Big Data for Smart cities: How do we go from Open Data to Big Data for Smart cities ..

I have been involved in working with Smart cities through the World Smart Capital program and last month, we announced the Big Data for Smart Cities conference.

In many cases, Smart cities are linked to the idea of 'data'. For instance, Open data and sensor data provide the foundation for making cities smarter by enabling new services and acting as a feedback loop for improving existing services. However, data itself does not constitute value. Indeed many governments are adopting Open data initiatives as a way to reduce cost – which is not the ideal motivation for many developers (who are expected to build new services for free)

In a nutshell, for cities - there are two ways to capture the value from data:

a) At a grassroots level i.e. by empowering citizens and hackers to create Long Tail apps (here we use the word 'apps' loosely) where a specific dataset or a collection of datasets are used to solve a specific problem. This was the idea behind Apps for Smart cities conference. This is the avenue for the 'hacker' – ***and I also include citizens here i.e. people who use grassroots technology and Open data to create new services to solve a specific problem***

b) The second avenue is through analytics. Here, as data from many sources such as sensors is collated, there is LOTs of data **but often no real problem defined** in advance. This is the realm of the Data scientist and Big Data which aims to make sense out of the data

In this article, leading up to the Big Data for Smart cities conference – I will elaborate on the **value proposition for Big Data and Smart cities** (i.e. the latter case) and also see how these ideas fit together

ROADMAP

In this article, I take the following approach:

a) I first look at Big data principles
b) Then, I look at comparison between Big Data and Data Warehouses through an article from @Barry Devlin
c) I next explore the idea of: 'How do we go from Open data to Big data?' I adapt the above model to cities with the idea of **Big data for Smart cities** – augmenting Open Government Data and sensor data with Social Media
d) I conclude with the value proposition i.e. the equivalent of 'nappies and beer' for Smart cities

BIG DATA – PRINCIPLES AND DEFINITIONS

Big Data is in the news but it is a fairly nebulous concept. Wikipedia's definition, as of December 2010, says: "*Big Data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes in a single data set*.". Thus, with references to 'commonly used software tools' and 'tolerable elapsed time' – this definition is a moving goalpost. Indeed, it is not a definition – but more an aspiration of a desire to be able to gain insights from massive volumes of data.

Another definition which I found useful comes from Michael Friedenberg- president and CEO of IDG Enterprise – "Simply put, it's about data sets so large - in volume, velocity and variety - that they're impossible to manage with conventional database tools" and that "Behind this explosive growth in data, of course, is the world of unstructured data.

Thus, we have **some common elements for Big Data to this aspirational definition**:
- Size of datasets
- Velocity of datasets (rate of change)
- Unstructured data sets
- The need to come back with responses to queries in reasonable time

In the book - **The Little Book of BIG DATA, 2012 Edition** by Noreen Burlingame – we have an idea of the statistics that are driving this data explosion:

- Every day of the week, we create 2.5 quintillion bytes of data is created.
- In the 11 years between 2009 and 2020, the size of the "Digital Universe" will increase 44 fold. That's a 41% increase in capacity every year.
- In addition, only 5% of this data being created is structured and the remaining 95% is largely unstructured, or at best semi-structured.
- Where does this data come from? Sensors, social media posts, pictures posted, videos posted, comments, transactions, GPS data etc.

UNDERSTANDING BIG DATA AND DATA WAREHOUSES

It is easier to contrast and understand Big Data when you compare to Data warehouses. Retailers, Telecom companies are others have always had to work with large data sets – mostly managed through data warehouses.

What has changed if companies like Telecom Operators and Retailers already had large datasets in their data warehouses?

Is Data Warehousing not 'Big data'?

Well not quite.
- Enterprise data is (mostly) about a specific enterprise
- An enterprise, no matter how big, is still small in comparison to social sites like Facebook or sites which work with social data
- Enterprise data is closed – both inbound and outbound. That's why 'Enterprise' version s of blogs, web 2.0, wikis etc never really took off. There is relatively less 'collective intelligence' within the confines of an enterprise (in comparison to the wider web)
- Enterprise data in traditional data ware houses is structured data (mostly transactional data)

Having said that, there are some similarities between Data Warehousing and Big Data.

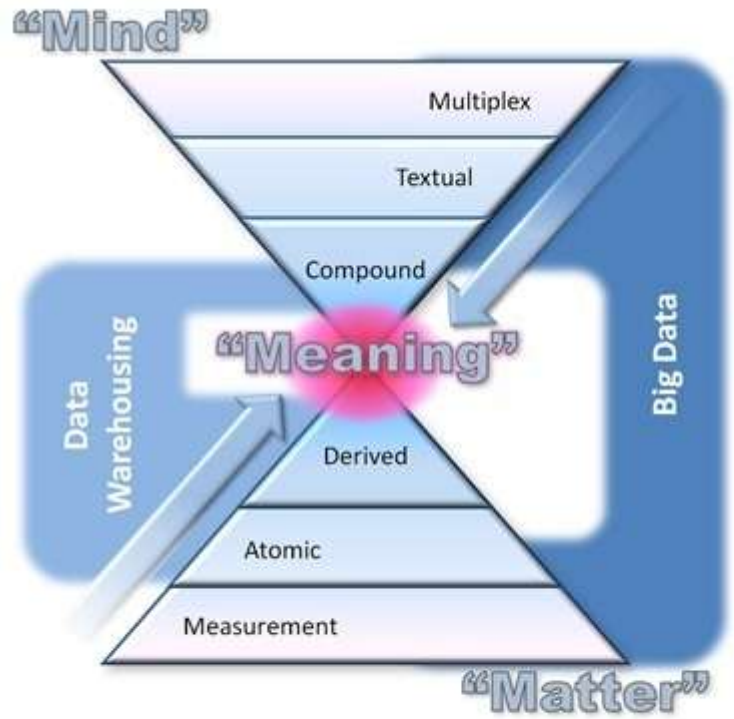
Data warehouses are a logical store of (mainly) transactional information. In some cases, the data in data warehouses is used to gain unpredictable insights from the data (as opposed to pre-defined reports)

Prior to working in mobile, I was involved in implementing large data warehouses, particularly around the Oracle database and for ERP systems. It is easier to explain the idea of Big Data in context of Data warehousing. Most people in technology are familiar with the idea of a Data Warehouse For more than 25 years, data warehousing provided the architecture for decision support systems.

Thus, when viewed as 'systems for decision support', Data warehouses are primarily concerned with specific business requirements. @Barry Devlin wrote an insightful article on the O Reilly radar

comparing data warehouses with Big Data. Big data Will data warehousing survive the advent of big data?. He proposes an approach which he calls 'mind over matter' in which he looks at the *"different types of data involved and, rather than focus on the actual data volumes, look to the scale and variety of processing required to extract implicit meaning from the raw data."*

In the figure below (source Barry Devlin/O Reilly): I summarise these ideas briefly below



source: http://radar.oreilly.com/2011/01/data-warehouse-big-data.html

- The bottom pyramid represents data gleaned primarily from the physical world, the world of matter ***measurement*** data - from sensors etc.

- ***Atomic* data** is thus comprised of physical events, meaningfully combined in the context of some human interaction. For example, a combined set of location, velocity and G-force measurements in a specific pattern and time from an automobile monitoring box may indicate an accident.

- ***Derived* data**, created through mathematical manipulation of atomic data, is generally used to create a more meaningful view of business information to humans. For example, banking transactions can be accumulated and combined to create account status and balance information.

In contrast, the information from the top of the pyramid is from the realm of the mind - information originating from the way we as humans perceive the world and interact socially within it. This is '*soft information*' — less well structured and requiring more specialized statistical and analytical processing.

- ***Multiplex* data**, image, video and audio information, often in smaller numbers of very large files and very much part of the big data scene.

- ***textual* data** — is more suited to statistical analysis and text analytics tools are widely used against big data of this type.

At the intersection of these two pyramids is **Compound data**. Compound data is a combination of hard and soft information "*a combination of hard and soft information, typically containing the structural, syntactic and model information that adds context and meaning to hard information and bridges the gap between the two categories*." Metadata is a very significant part of compound data – but compound data includes other components as well.

Compound data is the category of data of most current interest in big data. It contains much social media information — a combination of hard web log data and soft textual and multimedia data from sources such as Twitter, Facebook and so on.

In addition:

- The width of each layer in the pyramids corresponds loosely to data volumes and numbers of records in each category. And
- More importantly, the underlying reason we do data warehousing (more correctly, business intelligence, for which data warehousing is the architectural foundation) and analyze big data is essentially the same: we are searching for meaning in the data universe. And meaning resides at the conjoined apexes of the two pyramids.
- Both data warehousing and big data begin with highly detailed data, and approach its meaning by moving toward very specific insights that are represented by small data sets that the human mind can grasp.

You can see the source for more information on this approach - Big data Will data warehousing survive the advent of big data?

BIG DATA AND SMART CITIES – COMPOUND DATA AND AUGMENTING OPEN GOVERNMENT DATA WITH SOCIAL MEDIA

I think the above approach comparing Data warehousing to Big data with the idea of 'mind over matter' and compound data is accurate. ***The question is: How does this approach of 'compound data' apply to Smart cities and what is the implication for Smart city services?*** I previously discussed some ideas in the Apps for Smart cities conference manifesto. Many cities are currently co-relating Open data and Smart cities for example through Hackathons and for public institutions like the NHS . The second source of data is through sensors – for example – this was the idea behind the Apps for Smart cities event in Amsterdam

But the release and acquisition of data is just the starting point and is not 'Big data' i.e. Both Open data and sensor data are feeders to Big data.

We saw in the previous section the importance of compound data in the Big Data scenario i.e. the holy grail is to get to compound data which is at the intersection of these two pyramids i.e. Compound data is a combination of hard and soft information

So, how do we go from 'Open data to Big Data' for Smart cities?

How do we get 'compound data for Smart cities' i.e. combine hard and soft information?

We need a few more elements to avail the benefits of Big Data for example:

- The information should be held in a central place
- The information should combine social data and Smart city data (open data + sensor data)

- The problems to be solved will be complex and not known in advance (more on that later)
- We need a different skillset that of a data scientist for cities i.e. someone who understands city services, data analysis, co-relation of data etc (more on this below)
- We could be dealing with unstructured data

To achieve compound data for Smart cities, here are some thoughts/insights:

- **Open government data and social data have some specific characteristics** source Kalampokis
  Open Government Data
  - Come in various formats
  - Come in large numbers
  - Mostly are created by the public sector
  - Some are open
  - Are *objective*
  - Are used for new value-added services

  Social Media Data
  - Come in various formats
  - Come in large numbers
  - Mostly are created by Social Media users
  - Some are open
  - Are *subjective*
  - Are used for new value-added services and research (e.g. for *predictions*)

- **We need tighter integration between Open data and social media data for cities:** Andrea Di Maio of Gartner asks the question – Why do Governments Separate Open Data and Social Media Strategies? – He says: *governments provide public data and create avenues for people to engage, but do not consistently reach out to where people themselves create their own data (pictures, comments, ideas) or to online places and communities where people are willing to have conversations that may be of great relevance to policy-making as well as service improvement. When it comes to open government, agencies want to host data and communities, but do not think about being guests of somebody else's communities.* There is an asymmetry in Open data – and there is an assumption that citizens are mostly at the receiving end of Open Data. That asymmetry of Government 2.0 needs to change.

- **We need a classification scheme for Open government data:** The lack of classification scheme for open data and the integration of government and social media data has also been addressed in the position paper Augmenting Open Government Data with Social Media Data pdf : "*The idea behind the integration of government and social media data is that although they both may refer to the same real world entity, the former provides objective facts while the latter subjective thoughts and opinions. The integration of both could enable the composition of two complementary points of view of the same problem and thus enable a better and more in depth understanding of it.*"

- **There are many ways to achieve integration between Open data and Social data** – and they may be initial quick wins. For instance, you could co-relate the location attribute in tweets with a crime dataset from Data.gov.uk

- **Linked data:** Linked data could address some of the issues of co-relating social data and open data. Sir Tim Berners Lee is doing a lot of work to promote Linked data. Sir Tim

Berners-Lee proposed a five-star maturity model as follows (Berners-Lee, 2010) for Open data

1 star: publishing data on the web even in proprietary and desktop-centric formats
2 stars: publishing data in machine-readable formats such as spreadsheets documents
3 stars: publishing data in machine readable and non-proprietary formats using open standards, e.g., CSV
4 stars: publishing data using linked data principles
5 stars: linking the available data.

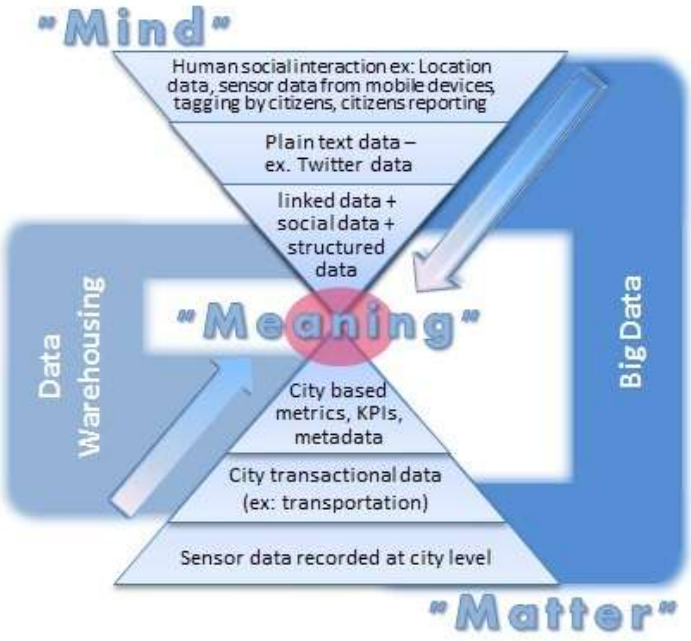
Four linked data principles as described by Sir Tim Berners-Lee (2010) are the following: (source – A classification scheme for open government data: towards linking decentralised data - Evangelos Kalampokis, Efthimios Tambouris, Konstantinos Tarabanis)

• All item should be identified using URIs
• All URIs should be dereferenceable, that is, using HTTP URIs allows looking up the item identified through the URI
• When looking up a URI it leads to more data, which is usually referred to as the follow your nose principle
• Links to other URIs should be included in order to enable the discovery of more data.

- **Categories of Open data:** To understand Open data, Social data and its co-relation for Smart cities, we have to understand categories of Open data. Surprisingly this was not easy! I finally found the below(translated from German) from Open data berlin site

Environmental data (particulate matter, CO2, pollen)
Markets (weekly, flea, Christmas markets)
events (festivals, concerts, long night of ..., sports events)
Disposal (appointment in my street, recycling centers, container sites, hazardous waste)
infrastructure (cycle paths, toilets, mailboxes, ATMs, telephones)
Traffic (construction sites, traffic jams, road closures)
transport (delays, cancellations, special trips)
opening times (libraries, museums, exhibitions)
Management (Forms, responsibilities, authorities, opening times)
consumer advice, debt counselling
Family (parental allowance, day nurseries, kindergartens)
Education (schools, community colleges, colleges and universities)
Housing (housing benefit, rent prices, real estate, land prices)
health (hospitals, pharmacies, emergency services, specialist counselling services, Blood donation)
Pets (veterinarians, animal shelter, animal care)
Control (bathing, food, restaurants, prices)
Legal (laws, regulations, guidance, arbitrator, evaluator)
Police Online (current events, investigation, crime Atlas)
City Planning (zoning, construction, transport, airports)
Population (number, regional distribution, demographics, purchasing power, Employment / unemployment, children)

Based on the above, here is an adapted diagram for Smart cities and Big Data when viewed in perspective of 'Mind and Matter'

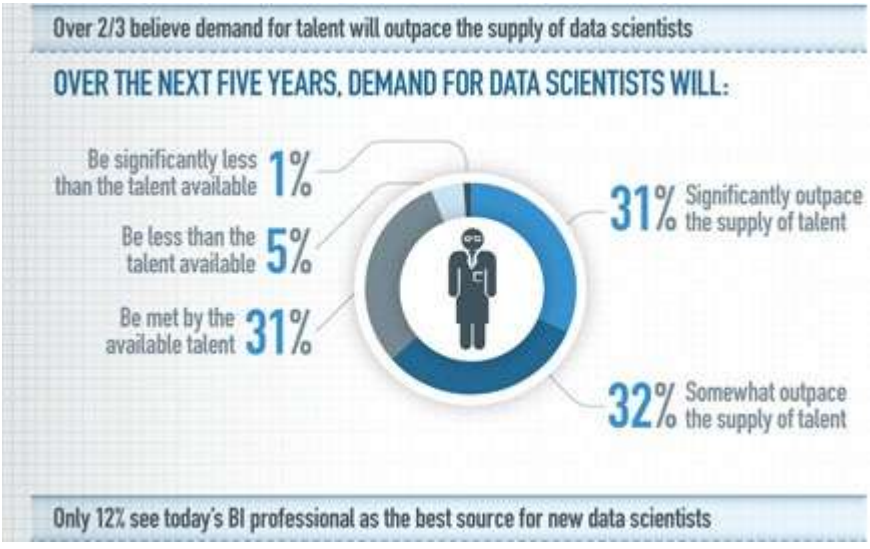


Big data for smart cities - http://www.opengardensblog.futuretext.com/archives/2012/08/big-data-for-smart-cities-for-hackers-data-scientists-and-citizens.html

DATA SCIENTSTS FOR SMART CITIES?

Who makes sense of all this data?

It is the job of the Data scientist - job widely regarded as the sexiest in the world for the next ten years(Data scientist). Based on the above, I believe that a Data scientist for the city will be a crucial role in the near future.



The Data scientist will not be the current 'Data warehouse/ Business Intelligence' expert.
They will have a different skill set (in contrast most current BI experts are domain experts in specific industries such as Retail where they are primarily working with structured data)

However, there is one important comparison between Data warehousing and Big Data.

In the world of Data Warehousing, there is the quest for the 'nappies and beer insight'

There is a story : *that a large supermarket chain, usually Wal-Mart, did an analysis of customers' buying habits and found a statistically significant correlation between purchases of beer and purchases of nappies (diapers in the US). It was theorized that the reason for this was that fathers were stopping off at Wal-Mart to buy nappies for their babies, and since they could no longer go down to the pub as often, would buy beer as well. As a result of this finding, the supermarket chain is alleged to have the nappies next to the beer, resulting in increased sales of both.*

So, the value for Big Data and Smart cities lies in finding the ' Nappies and Beer' equivalent for Smart cities – a role involving the Data scientist with a knowledge of Smart cities. It also involves working with many new technologies, ideas and large volumes of data.

This blog is already very long but from a technical perspective, here are two thoughts: (I will cover these in subsequent blogs)

a) It's hard to talk of Big Data without speaking of technologies like Hadoop see : Hadoop: What it is, how it works, and what it can do. However, Hadoop is meant for batch processing. Other variants of Hadoop for instance HOP (Hadoop Online Prototype (HOP) could have a role to play in a near Real time processing for city level services. HOP is essentially a modified MapReduce architecture that allows data to be pipelined between Operators – see this Berkeley paper on Hadoop Online Prototype

b) Secondly, there is the case of how sensor data integrates with Open Data. More specifically, the question of where processing could take place for sensor based data. Sensor data at the moment is encrypted in many cases. This means, it is not very useful as 'Open data' i.e. for example – in the potential as a mashup. This problem has been identified by companies like Pachube (now COSM) calling for Open sensor data. However, Open sensor data leads to it's own set of problems and could call for perhaps a distributed Open sensor data (as opposed to centralization of sensor data). In turn, that may call for greater local processing on the sensor / gateway. One could argue that sensor data should be decentralised with local and

distributed storage. As devices increase in processing and storage capacity, this is possible – indeed it may be needed especially to fulfil the vision of Billions of sensor devices with science fiction like possibilities (How low power can you go?)

CONTINUING THE DISCUSSION

To conclude, today, we are at the beginning of these profound changes – here's why:

- As citizens, we are only now beginning to understand the impact of Big Data on our lives – this will be more so for many of us who live in cities. The Pew Internet research foundation has an insightful report where they say that new forms of information analysis (which they call 'humanity's dashboard') are helping us be more nimble and adaptive, but they worry over humans' capacity to understand and use new tools well.
- Physicist Geoffrey West says data could save cities from outgrowing themselves. According to physicist Geoffrey West, the world's cities have what one might call a growing problem. As they grow bigger, their problems grow worse, which means it takes an ever-faster pace of innovation to keep things in check. Big data techniques might provide the answer.
- Finally, while not my regular reading, the economic collapse blog says - These 12 Hellholes Are Examples Of What The Rest Of America Will Look Like Soon. The sentiment is correct i.e. we are in the midst of fundamental change and the future of whole cities is at stake .. Cities will have to fight to attract the most productive citizens – who will gravitate to the most vibrant, well managed cities. Sadly, this will come at the expense of many cities will die.

So, the questions I leave you with are:

a) What is the role of the data scientist for Smart cities(based on the above analysis)?
b) What is the 'nappies and beer' equivalent for Smart cities?
c) How do we combine Open Data and Social media data for cities?
d) How do we integrate sensor data into the mix?
e) Do we need localised processing at sensors/gateways to manage data?
f) Will citizens get (and trust) 'humanity's dashboard'?
g) Will better management of some cities through data lead to a snowball effect at the expense of other cities?
h) Which technologies will play out here? I mentioned HOP (Hadoop Online Prototype) but that could be just the beginning. What else could play a role here?
i) Will Tim Berners Lee change the world once again through encouraging linked data?

This is a long blog but still work in progress! Happy to continue the conversation

- **If you are interested in discussing these ideas or speaking /sponsoring/hosting the event please contact me at ajit.jaokar at futuretext.com.**
- **You can join our meetup group (which we will adapt for Big Data and Smart cities http://www.meetup.com/Apps-for-Smart-Cities/).**
- **I am on the advisory board of the World Smart Capital along with some very clued on folk**
- **Following the successful Apps for Smart cities event in Amsterdam, we hope to hold events in Amsterdam, Berlin and Liverpool this year. Comments / questions – speaking at event/sponsoring event – please contact me at ajit.jaokar at futuretext.com**

By Ajit Jaokar