Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 52 (2015) 1004 – 1009

Procedia Computer Science

The 4th International Workshop on Body Area Sensor Networks (BASNet-2015)

Stream processing of healthcare sensor data: studying user traces to identify challenges from a big data perspective

Rudyar Cortés[a,*], Xavier Bonnaire[b], Olivier Marin[a], and Pierre Sens[a]

[a]*Université Pierre et Marie Curie, CNRS*
INRIA - REGAL, Paris, France.
[b]*Universidad Técnica Federico Santa María, Valparaíso, Chile.*

Abstract

The Internet of Things (IoT) generates massive streams of data which call for ever more efficient real time processing. Designing and implementing a big data service for the real time processing of such data requires an extensive knowledge of both input load and data distribution in order to provide a service which can cope with the workload. In this context, we study in this paper the challenges inherent to the real time processing of massive data flows from the IoT. We provide a detailed analysis of traces gathered from a well-known healthcare sport-oriented application in order to illustrate our conclusions from a big data perspective.

Keywords: Internet of Things; Healthcare Sport Services ; Big Data.

1. Introduction

One of the most important challenges in the big data era is to be able to collect and process massive and heterogeneous data flows. In this context, the Internet of Things (IoT) ecosystem[1] generates a huge amount of data from billion of internet-connected devices. The Cisco Internet Business Solutions Group (IBSG) predicts that, by 2020, the world will count 50 billion connected devices supporting various applications such as healthcare services, air pollution monitoring, transportation, energy, and so on[2].

From a big data perspective, these applications exhibit three main characteristics: (i) the data originates from millions of users/sources; (ii) users generate a massive volume of data composed of small pieces of sensing data with spatio-temporal properties; (iii) the data is mainly semi-structured. Such data can be seen as a massive stream of small pieces of spatio-temporal data generated by millions of users, which must be collected, stored, indexed, and processed in real time in many cases[3].

Real time processing of this massive flow of data generated by the IoT cannot be implemented with a traditional cluster based solution, as it obviously cannot scale to process billions or trillions of tuples on-the-fly. Therefore, fully

[*] Corresponding author. Tel.: +33 1 44 27 88 17
E-mail address: rudyar.cortes@lip6.fr

doi:10.1016/j.procs.2015.05.093

Table 1: Example of GPS tuple generated every Δt.

Parameter	Value
Longitude	−79.975199921716
Latitude	32.777506706718
Altitude	−14.42
Distance	0.12614343 [km]
Duration	51081 [ms]
Pace	8.44095 [min/km]

decentralized architectures are required. This raises new challenges in terms of data collection, data transportation and data processing.

In this paper, we present a detailed analysis of a well-known healthcare sport-oriented application in order to understand the data challenges inherent to the real time processing of IoT-generated data flows. We give an overview of these challenges according to the Big Data life cycle and from the point of view of the scalability, the heterogeneity, the timeliness and the privacy of the solutions.

The main contributions of this paper are:

- a study of a dataset from a popular IoT-oriented sport-tracking application[4],
- a discussion about the big data challenges and opportunities associated with the real time processing and the analysis of data flows from the IoT.

2. Study Case: Endomondo healthcare sports service

Healthcare services running on smartphones equipped with sensors are becoming extremely popular. Examples of such applications are *Endomondo*[5], *Nike+*[6], *RunKeeper*[7], and *Runtastic*[8] which currently track all daily activities like sport, sleep, and diet habits over sport oriented social networks. Our study case focuses on one of these healthcare services: the Endomondo sport tracker service. Endomondo is a popular sports-oriented application with nearly 30 million users worldwide[9]. It allows users to track and share their workout results with friends, and to publish them for everyone to see in the case of public profiles. Thus, public workouts are a rich source of real world traces of sports activities.

The application runs on smarthphones, uses sensors such as GPS and accelerometers in order to track the device along its route, and registers incremental parameters like distance, speed, duration and time. When the workout is finished, it can be enriched with user comments and shared via the Endomondo social network or via traditional social networks like Facebook and Twitter.

In order to acquire real world traces for our study, we gathered public information from the Endomondo web server. The data set covers the entire sports activities log of $15,090$ users picked at random. It covers a total of $333,689$ workouts obtained over five months in 2014. Every log file contains a *user profile*, a *workout summary*, and a *GPS trace*. The user profile contains general information regarding the user, such as *username*, *country*, *birthdate*, *postal code*, *sex*, *weight*, and *height*. The workout summary presents statistics assessed during the session, such as *maximum speed*, *average speed*, theoretical amount of burnt *calories*, and *weather*. The GPS trace contains all GPS points generated along the route of the workout session every Δt. The value of Δt is typically a few seconds and it depends on the type of sport. Table 1 shows an example of a GPS tuple generated along a route. Table 2 shows a summary of our data set. $6,009$ users (39.82%) did not provide information about their gender. Note that the tracking of a workout route generates a significant volume of data in terms of GPS tuples. A workout session averages 170 GPS tuples, which represents a size of 18.13 [KB].

Table 2: Sport tracker dataset summary.

Parameter	**Value**
Users	15,090
Workouts	333,689
Male	6,269
Female	2,812
Not specified	6,009
Types of sports	69
GPS tuples	56,851,893

3. Evaluation

This section presents our analysis of public workout traces extracted from the Endomondo sports tracker server. Our goal is to estimate the data distribution and workload of one of today's most popular healthcare sport services as a planning step to design a big data stream service.

3.1. Data Distribution

In order to show the spatial distribution of sport activities we measured the number of users of our sample dataset by continent and country and the activity which they generated.

The three continents with most sport users are Europe, which accounts for 60.84% of all users, followed by America (24.28%) and Asia (11.53%). The country with the highest number of registered users is the U.S., which accounts for 18.5% of all users, followed by Spain (13.8%) and the United Kingdom (8.42%). The most popular sport activity is running, which represents 42.6% of the workout publications, followed by cycling (22.87%), walking (19.95%), and others (14.53%).

Our results match the information released by Endomondo[10] about the sports popularity, which confirms the representativeness of the dataset we extracted as our study sample.

3.2. Measuring the current data flow

In this section we estimate the current volume and velocity of sensing data from sport activities. In order to conduct our evaluation we measure the data flow from our sample population and we scale it to the magnitude of the real user base. We use the available dataset from year 2013.

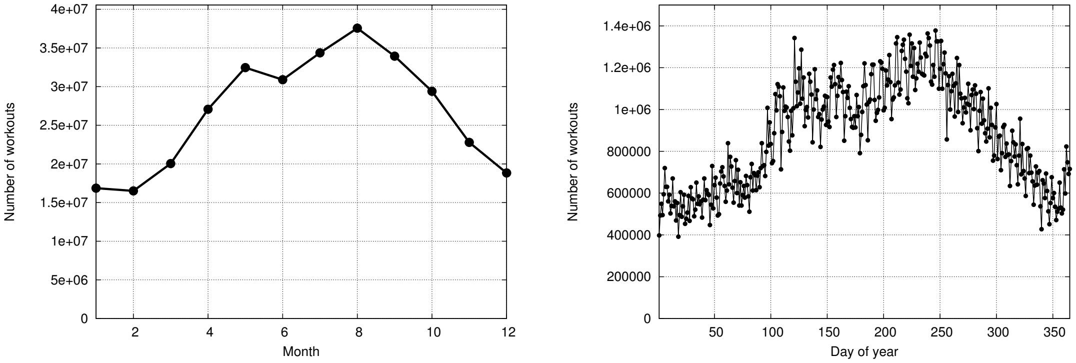
Figure 1a presents the estimated number of workouts generated per month by 30 million users. The maximum number of workouts was reached in August with 37,558,648 workouts and the minimum in February with 16,510,934 workouts. From our sample data we estimate that a single workout generates an average of 170 GPS tuples. Following this result, the estimated number of GPS tuples generated in a single month is between 2.8 and 6.3 billions.

Figure 1b shows the estimated number of workouts generated per day by 30 million users. We estimate that the peak of workout generation in a single day is close to 1.4×10^6 workouts, and that it is reached several times during periods of high activity. On the other hand, during periods of low activity, we estimate that the number of workouts generated is around 400,000 workouts per day.

In order to estimate the generation frequency of data from sport activities in a single day, we analysed the activity generated during days of either high or low activity in 2013. As our sample data provides no timezone information, we used the Google Timezone API[11] to assign a timezone to every user: we took a GPS point from a single workout and we associated it with a defined timezone. Then, we translated all workouts to the UTC timezone. Finally, we used the start time and the incremental time parameters inside every GPS tuple to calculate the number of tuples generated per second in a single day.

Figure 2a shows an estimation of the maximum insertion frequency encountered in the course of a single day. The maximum number of tuples generated per second is around 25,000. Figure 2b shows the minimum insertion frequency registered in a single day. In this case the minimum frequency was around 10,000 tuples/second.

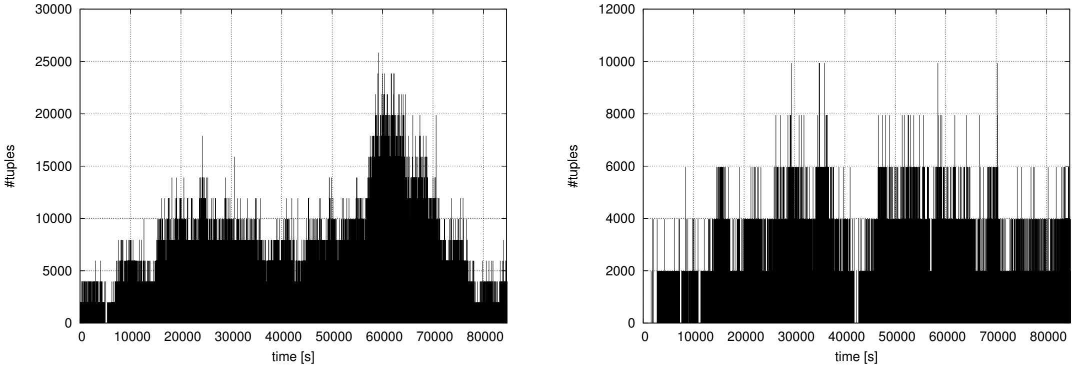
From this result we estimate that, following the example of the Endomondo healthcare sport service, a representative peak of workload for a big data stream service is around 25, 000 tuples/second.



(a) Estimation of the number of workouts generated per month by 30 million users

(b) Estimation of the number workouts generated per day by 30 million users

Fig. 1: Number of generated workouts



(a) Estimation of the maximum GPS data generation frequency registered in a single day by 30 million of users

(b) Estimation of the minimum GPS data generation frequency registered in a single day by 30 million of users

Fig. 2: GPS data generation frequency

4. Big data steps and challenges

The Endomondo application is not a big data application. Nevertheless, if we consider all of the available sport-tracking applications, the amount of tuples generated per second qualifies as a big data issue. Moreover, if we consider several applications from the IoT (or the whole IoT), then we face a huge volume of data with massive flows.

In this section we discuss the main big data challenges associated with the design and implementation of a big data service for real time processing of healthcare sensor data or other IoT applications. These challenges are present across all the big data lifecycle[3,12,13].

1. **Data acquisition**. The first step is to be able to acquire and filter the massive input stream generated by millions of sources from the IoT at an application-defined frequency. A key challenge here is to define online filters in order to discard redundant data without loss of useful information. These filters can be defined directly at the

source level, or at a higher level. For instance, in the context of healthcare services when a jogger stops to take a rest her sensor reads the same value at regular intervals. These values could be locally filtered in order to compress the input data set.
We showed that the input workload is continuous but that the flow rate varies over time. A key challenge is to design and implement a scalable way of supporting a variable number of connected objects in order to handle peaks of workload.

2. **Data cleaning**. Sensor data from smartphones is inherently erroneous and uncertain. The main factors are battery life, imprecision, and transmission failures. This problem is especially challenging when we consider stream processing. For instance, a smartphone can exhaust its battery life in the middle of the route or its GPS sensor can position it outside the route, which corrupts the resulting GPS trace. Addressing this problem requires detection and correction of this kind of data by performing *online data cleaning*.
3. **Data integration, aggregation and representation**. The integration of data streams coming from different applications allow to extract valuable information. A key challenge here is to define a common data representation that enables data aggregation from heterogeneous data streams.
The massive volume of sensor data coming from different applications of the IoT raises several issues in terms of data management. The main challenge here is how to efficiently organise this kind of data to extract value. The spatio-temporal properties of data must be considered. Another main challenge consists in implementing some early events detection, avoiding the latency of an off-line analysis.
4. **Query Processing, data Modeling, and analysis**. The three previous steps prepare the data and the storage structure for query processing and analysis. Alongside storage, the structure must also handle query processing and data mining in order to enhance knowledge discovery inside the dataset. For instance, in order to detect zones with high user concentration, real time sport activity processing requires online clustering of sensor data.
5. **Interpretation**. One of the key challenges is to start some data interpretation before the off-line processing step. An on-line (real-time) analysis of some events allows a better reactivity to local events, when the off-line analysis allows a more global view as well as making correlations among events at a large scale.

All these phases present many challenges that cover the whole big data process.

1. **Scalability**. This is one of the main challenges of big data and it is present across the whole big data pipeline presented above. We measured that the Endomondo healthcare service alone generates a peak data flow of $25,000$ tuples per second although it constitutes an incredibly small part of the whole IoT ecosystem. When we consider the integration of streams coming from all healthcare sport services with other IoT applications such as GPS sensors inside cars or air pollution sensors, the data flow can easily reach up to millions of tuples per second. Centralized servers cannot process flows of this magnitude in real time. Thus, the main challenge is to build a distributed system where every node has a local view of the data flow (i.e, every node performs the big data steps presented above with a small number of sources compared to the whole data flow). These local views must then be aggregated in order to build a global view of the data with an off-line analysis.
2. **Heterogeneity and incompleteness**. Analysis algorithms work well with structured data. However, the IoT ecosystem generates heterogeneous data flows coming from different types of applications and devices. Therefore, the main challenge here is to integrate and structure massive and heterogeneous data flows coming from the IoT to prepare their analysis in real time. For instance, streams from cyclists can be crosschecked with sensor data from cars to allow an early detection of potential accident liabilities when organizing cycling routes in cities.
3. **Timeliness**. Speed in big data is important in both input and output. The input is represented by a huge dataset coming from multiples sources that must be processed and structured for analysis. The output is represented by results of analysis or queries over the dataset. The main challenge here is how to implement a distributed architecture able to aggregate local views of data inside every node into a single global view of results with minimal communication latency between nodes.
4. **Privacy**. People generate and share personal data that are not always protected. According to a study of the IDC, only about half the information that needs protection effectively has protection[14]. Data generated from healthcare sport services contains sensitive personal information: for example, the starting point of a route could

reveal the user's home address). A key challenge here is to propose techniques that protect this kind of data before its analysis.

5. Related Work

The increasing popularity of healthcare sport services has recently attracted the research community to understanding and analyzing this kind of data. An initial discussion about the secondary use of sensor data and the emergence of a conceptual architecture is provided in[15]. The main focus of current research in sports-oriented data is in offline processing to understand its spatial distribution. A recent work[16] proposes an offline technique in order to identify popular sports areas in individual cities by using sensor data gathered from the Nokia Sport Tracker service. Another study[17] extracts and analyses sensor data from the MapMyFitness sport tracker server in order to describe the use of sports related facilities in Winston-Salem, USA. A full discussion about how this kind of data can be used to acquire an effective feedback on physical activity and how it can be used for global healthcare is provided in[18]. The evaluation of different architectures for continuous processing of GPS data is provided[19] . Some studies[16 17] provide general statistics regarding sports activity data sets. However, none of these studies focus on assessing the current workload generated by healthcare sport services.

6. Conclusion

This paper presents a study of a real world sport-tracking system as a preliminary analysis of the data flows generated by a typical IoT application. Our results show that the IoT requires real time data analysis services which can cope with huge amounts of data. These services raise new challenges from the Big Data point of view: we claim that fully distributed frameworks are required to achieve scalability.

We are currently working on an Internet-wide distributed architecture that fulfills the requirements given in this paper for the real time analysis of massive flows of data from the IoT.

References

1. Zaslavsky, A., Perera, C., Georgakopoulos, D.. Sensing as a service and big data. In: *Proceedings of the International Conference on Advances in Cloud Computing (ACC), Bangalore, India*. 2012, .
2. Cisco iot. Available: http://share.cisco.com/internet-of-things.html; 2014. Accessed: 2014-12-07.
3. Chen, M., Mao, S., Liu, Y.. Big data: A survey. *Mobile Networks and Applications* 2014;**19**(2):171–209.
4. Sport tracker dataset [online]. Available: https://cloud.lip6.fr/public.php?service=files&t=f095c275eba6fb05d56ed79aeb0f5bcf; ????
5. Endomondo sports tracking [online]. Available: http://www.endomondo.com; 2014. Accessed: 2014-12-07.
6. Nike+ sports tracking [online]. Available: http://nikeplus.nike.com; 2014. Accessed: 2014-12-07.
7. Runkeeper sports tracking [online]. Available: http://runkeeper.com; 2014. Accessed: 2014-12-07.
8. Runtastic sports tracking [online]. Available: https://www.runtastic.com/; 2014. Accessed: 2014-12-07.
9. Endomondo blog [online]. Available: http://blog.endomondo.com/press/; 2014. Accessed: 2014-12-07.
10. Endomondo infographics [online]. Available: http://blog.endomondo.com/tag/infographic/; 2014. Accessed: 2014-12-07.
11. Google timezone api [online]. Available: https://developers.google.com/maps/documentation/timezone/; 2014. Accessed: 2014-12-07.
12. Labrinidis, A., Jagadish, H.. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment* 2012;**5**(12):2032–2033.
13. Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., et al. Big data and its technical challenges. *Communications of the ACM* 2014;**57**(7):86–94.
14. Gantz, J., Reinsel, D.. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the Future* 2012;.
15. Clarke, A., Steele, R.. How personal fitness data can be re-used by smart cities. In: *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011 Seventh International Conference on*. IEEE; 2011, p. 395–400.
16. Ferrari, L., Mamei, M.. Identifying and understanding urban sport areas using nokia sports tracker. *Pervasive and Mobile Computing* 2013; **9**(5):616–628.
17. Hirsch, J.A., James, P., Robinson, J.R., Eastman, K.M., Conley, K.D., Evenson, K.R., et al. Using mapmyfitness to place physical activity into neighborhood context. *Frontiers in public health* 2014;**2**.
18. Atallah, L., Lo, B., Yang, G.Z.. Can pervasive sensing address current challenges in global healthcare? *Journal of epidemiology and global health* 2012;**2**(1):1–13.
19. Duarte, S., Navalho, D., Ferreira, H., Preguiça, N.. Scalable data processing for community sensing applications. *Mobile Networks and Applications* 2013;**18**(3):357–372.