Big data coming soon ...... to an NSI near you

John Dunne

Central Statistics Office (CSO), Ireland John.Dunne@cso.ie

Big data is beginning to be explored and exploited to inform policy making. However these new data sources provide significant challenges for the traditional data collection and processing environments of the typical National Statistical Institute. This paper considers mobile phone, electronic payments and energy utility data to explore new methods and solutions to the challenges faced. Among the solutions and methods considered are enhancements to the existing technical infrastructure, outsourcing (where the data processing requirement is large), downsampling techniques to make Big data small enough to manage. The focus of the paper is on those data sources that are highly structured and typically standardised across borders as these are the sources that will provide the more immediate benefits to NSIs in informing policy – with mobile phone and credit card data informing tourism policy being the obvious example. The objective of the paper is to explore some of the challenges that may be encountered when working with Big data for Official statistics.

Key Words: data collection and processing, outsourcing, downsampling

Background

CSO, Ireland is the National Statistical Institute (NSI) for a country of approximately 4.6m persons. Like many NSIs in the current economic environment it is challenged with doing a lot more with less. New data sources and in particular Big data[1] sources have the potential to deliver significant value in delivering a lot more with less. In particular, Big data sources that are based on some systematic monitoring or recording of transactions for defined unit of observation within a defined population appears to offer the more obvious benefits for Official statistics. Examples of such sources are provided in Table 1.

While it is acknowledged further exploration of these data sources is necessary to explore calibration and estimation methods for currently defined Official statistics (as with administrative data sources from Public authorities), the possibilities and benefits of generating new statistics directly from these sources should not be ignored, particularly where a high correlation can be identified with the statistics that users want. Users however need to be aware of the pitfalls of using alternative but highly correlated information when using such statistics. Does the change reflect a change in correlation or a change in the statistic they are interested in?

[1] Big data has been defined in terms of “high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making” , UNECE HLG (2013)

Table 1 Example Big data sources with summary descriptions and possible benefits

Data source	**Notes (guesstimates based on of 4.6m persons living in 1.6m households)**
1) Call Detail Records for mobile phone roaming customers in Ireland	Guesstimate of 5 million records per month, for approximately 0.5 million mobile phone users from abroad visiting Ireland. Obvious benefit, region visited by Country of Residence in Tourism statistics. Calibration with existing surveys other data sources. Small number of data providers (mobile operators) with highly formatted data. Unique identifier for each mobile phone across operators. Relative low volume, low but variable velocity.
2) Electricity consumption data	Guesstimate of 1.5 million to 2.2 billion records per month depending on whether data relates to monthly consumption or smart metering (half hourly consumption) figures. Benefits to the development of a household register and Population estimates. Possibly single data provider of highly structured data for approximately 1.5 million residential Electricity accounts/meters. Unique identifier for each account meter. Calibration possible through household surveys. Low to high volumes, low to medium and generally constant velocity.
3) Mobile position data	Guesstimate of 4 billion records per day. Possible benefits for mobility/ commuting type statistics. Small number of data providers with highly formatted data on approximately 8m phones with a unique identifier. High volume. High and generally constant velocity.
4) Electronic payments / transaction data	Guesstimate of 80 million records per day. Potential benefits to multiple statistical sources. Small number of data providers with highly formatted data on approximately 8 million accounts with a unique ID. High volume. High and variable velocity.

However, for CSO like many NSIs, these new data sources will significantly challenge the existing data collection and processing infrastructure. In fact, these data sources will require a change in mindset (and infrastructure with respect to how NSIs collect and process data. This paper explores some of the options available to overcome potential bottle necks and constraints with different parts of the statistical value chain that deal with data transmission/collection and processing of data. The options explored include enhancing the existing IT infrastructure, outsourcing and downsampling. The overall goal is to minimise processing overhead without compromising the outputs beyond an acceptable level.

Big data processing

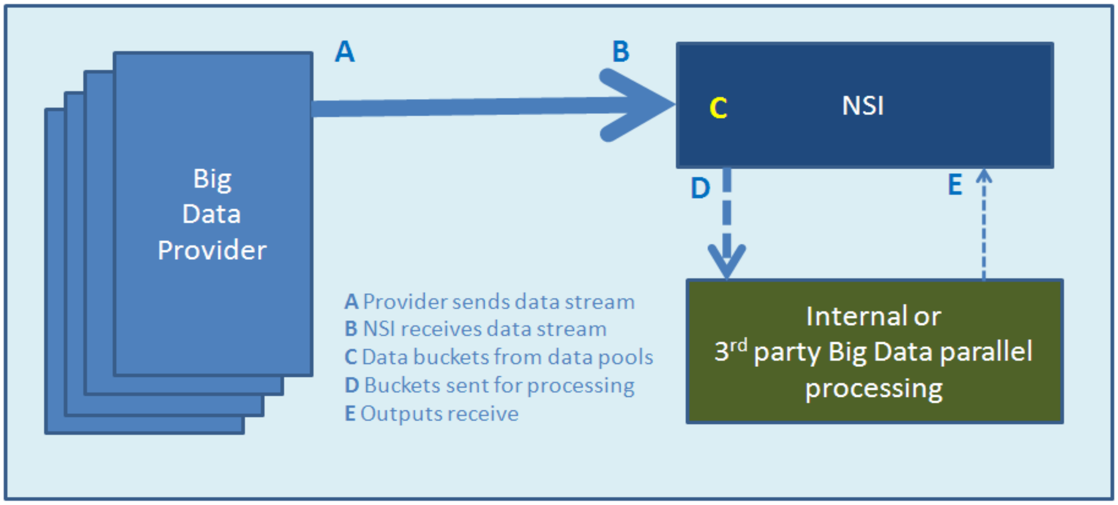
The typical solution to processing big data streams is to organise it into a large number of buckets or pools of data that can be processed in parallel using different nodes, such that, the outputs of that processing are easily manageable in the traditional processing environment. The outputs from the parallel processing can be considered as intermediate outputs that can act as inputs to other processes[2]. Basic parallel processing tasks include coding and number crunching such as aggregation. The Big data processing mindset needs to include the idea of single pass thru of data prior to organising into buckets, how to efficiently and effectively distribute data into buckets, what outputs should the parallel processing

[2] These other processes may also be Big data type processing tasks.

be focussed on, the efficient way to achieve these outputs, the capacity and number of the buckets available, processing capacities[3] and finally if there is a need to retain the original data streams once processed.

In considering what outputs are required from each bucket a good starting point are counts, sums, sum of squares, sum of cross products and min/max figures for key data groupings as these are the building blocks of many statistics. These building blocks require only a single pass through of the data thus eliminating the need to retain all data in memory.

Fig 1 Big data flows with different data flow points labelled A, B, C, D and E


A
B
C
NSI
Big Data Provider
D
E
A Provider sends data stream
B NSI receives data stream
C Data buckets from data pools
D Buckets sent for processing
E Outputs receive
Internal or 3rd party Big Data parallel processing

Enhancing existing IT infrastructure and capabilities

Data source 1 (CDR of tourists) from table 1 is identified as a good candidate project as could possibly be run with existing infrastructure and tools alongside the introduction of Big data methodology and tools. Therefore there is the comfort of parallel running with old and new technologies. A small low investment in new infrastructure to facilitate distributed computing/ parallel processing is envisaged as being sufficient for this type of project. Flow point C, where the data streamed is pooled and allocated to buckets is a critical point in terms of efficiency and effectiveness of processing. In this project it is important that all records from the same tourist get allocated to the same bucket, such that, parallel processing can be used to reduce the data to a single summary record per tourist for further processing. This can be achieved by ensuring the algorithm or hash function[4] for allocating data records is based on the mobile phone identifier in the data stream. It is, therefore, important that Big data flows contain the original identifier (or at least a common Protected Identifier Key) such that the same unit can be identified across data providers.

Furthermore, if Big data is to contribute to Official statistics, the author considers it strategically important that NSIs undertake a Big data project similar in nature to the one described here to acquire an understanding of the capabilities needed for exploiting Big data. Any IT infrastructure

[3] Number of processing nodes and memory size will be key determinants

[4] Hash functions are typically used in Big data techniques to allocate data records to buckets

enhancements implemented should also consider scalability and standardisation of methods and tools among other NSIs.

However if the volume and velocity of the data is significantly more than as described for data source (1), consideration needs to be given to the cost benefit of further enhancing the IT infrastructure once an understanding Big data processing has been developed. The key bottle neck points with respect to volume and velocity occur at point B, the capacity of the NSI to receive the data (bandwidth), at point C, the capacity of the NSI to catalogue and organise the data for Big data processing environment and at points D and E, the capacity of the Big data processing environment to process the buckets of data coming from C in a sufficiently timely manner. Two possible options for consideration of Big data processing are outsourcing and or downsampling.

Given the volumes of data involved, consideration should be given to whether it is necessary to retain data in any form other than after point E where the requirement for Big data processing techniques are no longer required. If there is a requirement to store the raw data, then some options that could be considered to reduce the volume of data to be stored include sampling or simply retaining a moving window of data (say the most recent 6 weeks of data).

Outsourcing

Outsourcing is considered at the commodity level only where processing capability is provided by a 3rd party[5] or co-op of Government organisations providing a private cloud. This maybe a viable option to enhance processing capacity however there may exist a constraint at point D, the velocity at which the buckets of data can be delivered from point C to the processing environment.

The reputation of the NSI in using a 3rd party processor and the risk of disclosure are significant considerations. Data masking techniques may be a consideration at data point C prior to allocating the data to a third party processor. Data masking techniques, if considered a requirement, can be a combination of some or all of the following

- Use of a Protected Identifier Key[6] (PIK) and removal of identifiable information
- Mapping each key grouping of data records to data buckets and removing that information prior to transmission for processing to the 3rd party (once the mapping is known the key groupings can be re-established at point E for each bucket once processing is complete)
- Linear mapping of numeric data (y=a+bx) for each bucket (and or numeric variable combination) prior to shipping. In a similar way once the linear mapping is known for each bucket many of the key statistics for x can be easily derived from those statistics for y once the mapping is known (ie., N, SUM, Sum of Squares, Sum of Cross product and Min Max).
- Adding buckets of redundant data to mask the volume of data being processed if volume is considered sensitive. Outputs from these buckets can then be excluded from final figures once these buckets can be properly identified.

If, however, the velocity or volume of data is too great to process in a timely manner then there may exist a requirement for downsampling to make the data volumes more manageable.

[5] See Elastic Compute Cloud (EC2) from Amazon at http://aws.amazon.com/ec2/ for an example of processing power being offered as a commodity. There are other providers of this type of service.

[6] Protected Identfier Key serves the purpose of hiding the original identifier while preserving the linking capabilities of the original identifier over time and across the sources it is deployed on.

Downsampling

Downsampling, a term used for sub sampling in signal processing, and used here to describe the activity of sampling data streams. The purpose of downsizing the volume of data for processing is to ensure the volumes and velocity can be managed within the given constraints (i.e., data reduction). The downsampling requirement is driven by the capacity constraints at flow points B, C and D where the NSI receives the data, pools it for allocating to buckets prior to sending buckets of data for processing. High volume and high velocity data such as data sources 2 and 3 are good candidates for the consideration of downsampling techniques. High volume, high velocity data such as that described in data source 4 (electronic payments data) is also considered a good candidate, however, this data source has the added complication that the velocity of the data can be highly variable (think about spending patterns on Christmas Eve).

The critical consideration with respect to a downsampling scheme is to ensure that the sampling is undertaken on the units that the records relate to (mobile phones[7], electricity meters, bank accounts) rather than the records themselves. An effective downsampling rate can be estimated as less than ***c/f*** where ***c*** is the minimum processing rate over each of the flow points B, C and D and ***f*** is the unconstrained velocity at which the data stream arrives to flow point B from each of the providers (flow point A). The sampling algorithm is based on allocating the relevant units in a uniform and random way to each of say ***B*** buckets such that all records related to a specific unit are contained in the same bucket and then subsequently processing not more than the first ***(c/f) B*** buckets. Note at the same time as allocation of records to buckets occur record based summary statistics (N, Sum, Sum of Squares, etc) can be compiled in an efficient manner for each bucket. Note if there are a large number of units around which the records are based then it is not feasible to collect definitive summary statistics (even the actual number of different units) around the units as this would require the significant information to be held in memory[8]. However, the information is available to extrapolate to the full data flow from the sample.

Note that if the flow rate ***f*** increases, the allocation to the ***B*** buckets happens faster, however, the proportion of buckets that are processed falls as ***c/f*** also falls.

The concept of stratification can also be introduced with the idea of separating streams in a systematic way. This is best illustrated by considering data source 3 mobile positioning data. There may exist a higher premium on tracking non domestic mobile phones (tourism statistics) in mobile positioning data streams. If each record contains an indicator of country of home network (typically a 2 letter code), this can then be used to separate streams and apply different sampling rates in each stream provided that constraint ***c*** is not violated.

If a down sampling scheme is to be implemented the optimal location to implement such a schema (from a transmission and processing perspective) is at flow point A prior to the transmission of data from the providers. This would ensure that bandwidth going from flow point A to flow point B is not used up unnecessarily. Significant co-operation would be required from the providers for such a co-ordinated down-sampling schema at data flow point A.

[7] It should be noted that with high velocity (say every 3 minutes) sensor data such as mobile positioning data, two stage sampling would be sensible whereby a sampling scheme based on position and time is taken for each mobile phone ID sampled.

[8] Although there exists tools and infrastructure that are designed for processing large amounts of data in memory.

Concluding Remarks

New data sources in the form of 'Big data' are coming to Official statistics. In the view of the author, the first wave of 'Big data' will include sensor type data and transactional type data. The underlying populations with sensor and transaction type data can usually be defined, in the way that the underlying population in administrative data sources can be defined. The volume and velocity of these data sources will challenge NSIs. The paper also outlined some solutions (outsourcing and downsampling) that can contribute to the NSI toolbox in meeting these challenges.

An opportunity for International statistics also exists with the data types explored in this paper. By their very nature, the data sources will be standardised (or easily standardised) across organisations, states and sources. For example, mobile companies share data across borders for billing purposes.

The Official statistics community is open and collaborative by nature. Traditionally, collaboration tends to happen in an organic or unstructured manner; pockets of innovation or best practice are identified and typically replicated in an organic way from organisation to organisation to improve or enhance existing systems. However this is changing. At present there exists an initiative[9] in the international community (HLG-BAS) to lead on the Modernisation of Official Statistics. There is a nexus for high level collaboration.

Big data and official statistics offers a 'green field' site for development. There is a significant opportunity for collaboration in this field in a more systematic and co-ordinated way across statistical organisation. There exists an opportunity to consider how to draw the big picture without the organisational, legal, technical constraints associated with existing systems. New and radical approaches are possible, for example, is there a place for a trusted third party to act on behalf of a number NSIs (or other public authorities) with respect to processing of Big data. Can a standard platform and toolkit (including sampling and other methodologies) be developed and prescribed.

NSIs, for their part, need to position themselves to engage with Big data. Activities that would facilitate this engagement include asserting the legal right of access for Official Statistics where it exists, undertaking pilot projects, partnering with universities to explore methodologies and other training activities that develop the relevant competencies and methodologies. NSIs also need to share their thoughts and experiences and engage with the broader statistical and academic communities.

References

Rajaraman A, Ullman J, (2011) "Mining of Massive Datasets" Cambridge Press

UNECE HLG (2013) "WHAT DOES 'BIG DATA' MEAN FOR OFFICIAL STATISTICS?", http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614 accessed on 3rd April 2013.

[9] http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Statistical+Production+and+Services