The Big Data riddle: what is the ideal solution?

Let's reflect about this for a while:can we figure out-on our own- the number of tweets posted every second, the number of hash-tags generated, the amount of blogs created, how many Facebook posts generated, the number and size of photos and videos shared on Facebook, the number of emails exchanged? It is staggering.It is next to impossible.

Suppose we were to wonder about this too: what individual people right now are saying about an organization across the Social Media,who is happy or disappointed with the services or products of a particular organization, and so on. We would end up considering the whole affair unthinkable.

And this: how many are currently carrying out fraudulent transactions using credit cards? Which product is currently trending in the market and why? What about the amount of data generated from sensors through satellite per second? Can we detect a meteorite that is just about to hit the earth? Which is the best stock to buy from the Nairobi Stock Exchange? That also will lead us to conclude the answer unimaginable.

If you are exhausted already, well, we have just been thinking about 'Big Data'! It is a term used to describe the "exponential growth and availability of data that exists in both structured and unstructured formats".

Much more clearly, this catch-phrase denotes a massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques we are currently aware of. In most enterprise scenarios, the data is too big or it moves too fast or it exceeds current processing capacity.

'Big data' can also be termed as a collection of data sets that are so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Elements of Big Data

An industry analyst Doug Laney articulated the now mainstream definition of Big Data as the three Vs of Big Data: volume, velocity and variety.

Volume: Volume is basically the size on storage space. Many factors contribute to the increase in data volume. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. In the past,

excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

Velocity: This is the speed at which the data streams in. Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

Variety: Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications.

Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with.

Big player of Big Data, SAS, also considers two additional dimensions when thinking about Big Data:

Variability. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

Complexity: Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

Why Big Data is the issue

The real issue is not that were are acquiring large amounts of data. It is what you do with the data that counts. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyze it to find answers that enable cost reductions, time reductions, new product development and optimized offerings, and smarter business decision making. For instance, by combining big data and high-powered 'analytics', it is possible to achieve the following feats:

To determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.
Optimize routes for many thousands of package delivery vehicles while they are on the road.

Analysis of millions of SKUs to determine prices that maximize profit and clear inventory,generate retail coupons at the point of sale based on the customer's current and past purchases,send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.

Others are to recalculate entire risk portfolios in minutes to quickly identify customers who matter the most as well as use 'click-stream analysis' and 'data mining' to detect fraudulent behavior.

But,in fact,the challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization.In other words, the challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation.

These technical challenges are common across a large variety of application domains. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data

The case for taming Big Data

But the trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.

The evidence

And as of 2012, data in order of millions of terabytes (exabytes) could be feasibly processed in a reasonable amount of time. IBM estimates 2.5 quintillion bytes of data being generated daily and ninety percent of data in the world is less than two years old.

Large data sets limit scientists in many areas, such as meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics.

Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks.

Further,the world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×1018) of data were created!

Big Data technologies

However, sophisticated technologies are coming up to help users cope with and handle the Big Data menace in a cost-effective manner. At a glance,some of them are as follows:

Column-oriented databases: Traditional, row-oriented databases are excellent for online transaction processing with high update speeds, but they fall short on query performance as the data volumes grow and as data become more unstructured. Column-oriented databases store data with a focus on columns, instead of rows, allowing for huge data compression and very fast query times. The downside to these databases is that they will generally only allow batch updates, having a much slower update time than traditional models.

Schema-less databases(NoSQL): There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

MapReduce: This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. Any MapReduce implementation consists of two tasks:the "Map" task, where an input dataset is converted into a different set of key/value pairs or 'tuples';

Another one is where several of the outputs of the “Map” task are combined to form a reduced set of ‘tuples’ (hence the name).

Hadoop: Hadoop is by far the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources, either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.

Hive: it is an “SQL-like” bridge that allows conventional Business Intelligence applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it’s a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for Business Intelligence users.

PIG: PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a “Perl-like” language that allows for query execution over data stored on a Hadoop cluster, instead of an “SQL-like” language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

WibiData: this is a combination of web analytics with Hadoop, being built on top of HBase, which is itself a database layer on top of Hadoop. It allows web sites to better explore and work with their user data, enabling real-time responses to user behavior, such as serving personalized content, recommendations and decisions.

PLATFORA: the greatest limitation of Hadoop is that it is a very low-level implementation of MapReduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional databases. PLATFORA is a platform that turns user’s queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

Storage Technologies: As the data volumes grow, so does the need for efficient and effective storage techniques. The main evolutions in this space are related to data compression and storage ‘virtualization’.

SkyTree: SkyTree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data. Machine learning, in turn, is an essential part of Big Data, since the massive data volumes make manual exploration, or even conventional automated exploration methods unfeasible or too expensive.