Medium Data Toolkit - A Case study on Smart Street Sensor Project

Balamurugan Soundararaj*[1], James Cheshire†[1] and Paul Longley‡[1]

[1]Department of Geography, University College London

January 31, 2019

Summary

Big data and its analysis promise huge benefits but also introduces numerous challenges. Moreover, Since all data are not big data, rather than directly adopting advanced big data tools and methods, it is imperative to understand the nature of the dataset and devise a toolkit uniquely suitable for it. This paper examines the size and complexity of a large, national level, sensor based dataset on pedestrian footfall and devises a toolkit suitable for its scale and complexity. Meanwhile, we also discuss the possibility of "medium data" toolkit or framework which could be useful in tackling data of similar nature.

KEYWORDS: Medium data toolkit, Smart Street Sensor, Not so big data.

1 Introduction

Developments in information technology, mobile communications and internet of things have made large assemblages of data readily available (Longley et al., 2018). These big data and their analysis promises huge benefits in terms of value realisation, cost reduction and new insights but when tackling these data at such scale, we also encounter equal amount of challenges in terms of added complexity and cost (Kitchin, 2013). Moreover not all large datasets are sufficiently "big" in all dimensions. This makes it imperative that, rather than directly moving to advanced big data methods and tools, we understand the nature of such datasets and choose suitable methods and tools to tackle them. We also observe that in the discipline of geography the existing methods and tools are already developed with large scale data (Miller and Goodchild, 2015). This along with improvements in computing hardware has made processing such datasets possible without major changes in methodology. At these scales, the challenge usually arises from the speed at which the data is generated and lack of structure. In this context, our approach needs to be finding the

*s.bala@ucl.ac.uk

†james.cheshire@ucl.ac.uk

‡p.longley@ucl.ac.uk

right tool for the right problems posed by these large datasets without losing the opportunity to extract valuable information. In this paper we look into data generated from the Smart Street Sensor project, evaluate it with respect to its big data characteristics and devise a toolkit which is best suited for it. While doing this we also make a case for a more general "medium data" toolkit or framework which uses well known tools and methods from computer science to solve the challenges.

2 Big Data Challenges

The first and foremost challenge in big data is its definition which can vary widely based on the discipline and perspective. Data also have various dimensions in which they can exhibit big data properties. Though it is defined in contrast to traditional data in general usage, it could be formally defined in three dimensions of volume, velocity and variety - the three Vs of big data (Laney, 2001). These can been extended to include veracity and visualisation to give us a comprehensive framework (Li et al., 2016; Gandomi and Haider, 2015). The second set of challenges arise while trying processing it in terms of data acquisition and recording, extraction cleaning and annotation, data integration & aggregation, modelling & analysis, and finally visualisation and interpretation. We encounter the need for distributed, crowdsourced methods for data acquisition and heavily parallelised computing and functional programming concepts for extraction, cleaning and annotation. The veracity of the big data introduces significant challenges in modelling and analysis of the data. Visualisation, being an indispensable tool in exploratory analysis, needs methods to maintain legibility and understandability. There is also the challenge of interoperability, which creates the need for consistent standards. We also have management challenges to consider such as privacy and security, data governance and ownership, data and information sharing, and cost. These challenges creates requirements for compliance with ethics and legislation, strict security by anonymisation and clear ownership model. Finally the are immense costs associated with collecting, moving, storing data at such scales. These challenges and the resulting requirements are unique to every dataset and hence it is imperative that we evaluate the data to understand its specific challenge.

3 Smart Street Sensor Project

The Smart Street Sensor project is one of the most comprehensive studies carried out on consumer volume and characteristics in retail areas across United Kingdom. The data is generated through a network sensors installed at shop-fronts around 1151 locations across 107 cities in United Kingdom as of Aug 2018. The sensors passively capture a series of public signals known as probe requests generated by WiFi enable devices. These 'probe requests' are then anonymised and sent to a central data store where they are processed into an estimated footfall for the corresponding locations. In base data collected each record is an individual probe request and when processed the record is a 5 minute The project was started on July 2015 and have been continuously collecting data for the past 3.5 years. The aim of the project is to provide value to researchers, occupiers, landlords, local authorities, investors and consumers within the retail industry. This large and unstructured dataset can be used not only to detect retail activity but also to measure all the activity around the sensors and can be linked to transport, work zones and demographic data, etc., to produce, for example, novel functional areas classifications.

4 How 'big' is the Smart Street Sensor Data?

In our aim to understand the smart street sensor data, the first question we want answered is that if it is big data? We look into each dimensions of the data and try to see if the data is "big" in the corresponding dimension. In terms of volume the base data from smart street sensors are probe requests at each location generated by the Wi-Fi enabled devices present around the area. This has generated around 2-3 TB of data on disk when encoded as text. Since it is beyond the scope of general desktop computing, but still could be processed with appropriate hardware, we can say that the data is "medium" in terms of the volume. In terms of velocity, we generate data at five minute intervals at the rate of 3-4MB which amounts to 1-2 GB per day. Again, compared to slow datasets such as census and fast datasets such as social media, this is a "medium" scale data in terms of velocity. Being a result of IEEE (IEEE, 2016) standard, the data shows almost no variety at all. It is in terms of veracity where the data shows true big data properties. Since the data collection is carried out in a passive unstructured manner, the resulting dataset is highly unstructured. There are numerous uncertainties in the dataset where the mobile devices randomise their identity regularly, the sensors being installed, uninstalled and failing at irregular intervals. Moreover the method being wireless has also very site specific spatial and temporal uncertainty. Though the data is not immensely complex to visualise, being a continuous study with a long time dimension, it does pose challenges in terms of being able to legibly conduct exploratory analysis. To summarise, As shown in Figure 1, the data generated by Smart Street Sensor can be confidently categorised as "medium" in most of the dimensions, except for its veracity. Though it cannot be tackled with traditional tools and methods, it doesn't need advanced big data tools with a focus on the volume and velocity aspects.

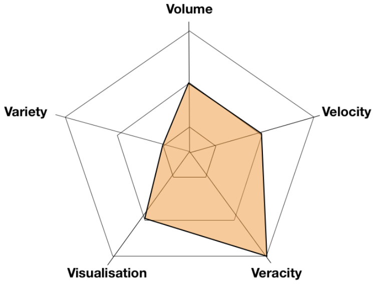

Figure 1: The footfall data from Smart Street Sensor is truly "big" only on the veracity dimension. Otherwise it is mainly a medium sized data.

5 A Medium Data Toolkit

The next towards devising a bespoke toolkit for the dataset is to conduct a survey of methods and tools available saving the above challenges. We looked at tools that can be used in terms of data collection, storage, processing and visualisation. In each aspect we looked at a spectrum of tools ranging from traditional tools to advanced big data tools and tried to choose the tool which is best suited for the challenges posed by the data. In terms of data collection a general purpose single board computer with software level collection is suitable for our purposes since traditional structured manual count is not possible at this scale and developing a micro controller based specialised IOT system might be expensive in terms of cost and time to develop. In terms of data storage, the tools range from traditional filesystem based data store to high performance distributed systems such as HDFS with database systems in the middle. Since the base data shows no variety and the processed data is very small, we chose a dual approach where the base data is stored in a filesystem and processed data in a relational database. This gives best of both worlds while avoiding the complexity of a HDFS cluster.

There have been a lot of advancements in recent years in terms of tools for processing big data. After ruling out traditional desktop computing for our dataset we looked into the tools such as MapReduce, Spark etc. We ruled them out since they were designed for data around a petabyte and upwards. They are also heavily optimised for large clusters. Though we would benefit from the parallelisation offered by these tools we do not need the complexity of the clusters. We solved this challenge using a combination of standard Unix tools such as grep, sed, awk, jq, gnupg along with high level languages such as R and Python, which are combined together with a scripting language such as bash with text being the format of data transfer between them. Finally we introduced parallelisation in the toolkit using GNU-Parallel (Tange, 2011). This toolkit, for a normal word count problem, give us a throughput of 540MB per minute without parallelisation and with parallelisation this can be improved to 2.5GB per minute. The visualisation aspect is done through tools such as R, d3 and Leftlet. The toolkit is outlined in Figure 2.

This “medium data” toolkit, consisting of standard, open source, free tools enables us to collect, store and process the Smart Street Sensor data with a throughput of 0.3GB per minute resulting in a highly granular footfall data at speeds close to being realtime.

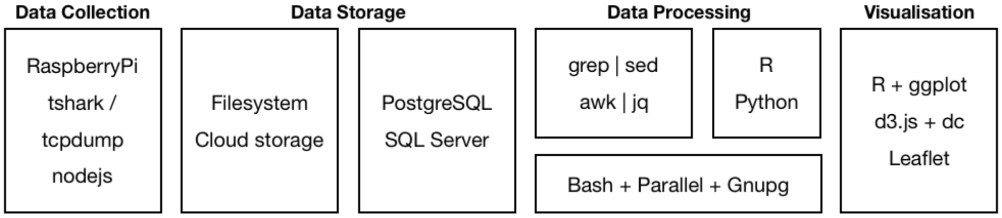

Figure 2: The ”Medium data” toolkit devised to be used with the Smart Street Sensor data

6 Acknowledgements

The data for this research has been provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC 109, ES/L011840/1; ES/L011891/1

7 Biography

Balamurgan Soundararaj is a PhD student at Dept. of Geography, University College London. He is currently working on measuring footfall in urban areas using WiFi based sensors and his research interests are Spatial analysis, Geo-informatics and Smart City Analytics.

References

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.

IEEE (2016). IEEE standard for information technology-telecommunications and information exchange between systems local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pages 1–3534.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography*, 3(3):262–267.

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., et al. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115:119–133.

Longley, P., Cheshire, J., and Singleton, A. (2018). *Consumer Data Research.* UCL Press.

Miller, H. J. and Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4):449–461.

Tange, O. (2011). Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47.