IQR LIVE PROJECT

FINAL REPORT

2015

Table of Contents

[OPEN SOURCE TOOLS 4](#_Toc415657847)

[Visualization applications and services 5](#_Toc415657848)

[BIG DATA 6](#_Toc415657849)

[Changing Data Management Landscape 6](#_Toc415657850)

[Big Data vs. Traditional Analytics 8](#_Toc415657851)

[New Approaches to Big Data Processing and Analytics 10](#_Toc415657852)

[1)Hadoop 10](#_Toc415657853)

[2) NoSQL 11](#_Toc415657854)

[3) Massively Parallel Analytic Databases 12](#_Toc415657855)

[Ten Big Data core technologies 12](#_Toc415657856)

[Big Data Landscape 13](#_Toc415657857)

[Big Data Vendor Landscape 13](#_Toc415657858)

[TCS Survey Findings: 15](#_Toc415657859)

[Functions of Big Data in various industries: 19](#_Toc415657860)

[Barrier 19](#_Toc415657861)

[Future Trends for 2015: 20](#_Toc415657862)

[Bigdata SV# 20](#_Toc415657863)

[Sources : 22](#_Toc415657864)

**OBJECTIVES OF THE REPORT:**

To study the major analytical tools and capabilities in Big Data:

1. To analyze the changing landscape from traditional static data to Big Data- 3 Vs.
2. To study the core technologies of Big Data.
3. To study the transformation in big data market segments.
4. To analyze geography wise and industry wise adoption of Big Data services.

Along with these objectives, certain visualization tools have also been covered.

**METHODOLOGY:** Secondary Research

**PROJECT HEAD:** Mr. Anand Maraiya

**SUBMITTED BY:**

Amit Bhatia- Batch of 2015

Ashdeep Kaur- Batch of 2016

International Management Institute

# OPEN SOURCE TOOLS

**Data cleaning**

Before you can analyse and visualize data, it often needs to be "cleaned." There might be some records with misspellings or numerical data-entry errors. The following two tools are designed to help get your data in tip-top shape to be analysed.

**[Data Wrangler](http://vis.stanford.edu/wrangler/" \t "_blank)**

This Web-based service from Stanford University's Visualization Group is designed for cleaning and rearranging data so it's in a form that other tools such as a spreadsheet app can use. It can be used on any web browser.

[**Google Refine**](http://code.google.com/p/google-refine/)

Google Refine can be described as a spreadsheet on steroids for taking a first look at both text and numerical data. Like Excel, it can import and export data in a number of formats including tab- and comma-separate text files and Excel, XML and JSON files.

Refine features several built-in algorithms that find text items that are spelled differently but actually should be grouped together. After importing your data, select which algorithm you want to use. After Refine runs, you decide whether to accept or reject each suggestion.

There are also numerical options that offer quick and easy overviews of data distributions. This functionality can reveal anomalies that might be the result of data input errors -- such as $800,000 instead of $80,000 for a salary entry, or it could expose inconsistencies -- such as differences in the way compensation data is reported from entry to entry, with some showing, say, hourly wages and others showing weekly pay or yearly salaries.Beyond data housekeeping, Google Refine offers some useful analysis tools, such as sorting and filtering.

**R**

R is a general statistical analysis platform that runs on the command line. Need to find means, medians, standard deviations, correlations? R can handle that and much more, including "linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing," according to the [project website](http://www.r-project.org/).

R also graphs, charts and plots results. There are numerous add-ons to this open-source project that significantly extend functionality. For users who prefer a GUI, Peter Aldhous, San Francisco bureau chief for *New Scientist* magazine, suggests [RExcel](http://rcom.univie.ac.at/download.html" \t "_blank), which offers access to the R engine through Excel.

There is a great deal of functionality in R, including quite a number of visualization options as well as numerical and [spatial analysis](http://www.spatstat.org/spatstat/).

## Visualization applications and services

[**Google Fusion Tables**](http://www.google.com/fusiontables/Home)

This is one of the simplest ways I've seen to turn data into a chart or map. You can upload a file in several different formats and then choose how to display it: table, map, heatmap, line chart, bar graph, pie chart, scatter plot, timeline, storyline or motion (animation over time). It's somewhat customizable, allowing you to change map icons and style info windows.

[**Tableau Public**](http://www.tableausoftware.com/public)

This tool can turn data into any number of visualizations, from simple to complex. You can drag and drop fields onto the work area and ask the software to suggest a visualization type, then customize everything from labels and tool tips to size, interactive filters and legend display.

Tableau Public offers a variety of ways to display interactive data. You can combine multiple connected visualizations onto a single dashboard, where one search filter can act on numerous charts, graphs and maps; underlying data tables can also be joined. And once you get the hang of how the software works, its drag-and-drop interface is considerably quicker than manually coding in JavaScript or R for most users, making it more likely that you'll try additional scenarios with your data set. In addition, you can easily perform calculations on data within the software.

[**Many Eyes**](http://www-958.ibm.com/software/data/cognos/manyeyes/)

A pioneer in Web-based data visualization, IBM's Many Eyes project combines graphical analysis with community, encouraging users to upload, share and discuss information. It's extremely easy to use and very well documented, including suggestions on when to use what kind of visual data representation. Many Eyes includes more than a dozen output options -- from charts, graphics and word clouds to treemaps, plots, network diagrams and some limited geographic maps.

You'll need a free account to upload and post data, although anyone can browse. Formatting is basic: For most visualizations, the data must be in a tab-separated text file with column headers in the first row.

**VIDI**

Although VIDI's website bills this as a tool for the Drupal content management system, graphics created by the site's visualization wizard can be used on any HTML page -- no Drupal required. Upload your data, select a visualization type, do a bit of customization selection, and your chart, timeline or map is ready to use via auto-generated embed code (using an iframe, not JavaScript or Flash).

**Zoho Reports**

What it does: One of the more traditional corporate-focused business analytics offerings in this group, Zoho Reports can take data from various file formats or directly from a database and turn it into charts, tables and pivot tables -- formats familiar to most spreadsheet users.You can schedule data imports from sources on the Web. Data can be queried using SQL and can be turned into visualizations, and the service is set up for Web publishing and sharing.

# BIG DATA

Universe Beginning to Explode

**Data is becoming the world’s new natural resource.**

Today, every discussion about changes in technology, business and society must begin with data. In exponentially increasing volume, velocity and variety, data is becoming a new natural resource. It promises to be for the 21st century what steam power was for the 18th, electricity for the 19th and hydrocarbons for the 20th.

Big data continues to grow. How big has it gotten? Consider a few facts:

* More than 2.5 quintillion bytes of data are created every day through a range of activities including social media posts, purchase transaction records, cell phone GPS signals, supply chain and logistics data, and digital videos, pictures, and audio recordings.
* Individuals generate more than 70 percent of all data; enterprises store and manage 80 percent of this data
* Hiring growth is part of this spending. It is estimated that for every IT job created in the United States to support big data efforts, three big data-related jobs outside of IT are being generated. This equals literally hundreds of thousands, potentially millions, of new jobs.
* Global spending on big data is growing at an average annual rate of nearly 30 percent and is expected to reach $114 billion in 2018.

Big data is changing our lives and changing the way we do business. Some examples include the following:

• Google uses big data to predict the next wave of influenza.

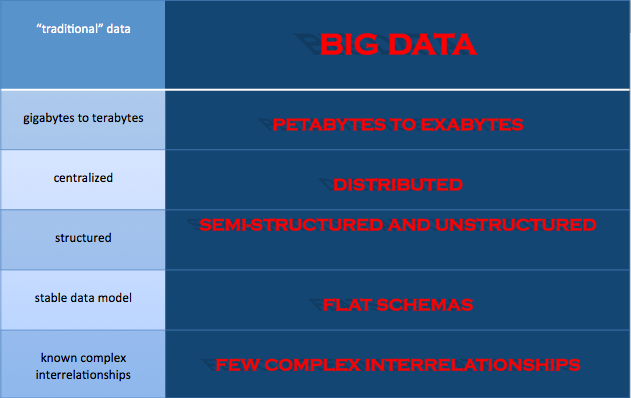
• IBM uses data to optimize traffic flow in the city of Stockholm and to get the best possible air quality.

• Dr. Jeffrey Brenner, a physician in New Jersey, uses medical billing data to map out hot spots where you can find his city’s most complex and costly healthcare cases as part of a program to lower healthcare costs.

• The National Centre for Academic Transformation is using data mining to help understand which college students are more likely to succeed in which courses.

## Changing Data Management Landscape

Traditional data management and business analytics tools and technologies are straining under the added weight of Big Data and new approaches are emerging to help enterprises gain actionable insights from Big Data. These new approaches – namely an open source framework called Hadoop, NoSQL databases such as Cassandra and Accumulo, and massively parallel analytic databases from vendors like EMC Greenplum, HP Vertica, and Teradata Aster Data – take a radically different approach to data processing, analytics and applications than traditional tools and technologies. This means enterprises likewise need to radically rethink the way they approach business analytics both from a technological and cultural perspective.



**The Data Deluge and Use**

Organizations are experiencing unprecedented data growth.Executives say they’ve seen the biggest increase in information variety and volume in the following areas in the last two years:

48% Customer information

34% Operations

33% Sales/marketing

Big data is a key to revenue growth. 93% of the executives believe they are losing revenue at an average rate of 14% annually. The present use of Big Data as follows:

1. Monitoring markets to identify sales opportunities

2. Creating personalized offers

3. Recognizing repeat customers at an early stage (churn)

4. Recruiting staff

5. Forecasting sales, for planning and control

6. Predicting wear and maintenance

7. Directing management, decision-making & control

8. Detecting fraud

9. Estimating financial risks part ii Your Big Data Potential

10. Recognizing cyber attacks

11. Improving products

12. Developing innovative products

Big data is misunderstood. Big data requires an industry-specific business strategy to improve the effectiveness of operational processes and the overall customer experience.

## Big Data vs. Traditional Analytics

The exponential advancement of social media, cloud and Mobile in today’s world is pressuring institutions to evolve into data centric organizations. There are several challenges in this regard that the enterprises are faced today owing to the limitations posed by usage of traditional data analytics. Some of these are:

• Unstructured data that could provide a real-time business decision support remains unused as they cannot be stored, processed or analyzed.

• Several data islands are created and it becomes difficult to generate meaningful information from those.

• Data models are non-scalable and data becomes unmanageable.

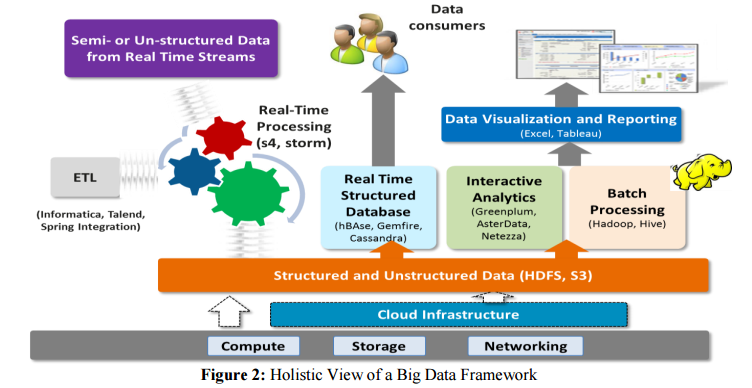
• The cost of managing the data increases exponentially with the growth of data.



These challenges can be overcome through Big Data as it turns the vast world of data into incredible opportunities and insights. Big data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings. Like traditional analytics, it can also support internal business decisions. It is based on massively parallel processing and NoSQL approach to data, but almost SQL compliant Hadoop framework (function to data model).

* Cost Reduction: Some organizations pursuing big data believe strongly that MIPS and terabyte storage for structured data are now most cheaply delivered through big data technologies like Hadoop clusters. One company’s cost comparison, for example, estimated that the cost of storing one terabyte for a year was $37,000 for a traditional relational database, $5,000 for a database appliance, and only $2,000 for a Hadoop cluster.
* Develop new big data based offerings: Companies can employ it in developing new product and service offerings based on data. The best example may be LinkedIn, which has used big data and data scientists to develop a broad array of product offerings and features, including People You May Know, Groups You May Like, Jobs You May Be Interested In, Who’s Viewed My Profile, and several others
* Support internal business decisions:Three major banks like Wells Fargo, Bank of America, and Discover are also using big data to understand aspects of the customer relationship that they couldn’t previously get at.
* Time reduction with rapid data capture, aggregation, processing, and analytics

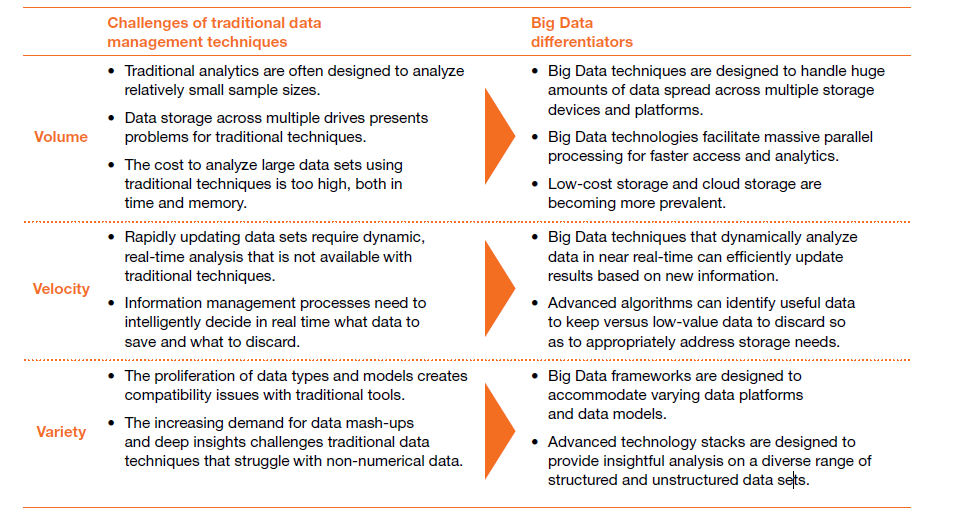
.



There are 3 key attributes of Big Data referred as 3 Vs. Traditional data management practices cannot accommodate these attributes. These 3 Vs are:

* **Volume:** The vast quantity of data available in terabytes.For example, Walmart collects more than 2.5 petabytes of data every hour from customer transactions alone.
* **Velocity:** Speed at which data can be stored and analysed.Real-time or near real-time information access enables organizations to be quicker in making decisions and executing moves than their competitors.
* **Variety:** Variation in type and sources of Big Data.It can come in the form of images posted on Facebook, email, text messages, GPS signals from mobile phones, tweets and other social media updates. It can be structured or non structured.

Keeping in mind the data deluge,its time that institutions gear to equip themselves for accomodating the 3 Vs



## New Approaches to Big Data Processing and Analytics

Wikibon has identified three Big Data approaches that are believed to be transforming the business analytics and data management markets.

### 1)Hadoop

Hadoop is an open source framework for processing, storing and analyzing massive amounts of distributed unstructured data. Originally created by Doug Cutting at Yahoo!, Hadoop was inspired by MapReduce, a user-defined function developed by Google in early 2000s for indexing the Web. It was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel.

Hadoop clusters run on inexpensive commodity hardware so projects can scale-out without breaking the bank. Hadoop is now a project of the Apache Software Foundation, where hundreds of contributors continuously improve the core technology. Fundamental concept: Rather than banging away at one, huge block of data with a single machine, Hadoop breaks up Big Data into multiple parts so each part can be processed and analyzed at the same time

**Pros and Cons**  
The main benefit of Hadoop is that it allows enterprises to process and analyze large volumes of unstructured and semi-structured data, heretofore inaccessible to them, in a cost- and time-effective manner. Because Hadoop clusters can scale to petabytes and even exabytes of data, enterprises no longer must rely on sample data sets but can process and analyze ALL relevant data. Data scientists can apply an iterative approach to analysis, continually refining and testing queries to uncover previously unknown insights. It is also inexpensive to get started with Hadoop. Developers can download the Apache Hadoop distribution for free and begin experimenting with Hadoop in less than a day.

The downside to Hadoop and its myriad components is that they are immature and still developing. As with any young, raw technology, implementing and managing Hadoop clusters and performing advanced analytics on large volumes of unstructured data requires significant expertise, skill and training. Unfortunately, there is currently a dearth of Hadoop developers and data scientists available, making it impractical for many enterprises to maintain and take advantage of complex Hadoop clusters. Further, as Hadoop’s myriad components are improved upon by the community and new components are created, there is, as with any immature open source technology/approach, a risk of forking. Finally, Hadoop is a batch-oriented framework, meaning it does not support real-time data processing and analysis.

### 2) NoSQL

A related new style of database called NoSQL (Not Only SQL) has emerged to, like Hadoop, process large volumes of multi-structured data. However, where as Hadoop is adept at supporting large-scale, batch-style historical analysis, NoSQL databases are aimed, for the most part (though there are some important exceptions) at serving up discrete data stored among large volumes of multi-structured data to end-user and automated Big Data applications. This capability is sorely lacking from relational database technology, which simply can't maintain needed application performance levels at Big Data scale.

In some cases, NoSQL and Hadoop work in conjunction. HBase, for example, is a popular NoSQL database modeled after Google BigTable that is often deployed on top of HDFS, the Hadoop Distributed File System, to provide low-latency, quick lookups in Hadoop.

NoSQL databases currently available include:

* HBase
* Cassandra
* MarkLogic
* Aerospike
* MongoDB
* Accumulo
* Riak
* CouchDB
* DynamoDB

The downside of most NoSQL databases today is that they traded ACID (atomicity, consistency, isolation, durability) compliance for performance and scalability. Many also lack mature management and monitoring tools. Both these shortcomings are in the process of being overcome by both the open source NoSQL communities and a handful of vendors -- such as DataStax, Sqrrl, 10gen, Aerospike and Couchbase -- that are attempting to commercialize the various NoSQL databases.

### 3) Massively Parallel Analytic Databases

Unlike traditional data warehouses, massively parallel analytic databases are capable of quickly ingesting large amounts of mainly structured data with minimal data modelling required and can scale-out to accommodate multiple terabytes and sometimes petabytes of data.

### Ten Big Data core technologies

The processing stages that apply to Business Intelligence applications also apply to Big Data, but demand extra technological effort to enable the complete process of data capturing, storage, search, sharing, analytics and visualization to occur smoothly. In his book entitled Big Data Glossary, Pete Warden discusses a total of sixty technological innovations and provides the following concise overview of Big Data concepts and tools.

1 **Data acquisition** (such as Google Refine, Needlebase, ScraperWiki, BloomReach, for example): For accessing various data sources, internal or external, structured or unstructured. Most interesting public data sources are poorly structured, full of contamination and difficult to open.

2 **Serialization** (such as json, bson, Thrift, Avro, Google Protocol Buffers, for example): At various points during processing, the data will be stored in files. These operations all require some sort of ranking.

3 **Storage** (such as Amazon S3, Hadoop Distributed File System, for example): Traditional file systems are unsuited to large-scale, distributed data processing, but nevertheless the Hadoop Distributed File System is actually suited to this function.

4 **Cloud** (such as Amazon ec2, Windows Azure, Google App Engine, Amazon Elastic Beanstalk, Heroku, for example): Hiring computers as virtual machines in a cloud environment is increasingly becoming standard procedure. In this way, you can make use of large processing capacity for Big Data applications at relatively low cost.

5 **Nosql** (such as Apache Hadoop, Apache Cassandra, Mongodb, Apache Couchdb, Redis, BigTable, hbase, Hypertable, Voldemor):Nosql is a broad class of management systems that deviates from the classical relational model.

6 **MapReduce** (such as Hadoop and Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum, for example): In traditional relational database environments, all processing takes place via a special query language after the structured information has been loaded. In contrast, MapReduce reads and writes unstructured data to all sorts of file formats. The interim results are passed on as files, and the processing is divided among many machines.

7 **Processing** (such as R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, Bigsheets, Tinkerpop; start-ups: Continuuity, Wibidata, Platfora, for example): Filtering concise, valuable information from an ocean of data is a challenge, but there are already many solutions that can help with such tasks.

8 **Natural Language Processing** (such as Natural Language Toolkit, Apache Opennlp, Boilerpipe, OpenCalais, for example): Natural Language Processing extracts meaningful information from untidy, humanbased language.

9 **Machine learning** (such as weka, Mahout, scikits.learn, Skytree, for example) Machine Learning systems automate and optimize decision-making. The recommendations given by Amazon, for instance, are well-know applications of this function.

10 **Visualization** (such as GraphViz, Protovis, Google Fusion Tables, Tableau Software, for example): This is one of the best ways to extract significance from data. Thanks to interactive graphics, the presentation and exploration of information blend together.

## Big Data Landscape

This is the early innings of Big Data market. Over the last couple of years, some promising companies failed (for example: Drawn to Scale), a number saw early exits (for example: Precog, Prior Knowledge, Lucky Sort, Rapleaf, Nodeable, Karmasphere), and a handful saw more meaningful outcomes (for example: Infochimps, Causata, Streambase, ParAccel, Aspera, GNIP, BlueFin labs, BlueKai).

Meanwhile, some companies reached significant scale and have raised spectacular amounts of money .For example, MongoDB has now raised over $230M, Palantir almost $900M, and Cloudera $1B. But overall, it is still early in the curve in terms of successful IPOs (Splunk or Tableau notwithstanding) and large exits. At the same time, the big companies are getting more acquisitive in the space (Oracle with BlueKai, IBM with Cloudant). In many segments, start-ups and large companies are jockeying for position and no obvious leader has emerged.Some pure Bigdata players [MapR](http://www.mapr.com/), [MemSQL](http://www.memsql.com/),[Databricks](http://databricks.com/), [Platfora](http://www.platfora.com/), [Splunk](http://www.splunk.com/), [Teradata](http://www.teradata.com/), [Palantir](http://www.palantir.com/), [Premise](http://www.premise.com/),[Datameer](http://www.datameer.com/), [Cloudera](http://www.cloudera.com/content/cloudera/en/home.html), [Hortonworks](http://hortonworks.com/), [MongoDB](http://www.mongodb.org/), and[Trifacta](http://www.trifacta.com/) have been shining in innovative space with big analytical companies like Dell,HP,IBM earning the highest revenue.

While big data is becoming less press worthy, the next couple of years are going to be hugely important for this market, as corporations start moving projects from experimentation to full production. While those deployments will lead to rapidly increasing revenues for some big data vendors, they will also test whether big data can truly deliver on its promise.

### Big Data Vendor Landscape

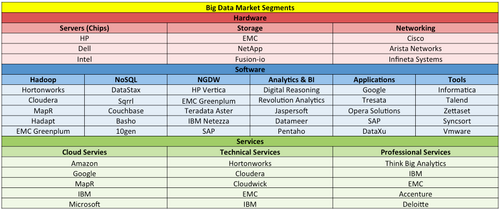
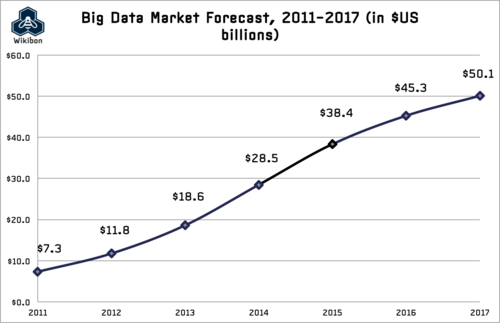
The Big Data vendor landscape is developing rapidly. The Big Data market is dynamic, with companies going public, others like Hadapt and Revelytix, Inc. [being acquired](http://www.teradata.com/News-Releases/2014/Teradata-Acquisitions-Accelerate-Big-Data-Innovation/) by Teradata Corp., and a lot of technology innovation happening. Large amounts of money have also been changing hands. For example, last year we saw Intel Corp., one of the industry’s biggest Big Data players,[invest $740 million](http://siliconangle.com/blog/2014/03/31/cloudera-raises-monster-round-900-million-740-million-from-intel-for-18/) into Cloudera, Inc. We also saw MapR Technologies, Inc.[complete a $110 million round](https://www.mapr.com/company/press-releases/mapr-closes-110-million-financing-led-google-capital) of financing. [](http://wikibon.org/vault/Special:FilePath/Manifesto-mksegBD.png)

Figure 4 - Big Data Vendor Landscape  
Source: Wikibon 2012

The Big Data market as measured by vendor revenue derived from sales of related hardware, software and services reached $18.6 billion in calendar year 2013. That represents a growth rate of 58% over [the previous year.](http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017)

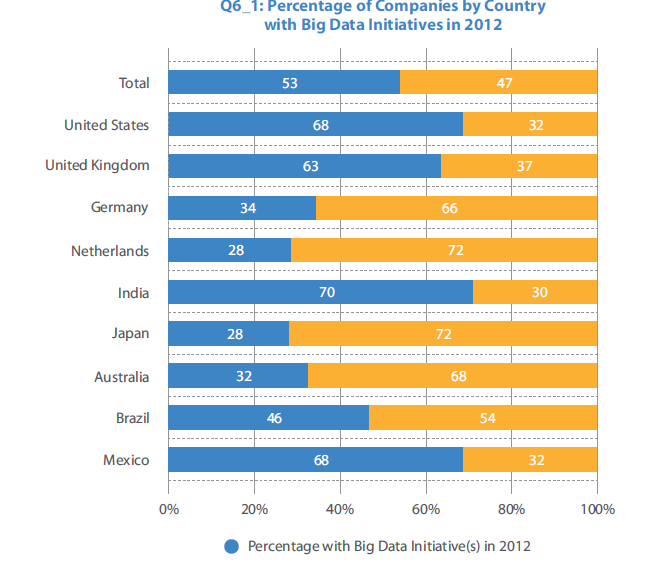
Broken down by type, Big Data-related services revenue made up 40% of the total market, followed by hardware at 38% and software at 22%. Such a breakdown is due in part to the open source nature of much Big Data software and related [business models of Big Data vendors](http://wikibon.org/wiki/v/Hadoop_Pure-Play_Business_Models_Explained), as well as the need for professional services to help enterprises identify Big Data uses cases, architect solutions and maintain performance.

[](http://wikibon.org/vault/Special:FilePath/BigDataMarketForecast2013.png)

Faster access to more relevant data and constant experimentation is creating a further gap between leaders and the rest of the organisations. It is also creating new challenges for IT and business leaders tasked with their organisation’s big data and analytics strategy and execution.

## TCS Survey Findings:

Across all four regions of the world that TCS surveyed, 53% of the 1,217 companies said they had undertaken at least one Big Data initiative in 2012. The U.S. was the leader among regions in Big Data use.



The investments these companies made in Big Data were sizable. Median spending on Big Data was $10 million, which was 0.14% of revenue. By the year 2015, companies across the surveyed regions expect to spend 75% more on Big Data, with Australia and U.K. companies projecting the highest spending per company.

**Highlights:**

U.S. companies are in the lead in using Big Data

Median per-company spending on Big Data in 2012 was $10 million

80% of the companies say they’ve improved business decisions as a result

Most of their data is structured (versus non-structured) and from internal sources.

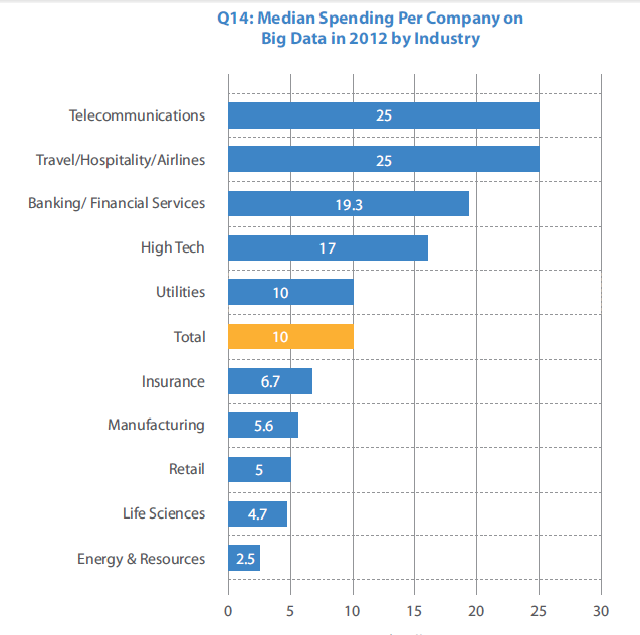
Only 27% of the companies sell their digital data, but those that did generated an average of $22 million from it in 2012

About 24% of the companies either expected a negative return or haven’t tried toquantify one for 2012

Companies that projected an ROI averaged a 46% ROI, which includes those projecting a negative return

The most difficult challenge in generating benefits from Big Data is organizational: getting organizational silos to share their data

Various industries have invested heavily in informal on technology over many decades, they have different levels of data intensity – that is, the volume, variety, and velocity with which digital data courses through their informal on systems. Retail chains, for instance, must process data on purchases of hundreds of millions of items from their stores. And they must do so rapidly for right pricing and product availability to regional or local demand. In contrast, auto manufacturers sell thousands of cars to independent dealers every week, and thus have a less onerous task of understanding sales trends.

T

There are industries in which only a small minority of companies started to pull ahead, thus industries where the majority need a wake-up call to increase their revenue. Through the survey, it was found a small minority of companies with projected ROIs onBig Data of more than 50% in four sectors:

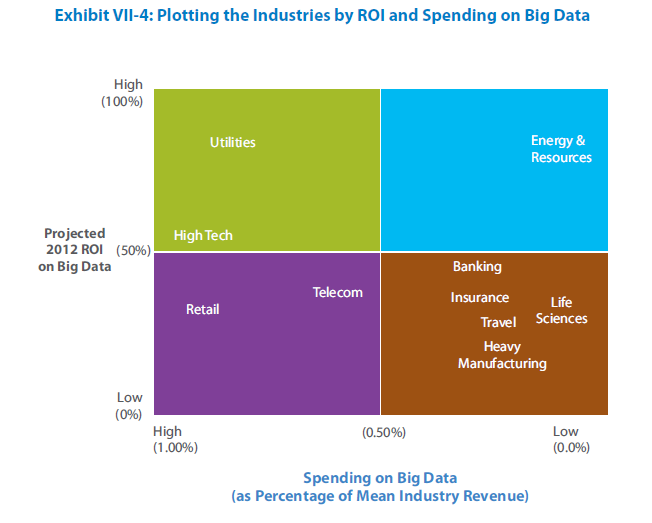
Consumer goods (in only 9% of these companies did functional managers reportROI >50%)

Utilities (15%)

Insurance (17%)

Media and entertainment (19%)

That’s a sign that the clear majority of companies in these industries have a long way togo to catch up to the leaders. Other industries are well listed in following plot.



Bigdata presents opportunities in three broad categories:

**Revenue**: These investments will most often fall in marketing, sales, service, andR&D/product development

**Risk reduction**: Finance and supply chain operations are full of opportunities todetect risk, from spotting customers whose credit rating is likely to suffer totransportation routes and locations that face heightened theft risks

**Operational effectiveness**: These investments enable companies to improve theway they engage with customers and make every interaction more effective. Thiscamp includes activities that may not directly lead to more revenue but whichare important to keeping customers satisfied. Consider investments such as getting tothe root of customer service problems, manufacturing process flaws, and supplychain bottlenecks.

Companies that create a portfolio of big data initiatives in these three categories and thenimplement them over time will reduce the chances of overlooking good opportunities, especially less obvious ones that don’t directly contribute to revenue.

## Functions of Big Data in various industries:

Big Data solutions help various companies to gain expertise in certain functions like

**Energy and Utilities:** Smart meter analytics, Asset management

**Digital Media:** Real-time ad targeting, Website analysis

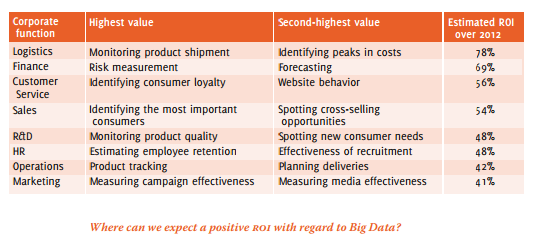
**Retail:** Omni-channel marketing, Click-stream analysis

**Law Enforcement:** Real-time multimodal surveillance**,** Cyber security detection

**Healthcare/Life Sciences:** Medical record text analytics, Genomic analytics

**Telecommunications:** Call detail record processing, Customer profile monetization

However, in order to gain revenue, following big data functions are of highest value to the organization:



Transformative Big Data initiatives begin with “magic moments”: by choosing a domain in which the organization wishes to excel, while taking into account the risks and side effects. Performance Big Data initiatives are directed to existing projects with the aim of improving the performance.

## Barrier

One of the most pressing barriers of adoption for Big Data in the enterprise is the lack of skills around Hadoop administration and Big Data Analytics skills, or Data Science. In order for Big Data to truly gain mainstream adoption and achieve its full potential, it is critical that this skills gap be overcome.

## Future Trends for 2015:

In few years big data has become a major disruptor in Analytics and it is going to push companies further in 2015.Data agility will become major focus for the organizations. In 2015, IT will embrace self-service big data to allow business users self-service to big data.Self-service empowers developers, data scientists and data analysts to conduct data exploration directly. In early 2013, Intel made a splash with the introduction of its own Hadoop distribution, saying that it would differentiate itself by taking a ground-up approach in which Hadoop was baked directly into its silicon. But just a year later, Intel ditched its distribution and threw its weight behind Hadoop distribution vendor Cloudera instead. At that time, Intel noted that customers were sitting on the sidelines to see how the Hadoop market would shake out. The number of Hadoop options was muddying the waters. So, it is believed Hadoop vendor consolidation will continue in 2015 as the also-rans discontinue their distributions and focus elsewhere in the stack.

 2015 will see the evolution of a new, more nuanced model of Open source software that combines deep innovation with community development.2015 will also see enterprise architects take center-stage as their improving understanding of the Hadoop technology stack leads to a better defined and more sophisticated statement of requirements for big data applications, including elements like high availability and business continuity.

2015 could be the year for Apache Spark:

Spark is an engine for analyzing data stored across a cluster of computers. Like Hadoop, Spark can be used to examine data sets that are too large to fit into a traditional data warehouse or a relational database. Also like Hadoop, Spark can work on unstructured data, such as event logs, that hasn't been formatted into database tables. Spark, however, goes beyond what [Hadoop can easily do](http://www.infoworld.com/article/2608723/hadoop/when-to-use-hadoop--and-when-not-to-.html), in that it can analyze streaming data as it is coming off the wire. In contrast to [Hadoop](http://en.wikipedia.org/wiki/Hadoop)'s two-stage disk-based [MapReduce](http://en.wikipedia.org/wiki/MapReduce) paradigm, Spark's in-memory primitives provide performance up to 100 times faster for certain applications. Spark has over 465 contributors in 2014, making it the most active project in the Apache Software Foundation and among [Big Data](http://en.wikipedia.org/wiki/Big_Data) open source projects.

### Bigdata SV#

In recent conference organized by silicon angle, IDC made the following predictions for the year 2015:

* Visual data discovery tools will grow 2.5x faster than rest of the business intelligence market, and by 2018 investing in this enabler of end user self-service will become a requirement for all enterprises.
* Over the next 5 years, spending on cloud-based Big Data analytics solutions will grow three times faster than spending for on-premise solutions, and hybrid on/off premise deployments will become a requirement.
* Shortage of skilled staff will persist, with almost 200,000 deep analytics jobs expected in the US alone by 2018.
* By 2017, unified data platform architecture will become the foundation of Big Data analytics strategy.
* Apps that incorporate advanced and predictive analytics, including machine learning, will accelerate in 2015 and will grow 65 per cent faster than apps without this functionality.
* 70 per cent of large organizations already purchase external data and 100 per cent will do so by 2019.
* More organisations will begin to monetize their data by selling them or providing value added content.
* Adoption of technology to continuously analyse streams of events will accelerate in 2015 as it is applied to IoT analytics, which is expected to grow at a 5-year CAGR of 30 per cent.
* Decision management platforms will expand at a CAGR of 60 per cent through 2019.
* Rich media analytics will at least triple in 2015 and emerge as the key driver for Big Data anlytics technology investment.
* By 2018, half of all consumers will interact with services based on cognitive computing on a regular basis.

 Kelly also predicted the next phase of Big Data to be moving towards real-time, near real-time, and the Internet of Things.

## Sources :

<http://www.computerworld.com/article/2507728/enterprise-applications-22-free-tools-for-data-visualization-and-analysis.html>

<http://vint.sogeti.com/wp-content/uploads/2013/11/Sogeti_NoMoreSecrets.pdf>

<http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf>

<http://www.tcs.com/SiteCollectionDocuments/Trends_Study/TCS-Big-Data-Global-Trend-Study-2013.pdf>

<http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017>

<http://www.csc.com/insights/flxwd/78931-big_data_growth_just_beginning_to_explode>

<http://www.atkearney.com/analytics/featured-article/-/asset_publisher/FNSUwH9BGQyt/content/beyond-big-the-analytically-powered-organization/10192>

<http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Big-Data-POV.PDF>

<http://venturebeat.com/2014/05/11/the-state-of-big-data-in-2014-chart/>

<http://www.ijarcsse.com/docs/papers/Volume_4/1_January2014/V4I1-0320.pdf>