Big Data:

A data-driven society ?

Roberto V. Zicari

Goethe University Frankfurt

Director Big Data Lab Frankfurt

http://www.bigdata.uni-frankfurt.de

Editor ODBMS.org and ODBMS Industry Watch

www.odbms.org

roberto@zicari.de

Stanford, October 29, 2014

Big Data *slogans*

"*Big Data: The next frontier for innovation, competition, and productivity*"

(McKinsey Global Institute)

"*Data is the new gold*"

Open Data Initiative, European Commission (aim at opening up Public Sector Information).

This is Big Data.

Every day, 2.5 quintillion bytes of data are created. This data comes from digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records, cell phone GPS signals to name a few.

What Data?
BIG DATA, OPEN DATA, Linked Data.

The term ”**Big Data**" refers to large amounts of different types of data produced with high velocity from a high number of various types of sources. Handling today's highly variable and real-time datasets requires new tools and methods, such as powerful processors, software and algorithms.

“The term ”**Open Data**" refers to a subset of data, namely to data made freely available for re-use to everyone for both commercial and non-commercial purposes”.

Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods.

Another Definition of Big Data

“Big Data” *refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze*” (McKinsey Global Institute)

- This definition is Not defined in terms of data size (data sets will increase)
- Vary by sectors (ranging from a few dozen terabytes to multiple petabytes)

  1petabyte is 1,000 terabytes (TB)

Big Data:
A data-driven economy ?

- the European Commission has just adopted (July 2014) its first strategy to promote a *data-driven economy in the EU* , as a response to the European Council's conclusions of October 2013, which focused on the *digital economy*, innovation and services as drivers for growth and jobs and called for EU action to provide the right framework conditions for a single market for *big data* and *cloud computing*.

https://ec.europa.eu/digital-agenda/en/news/communication-data-driven-economy

Big Data:
a new industrial revolution?

- *Big data technology and services are expected to grow worldwide to USD 16.9 billion in 2015 at a compound annual growth rate of 40%* – about seven times that of the information and communications technology (ICT) market overall.
- A recent study predicts that in the UK alone, the number of specialist big data staff working in larger firms will increase by more than 240% over the next five years.
- This global trend holds enormous potential in various fields, ranging from *health*, *food security*, *climate* and *resource efficiency* to *energy*, *intelligent transport systems* and *smart cities*

World Digital Economy?

- "Yet the European digital economy has been slow in embracing the *data revolution* compared to the USA and also lacks comparable *industrial capability*. *Research and innovation* (R&I) *funding on data* in the EU is sub-critical and the corresponding activities are largely uncoordinated.
- There is a *shortage of data experts* able to translate technology advances into concrete business opportunities.
- The *complexity of the current legal environment* together with the insufficient access to large datasets and enabling infrastructure create entry barriers to SMEs and stifle innovation.
- As a result, there are fewer successful data companies in Europe than in the USA where large players have recognized the need to invest in tools, systems and new data-driven processes.
- However, significant new opportunities exist in a number of sectors (from health and smart factories to agriculture) where the application of these methods is still in its infancy and global dominant players have not yet emerged". (European Commission, July 2014)

What to do with Big Data?

“In general, analyzing data means better results, processes and decisions.

It helps us generate new ideas or solutions or to predict future events more accurately.

As technology advances, entire business sectors are *being reshaped by systematically building on data analytics*.”

(European Commission, July 2014)

Let`s critically review this statement….

What is Big Data supposed to create?

"***Value***" (McKinsey Global Institute):

- Creating transparencies
- Discovering needs, expose variability, improve performance
- Segmenting customers
- Replacing/supporting human decision making with automated algorithms
- Innovating new business models,products,services

How Big Data will be used?

Combining Data together is the real value for corporations:

90% corporate data

10% social media data

Sensors data just begun (e.g. smart meters)

Key basis of competition and growth for individual firms
(McKinsey Global Institute).

Big Data Search

- “The most impactful Big Data Applications will be industry- or even organization-specific, leveraging the data that the organization consumes and generates in the course of doing business. *There is no single set formula for extracting value from this data; it will depend on the application*”
- “There are many applications where simply being able to comb through large volumes of complex data from multiple sources via interactive queries can give organizations new insights about their products, customers, services, etc. *Being able to* ***combine*** *these interactive* ***data explorations*** *with some* ***analytics*** *and* ***visualization*** *can produce new insights that would otherwise be hidden.* We call this **Big Data Search.**”

-- David Gorbet (MarkLogic)

Examples of BIG DATA USE CASES

- Log Analytics
- Fraud Detection
- Social Media and Sentiment Analysis
- Risk Modeling and Management
- Energy sector
- Politics?
- Security?

Big Data

can **generate financial value**(*)

across sectors, e.g.

- Health care
- Public sector administration
- Global personal location data
- Retail
- Manufacturing

(McKinsey Global Institute)

(*)Note: *but it could be more than that!*

Big Data: What are the consequences?

- The existence of datasets, be they distributed across different locations and sources, open or restricted, and possibly including personal data that needs special protection, **poses new challenges for the underlying infrastructure.**
- **Data analytics requires a secure and trusted environment** that enables operations across different cloud and high-performance computing infrastructures, platforms and services.
- Data-driven innovation brings vast ***new job opportunities***. However, it requires **multidisciplinary teams** with highly skilled specialists in data analytics, machine learning and visualisation as well as relevant legal aspects such as data ownership, licence restrictions and data protection. The **training of data professionals** who can perform in-depth thematic analysis, exploit machine findings, derive insight from data and use them for improved decision-making is crucial.

Data-driven innovation

„The term 'data-driven innovation' (DDI) refers to the capacity of <u>businesses</u> and <u>public sector bodies</u> to make use of information from improved data analytics to develop improved services and goods that facilitate everyday life of individuals and of organisations, including SMEs.“ (European Commission)

EU's Horizon 2020

„The EU's **Horizon 2020** (H2020) and national R&I funding programmes can address relevant technical challenges:

- from data creation and actuation through networks,
- storage and communication technology to large-scale analysis,
- advanced software tools and
- cyber security.

Finally, support to stimulate sector-specific entrepreneurship and innovation is important“. (European Commission)

Limitations

- Shortage of talent necessary for organizations to take advantage of big data.
- Very few PhDs.
  - Knowledge in statistics and machine learning, data mining.
  - Managers and Analysts who make decision by *using insights from big data.*

Source: McKinsey Global Institute

Issues

(source: McKinsey Global Institute)

- Data Policies
  - e.g. storage, computing, analytical software
  - e.g.new types of analyses
- Technology and techniques
  - e.g. Privacy, security, intellectual property, liability
- Access to Data
  - e.g. integrate multiple data sources
- Industry structure
  - e.g. lack of competitive pressure in public sector

Towards a data-driven economy
(source: European Commission)

- **Availability of data and interoperability**

  Availability of good quality, reliable and interoperable datasets and enabling infrastructure

- **Improved framework conditions that facilitate value generation from datasets**

- **A range of application areas where improved big data handling can make a difference**

- **Regulatory issues**

  ->Personal data protection and consumer protection

  Data-mining

  Security

  Ownership/transfer of data

Enabling infrastructure for a data-driven economy

->Cloud computing

- **E-infrastructures and High Performance Computing**
- **Networks/ Broadband /5G**

->Internet of Things (IoT)

- **Public Data Infrastructures**

Big Data:
What are the consequences?

“Any technological or social force that reaches down to affect the majority of society`s members is bound to produce a number of controversial topics” (John Bittner, 1977)

But, what are the “true” consequences of a society being reshaped by “systematically building on data analytics” ?

Big Data: Research Challenges

1. **Data**

2. **Process**

3. **Management**

Data Challenges

- **Volume: dealing with the size of it**

  In the year 2000, **800,000 petabytes (PB)** of data stored in the world (source IBM). Expect to reach **35 zettabytes** (ZB) by 2020. Twitter generates 7+ terabytes (TB) of data every day. Facebook 10TB.

- **Variety: handling multiplicity of types, sources and formats**

  Sensors, smart devices, social collaboration technologies.

  Data is not only **structured**, but **raw**, **semi structured**, **unstructured data** from web pages, web log files (click stream data), search indexes, e-mails, documents, sensor data, etc.

Variety (cont.)

Structured Data
Semi-structured Web data
Unstructured Data

- A/B testing, sessionization, bot detection, and pathing analysis all require powerful analytics on many petabytes of **semi-structured** Web **data**.
- Sensors data: separating signal to noise ratio

Data Challenges cont.

- **Data availability** – is there data available, at all?
- **Data quality** – how good is the data? How broad is the coverage? How fine is the sampling resolution? How timely are the readings? How well understood are the sampling biases?

  Determining the quality of data sets and relevance to particular issues (i.e., is the data set making some underlying assumption that renders it biased or not informative for a particular question).

  A good process will, typically, make bad decisions if based upon bad data.

  e.g. what are the implications in, for example, a Tsunami that affects several Pacific Rim countries? If data is of high quality in one country, and poorer in another, does the Aid response skew 'unfairly' toward the well-surveyed country or toward the educated guesses being made for the poorly surveyed one? (Paul Miller)

Data Challenges cont

- **Velocity** (reacting to the flood of information in the time required by the application) *Stream computing: e.g. "*Show me all people who are *currently* living in the Bay Area flood zone"- continuosly updated by GPS data in real time. (IBM)

- **Veracity (**how can we cope with uncertainty, imprecision, missing values, misstatements or untruths?)

- **Data discovery is a huge challenge** (how to find high-quality data from the vast collections of data that are out there on the Web).

- **Combining multiple data sets**

Data Challenges cont.

- **Data comprehensiveness** – are there areas without coverage? What are the implications?

- **Personally Identifiable Information** – much of this information is about people. Can we extract enough information to help people without extracting so much as to compromise their privacy? Partly, this calls for effective industrial practices.

  Partly, it calls for effective oversight by Government. Partly – perhaps mostly – it requires a realistic reconsideration of what privacy really means. (Paul Miller)

  "*right to be forgotten*". 1,000 a day ask Google to remove search links (145,000 requests have been made in the European Union covering 497,000+ web links)

Data Challenges cont.

– **Data dogmatism** – analysis of big data can offer quite remarkable insights, but we must be wary of becoming too beholden to the numbers. Domain experts – and common sense – must continue to play a role.

e.g. *It would be worrying if the healthcare sector only responded to flu outbreaks when Google Flu Trends told them to*. (Paul Miller)

Process Challenges

The challenges with deriving insight include

- **Capturing data**,
- **Aligning data from different sources** (e.g., resolving when two objects are the same),
- **Transforming the data into a form suitable for analysis**,
- **Modeling it**, whether mathematically, or through some form of simulation,
- **Understanding the output** — visualizing and sharing the results,

(Laura Haas, IBM Research)

Management Challenges

Data Privacy, Security, and Governance.

- ensuring that data is used correctly (abiding by its intended uses and relevant laws),
- tracking how the data is used, transformed, derived, etc,
- and managing its lifecycle.

*"Many data warehouses contain sensitive data such as personal data. There are **legal and ethical concerns** with accessing such data. So the data must be secured and access controlled as well as logged for audits"* (Michael Blaha).

Big Data: Research Opportunities.

Analytics "In the Big Data era the old paradigm of shipping data to the application isn`t working any more. Rather, *the application logic must "come" to the data* or else things will break: this is counter to conventional wisdom and the established notion of strata within the database stack.

Data management "With terabytes, things are actually pretty simple -- most conventional databases scale to terabytes these days. However, try to scale to petabytes and it`s a whole different ball game."

(Florian Waas, previously at Pivotal)

Confirms **Gray`s Laws of Data Engineering:**

Take the "Analysis" to the Data!

“Objects” in Space vs. “Friends” in Facebook.

- Alex Szalay- who knows about data and astronomy, having worked from 1992 till 2008 with the **Sloan Digital Sky Survey** together with Jim Gray – wrote back in 2004:

“Astronomy is a good example of the data avalanche. It is becoming a data-rich science. The computational-Astronomers are riding the Moore’s Law curve, producing larger and larger datasets each year.” [Gray,Szalay 2004]

“Data is everywhere, never be at a single location. Not scalable, not maintainable.”–Alex Szalay

Big Data Analytics

“ *In the old world of data analysis you knew exactly which questions you wanted to asked, which drove a very predictable collection and storage model.*

In the new world of data analysis your questions are going to evolve and change over time *and as such you need to be able to collect, store and analyze data without being constrained by resources.*

”— **Werner Vogels, CTO, Amazon.com**

How to analyze?

"It can take significant exploration to find the **right model for analysis**, and the ability to iterate very quickly and "fail fast" through many (possible throwaway) models **-at scale** - is critical." (Shilpa Lawande,HP Vertica)

Faster

“As businesses get more value out of analytics, it creates a success problem - they want the data available faster, or in other words, want **real-time analytics**. And they want more people to have access to it, or in other words, high user volumes.” (Shilpa Lawande, HP Vertica)

The Beckman Database Research Self-Assessment Meeting Report October 2013

Identified Five database research areas in Big Data:

- 1. *Scalable big/fast data* infrastructures;
- 2. Coping with *diversity* in the *data management* landscape;
- 3. *End-to-end processing* and *understanding* of data;
- 4. *Cloud services*; and
- 5. Managing the diverse roles of *people* in the data life cycle.

Daniel Abadi, Rakesh Agrawal, Anastasia Ailamaki, Magdalena Balazinska, Philip A. Bernstein, Michael J. Carey, Surajit Chaudhuri, Jeffrey Dean, AnHai Doan, Michael J. Franklin, Johannes Gehrke, Laura M. Haas, Alon Y. Halevy, Joseph M. Hellerstein, Yannis E. Ioannidis, H.V. Jagadish, Donald Kossmann, Samuel Madden, Sharad Mehrotra, Tova Milo, Jeffrey F. Naughton, Raghu Ramakrishnan, Volker Markl, Christopher Olston, Beng Chin Ooi, Christopher R´e, Dan Suciu, Michael Stonebraker, Todd Walter, Jennifer Widom

Beckman Center of the National Academies of Sciences & Engineering
Irvine, CA, USA
October 14-15, 2013

Scale and performance requirements strain conventional databases.

"The problems are a matter of the **underlying architecture**. **If not built for scale from the ground-up a database will ultimately hit the wall** -- this is what makes it so difficult for the established vendors to play in this space because **you cannot simply retrofit a 20+ year-old architecture to become a distributed MPP database over night**." **(**Florian Waas, previously Pivotal**)**

Scalability

Scalability has three aspects:

- Data Volume,
- Hardware Size, and
- Concurrency.

Seamless integration

“Instead of stand-alone products for **ETL, BI/ reporting and analytics** we have to think about **seamless integration**: **in what ways can we open up a data processing platform to enable applications to get closer?** What language interfaces, but also what resource management facilities can we offer? And so on.” (Florian Waas)

The debate: Which Analytics Platform for Big Data?

Mike Carey (EDBT Keynote 2012):

Big Data in the Database World (early 1980s till now)

- **Parallel Data Bases**. Shared-nothing architecture, declarative set-oriented nature of relational queries, divide and conquer parallelism (e.g. Teradata)
- **Re-implemention of relational databases** (e.g. HP/Vertica, IBM/Netezza, Teradata/ Aster Data,EMC/ Greenplum.)

Big Data in the Systems World (late 1990s)

- **Apache Hadoop** (inspired by Google GFS, MapReduce), (contributed by large Web companies.e.g. Yahoo!, Facebook
- **Google BigTable,**
- **Amazon Dynamo**.

Big Data Analytics

- **In order to analyze Big Data, the current state of the art is a parallel database or NoSQL data store, with a Hadoop connector.**
  - Concerns about performance issues arising with the transfer of large amounts of data between the two systems. The use of connectors could introduce delays, data silos, increase TCO.
  - What about existing **Data Warehouses?**

Which Analytics Platform for Big Data?

- NoSQL (document store, key-value store,…)
- NewSQL
- InMemory DB
- Hadoop
- Data Warehouses
- Plus… scripts, workflows, and ETL-like data transformations

….Are we going back to “*Federated* ” Databases?

This just seems like too many “moving parts”.

Build your own database…

Spanner: Google's Globally-Distributed Database

Spanner is Google's scalable, multi-version, globally- distributed, and synchronously-replicated database. It is the first system to distribute data at global scale and support externally-consistent distributed transactions.

Spanner: Google's Globally-Distributed Database
Published in the Proceedings of OSDI'12: Tenth Symposium on Operating System Design and Implementation, Hollywood, CA, October, 2012. Recipient of the Jay Lepreau Best Paper Award.

Google AdWords Ecosystem

One shared database backing Google's core AdWords business

Legacy DB: Sharded MySQ

Critical applications driving Google's core ad business

- 24/7 availability, even with data center outages
- Consistency required
  - ○ Can't afford to process inconsistent data
  - ○ Eventual consistency too complex and painful

Scale: 10s of TB, replicated to 1000s of machines

F1: A new database, built from scratch, designed to operate at Google scale, without compromising on RDBMS features.

Co-developed with new lower-level storage system, Spanner

- Better scalability
- Better availability
- Equivalent consistency guarantees
- Equally powerful **SQL query**

www.stanford.edu/class/cs347/slides/f1.pdf

Google F1 - A Hybrid Database

F1 - A Hybrid Database combining the

- Scalability of Bigtable
- Usability and functionality of ***SQL databases***
- Key Ideas
- Scalability: Auto-sharded storage
- Availability & Consistency: Synchronous replication
- High commit latency: Can be hidden
- ○ Hierarchical schema
  ○ Protocol buffer column types ○ Efficient client code

A scalable database without going NoSQL.

F1 - The Fault-Tolerant Distributed RDBMS Supporting Google's Ad Business
Jeff Shute, Mircea Oancea, Stephan Ellner, Ben Handy, Eric Rollins, Bart Samwel,
Radek Vingralek, Chad Whipkey, Xin Chen, Beat Jegerlehner, Kyle Littlefield, Phoenix Tong
SIGMOD May 22, 2012

Hadoop Limitations

Hadoop can give powerful analysis, but it is fundamentally a **batch-oriented** paradigm.

The missing piece of the Hadoop puzzle is accounting for **real time changes**.

*Apache™ Hadoop® **YARN** (MapReduce 2.0 (MRv2)) is a sub-project of Hadoop at the Apache Software Foundation that takes Hadoop beyond batch to enable broader data-processing.*

Research Stream: Replacing/Integrating with Hadoop

AMPLab UC Berkeley (https://amplab.cs.berkeley.edu)

Apache Spark is an open-source data analytics cluster computing framework originally developed in the AMPLab at UC Berkeley (https://spark.apache.org)

Databricks was founded out of the UC Berkeley AMPLab by the creators of Apache Spark. A unified platform for building Big Data pipelines – from ETL to Exploration and Dashboards, to Advanced Analytics and Data Products.

The ASTERIX project (UC Irvine- started 2009) http://asterix.ics.uci.edu

Four years of R&D involving researchers at UC Irvine, UC Riverside, and Oracle Labs. The **AsterixDB** code base currently consists of over 250K lines of Java code that has been co-developed by project staff and students at UCI and UCR.open-source Apache-style licence

"To distinguish AsterixDB from current Big Data analytics platforms – which query but don't store or manage Big Data – we like to classify AsterixDB as being a "Big Data Management System" (BDMS, with an emphasis on the "M")"–Mike Carey.

The Stratosphere project (TU Berlin, Humboldt University, Hasso Plattner Institute) (www.stratosphere.eu) contributes to **Apache Flink** is a platform for efficient, distributed, general-purpose data processing. flink.incubator.apache.org

Which Language for Analytics?

- *There is a trend in using **SQL** for analytics and integration of data stores.*

  (e.g. SQL-H, Teradata QueryGrid*)*

Is this good?

Research Stream:
SQL Query Engines for large volumes of data.

BlinkDB http://blinkdb.org Developer Alpha 0.2.0

(amplab UC Berkeley, MIT CSAIL, Uni. Michigan)

BlinkDB is a massively parallel, approximate query engine for running interactive SQL queries on large volumes of data.

BlinkDB is being developed by Sameer Agarwal, Henry Milner, Ariel Kleiner, Ameet Talwalkar, Aurojit Panda, Prof. Michael I. Jordan and Prof. Ion Stoica at the University of California, Berkeley

in collaboration with Prof. Barzan Mozafari at the University of Michigan and Prof. Samuel Madden at the Massachusetts Institute of Technology.

Shark: SQL and Rich Analytics at Scale (amplab at UC Berkeley)

Shark: SQL and Rich Analytics at Scale. Reynold Xin, Joshua Rosen, Matei Zaharia, Michael J. Franklin, Scott Shenker, Ion Stoica. *SIGMOD 2013*. June 2013.

N1QL effort at Couchbase

The SQL++ design from UCSD

Research Stream: Graphs and Big Data

GraphBuilder: A Scalable Graph ETL Framework
Systems Architecture Lab Intel Corporation

GraphBuilder: A Scalable Graph ETL Framework

Graph abstraction essential for many applications, e.g. finding a shortest path to executing complex machine learning (ML) algorithms like collaborative filtering.

Constructing graphs from relationships hidden within large unstructured datasets is challenging. Graph construction is a data-parallel problem, MapReduce is well-suited for this task.

GraphBuilder, an open source scalable framework for graph Extract-Transform-Load (ETL), for graph construction: graph construction, transformation, normalization, and partitioning.

GraphBuilder is written in Java, it scales using the MapReduce model

Large graphs should be partitioned over a cluster for storing and processing and partitioning methods have a significant impact on performance

Research Stream: Benchmarking NoSQL data stores

There is a scarcity of benchmarks to substantiate the many claims made of scalabilty of NoSQL vendors. NoSQL data stores do not qualify for the TPC-C benchmark, since they relax ACID transaction properties.

How can you then measure and compare the performance of the various NoSQL data stores instead?

YCSB: Results and Lesson Learned.

- **Result #1**. "*We knew the systems made fundamental decisions to optimize writes or optimize reads. It was nice to see these decisions show up in the results. Example: in a 50/50 workload, Cassandra was best on throughput. In a 95% read workload, PNUTS caught up and had the best latencies.*"
- **Result #2**. "*The systems may advertise scalability and elasticity, but this is clearly a place where the implementations needed more work. Ref. elasticity experiment. Ref. HBase with only 1-2 nodes.*"
- **Lesson.** "*We are in the early stages. The systems are moving fast enough that there is no clear guidance on how to tune each system for particular workloads.*"

(Adam Silberstein, Raghu Ramakrishnan, previuosly Yahoo! Research)

Research Stream : Hadoop Benchmarks

Quantitatively evaluate and characterize the Hadoop deployment through benchmarking

HiBench: A Representative and Comprehensive Hadoop Benchmark Suite
Intel Asia-Pacific Research and Development Ltd

THE HIBENCH SUITE

HiBench -- benchmark suite for Hadoop, consists of a set of Hadoop programs including both synthetic micro-benchmarks and real-world applications.

Micro Benchmarks : Sort, WordCount , TeraSort, EnhancedDFSIO

Web Search : Nutch Indexing, Page Rank

Machine Learning: Bayesian Classification, K-means Clustering

Analytical Query : Hive Join, Hive Aggregation

Big Data **Benchmark** - **AMPLab** – UC Berkeley

https://amplab.cs.berkeley.edu/benchmark/

This benchmark provides quantitative and qualitative comparisons of four sytems. (hosted on EC2)

- **Redshift - a hosted MPP database offered by Amazon.com based on the ParAccel data warehouse.**
- **Hive - a Hadoop-based data warehousing system. (v0.10, 1/2013 *Note: Hive v0.11, which advertises improved performance, was recently released but is not yet included*)**
- **Shark - a Hive-compatible SQL engine which runs on top of the Spark computing framework. (v0.8 preview, 5/2013)**
- **Impala - a Hive-compatible* SQL engine with its own MPP-like execution engine. (v1.0, 4/2013)**

What is being evaluated?

This benchmark measures response time on a handful of relational queries: scans, aggregations, joins, and UDF's, across different data sizes

Dataset and Workload

The input data set consists of a set of unstructured HTML documents and two SQL tables which contain summary information. It was generated using Intel's

Big Data Benchmarks

TPC launched **TPCx-HS**:

"*industry's first standard for benchmarking big data systems, is designed to provide metric and methodologies to enable fair comparisons of systems from various vendors*"

-- Raghunath Nambiar (CISCO), chairman of the TPC big data committee , August 18, 2014.

Big Data and the **Cloud**

– What about traditional enterprises?
– Very early adoption for analytics

In general people are concerned with the protection and security of their data.

Hadoop in the cloud: Amazon has a significant web-services business around Hadoop.

Big Data myth?

Marc Geall, Former Research Analyst, Deutsche Bank AG/ London, wrote in 2012 (later he joined SAP):

“ We believe that in-memory / NewSQL is likely to be the prevalent database model rather than NoSQL due to three key reasons:

1) the limited need of petabyte-scale data today even among the NoSQL deployment base,

2) very low proportion of databases in corporate deployment which requires more than tens of TB of data to be handles, and

3) lack of availability and high cost of highly skilled operators (often post-doctoral) to operate highly scalable NoSQL clusters.”

Big Data for the Common Good

- Very few people seem to look at how Big Data can be used for solving social problems. Most of the work in fact is not in this direction.

Why this?

Lack of obvious economic and personal incentives…

What can be done in the international research and development communities to make sure that some of the most brilliant ideas do have an impact also for social issues?

Big Data for the Common Good

"As more data become less costly and technology breaks barrier to acquisition and analysis, the opportunity to deliver actionable information for civic purposed grow.

This might be termed the "common good" challenge for Big Data."

(Jake Porway, DataKind)

Leveraging Big Data for Good: Examples

UN Global Pulse: an innovation initiative of the UN Secretary-General, harnessing today's new world of digital data and real-time analytics to gain a better understanding of changes in human well-being.www.unglobalpulse.org

Global Viral Forecasting: a not-for-profit whose mission is to promote understanding, exploration and stewardship of the microbial world.www.gvfi.org

Ushadi SwiftRiver Platform: a non-profit tech company that specializes in developing free and open source software for information collection, visualization and interactive mapping.

http://ushahidi.com

The Eric & Wendy Schmidt Data Science for Social Good fellowship is a University of Chicago summer program for aspiring data scientists to work on data mining, machine learning, big data, and data science projects with social impact.

What are the main difficulties, barriers hindering our community to work on social capital projects?

- **Alon Halevy (Google Research)**: "I don't think there are particular barriers from a technical perspective. **Perhaps the main barrier is ideas of how to actually take this technology and make social impact. These ideas typically don't come from the technical community, so we need more inspiration from activists."**

- **Laura Haas: (IBM Reserch)" Funding and availability of data are two big issues here**. Much funding for social capital projects comes from governments — and as we know, are but a small fraction of the overall budget. Further, the market for new tools and so on that might be created in these spaces is relatively limited, so it is not always attractive to private companies to invest. **While there is a lot of publicly available data today, often key pieces are missing, or privately held, or cannot be obtained for legal reasons, such as the privacy of individuals, or a country's national interests.** While this is clearly an issue for most medical investigations, it crops up as well even with such apparently innocent topics as disaster management (some data about, e.g., coastal structures, may be classified as part of the national defense). "

What are the main difficulties, barriers hindering our community to work on social capital projects?

- **Paul Miller (Consultant)** "Perceived lack of easy access to data that's unencumbered by legal and privacy issues? The large-scale and long term nature of most of the problems? It's not as 'cool' as something else? A perception (whether real or otherwise) that academic funding opportunities push researchers in other directions? Honestly, I'm not sure that there are significant insurmountable difficulties or barriers, if people want to do it enough. As Tim O'Reilly said in 2009 (and many times since), **developers should "work on stuff that matters." The same is true of researchers. "**
- **Roger Barga (Microsot Research)**: "The greatest barrier may be social. Such projects require community awareness to bring people to take action and often a champion to frame the technical challenges in a way that is approachable by the community. These projects will likely **require close collaboration between the technical community and those familiar with the problem."**

What could we do to help supporting initiatives for Big Data for Good?

- **Alon** : **Building a collection of high quality data that is widely available and can serve as the backbone for many specific data projects**. For example, data sets that include boundaries of countries/counties and other administrative regions, data sets with up-to-date demographic data. It's very common that when a particular data story arises, these data sets serve to enrich it.

- **Laura:** Increasingly, we see consortiums of institutions banding together to work on some of these problems. These Centers may **provide data and platforms for data-intensive work, alleviating some of the challenges mentioned above by acquiring and managing data, setting up an environment and tools, bringing in expertise in a given topic, or in data, or in analytics, providing tools for governance, etc**. My own group is creating just such a platform, with the goal of facilitating such collaborative ventures. Of course, lobbying our governments for support of such initiatives wouldn't hurt!

What could we do to help supporting initiatives for Big Data for Good?

- **Paul: Match domains with a need to researchers/companies with a skill/product**. Activities such as the recent Big Data Week Hackathons might be one route to follow – encourage the organisers (and companies like Kaggle, which do this every day) to run Hackathons and competitions that are explicitly targeted at a 'social' problem of some sort. **Continue to encourage the Open Data release of key public data sets.** Talk to the agencies that are working in areas of interest, and understand the problems that they face. Find ways to help them do what they already want to do, and build trust and rapport that way.
- **Roger: Provide tools and resources to empower the long tail of** research. Today, only a fraction of scientists and engineers enjoy regular access to high performance and data-intensive computing resources to process and analyze massive amounts of data and run models and simulations quickly. The reality for most of the scientific community is that speed to discovery is often hampered as they have to either queue up for access to limited resources or pare down the scope of research to accommodate available processing power. This problem is particularly acute at the smaller research institutes which represent the long tail of the research community**. Tier 1 and some tier 2 universities have sufficient funding and infrastructure to secure and support computing resources while the smaller research programs struggle.** Our funding agencies and corporations must provide resources to support researchers, in particular those who do not have access to sufficient resources.

Full report : "Big Data for Good", Roger Barga, Laura Haas, Alon Halevy, Paul Miller, Roberto V. Zicari. ODBMS Industry Watch June 5, 2012 www.odbms.org and www.odbms.org/blog

The search for meaning behind our activities.

“ All our activities in our lives can be looked at from different perspectives and within various contexts: our individual view, the view of our families and friends, the view of our company and finally the view of society- the view of the world. Which perspective means what to us is not always clear, and it can also change over the course of time. This might be one of the reasons why our life sometimes seems unbalanced. We often talk about work-life balance, but maybe it is rather an imbalance between the amount of energy we invest into different elements of our life and their meaning to us ”

--Eran Davidson, CEO Hasso Plattner Ventures.