Big Data: Challenges and Opportunities

Roberto V. Zicari

Goethe University Frankfurt

ODBMS.org

www.odbms.org

roberto@zicari.de

October 5, 2012

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.

Big Data:The story as it is told from the Business Perspective.

- "*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).

Big Data: A Possible Definition

"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)

Where is Big Data?

- (Big) Data is in every industry and business function and are important factor for production

  (McKinsey Global Institute)

- (estimated 7 exabytes of new data enterprises globally stored in 2010-MGI)

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?

- *Key basis of competition and growth for individual firms* (McKinsey Global Institute).
  - E.g. “retailer embracing big data has the potential to increase its operating margin by more than 60 percent”.

How to measure the value of Big Data?

- Consider only those actions that essentially depends on the use of big data. (McKinsey Global Institute)

Big Data can generate financial value across sectors

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

(McKinsey Global Institute)

Limitations

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

  (McKinsey Global Institute)

Issues

(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

Big Data: Challenges

Data, Process, Management

Data:

- **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.

Challenges cont.

Data:

- **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?

  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)

Challenges

Data

- **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, mis-statements 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).
- **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).
- **Combining multiple data sets**

Challenges cont.

Data

- **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)

Challenges cont.

Data:

- **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)

Challenges cont.

Process

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)

Challenges cont.

Management: 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).

Let`s take some time

to critically review this story.

Examples of BIG DATA USE CASES

- Log Analytics
- Fraud Detection
- Social Media and Sentiment Analysis
- Risk modeling and management
- Energy sector

Big Data: The story as it is told from the Technology Perspective.

What are the main technical challenges for big data 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.

"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)

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

Take the Analysis to the Data!

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)

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)

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)

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)

Semi-structured Web data.

- A/B testing, sessionization, bot detection, and pathing analysis all require powerful analytics on many petabytes of semi-structured Web data.

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.**

Scalability

Scalability has three aspects:

- data volume,
- hardware size, and
- concurrency.

Which Analytics Platform for Big Data?

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

Parallel database software stack (Michael J. Carey)

SQL-->**SQL Compiler**

Relational Dataflow layer (runs the query plans, orchestrate the local storage managers, deliver partitioned, shared-nothing storage services for large relational tables)

Row/Column Storage Manager (record-oriented: made up of a set of row-oriented or column oriented storage managers per machine in a cluster)

No open-source parallel database exists!

SQL is the only way into the system architecture.

Monolithic: Can`t safely cut into them to access inner functionalities

Hadoop software stack

(Michael J. Carey)

- *HiveQL. PigLatin, Jaql script*-->**HiveQL/Pig/Jaql** (High-level languages)
- *Hadoop M/R job*-->**Hadoop Map Reduce** Dataflow Layer/ (for batch analytics, applies Map ops to the data in partitions of an HDFS file, sorts, and redistributes the results based on key values in the output data, then performs reduce on the groups of output data items with matching keys from the map phase of the job).
- *Get/Put ops*-->**Hbase Key-value Store** (accessed directly by client app or via Hadoop for anylstics needs)
- **Hadoop Distributed File System** (byte oriented file abstraction- files appears as a very large contiguos and randomly addressable sequence of bytes

Hadoop Pros
(Michael J. Carey)

Hadoop Pros:

- open source
- non-monolithic
- support for access to file-based external data
- support for automatic and incremental forward-recovery of jobs with failed task
- ability to schedule very large jobs in smaller chunks
- automatic data placement and rebalancing as data grows and machines come and go.
- support for replication and machine fail-over without operation intervention

Hadoop Cons (Michael J. Carey)

Hadoop Cons:

- *questionable to layer a record-oriented data abstraction on top of a giant globally-sequenced byte-stream file abstraction. (e.g. HDFS is unaware of record boundaries. "broken records" instead of fixed-lenght file splits, i.e. a record with some of its bytes in one spit and some inn the next)*

- *questionable building a parallel data runtime on top of a unary operator model (map, reduce, combine). E.g. Performing joins with MapReduce.*

- *questionable building a key-value store layer with a remote query access at the next layer. Pushing queries down to data is likely to outperform pulling data up to queries.*

- *lack of schema information, today is flexible, but a recipe for future difficulties. E.g. Future maintainers of applications will likely have problems in fixing bugs related to changes or assumptions about the structure of data files in HDFS. (This was one of the very early lessons in the DB world).*

- *Not addressed single system performance, focusing solely on scale-out.*

“Big Data” stack

“Many academics are being too “shy” about questioning and rethinking the “Big Data” stack forming around Hadoop today, perhaps because they are embarassed about having fallen asleep in the mid-1990s (with respect to parallel data issues) and waking up in the world as it exists today” (Michael Carey, EDBT keynote 2012)

“Big Data” stack

“Rather that trying to modify the Hadoop code base to add indexing or data co-clustering support, or gluing open-source database systems underneath Hadoop`s data input APIs, we believe that database researchers should be asking “Why?” or “What if we`d meant to design an open software stack with records at the bottom and a strong possibility of a higher-level language API at the top?”

(Michael Carey, EDBT keynote 2012)

Two Research Projects

- The ASTERIX project (UC Irvine-started 2009) open-source Apache-style licence.
- The Stratosphere project (TU Berlin) (www.stratosphere.eu)

Apache Hadoop

- Apache Hadoop provides a new platform to analyze and process Big Data.
- Hadoop was inspired by Google`s **MapReduce** and **Google File System** (GFS) papers.

Hadoop is really an ecosystems of projects

Higher-level declarative languages for writing queries and data analysis pipelines

- Pig (Yahoo!) - relational-like algebra
  - (ca. 60% of Yahoo! MapReduce use cases)
- PigLatin
- Hive (Facebook) also inspired by SQL
  - (ca. 90% of Facebook MapReduce use cases)
- Jaql (IBM)
- Load
- Transform
- Dump and store

More

- Flume
- Oozie
- Etc.

Zookeeper

Lucene

Hbase

Avro

Apache Hadoop benefits

- The combination of **scale**, ability to **process unstructured data** along with the availability of machine learning algorithms and recommendation engines creates the opportunity to build new game changing applications.

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.

Hadoop Limitations

HDS has a centralized metadata store (NameNode), which represents a single point of failure without availability. When the NameNode is recovered, it can take a long time to get the Hadoop cluster running again.

Difficult to use

– Work is in progress to fix this from vendors of commercial Hadoop distributions (e.g. MapR, etc.) by re-implementing Hadoop components.

Big Data “Dichotomy”

- Analytics: MapReduce, Hadoop
- Developers of very large scale user-facing Web sites implemented **key-value stores**
  - Google Big Table
  - Amazon Dynamo
  - Apache Hbase (open source BigTable clone),
  - Apache Cassandra, Riak (open source Dynamo clones),

Hadoop

"By **not requiring a schema first**, Hadoop provides a great tool for exploratory analysis of the data, as long as you have the software development expertise to write Map Reduce programs.

Hadoop assumes that the workload it runs will be long running, so it makes heavy use of checkpointing at intermediate stages. This means parts of a job can fail, be restarted and eventually complete successfully.

There are no transactional guarantees.

(Shilpa Lawande**)**

Why Using Hadoop?

"We chose Hadoop for several reasons.

- First, it is the only available framework that could scale to process 100s or even 1000s of terabytes of data and scale to installations of up to 4000 nodes.
- Second, Hadoop is open source and we can innovate on top of the framework and inside it to help our customers develop more performant applications quicker.
- Third, we recognized that Hadoop was gaining substantial popularity in the industry with multiple customers using Hadoop and many vendors innovating on top of Hadoop.

Three years later we believe we made the right choice. We also see that existing BI vendors such as Microstrategy are willing to work with us and integrate their solutions on top of Elastic. MapReduce."

(**Werner Vogels, VP and CTO Amazon)**

Vertica (NewSQL) and Hadoop

“Vertica (*) and Hadoop are both systems that can store and analyze large amounts of data on commodity hardware.

The main differences are how the data gets in and out, how fast the system can perform, and what transaction guarantees are provided.

Also, from the standpoint of **data access, Vertica`s interface is SQL and data must be designed and loaded into a SQL schema for analysis. With Hadoop, data is loaded AS IS into a distributed file system and accessed programmatically by writing Map-Reduce programs**. “**(** Shilpa Lawande**)**

(*) columnar database engine including sorted columnar storage, a query optimizer and an execution engine, provides standard ACID transaction semantics on loads and queries

Analytics at eBay: technical challenges

Main technical challenges for big data analytics at eBay:

- ***I/O bandwidth***: limited due to configuration of the nodes.
- ***Concurrency/workload management****:* Workload management tools usually manage the limited resource. For many years EDW systems bottle neck on the CPU; big systems are configured with ample CPU making I/O the bottleneck. Vendors are starting to put mechanisms in place to manage I/O, but it will take some time to get to the same level of sophistication.
- ***Data movement (loads, initial loads, backup/restores****):* As new platforms are emerging you need to make data available on more systems challenging networks, movement tools and support to ensure scalable operations that maintain data consistency **(Tom Fastner)**

Analytics at eBay: Platforms for Analytics

3 different platforms for Analytics (**Tom Fastner**):

A) *EDW*: Dual systems for transactional (**structured data**); Teradata 6690 with 9.5 PB spinning disk and 588 TB SSD - the largest mixed storage Teradata system world wide; with spool, some dictionary tables and user data automatically managed by access frequency to stay on SSD. 10+ years experience; very high concurrency; good accessibility; hundreds of applications.

B) *Singularity*: deep Teradata system for **semi-structured data**; 36 PB spinning disk; lower concurrency that EDW, but can store more data; biggest use case is User Behavior Analysis; largest table is 1.2 PB with ~3 Trillion rows.

C) *Hadoop*: for **unstructured/complex data**; ~40 PB spinning disk; text analytics, machine learning; has the User Behavior data and selected EDW tables; lower concurrency and utilization.

Analytics at eBay: Scalability and Performance

- **DW**: We model for the unknown (close to 3rd NF) to provide a solid physical data model suitable for many applications, that limits the number of physical copies needed to satisfy specific application requirements. **A lot of scalability and performance is built into the database,** but as any shared resource **it does require an excellent operations team to fully leverage the capabilities of the platform**
- **Singularity:** The platform is identical to EDW, the only exception are limitations in the workload management due to configuration choices. But since we are leveraging the latest database release we are exploring ways to adopt new storage and processing patterns. Some new data sources are stored in a **denormalized form** significantly simplifying data modeling and ETL. On top **we developed functions to support the analysis of the semi-structured data.** It also enables more **sophisticated algorithms that would be very hard, inefficient or impossible to implement with pure SQL.** One example is the pathing of user sessions. However the size of the data requires us to focus more on best practices (develop on small subsets, use 1% sample; process by day),
- **Hadoop:** The emphasis on Hadoop is on optimizing for access. The reusability of data structures (besides “raw” data) is very low.

(**Tom Fastner)**

Analytics at eBay: Un-structured data

“Un-structured data is handled on Hadoop only. The data is copied from the source systems into HDFS for further processing. We do not store any of that on the Singularity (Teradata) system” **(Tom Fastner)**

Analytics at eBay: Use of Data management technologies

- **ETL**: AbInitio, home grown parallel Ingest system.
- **Scheduling**: UC4.
- **Repositories:** Teradata EDW; Teradata Deep system; Hadoop.
- **BI:** Microstrategy, SAS, Tableau, Excel.
- **Data modeling:** Power Designer.
- **Adhoc:** Teradata SQL Assistant; Hadoop Pig and Hive.
- **Content Management:** Joomla based.

(Tom Fastner)

Analytics at eBay:
Cloud computing and open source

“We do leverage internal cloud functions for Hadoop; no cloud for Teradata.

Open source: committers for Hadoop and Joomla; strong commitment to improve those technologies”

(Tom Fastner)

Analytics at eBay: use of analytics

“Ebay is rapidly changing, and analytics is driving many key initiatives like **buyer experience, search optimization, buyer protection or mobile commerce.** We are investing heavily in new technologies and approaches to leverage new data sources to drive innovation.” (**Tom Fastner**)

Hadoop users

- Advanced users of Hadoop are looking to go beyond batch uses of Hadoop to support real-time streaming of content.
  - How many advanced users?
- New users need Hadoop to become easier. Need it to be easier to develop Hadoop applications, deploy them and run them in a production environment.
  - Is there a real need for it?

Applicability of Hadoop

“Promises” from Hadoop vendors:

Product recommendations, ad placements, customer churn, patient outcome predictions, fraud detection and sentiment analysis are just a few examples that improve with **real time information**.

- Organizations are also looking to **expand Hadoop use cases** to include business critical, secure applications that easily integrate with file-based applications and products.
- With mainstream adoption comes the **need for tools that don`t require specialized skills and programmers**. New Hadoop developments must be simple for users to operate and to get data in and out. This includes direct access with standard protocols using existing tools and applications.

--> See also Big Data Myth later

Hadoop distributions challenges

- **Getting data in and out of Hadoop**. Some Hadoop distributions are limited by the append-only nature of the Hadoop Distributed File System (HDFS) that requires programs to batch load and unload data into a cluster.
- **Deploying Hadoop into mission critical business projects**. The lack of reliability of current Hadoop software platforms is a major impediment for expansion.
- **Protecting data against application and user errors**. Hadoop has no backup and restore capabilities. Users have to contend with data loss or resort to very expensive solutions that reside outside the actual Hadoop cluster.

Hadoop and the Cloud

- 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
  - What about traditional enterprises?

VoltDB (NewSQL) and Hadoop

"VoltDB is not focused on analytics. We believe they should be run on a companion data warehouse. Most of the warehouse customers I talk to want to keep increasing large amounts of increasingly diverse history to run their analytics over. The major data warehouse players are routinely being asked to manage petabyte-sized data warehouses.

VoltDB is intended for the OLTP portion, and some customers wish to run Hadoop as a data warehouse platform. To facilitate this architecture, VoltDB offers a Hadoop connector" **(Mike Stonebraker)**

Couchbase (NoSQL) and Hadoop

- In some applications Couchbase (NoSQL) is used to enhance the batch-based Hadoop analysis with real time information, giving the effect of a continuous process.

Couchbase (NoSQL) and Hadoop

- Hot data lives in Couchbase in RAM.
- Essentially move the data out of Couchbase into Hadoop when it cools off.
- Connector to Apache Sqoop (Top-Level Apache project since March of 2012): a tool designed for efficiently transferring bulk data between Hadoop and relational databases.

NoSQL and Hadoop

“In my opinion the primary interface will be via the real time store, and the Hadoop layer will become a commodity. That is why there is so much competition for the NoSQL brass ring right now.” --J. Chris Anderson.

Benchmarking SQL and 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?

Yahoo! YCSB benchmark

- Vendors are making a lot of claims about latency, throughput and scalability without much proof,

- **Yahoo YCSB benchmark** is source of good comparisons.

Benchmark for Cloud Serving Systems

- *A team of researchers composed of **Adam Silberstein, Brian F. Cooper, Raghu Ramakrishnan, Russell Sears, and Erwin Tam**, all at **Yahoo! Research Silicon Valley**, developed a new benchmark for Cloud Serving Systems, called YCSB.*

Measuring the scalability of SQL and NoSQL systems.

- *They open-sourced the benchmark about a year ago.*
- *The YCSB benchmark appears to be the best to date for measuring the scalability of SQL and NoSQL systems.*

NoSQL Performance

- There are many design decisions to make when building NoSQL systems, and those decisions have a huge impact on how the system performs for different workloads (e.g., read-heavy workloads vs. write-heavy workloads), how it scales, how it handles failures, ease of operation and tuning, etc.

YCSB vs. TPC-C

- At a high level, there is a lot in common with TPC-C and other OLTP benchmarks: Query latency and overall system throughput.

BUT

- Queries are very different. TPC-C contains several diverse types of queries meant to mimic a company warehouse environment. Some queries execute transactions over multiple tables; some are more heavyweight than others.
- **In contrast, the web applications YCSB is benchmarking tend to run a huge number of extremely simple queries**.

YCSB vs. TPC-C

Consider a table where each record holds a user`s profile information.

Every query touches only a single record, likely either reading it, or reading+writing it.

YCSB does include support for skewed workloads; some tables may have active sets accessed much more than others.

Focused on simple queries.

Ease of creating a new suite of benchmarks using the YCSB framework.

Parameters

- **Performance**: refers to the usual metrics of latency and throughput, with the ability to scale out by adding capacity.
- **Elasticity**: refers to the ability to add capacity to a running deployment “on-demand”, without manual intervention (e.g., to re-shard existing data across new servers).

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)

YCSB: Availability

- The authors open-sourced the benchmark about a year ago.

  It is available at:

  https://github.com/brianfrankcooper/YCSB

Big Data in Data Warehouse or in Hadoop?

- Data Warehouse: Structured data, Data "trusted"
- Hadoop: Semistructured and unstructured data. Data "not trusted".
  - Work in progress to develop tools

How easy is Hadoop?

“There are only a few Facebook-sized IT organizations that can have 60 Stanford PhDs on staff to run their Hadoop infrastructure. The others need it to be easier to develop Hadoop applications, deploy them and run them in a production environment.”-- John Schroeder.

Big Data Search

- There is no single set formula for extracting value from big 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.

Enterprise Search

Enterprise Search implies being able to search multiple types of data generated by an enterprise.

Apache Solr. There`s an ecosystem of tools that build on Solr,

- Solr support or implementing a proprietary full-text search engine ?

Big Data Search: Example

For example, real-time co-occurrence analysis new insights about how products are being used.

- It was analysis of social media that revealed that Gatorade is closely associated with flu and fever, and with the ability to drill seamlessly from high-level aggregate data into the actual source social media posts shows that many people actually take Gatorade to treat flu symptoms.
- Geographic visualization shows that this phenomenon may be regional.

(David Gorbert)

Big Data myth

“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.”

(Marc Geall, Research Analyst, Deutsche Bank AG/London)

Let`s take time

to review this story

Big Data: The other story

- 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?

- What can be done in the international research/development community 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)

World Economic Forum
Big Data, Big Impact: New Possibilities for International Development

“A flood of data is created every day by the interactions of billions of people using computers,GPS devices, cell phones, and medical devices. Many of these interactions occur through the use of mobile devices being used by people in the developing world, people whose needs and habits have been poorly understood until now. Researchers and policymakers are beginning to realize the potential for channeling these torrents of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises for the benefit of low-income populations. Concerted action is needed by governments, development organizations, and companies to ensure that this data helps the individuals and communities who create it.”

The United Nations Global Pulse initiative

The United Nations Global Pulse initiative is one example. Earlier this year at the 2012 Annual Meeting in Davos, the World Economic Forum published a white paper entitled

“Big Data, Big Impact: New Possibilities for International Development“.

The WEF paper lays out several of the ideas which fundamentally drive the Global Pulse initiative and presents in concrete terms the opportunity presented by the explosion of data in our world today, and how researchers and policymakers are beginning to realize the potential for leveraging Big Data to extract insights that can be used for Good, in particular for the benefit of low-income populations.

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

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

- **Alon Havely (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 Barca, 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.

Acknowledgements

I thank Michael Blaha, Rick Cattell, Michael Carey, Akmal Chaudhri, Tom Fastner, Laura Haas, Alon Halevy, Volker Markl, Dave Thomas, Duncan Ross, Cindy Saracco, Justin Sheehy, Miguel-Angel Sicilia, Mike OSullivan, Steve Vinoski, for their feedback on this presentation.