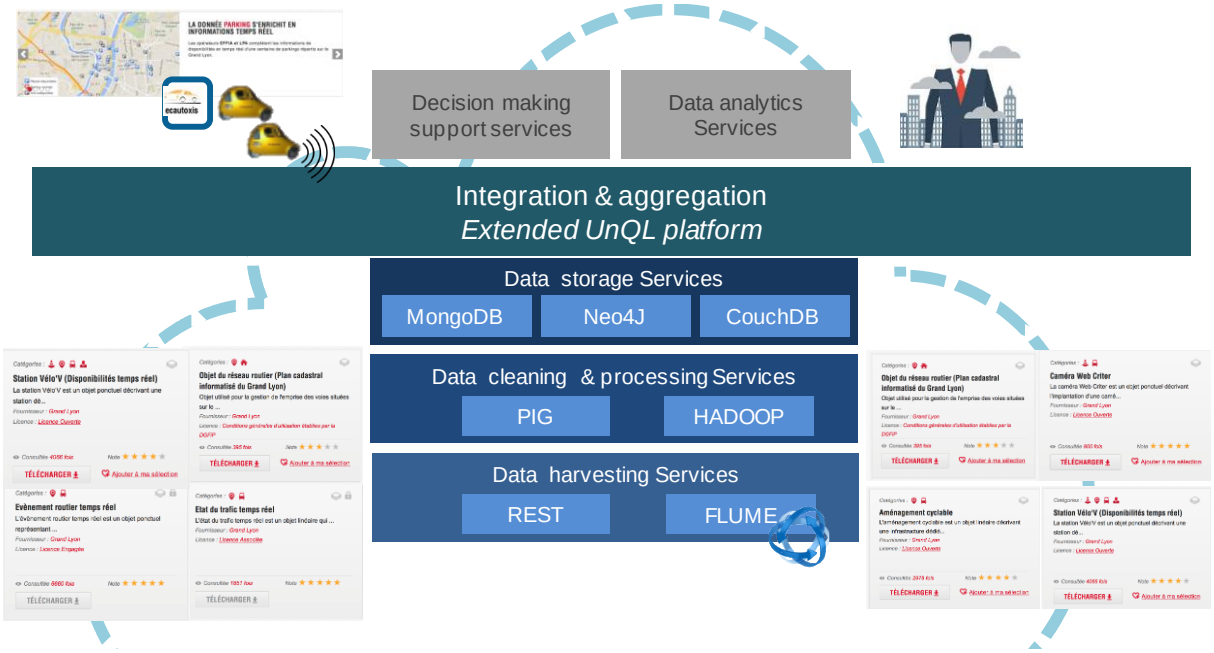

Figure 1: Big data services

3.1 Data Harvesting Services

The first step of a Big Data infrastructure is collecting the data. This is basically hardware and infrastructure services that produce data consumed by services and archive them in different NoSQL data stores according to their characteristics. Data are acquired by the vehicles, users, and sensors deployed in cities (e.g. roads, streets, public spaces) according to different strategies such as collective explicit and implicit crowdsourcing, push continuous data production performed by sensors. This is done by companies and entities such as town or companies managing certain public spaces, which have data collecting facilities.

Raw data from sensors cannot be used directly to perform analytics processes, since it can be incomplete, it can contain noise, and there is little knowledge about its structure and the conditions in which it has been produced. To be able to exploit the data, the analyst needs information about the data structure (variable, graph or table …) and about its content like the distribution of values of every attribute, possible dependencies among attributes,

possible missing values or erroneous ones, provenance. Information extraction and cleaning services are tools to extract comprehensive views of the data. These tools include mostly statistical exploratory methods that are very good to provide a comprehensive view of the data, such as principal component analysis [29]. Views produced by these information extraction services can be used by data scientists to determine how to clean them and produce collections that can be used to perform analytics.

3.2 Integration and Aggregation Services

The objective of big data analytics is to extract new knowledge by searching for example for patterns within proper and representative data collections. This means that heterogeneous data stemming from different providers has to be integrated into a usable format for the analytics tools to use. Integration and aggregation services propose algorithms for real-time data aggregation and historical data aggregation.

The real-time data aggregation service gets the data produced by the real-time data acquisition services and generates a database that integrates data from all the data acquisition services. Thus, an integrated database could aggregate data from the city, states of recharging stations having, location of people for example based on their time stamp.

The historical data aggregation service has to find a way to do this action. Importing all the data into a new huge data store would be redundant on already existing resources making this service 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.

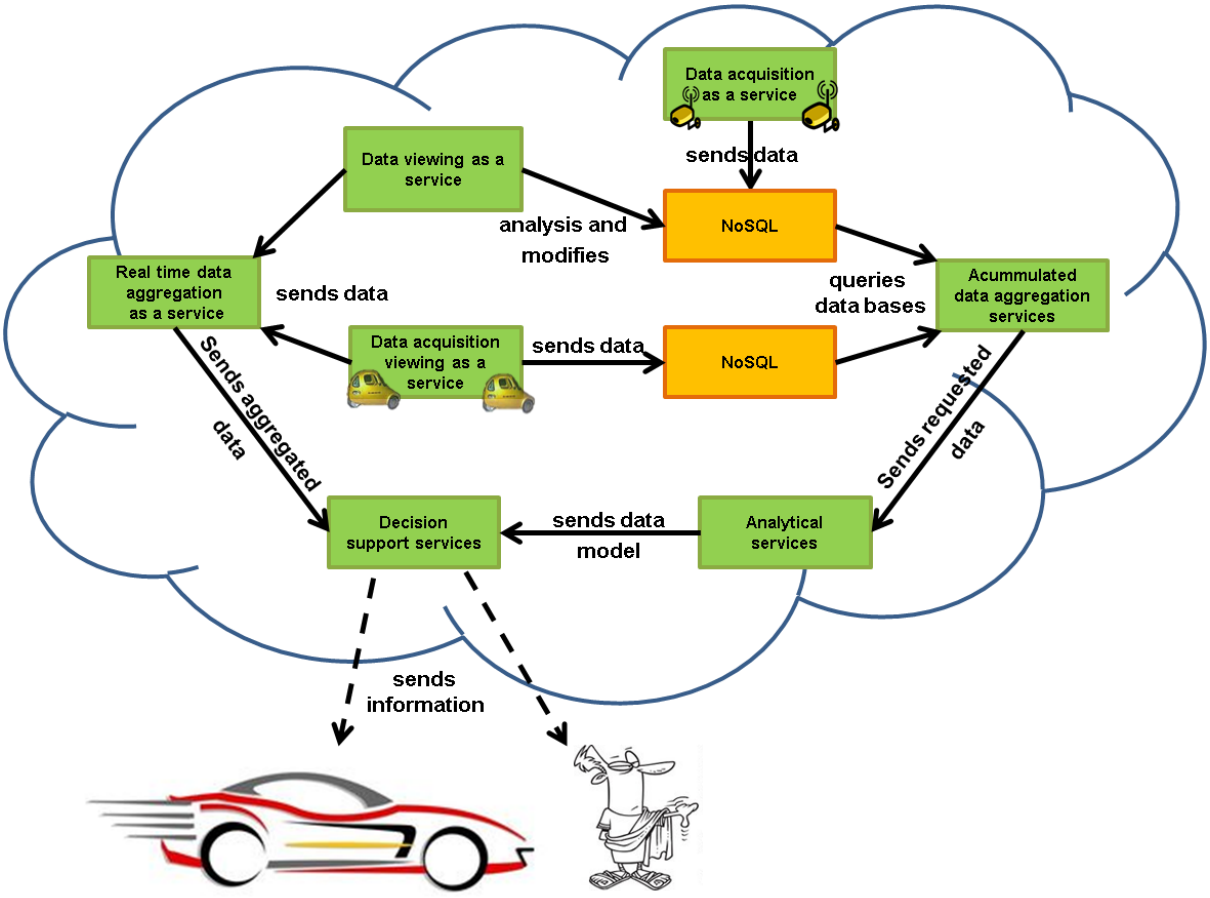

Figure 2: Service oriented big data architecture

3.3 Big Data Analytics and Decision Support Services

The whole point of Big Data is to identify and extract meaningful information. Predictive tools can be developed to anticipate the future. The role of the Big Data analytics and decision support services in our infrastructure is to provide data analytics solutions. These solutions can be used for predicting events or for decision making tasks by composing several services. For example, regularly observing an increase in the population in one place and traffic jams 30 minutes later we can deduce cause and effect situations and intervene in future situations so the taxis avoid and evacuate that area. Data on decisions made by strategists are stored to be used in future tasks. For example, provide advice to the vehicle for

optimal economic driving based on the driving conditions and on previous recommendations given in similar situations. The next section presents the implementation and the architecture of the data collection services and the data storage strategies used for storing data.

4. IMPLEMENTING DATA HARVESTING AND STORAGE SERVICES

This section presents implementation issues of the collection services used for developing an experimental testbed of our approach. We programmed REST[2] services that export streaming functions to return public statuses that match one or more filters about traffic. The filters used for our application are "track" and "location". Collected data are archived in NoSQL stores using different data sharding (i.e., horizontal fragmentation) techniques that ensure data availability.

4.1 Data Storage Service

The general architecture of a data storage service consists of three layers: the Software Development Kit (SDK), Retrieving Storing API and Collector and Database Proxy. The SDK layer communicates with other layers using the HTTP requests and responses. The service is hosted by a server deployed on the layer exporting the interface Retrieving Storing API. Data collected and returned by the Collector are transformed into the target NoSQL store data model. To store the documents into the NoSQL database we use the Database Proxy. In this module we implemented all the CRUD operations ("Create, Read, Update and Delete).

Based on this architecture we implemented the Grand Lyon REST service for collecting data produced by the city. We built it as a RESTful Web service following the parameters shown in Table 1. We used MongoDB NoSQL database to store the retrieved data according to different sharding strategies for organizing data and ensuring availability.

Table 1: Grand Lyon service parameters

Parameters	Description
URI **/traffic/event**	Event streaming to collect and store documents from Grand Lyon
URI **/traffic/state**	State streaming to collect and store documents from Grand Lyon
URI **/traffic/status**	Traffics received since the beginning of streaming
type ***Necessary***	Type of streaming **Example:** *event or state*
URI **/traffic/stop**	Stop streaming. Response will contain all traffics received since the beginning of stream
type ***Necessary***	Type of streaming **Example:** *event or state*

Figure 3 shows a sequence diagram specifying the collect and store functions. The different actors will interact through the medium of different functions and reach the "Grand Lyon Collector". We decided not to show the real URI for the GET request from the collector to the Grand Lyon REST API to make this diagram simple and comprehensible. The "Collector" will retrieve the different documents. A formatting has to be made to avoid the duplications and build a strategy for MongoDB. The formatted document will be stored into MongoDB by sending a PUT request on the MongoDB service. This sequence will be repeated each minute since Grand Lyon updates traffic data each minute. We can also follow the sequence to get the status and to stop the service.

The diagram for the "state" service will be the same as the one in Figure 3 by replacing "event" by "state".

4.2 Experimentation

Our preliminary tests collected a total of 43 MB based on these services implemented. To store them we used our MongoDB service. This service provides two functions. The first one is to list all documents stored into a collection, and the second one to store a document into a collection. It was implemented using the Mongodb module from NodeJS. In this scenario we use the traffic "events" collection. The UML class diagram in Figure 4 is a simplified version of the elements inside each different type of documents.

The retrieved documents are stored into a MongoDB architecture deployed in the Openstack Cloud infrastructure, at the Lyon 1 University. We use for this purpose three routers (one for each service) to redirect each service to its corresponding router and avoid the saturation from a router. It contains three

[2] https://dev.twitter.com/rest/public

configuration servers for security reasons and three replica sets with each three members converted into three shards.

Two types of sharding strategies [16] were tested for organizing our data collection in distributed storage supports (Hashed and Ranged sharding). For the "hashed" strategy we used as shard key the "_id" field from the collection. For the "ranged" strategy, the "full_location" (element containing the address) shard key was used. On Figure, we analyze data's distribution evolution in function of the size of the collection.

The "hashed" sharding strategy distributes the data in a more homogeneous way (delta of 22.48% for maximum size), whereas the ranged collection has a difference of more than 79.32% between shards.

To test the performance of the strategies we ran some tests using a query into the "tweets" collections. The query did not take into account the "shard key" to test the performance of the distribution itself. Results can be seen on Figure. In this test, for a database size of 10.449 MB there is a significant performance difference from the "hashed" the "ranged" strategy. When the size of the database increases four times (10.449 to 43.039) the "ranged" strategy analyses 41 documents/ms faster than the "hashed" strategy. Even if the distribution of data into the shards is more homogeneous with a "hashed" collection it is not more efficient. To confirm this theory we decided to make an insertion test (Figure).

When the size of the database increases from 10MB to 43MB, 6824 documents were inserted into the "tweets" collection. This test showed a change in the performance of each strategy. The theory that "hashed" strategy is less efficient than the "ranged" strategy for our collection is verified thanks to the insertion test. The "ranged" strategy inserted 40 times more documents per second than the "hashed" strategy.

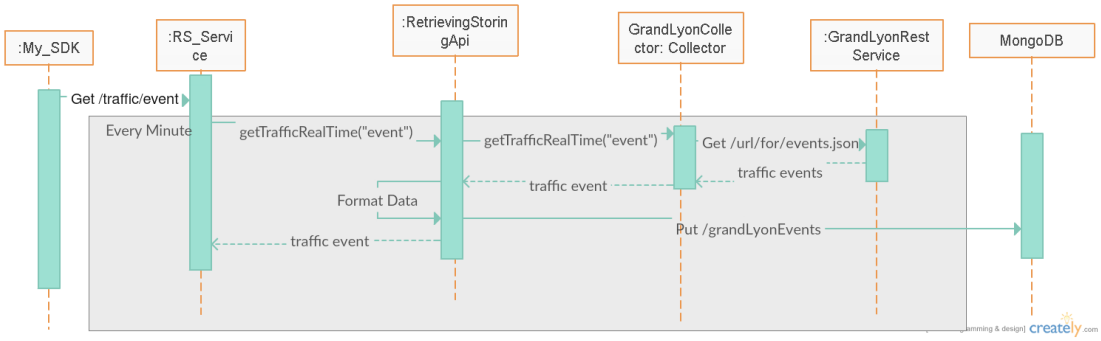

Figure 3: Collecting and storing using Grand Lyon Service

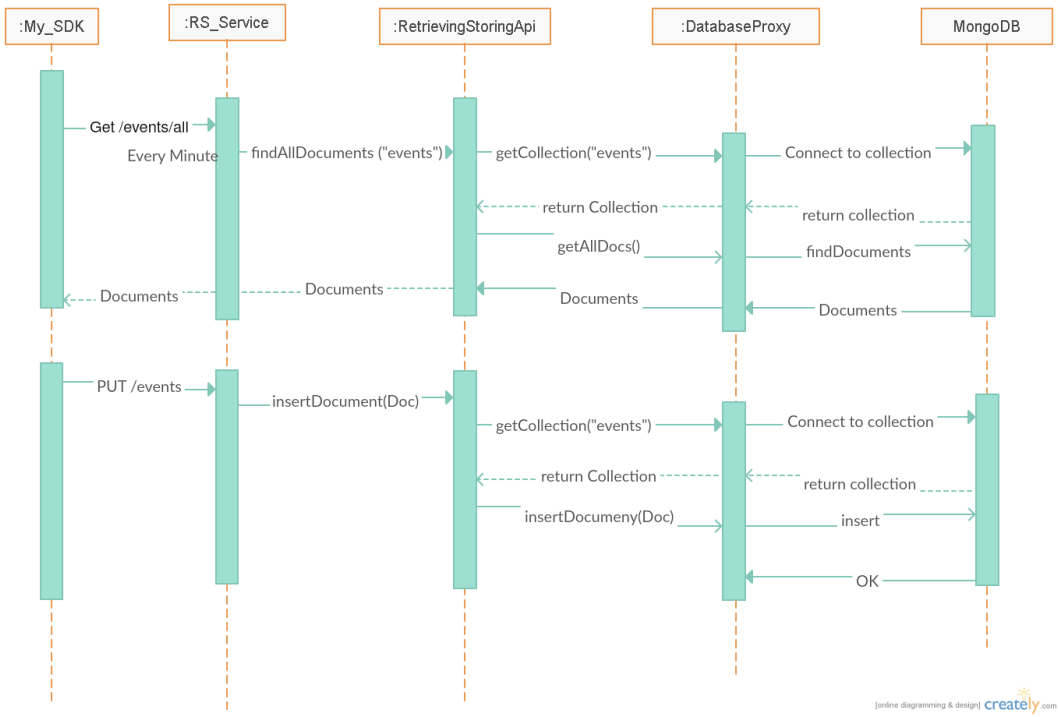

Figure 4: MongoDB implemented Service

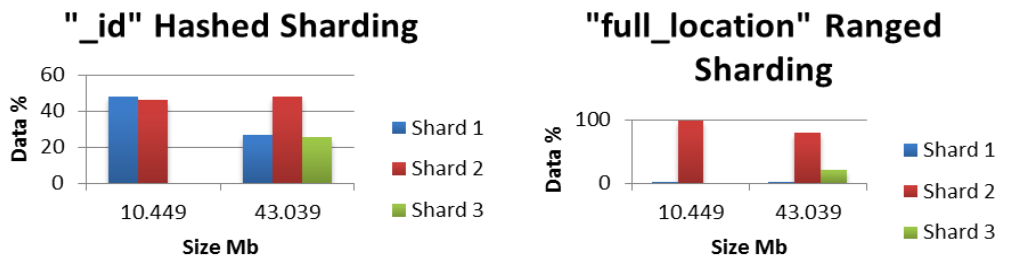

Figure 5: Strategy shard distribution analysis

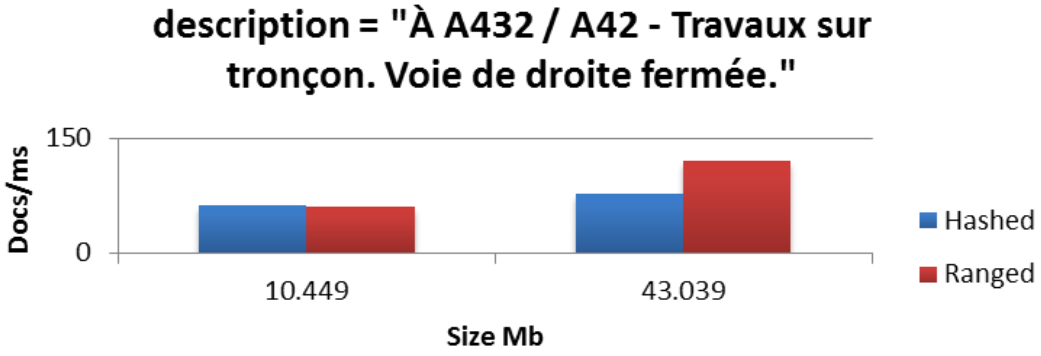

Figure 6: Query performance test

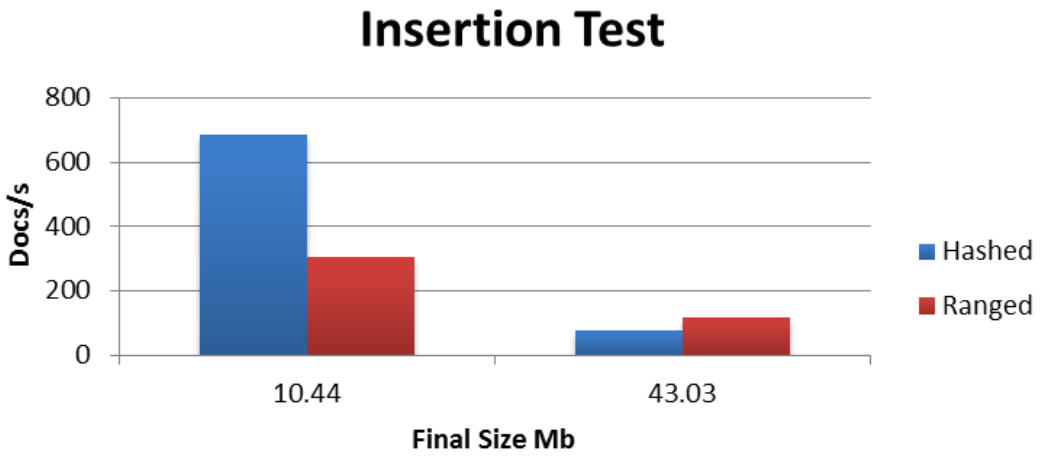

Figure 7: Insertion strategy performance

5. CONCLUSION

This paper introduced our service oriented cloud architecture for hosting several services for each step of big data analytics. This paper focuses on the cycle collection-storage of data. According to our general architecture we built services to collect data from REST services and store them in a MongoDB database by means of a storage service. The storage service enables the use of different sharding strategies to organize data and increase read/write performances. We used collected data to run experiments using different keys and sharding techniques. For our database the ranged strategy seems more efficient than a hashed strategy.

For the time being our storage service concentrates 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 region Rhône-Alpes who finances the thesis work of Gavin Kemp by means of the ARC 7 program (http://www.arc7-territoires-mobilites.rhonealpes.fr/).

REFERENCES

1. Abramova, Veronika and Jorge Bernardino, 2013, "NoSQL databases: A step to database scalability in web environment," Proceedings of the International C* Conference on Computer Science and Software Engineering - C3S2E '13, pp. 14-22. Retrieved (http://dl.acm.org/citation.cfm?id=2494444.2494447).
2. Anon. n.d. "Data Universe Explosion & the Growth of Big Data | CSC." Retrieved October 29, 2015 (http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode).
3. Artikis, A., Weidlich, M., Gal, A., Kalogeraki, V. and Gunopulos, D., 2013, "Self-adaptive event recognition for intelligent transport management," *2013 IEEE International Conference on. IEEE*, pp. 319-25.
4. Bao, J., Zheng, Y. and Mokbel, M. F., 2012, "Location-based and preference-aware recommendation using sparse geo-social networking data," *In Proceedings of the 20th International Conference on Advances in Geographic Information Systems, ACM*, pp. 199-208. SIGSPATIAL '12. New York, New York, USA: ACM Press. Retrieved March 4, 2016 (http://dl.acm.org.gate6.inist.fr/citation.cfm?id=2424321.2424348).
5. Berson, A., Smith, S. and Thearling, K., 2004, "An overview of data mining techniques," Data Mining Application for CRM, pp. 1-49. Retrieved (http://www.stat.ucla.edu/~hqxu/stat19/dm-techniques.pdf).
6. Borgnat, P., Fleury, E., Robardet, C. and Scherrer, A., 2009, "Spatial analysis of dynamic movements of V{é}lo'v, Lyon's shared bicycle program," *In ECCS*.
7. Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A. L., 2008, "Uncovering individual and collective human dynamics from mobile phone records. " *Journal of Physics A: Mathematical and Theoretical*, Vol. 41, No. 22, pp.224015.
8. Chen, X., Vo, H., Aji, A. and Wang, F., 2014, "High performance integrated spatial big data analytics," *In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data - BigSpatial '14*, New York, New York, USA: ACM Press, pp. 11-14.
9. Das, S., Haas, P. J. and Beyer, K. S., 2000, "Ricardo : Integrating R and Hadoop categories and subject descriptors," pp. 987-998.
10. Demirkan, H. and Dursun D., 2013, "Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud," *Decision Support Systems*, Vol. 55, No.1, pp. 412-21. Retrieved (http://dx.doi.org/10.1016/j.dss.2012.05.048).
11. Demiryurek, U., Banaei-Kashani, F. and Shahabi, C., 2010, "TransDec: A spatiotemporal query processing framework for transportation systems," *IEEE*, pp. 1197-1200.
12. Ge, Y., Xiong, H., Tuzhilin, A., Xiao, K., Gruteser, M., & Pazzani, M., 2010. "An energy-efficient mobile recommender system. " *In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 899-908, ACM.
13. Grance, T. and Mell, P., 2008, "The NIST definition of cloud computing recommendations of the national institute of standards and technology,".
14. Haklay, M. and Weber, P., 2008, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Computing*, Vol. 7, No. 4, pp. 12-18. Retrieved March 4, 2016 (http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4653466).
15. Hipgrave, S., 2013, "Smarter fraud investigations with big data analytics," *Network Security*, 2013(12):7-9. Retrieved May 20, 2015 (http://www.sciencedirect.com/science/article/pii/S1353485813701351).
16. Inc., MongoDB, 2014, "Sharding and MongoDB," pp. 1-80.
17. Jagadish, H. V., et al., 2014, "Big data and its technical challenges," *Communications of the ACM*, Vol. 57, No.7, pp. 86-94.
18. Lee, J., Jia, Y., Shu, Z. and Zhu, X., 2008, "Improved design of communication platform of distributed traffic information systems based on SOA," *In 2008 International Symposium on Information Science and Engineering*, *IEEE*, Vol. 2, pp. 124-128.
19. Kemp, G., Vargas-Solar, G., da Silva, C. F., Ghodous, P. and Collet, C., 2015, "Aggregating and managing big realtime data in the cloud : Application to intelligent transport for smart cities," *In VEHITS 2015. Lisbon*.
20. Lathia, N. and Capra, L., 2011, "Mining mobility data to minimise travellers' spending on public transport," *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, pp. 1181-1189. New York, New York, USA: ACM Press. Retrieved March 4, 2016

(http://dl.acm.org.gate6.inist.fr/citation.cfm?id=2020408.2020590).
21. Lee, D. H., Wang, H., Cheu, R. and Teo, S., 2004, "Taxi dispatch system based on current demands and real-time traffic conditions," *Transportation Research Record*, Vol. 1882, pp. 193-200.
22. Li, Z., et al., 2015, "Enabling big geoscience data analytics with a cloud-based, MapReduce-enabled and service-oriented workflow framework," *PloS one*, 10(3):e0116781. Retrieved July 2, 2015 (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0116781).
23. Lim, Sejoon. 2008, "Scalable SQL and NoSQL Data Stores," Statistics.
24. Lin, J. and Ryaboy, D., 2013, "Scaling big data mining infrastructure : The twitter experience," *ACM SIGKDD Explorations Newsletter*, Vol. 14, No. 2, pp. 6.
25. Rana, R., Chou, C. T., Bulusu, N., Kanhere, S. and Hu, W., 2015, "Ear-Phone: A context-aware noise mapping using smart phones," *Pervasive and Mobile Computing*, Vol. 17, pp. 1-22. Retrieved March 3, 2016 (http://www.sciencedirect.com/science/article/pii/S1574119214000273).
26. Schadt, Eric E., Michael D. Linderman, Jon Sorenson, Lawrence Lee, and Garry P. Nolan., 2011. "Cloud and Heterogeneous Computing Solutions Exist Today for the Emerging Big Data Problems in Biology." Nature reviews. Genetics 12(3):224. Retrieved July 27, 2015 (http://www.nature.com.gate6.inist.fr/nrg/journal/v12/n3/full/nrg2857-c2.html).
27. Schaller, R. R., 1997, "Moore's law: Past, present and future." *Spectrum, IEEE*, Vol.34, No.6, pp.52-59.
28. Talia, Domenico., 2013. "Clouds for scalable big data analytics." *Computer*, Vol.46, No.5, pp.98-101.
29. Tannahill, B. K., and Jamshidi, M., 2014, "System of Systems and Big Data analytics–Bridging the gap." *Computers & Electrical Engineering*, Vol.*40*, No.1, pp.2-15. (http://www.sciencedirect.com/science/article/pii/S004579061300298X).
30. Tavakoli, S., and Mousavi, A., 2008, "Adopting user interacted mobile node data to the flexible data input layer architecture." In *Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008. International Conference on* , pp. 533-538, IEEE. (http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4762044).
31. Thompson, Dale, Gene McHale, and Randy Butler. 2014. "RITA." Retrieved (http://www.its.dot.gov/data_capture/data_capture.htm).
32. Yan, W., Brahmakshatriya, U., Xue, Y., Gilder, M., and Wise, B., 2013, P-PIC: Parallel power iteration clustering for big data., *Journal of Parallel and Distributed computing*, Vol. *73*, No.3, pp.352-359. (http://www.sciencedirect.com/science/article/pii/S0743731512001487)
33. Yu, J., Jiang, F., and Zhu, T., 2013, RTIC-C: A big data system for massive traffic information mining." In *Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on*,pp. 395-402, IEEE. (http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6821021)
34. Yuan, J., Zheng, Y., Xie, X., and Sun, G., 2011, "Driving with knowledge from the physical world." In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 316-324, ACM. (http://dl.acm.org/citation.cfm?doid=2020408.2020462)
35. Yuan, N. J., Zheng, Y., Zhang, L., and Xie, X., 2013, "T-finder: A recommender system for finding passengers and vacant taxis." *Knowledge and Data Engineering, IEEE Transactions on*, Vol.*25*, No.10, pp.2390-2403.
36. Zheng, Z., Zhu, J., and Lyu, M. R., 2013, "Service-generated big data and big data-as-a-service: an overview." *In Big Data (BigData Congress), 2013 IEEE International Congress on*, pp. 403-410, IEEE.

ABOUT THE AUTHORS

Gavin KEMP is PhD student at the Computer Science Department of the Claude Bernard Lyon 1 University (France), and joined the Service Oriented Computing team of the LIRIS lab in 2014. His current research interests include cloud computing services, big data and intelligent transport systems.

Pedropablo López Amaya was a intern who joined the Service Oriented Computing team of the LIRIS lab in 2015 during which he participated in the AMBED project in developing a data collection service. He currently works for Worldline, an IT service company.

Catarina Ferreira Da Silva is Associate Professor at the Computer Science Department of the Institute of Technology of the University Claude Bernard Lyon 1 University (France), and joined the Service Oriented Computing team of the LIRIS lab in 2012. Previously she worked at the Centre for Informatics and Systems of the University of Coimbra (Portugal). She obtained her PhD thesis in computer science (2007) from the University of Lyon 1. Her current research interests include Cloud

Computing Services, Business Services, Linked Data, Semantic Web and Interoperability.

Genoveva Vargas Solar is a Senior Scientist of the French Council of Scientific Research (CNRS) and Deputy Director of the Franco-Mexican Laboratory of Informatics and Automatic Control (LAFMIA, UMI 3175). She is also member of the Informatics Laboratory of Grenoble (France) and invited research fellow of the Data and Knowledge Management Group at Universidad de las Américas Puebla. Her research contributes to the construction of service based database management systems. The objective is to design data management services guided by Service Level Agreements (SLA). She proposes methodologies, algorithms and tools for integrating, deploying and executing a service composition for programing data management functions. The results of her research are validated in the context of grids, embedded systems and clouds.

Parisa Ghodous is currently full professor in computer science department of University of Lyon I. She is head of cloud computing theme of LIRIS UMR 5205 (Laboratory of Computer Graphics, Images and Information Systems). Her research expertise is in the following areas: Cloud Computing, Interoperability, Semantic Web, Web services, Collaborative modeling, Product data exchange and modeling and Standards. She is in editorial boards of CERA, ICAE and IJAM journals and in the committees of many relevant international associations such as concurrent engineering, ISPE, Interoperability.

Christine Collet is currently full Professor of Computer Science at the Grenoble Institute of Technology (www.grenoble-inp.fr), Grenoble – France. She is the head of the database management group (HADAS http://hadas.imag.fr/) of the Grenoble Informatics Laboratory – UMR 5217. Her research domain concerns databases and their evolution in terms of data models, languages and architectures. More precisely she contributed or still contributes to research activities on: Object and Active Database management systems; Multi-scale distributed and heterogeneous data(base) management systems; Adaptive Optimization and evaluation of hybrid queries (on large data sets and streams); Distributed Composite Event Management ; Data (service) mediation and integration. She is the author or co-author of around 100 publications: books, chapters and articles in national and international journals and conferences (see the DBLP and Google Scholar servers). She directed more than 20 PhD theses. She is currently Presidente of the EDBT association (http://www.edbt.org/).

(Received March 2016, revised March 2016, accepted March 2016)