

Emergence and Taxonomy of Big Data as a Service

Benoy Bhagattjee

Working Paper CISL# 2014-06

May 2014

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E62-422
Massachusetts Institute of Technology
Cambridge, MA 02142

Emergence and Taxonomy of Big Data as a Service

By

Benoy Bhagattjee

MS, Texas A & M University, 2008

MCA, University of Mumbai, 2005

BSc, University of Mumbai, 2002

SUBMITTED TO THE SYSTEM DESIGN AND MANAGEMENT PROGRAM IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN ENGINEERING AND MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
FEBRUARY 2014

© 2014 Benoy Bhagattjee. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part.

Signature of the Author:

Benoy Bhagattjee
System Design and Management Program
February 2014

Certified and Accepted by:

Stuart Madnick
John Norris Maguire Professor of Information Technology
Thesis Supervisor
MIT Sloan School of Management
MIT School of Engineering

Certified and Accepted by:

Patrick Hale
Director
System Design and Management Program

This page left intentionally blank

Emergence and Taxonomy of Big Data as a Service

By

Benoy Bhagattjee

Submitted to the System Design and Management Program on Feb 30, 2014 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Engineering and Management

Abstract

The amount of data that we produce and consume is growing exponentially in the modern world. Increasing use of social media and new innovations such as smartphones generate large amounts of data that can yield invaluable information if properly managed. These large datasets, popularly known as Big Data, are difficult to manage using traditional computing technologies. New technologies are emerging in the market to address the problem of managing and analyzing Big Data to produce invaluable insights from it. Organizations are finding it difficult to implement these Big Data technologies effectively due to problems such as lack of available expertise.

Some of the latest innovations in the industry are related to cloud computing and Big Data. There is significant interest in academia and industry in combining Big Data and cloud computing to create new technologies that can solve the Big Data problem. Big Data based on cloud computing is an upcoming area in computer science and many vendors are providing their ideas on this topic. The combination of Big Data technologies and cloud computing platforms has led to the emergence of a new category of technology called Big Data as a Service or BDaaS.

This thesis aims to define the BDaaS service stack and to evaluate a few technologies in the cloud computing ecosystem using the BDaaS service stack. The BDaaS service stack provides an effective way to classify the Big Data technologies that enable technology users to evaluate and choose the technology that meets their requirements effectively. Technology vendors can use the same BDaaS stack to communicate the product offerings better to the consumer.

Thesis Advisor: Stuart Madnick

Title: John Norris Maguire Professor of Information Technologies, MIT Sloan School of Management & Professor of Engineering Systems, MIT School of Engineering

This page left intentionally blank

Acknowledgments

I would like to express my sincere gratitude to my thesis advisor Professor Stuart A. Madnick for his guidance and help. Professor Madnick showed immense understanding and patience for my thesis during my difficult times, especially when I was pivoting around topics.

Thanks to my colleagues and management at Talend, Inc. who gave me the time and the flexibility to complete the SDM program. Thanks to Gowri for her support, encouragement and patience.

Thanks to my family especially Jonna, Natu, Oamu, and Anu for your support and keeping my spirits up

Deepest gratitude and thanks to my mother Anjali, Your love, prayers, support and guidance have made me who I am today.

This page left intentionally blank

Table of Contents

Chapter 1. Introduction	9
1.1 Data Analytics	9
1.2 Introduction to Cloud Computing	10
1.3 Introduction to Big Data	11
1.4 Research Objectives and Thesis Statement	13
1.5 Approach and Thesis Structure	15
Chapter 2. Big Data	17
2.1 Scalability in Big Data	17
2.2 MapReduce	18
2.3 Real-time and Batch Processing	22
2.4 Properties of Big Data	24
2.5 Big Data in a Software Ecosystem	28
2.6 Big Data in Industry	30
Chapter 3. Overview of Business Intelligence	32
3.1 Collecting Data	33
3.2 Storing Data	34
3.3 Analyzing Data	36
3.4 Retrieving and Presenting Information	37
Chapter 4. Big Data as a Service	39
4.1 Cloud and Big Data	39
4.2 BDaaS Framework	42
4.3 Essential Tenets	45
4.4 Big Data Concerns	48
4.5 Advantages and Disadvantages of BDaaS	52
4.6 Industry Perspectives of BDaaS	53
Chapter 5. Case Studies of BDaaS	57
5.1 Politics	58
5.2 Media and Entertainment	59
5.3 Retail	60
Chapter 6. Use of the BDaaS Framework	62

6.1	Amazon EC2	62
6.2	Amazon Dynamo	65
6.3	Amazon Elastic MapReduce	67
6.4	Google BigQuery	69
6.5	Splunk Storm	71
6.6	Microsoft Azure-HDInsight	73
6.7	Tibco Silver Spotfire	75
6.8	QuBole	76
Chapter 7. Conclusion		79
7.1	Future Outlook	79
7.2	Role of BDaaS Framework	80
Bibliography		82

Chapter 1. Introduction

Data is produced and consumed as an intrinsic part of our everyday activities. Information is the key factor that drives the modern world, enabling activities from checking the weather to making complex business decisions based on stock prices. As the world grows more interconnected, the number of sources producing and consuming information grows as well. The increasing use of mobile devices has contributed significantly to the massive increase in the amount of data produced. Unstructured communication such as instant messages, twitter feeds, GPS data, and user videos generate massive amounts of data every day. For example, the Sloan Digital Sky Survey project dedicated to studying outer space produced more data in a few weeks than had been produced in the prior history of astronomy. Wal-Mart generates over a billion data entries in its systems to store transactions, manage inventory, and prepare financial reports. This data helps Wal-Mart identify consumer behavior and purchase patterns for decision making. AT&T has project Daytona to address data management problems within the organization's multi-terabyte database. Daytona manages the massive multi-terabyte data repository of call records, which is used for criminal investigations and identifying user call patterns.

According to technology consulting company CSC, data production will be 44 times greater in 2020 than it was in 2009. According to a report generated by the University of Southern California,¹ the world has stored over 295 exabytes of information till now. In 1986, the storage of data on paper-based systems was 33% of the total in that year. It dropped to less than 0.007% of the total in 2007. The price of data storage has dropped from an average of \$1,120 per GB in 1995 to less than \$0.05 in 2013,² which has reduced the costs of storing large amounts of data. Currently more than 95% of the total persistent data in the world is in digital form. Conventional data analysis tools, due to their architectural limitations, fall short in storing and analyzing these growing datasets. A large set of data that is difficult to manage using conventional tools but can yield valuable information if effectively analyzed is commonly known as Big Data.

1.1 Data Analytics

Businesses leverage data assets to derive information, measure performance, and strategize business decisions effectively. Organizations can use the information derived from analysis of data to identify the trends and measure the performance of the business. The information derived from data can also be used to predict future trends and to enable risk analysis and fraud detection. For example, a retail store can quickly identify items commonly purchased together and place them in a nearby location in the store to maximize sales. An online retailer can make suggestions

¹ <http://news.usc.edu/#!/article/29360/How-Much-Information-Is-There-in-the-World>.

² <http://www.statisticbrain.com/average-cost-of-hard-drive-storage/>.

depending on which item the customer is considering purchasing. Credit card companies can analyze transactions in real time to find suspicious transactions and automatically take measures to protect customers. Consumer location can be tracked by analyzing trends and any aberration in purchasing behavior can be highlighted as a possible fraud and appropriate action taken. For example, if a consumer uses a credit card in New York and if the same credit card is used to make a purchase in San Francisco at that moment, the real-time fraud detection system can analyze the geographical distance and the time duration and send an alert. Banks and financial organizations also use stock price data and perform statistical calculations such as Monte Carlo simulations to predict risk and expected stock prices for investment.

Data analytics is a very powerful tool that has enabled organizations to gather information regarding consumer preferences as well as their demographics. For example, most grocery stores have discount cards, which enable customers to receive discounts on some commonly used items. In return for the shopping discounts offered to the customer, the retailer gathers valuable data on shopping behavior, highlighting shopping trends and buying preferences, which ultimately results in better marketing for the retailer. By leveraging data and the subsequent business analytics, Target was able to increase its sales from \$44 billion in 2002 to \$67 billion in 2010. The total cumulative inflation for the same time period was around 22.3%, but sales increased by over 65%.

Harrah's entertainment, one of the world largest casino entertainment companies with over 40 million customers, implemented a comprehensive program to leverage its data assets to increase its profit margins. Harrah's launched a "Total Rewards" program that enabled it to collect data on customer spending habits, preferences, and gaming patterns. The information extracted from the data allowed Harrah's to design custom-made vacation packages for individual customers based on their preferences and spending patterns. This enabled Harrah's to identify the most frequently used gaming services, reduce operational expenses, and increase customer base. Since this program was implemented, customers' discretionary spending at the casino increased from 36% to 46% of their gambling budget. This new system also added some benefits to the customer experience, such as access to tax forms for game winnings through the player account.

1.2 Introduction to Cloud Computing

Cloud computing is the use of external virtualized resources that are dynamically scalable and that are used to provide services over the Internet. Cloud computing has its origins in the 1960s, when J. C. R. Licklider, who played a significant role in the development of ARPANET, envisioned a global computing framework. He envisioned a network in which computing services would be delivered as a public utility. Professor Ramnath Chellappa introduced the

word “Cloud Computing” at the 1997 INFORMS meeting.³ Telecom providers began offering VPN (Virtual Private Network) Services, helping to grow the infrastructure needed for cloud computing. This allowed effective load balancing and lowered the cost of services.

Amazon has been in the forefront of developing and maturing the cloud computing paradigm. The average utilization of Amazon data centers in the 1990s was around 10%, which was improved significantly by the introduction of the new cloud architecture. Amazon began selling unused computing capacity at its data centers as a part of the Amazon Web Services (AWS) project. The AWS architecture allowed people quickly to instantiate virtualized resources to add and develop applications. In 2003, Pinkham and Black proposed selling the cloud-based virtualized servers as services to external customers.⁴ In 2006, Amazon launched the elastic compute cloud (EC2) platform to provide virtualized services over the cloud platform. Organizations are increasingly looking to implement cloud computing architecture to supplant the existing infrastructure due to its flexibility and ease of use.

1.3 Introduction to Big Data

Fremont Rider, in the paper “The Scholar and Future of Research Library,” stated that the size of American university libraries was doubling every year to illustrate the growing volume of data (Rider 1944). For example, he predicted that Yale library in 2040 would comprise of 200 million volumes, requiring over 6000 miles of shelves. David Ellsworth and Michael Cox published a paper, “Application Controlled Demand Paging for Out-of-core Visualization,” which first mentioned the Big Data problem (Ellsworth, Cox 1997). They stated that the requirements for data analysis were quite large even for memory or disk space to accommodate. Francis Diebold presented a paper “On the Origin(s) and Development of the Term Big Data” that states that science has much to benefit from Big Data (Diebold, 2012). The three dimensions of Big Data were first published by Doug Laney in his article “3D Data Management: Controlling Data Volume, Velocity and Variety” (2012).

Big Data is used to describe data which is too large in size, which makes it difficult to analyze in traditional ways (Davenport 2012). Gartner refers to data sets that have “high variety, volume and velocity as Big Data” (Heudecker 2013). The term “Big Data” is surrounded by a lot of hype, where many software vendors claim to have the ability to handle Big Data with their products. Organizations should invest in Big Data technologies after they have identified the business need, not the other way around. Organizations need to ensure that new Big Data

³ <http://blog.b3k.us/2009/01/25/ec2-origins.html>.

⁴ <http://www.meronymy.com/SPARQL-Database-Server/Faq/advantages-of-a-dbms.aspx>.

technologies address their current business requirements to see their effectiveness (Heudecker 2013).

Innovations in hardware technology such as those in network bandwidth, memory, and storage technology have helped the stimulation of Big Data technology. The new innovations coupled with the latent need to analyze the massive unstructured data that stimulated their development (Beyer et al. 2012). As Big Data is an emerging technology, it is still at the “peak of inflated expectations” of the Gartner hype cycle. Hadoop⁵ is an open-source framework for distributed computing that enables processing of large datasets through horizontal scalability. Organizations such as Yahoo have invested in and nurtured the Apache Hadoop project to address their Big Data needs. Yahoo launched Hadoop as a science project to enable processing of large datasets in a distributed environment.⁶ Yahoo uses Hadoop to screen out the spam messages that can potentially clutter its servers. Increasingly software organizations are providing these technologies over a cloud platform, leading to a new paradigm, Big Data as a Service (BDaaS), which is delivering Big Data technologies over a cloud-based platform.

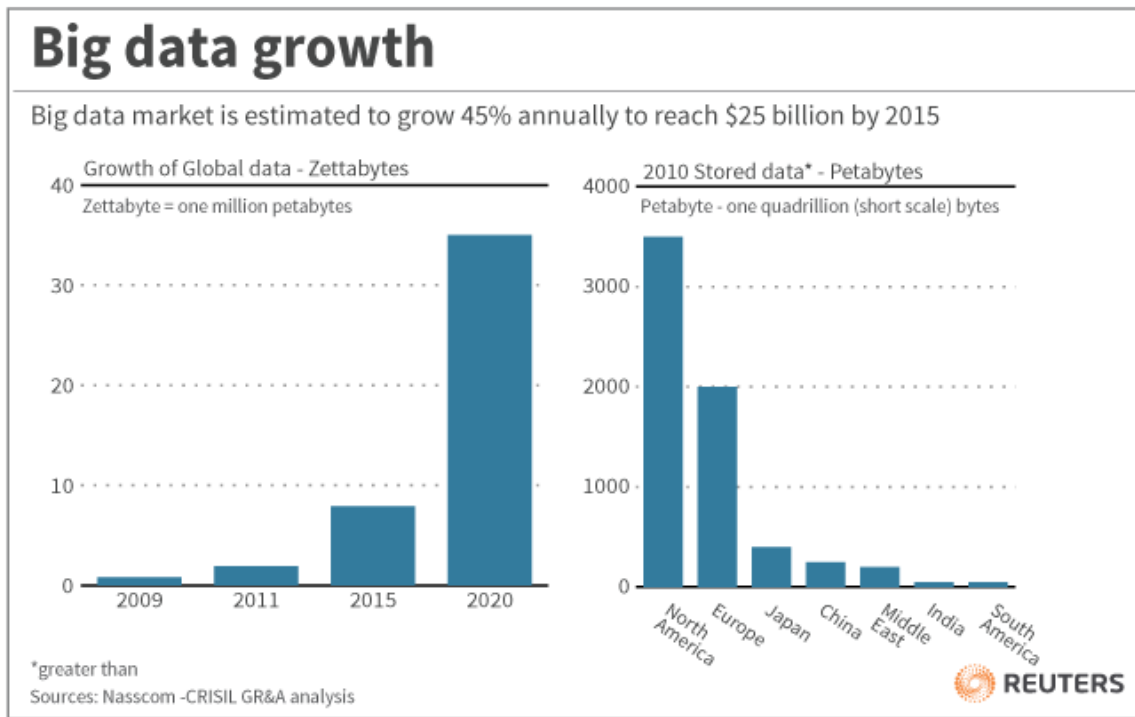


Figure 1.1: Projected growth of data⁷

⁵ <http://hadoop.apache.org/>.

⁶ <http://readwrite.com/2010/05/05/yahoo-rebuilding-its-future-on#awesm=~oj73tfBjDWp5pq>.

⁷ <http://blog.thomsonreuters.com/index.php/Big-Data-graphic-of-the-day/>.

According to Gartner reports (see Figure 1.1) the market for Big Data technologies is expected to grow by almost 45% annually till 2015. The report also expects that the total data in the world to grow almost 400% to approximately 35 zettabytes in 2020. The growth in global data represents the total amount of data generated, which may or may not be stored as voice calls or live video. The stored data represents the amount of data in persistent storage in the region specified.

1.4 Research Objectives and Thesis Statement

Organizations implementing Big Data systems face significant costs in terms of setting up infrastructure and obtaining specialized manpower. Currently many organizations are spending substantial resources to evaluate Big Data technologies to see if they meet their business requirements. An industry survey sponsored by Gartner (Kart, Heudecker, Buytendijk 2013; see Figure 1.2) indicates that a major challenge in Big Data implementations is determining their business value. According to the survey, over 56% of the technology leaders were trying to determine how to derive value for Big Data. In the same survey, 29% of the respondents mentioned that setting up and managing the infrastructure to manage Big Data was a major challenge.

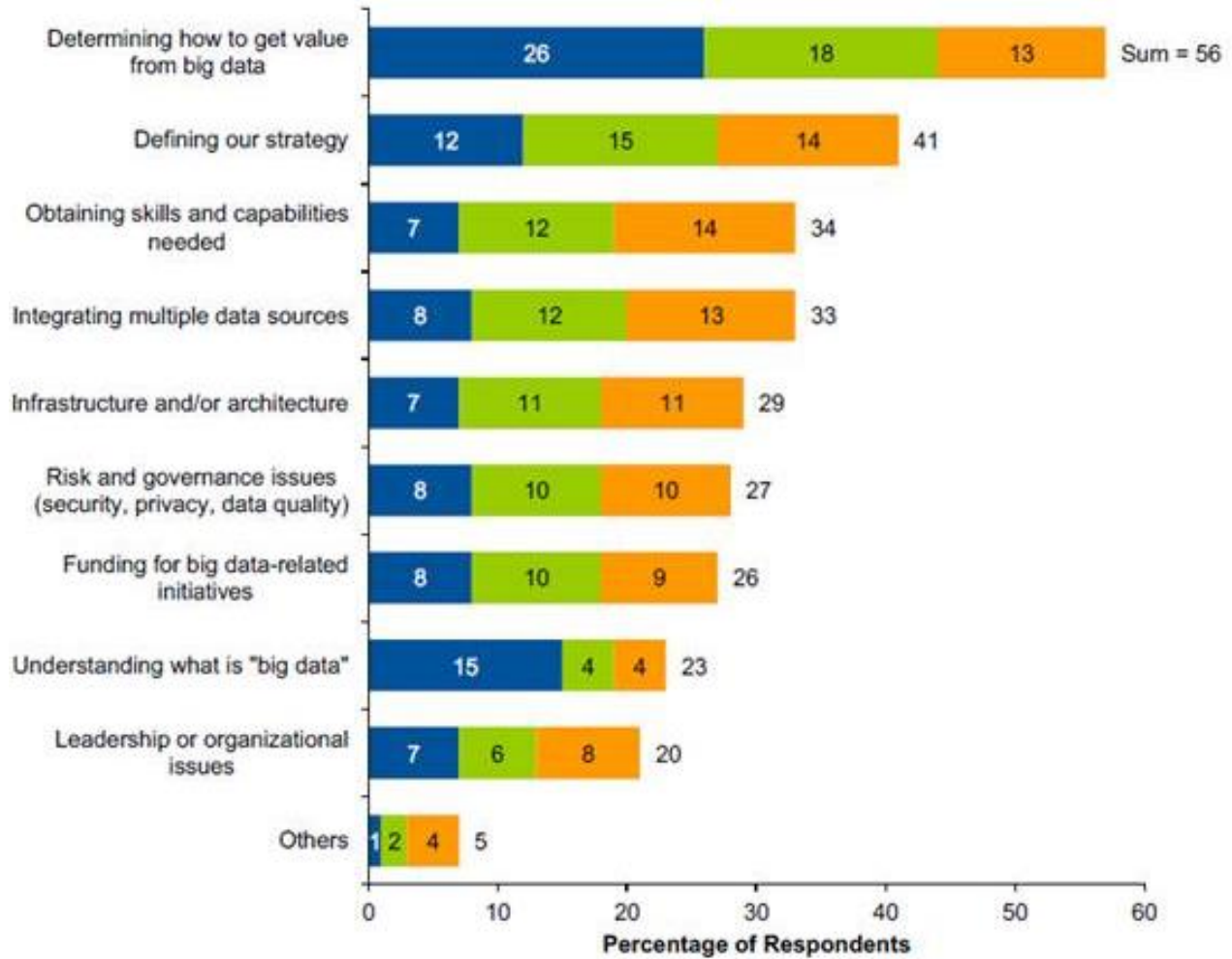


Figure 1.2: Top Big Data Challenges (Kart, Heudecker and Buytendijk, 2013).

There is also an ‘analytical capability gap’ between smaller and larger organizations where the latter have resources to analyze their data leading to better decision making. The smaller organizations lack that capability and hence they are at a severe disadvantage, which can be detrimental to their growth. The use of technologies in the BDaaS framework would help smaller organizations to close this capability gap. These technologies would ultimately enable smaller organizations to achieve Big Data capability similar to that of the larger organizations. Technologies in the BDaaS framework enable organizations to have scalable, on-demand computing resources that would reduce overall cost for the company and provide flexibility. Scalability of BDaaS enables organizations to request resources on demand and frees up organizations from infrastructure constraints. For example, a small business may not require frequent analysis of its large datasets to warrant an investment in infrastructure. For occasional analysis of its datasets it may want to have the capability “on-demand” and hence it might host its Big Data framework using a third party cloud system. This service-based framework would

enable the smaller organization to save on the infrastructure costs and maintenance when it does not require it.

BDaaS is defined as a framework including cloud-based, distributed computing technologies that are horizontally scalable and are designed to handle large datasets or Big Data. The multilayered framework serves as an effective classification scheme for both BDaaS service consumers and providers. Big Data integrated with a cloud computing platform offers both the scalability and the service model of cloud-based software. It also enables organizations to implement infrastructure quickly without requiring specialized skills. Technologies in a BDaaS framework are easy to implement and test as compared to non-cloud-based systems. As the with the cloud paradigm, these systems can scale up easily and users pay only for the services consumed.

With the current hype of Big Data technologies, both technology providers and consumers are facing a problem of identifying the right technology for their needs. There are many technology vendors launching new Big-Data-enabled products into the market without appropriately defining them, causing confusion and unmet expectations for the end consumer. Technology providers need to communicate their product benefits effectively to end users. For consumers it is important to get the right technology that meets their needs. The intent of this thesis is to formulate a BDaaS framework that will help technology producers and consumers to identify and classify Big Data technologies based on a cloud computing platform. This BDaaS framework includes a classification scheme using service stack layers which will help consumers evaluate different technologies in the Big Data ecosystem. The BDaaS framework can help technology consumers understand available products that meet their requirements as well as illustrating basic characteristic of each product in its layer.

1.5 Approach and Thesis Structure

The main contribution of this thesis is the development of the BDaaS framework and the classification of technologies under it. This framework gives a better picture of cloud-based Big Data technologies. As of 2014, Big Data technologies are still evolving and it is the latest industry buzzword. As this topic is in an evolving field, this thesis evaluates reports from leading industry research firms such as Gartner and Forrester.

Chapter 2 presents the scalability issue in Big Data along with an introduction to the MapReduce algorithm. It also illustrates the main attributes of Big Data, namely velocity, volume, and variety according to the popular definition. As newer technologies to address Big Data are being introduced in the market, it is important to identify their place in the software

ecosystem. There is a good market demand for Big Data technologies and this section also highlights some of the current market trends.

Chapter 3 presents the end-to-end flow of data within the organization. It is important to understand the different stages of the data during this flow. This section illustrates how data is generated and consumed at each stage and the different technologies involved.

Chapter 4 presents the main topic of the BDaaS framework along with its essential tenets. It also illustrates the industry perspective and apprehensions of Big Data.

Chapter 5 presents some case studies in politics, media, and retail where the BDaaS framework could be used. The Obama campaigns used some Big Data technologies, which helped increasing the effectiveness of the presidential campaigns. The New York Times has effectively used Big Data in converting its historical documents.

Chapter 6 illustrates how the BDaaS stack could be used to classify and evaluate new technologies in the Big Data software ecosystem.

Chapter 7 concludes the thesis with insights gained from this research. This section also highlights the future growth of the industry.

Chapter 2. Big Data

Big Data is not just about handling large datasets and distributed computing; it also represents a paradigm shift in the way data is managed and analyzed (see Figure 2.1). In the case of traditional business intelligence (BI), analytics operations are performed on transformed data from raw sources. Source data is transformed in various ways to be structured before it can be used by analytics tools. Transformation operations can potentially result in loss of valuable information from the data. For example, transaction data from retail stores is collected and extracted from operational databases using Extract Transform Load (ETL)/integration systems. The integration and transformation operations do not handle unstructured data well. With the highly scalable Big Data technologies and the resultant increase in computing power, users have the increased capability to work with raw data. The analytics on raw data may yield valuable information that might not have been discovered before. Using Big Data technologies, it has been possible to apply machine learning and natural language processing tools to large datasets. This also enables effective analysis of unstructured data, which represents 80% of total data stored in the world. With Big Data, new innovations such as MapReduce have been invented to aggregate and analyze large data sets to derive information that was previously unavailable.

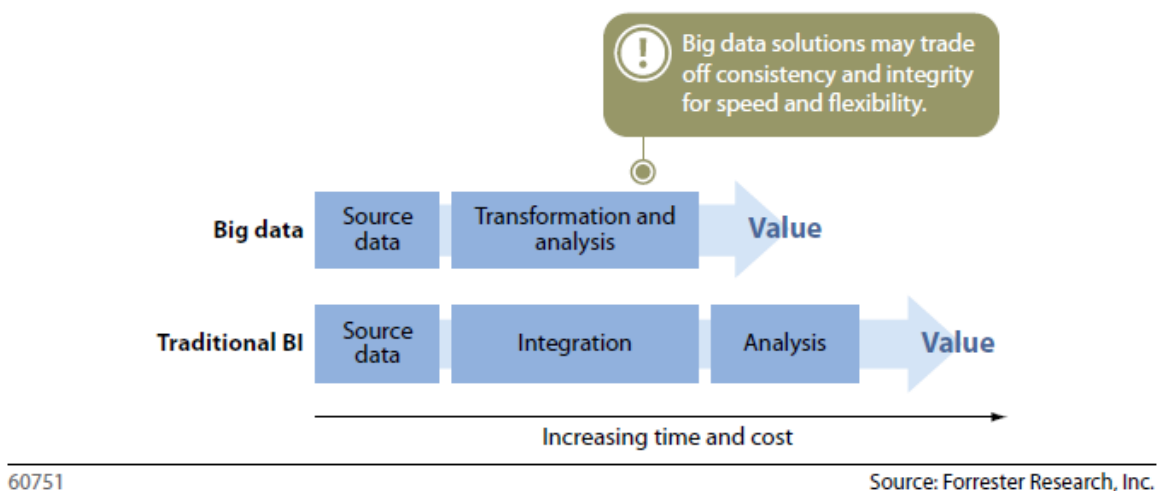


Figure 2.1: Comparing Traditional BI with Big Data (Hopkins and Evelson 2012).

2.1 Scalability in Big Data

Most Big Data technologies are based on distributed computing architecture, which is horizontally scalable to handle large datasets. Horizontal scaling means adding multiple autonomous computers that work together to increase the overall capacity of the system (Figure 2.2). Vertical scaling means adding more processors or memory, thereby increasing capability of

the system. Most of the traditional systems have to scale up vertically so that they can process and analyze larger data sets. Vertical scaling relies mainly on technological innovations to achieve higher performance. For example, if the computing power of a single processor machine is doubled, then the single processor has to be replaced with another processor twice the original capability, which may not be technically feasible. In the case of horizontal scaling, two processors of the same capacity can be added to increase the capacity. This is a major limitation, as most existing systems may not scale up to process and analyze large datasets. Big Data analytics generally refers to distributed data processing across multiple nodes in a horizontally scaled architecture (EMC Solutions Group 2012). For example, using Hadoop, users can add more ‘nodes’ or computer systems to the existing architecture to scale up. The capacity of individual machines need not be increased to for scaling up in this case. With horizontal scaling the ability to scale up is much more than vertical scaling, as the later depends much more on technical innovations to push the boundaries forward.

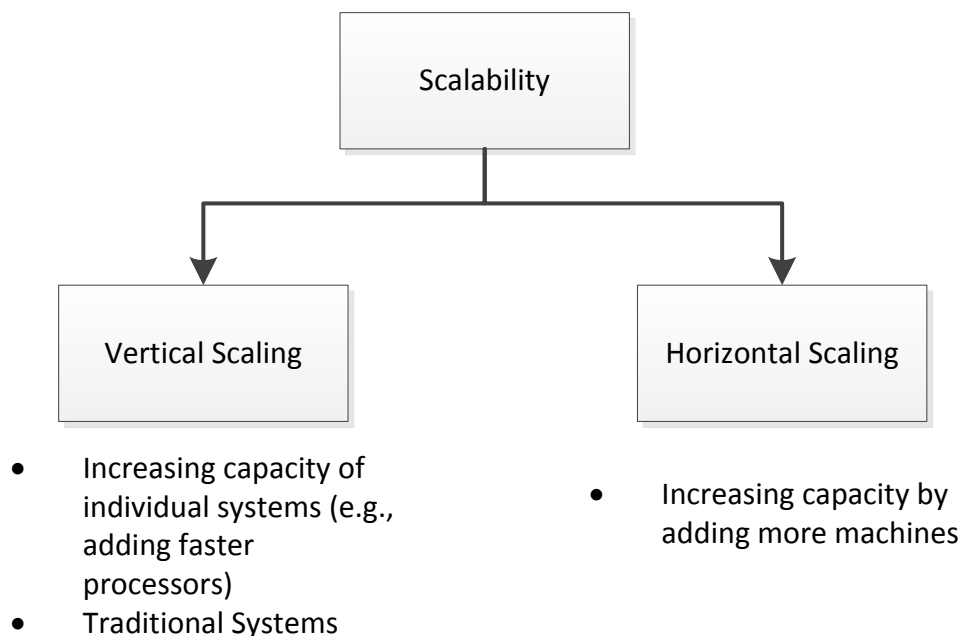


Figure 2.2: Vertical and Horizontal Scaling

2.2 MapReduce

MapReduce is an architectural model for parallel processing of tasks on a distributed computing system. This algorithm was first described in a paper “MapReduce Simplified Data Processing on Large Clusters,” by Jeffery Dean and Sanjay Ghemwat from Google. This algorithm allows splitting of a single computation task to multiple nodes or computers for distributed processing.

As a single task can be broken down into multiple subparts, each handled by a separate node, the number of nodes determines the processing power of the system. There are various commercial and open-source technologies that implement the MapReduce algorithm as a part of their internal architecture. A popular implementation of MapReduce is the Apache Hadoop, which is used for data processing in a distributed computing environment. As MapReduce is an algorithm, it can be written in any programming language.

Function of the MapReduce Algorithm

The initial part of the algorithm is used to split and ‘map’ the sub tasks to computing nodes. The ‘reduce’ part takes the results of individual computations and combines them to get the final result. In the MapReduce algorithm, the mapping function reads the input data and generates a set of intermediate records for the computation. These intermediate records generated by the map function take the form of a (key, data) pair. As a part of mapping function, these records are distributed to different computing nodes using a hashing function. Individual nodes then perform the computing operation and return the results to the reduce function. The reduce function collects the individual results of the computation to generate a final output.

As shown in Figure 2.3, we are calculating the number of occurrences of each word in an input text file. The mapping function takes the input file, separates the records, and sends them to different nodes or mapping instances for processing. The mapping function then splits the document into words and assigns a digit “1” to them to form a key-value pair for further computation. The intermediate output is in the form of (word, 1) and is sorted and grouped into individual nodes to calculate the frequency. The resultant output from the sort operation is then fed to the reduce function, which sums up the outputs from different nodes and generates the final output containing the frequency.

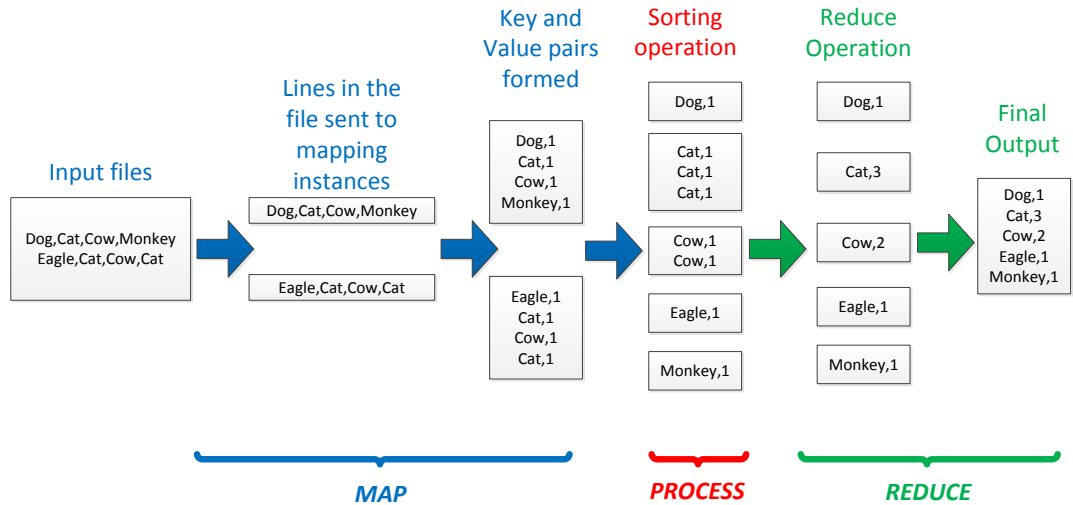


Figure 2.3: Illustration of MapReduce Algorithm⁸

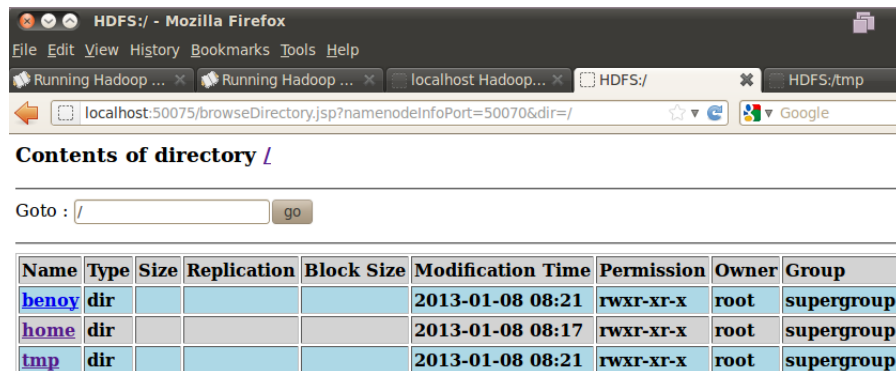
Example of MapReduce with Hadoop

The Hadoop architecture consists of a master server which directs tasks to underlying worker machines. It consists of a bundle of components commonly called the “Hadoop Common Package.” This package is made up of components such as the Hadoop Distributed File System (HDFS), MapReduce engine, programming libraries, and scripts to manage the Hadoop installation. Data in the Hadoop framework is stored in the HDFS, which may be stored across multiple nodes. HDFS replicates data three times on separate nodes by default for protection against failure. A checksum for the data blocks is periodically calculated to ensure their integrity. Programs to perform the distributed computing tasks utilize the programming libraries, which implement the MapReduce algorithm. These components work together to process a large computing task in a batch processing mechanism.

HDFS enables data to be stored on a distributed system which is interfaced externally as a unified file namespace. For example, data files might be physically stored on different machines, but appear as though they are stored on a single system. Internally a file is split into multiple blocks and stored on different machines using data nodes. HDFS is composed of name nodes and data nodes, which are installed on the master server and the worker servers respectively. The name node on the master server manages data using data nodes on worker machines. The name node on the master server performs file management tasks such as opening and closing files as well as mapping those files to data nodes. The data nodes on the worker machines perform read and write requests as required. Programs can utilize this file system and code libraries to create

⁸ <http://blog.trifork.com/2009/08/04/introduction-to-hadoop/>.

distributed programming tasks. Name node can be a single point of failure in an HDFS system, as it manages the data nodes.



Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
benoy	dir				2013-01-08 08:21	rwxr-xr-x	root	supergroup
home	dir				2013-01-08 08:17	rwxr-xr-x	root	supergroup
tmp	dir				2013-01-08 08:21	rwxr-xr-x	root	supergroup

Figure 2.4 HDFS File System

In Figure 2.4, we can see the HDFS file system in the Hadoop administration console and the directory structure of the input files. In Figure 2.5, a program using Hadoop is executed to run the word count program on the input files stored in the HDFS system. Pointers 1 and 2 indicate input and output directories on the HDFS file system. Pointer 3 shows the execution of the MapReduce job, which takes the input files from the HDFS system, splits and processes them, and then stores the result back in the directory, as indicated pointer 2. Pointer 4 shows the execution statistics of the MapReduce job.

```

root@ubuntu: /home/hadoop/hadoop-1.1.1
File Edit View Terminal Help

root@ubuntu: /home/hadoop/hadoop-1.1.1# bin/hadoop jar hadoop*examples*.jar wordcount /user/root/benoy/sourcefiles /user/root/benoy/sourcefiles-output
Warning: SHADOOP_HOME is deprecated.

13/01/09 09:39:02 INFO input.FileInputFormat: Total input paths to process : 3
13/01/09 09:39:02 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/01/09 09:39:02 WARN snappy.LoadSnappy: Snappy native library not loaded
13/01/09 09:39:02 INFO mapred.JobClient: Running job: job_201301090655_0002
13/01/09 09:39:03 INFO mapred.JobClient: map 0% reduce 0%
13/01/09 09:39:13 INFO mapred.JobClient: map 33% reduce 0%
13/01/09 09:39:14 INFO mapred.JobClient: map 66% reduce 0%
13/01/09 09:39:19 INFO mapred.JobClient: map 100% reduce 0%
13/01/09 09:39:23 INFO mapred.JobClient: map 100% reduce 33%
13/01/09 09:39:26 INFO mapred.JobClient: map 100% reduce 100%
13/01/09 09:39:28 INFO mapred.JobClient: Job complete: job_201301090655_0002
13/01/09 09:39:28 INFO mapred.JobClient: Counters: 29
13/01/09 09:39:28 INFO mapred.JobClient: Job Counters
13/01/09 09:39:28 INFO mapred.JobClient:   Launched reduce tasks=1
13/01/09 09:39:28 INFO mapred.JobClient:   SLOTS_MILLIS_MAPS=25054
13/01/09 09:39:28 INFO mapred.JobClient:   Total time spent by all reduces waiting after reserving slots (ms)=0
13/01/09 09:39:28 INFO mapred.JobClient:   Total time spent by all maps waiting after reserving slots (ms)=0
13/01/09 09:39:28 INFO mapred.JobClient:   Launched map tasks=3
13/01/09 09:39:28 INFO mapred.JobClient:   Data-local map tasks=3
13/01/09 09:39:28 INFO mapred.JobClient:   SLOTS_MILLIS_REDUCE=13033
13/01/09 09:39:28 INFO mapred.JobClient: File Output Format Counters
13/01/09 09:39:28 INFO mapred.JobClient:   Bytes Written=880838
13/01/09 09:39:28 INFO mapred.JobClient: FileSystemCounters
13/01/09 09:39:28 INFO mapred.JobClient:   FILE BYTES READ=2214849
13/01/09 09:39:28 INFO mapred.JobClient:   HDFS BYTES READ=3671896
13/01/09 09:39:28 INFO mapred.JobClient:   FILE BYTES WRITTEN=3784947
13/01/09 09:39:28 INFO mapred.JobClient:   HDFS BYTES WRITTEN=880838
13/01/09 09:39:28 INFO mapred.JobClient: File Input Format Counters
13/01/09 09:39:28 INFO mapred.JobClient:   Bytes Read=3671517
13/01/09 09:39:28 INFO mapred.JobClient: Map-Reduce Framework
13/01/09 09:39:28 INFO mapred.JobClient:   Map output materialized bytes=1474341
13/01/09 09:39:28 INFO mapred.JobClient:   Map input records=77932
13/01/09 09:39:28 INFO mapred.JobClient:   Reduce shuffle bytes=1474341
13/01/09 09:39:28 INFO mapred.JobClient:   Spilled Records=255962

```

Figure 2.5: Hadoop MapReduce Program Execution

2.3 Real-time and Batch Processing

In traditional data storage systems such as relational database management systems (RDBMS), data is calculated according to the query submitted. These systems do not scale up effectively as the database size increases. Having pre-computed results from data does help in speeding up processing, but it may not give correct results as it is essentially a batch data processing mechanism. For example, if the underlying data changes after the result is pre-computed, then users do not have the latest data (Figure 2.6). Real-time Big Data systems address this problem by processing data as it is available.

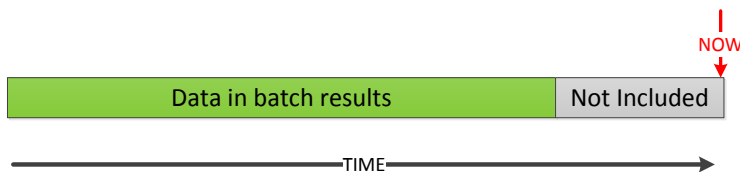


Figure 2.6: Missing Data in Batch Processing Systems

Most Big Data systems such as those using the Hadoop framework are essentially batch processing systems. In batch processing systems, the jobs are run across the entire datasets. If there is any change to the datasets, then the jobs have to be executed again to incorporate the changes in the input. Big Data systems take data in bulk and then submit it as a batch job which

splits and processes it in multiple nodes for processing and stores a computed result for consumption. It essentially speeds up the processing operation, but if the source dataset changes then the computation has to be done again. For example, if a financial organization is analyzing a large dataset related to stock prices at a point in time during the day when the market is stable, using a batch data system might not be the best solution. The batch processing Big Data system will work on the snapshot of the data when it is submitted. If there is an unexpected change in the dataset, such as a crash in the stock market, after the batch processing job has executed, then it will not incorporate the latest information in the analysis. In Figure 2.7, the MapReduce workflow computes the results when the distributed processing jobs are executed. Information consumers may be reporting or analytical systems which use the results of the computation from that batch.

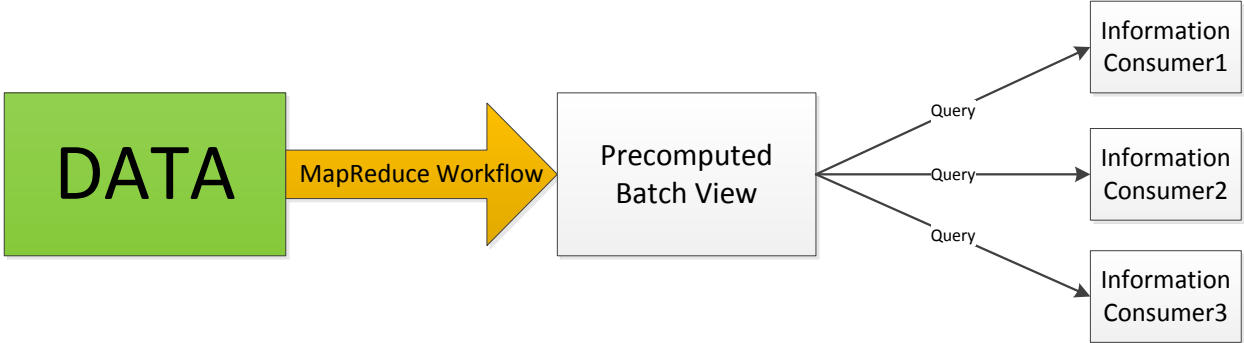


Figure 2.7: Batch Processing of Data Using Hadoop

Real-time processing of Big Data is based on data stream processing rather than the task execution deadlines. The current version implementation is the Twitter Storm framework, which provides data streaming capability. An application using the Storm framework consists of a topology of “spouts” and “bolts” which provide stream data processing (see Figure 2.8). A spout is a source of a data stream for computation which may be from a messaging queue or it might generate on its own. A bolt is a computation unit that processes any number of input streams to produce output streams that may be consumed by other bolts or direct consumers. Most of the logic in bolts is related to data operations such as filters, joins, and aggregations. So essentially, the real time Big Data is processed from the data streams as they come along.

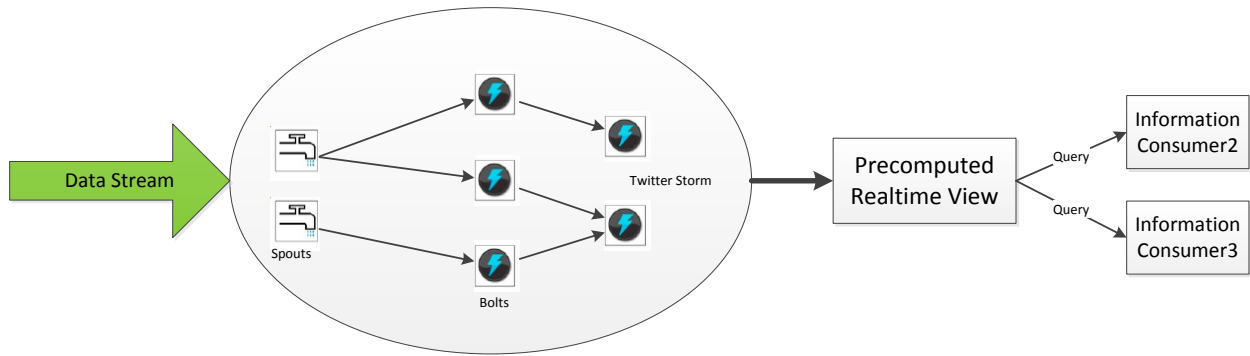


Figure 2.8 Real-time/Stream Processing of Data Through the Twitter Storm Framework

2.4 Properties of Big Data

Big Data is commonly described through three dimensions: velocity, variety, and volume, which have a different scale than traditional datasets (Figure 2.9). The data variety has proliferated with the rise in portable computing devices such as smartphones and tablets.

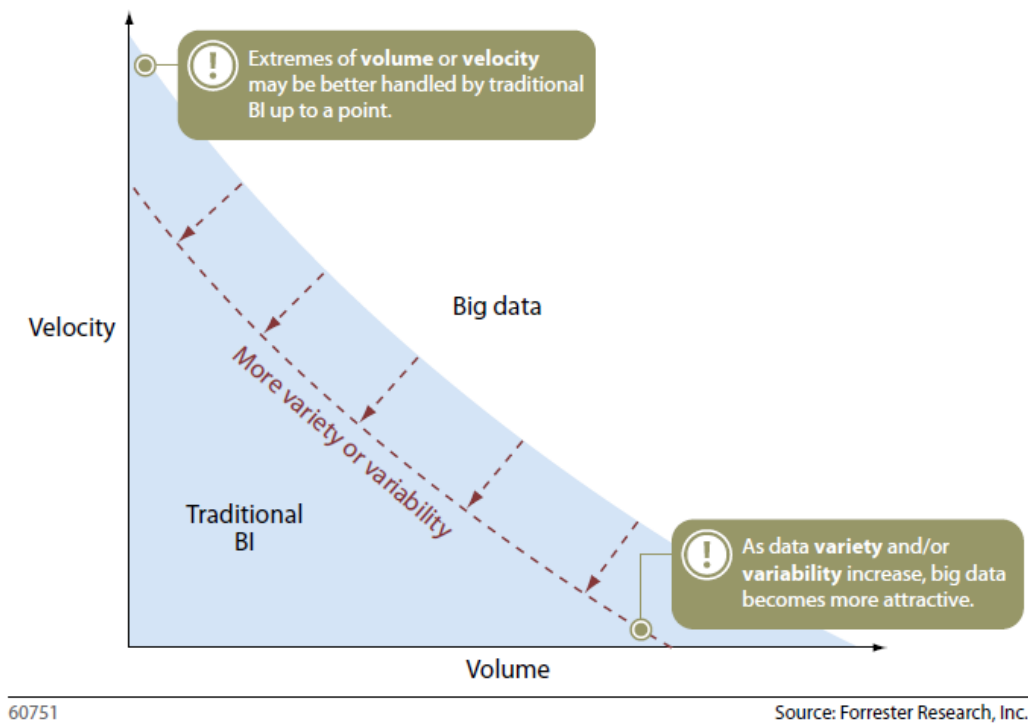


Figure 2.9: Three dimensions of Big Data (Hopkins and Evelson 2012)

- *Velocity* : This refers to the rate at which data is flowing through the system. Velocity is an important dimension, which may determine the usefulness of information. For example, industries such as investment banking develop systems such as those used in high frequency trading systems (HFTs) to leverage the fast data movement for business

advantage. One of the important considerations in velocity is complex event processing (CEP), which refers to computing incoming data to generate information on real-time events. CEP is used in applications that provide real-time, situational awareness. For applications such as just-in-time inventory and time-based computation and financial markets, CEP is vital in organizational applications. Fast-moving data, also known as “streaming data,” has two important considerations that need to be addressed for effective use: storage and response.

- *Data Storage:* If the storage system receives more data than it can write effectively to persistent storage, it can cause problems. The incoming flow of data should not overwhelm storage systems, but should be manageable so that it can be analyzed. For example, experiments at the CERN hadron collider generate over 6GB of data per second, which is challenging to store and analyze. CERN uses Hadoop to store and manage data.⁹ The HDFS enables data storage in multiple locations, minimizing the possibility of data loss. Helix is a cloud computing initiative to address the growing IT requirements of European research organizations. The Helix initiative aims to leverage partnership among research organizations to create a sustainable cloud computing infrastructure to meet the demand for research organizations. CERN has announced that it will join the Helix initiative, which will help to increase the computing capability at its disposal. Currently CERN is a part of a Worldwide Large Hadron Collider computing grid, which enables it to have the computing power of over 150K processors, but it hopes that the Helix initiative will more than double that capability.
- *Immediate response:* In some situations, an immediate response to the incoming flow of data is required. Organizations need to process the fast data flow which allows them to use real-time decision-making to gain competitive advantage. To store, analyze, and process a fast-moving dataset, new technologies such as NoSQL have been invented. Key-value store databases such as HBase and Cassandra enable faster retrieval of large sets of data to facilitate response time. For example, Facebook uses HBase, a key-value store database, to handle messaging amongst 350 million users with over 15 billion messages per month delivered in real time. Standard relational databases would not easily scale up to deliver the performance horizontally scalable key-value databases provide.

⁹ Using the Hadoop/MapReduce approach for monitoring the CERN storage system and improving the ATLAS computing model.

- *Volume:* Volume refers to the size of a typical dataset, which can be as high as multiple petabytes. The data volume is higher when unstructured information such as video and audio are stored. For example, over 100 hours of video are uploaded every minute on popular websites such as YouTube and the volume is growing at the rate of 50% per year.¹⁰ Due to the volume of Big Data, storage methods need to be focused on scalability and distributed processing. As most of the Big Data technologies are based on distributed computing principles, they process data by partitioning it into distinct sets. Large data sets are harder to query and hence are problematic to analyze, which necessitates new distributed storage architecture to address them. The more data points are available for analysis, the better the accuracy in both predictive and descriptive analytics. Data-processing options can range from large data warehouse appliances such as Netezza or Oracle Exadata to open-source HDFS based systems. The Big Data appliances such as Netezza and Exadata are based on massive parallel internal processing architecture, which is based on traditional relational structures. These propriety systems abstract the distributed processing architecture for end users, but are much more expensive than open-source systems. Systems based on technologies like the Hadoop stack are more flexible but difficult to implement.
- *Variety:* The variety represents the increasing mixture of structured and unstructured data in the real world. Data can be stored in various formats such as text, audio, video, and images. Traditional analytics can handle data that can be easily represented by a relational or hierarchical structure. Over 80% of the data generated in the world is unstructured, which makes it difficult for traditional tools to handle and analyze. Managing the variety of data vastly increases the ROI on the Big Data infrastructure. Normally data is not available readily in consumable form, but has to be transformed to be interpreted and consumed by external systems. One of the main problems of data analysis systems is that the data is diverse and that it has to be converted to a usable form and structure to yield information. For example, the data can range from messages on social networks to transaction data from retail stores, but it has to be formatted and structured before it is analyzed to yield information. Big Data technologies enable processing of unstructured data to extract order or meaning for consumption.

According to a Gartner survey (AUTHOR(S) 2013), the largest sources of data that organizations are currently processing are transaction data and log data, as they are in relatively structured form. Unstructured data such as emails, social media, and free-form data form the next largest

¹⁰ <http://www.youtube.com/yt/press/statistics.html>.

segments of data. Video and audio data is hard to organize and classify, hence it is relatively less used as compared to the other forms. Figure 2.10 represents variety of data analyzed by the respondents in the Gartner survey.

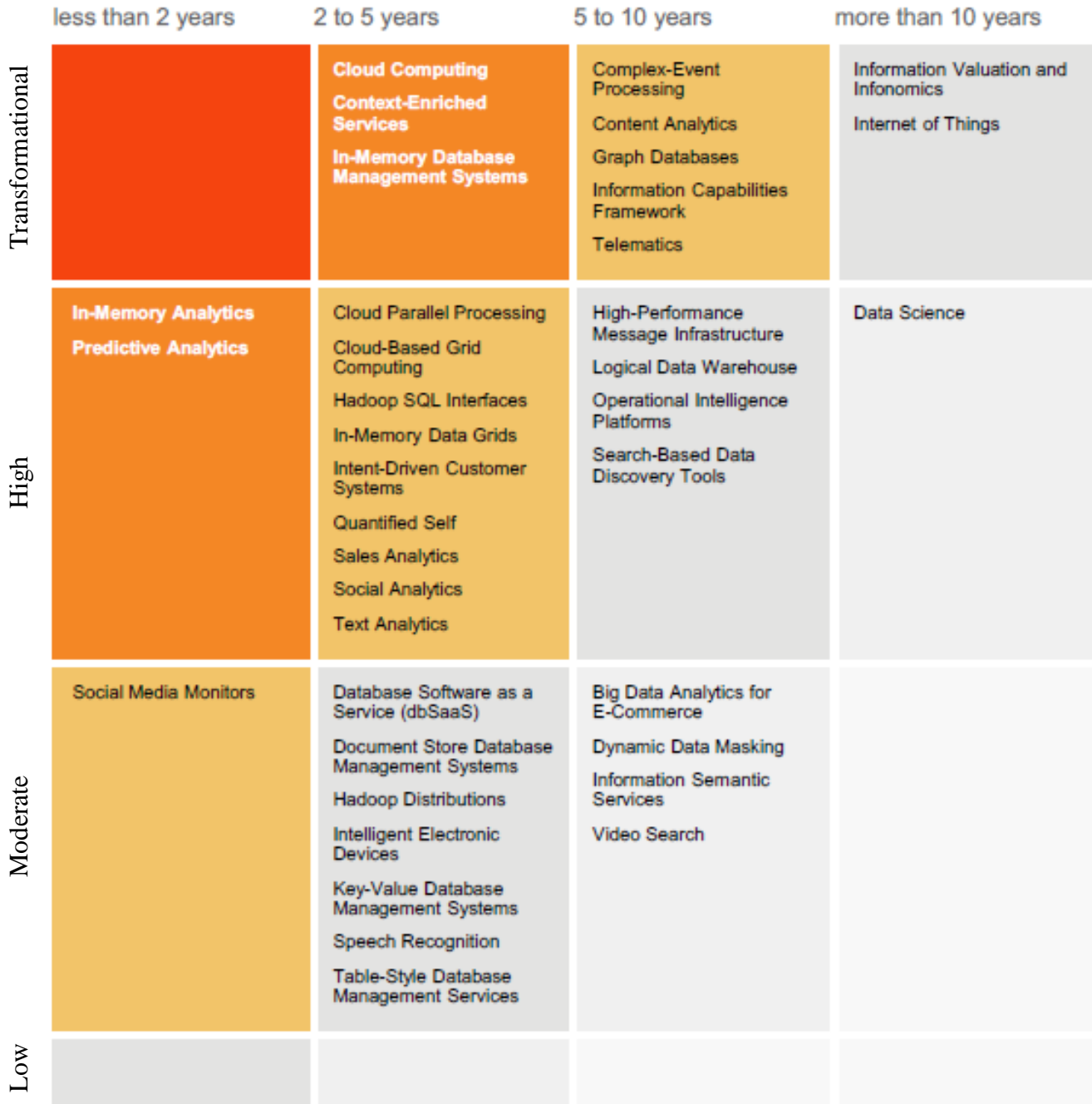
	Manufacturing and Natural Resources	Media/ Communications	Services	Government	Education	Retail	Banking	Insurance	Healthcare	Transportation	Utilities
Transactions	73%	62%	67%	67%	54%	93%	83%	81%	75%	79%	80%
Log data	44%	57%	58%	59%	54%	40%	66%	61%	33%	71%	60%
Machine or sensor data	53%	38%	35%	33%	31%	27%	27%	48%	42%	50%	40%
Emails /documents	27%	43%	43%	41%	46%	27%	34%	39%	17%	29%	20%
Social media data	32%	52%	39%	26%	54%	73%	27%	13%	-	50%	-
Free-form text	17%	24%	26%	30%	31%	20%	34%	35%	67%	21%	40%
Geospatial data	27%	14%	19%	19%	38%	27%	27%	26%	8%	29%	40%
Images	19%	24%	17%	11%	38%	13%	5%	16%	25%	7%	-
Video	8%	29%	12%	7%	31%	13%	-	6%	8%	7%	-
Audio	10%	19%	8%	4%	8%	-	-	6%	-	-	-
Other	8%	14%	13%	15%	8%	7%	10%	16%	42%	14%	-
n =	59	21*	127	27*	13*	15*	41	31	12*	14*	5*

Note: Highlighted cells indicate the top three data types by industry.
Multiple responses allowed

Source: Gartner (September 2013)

Figure 2.10: Gartner Survey Results on the Type of Data Analyzed According to Industry (Kart, Heudecker, and Buytendijk 2013)

2.5 Big Data in a Software Ecosystem



As of July 2013

Figure 2.11: Gartner Priority Matrix for Big Data (Heudecker 2013).

According to the priority matrix in Figure 2.11, the business demand for information is going to increase and new sources of information will need to be analyzed. Data analytics has the highest impact and organizations should invest in it as such. The implementation of data analytics

software will be increasingly driven by Big Data technologies which are based on the cloud platform. For example, TIBCO Silver Spotfire (section 6.7) is a cloud-based analytics platform that allows users to leverage analytical capability quickly without large infrastructure costs. Organizations that have the capability to analyze different sources of information for strategic growth will have a better competitive advantage. Organizations need to find different ways such as using cloud computing platform to effectively leverage Big Data capability. Cloud computing is going to be transformational for Big Data implementations in the next two to five years.

Organizations can expect greater benefit from predictive and in-memory analytics than social media monitors such as Big Data. Predictive analytics can actively assist in decision making, whereas social media monitoring is more reactive in approach. For example, using predictive analytics, retailers can estimate their holiday sales and hence stock up their inventory accordingly. Using social media monitoring, such as Twitter data analysis, organizations can see consumer sentiment. It's important that a relative value or risk should be assigned to a given piece of information asset. An information asset is any piece of corporate data or a physical dataset. Information should not only be considered as an asset, but also should be treated and quantified as one. Data storage and management storage can have significant costs if there is no business value that is realized from it. There is no such thing as 'complete information'; organizations must be aware of the amount of information they need to make a decision. Big Data initiatives in large companies are still in an exploratory phase because of such implementations require a large variety of skills.

- Data-management skills in combination with teamwork and effective communication are required to understand business requirements.
- Analytical skills are required to create data models for predictive and confirmatory analytics.

The main reason why business have not fully implemented Big Data projects on a wider scale is because it is difficult to articulate the benefits of such projects to all stakeholders. Implementation is complicated because of lack of relevant skills, which makes it difficult to get the project off the ground. Gartner mentions that software vendors and service providers have cited a lack of skills as a major impediment in developing Big Data initiatives (Kart, Heudecker, Buytendijk, 2013)

2.6 Big Data in Industry

According to the Gartner survey (Beyer and Friedman 2013), telecom and media industries tend to show the maximum interest in implementing Big Data solutions, especially on analyzing social media data. Insurance companies already have a lot of Big Data projects underway for fraud detection and new implementations. Healthcare providers are just starting to implement Big Data projects and are more likely to be customers of BDaaS. According to Gartner research (Beyer and Friedman 2013; see Figure 2.12), business and consumer services organizations have the maximum projected implementation of Big Data followed by financial services. Media and retail are highly influenced by public opinion and hence there is a need to analyze the vast volume of unstructured data from social networks.

Healthcare providers are generally slow to adopt new technologies and hence they are predicted to be late adopters of Big Data technology. Amongst the type of data which will be analyzed by Big Data industries, unstructured social media data is the most common. Machine-generated data in the case of the telecom industry has a large volume, but it may not be as unstructured as the data in social media. The maximum adoption of Big Data technology is expected to be in the media and telecom industry. The government sector is expected to launch Big Data initiatives to process its large document database. In 2012, the U.S. government under the Obama administration launched a Big Data initiative and allocated over \$200M for new research. This initiative was launched to acquire the capability to analyze the massive government databases across federal agencies.

	Low Labels	Social Profile Data	Social Chat Data	Documents	Text/Email	Video	Image	Audio	OT/Machine Data	Average
Business or Consumer Services	70%	60%	20%	40%	10%	20%	20%	30%	34%	
Financial Services (not including Insurance)	31%	34%	38%	41%	6%	22%	9%	16%	25%	
Government (Federal, National or International)	21%	14%	50%	36%	29%	36%	29%	36%	31%	
Healthcare Providers	13%	13%	44%	56%	13%	19%	13%	38%	26%	
Insurance	20%	10%	30%	60%	0%	10%	0%	10%	18%	
IT Service Providers	45%	35%	20%	40%	5%	25%	15%	35%	28%	
Media	85%	62%	31%	15%	31%	31%	23%	54%	41%	
Retail	54%	49%	22%	41%	5%	27%	3%	38%	30%	
Telecommunications	63%	58%	37%	47%	11%	11%	16%	79%	40%	
Average	45%	37%	32%	42%	12%	22%	14%	37%	30%	

* In the next 12 months (beginning December 2012)

The figure represents only those verticals where a minimum of 10 respondents represented the given industry. The chart shows areas of greatest adoption in red, medium or undetermined in yellow, and low probability for adoption in blue.

OT = operational technology

Figure 2.12: Big Data Adaption across Industry Verticals in 2012 in the HEAT Map (Beyer and Friedman 2013).

Chapter 3. Overview of Business Intelligence

This section illustrates how data is managed end-to-end within the organization. Data has to be collected, transformed, and analyzed to extract information to derive value. The analysis can yield useful information, such as predictions of future sales, which can help managers make better decisions. For example, point of sale (POS) terminals store transaction data in data repositories. These data repositories may be analyzed to quantify sales performance for better decision making. Predictive and descriptive analytics operations can be performed on the data to forecast future sales. The first task in extracting information from data is to ascertain the information being sought. The key to retrieving information from data is to know what one is looking for and to construct the framework for data analysis accordingly. While analyzing data for information it is important to establish the originator and the destination or audience of data. Identifying the origination is useful in identifying and qualifying the quality of the data. This analysis also helps in identifying important entities in the data set such as stakeholders (“who is the beneficiary”); the ‘what’ indicates the data can be raw or processed depending on whether it has been subject to any manipulation after it is generated.

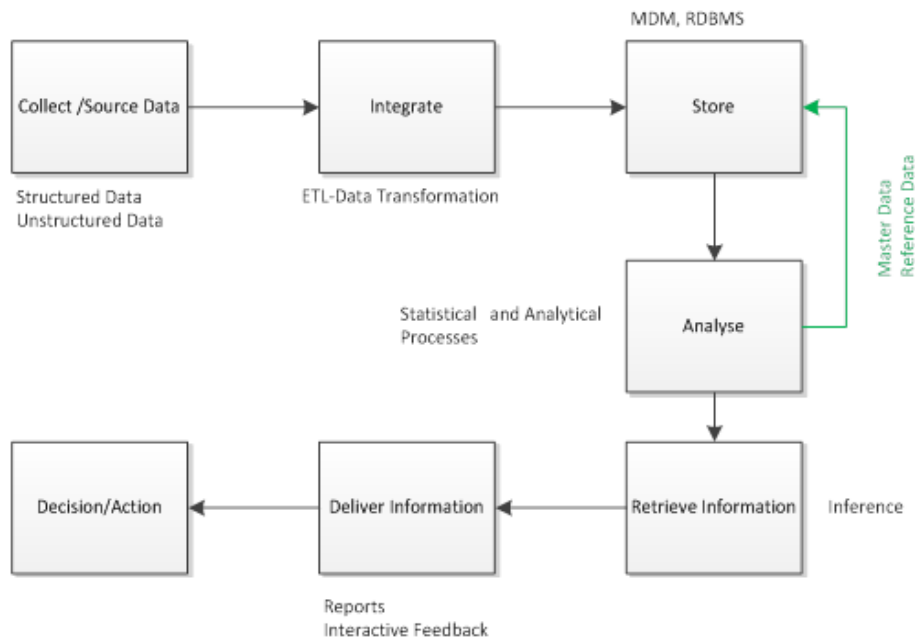


Figure 3.1 Transforming Raw Data into Information

3.1 Collecting Data

Raw data collected from heterogeneous sources is analyzed to yield information (see Figure 3.1). Data can be sourced from databases or files as well as non-traditional systems such as digital sensors. For example, when a customer purchases a product, transaction data is recorded in the information repository. Data contains information regarding the items purchased as well as the financial transaction details. Large stores such as Wal-Mart collect over 2 petabytes of data every day. Whenever an item is sold in Wal-Mart, the information becomes available on an extranet system called “RetailLink,” which the suppliers can monitor. This allows the retailers to restock the items on the Wal-Mart stores if need be. Telecom companies such as AT&T and Verizon have petabytes of data generated through call logs and voice data. The voicemail and text data has to be instantaneously accessible by subscribers. AT&T has its inbuilt project – Daytona, which manages more than 2 trillion records with over 900 billion of them in a single table. The system uses a high-level query language, “Cymbal,” which is translated into an object code through intermediate parsing in “c” to object code. This data includes information on customer call records and billing details, which is essential for business operations. The information from these records allows them to perform business expansion activities such as advertising services for high-value customers. For example, if a customer is identified to be an extensive user of a particular service, then an automatic upgrade is suggested to the consumer for that part.

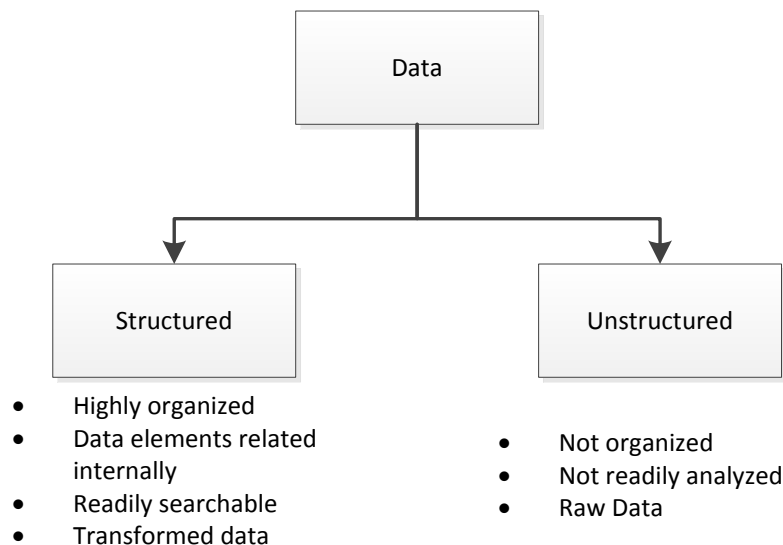


Figure 3.2: Types of Data

Data can be structured or unstructured, depending on its form and internal organization (Figure 3.2). If data has a structure or methodology for its internal organization, it is considered structured. For example, data in RDBMS is stored in tables composed of rows and columns with

keys that relate the individual elements together. If the data cannot be organized into a specific internal structure, then it is considered unstructured data. Unstructured data is inherently difficult to analyze to extract information and requires new technologies and methodologies for handling it.

Organizations generally have multiple systems that act as producers and consumers of information. For example, a point of sale system generates data that is then consumed by the reporting system. As the structure and the intent of the system are different, it is important to integrate the data so that it can be stored and analyzed. In most cases, the data is extracted from various operational systems and transformed before being loaded into a data warehouse, which acts like a central repository. The data warehouse generally stores the data history using techniques such as slowly changing dimensions (SCD). Although the system works effectively for smaller data sets, its performance is highly dependent on a single system. Increase in data volumes requires that the single system be vertically scalable. Large data repositories, such as those for social networking sites, require different data storage techniques such as key-store databases, which have higher performance. The integration amongst various systems is generally done using ETL tools, which are used to transfer and process data. They enable data to be transformed before being loaded into the target system.

3.2 Storing Data

In traditional systems, the data is stored in files or database management systems (DBMS) in various ways (Figure 3.3). The data can be stored in files in different formats depending on its structure. There are different categories of DBMS depending on the internal structure and representation of the data.

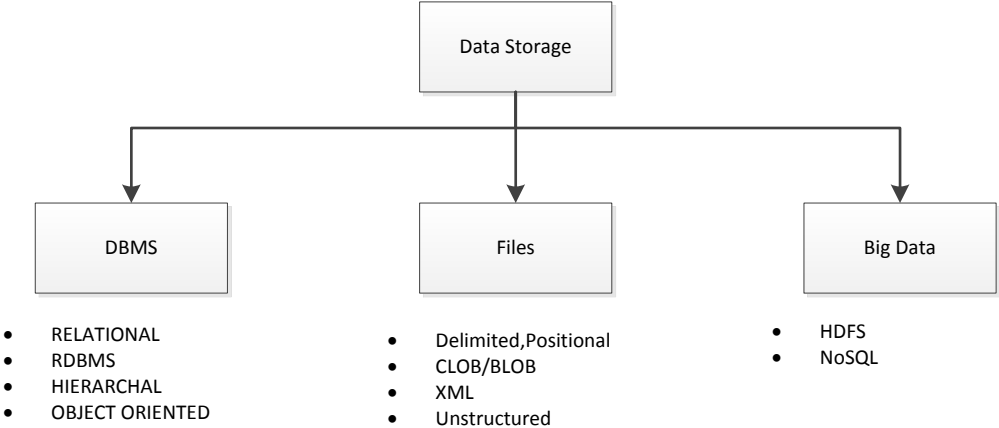


Figure 3.3: Storage of Data

All data within most computer systems at some level is stored in files. The files can have unstructured or structured data. For example, positional and delimited files essentially store data in a table structure where the delimiter in the latter indicates the columns. In the case of a positional file, the location of each column is marked by the position number in reference to a starting point in the file. XML files have a hierarchical structure and the data within them is tagged to represent the elements within them. Examples of unstructured data can be user-written files, e-mails, and other miscellaneous collections of data that do not have a specified structure. The DBMS allows the users to collect, organize, and retrieve the data effectively. The DBMS allows the user to manage and access the data in a structured way. Some of the different types of databases are as are relational (RDBMS), hierarchical, and NoSQL.

RDBMS store data in the form of tables along with their relations, attributes, and keys. They are based on the relational model described by E. F. Codd(1970). The majority of database systems today are RDBMS. Tables in RDBMS can have primary keys to enforce unique identifiers, which prevents duplication of data. A 'relation' is a set of rows that have the same attributes. Foreign keys in RDBMS are used to indicate relations among the elements. In a hierarchical database model, the data is stored in a tree structure, where the data elements are stored in a parent-child relationship. The connected records are arranged in a hierarchical system. The parent or root record is on top of the hierarchy and all the related child records originate from it. A record in a hierarchical database is composed of fields, each of which has one value. Unlike a network model, in a hierarchical model the data is organized as trees. The parent node may have a link to a child node, but the child node should have a link to the parent node. The root instance is a dummy node, whereas all the other children are actual instances consisting of a pointer indicating hierarchy and the record value. These databases enable rapid access of data and move down from the root record. In hierarchical databases, the structure is defined in advance and enables rapid access of data but it has some drawbacks. There is no relationship defined among child elements within a hierarchical database.¹¹ Such databases are also rigid, so adding a new field requires the whole database to be redesigned.

Object oriented databases or OODBMS are designed to handle both unstructured and structured data, unlike the other database model. They can store data from a variety of media sources as objects. The OODBMS store data as reusable objects, which consist of methods and data. The methods indicate what can be done with the data and the data is the actual piece of information such as video, text, or sound. The OODBMS are expensive to develop but offer the ability to handle a variety of media.

¹¹ <http://stat.bell-labs.com/S/>.

NoSQL/KeyStore databases are less constrained by consistency models but focus more on better performance and horizontal scaling capability. Large data repositories such as those with Facebook use key-value store databases for faster retrieval of information. A Key-value store consists of a unique “key” and a “value” which is a single record for all the data. They are much faster than RDBMS as they do not store relations between entities or have as much query capability. They are very useful in storing high volumes of unstructured data, as most of the records are stored “as-is” in a values section with a key associated with it for retrieval.

3.3 Analyzing Data

Data analysis is the process of extracting information from data, which can be used for decision making and other purposes. It uses a combination of techniques from various fields such as operational research, statistics, and computer programming to derive information from data. Figure 3.4 indicates the main categories of data analysis according to their purpose.

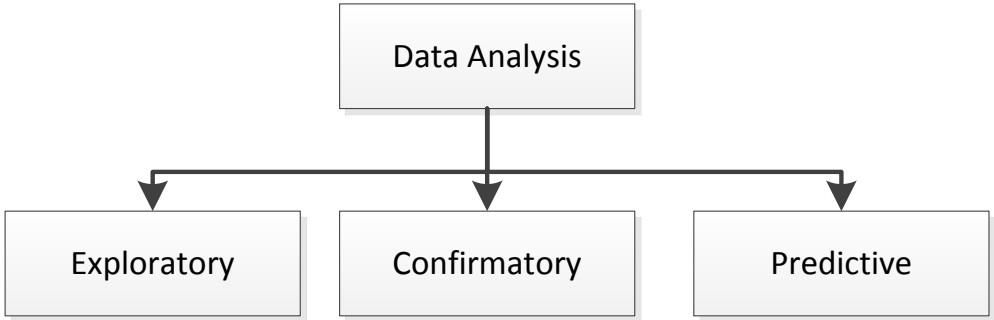


Figure 3.4: Types of Data Analysis

- *Exploratory*: Exploratory data analysis refers to finding information about data which was not known before. For example, data scientists can use exploratory analytics to find patterns in consumer purchases. The outcome of this analysis is generally a graph or other visual tool that can quickly help identify patterns and other relevant information. The main purpose of exploratory data analytics is to see what the data can tell us beyond the existing obvious data. The ‘S’ statistical programming language at Bell Labs was one of the first programming languages designed to organize and analyze data.¹² This programming language later evolved into newer, more capable languages such as ‘R’ and ‘S-Plus’. These statistical programming languages enabled analysts to analyze, identify, and visualize patterns and trends in the data. Exploratory data analysis, unlike confirmatory analysis, does not have a predefined hypothesis on which data is analyzed.

¹² <http://www.crmbuyer.com/story/71081.html>.

- *Confirmatory*: Confirmatory data analysis is related to statistical hypothetical testing, which is the science of making decisions based on data analysis. In statistical hypothetical testing, the results are said to be statistically significant if the results could not have had occurred by chance alone. For example, people may buy wine and cheese in the same shopping transaction as they are generally paired up. A grocery retailer can analyze the transaction data for purchasing habits to see if this phenomenon is indeed true and it might help to advertise those products together or to keep them close to each other.
- *Predictive*: Predictive analytics refers to predictions about the future through analysis of data. It uses algorithms to recognize patterns in data to determine future possible values. Predictive analytics are used in various fields such as banking, marketing, retail, and travel. It deals with identifying unknown events of interest in the past, present, or future. The predictive analytics can be used to gauge future events such as predicting the stock price of a company by analyzing past data. Predictive analytics can also be used to analyze past events such as analyzing spending patterns on a credit card to determine fraud. Predictive models are used to analyze consumer behaviors by determining their past actions.

3.4 Retrieving and Presenting Information

After the data has been analyzed, the next step is to generate reports to communicate the information. These reports help to consolidate and represent the information in a visual way so that they can be easily interpreted. Smartphones have the processing and display capability to access information from information systems. For example, a major provider of supply chain service to healthcare providers across the country launched the VHA SupplyLYNX platform in 2001, which enabled healthcare providers to optimize their supply chain. The company launched VHA PriceLYNX™ mobile¹³ to enable healthcare professionals at hospitals to gain insight into current product prices. The platform was developed for mobile devices and was connected to the VHA's main SupplyLYNX system to get the results of analytics. Figure 3.5 shows the price index trend for a medical devices manufacturer which can help hospital managers to gauge the price trends from their suppliers.

¹³ <http://www.rba.co.uk/wordpress/2010/03/27/google-public-data-explorerer-fine-as-far-as-it-goes/>.

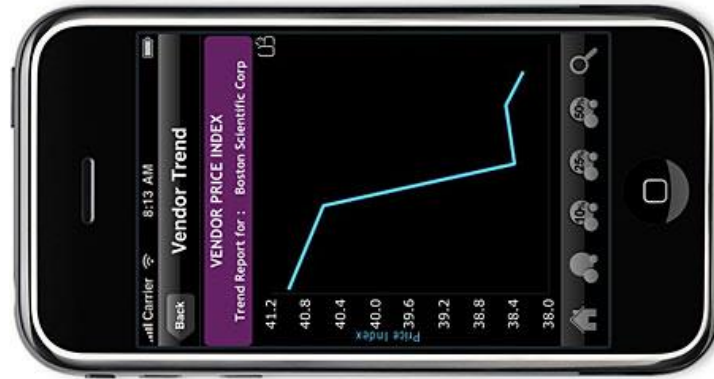


Figure 3.5: PriceLYNX iPhone App (VHA-INC)¹⁴

Electronic dashboards allow users to view analytical reports over a computer either through a website or a dedicated application. An excellent example of BDaaS reporting would be the Google public data platform. The Google public data explorer enables users to access, analyze, and view the large publicly available datasets from governments and international organizations. For example, Figure 3.6 shows that the user can access an interactive dashboard that consolidates the records from data that is publicly available from sources such as the World Bank, Eurostat, and Data.gov to show a graphical view. This interactive graphical view of the dataset enables users to perform visual analysis of the data such as spotting trends.

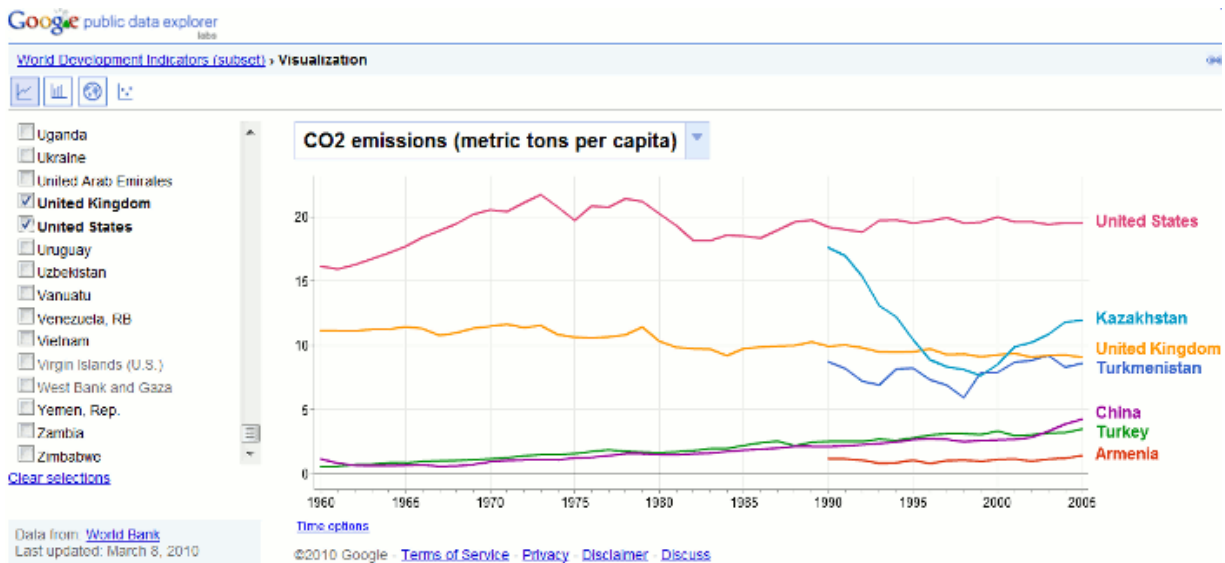


Figure 3.6: Google Public Data Analytics and Reporting (Davenport 2012)

¹⁴ https://www.vha.com/Solutions/Analytics/Pages/PriceLYNX_iPhoneApp.aspx.

Chapter 4. Big Data as a Service

According to IBM,¹⁵ 90% of the data in the world, which includes both structured and unstructured information, has been created in the last two years. According to Gartner, by 2016, 50% of the data in organizations will be stored in cloud-based systems (Smith 2013). Vendors are combining distributed computing architecture capable of addressing Big Data problems with cloud computing frameworks resulting in a new category of technology called BDaaS.

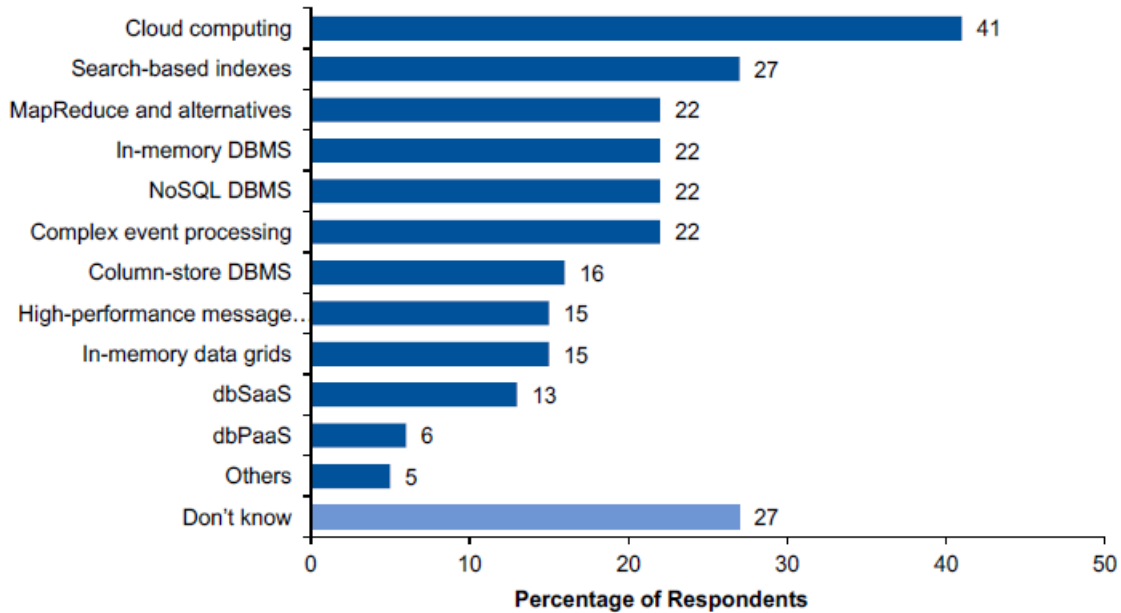
When organizations initiate data analytics projects, their computing requirements vary widely (Davenport 2012). In these cases it is easier to implement the solution using cloud-based systems (Davenport 2012). Predictive analytics on large datasets use a lot of computing power in short “bursts” and cloud computing enables the architecture to leverage it as well (Davenport 2012). BDaaS on the cloud platform allows organizations to have flexible deployment architecture by allowing them to instantiate computer clusters on demand as required. As the clusters are preconfigured, users can run analytical jobs as required and then shut them down after the task is complete.

Leveraging the ubiquitous access capabilities of the cloud computing platform with Big Data analytics will enable researchers and organizations to implement analytical platforms on demand and utilize them as required. Hence, when Big Data analytics is delivered as a service (BDaaS), it enables small- and medium-scale industries to implement the framework effectively by minimizing cost and resource demands. Traditionally, most of the data within an organization is stored on a non-cloud system, but with increasing adoption of service-based applications, organizations are increasingly moving to cloud systems.

4.1 Cloud and Big Data

In 2013, Gartner conducted a survey (Kart, Heudecker and Buytendijk 2013) of over 720 IT executives from organizations across various sectors to identify the key drivers in the implementation of Big Data technologies. According to the survey, Big Data is not replacing the traditional systems but is supplementing it with additional capability to handle large datasets. Along with increasing demand for infrastructure capability, IT decision makers are facing a problem of understanding which technologies are required and how they fit into the technology landscape (AUTHOR(S) 2013; see Figure 4.1). Organizations are looking towards cloud-enabled Big Data technologies to reduce costs and time to deployment for Big Data projects. In the survey, over 41% of respondents were looking towards cloud-based Big Data solutions for their projects

¹⁵ <http://www.theatlantic.com/sponsored/ibm-cloud-rescue/archive/2012/09/Big-Data-as-a-service/262461/>.



N = 465 (multiple responses allowed)

Figure 4.1: Technologies to Derive Value from Big Data Projects (Kart, Heudecker and Buytendijk 2013)

The survey indicated that organizations are increasing their data by almost 40% annually, which would require them to double their infrastructure capacity every year. Cloud computing offers a promising and a flexible way to address the challenge of growing data sets. To address the challenges of storing and managing Big Data, it is necessary to have a balance in cost effectiveness, scalability, and performance. Organizations may use network systems such as storage area network (SAN) and direct attached storage (DAS) to meet the Big Data storage requirements, but these systems are expensive and not as scalable as cloud-based systems.

A growing segment of the cloud computing market is system infrastructure services such as storage and basic computing. Leading organizations such as Netflix and Dropbox use cloud-based Big Data storage services such as Amazon’s simple scalable storage (S3) to store data in object form. According to Gartner (David 2013), the compound annual growth rate from 2012 to 2017 is likely to be over 30% and the organizations are likely to increase their cloud storage spending from \$2.4 billion to \$8.7 billion. The National Institute for Standards and Technology (NIST) defines cloud computing as follows:

“Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”. (Mell and Grance 2011)

Cloud computing technologies can be classified according to the level of services they provide. There are three main provisioning models of the cloud computing stack.

- *Infrastructure as a service (IaaS)*: In this model, the cloud computing service provides the basic technology infrastructure in terms of virtualized operating systems, networks, and storage devices. Amazon EC2 is an excellent example of IaaS where users can instantiate computing machines on demand for their use.
- *Platform as a service (PaaS)*: Cloud computing service in this provisioning model includes both the underlying infrastructure and the applications deployment platform. Cloud services in this model can include features to promote application design, development, and deployment. Some examples of PaaS vendors are Heroku and Force.com.
- *Software as a Service (SaaS)*: In this provisioning model, application software is included so that users can access it on demand. In this case, users do not have to worry about setting up or maintaining applications as they are managed by the service provider. Users can use the software directly without much setup required.

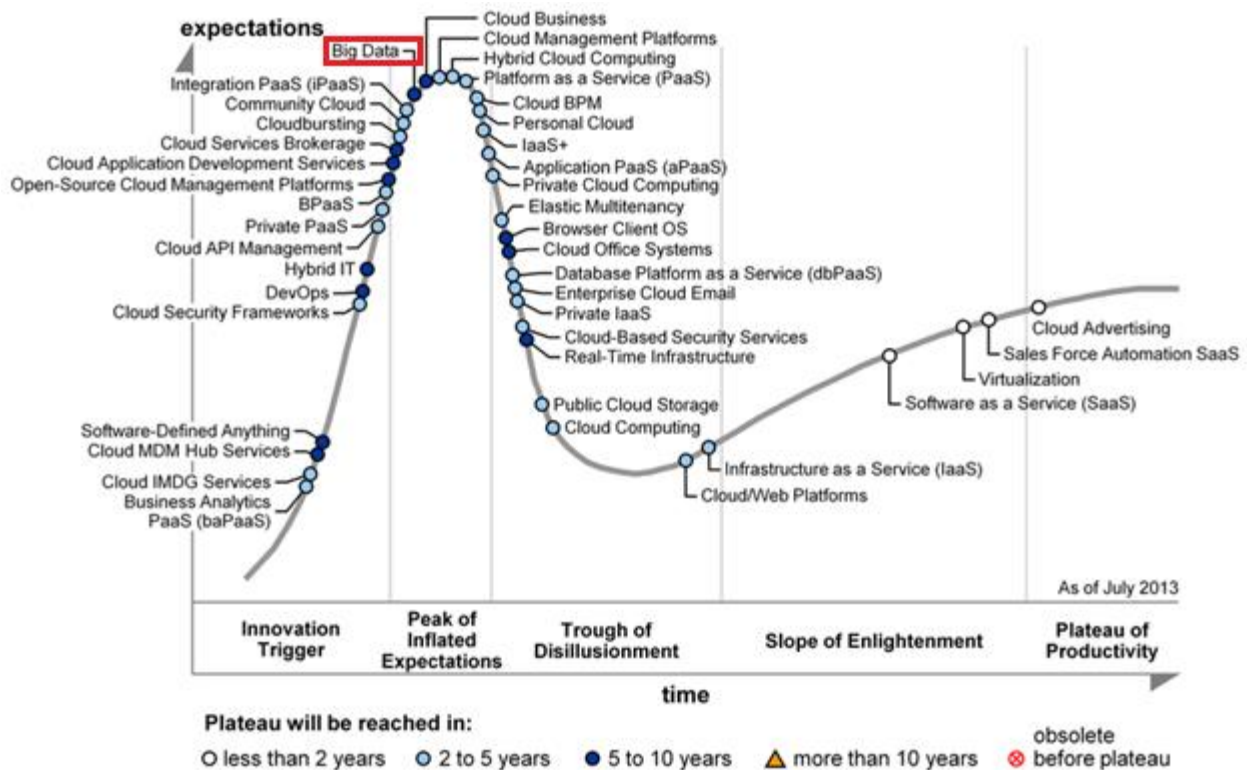


Figure 4.2: Hype Cycle for Cloud Computing 2013 (David 2013).

As we can see from the Gartner's hype cycle for cloud computing, published in 2013 (Figure 4.2), Big Data is one of the most highly anticipated topics and among those with the most innovations. Cloud computing-enabled Big Data solutions offer on-demand scalability and ease of use to address some of the issues successfully. Dropbox and Netflix have successfully demonstrated how small, high-growth organizations can effectively leverage Big Data technologies through the cloud computing platform for data storage and management of data. For example, Dropbox stores user data in an encrypted form on the Amazon S3 system (KPMG International 2013). In cloud computing-enabled BDaaS technologies, the cloud service company takes care of the complexity of implementation, hence easing implementation for the user. Within the cloud computing platform, the Big Data analytics can be thought of as a service rather than a physical implementation. So customers can request a preconfigured computing environment based on the Big Data framework from the cloud service company. Cloud computing is a proven way to deliver services and infrastructure on demand, enabling organizations to grow and scale up as required. Big Data combined with cloud computing promises to leverage the flexibility of cloud computing coupled with the ability to analyze large datasets. Big Data still has very high expectations from the technology market and the concept will reach maturity within a few years. The development of Big Data on the cloud platform depends on the maturity of the cloud deployment models such as IaaS and PaaS.

4.2 BDaaS Framework

The BDaaS service stack is composed of layers that group technology types according to the function they perform (Figure 4.3). For example, the data analytics layer includes technologies such as Tibco Spotfire Silver, which provides a cloud-based analytical platform. The data-storage layer has Amazon S3, which provides the storage services. The lower layers of the BDaaS service stack are closer to the IaaS platform of cloud computing. The upper layers of the BDaaS stack have a presentation layer, which enables users to access the services. The ease of use of the service stack increases at higher layers. Each layer has a specific use and abstracts the complexity of distributed Big Data processing from the end users.

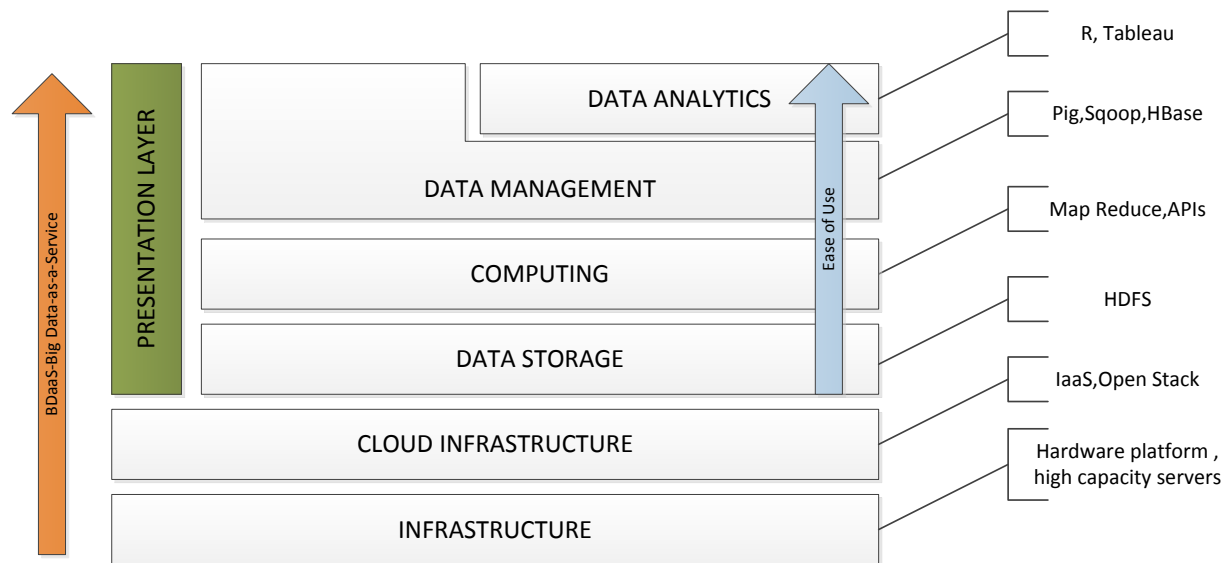


Figure 4.3: Big Data as a Service (BDaaS) Stack

- Data Analytics:* This layer includes high-level analytical applications similar to R or Tableau delivered over a cloud computing platform which can be used to analyze the underlying data. Users can access these technologies in this layer through a web interface where they can create queries and define reports that will be based on the underlying data in the storage layer. Technologies in the data analytics layer abstract complexities of the underlying BDaaS stack and enable better utilization of data within the system. The web interface of those technologies may have wizards and graphical tools that enable the user to perform complex statistical analysis. For example, if data in the text file is to be analyzed, then the user can upload this text file to a Tableau cloud platform and then run analysis by starting a ‘wizard,’ which will have step-by-step instructions to achieve the task. Technologies in this layer can further be specialized according to the industry. For example, a BDaaS provider can have an analytical layer specialized for the financial industry, where the functions and the capabilities of the analytical and the presentation layer are geared towards common operations in the industry. The layer will have functions for monitoring risk, performing banking operations, and simulating stock prices. This ability to specialize the data analytics layer within the Big Data stack makes it usable and adaptable to most organizations
- Data Management:* In this layer, higher level applications such as Amazon Relational Database Service (RDS) and DynamoDB (see Chapter 6) are implemented to provide distributed data management and processing services. Technologies contained in this layer provide database management services over a cloud platform. For example, if a user needs to have an Oracle database over the cloud, then using Amazon RDS such a service

can be instantiated. As data-management services are managed by the service provider, such as taking periodic backups of data, this eases deployments and reduces resources requirements for their upkeep.

- *Computation Layer:* This layer is composed of technologies that provide computing services over a web platform. For example, using Amazon Elastic MapReduce (EMR), users can write programs to manipulate data and store the results in a cloud platform. This layer includes the processing framework as well as APIs and other programs to help the programs utilize it.
- *Data Storage Layer:* This data storage layer is typically distributed in a portable file system, which can be scalable on demand. This can be a typical HDFS file system composed of a name node and a cluster of data nodes, which can be scaled up according to demand. This layer has high availability build into the system. The data storage layer will have a presentation component where users can directly interact with the HDFS file system to upload data for analytics.
- *Cloud Infrastructure:* In this layer cloud platforms such as open stack or VMware ESX server provide the virtual cloud environment that forms the basis of the BDaaS stack. This layer can also perform usage monitoring and billing for the services rendered. Although this layer is a part of BDaaS, it is not accessible directly through external applications. For example, this layer does not have a presentation option and hence cannot be accessed in the same way the AWS layer would be.
- *Data Infrastructure:* This layer is composed of the actual data center hardware and the physical nodes of the system. Data centers are typically composed of thousands of servers connected to each other by a high-speed network line enabling transfer of data. The data centers also have routers, firewalls, and backup systems to insure protection against data loss. This data center provides computing and storage power to multiple consumers of the BDaaS service. The costs for building a datacenter can be immense, with over \$1.5m for each 1000 sq. ft. of area.

Moving Big Data analytics to a cloud computing platform enables small organizations or individuals to ramp up the infrastructure quickly on demand without incurring prohibitive setup and maintenance costs (Edala 2012). Most organizations are not looking for predictive or business cloud-based solutions; their requirement pertains to a solution for a specific business challenge (Davenport2012). The cloud-based solution allows the organizations to deploy easily and is cost-effective as well (Davenport 2012). A lot of proofs of concepts (POCs) start by

having the system deployed on a cloud-based platform. This not only reduces the development time, but also enhances access and collaboration (Davenport 2012).

There are a lot of software products in the market today that claim to have Big Data and cloud capability. This lack of clarity results in customers sometimes buying software that does not meet their requirements. Sometimes organizations end up trying to force-use the software because they have purchased it. It is not uncommon for organizations first to buy the software, as they do not want to get left out of the technology wave or hype, and later to try to fit the software somewhere into their projects. For technologies to be in the BDaaS stack, they have to follow certain essential characteristics, which make them part of a layer or multiple layers. This layering helps organizations to place their products correctly in a particular layer for product placement and marketing. It also helps customers to identify a product that best meets their needs.

4.3 Essential Tenets

Each layer in the BDaaS has particular characteristics as defined below. Technologies belonging to a particular stack should embody those characteristics.

- 1) Technologies should be clearly defined to be in a particular stack or multiple stacks.
- 2) They should have some interconnectivity to other technology in the stack above and below them.
- 3) They should provide the same basic set of services as others in the same stack.
- 4) They should be scalable and delivered over the cloud platform.

For example, Amazon S3 is a technology in the data-storage layer as it is scalable storage for software objects such as large files delivered over a cloud platform. As this technology is in the data storage layer, other technologies in the computing layer should be able to interface with those in the data management layer. Consumers wishing to buy scalable cloud-based storage for Big Data can evaluate just those technologies in the data-storage layer for effective evaluation and comparison of related technologies. Technologies can be a part of multiple stacks where they provide much greater value to consumers. For example, Amazon EMR enables organizations to create Hadoop clusters on the fly. It provides the capability of both the computing and data storage layers.

As technologies in the BDaaS stack are based on a cloud computing platform, most BDaaS products span over cloud and data infrastructure layers. As organizations specialize their products and add more capability, the products grow up the technology stack. The best products

in the BDaaS space are those that work across multiple layers. Organizations providing software products, especially in the data storage layer, will soon find that their services are being increasingly commoditized. The lower layers in the stack, such as cloud computing or data storage, can be differentiated according to cost of service and performance, whereas the upper layers can be differentiated according to domain or specialization. For example VeevaCRM is an add-on technology on top of Salesforce.com CRM that enables users to have specialized functions for the pharmaceutical industry. Their clients include some of the largest pharmaceutical industries in the world. As the technologies move up the stack and become easier to use, they require domain knowledge in a particular field to be used effectively. For example, as mentioned earlier, VeevaCRM, which is focused on the life sciences industry, requires users to have some domain knowledge to work effectively with it.

According to Gartner (Kart , Heudecker and Buytendijk 2013), in 2013 most organizations were experimenting with and deploying Big Data systems, with fewer than 8% of the organizations claiming to have deployed Big Data solutions. Almost 20% of organizations are still evaluating various Big Data technologies in the market, with over 18% still in the phase of gathering knowledge and developing strategy (Figure 4.4). The BDaaS framework may help these organizations to move quickly from the knowledge-gathering to the deployed phase. As BDaaS technologies are cloud based, the time taken for piloting and experimenting can also be reduced, resulting in substantial savings in time and expenses. In the knowledge-gathering phase, organizations try to get information about new technology in the market that might be applicable to their needs. In this stage the investment in the new technology is not much. In the developing strategy stage, organizations try to identify technologies that can help solve a specific business problem. In the piloting and experimentation stage, various technology vendors may be contacted to develop prototypes for evaluation. In this stage, the technology is selected and it is then used for deployment in the next stage (Figure 4.5).

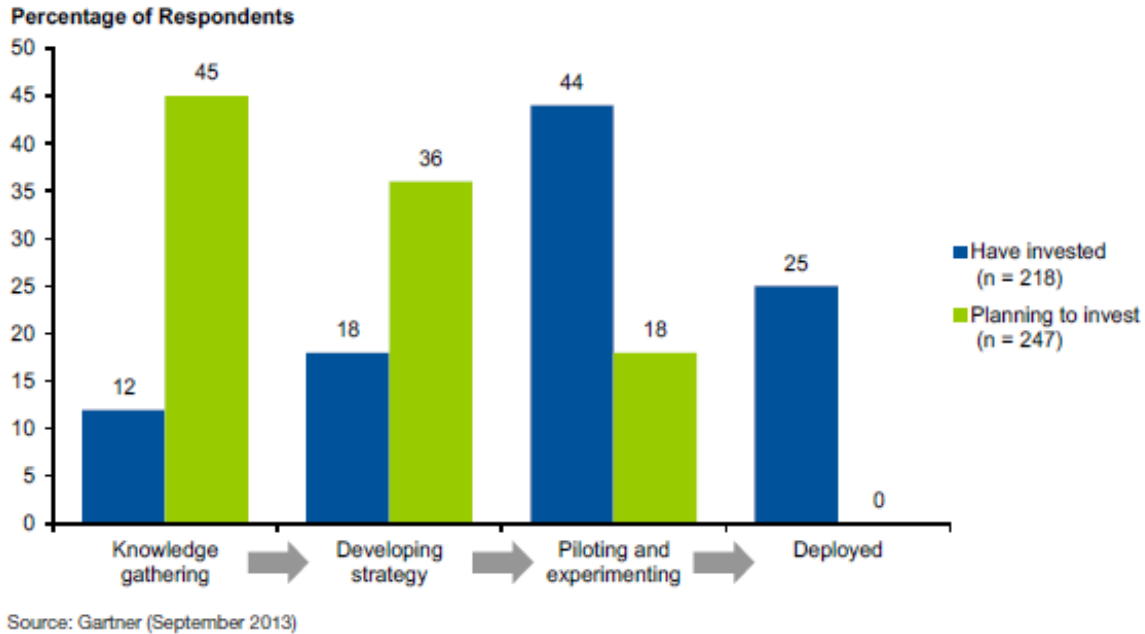


Figure 4.4: Big Data Adoption by Investment Stage (Kart, Heudecker and Buytendijk 2013)

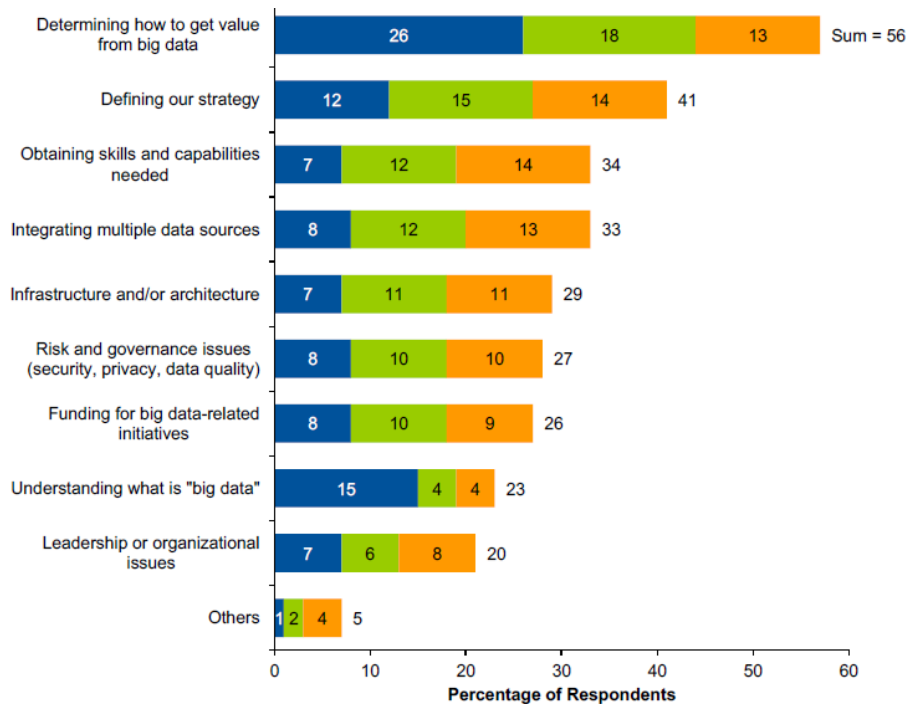


Figure 4.5: Top Big Data Challenges (Kart, Heudecker and Buytendijk 2013)

The main issue with organizational management from sponsoring and implementing Big Data projects is to determine the value tied to Big Data initiatives. Over 56% of the respondents of the survey (Kart, Heudecker and Buytendijk 2013) cited the business value derived by implementing Big Data technologies as their main concern. Using the BDaaS framework, organizations can

quickly identify the technologies that meet their requirements and accelerate development of prototypes using those technologies, which can help the organizations to understand the true value of Big Data. For example, if an organization is looking for distributed storage systems and the capability to write MapReduce jobs, it can look at the technologies in the data-management and computation layers such as Amazon S3 and EMR and implement them in projects, accelerating development and reducing costs.

If organizations are looking to increasing their sales, they might need new data analytics platforms to do predictive analytics on their business data. Using the predictive analytics provided by technologies in the data-analytics layer, organizations can quickly create reports that might help them in decision making, such as estimating inventory or identifying the hottest products for sales growth.

4.4 Big Data Concerns

Although Big Data is a popular buzzword that is interpreted in different ways, there are certain misconceptions associated with it. There are new challenges and apprehensions that organizations face when implementing Big Data technologies which need to be addressed. Big Data is considered synonymous with MapReduce technology, but the latter is an algorithm based on distributed data processing that helps in addressing the Big Data problem (Heudecker 2013). A Hadoop framework is based on the MapReduce algorithm and is used by major organizations to process large datasets or Big Data and hence it is commonly associated with all Big Data implementations. Although Big Data is commonly referred to when there is a large volume of data, the variety and velocity of the data are also equally important factors to consider. A high variety and volume of data can overwhelm the infrastructure as well. Big Data is not just the Hadoop framework and MapReduce technology; although they are the most commonly used implementations, there are alternatives. Some of the major issues challenges with Big Data are:

a) Secure computations in distributed processing

In distributed processing, where multiple processes operate on separate datasets in parallel, identifying and securing security loopholes is difficult. For example, the earlier Hadoop versions had data privacy issues regarding user authentication.¹⁶ They were designed to run all users' programs and relied on OS-level authentication for security. The auditing and authorization mechanism in the earlier versions of HDFS system could be overridden by user impersonation using a command line switch. As all users had the same access level to the data in the Hadoop cluster, anyone could write a program to read any data set in the cluster. Due to lack of effective

¹⁶ <http://www.infoq.com/articles/HadoopSecurityModel/>.

authentication and authorization in earlier Hadoop versions, users could override the priorities of other running programs as well.

When Hadoop was launched, it was limited to private networks behind company firewalls; hence security was not as much of an issue. When these distributed computing frameworks such as Hadoop are integrated with a cloud computing network for BDaaS, then these security flaws can be very serious. Malicious users can execute programs in the cloud environment that exploit the vulnerabilities of these distributed frameworks. Technologies in the lower layers of the BDaaS stack such as infrastructure or computing are more vulnerable than those above them, as they give more accessibility and are closer to the data being manipulated.

b) Portability

Portability of data to a cloud computing environment using BDaaS technologies is difficult. When the Big Data infrastructure is moved to the cloud, there are two main issues that organizations have to address: regulatory and integration.

- *Regulatory Issues*

When organizations use BDaaS technologies, they essentially move the data to an offsite location. When data is in an offsite location, out of direct control of the owner, there might be various regulatory and internal policies affecting it. Most banks do not allow their internal data to be moved to a public cloud. Organizations having restrictions on moving data to a public cloud may still use a private cloud and establish an interdepartmental BDaaS system to cater to their requirements. The federal government has launched a federal risk and authorization management program (FedRAMP) to standardize and mitigate risks for cloud computing environments in government implementations.¹⁷ It provides a security risk model that can be used across federal agencies to implement cloud-based technologies.

- *Integration with existing architecture*

Organizations have large data sets that are difficult to move to BDaaS environment. This requires alternate ways physically to migrate the data in an efficient way to the cloud environment. Amazon has an AWS import/export facility so that users can use portable devices for transporting large datasets. Amazon then uses those portable devices to upload the data directly into the cloud system. This service also helps in backing up data as well as content distribution. In some cases, data may have to be transformed before BDaaS technologies can use

¹⁷ http://www.gsa.gov/portal/category/102371?utm_source=OCSIT&utm_medium=print-radio&utm_term=fedramp&utm_campaign=shortcuts.

it. For example, if an organization has a data warehouse that it wants to migrate to the cloud for data analytics services provided by the technologies BDaaS stack, it may have a particular data model that may not map exactly to the one in the cloud environment. Organizations wishing to move to BDaaS technologies have to ensure that the new environment integrates with existing architecture. The problem of portability persists even if the data has been uploaded to the cloud environment. If an organization is using a particular BDaaS service and it decides to move over to another cloud provider, then the data may have to be extracted into a non-cloud environment before being uploaded into to the new BDaaS service. This is a very expensive and disruptive operation and it increases the lock-in effect of BDaaS vendors, especially those in the upper layer of the stack.

c) Reliability and Availability

The cloud computing providers advertise greater than 99% availability for services but yet, lack of direct control of infrastructure is a major deterrent to moving to a cloud architecture. When a BDaaS service is used, reliability and availability become major issues. For example, if an organization wants to run a single ad-hoc computing task that requires over 1000 machines, then the service provider must be able to do so. In the case of the services in the lower layers of the BDaaS stack, such as Amazon EC2, users have various choices such as on-demand, reserved, or spot instances that determine the availability as well as the cost. In the case of on-demand instances, users pay for the computing services and do not have upfront payments. It's generally more expensive but has the highest reliability. In the case of reserved instances, the users make an up-front payment to reserve computing services for a particular duration, usually a few years. This is cheaper than the on-demand instances and is suitable when the services have predictable demand. The last is the spot instances, where the users can specify the maximum price per hour they are willing to pay and depending on how the spot proceed fluctuates, the users may or may not get the machines. Spot computing is the cheapest option, but the most vulnerable to availability problems. It is more useful when large amounts of computing resources are required instantaneously at a low price. From Figure 4.6, we can see that the prices for on-demand instances are generally higher in all machine categories than reserved or spot instances.

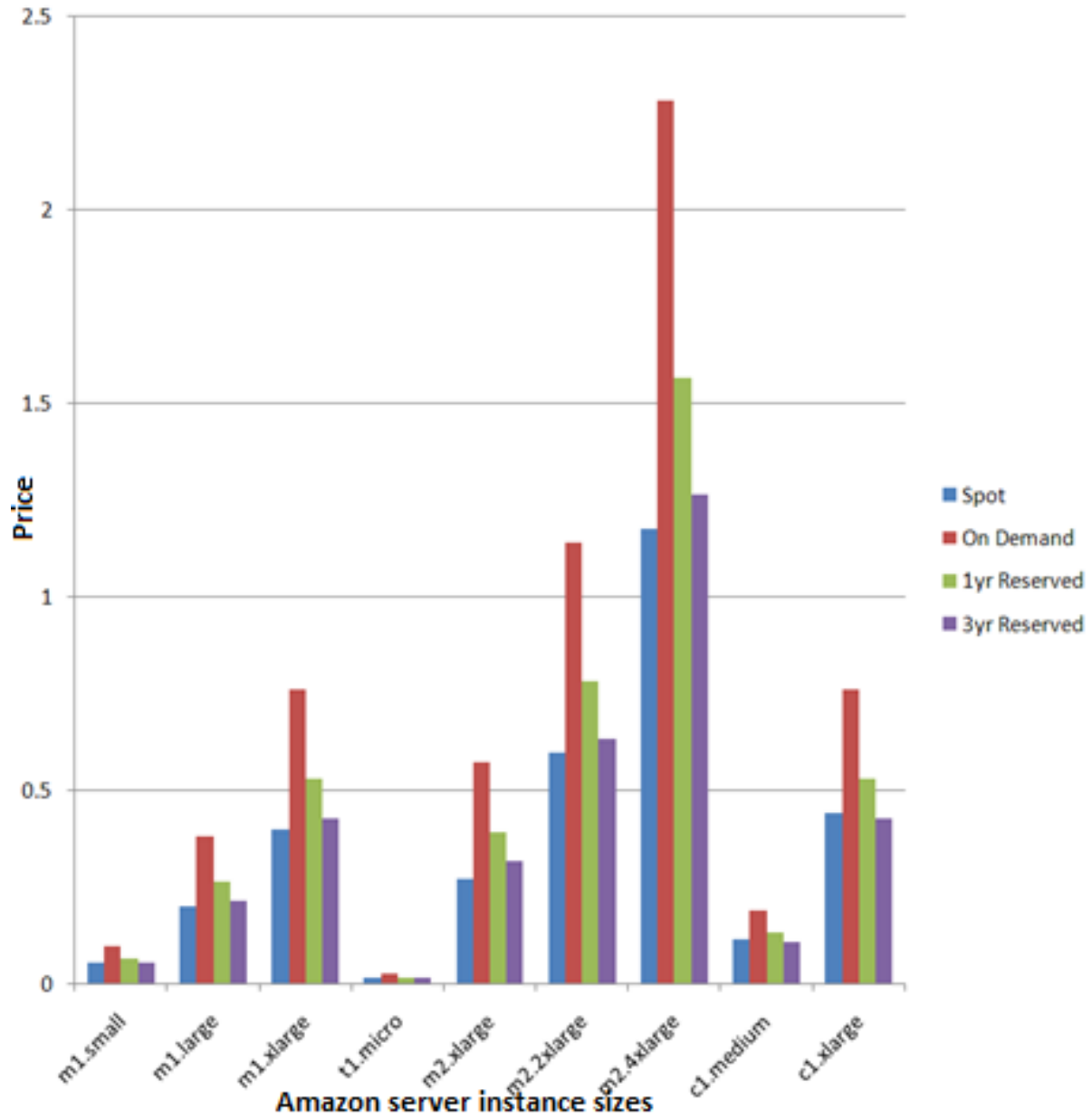


Figure 4.6: Comparing Amazon Spot, Reserved, and On-demand Instance Average Prices¹⁸

¹⁸ <http://cloudcomments.net/2011/05/16/dont-forget-spot-instances-on-aws/>.

4.5 Advantages and Disadvantages of BDaaS

Table 4.1 shows the advantages and disadvantages of each layer in the BDaaS stack. Most of the technologies evaluated in the BDaaS stack span more than one layer. The data analytics layer is more useful for analysts and business users, as they perform analytical tasks and create business reports. The technologies in the data-management and computing layers require significant technical and programming expertise for implementation. The technologies in the cloud and infrastructure layers mostly deal with the underlying computing resources such as virtual machines or hardware. Most of the technologies require significant technical expertise for operations. There are skills required to work with the technologies in the data management, computing, and data-storage areas, which are particularly sought after. Lack of the required skills is one of the most pressing challenges in implementing Big Data systems.

	Advantages	Disadvantages	User Base
<i>Data Analytics</i>	<ul style="list-style-type: none"> • Users can readily access analytics services without the hassle of data or infrastructure management 	<ul style="list-style-type: none"> • No direct access to data • Analytics limited to the data exposed in the layer 	<ul style="list-style-type: none"> • Business users • Data scientists • Business analysts
<i>Data Management</i>	<ul style="list-style-type: none"> • Direct access to data • Ability to analyze or modify complex data sets 	<ul style="list-style-type: none"> • Requires programming knowledge to operate 	<ul style="list-style-type: none"> • Database administrators • Programmers
<i>Computing</i>	<ul style="list-style-type: none"> • Most flexibility as programmers can write programs to manipulate data 		<ul style="list-style-type: none"> • Programmers
<i>Data storage</i>	<ul style="list-style-type: none"> • Access to raw data in the distributed storage 		

<p style="text-align: center;"><i>Cloud infrastructure</i></p>	<ul style="list-style-type: none"> • Can be used to instantiate IT infrastructure • Ability to determine the capability of the overlying infrastructure 	<ul style="list-style-type: none"> • Requires infrastructure knowledge to operate 	<ul style="list-style-type: none"> • Infrastructure engineers • Programmers
<p style="text-align: center;"><i>Infrastructure</i></p>	<ul style="list-style-type: none"> • Basic hardware on which computing infrastructure is based 		<ul style="list-style-type: none"> • Infrastructure engineers

Table 4.1: Advantages and Disadvantages of Each Layer for Users

4.6 Industry Perspectives of BDaaS

Leading industry experts predict that owing to recent awareness of the Big Data phenomenon, there will be increased demand for talent with Big Data analytics skills. By 2018 the skills gap between existing talent and demand will grow to over 1.7 million professionals. There is a limited supply of available talent and organizations are sometimes reluctant to develop these talent pools internally. It is expected that a lack of the analytical and managerial talent necessary to work with Big Data will be one of the most pressing challenges that organizations must address. As the BDaaS technologies are managed by the service provider, they can help organizations meet their need for resources skilled in Big Data technologies. For example, if an organization needs to have a distributed data storage system such as HDFS, then it has to hire infrastructure specialists who have those specialized skills sets. To maintain the new Big Data infrastructure is an added cost because of the expensive skills that need to be hired. Implementing Big Data is an expensive proposition for large organizations, as the skills required to implement such projects are rare. It is estimated that the service cost to implement Big Data projects is almost 20 times that of the software (Beyer et al. 2012).

Using technologies in BDaaS, organizations can get a service set up using Microsoft Azure or QuBole, which will provide the required infrastructure without necessitating expensive hires with specialized skillsets. The service provider companies will take care of maintenance and upkeep of such technologies and service consumers can focus on using services to address their business requirements. Technological advantages and reducing hardware prices have led organizations to accumulate massive amounts of data, which needs to be analyzed. Organizations expect data analyst professionals to be capable of analyzing data, identifying patterns, and

communicating the findings to senior management. The demand for analytics professionals, popularly known as data scientists, who can master cutting-edge technologies such as Hadoop, is increasing. Data scientists typically have statistical and data modeling skills, which help them to do predictive analytics for decision making.

The Gartner ascendancy model for analytics (Maoz 2013; see Figure 4.7) indicates the maturity level of data analytics used by the organizations in their operations. The model also indicates the progression of an organization from descriptive analytics to prescriptive analytics as its analytics capability increases. As organizations grow up in the ascendancy model, the need for data analytics professionals is going to increase as predicted. Organizations that grow faster than their counterparts, according to the analytics ascendancy model, are more likely to have a competitive edge derived from their analytical capability. BDaaS technologies will enable organizations to reach the higher levels of the ascendancy models faster. For example, an organization with basic reporting capability can do descriptive analytics using simple tools such as Excel, but for predictive and prescriptive analytics it may need advanced specialized analytical tools such as Tableau or Spotfire. Using the cloud-based analytical tools in the BDaaS analytics layer, organizations can quickly implement such tools and reduce their implementation costs, especially if the need is infrequent.

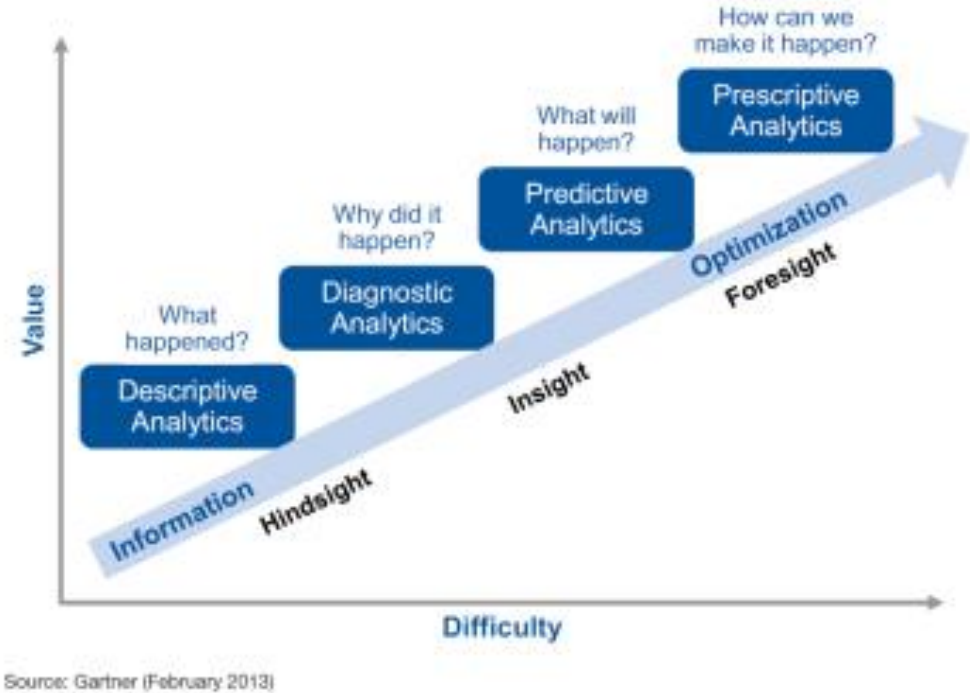


Figure 4.7: Gartner Analytic Ascendancy Model (Maoz 2013)

The InformationWeek Group (2012) conducted a survey of over 660 business intelligence and analytics professionals to understand the skills gap in industry (Figure 4.8). The survey indicated a growing demand for data analytics talent across all sectors. This need for specialized talent impedes organizations from growing their analytics practices. IT salaries for the BI/analytics group tend to be near top of the salary spectrum owing to the demand. The rising salary of these professionals results in a significant cost increase in developing and manning a dedicated staff to handle analytics requirements for the organizations. Technologies in the BDaaS stack will enable these organizations to ease the implementation, reducing the demand and hence the cost of acquiring Big Data capability.

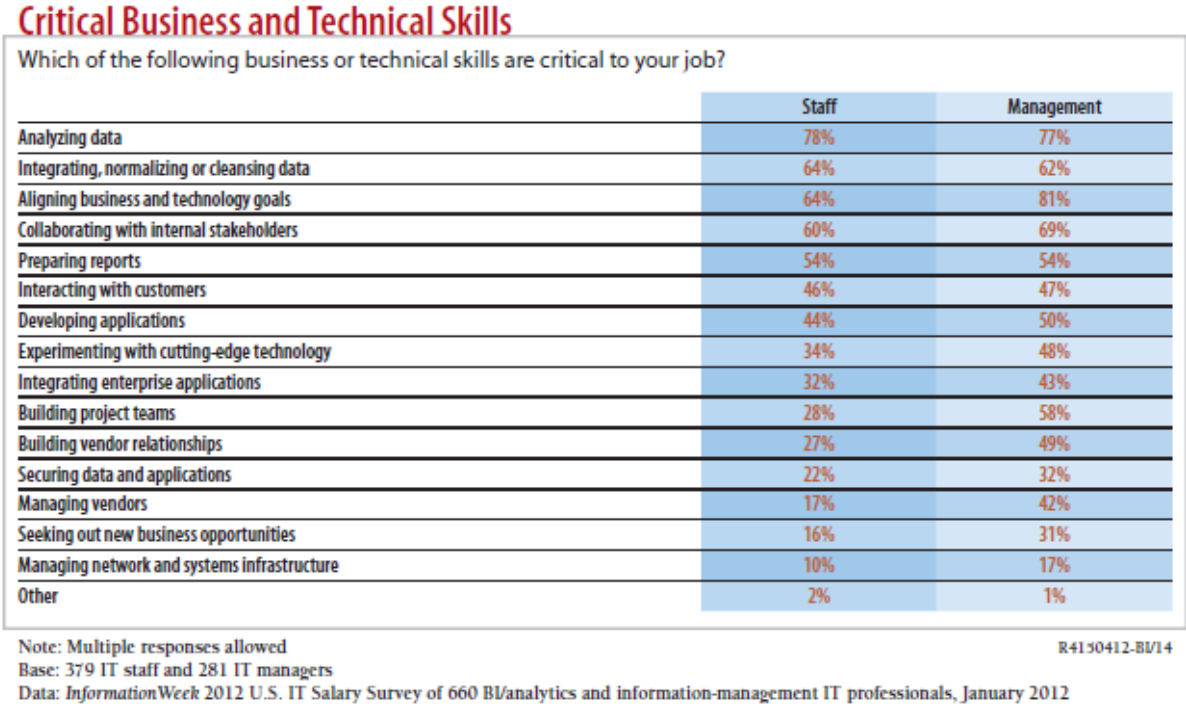


Figure 4.8: Critical Business Skills

The BDaaS platform reduces the need for people to have advanced knowledge in statistics and computer science to be able to perform analytical tasks better. For example, organizations may have a team of analysts to perform data-mining and analysis tasks, but may not have skilled resources to set up the Big Data infrastructure. In this case, the organizations can use technologies in the BDaaS stack, where the hosting and maintenance of Big Data infrastructure is managed by the service provider and the service consumers (in this case the organizations) can focus on the tasks directly relevant to their business without the need to hire specialized talent. For example, an organization might have a team of business analysts who have the required skills for analyzing data, but they perhaps lack the computer science skills that are required to set

up an infrastructure for Big Data processing. Using BDaaS, analysts can focus on using the analytical tools without worrying about the infrastructure setup for Big Data, which requires computer science skills. They can use cloud-based analytical tools such as Tibco Silver Spotfire, where they do analytics operations and the service provider takes care of the infrastructure setup.

Chapter 5. Case Studies of BDaaS

BDaaS has the potential to be the technology disruptor that helps smaller organizations to have the data analytics capability that currently only the larger ones do. BDaaS will enable traditional cloud service providers to differentiate their services and prevent commoditization. Table 5.1 shows some of the impacts of BDaaS on the traditional scenario.

Traditional Environment	Traditional Big Data	Big Data as a Service
Lack of resources such as computational power and storage capacity.	Scalability in processing and storage achieved through distributed architecture	Scalability on demand through a combination of cloud computing and distributed architecture
Integrated hard data storage such as NAS, SAN, and traditional disks	Data storage on HDFS or distributed platform	Virtualized data storage on a distributed platform.
Structured data	Structured and unstructured data	Structured and unstructured data on cloud environment
Reporting using tools such as OLAP	Advanced analytics functions	Advanced analytics functions with on-demand computing power
Limited accessibility	Limited accessibility	Ubiquitous accessibility
Analytical capability derived through custom coding	Analytical capability derived through custom coding	Analytical capability derived from out-of-box domain-specific algorithms along with custom coding.

Table 5.1: Differences between BDaaS and Traditional Big Data

BDaaS has a tremendous capability to transform lives and processes in various sectors such as education, finance, insurance, research, and healthcare in emerging economies. These sectors may not have the capability to develop data analytics resources. In the case of education, scientific research, and politics the use of BDaaS is even more pronounced.

5.1 Politics

In 2008 the Democratic Party used Big Data to analyze the public sentiment, which helped it with favorable results in the election. It analyzed vast public data and engineered social, television, and other media outlets to create a targeted campaign to persuade young voters for the elections. The campaign proved effective in swing states, where Democrats won a resounding victory. The Big Data analysis also allowed Democrats to connect to campaign voters, which enabled them to generate over \$1 billion in revenue. During the initial phases of the campaign, the data analytics team realized that the various departments such as the campaign office, website department, and area departments were working from different sets of data. The data was not effectively shared to be truly effective in analyzing potential voters. The fundraising lists differed from the “get out the vote” lists, causing problems for the campaign office as well as voters. The analytics team helped to create a massive single system that could act as a central store for data. This data store enabled the Democrats to collect data from fieldworkers, fundraisers, and public consumer databases for analysis. This centralized store helped the campaign office to find voters and to create targeted campaigns to get their attention. Analytics on the vast datasets allowed the campaign office to find out what exactly appealed to the voter in a particular segment. It allowed campaigners to predict which voters were likely to give online. It enabled them to see who had canceled their subscriptions from their campaign lists, which indicated voters which may have switched to their political rivals. It also allowed them to evaluate things such as how the people would react to a local volunteer making a call as opposed to someone from a non-swing state. The data indicated that the people who had signed up for the quick donate program were four times more likely to give than others. This information enabled them to create a better donation system, where people could contribute with much less hassle, leading to increased funds.

The Obama campaign had a dedicated staff of over 100 people with over 50% of them in a special analytics department to analyze the data and 30% in the field to interpret the results. With regards to technology, they evaluated Hadoop for driving the Big Data analytics engine, but they were unable to do so as it required highly specialized skills to develop applications to interpret the Big Data. Another problem that they faced was that Hadoop, in its initial versions, was not designed to handle real-time queries. The team ultimately used Vertica, a large Big Data appliance that was scalable and easy to implement. Vertica is a column-oriented database that provides a standardized interface and SQL tools to access the data, hence existing tools and users can easily work with it without specialized skillsets. Hadoop is an open-source framework for Big Data which is complex to implement as compared to Vertica and requires specialized

skillsets. The central data repository for the campaign was created on Vertica, which enabled the analysts to get a 360-degree view.

The difficulties that the Democratic campaign faced with technology are very common with other political campaigns around the world. People may not have the resources to construct an analytical engine similar to that used by the Democratic campaign. All over the world, people may have data that they wish to use, not just to persuade voters, but also to identify problem areas for their respective constituencies. Hiring a large analytics team and developing a data computation facility is not feasible in most cases. When BDaaS is used, it enables people without specific Big Data skills such as Hadoop and IaaS to work with the data at hand. Specialized service providers working in the upper layers of the stack are able to provide domain-specific packages for managing campaigns. As the service is on-demand, the whole infrastructure can be ramped up as required.

For developing countries such as India, use of data in existing political campaigns is largely absent. India is heading for a general election in 2014 and it is important to see how the results of Big Data analytics help the political parties to reach out to their voters. A data-driven approach may help the political parties to make more informed decisions regarding the people's sentiments when they nominate their leaders. A large number of Indians are under 30 years of age and are technology savvy, with increasing exposure to social media. Collecting information from this demographic can be of immense importance to political parties to understand the most pressing issues of their constituencies.

BDaaS can help the regional and national parties to set up an infrastructure on demand when elections are being conducted and shut down when the elections are done. The services defined in the upper layers of the stack can enable the analytics to work with data, using some prebuilt tools for analysis. The political parties need not hire expensive resources, both hardware and technical, to set up the system.

5.2 Media and Entertainment

Established in 1851, the New York Times (NYT) is one of the oldest and the most popular news publication houses in the world. Its website attracts over 30 million unique visitors each month and it delivers news using traditional newspapers as well as over the Internet through its website and applications on smartphones. In an effort to modernize the news delivery mechanism and increase its web presence, the NYT had planned to digitize its news archives. One of the main initiatives was to launch the "timesmachine," which is a searchable content of news archives

from 1851 to 1922. NYT used Big Data technology and cloud computing services to deliver the historical archives on the web.

The project was to convert the existing historical archive into a data format that would be suitable for better delivery over the Internet. The existing archives were in an older file format, which was bulky and hence could not be used. The historical articles could have been converted dynamically, but pre-converting those articles would allow better scalability to the system. The “timesmachine” service had few potential subscribers when it was first launched and the subscriber base was expected to grow significantly. The dynamic conversion of documents takes significant computing resources as compared to batch processing the articles in bulk. If the subscriber base was to increase significantly, then it would result in high resource costs, due to expensive infrastructure, to have the dynamic conversion features. In addition the historical archives do not change; hence there was no need to have the dynamic capability on the system. As archive conversion was a one-time task, setting up an infrastructure for one-off use would have been a waste of resources. The historical archives with over 4GB of data were uploaded to an Amazon S3 system to act as source and target for the conversion process programs. The conversion programs were written in Hadoop, which was installed on multiple Amazon EC2 instances which provided the computation power. NYT used 100 Amazon EC2 instances for the conversion process which resulted in 11 million web-ready articles in just less than 24 hours. The process converted 4.3 terabytes of source data into 1.5 terabytes of output.

Although this project happened when most of the technologies in the BDaaS stack were still evolving, the BDaaS stack could have had helped the IT team at NYT to identify technologies that met their needs better more easily. In this case, there were two technologies, Amazon EC2 and S3, which lie in the data-storage and computing layer respectively. Using the BDaaS framework, technologies that span both layers, such as QuBole or Microsoft Azure, could have been used if they were available.

5.3 Retail

Small and medium businesses (SMBs) produce data although it may not be in the same range as other large retail business such as Wal-Mart. Although they may not have terabytes of data to be analyzed, it is still substantial enough that it needs dedicated resources to handle it. The main problem of SMBs is that they may not have the adequate infrastructure and the required human resources to analyze the data properly. Implementing infrastructure to analyze Big Data is expensive and mostly out of reach of SMBs. Smaller businesses can use cloud-based data and analytical services as defined in the upper layer of the BDaaS stack to address this situation. This can enable them to implement infrastructure quickly to handle their data requirement needs.

Also, the BDaaS service providers can have dedicated tools for particular industries, which can ease implementation and usage for SMBs. For example, a small chain of coffee shops can store data in a BDaaS system, where service providers have dedicated domain-specific applications to help users. Applications on the platform may be used to calculate the frequency of customers to the coffee shops. Another problem for implementing data analytics for SMBs is from a cultural standpoint. The ease of use provided by BDaaS systems can help SMBs to understand the value of data analytics, which can help in increasing its implementation.

BDaaS can also help in organizing data within SMBs in the form of a master repository of information. This will enable them to collect data from different units and prevent data silos. Currently most small businesses use a fragmented set of tools, which can prevent effective data analysis. Having a consistent platform, such as the platform provided by BDaaS, can help in data collection, storage, and analytics. It can also help SMBs to integrate sources of social media, which may help streamline the analytical process. The flexibility coupled with the ease of use of BDaaS may enable SMBs to create analytical algorithms to test their hypotheses rapidly. For example, a BDaaS service for a small retailer would have an inbuilt set of tools for calculating the time of day when maximum sales are likely to take place.

Chapter 6. Use of the BDaaS Framework

In this section some of the Big Data technologies are evaluated using the BDaaS stack. Various products in the market are analyzed with regards to their architecture followed by their classification in the BDaaS stack. This section is intended to illustrate how cloud-based Big Data technologies can be classified according to their function, which makes it easier for users to evaluate or compare them. For example, a research scientist at an educational institute has large datasets or Big Data of research data that needs to be analyzed. Due to resource constraints and time limitations, the research scientist may not have the inclination to set up a full-scale Hadoop infrastructure and then write MapReduce programs to analyze those datasets. Setting up a full-scale infrastructure is also wasteful, as the infrastructure is not required beyond the duration of the research project. Using the technologies in BDaaS, the researcher can identify technologies that are required such as those providing storage, computing, and analytical capabilities and just focus on the actual task, rather than worrying about the details or maintenance of the infrastructure setup to analyze the Big Data problem. This will result in better evaluation of Big Data technologies by focusing on only those in the appropriate layer. Users can refer to only those technologies in the BDaaS stack that meet their requirements rather than evaluating every popular technology in the market. For example, if the requirement is distributed storage with computing capability, then users can look at technologies that span the appropriate layers.

6.1 Amazon EC2

Introduction: Amazon EC2 is the principal Amazon cloud service and it was launched in 2006. EC2 is an IaaS offering that enables its users to access Amazon's cloud computing infrastructure over a cloud platform. Using an IaaS platform, EC2 users can instantiate virtual resources on the cloud computing platform and then build applications on top of it. EC2 enables users to have their own virtual computing machines, which are billed according to their usage and service plan type, such as spot, reserved, or on-demand. Amazon charges users depending on usage of resources in terms of capacity, data transfer, uptime, and processor usage for the virtual machines. Amazon uses Elastic Compute Unit (ECU) as an abstraction of computing resources. Each EC2 compute unit is calculated to be equivalent to a machine with a 1-1.2 GHz processor. The platform is scalable; hence users have the flexibility to manipulate the capabilities of their machines depending on their requirements. Amazon also enables users to choose the location of the virtual server resources that they may rent. The geographical location of the Amazon datacenters that host these services may have implications other than just the network issues. In some cases, hackers use EC2 machines located in other geographic zones to launch malicious attacks on organizations' infrastructures. In other cases, it might be that the user data that is

stored on the cloud may be subject to regulatory actions. For example, a government-affiliated organization in the United States may not be able to store data in an offshore datacenter location in Europe. European countries have stringent rules about data protection, especially when EU citizens' personal data is stored. American companies having their data center in Europe may be subject to such laws. According to the Patriot Act, American companies have to comply with data release requests if need be. Now if the data center is located in Europe, then it complicates the situation, as it might be subject to both American and European Union laws, which may conflict with each other.

The Amazon EC2 platform implementations are classified according to regions and divisions. The regions are isolated from one another with the intention of minimizing the network distance between the consumer and the services. Worldwide there are eight regions, including three in North America, three in Asia, and one each in Europe and South America. Communication between different regions is over the public Internet; hence users have to select an appropriate encryption method to protect their data. EC2 users can store the template for their virtual instances in an Amazon machine image (AMI). The AMI stores the software configuration of the virtual resource and allows users to make identical copies of the resources based on the AMI template. According to Huan Liu, co-founder of a technology startup 'Jamo', Amazon EC2 infrastructure is made up of approximately 450K servers as of 2012.¹⁹

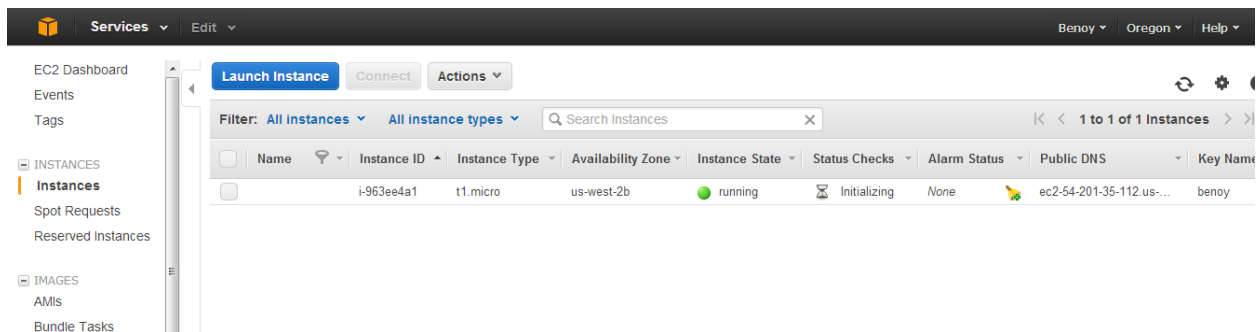


Figure 6.1: Amazon EC2 Console Used to Manage the Virtual Computing Machines

Architecture: Amazon EC2 uses a modified version of Xen bare metal hypervisors to run the virtual machines. Xen is a native hypervisor that allows guest virtual machines to run directly on top of server hardware. It is open-source software under GNU, which was started as a research project at the University of Cambridge. The first public release of Xen was in 2003 and it has been supported by other organizations such as Oracle, Amazon, and RedHat.

¹⁹ <http://huanliu.wordpress.com/2012/03/13/amazon-data-center-size/>

Amazon EC2 systems are among the largest Xen installations in the world. Every virtual machine on EC2 runs an instance of an operating system. The Xen hypervisor is used to schedule the virtual machines and assign them resources (Figure 6.1). In EC2, there could be multiple virtual machines running on a physical server (Figure 6.2). There are two types of storage: instance storage for non-persistent storage and elastic block store (EBS) for persistent network-based storage.

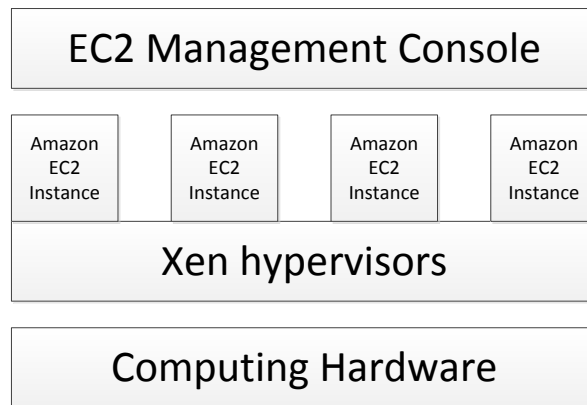


Figure 6.2: Amazon EC2 Architecture.

As Amazon's EC2 is primarily an IaaS system without any specialized Big Data capabilities, it can be positioned in the lower layers of the BDaaS stack (Figure 6.3). There are no inherent analytical and Big Data capabilities within the EC2 system and hence the users have to build those into it. As it is a basic infrastructure service, users have to create instances of virtual machines and then install distributed computing software such as HDFS to implement Big Data systems. Basic IaaS gives users flexibility to customize their systems, but it requires a specialized skillset to do so. To enable EC2 systems for Big Data, users have to create instances of virtual machines and then install the required architecture. For example, users can set up a distributed Big Data system such as Hadoop on an Amazon EC2 system by creating AMI templates for the servers and adding instances when necessary. So when the user wants to scale up an EC2-enabled Big Data system he or she can add preconfigured data nodes using AMI templates. EC2 provides the essential cloud capabilities upon which the Big Data systems can be built.

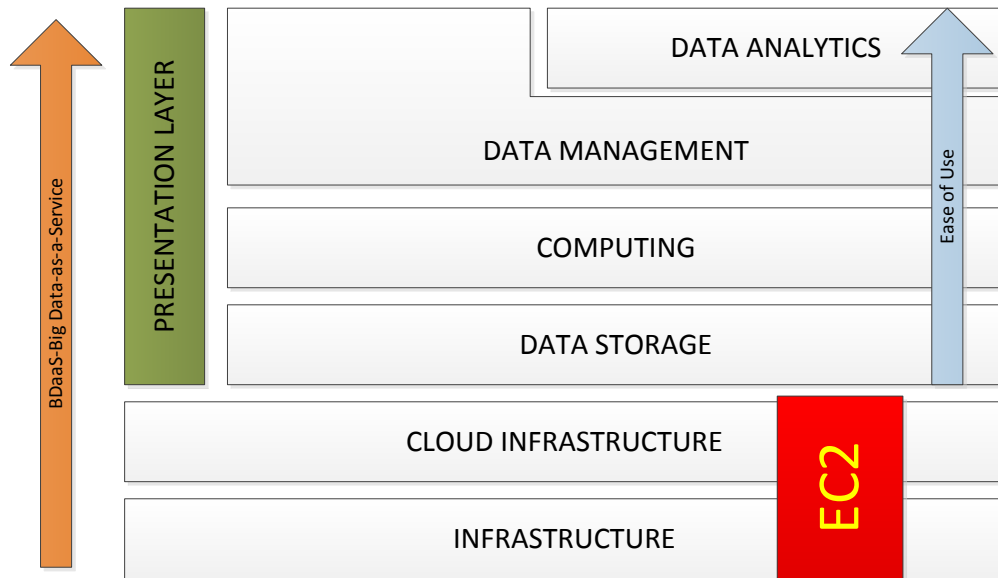


Figure 6.3: Amazon EC2 According to BDaaS

6.2 Amazon Dynamo

Introduction: Amazon Dynamo is a cloud-based NoSQL database service that enables users to store and retrieve large volumes of data. NoSQL databases have fewer constraints on the consistency models, such as lesser adherence to atomicity, consistency, isolation, and durability (ACID) properties than relational databases such as Oracle. ACID properties ensure that a data storage system can reliably manage and process transactions in the system reliably. Atomicity means each transaction is a single unit of work so it is basically “all or nothing.” If part of the transaction fails, then the entire transaction fails. Consistency means that transactions will keep the database in a valid state. The data written to the database should not violate any of the database rules or keep it in an inconsistent state. Isolation means that when multiple transactions are executed simultaneously on a system, the transactions are isolated from one another so that one transaction may not see the other. Durability refers to the ability to persist the transaction in permanent storage, even in the case of unexpected events such as system crashes. Typically NoSQL databases are optimized, key-value stores that are intended for fetching and appending data. NoSQL databases are usually considered a part of the Big Data technologies due to their capacity to handle large and unstructured datasets.

Architecture: Amazon’s DynamoDB is based on the Dynamo framework for key-value based distributed storage. The Amazon DynamoDB is different from other Amazon services, as it charges users based on throughput rather than storage. For example, Amazon enables users to request capacity for DynamoDB in terms of reads and writes and then charges a flat hourly rate accordingly. After the user specifies the desired capability, Amazon provisions the resources to

meet that capacity. To scale up, users can purchase additional throughput and the DynamoDB will automatically distribute the traffic over multiple servers. To create and instantiate a NoSQL table in DynamoDB, users have to specify the throughput capacity in terms of read capacity. Read capacity enables the user to perform a consistent read for 4kb of data per second. Users can add data to the table using a 'put' operation and a 'get' operation is used to perform queries on the table. The get operations only search on the hash key or the key identifier that has been specified on the table. All the other columns are combined into a single value store. Figure 6.4 shows the Amazon DynamoDB console, which shows the table metadata such as the table name and the provisioned read/write capability of the resource. Amazon uses this provisioned read/write capability of the resource to bill for the services.

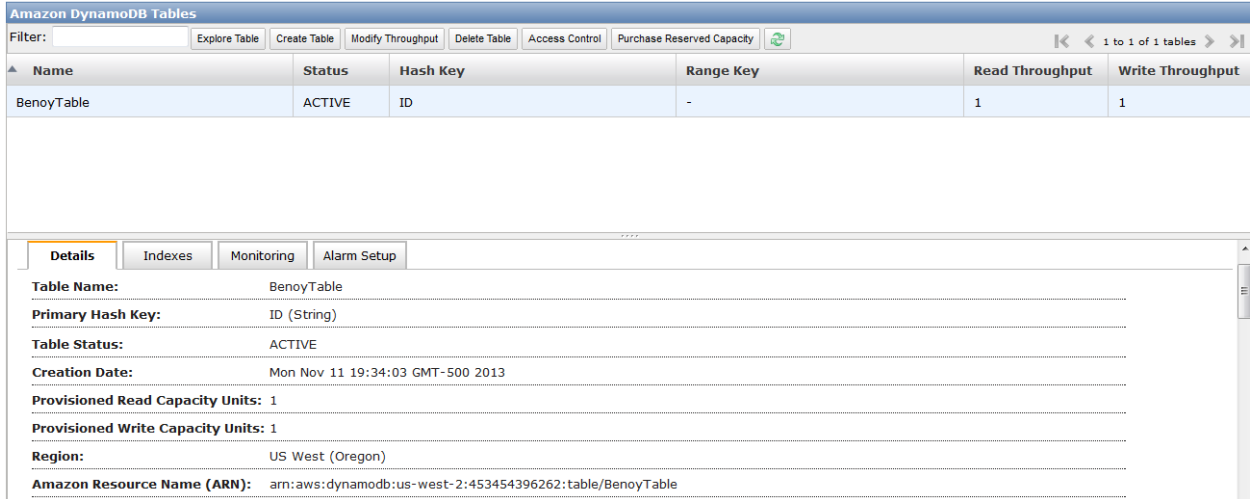


Figure 6.4: Amazon DynamoDB Console.

Dynamo DB is a scalable data storage device service from Amazon and hence it spans three layers of the BDaaS stack (Figure 6.5). DynamoDB abstracts the complexity of setting up a NoSQL data store and enables users to access the data through a presentation interface and scales on demand. Dynamo DB is similar to other key-value stores such as Cassandra or MongoDB, but is offered as a service by Amazon.

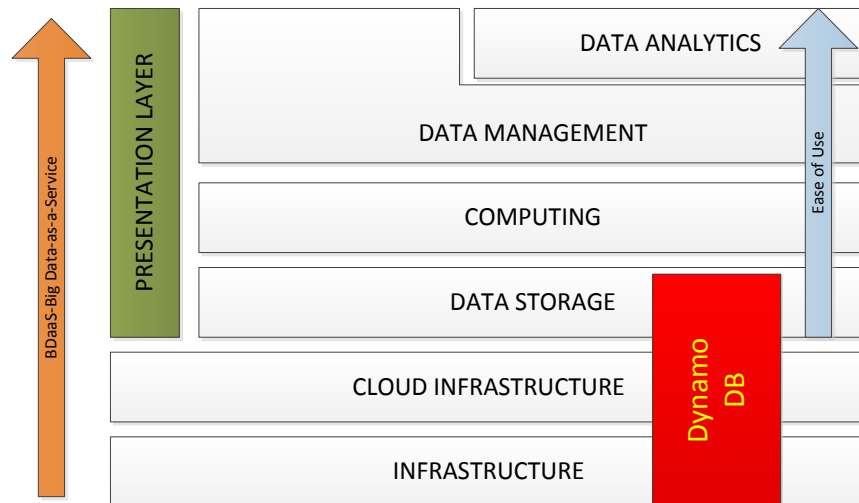


Figure 6.5: Amazon DynamoDB According to BDaaS

6.3 Amazon Elastic MapReduce

Introduction: Amazon EMR is a service that uses distributed processing framework such as Hadoop to process large data sets efficiently. EMR enables users to instantiate and use Hadoop infrastructure in a user-friendly way without having to know how to set up or manage the infrastructure. As the EMR uses a distributed framework, it quickly and effectively processes vast amounts of data across a scalable cluster of Amazon EC2 instances through horizontal scaling. The EMR is used in multiple areas such as log analysis, data warehousing, and machine learning. The instantiation of a Hadoop cluster and the management of MapReduce jobs, along with data transfer, are automated in EMR. Apache Users can add applications such as Hive to access the data processed by the EMR job flow. Hive is a data warehousing application that is built on the Hadoop framework and has an SQL like language, HiveQL, which enables users to perform data analysis. The Hive engine breaks down Hive statements into MapReduce jobs and executes them across the Hive cluster.

Architecture: The EMR uses an Amazon S3 service and customized Hadoop or MapR frameworks to conduct operations. Users have to upload the data they wish to process in an Amazon S3 bucket. Users can upload the data using APIs or using the Amazon Export/Import facility, where users physically ship hardware containing the data to Amazon. Users create buckets by specifying regions and these buckets are distributed to insure reliability. When data is stored in Amazon S3, data is stored on multiple devices in the region. Also there are additional data-protection mechanisms such as checksums and write protection to ensure consistency. For example, when a ‘write’ operation occurs, Amazon copies the data in multiple locations and only after it is done does it send a “success” message to the process performing data write.

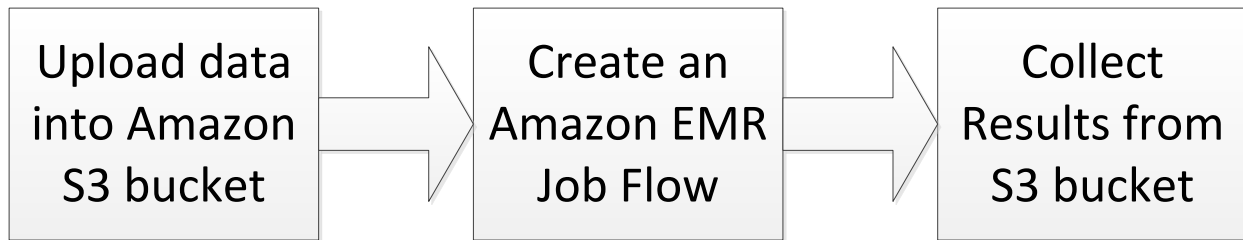


Figure 6.6: Amazon EMR Process Flow

Once the data is loaded in the source S3 bucket, EMR can be used to specify the MapReduce job that does the processing of data. For example, if data in a text file is to be aggregated, then the user can upload it to an Amazon S3 bucket and then execute a custom program on Amazon EMR, which performs the computation operation and stores the output back to another bucket (Figure 6.6). In Figure 6.7, we can see the input and output location for the EMR program. While specifying the parameters for the job flow, users can specify data inputs, outputs, and number of EC2 instances and create the job flow. The user can then run the MapReduce job and collect data in a target S3 bucket. The screenshot of the EMR console shows the user-generated program ‘wordSplitter.py’ which uses the MapReduce algorithm to process the data in the S3 buckets.

Create a New Job Flow Cancel X

DEFINE JOB FLOW SPECIFY PARAMETERS CONFIGURE EC2 INSTANCES ADVANCED OPTIONS BOOTSTRAP ACTIONS **REVIEW**

Please review the details of your job flow and click "Create Job Flow" when you are ready to launch your Hadoop Cluster.

Job Flow Name: My Job Flow
Type: Word Count (Streaming) [Edit Job Flow Definition](#)

Input Location: s3n://us-west-2.elasticmapreduce/samples/wordcount/input
Output Location: s3n://benoybucket.s3-website-us-west-2.amazonaws.com/wordcount/output/2013-11-14
Mapper: s3n://us-west-2.elasticmapreduce/samples/wordcount/wordSplitter.py
Reducer: aggregate
Extra Args: [Edit Job Flow Parameters](#)

Master Instance Type: m1.small **Instance Count:** 1
Core Instance Type: m1.small **Instance Count:** 2 [Edit EC2 Configs](#)

Amazon EC2 Key Pair:
Amazon Subnet Id:
Amazon S3 Log Path:
Enable Debugging: No **Keep Alive:** No
Termination Protected: No **Visible To All Users:** No [Edit Advanced Options](#)

Bootstrap Actions: No Bootstrap Actions created for this Job Flow [Edit Bootstrap Actions](#)

< Back **Create Job Flow**

Note: Once you click "Create Job Flow," instances will be launched and you will be charged accordingly.

Figure 6.7: Amazon EMR Console Indicating the Input and Output of the MapReduce Jobs

As the Amazon EMR uses S3 to store data and has the capability to process it using MapReduce frameworks such as MapR and Hadoop, it is mapped to the computing layer of the BDaaS stack. Amazon EMR has the option for users to use the Hadoop or MapR frameworks for data-processing jobs; hence it has the computing capability. MapR is a software vendor which provides proprietary software based on Hadoop. The MapR software is supposed to be more secure and have better performance than the standard Hadoop distributions from the open-source community. The Hive interface is optional and hence it does not have a true data-management capability, so the EMR spans from data infrastructure to the computing stage (Figure 6.8).

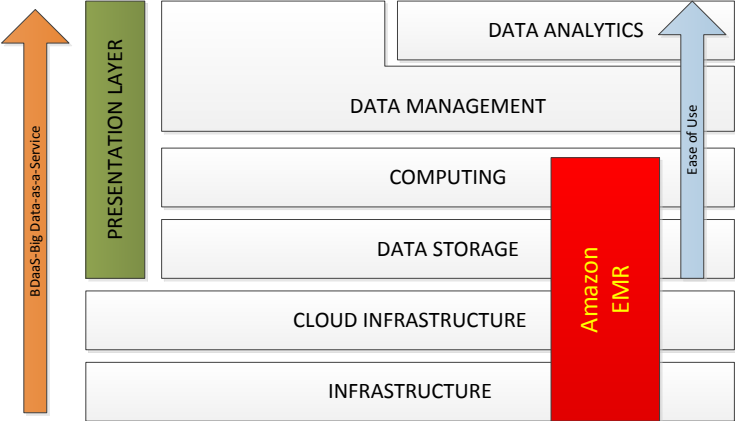


Figure 6.8: Amazon EMR According to BDaaS

6.4 Google BigQuery

Introduction: Google BigQuery enables users to have real-time access to massive data sets without having to set up expensive infrastructure. This enables the users to save on resources while empowering them with a user-friendly scalable cloud platform to query their datasets. Google BigQuery is accessible through a web-based or standalone UI to run queries. Google also provides a representational state transfer (REST) interface through which users can access the datasets and run operations. REST is an architectural framework for networked applications where users can use the simple hypertext transfer protocol (HTTP) to make service calls between machines rather than relying on complex protocols. For example, REST service calls can be made using a simple HTTP URL rather than using a complex XML document. As the platform is cloud based, users pay only for the resources they consume. The BigQuery is integrated with other Google platform products such as Google spreadsheets. Google BigQuery can be beneficial for organizations who would like to process large datasets efficiently without much cost overhead.

Architecture: Google BigQuery is based on Dremel, which is a scalable, ad-hoc query system to mine large datasets. Dremel uses multilevel execution trees and has columnar layout. Dremel is a

query service that users can use to run SQL-like queries to execute aggregation queries on datasets of over a trillion rows instantaneously. Dremel and MapReduce are complementary technologies that are not substitutes of each other. Dremel is an interactive query system that uses multi-level tree structure execution and column-oriented structures to perform ad-hoc query analysis.

Dremel is designed for data analysis, it is used to run queries quickly on massive data sets, and it supports SQL-like syntax. The data is organized in columnar format, which enables faster querying of datasets. Google BigQuery is the public implementation of Dremel, which is accessible primarily through RESTful API. Dremel and Google BigQuery are different from Hadoop Hive, which turns SQL queries into MapReduce operations. Also, Hive uses internal table indexes to improve the query performance in conjunction with MapReduce, instead of using the columnar format of Dremel. Hadoop, in conjunction with the Hive data warehouse software, also allows data analysis for massive datasets using a SQL-style syntax. Hive essentially turns queries into MapReduce functions and executes them. The HiveQL queries are compiled into MapReduce jobs, which are then executed on Hadoop. The data represented through Hive is arranged in traditional structures such as tables, rows, and columns.

When dealing with Big Data, if there is a single query to the data then MapReduce jobs would be sufficient. In most cases, data analysis requires results to be delivered in a format similar to the output from an OLAP (Online Analytical Processing Systems) for relational databases. Dremel, which is based on similar concepts such as columnar storage and query execution engines, supplements the traditional MapReduce jobs and helps facilitate better response for data analytics. The data consumed by Google Big Query has to be loaded in Google cloud storage for consumption. The screenshot in Figure 6.9 shows the execution speed of a query to count the number of distinct articles in dataset over 6.5GB. Google BigQuery uses Google's cloud storage service to store the data.

New Query

```
1 select count(distinct(title)) from publicdata:samples.wikipedia desc limit 10
```

RUN QUERY Save Query Enable Options Query complete (2.2s elapsed, 6.79 GB processed)

Query Results 2:38pm, 14 Nov 2013

Row	f0_
1	19996639

Figure 6.9: Running an Aggregate Operation on a Large Dataset

As Google BigQuery offers data computation and analytics, it spans across the computing and data management stack. The data storage for Google BigQuery is handled by Google Cloud storage, which is a similar concept to Amazon S3, where both the services provide mass data storage on cloud platforms. Google Big Query does not have extensive inbuilt presentation and analytics capabilities; hence it does not cross the data analytics layer unlike Amazon AWS. Users need to have knowledge of RESTful APIs and SQL queries to manage and analyze the data. Figure 6.10 shows the placement of BigQuery on the BDaaS stack.

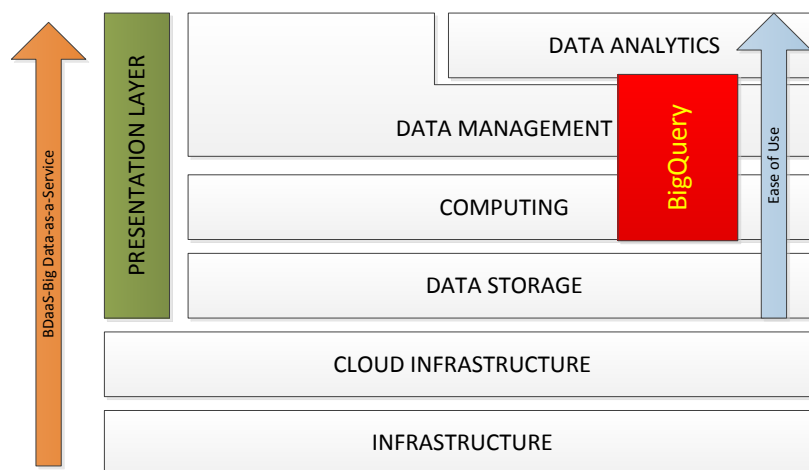


Figure 6.10: Google BigQuery According to BDaaS

6.5 Splunk Storm

Introduction: Splunk Storm is a data analytics platform for monitoring and analyzing machine-generated log data. It enables users to create a data repository which is then indexed to facilitate

the generation of graphs, dashboards, and other visual interfaces providing real-time information. Splunk identifies data patterns and helps disseminate information within the organization using a web-based interface. The software is available in both standalone installations and cloud-based platforms. Splunk cloud enables users to run customized queries and create dashboards from the existing datasets, where users upload the data to the cloud system. As with all other cloud-based systems, Splunk cloud offers scalability, ease of implementation, and pay-per-use features of the cloud platform. As Figure 6.11 shows, users have a web interface to perform analysis such as searching for strings or creating graphs. The console shows the source data and has an input field to perform string searches. This gives users the ability to extract information from log files without writing customized programs to do so. For example, if the infrastructure engineers want to search for an error in the log files that is identified by a particular identifier, then using this interactive dashboard, they can easily do so. The interactive console allows users to configure data sources as well as creating visual reports for problem investigations.

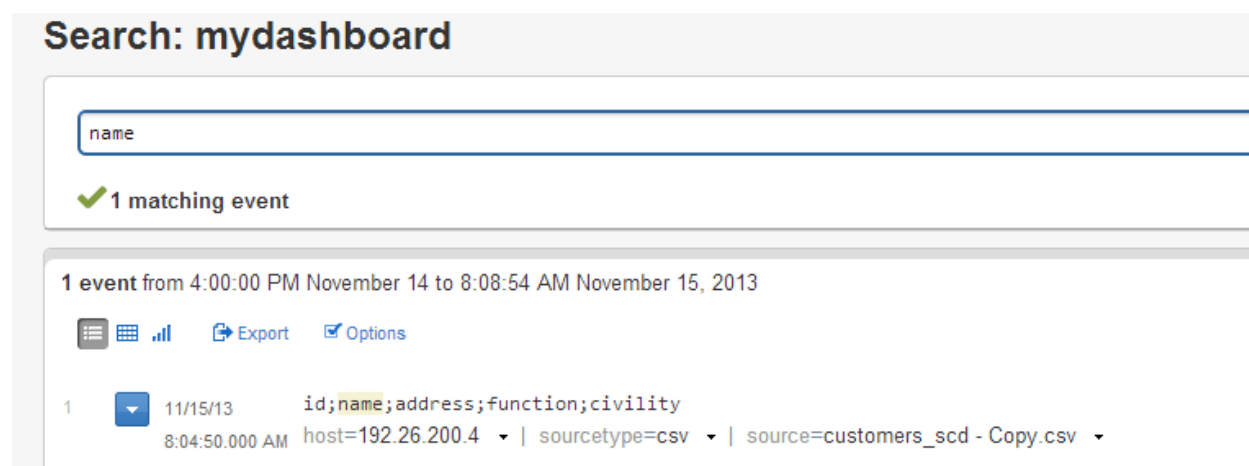


Figure 6.11: Splunk Interactive Console

Architecture: Splunk uses a propriety-specific database called 'index' to store the user data. This data is internally classified into two types of files: raw data and the indexes that point to the raw data that is stored on the cloud platform. The Splunk enterprise cloud is hosted on an Amazon AWS platform and runs on 64-bit Linux machines. Splunk consists of two major components. The 'splunkd' is a distributed server engine that processes and indexes incoming data and handles search requests. The data is processed using data pipelines, which are single threads inside the splunkd process. 'Processors' are functions that act on the stream of data passing through the pipeline. The other major component of the Splunk installation is the 'splunkweb,' which is a web-server that provides a web interface for users to access the system.

Splunk offers computing and analytics, but the underlying cloud infrastructure uses Amazon AWS services (Figure 6.12). The data analytics layer is specialized for managing and processing log files. The pricing is determined by the amount of data that the user indexes into the Splunk system for analytics per day. There are no extra charges for the number of data sources or users. As the software stores analyzed data as indexes and has interactive consoles for accessing information, it spans from data storage to the analytics layer. Liverpool Football Club uses Splunk storm service to maintain its IT infrastructure. The club has a website that gives access to the latest news and shopping as well as mobile games. The website is also acts like a community, where fans can discuss events and connect with each other. This website faces peak demand during games and it is essential to monitor the infrastructure to insure continuous operation.

The club uses Splunk storm to increase visibility across infrastructure and to optimize content delivery. The infrastructure team at the club sends logs and other machine-related data to the cloud-based Splunk storm, which enables report generation on recent activities. It also gives visibility to gauge the infrastructure capacity, which helps to determine if the underlying website capacity is sufficient to handle the demand.

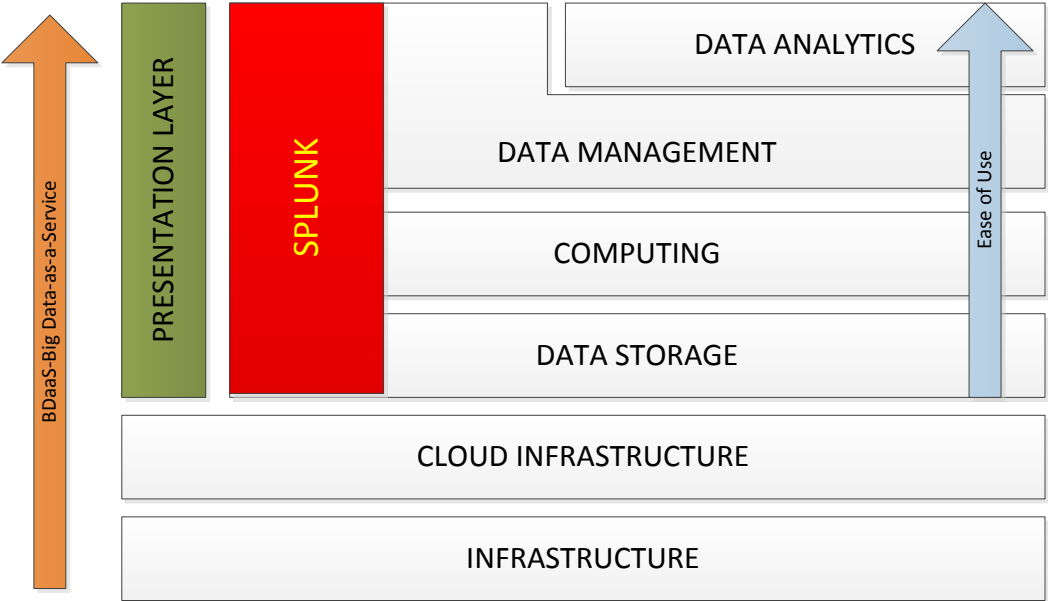


Figure 6.12: Splunk Storm According to BDaaS

6.6 Microsoft Azure-HDInsight

Introduction: The HDInsight is a Microsoft BDaaS solution that allows instantiation of Hadoop clusters in a cloud environment. HDInsight provides Apache Hadoop as a service, which enables end users to have a more scalable and cost-efficient environment. HDInsight is the flagship Microsoft solution for cloud-based Big Data analytics.

Architecture: The HDInsight combines the Microsoft cloud platform with the distributed processing architecture of the Hadoop framework to provide service consumers an on-demand infrastructure to handle large data sets. Microsoft cloud platform is a cloud computing service similar to Amazon AWS. HDInsight is a much newer service launched by Microsoft and it is similar to Amazon's. Amazon cloud services providing Big Data capability are better integrated into each other than HDInsight. The pricing model for Azure is similar to that of Amazon, but it offers a free tier for business to try out services for a year before being charged. Users also have open database connectivity (ODBC) drivers, which enable the users to integrate business intelligence and integration tools to access and analyze data. ODBC is a middleware programming interface that allows users to connect to any type of database independent of platform. Because of the ODBC driver, Microsoft platform users can connect to the Azure Hadoop. It integrates easily with other Microsoft products such as Excel and power BI. Without this ODBC driver, users would have to write programs to perform this connectivity. ODBC drivers also allow users application portability. For example, if the analytics team is using Excel for generating reports, then there is an ODBC driver that allows direct connectivity to export the results. If the user wants to change the target data source to something other than Excel, then the same ODBC driver can be used with the new source. The Azure HDInsight uses Windows Azure Blob storage to store the underlying data within the system. The HDInsight also allows users to store metadata of the table definitions in an SQL server. This allows users to retain data models and other information when the cluster is shut down. As the HDInsight has basic computing storage service over an IaaS cloud infrastructure, it spans from the computing to the data infrastructure stage (Figure 6.13). It provides users with both storage and computing services, which are hosted on Microsoft datacenters.

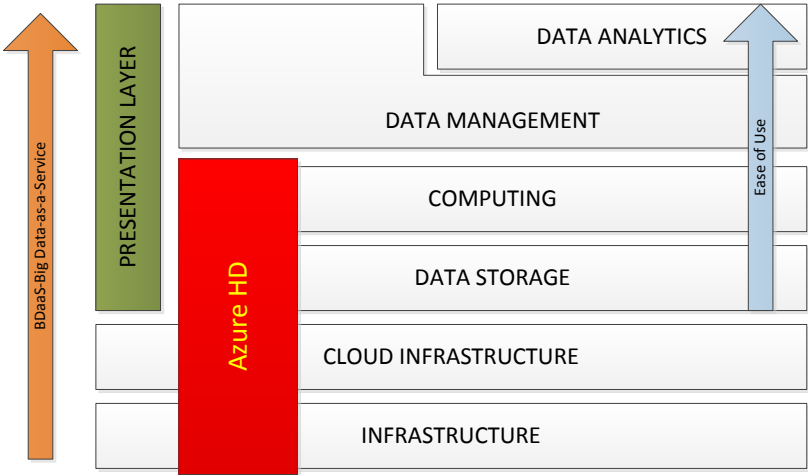


Figure 6.13: Azure HDInsight According to BDaaS

6.7 Tibco Silver Spotfire

Introduction: Tibco Spotfire, a data analytics platform, enables users to perform complex analytics tasks and generate interactive reports. Users can develop dynamic analytical reports that can be accessed through a web console from the data analyzed by the Spotfire platform. Spotfire can be integrated with other Tibco products such as Master Data Management (MDM) and Rendezvous. The MDM software allows users to collect data from heterogeneous sources for data de-duplication and acts as a central information store. The Tibco Rendezvous is an ‘enterprise application interface,’ which provides a software message bus that various applications can use to communicate and transfer data to each other.

Architecture: The application is composed of multiple interconnected components that enable clustered deployment. In a standardized Spotfire installation, all the components are installed on user infrastructure. In the cloud-based Silver Spotfire, users have a local copy of the Spotfire client with which they can create visualizations and upload the data to the Tibco server. There are size limitations to the amount of data uploaded. After the analytics results are uploaded to the Tibco servers, they are not connected to the underlying data from which they have been derived. So if the underlying data changes, then the results have to be uploaded again.

As the Silver Spotfire is a cloud-based analytical platform, it spans from analytics to the data-management layer on the BDaaS stack (Figure 6.14). As the raw data which is analyzed still resides on the user’s machine and only the results of the analytics are uploaded, the application does not span over the storage layer. As the data analytics is done by the Spotfire client, which is not based on a cloud platform, it does not span the computing layer. As compared to the other pure Big Data systems such as Amazon EMR or Google Big query, it has a well-developed user interface in the presentation layer, with which the users can interact directly.

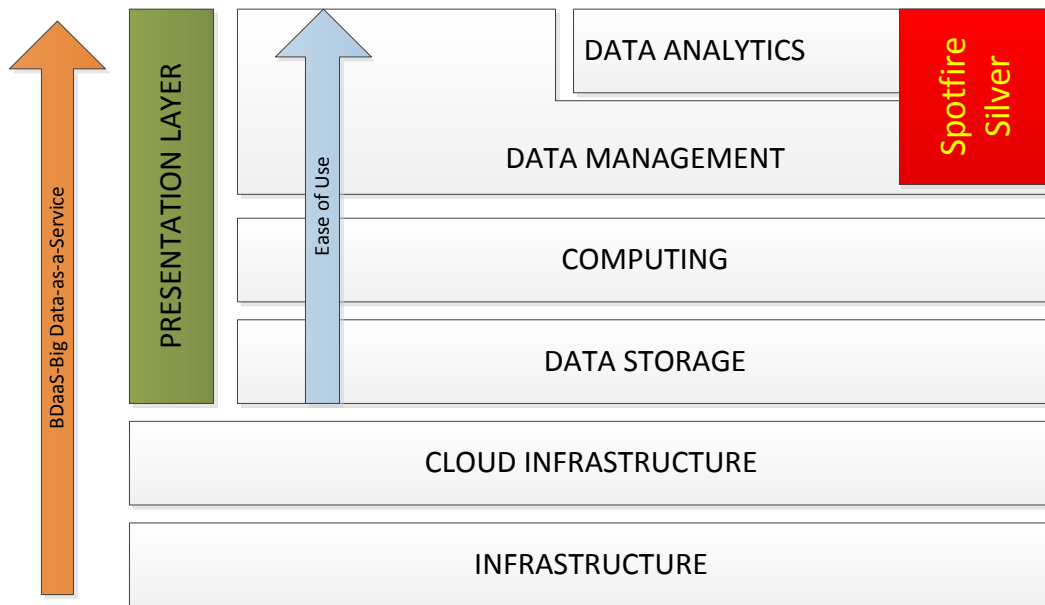


Figure 6.14: Tibco SpotFire Silver According to BDaaS

6.8 QuBole

Introduction: QuBole is a BDaaS company that provides a Hadoop framework on a cloud-based platform. It has features such as auto scaling as well as GUIs for managing the Big Data implementation on a cloud platform. It abstracts the complexity of managing Hadoop clusters and provides real-time scalability. It allows users to connect to data sets stored on a cloud platform such as Amazon S2.

Architecture: The platform has an inbuilt Hadoop engine that uses daemons and optimizes resource allocations for faster execution. Users can upload data directly from their Amazon S3 buckets and create tables in Hive using the GUI interface. The Hadoop clusters are instantiated on customers' AWS accounts. When customers create accounts on QUBole, they have to register their AWS accounts, which the QUBole service uses to instantiate the clusters. Users can use both AWS reserved and spot instances while working with QUBole.

The results of the computation can also be directed back to user-specified data storage. QuBole enables users to reduce costs by instantiating nodes only when jobs are started and the platform automatically scales to meet the demand. The platform does not require changes to the user infrastructure, but allows users to work on their data on any platform, such as cloud or local systems. For example, if users have their data stored on Amazon S3 systems, then they can use the QuBole service to instantiate Hadoop clusters and use those S3 systems as the source and target. Figure 6.15 shows that users can create a table by specifying the source location as an Amazon S3 bucket. The command creates a new table with the name 'benoytable' with the field

specifications in terms of column names and their data-types. The attributes at the end of the query specify the format of the source in the S3 bucket, in this case its 'row format.'

```
CREATE EXTERNAL TABLE benoytable (`site` STRING, `ts` STRING, `phr` STRING, `lnks` ST  
RING) PARTITIONED BY (`monthly` STRING) ROW FORMAT SERDE 'org.apache.hadoop.hive.cont  
rib.serde2.JsonSerde' LOCATION 's3n://paid-qubole/default-datasets/memetracker/';
```

Figure 6.15: Creating a Table Using Hive and S3 Buckets on QuBole

Figure 6.16 shows the web-based query interface with which users can query data for analytics. Users can use this interface to write queries in Hive on the tables indicated on the left side of the screen. The table explorer in the screenshot indicates the tables as well as their structures.

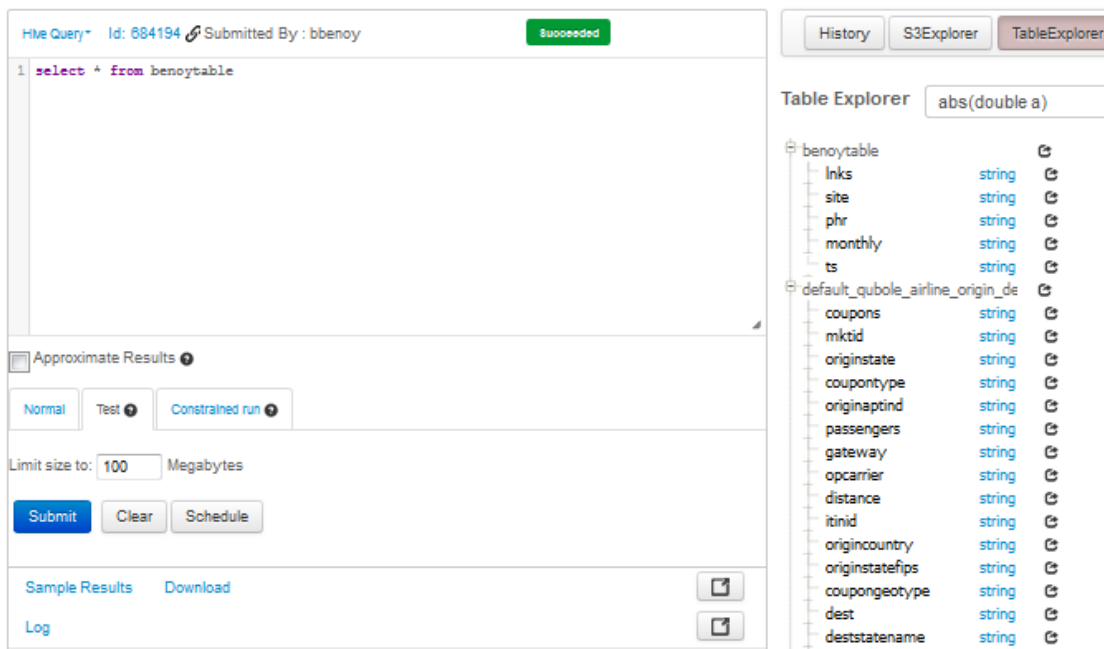


Figure 6.16: QuBole Hive Interface and Query Processing

As the QuBole platform primarily offers data-processing services using the Apache product, it spans from the computing to the data management layer (Figure 6.17). The platform has a presentation layer in the form of a GUI with which users can interact with the data, but it lacks advanced statistical and analytics functions for the data analytics layer. As QuBole primarily uses Amazon S3 for data storage, it does not span over the data storage layer. As it is not a pure analytics platform like Spotfire or Splunk, it does not include the data analytics layer.

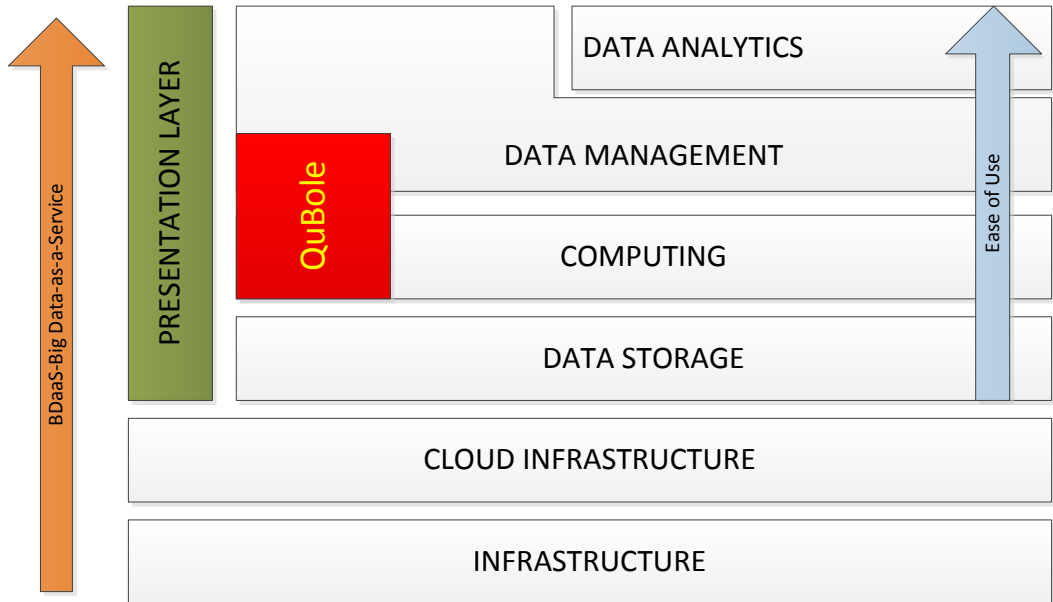


Figure 6.17: QuBole Foresight According to BDaaS

Chapter 7. Conclusion

The concept of Big Data is being increasingly well defined and transforming from an idea to a well-defined concept with real-world implementations. New innovations in Big Data technologies are helping increase the adoption rate across various industries. Newer and more capable BDaaS technologies appear in the market every day, which add to the BDaaS stack. Over time, due to awareness and simplification of interfaces, these sophisticated technologies in this BDaaS service stack will become easier to use, but they will require coordination amongst business teams and engineers to be truly effective.

7.1 Future Outlook

Cloud computing is increasingly seen by organizations as a viable way to reduce costs and increase implementation efficiency. Vendors are increasingly launching cloud-enabled versions of their transitional Big Data technologies. This trend indicates that most of the Big Data technologies that we see in the industry today will be cloud-enabled in the upcoming years. Gartner also predicts that by 2016, almost 80% of organizations will use a maximum of two cloud service providers (David 2013). Consolidation of cloud service providers indicates that technologies implemented will be increasingly capable, perhaps spanning most of the layers of the BDaaS stack. As such technologies emerge, they will increasingly encompass functions of the higher layers, effectively providing an “analytics package” over the cloud.

Big Data will not exist as a standalone market, but will be a composite market along with traditional IT markets. In the additive phase, new technologies emerge that are implemented by highly skilled IT services professionals. As of 2013, skills for Big Data are highly sought after and demand a high compensation from the market, indicating the current market phase for Big Data (Information Week 2012). According to Gartner, Big Data is in the additive phase of composite markets, where demand-driven spending is higher (Beyer et al. 2012). The proportion of spending on software products for IT services is significantly more in the additive phase of composite markets than in others.

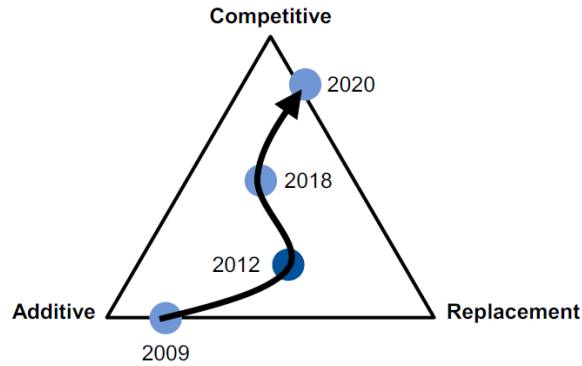


Figure 7.1: Current State of the Big Data Market (Beyer et al. 2012)

Figure 7.1 indicates the current state of the market for Big Data products, according to Gartner research (Beyer et al. 2012). In the additive phase, a solution is considered a new purchase and is typically used in beta projects. Currently organizations continue to purchase traditional technologies as the market is approaching replacement stage, but they are actively evaluating Big Data technologies for future needs. In the competitive phase, organizations are considering the new technology as an alternative. Upgrade and replacement decisions are made by comparing the new technologies with old ones. As more mature Big Data products are introduced, the market is expected to go into the competitive phase, where there will be multitudes of vendors offering similar products. The Big Data market will enter a competitive phase in 2020, where the revenue from IT services due to implementation will reduce in proportion to software. As more Big Data technologies are introduced, consumers can use the BDaaS service stack for their technology selection.

According to Gartner, revenue from Big Data enterprise software will increase to \$6.5 billion in 2016 from \$3 billion in 2013(Thomas 2012). IT spending driven by Big Data projects will reach \$55 billion. Big Data is driving functional demands across all organizations and was expected to reach \$28 billion of total spending in 2012 with \$4.3 billion in software sales for Big Data products (Beyer et al. 2012). As the market develops for Big Data technologies, organizations will have them embedded into their current platforms (Beyer et al. 2012). Mainstream technologies such as cloud computing will be merged with Big Data technologies. Organizations that do not address the Big Data wave early may find themselves at a strategic disadvantage.

7.2 Role of BDaaS Framework

Big Data is a major challenge which organizations are trying to overcome by leveraging the power of cloud computing. As BDaaS technologies are delivered over the cloud computing platform, they address the challenges of the traditional Big Data technologies. Use of BDaaS

technologies can help organizations in accelerating their Big Data implementations, as organizations need not have their own infrastructure and can use that of the service provider. This also reduces the resource requirements for pilot projects that are launched to evaluate and determine the value of Big Data technologies to the respective organizations.

Using technologies in the BDaaS stack, organizations can quickly ramp up infrastructure to evaluate different technologies without investing much in their infrastructure. Currently there is no specific organizing work that would define a framework of BDaaS. This classification would enable faster and more effective evaluation of technologies, leading to reduced costs. For service providers, the BDaaS framework will help them to communicate their products effectively to consumers, leading to increased sales. It can also help them identify their competitors in their product space. So using the BDaaS framework, both producers and consumers of Big Data technology products would derive benefit.

Bibliography

- Beyer, Mark, and Friedman Ted. 2013. "Big Data adoption in Logical Data Warehouse." Gartner G00248456.
- Beyer, Mark, Lovelock John, Sommer, Dan , and Merv Adrian. 2012. "Big Data Drives Rapid Changes in Infrastructure and \$232 Billion in IT Spending Through 2016." Gartner G00245237.
- Davenport, Thomas. 2012. "Enterprise Analytics Optimize Performance, Process and Decisions through Big Data." *FT Press*: 30-45.
- Dean, Jeffery, and Ghemawat Sanjay. 2004. "MapReduce: Simplified Data Processing on Large Clusters." Google.
<http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf>
- Diebold, Francis. 2012. Diebold, Francis X., On the Origin(s) and Development of the Term 'Big Data' ,PIER Working Paper No. 12-037:1-5
- Edala, Seshu. 2012. "Big Data Analytics: Not Just for Big Business Anymore." *Forbes*.
<http://www.forbes.com/sites/ciocentral/2012/12/28/big-data-analytics-not-just-for-big-business-anymore/>
- E. F. Codd. 1970. A relational model of data for large shared data banks. *Communications of ACM* 13, 6 (June 1970), 377-387
- Ellsworth, David, and Michael Cox. 1997. "Application Controlled Demand Paging for Out-of-core Visualization." *NASA Ames Research Center ,Report NAS-97-010*: 1-5.
- EMC Solutions Group. 2012. "Big Data as a Service-Market and Technology Perspective." <http://www.emc.com/collateral/software/white-papers/h10839-big-data-as-a-service-perspt.pdf>
- Heudecker, Nick. 2013. "Hype Cycle for Big Data." Gartner G00252431
- Hopkins, Brian, and Boris Evelson. 2012. "Expand Your Digital Horizon With Big Data." Forrester.
- Information Week. 2012. "Big Data Widens Analytic Talent Gap." *Information Week* April.

- KPMG International. 2013. "Breaking through the cloud adopting barriers: KPMG Cloud Providers Survey."
- Kart Lisa, Heudecker Nick, Buytendijk Frank, 2013 "Survey Analysis: Big Data Adoption in 2013 Shows Substance Behind the Hype" Gartner G00255160
- Laney, Doug. 2012. "3D Data Management: Controlling Data Volume, Velocity and Variety." <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
- Lyman, Peter, and Hal Varian. 2000. "How Much Information." University of California at Berkeley: 10-34.
- Maoz, Michael. 2013. "How IT Should Deepen Big Data Analysis to Support Customer-Centricity." Gartner G00248980
- Mell, Peter, and Grance Timothy. 2011. "The NIST Definition of Cloud Computing." NIST Special Publication 800-145. <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>.
- Price, Derek. 1961. "VIS 97 Application-controlled demand paging for out-of-core visualization." Proceedings of the 8th conference on Visualization '97, IEEE Computer Society Press
- Rider, Fremont. 1944. "The Scholar and Future of Research Library." *Hadham Press* 297-325.
- Smith, David. 2013. "Hype Cycle for Cloud Computing, 2013." Gartner G00252159: 5-35