3

Big Data: Challenges and Opportunities

Roberto V. Zicari

CONTENTS

Introduction

> "Big Data is the new gold" (Open Data Initiative)

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

There is no doubt that Big Data and especially *what we do with it* has the potential to become a driving force for innovation and value creation. In this chapter, we will look at Big Data from three different perspectives: the business perspective, the technological perspective, and the social good perspective.

The Story as it is Told from the Business Perspective

Now let us define the term *Big Data.* I have selected a definition, given by McKinsey Global Institute (MGI) [1]:

> "Big Data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.

This definition is quite general and open ended, and well captures the rapid growth of available data, and also shows the need of technology to "catch up" with it. This definition is not defined in terms of data size; in fact, data sets will increase in the future! It also obviously varies by sectors, ranging from a few dozen terabytes to multiple petabytes (1 petabyte is 1000 terabytes).

(Big) Data is in every industry and business function and is an important factor for production. MGI estimated that 7 exabytes of new data enterprises globally were stored in 2010. Interestingly, more than 50% of IP traffic is non-human, and M2M will become increasingly important. *So what is Big Data supposed to create*? Value. But what "value" exactly? Big Data *per se* does not produce any value.

David Gorbet of MarkLogic explains [2]: "the increase in data complexity is the biggest challenge that every IT department and CIO must address. Businesses across industries have to not only store the data but also be able to leverage it quickly and effectively to derive business value."

Value comes only from what we infer from it. That is why we need *Big Data Analytics*.

Werner Vogels, CTO of Amazon.com, describes Big Data Analytics as follows [3]: "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 changeover time and as such you need to be able to collect, store and analyze data without being constrained by resources."

According to MGI, the "value" that can be derived by analyzing Big Data can be spelled out as follows:

- Creating transparencies;
- Discovering needs, exposing variability, and improving performance;
- Segmenting customers; and
- Replacing/supporting human decision-making with automated algorithms—Innovating new business models, products, and services.

"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" explains David Gorbet.

"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" says David Gorbet.

Gorbet's concept of "Big Data Search" implies the following:

- 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.

Gorbet gives an example of the result of such Big Data Search: "it was analysis of social media that revealed that Gatorade is closely associated with flu and fever, and our 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. Our ability to sift through all this data in real time, using fresh data gathered from multiple sources, both internal and external to the organization helps our customers identify new actionable insights."

Where Big Data will be used? According to MGI, Big Data can generate financial value across sectors. They identified the following key sectors:

- Health care (this is a very sensitive area, since patient records and, in general, information related to health are very critical)
- Public sector administration (e.g., in Europe, the Open Data Initiative—a European Commission initiative which aims at opening up Public Sector Information)
- Global personal location data (this is very relevant given the rise of mobile devices)
- Retail (this is the most obvious, since the existence of large Web retail shops such as eBay and Amazon)
- Manufacturing

I would add to the list two additional areas

- Social personal/professional data (e.g., Facebook, Twitter, and the like)

What are examples of *Big Data Use Cases?* The following is a sample list:

- Log analytics
- Fraud detection
- Social media and sentiment analysis
- Risk modeling and management
- Energy sector

Currently, the key *limitations* in exploiting *Big Data*, according to MGI, are

- Shortage of talent necessary for organizations to take advantage of Big Data
- Shortage of knowledge in statistics, machine learning, and data mining

Both limitations reflect the fact that the current underlying technology is quite difficult to use and understand. As every new technology, Big Data Analytics technology will take time before it will reach a level of maturity and easiness to use for the enterprises at large. All the above-mentioned examples of values generated by analyzing Big Data, however, do not take into account the possibility that such derived "values" are *negative*.

In fact, the analysis of Big Data if improperly used poses also *issues*, specifically in the following areas:

- Access to data
- Data policies
- Industry structure
- Technology and techniques

This is outside the scope of this chapter, but it is for sure one of the most important nontechnical challenges that Big Data poses.

The Story as it is Told from the Technology Perspective

The above are the business "promises" about Big Data. But what is the reality today? Big data problems have several characteristics that make them *technically challenging*.

We can group the challenges when dealing with Big Data in three dimensions: *data, process, and management*. Let us look at each of them in some detail:

Data Challenges

Volume

The volume of data, especially machine-generated data, is exploding, how fast that data is growing every year, with new sources of data that are emerging. For example, in the year 2000, 800,000 petabytes (PB) of data were stored in the world, and it is expected to reach 35 zettabytes (ZB) by 2020 (according to IBM).

> Social media plays a key role: Twitter generates 7+ terabytes (TB) of data every day. Facebook, 10 TB. Mobile devices play a key role as well, as there were estimated 6 billion mobile phones in 2011.

The challenge is how to deal with the size of Big Data.

Variety, Combining Multiple Data Sets

More than 80% of today's information is unstructured and it is typically too big to manage effectively. What does it mean?

David Gorbet explains [2]:

> It used to be the case that all the data an organization needed to run its operations effectively was structured data that was generated within the organization. Things like customer transaction data, ERP data, etc. Today, companies are looking to leverage a lot more data from a wider variety of sources both inside and outside the organization. Things like documents, contracts, machine data, sensor data, social media, health records, emails, etc. The list is endless really.
>
> A lot of this data is unstructured, or has a complex structure that's hard to represent in rows and columns. And organizations want to be able to combine all this data and analyze it together in new ways.
>
> For example, we have more than one customer in different industries whose applications combine geospatial vessel location data with weather and news data to make real-time mission-critical decisions.
>
> Data come from sensors, smart devices, and social collaboration technologies. Data are not only structured, but raw, semistructured, unstructured data from web pages, web log files (click stream data), search indexes, e-mails, documents, sensor data, etc.
>
> Semistructured Web data such as A/B testing, sessionization, bot detection, and pathing analysis all require powerful analytics on many petabytes of semistructured Web data.

The challenge is how to handle multiplicity of types, sources, and formats.

Velocity

> Shilpa Lawande of Vertica defines this challenge nicely [4]: "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."

One of the key challenges is how to react to the flood of information in the time required by the application.

Veracity, Data Quality, Data Availability

Who told you that the data you analyzed is good or complete? Paul Miller [5] mentions that "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?"

There are several challenges:

- How can we cope with uncertainty, imprecision, missing values, misstatements or untruths?
- 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?
- Is there data available, at all?

Data Discovery

This is a huge challenge: how to find high-quality data from the vast collections of data that are out there on the Web.

Quality and Relevance

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

Data Comprehensiveness

Are there areas without coverage? What are the implications?

Personally Identifiable Information

Much of this information is about people. 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 [5])

Can we extract enough information to help people without extracting so much as to compromise their privacy?

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.

For example, "It would be worrying if the healthcare sector only responded to flu outbreaks when Google Flu Trends told them to." (Paul Miller [5])

Scalability

Shilpa Lawande explains [4]: "techniques like social graph analysis, for instance leveraging the influencers in a social network to create better user experience are hard problems to solve at scale. All of these problems combined create a perfect storm of challenges and opportunities to create faster, cheaper and better solutions for Big Data analytics than traditional approaches can solve."

Process Challenges

"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 throw away) models—at scale—is critical." (Shilpa Lawande)

According to Laura Haas (IBM Research), process challenges with deriving insights include [5]:

- 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, think for a second how to display complex analytics on a iPhone or a mobile device

Management Challenges

"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)

The main management challenges are

- Data privacy
- Security

- Governance
- Ethical

The challenges are: Ensuring that data are used correctly (abiding by its intended uses and relevant laws), tracking how the data are used, transformed, derived, etc., and managing its lifecycle.

Big Data Platforms Technology: Current State of the Art

The industry is still in an immature state, experiencing an explosion of different technological solutions. Many of the technologies are far from robust or enterprise ready, often requiring significant technical skills to support the software even before analysis is attempted. At the same time, there is a clear shortage of analytical experience to take advantage of the new data. Nevertheless, the potential value is becoming increasingly clear.

In the past years, the motto was "rethinking the architecture": 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," says Florian Waas of EMC/Greenplum [6].
>
> "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)

This confirms Gray's Laws of Data Engineering, adapted here to Big Data:

Take the Analysis to the Data!

In order to analyze Big Data, the current state of the art is a parallel database or NoSQL data store, with a Hadoop connector. *Hadoop* is used for processing the *unstructured* Big Data. Hadoop is becoming the standard platform for doing large-scale processing of data in the enterprise. Its rate of growth far exceeds any other "Big Data" processing platform.

What Is Apache Hadoop?

Hadoop provides a new open source platform to analyze and process Big Data. It was inspired by Google's MapReduce and Google File System (GFS) papers. It is really an ecosystems of projects, including:

Higher-level declarative languages for writing queries and data analysis pipelines, such as:

- Pig (Yahoo!)—relational-like algebra—(used in ca. 60% of Yahoo! MapReduce use cases)
- PigLatin
- Hive (used by Facebook) also inspired by SQL—(used in ca. 90% of Facebook MapReduce use cases)
- Jaql (IBM)
- Several other modules that include Load, Transform, Dump and store, Flume Zookeeper Hbase Oozie Lucene Avro, etc.

Who Are the Hadoop Users?

A simple classification:

- Advanced users of Hadoop.

  They are often PhDs from top universities with high expertise in analytics, databases, and data mining. They are looking to go beyond batch uses of Hadoop to support real-time streaming of content. 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.

  How many of such advanced users currently exist?

  "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." (JohnSchroeder [7])

  So, not that many apparently.
- New users of Hadoop

  They need Hadoop to become easier. Need it to be easier to develop Hadoop applications, deploy them, and run them in a production environment.

  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 do not require specialized skills and programers. 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.

Is there a real need for it? See also Big Data Myth later.

An Example of an Advanced User: Amazon

"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 [3])

Big Data in Data Warehouse or in Hadoop?

Roughly speaking we have:

- *Data warehouse*: structured data, data "trusted"
- *Hadoop*: semistructured and unstructured data. Data "not trusted"

An interesting historical perspective of the development of Big Data comes from Michael J. Carey [8]. He distinguishes between:

Big Data in the Database World (Early 1980s Till Now)

- Parallel Databases. Shared-nothing architecture, declarative set-oriented nature of relational queries, divide and conquer parallelism (e.g., Teradata). Later phase re-implementation of relational databases (e.g., HP/Vertica, IBM/Netezza, Teradata/Aster Data, EMC/Greenplum, Hadapt)

and

Big Data in the Systems World (Late 1990s Till Now)

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

The Parallel database software stack (Michael J. Carey) comprises

- 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)

Note: no open-source parallel database exists! SQL is the only way into the system architecture. Systems are monolithic: Cannot safely cut into them to access inner functionalities.

The Hadoop software stack comprises (Michael J. Carey):

- HiveQL. PigLatin, Jaql script → HiveQL/Pig/Jaql (High-level languages)
- Hadoop M/R job → Hadoop MapReduce 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 analytics needs)
- Hadoop Distributed File System (byte oriented file abstraction—files appears as a very large contiguous and randomly addressable sequence of bytes

Note: all tools are open-source! No SQL. Systems are not monolithic: Can safely cut into them to access inner functionalities.

A key requirement when handling Big Data is *scalability.*

Scalability has three aspects

- data volume
- hardware size
- concurrency

What is the trade-off between *scaling out* and *scaling up*? What does it mean in practice for an application domain?

Chris Anderson of Couchdb explains [11]: "scaling up is easier from a software perspective. It's essentially the Moore's Law approach to scaling—buy a bigger box. Well, eventually you run out of bigger boxes to buy, and then you've run off the edge of a cliff. You've got to pray Moore keeps up.

Scaling out means being able to add independent nodes to a system. This is the real business case for NoSQL. Instead of being hostage to Moore's Law, you can grow as fast as your data. Another advantage to adding independent nodes is you have more options when it comes to matching your workload. You have more flexibility when you are running on commodity hardware—you can run on SSDs or high-compute instances, in the cloud, or inside your firewall."

Enterprise Search

Enterprise Search implies being able to search multiple types of data generated by an enterprise. There are two alternatives: Apache Solr or implementing a proprietary full-text search engine.

There is an ecosystem of open source tools that build on Apache Solr.

Big Data "Dichotomy"

The prevalent architecture that people use to analyze structured and unstructured data is a two-system configuration, where *Hadoop* is used for processing the unstructured data and a relational database system or an NoSQL data store is used for the structured data as a front end.

NoSQL data stores were born when 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), etc.

There are concerns about performance issues that arise along with the transfer of large amounts of data between the two systems. The use of connectors could introduce delays and data silos, and increase Total Cost of Ownership (TCO).

Daniel Abadi of Hadapt says [10]: "this is a highly undesirable architecture, since now you have two systems to maintain, two systems where data may be stored, and if you want to do analysis involving data in both systems, you end up having to send data over the network which can be a major bottleneck."

Big Data is not (only) Hadoop.

"Some people even think that 'Hadoop' and 'Big Data' are synonymous (though this is an over-characterization). Unfortunately, Hadoop was designed based on a paper by Google in 2004 which was focused on use cases involving unstructured data (e.g., extracting words and phrases from Web pages in order to create Google's Web index). Since it was not originally designed to leverage the structure in relational data in order to take

short-cuts in query processing, its performance for processing relational data is therefore suboptimal" says Daniel Abadi of Hadapt.

Duncan Ross of Teradata confirms this: "the biggest technical challenge is actually the separation of the technology from the business use! Too often people are making the assumption that Big Data is synonymous with *Hadoop*, and any time that technology leads business things become difficult. Part of this is the difficulty of use that comes with this.

It's reminiscent of the command line technologies of the 70s—it wasn't until the GUI became popular that computing could take off."

Hadoop and the Cloud

Amazon has a significant web-services business around Hadoop.

But in general, people are concerned with the protection and security of their data. *What about traditional enterprises?*

Here is an attempt to list the *pros* and *cons* of Hadoop.

Hadoop Pros

- Open source.
- Nonmonolithic 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.
- The combination of scale, ability to process unstructured data along with the availability of machine learning algorithms, and recommendation engines create the opportunity to build new game changing applications.
- Does not require a schema first.
- Provides a great tool for exploratory analysis of the data, as long as you have the software development expertise to write MapReduce programs.

Hadoop Cons

- Hadoop is difficult to use.
- 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 file system (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.
- Hadoop assumes that the workload it runs will belong 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.

Current 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.
- 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.

There is work in progress to fix this from vendors of commercial Hadoop distributions (e.g., MapR, etc.) by reimplementing Hadoop components.
It would be desirable to have 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)

Daniel Abadi: "A lot of people are using Hadoop as a sort of data refinery. Data starts off unstructured, and Hadoop jobs are run to clean, transform, and structure the data. Once the data is structured, it is shipped to SQL databases where it can be subsequently analyzed. This leads to the raw data being left in Hadoop and the refined data in the SQL databases. But it's basically the same data—one is just a cleaned (and potentially aggregated) version of the other. Having multiple copies of the data can lead to all kinds of problems. For example, let's say you want to update the data in one of the two locations—it does not get automatically propagated to the copy in the other silo. Furthermore, let's say you are doing some analysis in the SQL database and you see something interesting and want to drill down to the raw data—if the raw data is located on a different system, such a drill down

becomes highly nontrivial. Furthermore, data provenance is a total nightmare. It's just a really ugly architecture to have these two systems with a connector between them."

Michael J. Carey adds that is:

- 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-length file splits, i.e., a record with some of its bytes in one split and some in 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.

Technological Solutions for Big Data Analytics

There are several technological solutions available in the market for Big Data Analytics. Here are some examples:

An NoSQL Data Store (CouchBase, Riak, Cassandra, MongoDB, etc.) Connected to Hadoop

With this solution, an NoSQL data store is used as a front end to process selected data in real time data, and having Hadoop in the back end processing Big Data in batch mode.

> "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" says J. Chris Anderson of Couchbase (an NoSQL datastore).

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

The process consists of essentially moving the data out of Couchbase into Hadoop when it cools off. CouchDB supplies a connector to Apache Sqoop (a Top-Level Apache project since March of 2012), a tool designed for efficiently transferring bulk data between Hadoop and relational databases.

An NewSQL Data Store for Analytics (HP/Vertica) Instead of Hadoop

Another approach is to use a NewSQL data store designed for Big Data Analytics, such as HP/Vertica. Quoting Shilpa Lawande [4] "Vertica was designed from the ground up for analytics." Vertica is a columnar database engine including sorted columnar storage, a query optimizer, and an execution engine, providing standard ACID transaction semantics on loads and queries.

With sorted columnar storage, there are two methods that drastically reduce the I/O bandwidth requirements for such Big Data analytics workloads. The first is that Vertica only reads the columns that queries need. Second, Vertica compresses the data significantly better than anyone else. Vertica's execution engine is optimized for modern multicore processors and we ensure that data stays compressed as much as possible through the query execution, thereby reducing the CPU cycles to process the query. Additionally, we have a scale-out MPP architecture, which means you can add more nodes to Vertica.

All of these elements are extremely critical to handle the data volume challenge. With Vertica, customers can load several terabytes of data quickly (per hour in fact) and query their data within minutes of it being loaded—that is real-time analytics on Big Data for you.

There is a myth that columnar databases are slow to load. This may have been true with older generation column stores, but in Vertica, we have a hybrid in-memory/disk load architecture that rapidly ingests incoming data into a write-optimized row store and then converts that to read-optimized sorted columnar storage in the background. This is entirely transparent to the user because queries can access data in both locations seamlessly. We have a very lightweight transaction implementation with snapshot isolation queries can always run without any locks.

And we have no auxiliary data structures, like indices or materialized views, which need to be maintained postload. Last, but not least, we designed the system for "always on," with built-in high availability features. Operations that translate into downtime in traditional databases are online in Vertica, including adding or upgrading nodes, adding or modifying database objects, etc. With Vertica, we have removed many of the barriers to monetizing Big Data and hope to continue to do so.

"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 get 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 an 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 [4])

A NewSQL Data Store for OLTP (VoltDB) Connected with Hadoop or a Data Warehouse

With this solution, a fast NewSQL data store designed for OLTP (VoltDB) is connected to either a conventional data warehouse or Hadoop.

"We identified 4 sources of significant OLTP overhead (concurrency control, write-ahead logging, latching and buffer pool management).

Unless you make a big dent in ALL FOUR of these sources, you will not run dramatically faster than current disk-based RDBMSs. To the best of my knowledge, VoltDB is the only system that eliminates or drastically reduces all four of these overhead components. For example, TimesTen uses conventional record level locking, an Aries-style write ahead log and conventional multi-threading, leading to substantial need for latching. Hence, they eliminate only one of the four sources.

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.

VoltDB supports standard SQL. Complex joins should be run on a companion data warehouse. After all, the only way to interleave 'big reads' with 'small writes' in a legacy RDBMS is to use snapshot isolation or run with a reduced level of consistency. You either get an out-of-date, but consistent answer or an up-to-date, but inconsistent answer. Directing big reads to a companion DW, gives you the same result as snapshot isolation. Hence, I do not see any disadvantage to doing big reads on a companion system.

Concerning larger amounts of data, our experience is that OLTP problems with more than a few Tbyte of data are quite rare. Hence, these can easily fit in main memory, using a VoltDB architecture.

In addition, we are planning extensions of the VoltDB architecture to handle larger-than-main-memory data sets." (Mike Stonebraker [13])

A NewSQL for Analytics (Hadapt) Complementing Hadoop

An alternative solution is to use a NewSQL designed for analytics (Hadapt) which complements Hadoop.

Daniel Abadi explains "at Hadapt, we're bringing 3 decades of relational database research to Hadoop. We have added features like indexing, co-partitioned joins, broadcast joins, and SQL access (with interactive query response times) to Hadoop, in order to both accelerate its performance for queries over relational data and also provide an interface that third party data processing and business intelligence tools are familiar with.

Therefore, we have taken Hadoop, which used to be just a tool for super-smart data scientists, and brought it to the mainstream by providing a high performance SQL interface that business analysts and data analysis tools already know how to use. However, we've gone a step further and made it possible to include both relational data and non-relational data in the same query; so what we've got now is a platform that people can use to do really new and innovative types of analytics involving both unstructured data like tweets or blog posts and structured data such as traditional transactional data that usually sits in relational databases.

What is special about the Hadapt architecture is that we are bringing database technology to Hadoop, so that Hadapt customers only need to deploy a single cluster—a normal Hadoop cluster—that is optimized for both structured and unstructured data, and is capable of pushing the envelope on the type of analytics that can be run over Big Data." [10]

A Combinations of Data Stores: A Parallel Database (Teradata) and Hadoop

An example of this solution is the architecture for Complex Analytics at eBay (Tom Fastner [12])

The use of analytics at Ebay is rapidly changing, and analytics is driving many key initiatives like buyer experience, search optimization, buyer protection, or mobile commerce. EBay is investing heavily in new technologies and approaches to leverage new data sources to drive innovation.

EBay uses three different platforms for analytics:

1. "*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 worldwide; 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.
2. "*Singularity*": deep Teradata system for *semistructured 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.

3. *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.

The main technical challenges for Big Data analytics at eBay are

- *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 bottleneck 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 sometime 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.

Scalability and Performance at eBay

- *EDW*: models for the unknown (close to third NF) to provide a solid physical data model suitable for many applications, which 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 they are leveraging the latest database release, they 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 they developed functions to support the analysis of the semistructured 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 them 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. There usability of data structures (besides "raw" data) is very low

Unstructured Data

Unstructured data are handled on Hadoop only. The data are copied from the source systems into HDFS for further processing. They do not store any of that on the Singularity (Teradata) system.

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
- *Ad hoc*: Teradata SQL Assistant; Hadoop Pig and Hive
- *Content Management*: Joomla-based

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, Principal Architect at eBay)

Big Data Myth

It is interesting to report here what Marc Geall, a research analyst at Deutsche Bank AG/in London, writes about the "Big Data Myth," and predicts [9]:

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

Time will tell us whether this prediction is accurate or not.

Main Research Challenges and Business Challenges

We conclude this part of the chapter by looking at three elements: data, platform, and analysis with two quotes:

Werner Vogels: "I think that sharing is another important aspect to the mix. Collaborating during the whole process of collecting data, storing

it, organizing it and analyzing it is essential. Whether it's scientists in a research field or doctors at different hospitals collaborating on drug trials, they can use the cloud to easily share results and work on common datasets."

Daniel Abadi: "Here are a few that I think are interesting:

1. *Scalability of non-SQL analytics*. How do you parallelize clustering, classification, statistical, and algebraic functions that are not 'embarrassingly parallel' (that have traditionally been performed on a single server in main memory) over a large cluster of shared-nothing servers.
2. Reducing the cognitive complexity of 'Big Data' so that it can fit in the working set of the brain of a single analyst who is wrangling with the data.
3. Incorporating graph data sets and graph algorithms into database management systems.
4. Enabling platform support for probabilistic data and probabilistic query processing."

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)

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?

In the following, I will list some relevant initiatives and selected thoughts for Big Data for the Common Good.

World Economic Forum, 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 policy-makers 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.

"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."

Three examples are cited in WEF paper:

- *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 better understanding of changes in human well-being (www.unglobalpulse.orgGlobal).
- *Viral Forecasting*: a not-for-profit whose mission is to promote understanding, exploration, and stewardship of the microbial world (www.gvfi.orgUshadi).
- *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?

I have listed below some extracts from [5]:

- 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 Research): "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)."

- 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?

I have listed below some extracts from [5]:

- Alon Havely (Google Research): "Building a collection of high quality data that is widely available and can serve as the backbone for many specific data projects. For example, datasets 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 Haas (IBM Research): "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!"

- Paul Miller (Consultant): "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 Barga (Microsot Research): "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's peed 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."

Conclusions: The Search for Meaning Behind Our Activities

I would like to conclude this chapter with this quote below which I find inspiring.

> "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**

Acknowledgments

I would like to 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, Mike OSullivan, Martin Verlage, and Steve Vinoski for their feedback on an earlier draft of this chapter.

But all errors and missing information are mine.

References

1. McKinsey Global Institute (MGI), Big Data: The next frontier for innovation, competition, and productivity, Report, June, 2012.
2. Managing Big Data. An interview with David Gorbet ODBMS Industry Watch, July 2, 2012. http://www.odbms.org/blog/2012/07/managing-big-data-an-interview-with-david-gorbet/
3. On Big Data: Interview with Dr. Werner Vogels, CTO and VP of Amazon.com. *ODBMS Industry Watch,* November 2, 2011. http://www.odbms.org/blog/2011/11/on-big-data-interview-with-dr-werner-vogels-cto-and-vp-of-amazon-com/
4. On Big Data: Interview with Shilpa Lawande, VP of Engineering at Vertica. *ODBMs Industry Watch*, November 16, 2011.
5. "Big Data for Good", Roger Barca, Laura Haas, Alon Halevy, Paul Miller, Roberto V. Zicari. *ODBMS Industry Watch*, June 5, 2012.
6. On Big Data Analytics: Interview with Florian Waas, EMC/Greenplum. *ODBMS Industry Watch*, February 1, 2012.
7. Next generation Hadoop—interview with John Schroeder. *ODBMS Industry Watch*, September 7, 2012.
8. Michael J. Carey, EDBT keynote 2012, Berlin.
9. Marc Geall, "Big Data Myth", Deutsche Bank Report 2012.
10. On Big Data, Analytics and Hadoop. Interview with Daniel Abadi. *ODBMS Industry Watch*, December 5, 2012.
11. Hadoop and NoSQL: Interview with J. Chris Anderson. *ODBMS Industry Watch*, September 19, 2012.
12. Analytics at eBay. An interview with Tom Fastner. *ODBMS Industry Watch*, October 6, 2011.
13. Interview with Mike Stonebraker. *ODBMS Industry Watch*, May 2, 2012.

Links:

ODBMS.org www.odbms.org
ODBMS Industry Watch, www.odbms.org/blog