Aggregating and Managing Big Realtime Data in the Cloud
Application to Intelligent Transport for Smart Cities

Gavin Kemp[1], Genoveva Vargas-Solar[2,3], Catarina Ferreira Da Silva[1],
Parisa Ghodous[1] and Christine Collet[2]
[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*

Keywords: ITS, Big Data, Cloud Services, NoSQL.

Abstract: The increasing power of computer hardware and the sophistication of computer software have brought many new possibilities to information world. On one side the possibility to analyse massive data sets has brought new insight, knowledge and information. On the other, it has enabled to massively distribute computing and has opened to a new programming paradigm called Service Oriented Computing particularly well adapted to cloud computing. Applying these new technologies to the transport industry can bring new understanding to town transport infrastructures. 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. Proposing big data management strategies for data produced by transport infrastructures, whilst maintaining cost effective systems deployed on the cloud, is a promising approach.

1 INTRODUCTION

During the last five years, the problem of providing intelligent real time data management using cloud computing technologies has attracted more and more attention from both academic researchers, e.g. P. Valduriez team in France (Gulisano et al. 2012), Freddy Lecue's work at Ireland IBM Research Lab (Lecue et al. 2014), Big Data Initiative CSAIL Laboratory in MIT, USA, Cyrus Shahabi's team University of Southern California in USA (Demiryurek et al. 2010) 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 (Jagadish et al. 2014): 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.). This 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 theses 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

Kemp G., Vargas-Solar G., Ferreira Da Silva C., Ghodous P. and Collet C..
Aggregating and Managing Big Realtime Data in the Cloud - Application to Intelligent Transport for Smart Cities.
DOI: 10.5220/0005491001070112
In *Proceedings of the 1st International Conference on Vehicle Technology and Intelligent Transport Systems* (VEHITS-2015), pages 107-112
ISBN: 978-989-758-109-0

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 (Grance 2008). 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.

2 RELATED WORK

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

Transdec (Demiryurek et al. 2010) is a project of the University of California to create a big data infrastructure adapted to transport. It's built on three tiers comparable to the MVC (Model, View, Controller) model for transport data. The presentation tier, based on Google™ Map, provides an interface to create the 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 (Artikis et al. 2013) is a European 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's an accident or not. When there is an ambiguity they rely on crowdsourcing to get further information. The RITA (Thompson et al. 2014) project 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. (Jian et al. 2008) 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 interesting decision support systems.

(Jagadish et al. 2014) 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. (Chen et al. 2014) use Hadoop-gis to get information on demographic composition and health from spatial data. (Lin & Ryaboy 2013) 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 throw Service-Oriented Architecture (SOA) and using services developed by specialist of each service.

(Yuan et al. 2013), (Ge et al. 2010), (Lee et al. 2004) worked a transport project to help taxi companies optimize their taxi usage. They work on optimising 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 don't 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.

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.

3 MANAGING TRANSPORT BIG 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 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:

- *Data acquisition service*: 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).
- *Information extraction and cleaning servic*e: receives the data from the data acquisition services. At this level information is extracted from the unstructured data (e.g. video stream).
- *Integration and aggregation services*: combines different types of data (e.g. video, sensor, weather data) formatting the data into a standard form to make simpler querying.
- *Big data analysis service*: to process collected data and have a comprehensive vision of it. For instance, identifying value trends of different measures within data or classifying data according to its values. The consumption of taxis during rush hours in commercial zones of a city when it rains. Keep track the on-going situation of traffic in the town and predict future bottlenecks in specific zones.
- *Decision support service* exports an end user application exposing information on, for instance, the town transport systems and on new destinations to the taxi drivers, and information regarding the best moments in which a taxi should search for energy or events where clients may need a taxy transportation.

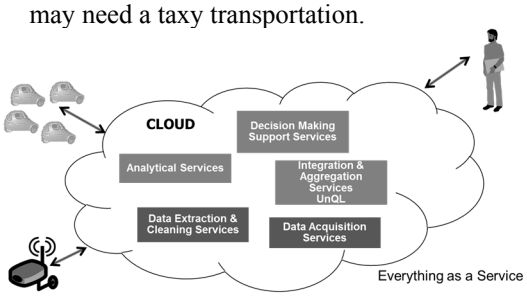

Figure 1: Big Data as a Service.

In order to let these services work in the best conditions, it is necessary to propose an approach for integrating and aggregating collected data and correlate it with data streaming from other data providers. Indeed, data must be cleaned to complete them and remove heterogeneity and then prepared (stored on disk, memory or cache) to support decision-making actions. This paper proposes a data store as a service that is a horizontal support for the services described in the previous lines.

4 TRANSPORT DATA STORE AS A SERVICE

Figure 2 presents the general architecture of the transport data store as a service that we propose. It adopts a polyglot persistence approach that combines several NoSQL systems for providing a storage support. Profiting from the cluster-oriented architecture of these systems our store uses a multi-cloud cluster based storage layer, which uses tools such as Swift (openstack 2015) and Amazon S3 (Amazon 2015).

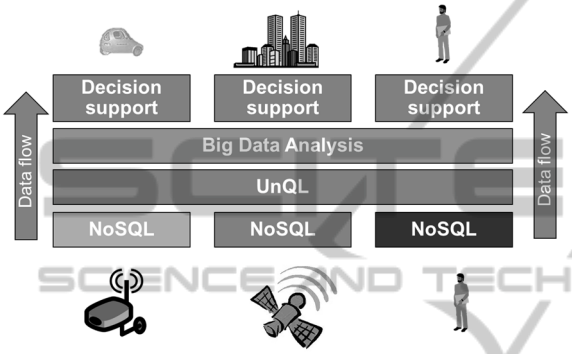

Figure 2: Data store as a service.

Our service extends an UnQL (Buneman et al. 2000) layer with data processing operators including joins and filters for storing and retrieving data in an homogeneous way.

Furthermore, it exploits the sharding strategies of the NoSQL (Cattell 2011) systems for distributing and duplicating data, and ensuring availability. Shards are organized according to ranges of values of given attributes, or to hash functions and tags related to geographic zones. This induces request balancing and ensures better performance when data must be inserted and retrieved.

The service exploits also the persistence supports of clients (disk and cache) installed in mobile devices in order to distribute data processing and ensure data availability. For example, in our transport scenario, the service uses storage provided by devices used by taxis and users to process and manage data necessary for ITS services described in the previous section. In this way it avoids data transfer that can be penalizing in terms of response time and economic cost for accessing 3G or 4G networks.

Our data store service provides a global access to clusters providing NoSQL and relational support, and enables applications designers to configure their resources provision and non-functional properties according to given requirements and cloud subscriptions. An application defines the data structures that must persist in the UML eclipse plugin and then the tool Model2Roo (https://code.google.com/p/model2roo/) generates the necessary bindings to interact with different NoSQL stores. The application designer according to a profiling phase executed using the QDB benchmark (https://github.com/qdb-io) chooses the NoSQL stores.

4.1 Designing Transport Data Collections

The data collections “Evènement routier temps reel”, “Etat du trafic temps reel”, “Borne Criter”, “Tronçon Web Criter”, “Trafic historique”, “plan Lyon”, “Aménagement cyclable”, “Caméra Web Criter” and “Station Velo'v”, provided by the project Grand Lyon (GrandLyon 2015), are sought and stored by our service in order to be able to correlate collected data with data describing the city and its infrastructures (parks, roads, commercial zones, river). This data is highly heterogeneous in format, information and update rates. There are images in JPG, JSON, XML, and PDF formats. The data is also updated at varying rates going from yearly updates to real-time data passing by daily and minutely updates. GPS and location data in devices and vehicles are seen by our service as continuous data that can be correlated to other collected data useful for performing some decision making requests, such as *which is the closest taxi (considering distance and time) to a client?*, *according to traffic and taxi-energy level, which are the possible destinations it can accept?* Data are sharded by our service to perform this type of requests that require computing resources. Our service uses a MongoDB cluster to store these data.

Data stemming from social networks particularly Twitter and Waze of taxi users are collected and stored in NeO4J. This collection provides a real-time view of the traffic, road and zones status and events. Data are sharded thanks to our storage service locally on mobile devices and on NeO4J instances deployed in the cloud.

4.2 Making Global Transport Decisions

We conducted an experimental validation of our transport data store as a service for the scenario we described in Section 3. The experiment implements a polyglot multi-database that contains data

collected from the French city Lyon. These data are retrieved by applications and infrastructure integrated by the project Grand Lyon. We then implement some important operations of the decision making cycle of the scenario (see Figure 3). The decision making cycle consists in:

- Collecting data streams from taxis and users that are mobile data providers evolving in Lyon and feeding the data store service.
- Computing recommendations, such as taxy itineraries and battery charging, according to the traffic status.

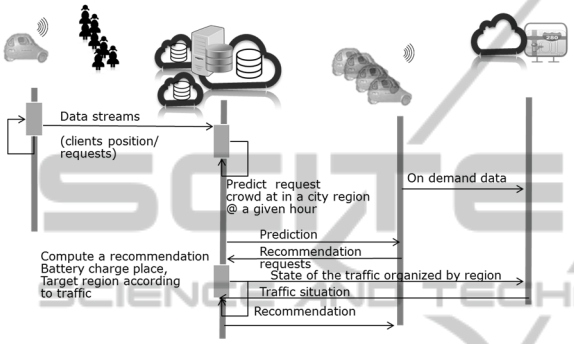

Figure 3: UML sequence diagram of the decision making process.

We focus particularly in three operations that use the transport data storage as a service approach, which are dissemination of events, optimization of energy recharging and scaling taxi provision of exceptional situations. We describe this use cases in the following sections.

4.2.1 Disseminating Events

The applications deployed in taxis and users can be used for disseminating exceptional situation events, for example, unexpected dangers (Figure 4). In our scenario a pedestrian is about to cross the road. “Vehicle A” is arriving in the same place but has no line of sight. “Vehicle B” in the area “sees” the pedestrian. The video stream from “Vehicle B” is analysed by the “Information extraction and cleaning services” and compared with topological information of the area. The pedestrian detection information is stored on a NoSQL database. As the vehicle comes in the area, the vehicle computer will make query to that database informing on potential dangers and then notifies the driver of the danger.

Depending on the nature of the danger the data store will make decisions on how long to keep that information and during which period it will re-execute the dissemination to taxis getting close to the zone.

Figure 4: Disseminating exceptional events.

4.2.2 Optimising Battery Recharging

Part of the objective of taxi companies is use only electric vehicles. Unfortunately the lack of data makes it complicated to make good strategic solutions on the locations of the recharging stations that are also mobile (Figure 5).

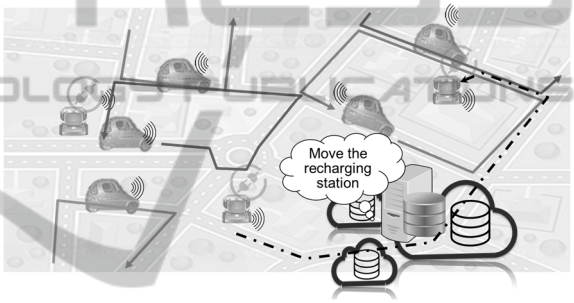

Figure 5: Optimising battery recharging.

Using UnQL queries, the historical data stored in the NoSQL databases is periodically uploaded into a Hadoop (Zikopoulos et al. 2012) based application designed to extract information to build a classification model and regression model for the real time data. Using this model and real-time data the system will make predictions on the location of taxi users and the traffic. As decision makers take decisions, this information is feed into the model to help the decision maker optimise the number of operational taxi, the location of these taxis and the location of the recharging stations by exposing the consequence previous similar decisions had.

5 CONCLUSIONS

This paper proposes a transport data store as a service that implements a distributed storage approach. Our approach uses NoSQL systems deployed in a multi-cloud setting and makes sharding decisions for ensuring data availability.

The transport data store service is validated in a scalable and adaptable ITS for 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 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.

ACKNOWLEDGEMENTS

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

Amazon, 2015. Amazon Simple Storage Service (Amazon S3). Available at: http://aws.amazon.com/s3/.

Artikis, A. et al., 2013. Self-Adaptive Event Recognition for Intelligent Transport Management. , pp.319–325.

Buneman, P., Fernandez, M. & Suciu, D., 2000. UnQL: a query language and algebra for semistructured data based on structural recursion. *The VLDB Journal*, 9(1), p.76.

Cattell, R., 2011. Scalable SQL and NoSQL data stores. *ACM SIGMOD Record*, 39(4), p.12.

Chen, X. et al., 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.

Demiryurek, U., Banaei-Kashani, F. & Shahabi, C., 2010. TransDec:A Spatiotemporal Query Processing Framework for Transportation Systems. *IEEE*, pp.1197–1200.

Ge, Y. et al., 2010. An energy-efficient mobile recommender system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*. New York, New York, USA: ACM Press, p. 899.

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

GrandLyon, 2015. Smart Data. Available at: http://data.grandlyon.com/.

Gulisano, V. et al., 2012. StreamCloud: An elastic and scalable data streaming system. *IEEE Transactions on Parallel and Distributed Systems*, 23, pp.2351–2365.

Jagadish, H.V. et al., 2014. *Big Data and Its Technical Challenges*,

Jian, L. et al., 2008. Improved Design of Communication Platform of Distributed Traffic Information Systems Based on SOA. In *2008 International Symposium on Information Science and Engineering*. IEEE, pp. 124–128.

Lecue, F. et al., 2014. STAR-CITY. In *Proceedings of the 19th international conference on Intelligent User Interfaces - IUI '14*. New York, New York, USA: ACM Press, pp. 179–188.

Lee, D.-H. et al., 2004. Taxi Dispatch System Based on Current Demands and Real-Time Traffic Conditions. *Transportation Research Record*, 1882, pp.193–200.

Lin, J. & Ryaboy, D., 2013. Scaling big data mining infrastructure : The twitter Experience. *ACM SIGKDD Explorations Newsletter*, 14(2), p.6.

openstack, 2015. swift. Available at: http://docs.openstack.org/developer/swift/.

Thompson, D., McHale, G. & Butler, R., 2014. RITA. Available at: http://www.its.dot.gov/data_capture/data_capture.htm.

Yuan, N.J. et al., 2013. T-finder: A recommender system for finding passengers and vacant taxis. *IEEE Transactions on Knowledge and Data Engineering*, 25, pp.2390–2403.

Zikopoulos, P., Eaton, C. & DeRoos, D., 2012. *Understanding big data*, Available at: http://www.lavoisier.fr/livre/notice.asp?ouvrage=2609842.