NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS
DEPARTMENT OF ECONOMICS

THESIS

Big Data

Georgios K. Gaitanis

Supervisor: Dr. **Katsianis Dimitrios**

ATHENS

OCTOBER 2017

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μεγάλα Δεδομένα

Γεώργιος Κ. Γαϊτάνης

Επιβλέπων: Δρ. **Δημήτριος Κατσιάνης**

ΑΘΗΝΑ

ΟΚΤΩΒΡΙΟΣ 2017

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μεγάλα Δεδομένα

Γεώργιος Κ. Γαϊτάνης
Α.Μ.: ΜΟΠ479

ΕΠΙΒΛΕΠΟΝΤΕΣ: Δρ. Δημήτριος Κατσιάνης

ΠΕΡΙΛΗΨΗ

Στη σημερινή εποχή, ο όρος "μεγάλα δεδομένα" προσελκύει ιδιαίτερα μεγάλη προσοχή, τόσο από την επιχειρηματική όσο και την προσωπική οπτική γωνία και προοπτική. Για δεκαετίες, οι επιχειρήσεις λαμβάνουν επιχειρηματικές αποφάσεις μέσω του τμήματος Business Intelligence, οι οποίες στηρίζονται στα δεδομένα συναλλαγών που αποθηκεύονταν στις γνώριμες σχεσιακές βάσεις δεδομένων. Ωστόσο, η κανονιστική συμμόρφωση, ο αυξημένος ανταγωνισμός και διάφορες άλλες "πιέσεις" έχουν δημιουργήσει στις εταιρείες μια άνευ προηγουμένου ανάγκη να συσσωρεύουν και να αναλύουν ταχέως, μεγάλες ποσότητες δεδομένων τα οποία ξεπερνούσαν τα σημαντικά εσωτερικά δεδομένα. Αυτά τα αδόμητα ή ημι-δομημένα δεδομένα, είναι ένας πιθανός θησαυρός μη παραδοσιακών δεδομένων για τις εταιρείες. Λιγότερο δομημένα δεδομένα: weblogs, κοινωνικά μέσα, ηλεκτρονικό ταχυδρομείο, αισθητήρες και φωτογραφίες μπορούν επίσης να χρησιμοποιηθούν για την εξόρυξη χρήσιμων πληροφοριών. Οι νέες τεχνολογίες του cloud έχουν συμβάλει τα μέγιστα στη μείωση του κόστους τόσο της αποθήκευσης όσο και της υπολογιστικής ισχύος, καθιστώντας εφικτή τη συλλογή αυτών των δεδομένων - τα οποία θα είχαν μάλλον διαγραφεί πριν από λίγα χρόνια. Τα παραπάνω έχουν σημαντικό αντίκτυπο στην αγορά, καθώς όλο και περισσότερες εταιρείες προσπαθούν να συμπεριλάβουν αυτά τα μη "παραδοσιακά" αλλά δυνητικά πολύτιμα δεδομένα.

Με αυτό τον τρόπο, οι εταιρείες εξελίσσουν το παραδοσιακό BI καθώς χρησιμοποιούν δεδομένα που παλαιότερα θεωρούνταν μη χρήσιμα.
Η επιτυχής λήψη αποφάσεων θα οδηγείται ολοένα και περισσότερο από τα συμπεράσματα των analytics που προκύπτουν από τους αναλυτές. Από τα πιο χαμηλά επίπεδα μιας εταιρείας, εξυπηρετητές δικτύου, έως τα υψηλού επιπέδου συστήματα υποστήριξης επιχειρήσεων (Τηλεπικοινωνιακή προοπτική της υπηρεσίας) αλλά και από τις εφαρμογές του ιστού που οδηγούν στα εξατομικευμένα δεδομένα (Προοπτική προσώπων). Σε όλα αυτά τα επίπεδα, θα υπάρχει η ευκαιρία να αξιοποιηθούν οι πληροφορίες που συγκεντρώθηκαν από την προσεκτική ανάλυση όλων των σχετικών δεδομένων (Αρχεία καταγραφής, Call detailed records (CDR), αναζήτηση Google, δεδομένα Facebook κλπ).

Τα μεγάλα δεδομένα αποτελούν τον πυρήνα αυτής της μεγάλης ευκαιρίας για το Business Intelligence και απαρχαιωμένη προοπτική που προσφέρουν οι μέθοδοι του παραδοσιακού BI. Οι εταιρείες οφείλουν να επικεντρώσουν τις προσπάθειες τους στις καινούριες τεχνολογίες που προσφέρουν τα μεγάλα δεδομένα. Υπάρχει μεγάλη δυνατότητα για αύξηση της δυναμικής τους και να δημιουργήσουν καινούργια αξία στην υπάρχουσα αγορά.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μεγάλα Δεδομένα

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Μεγάλα δεδομένα, ταχύτητα, όγκος, δομημένα, ημι-δομημένα, μη δομημένα, ανάλυση δεδομένων.

ABSTRACT

Nowadays, term big data, draws a lot of attention, both for Business and person perspective. For decades, companies have been making business decisions through its Business Intelligence department, based on transactional data which were basically stored in relational databases. However, regulatory compliance, increased competition, and other pressures have created an insatiable need for companies to accumulate and analyze large, fast-growing quantities of data that was beyond the critical data. Those unstructured – semi structured data, is a potential treasure trove of non-traditional, less structured data: weblogs, social media, email, sensors, and photographs that can be mined for useful information. Decreases in the cost of both storage and compute power, using the new cloud and Saas services, have made it feasible to collect this data - which would have been thrown away only a few years ago. As a result, more and more companies are looking to include non-traditional yet potentially very valuable data with their traditional enterprise data in their business intelligence analysis.

On contrast with today, successful decision-making will increasingly be driven by analytics-generated insights. From the lowest-level network enablers to high-level business support systems (Telco service perspective), and from web applications to personalized data (Person perspective), there will be an opportunity to utilize insights gathered from careful analysis of all relevant data (Network logs, CDRs, Google search, Facebook data etc).

Big data is at the core of this opportunity for Business Intelligence and the traditional view of BI methods are not enough. We should focus on what technology big data brings and we have to look at what value it can create.

SUBJECT AREA: Big Data

KEYWORDS: Big Data, velocity, volume, structured, semi-structured, unstructured analytics, Teradata, hadoop, oracle, Security, SIEM

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

People and devices are constantly generating data. While streaming a video, playing the latest game with friends, or making in-app purchases, user activity generates data about their needs and preferences, as well as the quality of their experiences. Even when they put their devices in their pockets, the network is generating location and other data that keeps services running and ready to use.

In 2014, estimates put worldwide data generation at a staggering 7ZB [1], and by 2018 each smartphone is expected to generate 2GB of data every month [2]. At the same time, the big data technology and services market is expected to grow at a 40 percent compound annual growth rate (CAGR) – about seven times the rate of the overall ICT market – with revenues expected to reach USD 16.9 billion in 2015 [3]. Clearly, the age of big data has begun

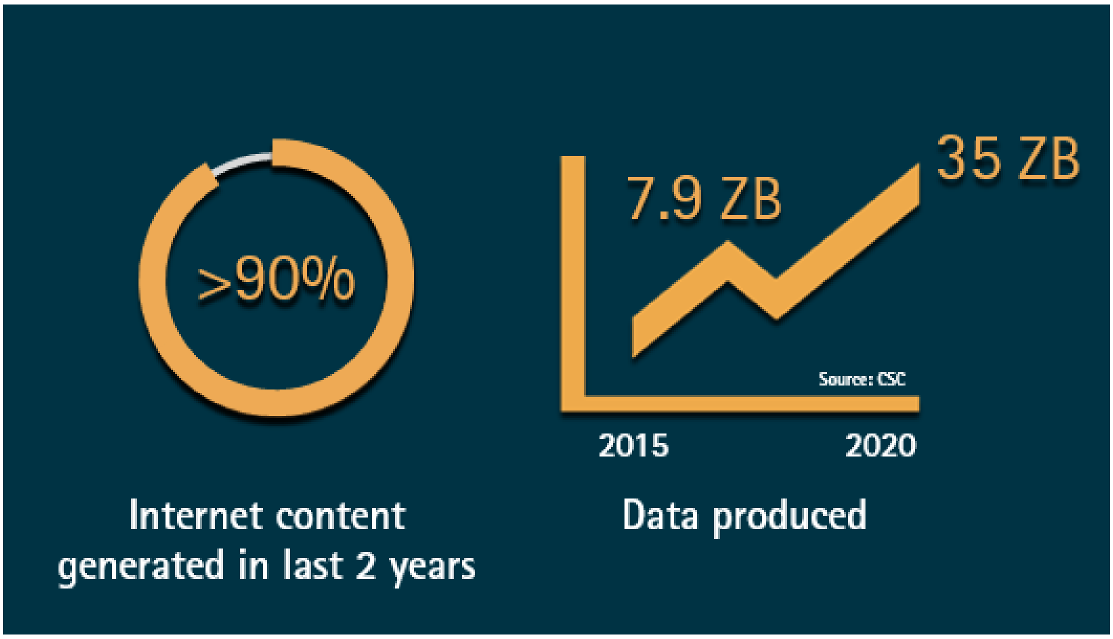


Figure 1: Data Depiction

Communication service providers (CSPs) can make use of this big data to drive a wide range of important decisions and activities, such as: designing more competitive offers, prices and packages; recommending the most attractive offers to subscribers during the shopping and ordering process; communicating with users about their usage, spending and purchase options; configuring the network to deliver more reliable services; and monitoring QoE to proactively correct any potential problems. All these activities enable improved user experience, increased loyalty, the creation of smarter networks, and extended network functionality to facilitate progress toward the Networked Society.

The profound impact that increased broadband networking will have on society will also create business opportunities in new areas for CSPs. With improved real-time connectivity and data management comes the possibility to create tailored data sets, readily available for analysis and machine learning. This would be the core ingredient in data-driven efficiency improvements in a number of business areas – for example, transport, logistics, energy, agriculture and environmental protection. Furthermore, decision-making in business and society would be facilitated by access to insights based on more accurate and up-to-date data.

In the past, CSPs have been prevented from benefiting from the value of big data on account of its sheer weight. The volume, velocity and variety – or the three Vs – of big data were simply overwhelming. Those data-handling challenges have now largely been met by a variety of easily obtained tools. Distributed databases, complex event-processing frameworks, analytics libraries and so on have been developed in the open-source community and are readily available to CSPs.

But like a blank spreadsheet, these tools are simply a platform for data handling. The real value comes from knowing which combination of the vast array of data elements reveals the desired insights. That is where deep network and operational expertise are of paramount value. Only when these key relationships are understood can the necessary insights – such as user behavior, network performance and causes of experience issues – be gained.

1. UNDERSTANDING BIG DATA

Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making.

Some of this data is held in transactional data stores – the by-product of fast-growing online activity. Machine-to-machine interactions, such as metering, call detail records, environmental sensing and RFID systems, generate their own tidal waves of data. All these forms of data are expanding, and that is coupled with fast-growing streams of unstructured and semi structured data from social media.

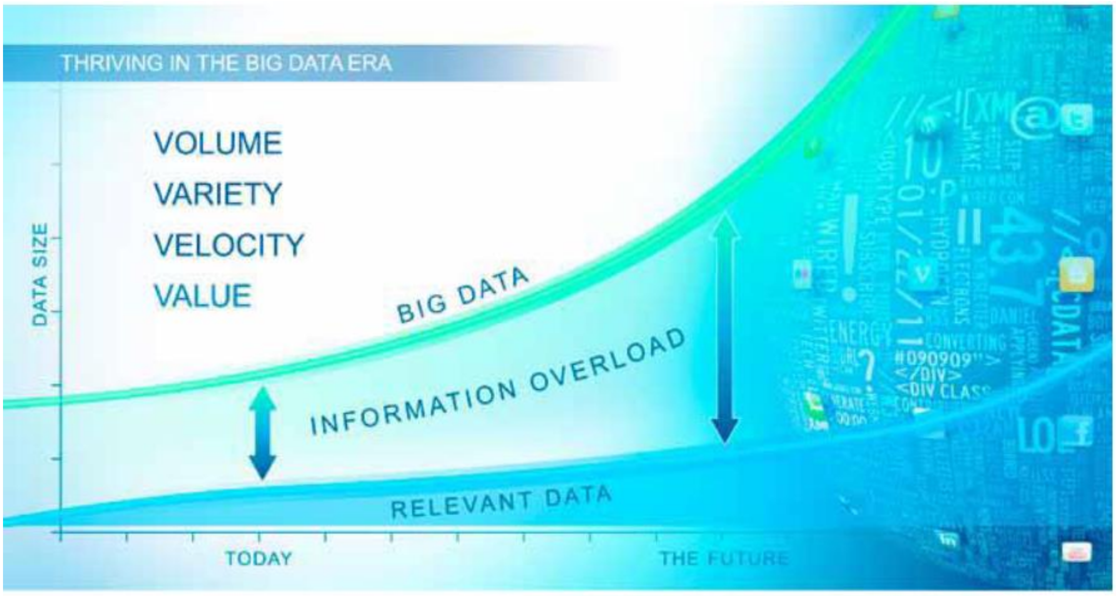


Figure 2: Determining relevant data is key to delivering value from massive amounts of data

The virtues of big data have been touted in hundreds of articles and reports during the past few years. Yet the benefits have proven elusive for a lot of companies. Indeed, some analysts already see a considerable level of disillusionment regarding big data — an umbrella term encompassing the new methods and technologies for collecting, managing, and analyzing in real time the vast increase in both structured and unstructured data — because too many efforts to implement the technology have not lived up to the high expectations triggered by the hype. This is particularly true in the telecom sector. Most operators conduct analytics programs that enable them to use their internal data to boost the efficiency of their networks, segment customers, and drive profitability with some success. But the potential of big data poses a different challenge: how to combine much larger amounts of information to increase revenues and profits across the entire telecom value chain, from network operations to product development to marketing, sales, and customer service — and even to monetize the data itself. The typical advice offered to telecom operators — indeed, to companies in every industry — is to take a top-down approach by focusing on specific business problems that big data might solve, and then gathering the data needed to solve them. But the challenge in this strategy is twofold: First, the business problem often exceeds the capacity of the available data to solve it, and second, the process of gathering the right data to help solve the problem is poorly understood by many companies. To circumvent this problem, companies should begin with the inverse approach, viewing the opportunity from the bottom up. In this scenario, you examine the data currently available, and only then determine the business problems the data might help solve, with the help of any additional structured or unstructured data that might be needed (see Exhibit 1, next page). We believe the best way to get started with this approach is through pilot programs. Keeping initial expectations reasonable, a dedicated team gathers all available data, analyzes it to allow new and unexpected opportunities to reveal themselves, and then tests the efficacy of the results in solving one or more real business problems. This tactic offers telecom operators and others a concrete starting point, a more realistic assessment of the benefits of big data, and a better understanding of what is actually needed to achieve those benefits in the long term.

1.1. Defining Big Data

Big data typically refers to the following types of data:

- ✓ Traditional enterprise data – includes customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data.
- ✓ Machine-generated /sensor data – includes Call Detail Records (“CDR”), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), trading systems data.
- ✓ Social data – includes customer feedback streams, micro-blogging sites like Twitter, social media platforms like Facebook

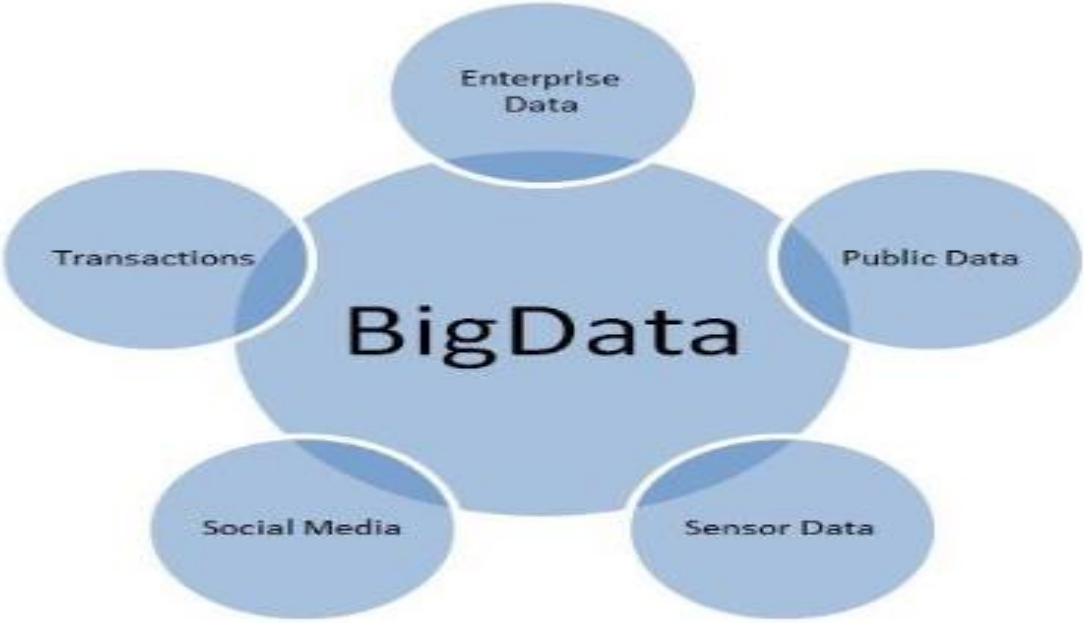


Figure 3: Source data

The McKinsey Global Institute estimates that data volume is growing 40% per year, and will grow 44x between 2009 and 2020. But while it’s often the most visible parameter, volume of data is not the only characteristic that matters. In fact, there are four key characteristics that define big data:

Volume. Machine-generated data is produced in much larger quantities than non-traditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes. Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes, compounding the problem.

Velocity. Social media data streams – while not as massive as machine-generated data – produce a large influx of opinions and relationships valuable to customer relationship management. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day). Initiatives such as the use of

RFID tags and smart metering are driving an ever greater need to deal with the torrent of data in near real time. This, coupled with the need and drive to be more agile and deliver insight quicker, is putting tremendous pressure on organizations to build the necessary infrastructure and skill base to react quickly enough.

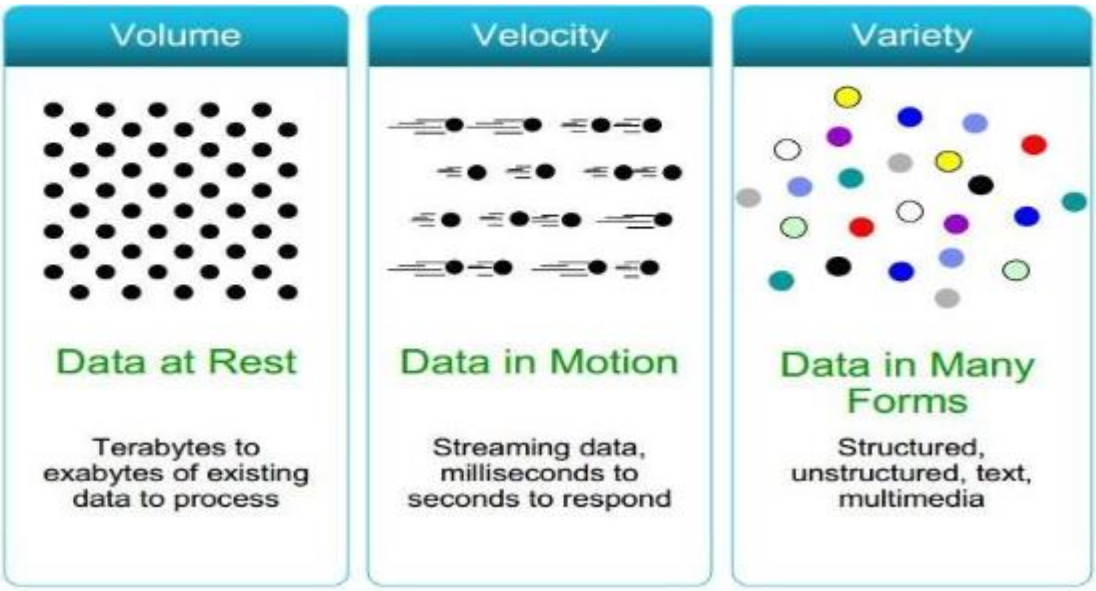

Volume
Velocity
Variety
Data at Rest
Data in Motion
Data in Many Forms
Terabytes to exabytes of existing data to process
Streaming data, milliseconds to seconds to respond
Structured, unstructured, text, multimedia

Figure 4: Categories

Variety. Traditional data formats tend to be relatively well defined by a data schema and change slowly. In contrast, non-traditional data formats exhibit a dizzying rate of change. As new services are added, new sensors deployed, or new marketing campaigns executed, new data types are needed to capture the resultant information. Up to 85 percent of an organization’s data is unstructured – not numeric – but it still must be folded into quantitative analysis and decision making. Text, video, audio and other unstructured data require different architecture and technologies for analysis. In addition to the speed at which data comes your way, the data flows can be highly variable – with daily, seasonal and event-triggered peak loads that can be challenging to manage. Difficulties dealing with data increase with the expanding universe of data sources and are compounded by the need to link, match and transform data across business entities and systems. Organizations need to understand relationships, such as complex hierarchies and data linkages, among all data.

Value. The economic value of different data varies significantly. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is identifying what is valuable and then transforming and extracting that data for analysis.

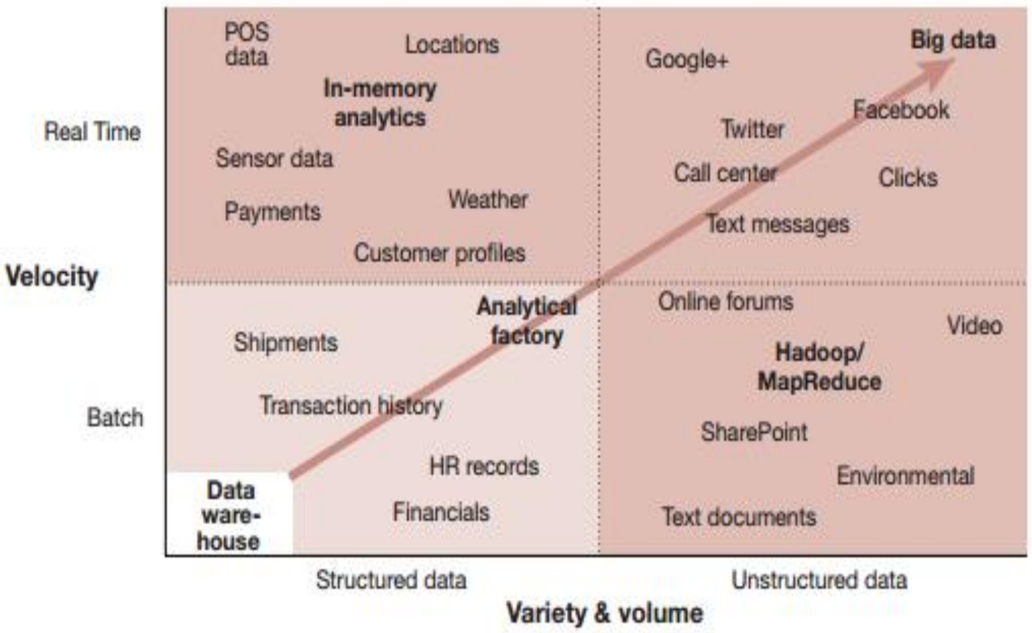


Figure 5: Expectations

To make the most of big data, enterprises must evolve their IT infrastructures to handle these new high-volume, high-velocity, high-variety sources of data and integrate them with the pre-existing enterprise data to be analyzed.

A data environment can become extreme along any of the above dimensions or with a combination of two or all of them at once. However, it is important to understand that not all of your data will be relevant or useful. Organizations must be able to separate the wheat from the chaff and focus on the information that counts – not on the information overload

1.2. Importance of Big Data

When big data is distilled and analyzed in combination with traditional enterprise data, enterprises can develop a more thorough and insightful understanding of their business, which can lead to enhanced productivity, a stronger competitive position and greater innovation – all of which can have a significant impact on the bottom line.

For example, in the delivery of healthcare services, management of chronic or long-term conditions is expensive. Use of in-home monitoring devices to measure vital signs, and monitor progress is just one way that sensor data can be used to improve patient health and reduce both office visits and hospital admittance.

Manufacturing companies deploy sensors in their products to return a stream of telemetry. In the automotive industry, systems such as General Motors' OnStar or Renault's R-Link, deliver communications, security and navigation services. Perhaps more importantly, this telemetry also reveals usage patterns, failure rates and other opportunities for product improvement that can reduce development and assembly costs.

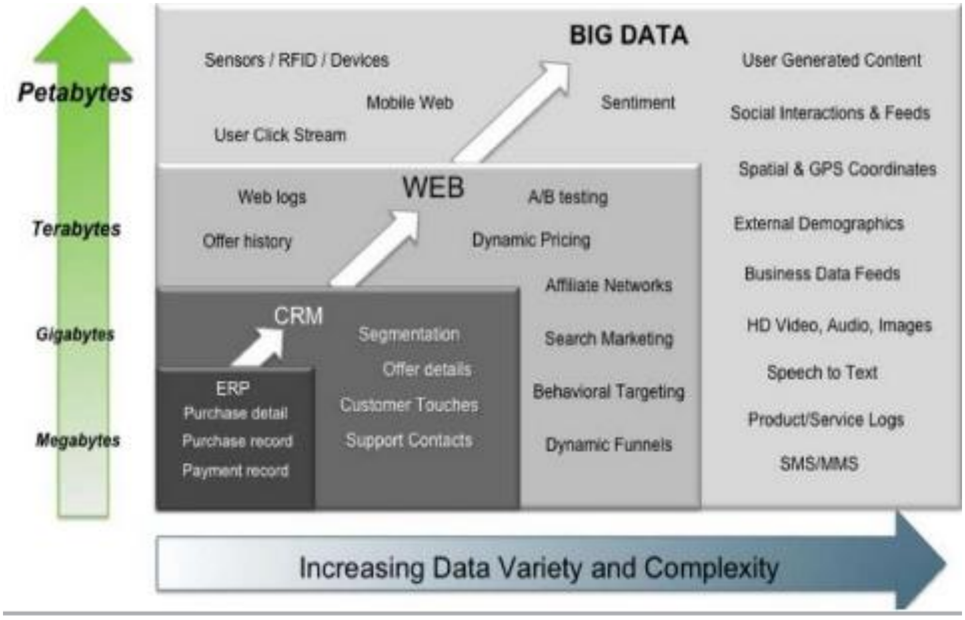


Figure 6: Defining, Big data = Transactions + Observations + Interactions

The proliferation of smart phones and other GPS devices offers advertisers an opportunity to target consumers when they are in close proximity to a store, a coffee shop or a restaurant. This opens up new revenue for service providers and offers many businesses a chance to target new customers.

Retailers usually know who buys their products. Use of social media and web log files from their ecommerce sites can help them understand who didn’t buy and why they chose not to, information not available to them today. This can enable much more effective micro customer segmentation and targeted marketing campaigns, as well as improve supply chain efficiencies through more accurate demand planning.

Finally, social media sites like Facebook and LinkedIn simply wouldn’t exist without big data. Their business model requires a personalized experience on the web, which can only be delivered by capturing and using all the available data about a user or member.

2. ANALYTICS

Except for gathering those Data we have also to identify those that is useful and meaningful for Business. There are specific rules that are defined from Business users in order to gather all data related with specific pattern.

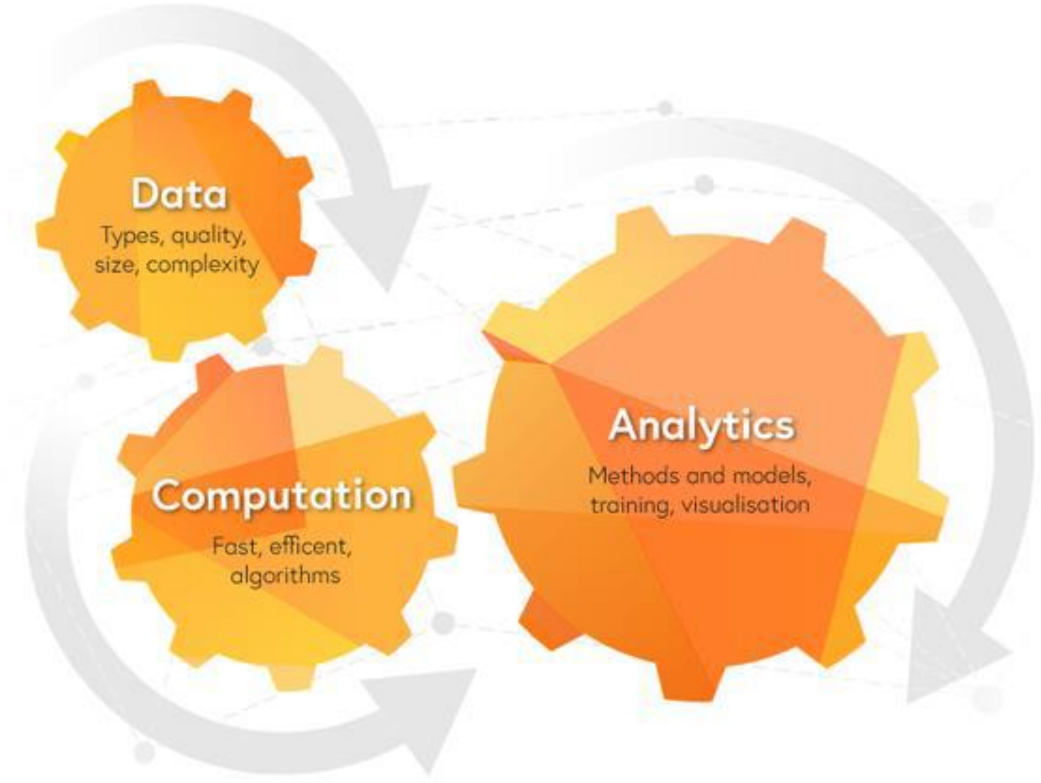

Data
Types, quality, size, complexity
Computation
Fast, efficent, algorithms
Analytics
Methods and models, training, visualisation

Figure 7: New Growth Opportunity for Software Vendors is Data Analytics

2.1. Data Science

Twenty years ago, the business transaction processing software market was in transition. Instead of building and maintaining ERP, CRM and other applications internally using large, expensive IT teams, companies began purchasing commercial off-the-shelf (COTS) applications from new vendors like SAP, JD Edwards and Peoplesoft. A multi-billion dollar software business was born.

Ten years later, in the early 2000s, the trend was further transformed by Salesforce.com and others who innovated the Software as a Service (SaaS) model and enabled companies to implement new applications without traditional data center costs or time-consuming cap-ex budget approval processes.

The major market opportunity today for enterprise software vendors and SaaS companies is in the analytic application arena.

The funding for these initiatives is available (as evidenced by the billions spent annually on custom data warehousing technology and services), and software companies can profit by putting customers on a faster path to better data-driven decision making. Such solutions enable customers to gain insights for a competitive edge, to reduce risk exposure, and/ or increase profitability.

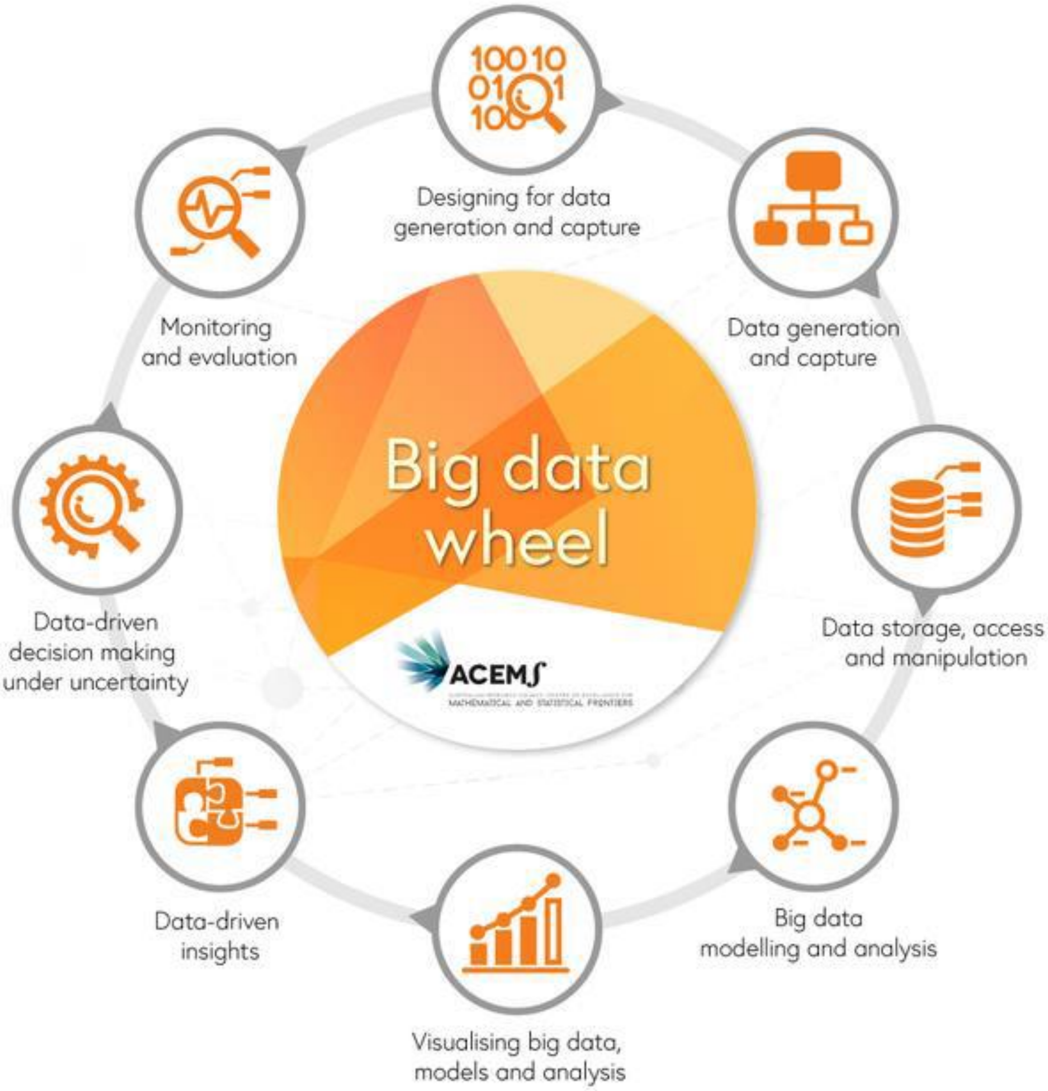

10010
01 1
10
Designing for data generation and capture
Data generation and capture
Data storage, access and manipulation
Big data modelling and analysis
Visualising big data, models and analysis
Data-driven insights
Data-driven decision making under uncertainty
Monitoring and evaluation
Big data wheel
ACEMS

Figure 8: Big Data wheel

Let us examine above wheel in order to identify the path has to be followed in order to create meaningful result for company .

1) Designing for data generation and capture . Business has to ask be low questions :
    a. What is the purpose of data ?
    b. What will we use it for ?
    c. What question do we have ?
    d. What will the data say ?

Answers to those questions will lead us to which data we really need in order to create an optimal data collection or generation strategies .

2) Data generation and capture
    a. It may be the case that we need to generate Business's data or we need to obtain it from different source
    b. If we are generating Business's data we need to accurate and efficient ways of generation
    c. If we are using other sources Business need to consider quality and relevance of source. Also how to identify data gaps and how to fill them.

Also Business need to consider, best ways to capture these data ?

Answering these Questions could not only save time and money, but can mean difference between having a wealth of useful information and having massive amount of useful junk.

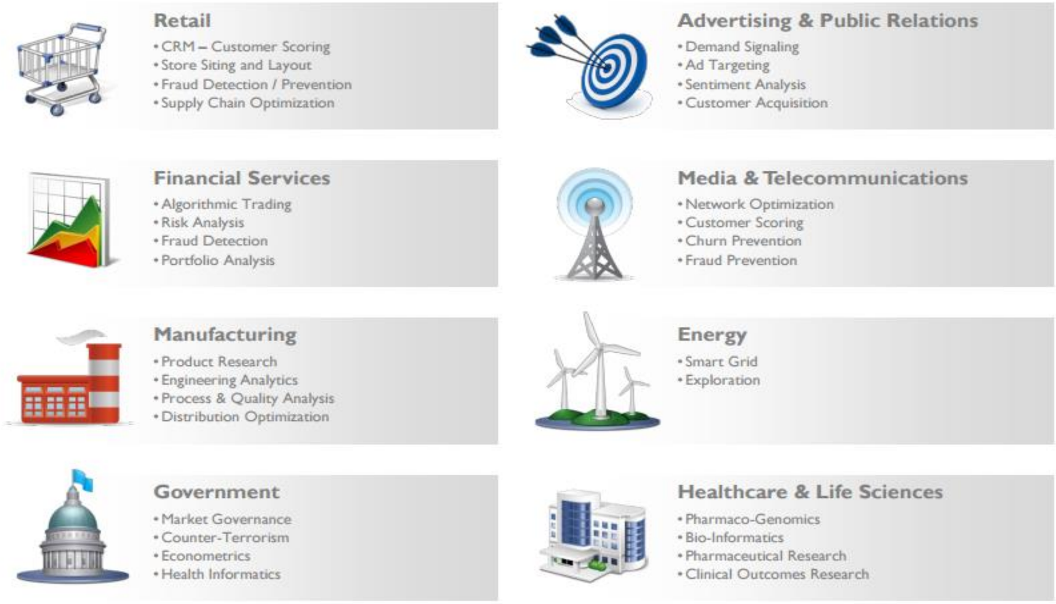
3) Data management
    a. Data management involves storage, access and manipulation
    b. In house or third party
    c. If it is third party, the facility should allow fast access and manipulation techniques for data compression and data handling.
4) Data use
    a. Reliable, insightful tools for analysis.
    b. Much more than simple summary, Graphs and Numbers.
5) Description of a problem, prediction, clustering, profiling, ranking, comparing and so on.
    a. Not only estimates but feedback is being generated for relevant estimations.
    b. Algorithm for model.
    c. Algorithms must be fast, resource efficient and scalable to large problems.
6) Visualization, Visual Analytics.
    a. Visualization gives more insight into the problem.
    b. Critical because data is BIG.
    c. Patterns easier to comprehend.
7) Data Driven Insights
    a. Getting to the basics of the problem.
    b. Requires expertise and experience.
8) Data Driven Decision Making Under Uncertainty
    a. Data driven interpretation to data driven decision making or evidence based decision.
    b. With statistical certainly.
    c. Cost based analysis along with evidence.
9) Monitoring and evaluation
    a. Continuous and iterative process.

In general Big Data is a bridge between Data on the one side and decision making on the other side. There are several vendors that offers software related with Business needs.

Enterprise software and SaaS vendors have benefited greatly from underlying database management systems allowing architects and developers to focus on their core competencies at the application level without worrying about the underlying data management. Historically, these traditional OLTP row-oriented engines could handle the transactional data entry nature of most applications. However, today’s analytic applications must continuously and simultaneously load and query against massive volumes of information.

2.2. Use Cases

There are several domains that Big Data may find use.



Retail
- CRM – Customer Scoring
- Store Siting and Layout
- Fraud Detection / Prevention
- Supply Chain Optimization

Advertising & Public Relations
- Demand Signaling
- Ad Targeting
- Sentiment Analysis
- Customer Acquisition

Financial Services
- Algorithmic Trading
- Risk Analysis
- Fraud Detection
- Portfolio Analysis

Media & Telecommunications
- Network Optimization
- Customer Scoring
- Churn Prevention
- Fraud Prevention

Manufacturing
- Product Research
- Engineering Analytics
- Process & Quality Analysis
- Distribution Optimization

Energy
- Smart Grid
- Exploration

Government
- Market Governance
- Counter-Terrorism
- Econometrics
- Health Informatics

Healthcare & Life Sciences
- Pharmaco-Genomics
- Bio-Informatics
- Pharmaceutical Research
- Clinical Outcomes Research

Figure 9: Use cases in several domains

2.2.1. Web Analytics

Web analytics measure, collect, analyze, and report on Internet data for the purpose of understanding and optimizing web usage.1

As a creator or owner of websites and Internet content, you are faced with the daunting task of understanding how your customers interact with your website or content. Observing your customers on the Internet is very data intensive, and you can choose from many tools to analyze and report on their web usage. However, you also need to look at how you can integrate data from many, different sources, such as customer relationship management (CRM) systems, visitor and ad logs, and data from social media sites.

This paper discusses web analytics, including:

✓An overview of web analytics
✓The importance of integrating clickstreams with additional data sets
✓The challenges in moving beyond metrics to insights
✓How the HP Vertica Analytics Platform can help you address web analytics

Over the past 20 years, standards have evolved for collecting, measuring, and analyzing data to optimize web usage. Companies such as Adobe® Omniture and comScore provide comprehensive suites of fee-based tools that make it easy for large customers to manage their reporting needs. Smaller sites can also perform web analytics now, as companies such as Google™ and Yahoo have launched free services that include rich reporting suites with advanced features, such as heat maps and conversion funnels.

Free tools have helped level the playing field. The owner of a start-up can generate metrics and reports to optimize his or her company’s website in the same way as the largest companies on the Internet. To stay ahead of the competition, you need to look beyond simple web metrics to gain deeper insights into how your customers use your site. This involves using sophisticated analytics to create and categorize visitor segments and then deriving value based on revenue or less tangible factors, such as loyalty.

To understand how well your website is working, you need to collect data on visitor behavior. First, you need to address user privacy by adopting a privacy policy that protects your customers' identity.

You also need three pieces of information to track your website visitors:

- Page URL—The web address of the page that your customer visited
- Referrer—The web address of the previous page that your customer visited
- Cookie—A text file that can be stored within your customer's browser to identify the site visit uniquely

You may also want to track user-submitted information, such as an email address, or information inferred from an IP address, such as location.

You can collect data in several ways, including:

- JavaScript tags—A script tracks the activity on a page and then forwards the data to a server for analysis. This is how companies such as Adobe Omniture and Google Analytics collect data.
- Server log file analysis—Web servers create the transaction data and log files contain detailed data such as visitor IP address, time of visit, and status code.

2.2.1.1. Common Data Metrics

Once you collect data, you need to aggregate it, using a common set of metrics, to get a high-level overview of visitor behavior. Common metrics include:

- Visits—By default, a visit or session is the period of time during which visitors interact with your site and inactivity is less than 30 minutes.3
- Unique visitors—This represents the number of unduplicated (counted only once) visitors to your website over a specific period of time.3 You can determine a unique visitor using cookies or an IP address.
- Page views—A page is any file or content delivered by a web server and is generally considered to be a web document.3
- Page views per visit—This is the average number of pages viewed per website visit.
- Bounce rate—This is the percentage of single-page visits or visits in which a viewer left your site from the entrance or landing page. Use this metric to measure visit quality. A high bounce rate generally means that your site entrance pages are not relevant to your visitors.3
- Conversion rate—This represents the percentage of visitors who take a meaningful action, such as signing up for an online newsletter. E-commerce site conversions may include tracking keywords, banner ads, landing pages, and pages in the purchase process.

You may need to include advertising as part of conversion reporting when clickstreams cannot address all of your web reporting needs. Many analytics tools have sophisticated conversion-tracking capabilities that can trace each step of your website visitors, from browsing to buying. However, tying advertising to conversions requires in-house or third-party ad server logs to complement clickstreams. You will need to integrate a variety of data types and sources as a result.

2.2.1.2. Moving from metrics to insights

This type of data and the simple reports generated by web analytics platforms is a start, but most companies may need additional reporting beyond conversion tracking. Many companies are starting to ask questions such as:

a. If my conversion rate is five percent, what happened to the rest of my customers?
b. What is the revenue generated per page view, and how do we improve it?
c. What segment of my customers is the most loyal and engaged?

Answering these questions requires multiple data sources. You will likely need to combine web logs and clickstream data such as:

- ✓ Advertising data logs
- ✓ Data extracts from CRM systems such as Salesforce.com
- ✓ Customer survey results
- ✓ Results from multivariate or A/B testing
- ✓ Visitor comments from social media channels, such as Twitter, LinkedIn, and Facebook

2.2.1.3. Fraud Detection

There are several markets that invests in fraud detection in order to ensure revenues. Specifically:

A. Telecom fraud is estimated to$40 B globally and it is the single biggest cause of revenue loss for operators, costing them between 3% and 5% of their annual revenue. With rising competition and extremely low average revenue per user (ARPU), detecting fraud and plugging revenue leaks have become extremely important to reduce costs.
B. One study reports that the internal fraud (40.3%), roaming fraud (11.4%), pre-paid (10.8%), subscription (11.6%) and premium (13.1%) are the most important in terms of losses by values.
C. Fraud connected to prepaid accounts is much easier to commit and harder to combat, since there is very little information on the subscriber, unlike postpaid accounts, where a credit check is usually done. Entry-level fraudulent activities such as subscription and impersonation are very serious since the cost is coming straight from the bottom line in the form of commissions and incentives.
D. The fraud management becomes more and more important as the new methods of access become available such as Cable networks, Wireless networks, DSL, Satellite, Metropolitan Optical Networks running Ethernet, Broadband Wireless Systems (radio, microwave, or infrared).
E. Although there is an abundance of data generated by various devices and systems most of the data is either not processed in real time or not processed at all.
F. Telco would like to detect critical events across all its data sources in real time, perform advanced analysis in a speedy manner, store this data on a more efficient thus providing better service to its customer and reducing the financial loses.

2.2.1.4. Problem Statements

1) The top fraud issues faced by communication service providers are as follows:
2) Lack of visibility to data to detect fraud by SIM card cloning in real time

3) Lack of analytical data to detect ‘Subscription Fraud’ using fake identity.
4) Delay in getting Customer Roaming Call Data from roaming partner paves the way for fraudsters to make roaming calls resulting in financial loss for the CSP.

5) Lack of an efficient system in place to detect any internal fraud activity, such as adding some paid service without consumer knowledge by a reseller or an employee resulting in unwanted cost and dissatisfaction for consumer.
6) Lack of Real Time Rating System to detect and prevent frauds related to prepaid telecom services.

2.2.1.5. Problem Description

First issue may be summarized in lack of visibility of data to detect fraud by SIM Card Cloning:

i. Currently no efficient way to detect SIM cloning fraud in near-real time.

ii. CSPs have to perform manual checks to detect cloning, which is not efficient.

iii. Billing system uses token management system wherein it opens a token for an initiating call, and if another call happens in the overlapping time, a system generated alert may be raised to internal monitoring representative(depends on CSP).But as of now, nothing in near real time stops 2 calls from same SIM(one cloned) from happening at overlapped times.

iv. When detected, fraudsters might have already exploited the service for many days, thereby causing financial loss to the consumer and CSP.

v. Legitimate customer come to know of the fraud when the bill is received at month-end.

vi. SIM cloning can also be an indicator of other kinds of threat such as terrorism, etc. So its early detection can be extremely valuable. For e.g. Detecting calls from a cloned SIM to a terrorist-prone area/country can help to point out the people involved in the crime.

Second issue is the lack of analytical data to detect Subscription Fraud using fake identity:

a. Subscription fraud involves the acquisition of telecommunications services using stolen or false credentials and/or identity with no intention of paying. With subscription fraud, service providers lose revenue.

b. Individual consumers are vulnerable to having their identity stolen and credit rating tarnished.

c. Posing as a credit worthy person or company, the subscription fraudster can gain access to any network, anywhere—1G, 2G or 3G.

d. A subscription fraudster typically gains access to a home network or creates subscriber accounts by help of rouge dealers or internal people within provider company, so as to appear as legitimate user. Currently, no efficient means to detect the same near-real time to minimize losses.

e. To handle the subscription fraud, it would need intelligent analytics on the data which is missing at present.

Third significant issue is the delay in getting Customer Roaming Call Data from roaming partner to home network paves the way for fraudsters to make roaming calls resulting in financial loss for the CSP, which is categorized a roaming fraud.

1) Roaming fraud causes severe loss to Telco. Acc Global Fraud Loss Survey report, fraud losses are typically 4% of revenues. For mobile service providers, approximately 25% of their total fraud is roaming fraud. Losses per handset range from $100's per handset if real time record checking is in place to $10,000 with HUR (High Usage Record) systems and $50,000 per handset for carriers that rely on clearing house data.
2) Roaming Fraud is prevalent among customers who get new service and immediately thereafter use service as a roamer on another network. Stealing mobile phones belonging to roamers, usually in vacation destinations.
3) Roaming cloning fraud—where subscriber identity numbers are used in another market—has been another widespread type of this kind of fraud.
4) Delay in the home provider receiving roamer call data from roaming networks, which can be anywhere from one to several days, which can be a problem in case of a roaming data fraud. For example, presently, roaming data usage info from roaming network to home network is currently transmitted in the from of TAPIN/TAPOUT files, whose frequency is once a day, every 4 hours, etc. depending on the service provider. Any unauthorized roaming data usage in between that time interval can't be detected near real-time, but only after 4 hours (or after an entire day).

Fourth is the lack of an efficient system in place to detect and act upon any internal fraud activity

a) A retail agent or call center agent may attach a value-added service (VAS) to an unsuspecting subscriber. For example, a ringtone can be added without the customer's knowledge or permission, resulting in a commission for the agent. Currently there is no effective means to analyze the data and detect such kind of activity.

b) Lack of awareness amongst several consumers regarding VAS make it further difficult to detect the fraud.

c)Employees can also pose as vendors. An effective means to reconcile vendor address against list of employee addresses(or vice-versa) near–real time is required.

d) An efficient way to detect in near-real time invalid/malicious vendor addresses, such as PO box, several vendors with same address, same vendor with multiple addresses, addresses associated with fraud in the past, etc is also currently missing.

Last but not least, lack of real time rating system to detect and prevent frauds related to prepaid telecom services

✓Internal manipulation of the system: By gaining access to the HLR (where all service oriented customer data is stored) the prepaid account parameter can be set to post-paid. In this case, the calls are not handled by the prepaid billing platform, but by the post-paid platform instead. This results in un-billable calls in the post-paid billing system.

✓Account adjustments via Voucher Administration Terminal: It is possible for inside personnel, e.g. customer care personnel, to manually provide credit to an account. It is common use that CC representatives are able to upgrade accounts without using an HRN. This can be done using a Voucher Administration Terminal. This situation is normally related to customers who found themselves to have purchased ‘flat’ vouchers. This functionality can be misused by a CC to illegally transfer credit to an account.

✓Tampering with billing/rating systems: Fraudsters(via insiders) can tamper with the billing/rating info.

3. NEW TECHNOLOGIES

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business.

To harness the power of big data, it is required infrastructure that can manage and process huge volumes of structured and unstructured data in realtime and can protect data privacy and security.

There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we are to examine companies that provides such high performance deliverables.

3.1. Operational Big Data

This include systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored

NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.

Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

3.2. Analytical Big Data

This includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data.

MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines.

These two classes of technology are complementary and frequently deployed together

.

3.3. Building a Big Data Platform

As with data warehousing, web stores or any IT platform, an infrastructure for big data has unique requirements. In considering all the components of a big data platform, it is important to remember that the end goal is to easily integrate your big data with your enterprise data to allow you to conduct deep analytics on the combined data set.

3.3.1. Infrastructure Requirements

The requirements in a big data infrastructure span data acquisition, data organization and data analysis.

3.3.2. Acquire Big Data

The acquisition phase is one of the major changes in infrastructure from the days before big data. Because big data refers to data streams of higher velocity and higher variety, the infrastructure required to support the acquisition of big data must deliver low, predictable latency in both capturing data and in executing short, simple queries; be able to handle very high transaction volumes, often in a distributed environment; and support flexible, dynamic data structures. NoSQL databases are frequently used to acquire and store big data. They are well suited for dynamic data structures and are highly scalable. The data stored in a NoSQL database is typically of a high variety because the systems are intended to simply capture all data without categorizing and parsing the data into a fixed schema. For example, NoSQL databases are often used to collect and store social media data. While customer facing applications frequently change, underlying storage structures are kept simple. Instead of designing a schema with relationships between entities, these simple structures often just contain a major key to identify the data point, and then a content container holding the relevant data (such as a customer id and a customer profile). This simple and dynamic structure allows changes to take place without costly reorganizations at the storage layer (such as adding new fields to the customer profile).

3.3.3. Organize Big Data

In classical data warehousing terms, organizing data is called data integration. Because there is such a high volume of big data, there is a tendency to organize data at its initial destination location, thus saving both time and money by not moving around large volumes of data. The infrastructure required for organizing big data must be able to process and manipulate data in the original storage location; support very high throughput (often in batch) to deal with large data processing steps; and handle a large variety of data formats, from unstructured to structured. Hadoop is a new technology that allows large data volumes to be organized and processed while keeping the data on the original data storage cluster. Hadoop Distributed File System (HDFS) is the long-term storage system for web logs for example. These web logs are turned into browsing behavior (sessions) by running MapReduce programs on the cluster and generating aggregated results on the same cluster. These aggregated results are then loaded into a Relational DBMS system.

3.3.4. Analyze Big Data

Since data is not always moved during the organization phase, the analysis may also be done in a distributed environment, where some data will stay where it was originally stored and be transparently accessed from a data warehouse. The infrastructure required for analyzing big data must be able to support deeper analytics such as statistical analysis and data mining, on a wider variety of data types stored in diverse systems; scale to extreme data volumes; deliver faster response times driven by changes in behavior; and automate decisions based on analytical models. Most importantly, the infrastructure must be able to integrate analysis on the combination of big data and traditional enterprise data. New insight comes not just from analyzing new data, but from analyzing it within the context of the old to provide new perspectives on old problems. For example, analyzing inventory data from a smart vending machine in combination with the events calendar for the venue in which the vending machine is located, will dictate the optimal product mix and replenishment schedule for the vending machine.

3.3.5. Solution Spectrum

Many new technologies have emerged to address the IT infrastructure requirements outlined above. At last count, there were over 120 open source key-value databases for acquiring and storing big data, while Hadoop has emerged as the primary system for organizing big data and relational databases maintain their footprint as a data warehouse and expand their reach into less structured data sets to analyze big data. These new systems have created a divided solutions spectrum comprised of:

- Not Only SQL (NoSQL) solutions: developer-centric specialized systems
- SQL solutions: the world typically equated with the manageability, security and trusted nature of relational database management systems (RDBMS) NoSQL systems are designed to capture all data without categorizing and parsing it upon entry into the system, and therefore the data is highly varied. SQL systems, on the other hand, typically place data in well-defined structures and impose metadata on the data captured to ensure consistency and validate data types.

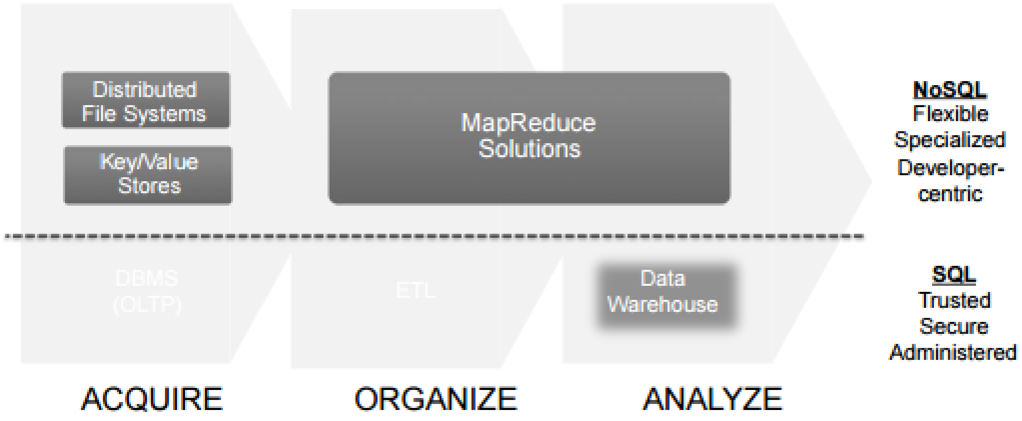


Figure 10: Divided solution spectrum

Distributed file systems and transaction (key-value) stores are primarily used to capture data and are generally in line with the requirements discussed earlier in this paper. To interpret and distill information from the data in these solutions, a programming paradigm called MapReduce is used. MapReduce programs are custom written programs that run in parallel on the distributed data nodes. The key-value stores or NoSQL databases are the OLTP databases of the big data world; they are optimized for very fast data capture and simple query patterns. NoSQL databases are able to provide very fast performance because the data that is captured is quickly stored with a single indentifying key rather than being interpreted and cast into a schema. By doing so, NoSQL database can rapidly store large numbers of transactions. However, due to the changing nature of the data in the NoSQL database, any data organization effort requires programming to interpret the storage logic used. This, combined with the lack of support for complex query patterns, makes it difficult for end users to distill value out of data in a NoSQL database. To get the most from NoSQL solutions and turn them from specialized, developer-centric solutions into solutions for the enterprise, they must be combined with SQL solutions into a single proven infrastructure that meets the manageability and security requirements of today’s enterprises.

3.4. Vendors

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

3.4.1. Hadoop Introduction

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

3.4.2. Hadoop Architecture

Hadoop framework includes following four modules:

- Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provides filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
- Hadoop YARN: This is a framework for job scheduling and cluster resource management.
- Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
- Hadoop MapReduce: This is YARN-based system for parallel processing of large data sets.

We can use following diagram to depict these four components available in Hadoop framework.

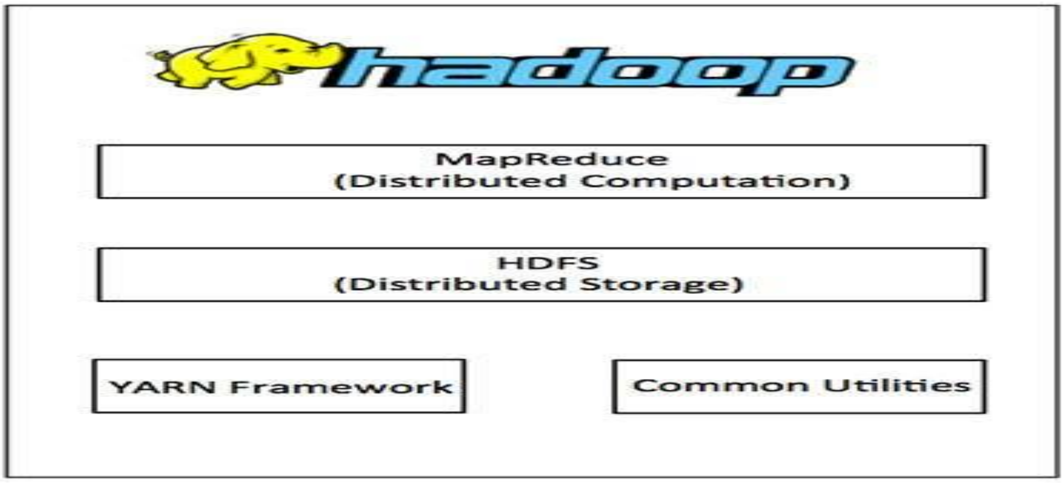
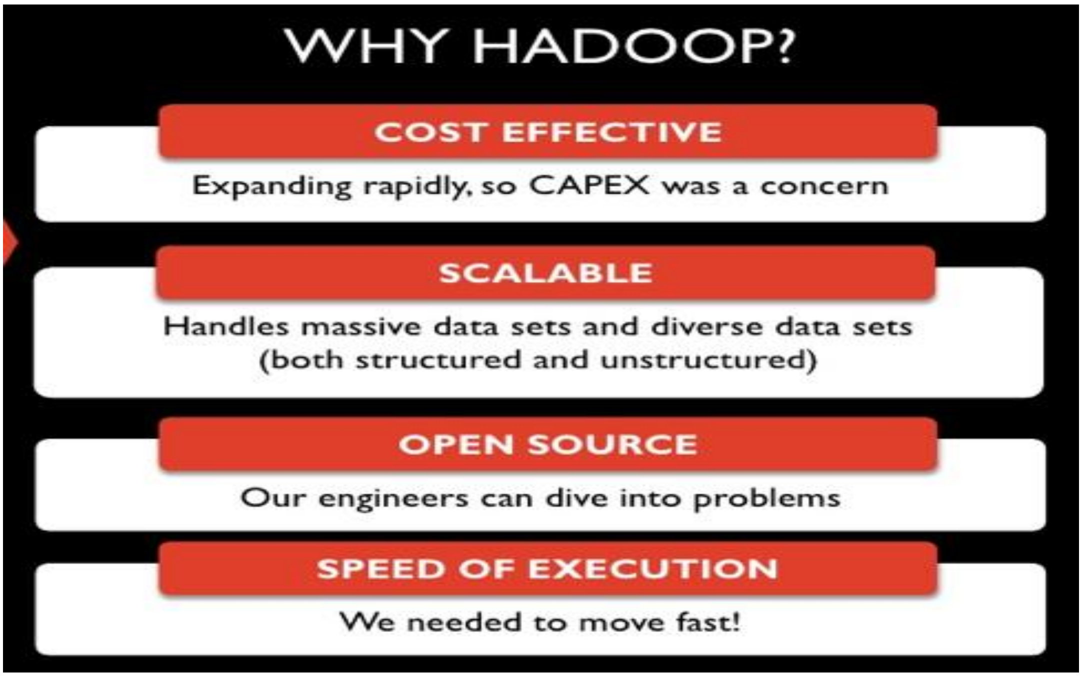


Figure 11: Depiction of Hadoop Architecture

3.4.3. Why Hadoop?

Hadoop is an open source project that offers a new way to store and process big data. The software framework is written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.

While large Web 2.0 companies such as Google and Facebook use Hadoop to store and manage their huge data sets, Hadoop has also proven valuable for many other more traditional enterprises based on its five big advantages.



WHY HADOOP?

COST EFFECTIVE

Expanding rapidly, so CAPEX was a concern

SCALABLE

Handles massive data sets and diverse data sets (both structured and unstructured)

OPEN SOURCE

Our engineers can dive into problems

SPEED OF EXECUTION

We needed to move fast!

Figure 12: Summary of Hadoop Advantages

3.4.3.1. Scalable

Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data.

3.4.3.2. Cost effective

Hadoop also offers a cost effective storage solution for businesses' exploding data sets. The problem with traditional relational database management systems is that it is extremely cost prohibitive to scale to such a degree in order to process such massive volumes of data. In an effort to reduce costs, many companies in the past would have had to down-sample data and classify it based on certain assumptions as to which data was the most valuable. The raw data would be deleted, as it would be too cost-prohibitive to keep. While this approach may have worked in the short term, this meant that when business priorities changed, the complete raw data set was not available, as it was too expensive to store. Hadoop, on the other hand, is designed as a scale-out architecture that can affordably store all of a company's data for later use. The cost savings are staggering: instead of costing thousands to tens of thousands of pounds per terabyte, Hadoop offers computing and storage capabilities for hundreds of pounds per terabyte.

3.4.3.3. Flexible

Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. This means businesses can use Hadoop to derive valuable business insights from data sources such as social media, email conversations or clickstream data. In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection.

3.4.3.4. Fast

Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing. If you're dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes, and petabytes in hours.

3.4.3.5. Resilient to failure

A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use.

The MapR distribution goes beyond that by eliminating the NameNode and replacing it with a distributed No NameNode architecture that provides true high availability. Our architecture provides protection from both single and multiple failures.

When it comes to handling large data sets in a safe and cost-effective manner, Hadoop has the advantage over relational database management systems, and its value for any size business will continue to increase as unstructured data continues to grow.

Main question is how does Hadoop Works. Below there is a depiction:

3.4.3.6. Stage 1

A user/application can submit a job to the Hadoop (a hadoop job client) for required process by specifying the following items:

1. The location of the input and output files in the distributed file system.
2. The java classes in the form of jar file containing the implementation of map and reduce functions.
3. The job configuration by setting different parameters specific to the job.

3.4.3.7. Stage 2

The Hadoop job client then submits the job (jar/executable etc) and configuration to the JobTracker which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client.

3.4.3.8. Stage 3

The TaskTrackers on different nodes execute the task as per MapReduce implementation and output of the reduce function is stored into the output files on the file system

3.4.3.9. Advantages of Hadoop

- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based

Hadoop as we have already stated is an open source framework. However there are several companies that offer solution regarding Business needs. Below we are to analyze and describe relevant offers.

3.5. Teradata

Vast amounts of data are created every day by machines as well as billions of peoples using computers, smart phones, tablets, and other personal electronics. Client are now capturing trillions of bytes of information about their customers, suppliers, and operations. Networked sensors in devices such as mobile phones, smart energy meters, automobiles, and industrial machines sense, create, and communicate data constantly.

Client need to capture, store and analyze not only structure data, but new forms of multi-structured data such as web logs, social media, test, graphics, email, audio, machine-generated data, as much more. Their users need powerful analytics to discover patterns in this data using the skills and tools they already possess.

Until now, these huge amounts of data have been ignored or underutilized simply because the tools didn’t exit to make sense of it all.

Teradata Unified Data architecture is a new innovation that lets organization can leverage all their data for new insights and new business opportunity. Teradata Unified Data Architecture is the most Powerful and complete analytics solution. By integrating the Teradata data warehouse, Aster discovery platform, and open-source Hadoop into Teradata Unified Data Architecture bridges the gap between the business language of SQL and the emerging popularity of MapReduce. The result is a unified, high-performance analytics environment.

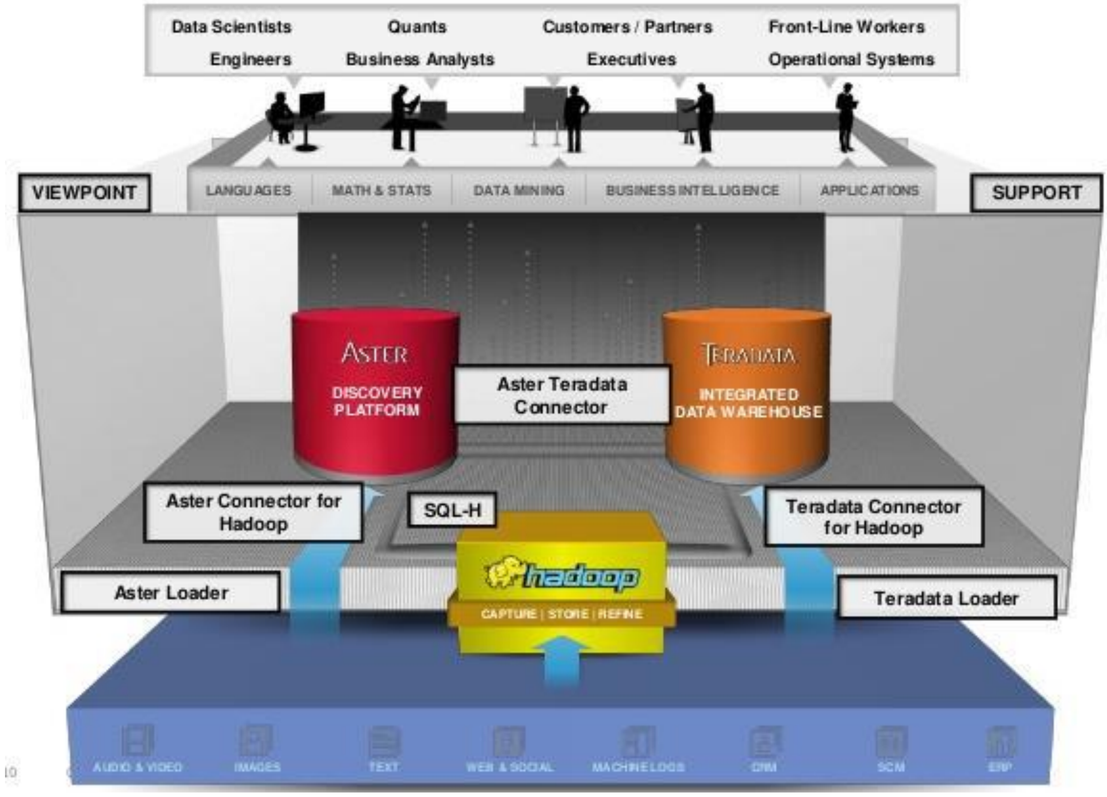
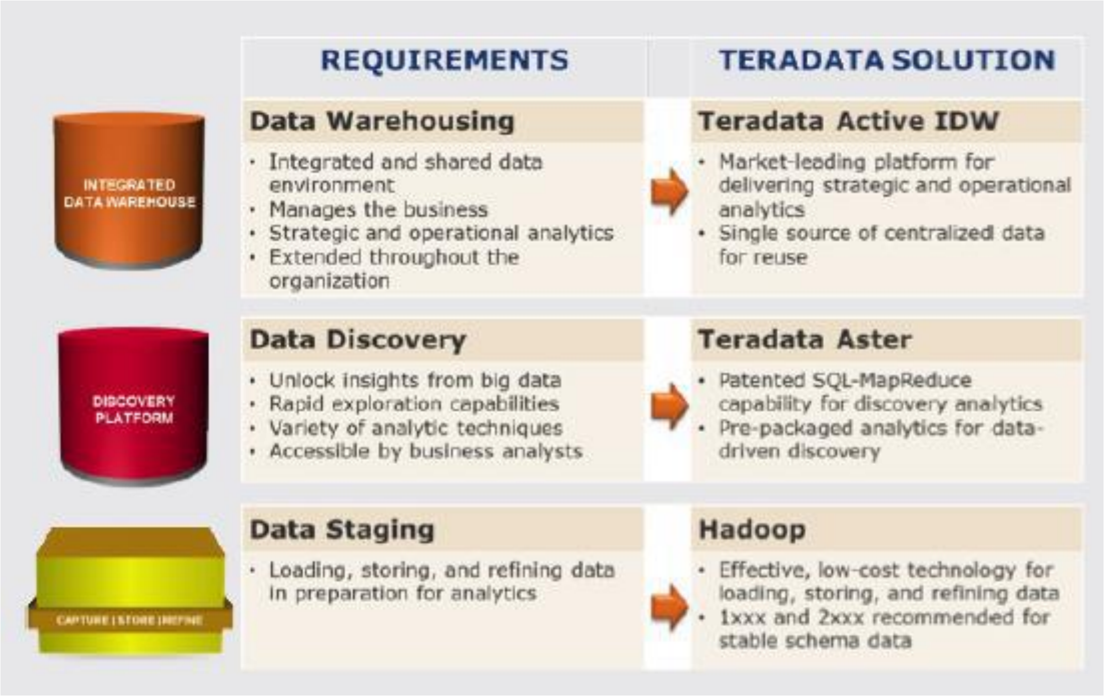


Figure 13: Key components of T-UDA are data warehousing, data discovery and data staging.

The Teradata database is the market-leading platform for delivering strategic and operational analytics throughout your organization so users can access a single source of consistent, centralized, integrated data. Teradata's approach to integrated data supports the highest business value through cross-functional analysis. With more than 30 years, Teradata DB runs the world's leading data warehouses.

The Teradata Aster database brings the power of MapReduce to business users. Patented Aster SQL-MapReduce empowers business users to run MapReduce functions using SQL, thus enabling data discovery through iterative analytics against both multi-structured data as well as structured data.

For Client that requires an open source Hadoop solution, Teradata delivers Hadoop appliance that goes beyond standard ones. Teradata integrated Hortonworks Hadoop with robust tools for system management, data access and supports for all Teradata products.



	REQUIREMENTS	TERADATA SOLUTION
INTEGRATED DATA WAREHOUSE	**Data Warehousing** • Integrated and shared data environment • Manages the business • Strategic and operational analytics • Extended throughout the organization	**Teradata Active IDW** • Market-leading platform for delivering strategic and operational analytics • Single source of centralized data for reuse
DISCOVERY PLATFORM	**Data Discovery** • Unlock insights from big data • Rapid exploration capabilities • Variety of analytic techniques • Accessible by business analysts	**Teradata Aster** • Patented SQL-MapReduce capability for discovery analytics • Pre-packaged analytics for data-driven discovery
CAPTURE | STORE | REFINE	**Data Staging** • Loading, storing, and refining data in preparation for analytics	**Hadoop** • Effective, low-cost technology for loading, storing, and refining data • 1xxx and 2xxx recommended for stable schema data

Figure 14: Requirements and Solution depiction

T-UDA brings thee powerful technologies together and integrates them with value-add software, staging and support.

Teradata integrates key value-add enabling technologies such as Teradata BYNET that unify the solution; allowing client to focus their efforts on extracting business value from their analytics rather than trying to make together. The key value-add-enabling technologies are:

A)Transparent Access-

- SQL- H provides a robust interface for run-time data access from Aster Database to Hadoop, with Teradata Database and Aster Database.
- SQL Assistant provides a use-friendly SQL creation front-end for a consistent express across Teradata database and Aster Database.
- Unity Director automatically routes users and queries between Teradata systems based on context of the query and system availability.

B)Seamless Data Movement

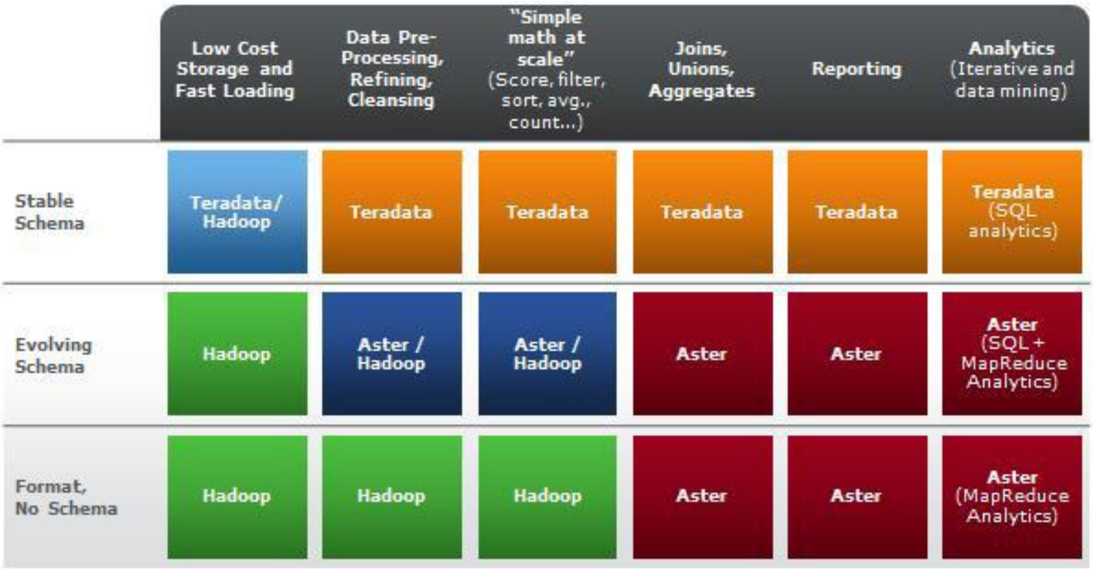
- Connectors provide easy to use, high-speed data movement between Teradata Database, Aster Database and Hadoop, and Teradata Database and Hadoop.
- Smart Loader for Hadoop gives users and administrators a friendly drag and drop interface for data movement Teradata Database and Hortonworks Hadoop.

- Unity Data Mover delivers intelligent, high speed data movement between Teradata systems. It combination of command line or GUI driven interface couple with its automatic selection of load utility gives users and administrators a powerful tool for data movement.

Bringing these technologies together allows organizations to quickly run iterative analytics against a broad, deep set of data using SQL, SQL- MapReduce, non-SQL languages and tools.

C)Single Operational View for Management

- Teradata vital infrastructure brings one-stop support to the Teradata workload-specific Platform Family, Aster Big Analytics Appliance. And Automated monitoring and fault escalation for all the these technologies delivered from as single source
- Unity Ecosystem Manager brings end –to-end monitoring of process, components, and data across Teradata systems.
- Unity Director makes managing multiple systems running the Teradata Database easy by intelligently applying database management commands to all participating Teradata systems.



	Low Cost Storage and Fast Loading	Data Pre-Processing, Refining, Cleansing	"Simple math at scale" (Score, filter, sort, avg., count...)	Joins, Unions, Aggregates	Reporting	Analytics (Iterative and data mining)
Stable Schema	Teradata/ Hadoop	Teradata	Teradata	Teradata	Teradata	Teradata (SQL analytics)
Evolving Schema	Hadoop	Aster / Hadoop	Aster / Hadoop	Aster	Aster	Aster (SQL + MapReduce Analytics)
Format, No Schema	Hadoop	Hadoop	Hadoop	Aster	Aster	Aster (MapReduce Analytics)

Figure 15: Vendor comparison

3.6. Oracle

Increasingly, data lives in many places. There are distinct advantages to relational databases, but there are also real benefits to data stores such as Hadoop and noSQL databases. At Oracle, we view the relational database as a part of what runs your business. It's secure, handles a multitude of workloads, and handles mission-critical tasks all over the world. However, organizations are finding ways to change their business through new data using Hadoop. The low-cost and flexible storage of Hadoop allows business to find value in data that was once cost-prohibitive to store. Similarly, NoSQL technologies are allowing businesses to economically scale out simple data to optimize the costs of operating in the information age.

As the world's premier data management provider, Oracle firmly believes customers should use the right tool for the job whenever possible. This ensures that data serves the business, and Oracle's data management portfolio maximizes the power of AND, delivering faster time to value and easier integration.

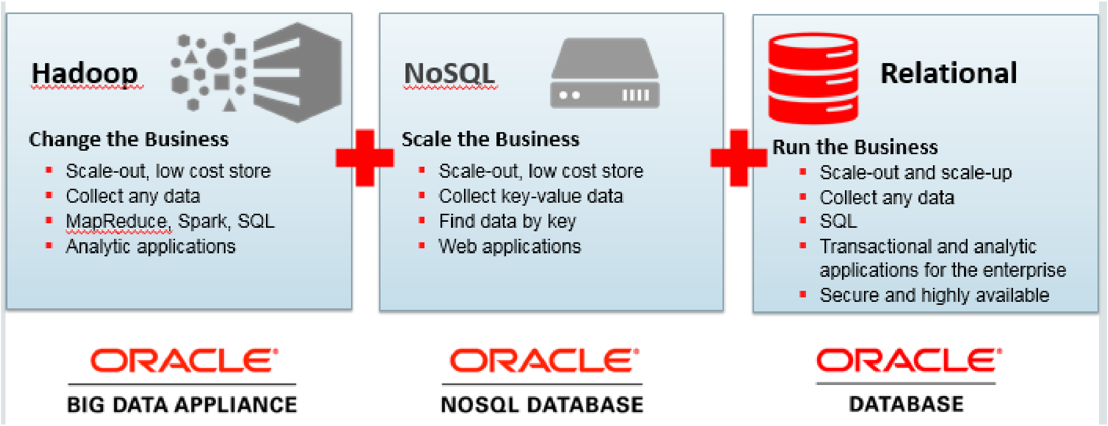


Figure 16: Oracle's Big Data Solution

Oracle is the first vendor to offer a complete and integrated solution to address the full spectrum of enterprise big data requirements. Oracle's big data strategy is centered on the idea that you can extend your current enterprise information architecture to incorporate big data. New big data technologies, such as Hadoop and Oracle NoSQL database, run alongside your Oracle data warehouse to deliver business value and address your big data requirements.

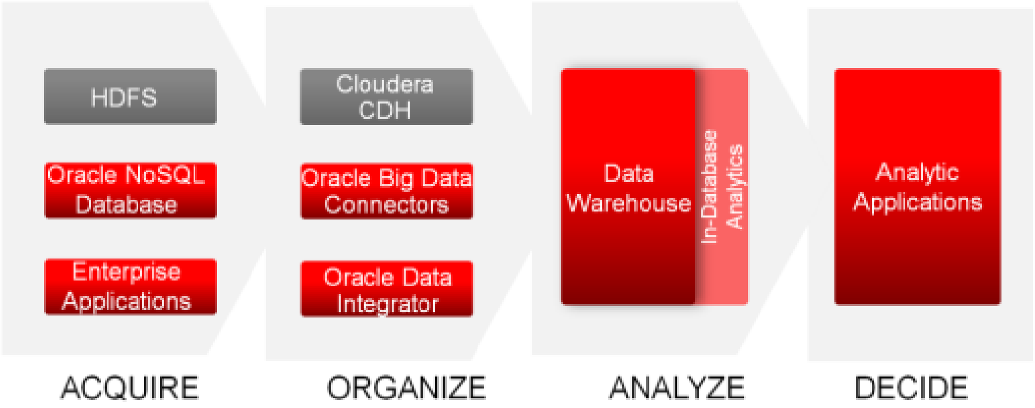


Figure 17: Oracle's solution

Oracle Big Data Appliance Oracle Big Data Appliance is an engineered system that combines optimized hardware with a comprehensive big data software stack to deliver a complete, easy-to-deploy solution for acquiring and organizing big data. Oracle Big Data Appliance comes in a full rack configuration with 18 Sun servers for a total storage capacity of 648TB. Every server in the rack has 2 CPUs, each with 8 cores for a total of 288 cores per full rack. Each server has 64GB1 memory for a total of 1152GB of memory per full rack.

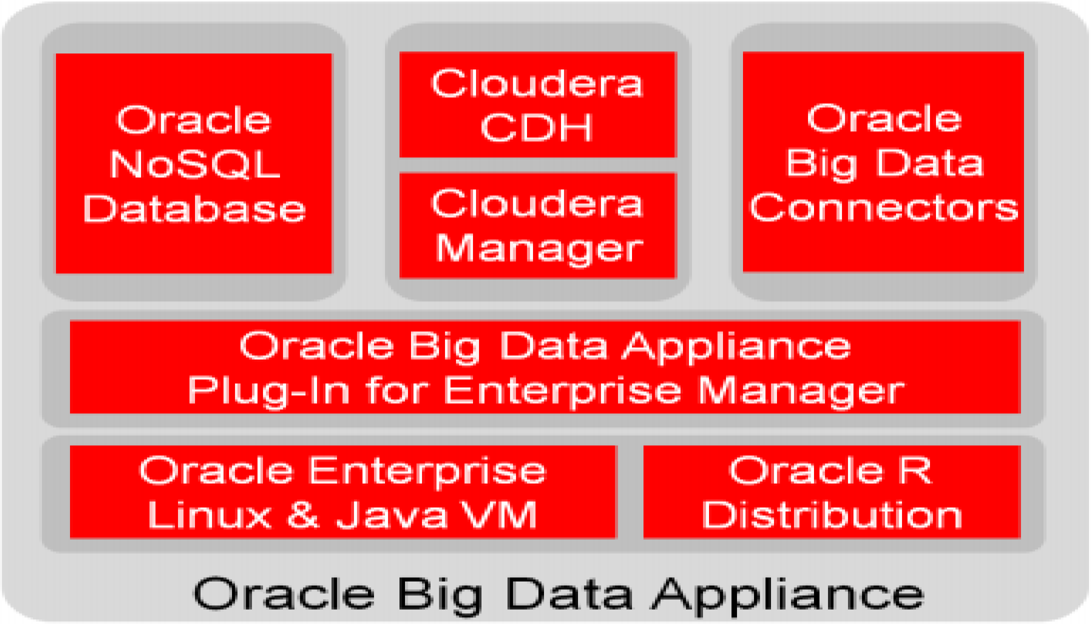

Oracle NoSQL Database
Cloudera CDH
Cloudera Manager
Oracle Big Data Connectors
Oracle Big Data Appliance Plug-In for Enterprise Manager
Oracle Enterprise Linux & Java VM
Oracle R Distribution
Oracle Big Data Appliance

Figure 18: High Level of software on Big Data Appliance

Oracle Big Data Appliance includes a combination of open source software and specialized software developed by Oracle to address enterprise big data requirements. The Oracle Big Data Appliance software includes:

- Full distribution of Cloudera's Distribution including Apache Hadoop (CDH4)
- Oracle Big Data Appliance Plug-In for Enterprise Manager
- Cloudera Manager to administer all aspects of Cloudera CDH
- Oracle distribution of the statistical package R
- Oracle NoSQL Database Community Edition2
- And Oracle Enterprise Linux operating system and Oracle Java VM

Oracle NoSQL Database is a distributed, highly scalable, key-value database based on Oracle Berkeley DB. It delivers a general purpose, enterprise class key value store adding an intelligent driver on top of distributed Berkeley DB. This intelligent driver keeps track of the underlying storage topology, shards the data and knows where data can be placed with the lowest latency. Unlike competitive solutions, Oracle NoSQL Database is easy to install, configure and manage, supports a broad set of workloads, and delivers enterprise-class reliability backed by enterprise class Oracle support.

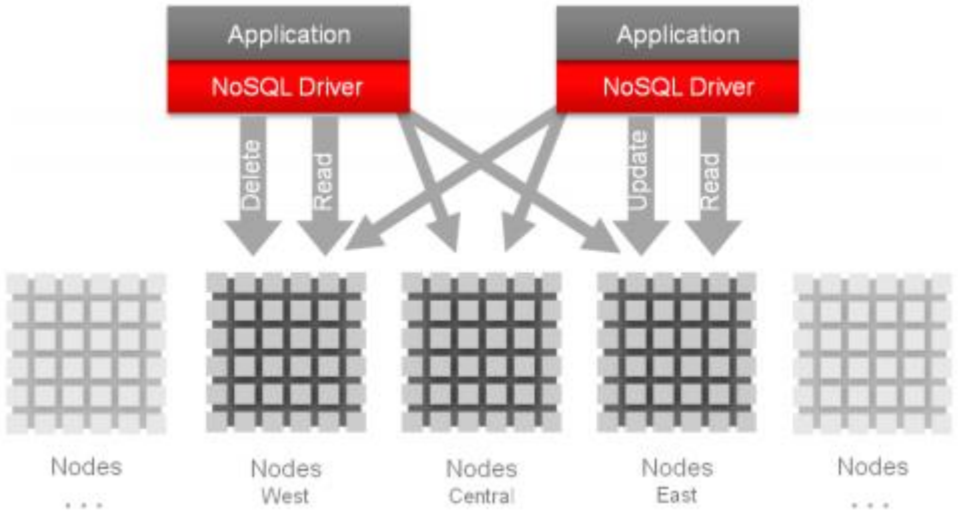


Figure 19: NoSQL Database Architecture

The primary use cases for Oracle NoSQL Database are low latency data capture and fast querying of that data, typically by key lookup. Oracle NoSQL Database comes with an easy to use Java API and a management framework. The product is available in both an open source community edition and in a priced enterprise edition for large distributed data centers. The former version is installed as part of the Big Data Appliance integrated software.

- Oracle Big Data Connectors

Where Oracle Big Data Appliance makes it easy for organizations to acquire and organize new types of data, Oracle Big Data Connectors tightly integrates the big data environment with Oracle Exadata and Oracle Database, so that you can analyze all of your data together with extreme performance. The Oracle Big Data Connectors consist of four components:

- Oracle Loader for Hadoop

Oracle Loader for Hadoop (OLH) enables users to use Hadoop MapReduce processing to create optimized data sets for efficient loading and analysis in Oracle Database 11g. Unlike other Hadoop loaders, it generates Oracle internal formats to load data faster and use less database system resources. OLH is added as the last step in the MapReduce transformations as a separate map – partition – reduce step. This last step uses the CPUs in the Hadoop cluster to format the data into Oracle's internal database formats, allowing for a lower CPU utilization and higher data ingest rates on the Oracle Database platform. Once loaded, the data is permanently available in the database providing very fast access to this data for general database users leveraging SQL or business intelligence tools.

- Oracle SQL Connector for Hadoop Distributed File System

Oracle SQL Connector for Hadoop Distributed File System (HDFS) is a high speed connector for accessing data on HDFS directly from Oracle Database. Oracle SQL Connector for HDFS gives users the flexibility of querying data from HDFS at any time, as needed by their application. It allows the creation of an external table in Oracle Database, enabling direct SQL access on data stored in HDFS. The data stored in HDFS can then be queried via SQL, joined with data stored in Oracle Database, or loaded into the Oracle Database. Access to the data on HDFS is optimized for fast data movement and parallelized, with automatic load balancing. Data on HDFS can be in delimited files or in Oracle data pump files created by Oracle Loader for Hadoop

- Oracle Data Integrator Application Adapter for Hadoop

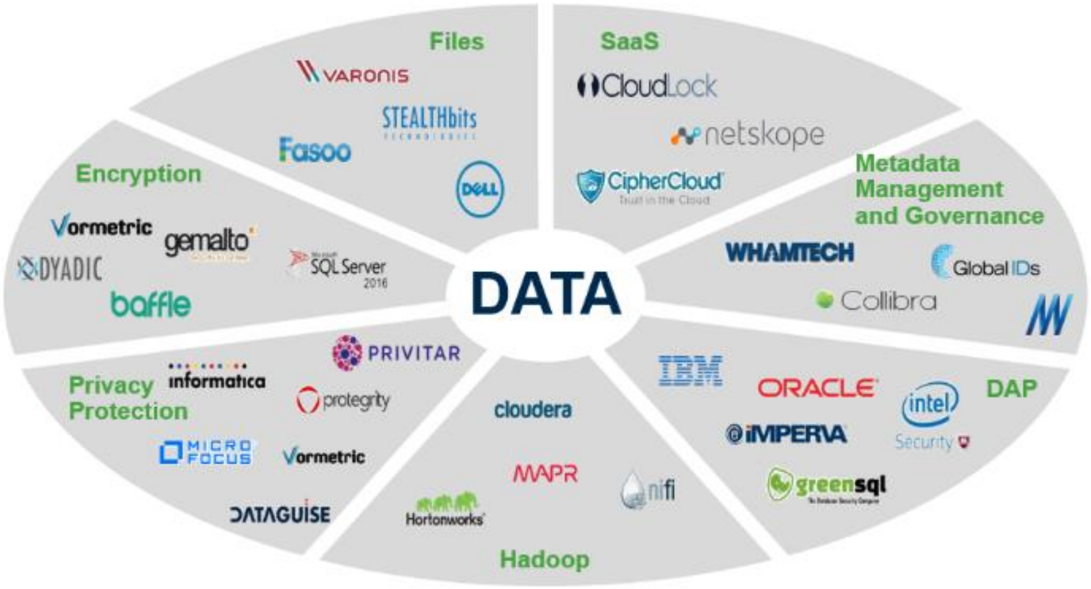
Oracle Data Integrator Application Adapter for Hadoop simplifies data integration from Hadoop and an Oracle Database through Oracle Data Integrator's easy to use interface. Once the data is accessible in the database, end users can use SQL and Oracle BI Enterprise Edition to access data. Enterprises that are already using a Hadoop solution, and don't need an integrated offering like Oracle Big Data Appliance, can integrate data from HDFS using Big Data Connectors as a standalone software solution.

- Oracle R Connector for Hadoop

Oracle R Connector for Hadoop is an R package that provides transparent access to Hadoop and to data stored in HDFS. R Connector for Hadoop provides users of the open-source statistical environment R with the ability to analyze data stored in HDFS, and to scalably run R models against large volumes of data leveraging MapReduce processing – without requiring R users to learn yet another API or language. End users can leverage over 3500 open source R packages to analyze data stored in HDFS, while administrators do not need to learn R to schedule R MapReduce models in production environments. R Connector for Hadoop can optionally be used together with the Oracle Advanced Analytics Option for Oracle Database. The Oracle Advanced Analytics Option enables R users to transparently work with database resident data without having to learn SQL or database concepts but with R computations executing directly in-database.

4. BIG DATA IN SECURITY

The data security ecosystem is bigger than most. The increased use of data silos and adoption of cloud has led to an increase in the amount of unstructured data, leading to an increase in use cases. The result is an environment with more tools and knobs to turn to create a more secure data environment.



© 2017 Gartner, Inc.

Figure 20: Data ecosystem

However, with the increase in cyberthreats, and the criticality and value of data, data protection products need to continue to evolve rather than add new tools and knobs. As end users continue moving toward digital business and adopting cloud services, particularly unapproved and untracked ones (shadow IT), their data becomes even more at risk. In addition, they need products and tools that provide stronger controls over access, visibility and monitoring, as these are the main ingredients in the future of data security.

4.1. Analysis

However, with the increase in cyberthreats, and the criticality and value of data, data protection products need to continue to evolve rather than add new tools and knobs. As end users continue moving toward digital business and adopting cloud services, particularly unapproved and untracked ones (shadow IT), their data becomes even more at risk. In addition, they need products and tools that provide stronger controls over access, visibility and monitoring, as these are the main ingredients in the future of data security.

First, as part of a people-centric data security approach, data access governance (DAG) products must be in place. DAG's primary purpose is to answer questions about who has access to what data residing in an organization's repositories, how that data is classified, and what is the history of accessing that data. This means closer integration between DLP, DCAP, DAG, privileged access management (PAM), and identity governance and administration (IGA) products.

Second, Gartner advises security and risk managers and CISOs to adopt a CARTA approach. This approach extends beyond the adaptive security architecture (ASA) of minimizing the risk of loss from external intruders, accidental and intentional internal theft, and misconfigurations. Examples include Verizon's recent 6 million user data leak , and Dow Jones exposing 2.2 million customers' information . Existing trust models, and access and control are flawed in their current state (see "Use a CARTA Strategic Approach to Embrace Digital Business Opportunities in an Era of Advanced Threats"). From a data perspective, the technology aspects of CARTA combine the need for adaptive access control, and user and entity behavior analytics (UEBA) monitoring for legacy applications.

For years, security and risk managers and CISOs have been investing in point solutions and tools such as DLP, encryption, and network and endpoint security products. Meanwhile, they have been underinvesting in the areas of detection and response, such as endpoint detection and response (EDR), security information and event management (SIEM), and UEBA. They will continue to invest in preventative controls — DLP; enterprise digital rights management (EDRM); DCAP; products to secure the Internet of Things (IoT), endpoints and networks; and cloud data security. However, these tools need to evolve to include enhanced technologies so organizations will be more successful in their implementations and better positioned for more predictive capabilities.

Technology strategic partners should help end-user organizations apply a continuous risk- and trust-based assessment approach to data by shaping product capabilities with CARTA in mind.

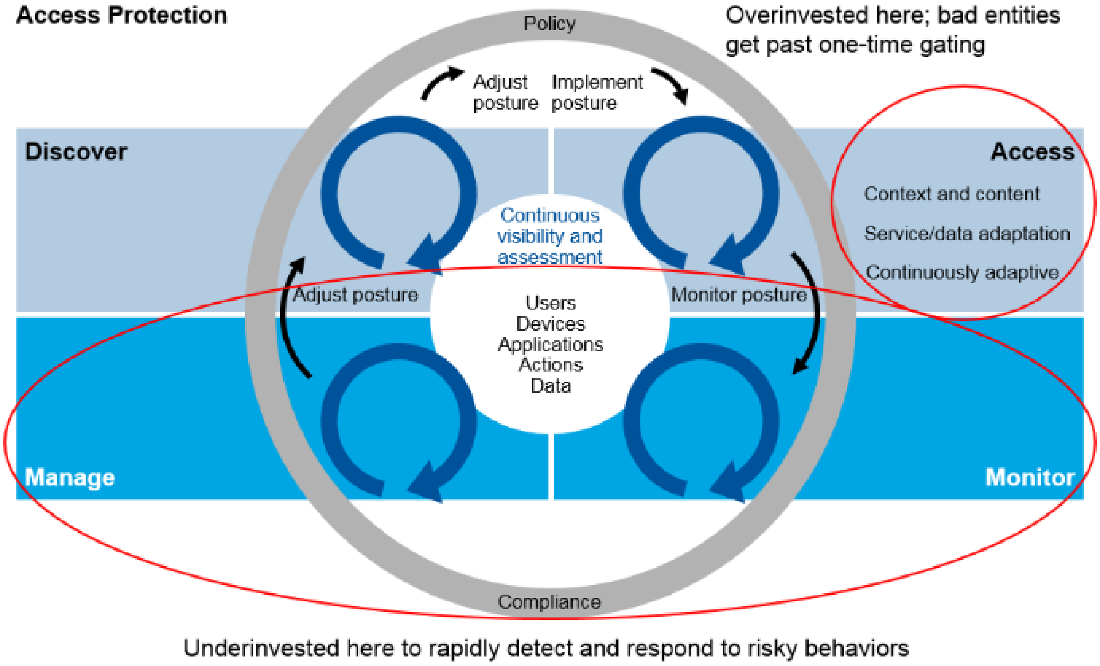


Figure 21: Gartner's Continuous Adaptive Risk and Trust Assessment Architecture

In the following sections, we outline "nontraditional" and alternative disruptive technologies that are in the very early stages of being part of the data security ecosystem. Technology strategic planners can use these emerging approaches to better position their customers for success and to enhance their products through the use of advanced analytics, ML/AI, blockchain and multiparty computing (see "Hype Cycle for Data Security, 2017"). We believe technology strategic planners need to think outside of the traditional technology box by investigating integration with or developing alternate disruptive technologies, and creating acquisition and integration strategies.

4.2. Emerging Technologies and Approaches

In below schema there is a depiction of technologies that aim to improve the protection of data within three main themes (intelligent, digital and mesh) for competitive advantage and technologies not discussed in this document (such as virtual reality and intelligent apps). The technologies outlined below are either in the very early stages of integration with data security products or not at all. They can be beneficial to fill in the gaps that current product sets lack for stronger data security.[16]

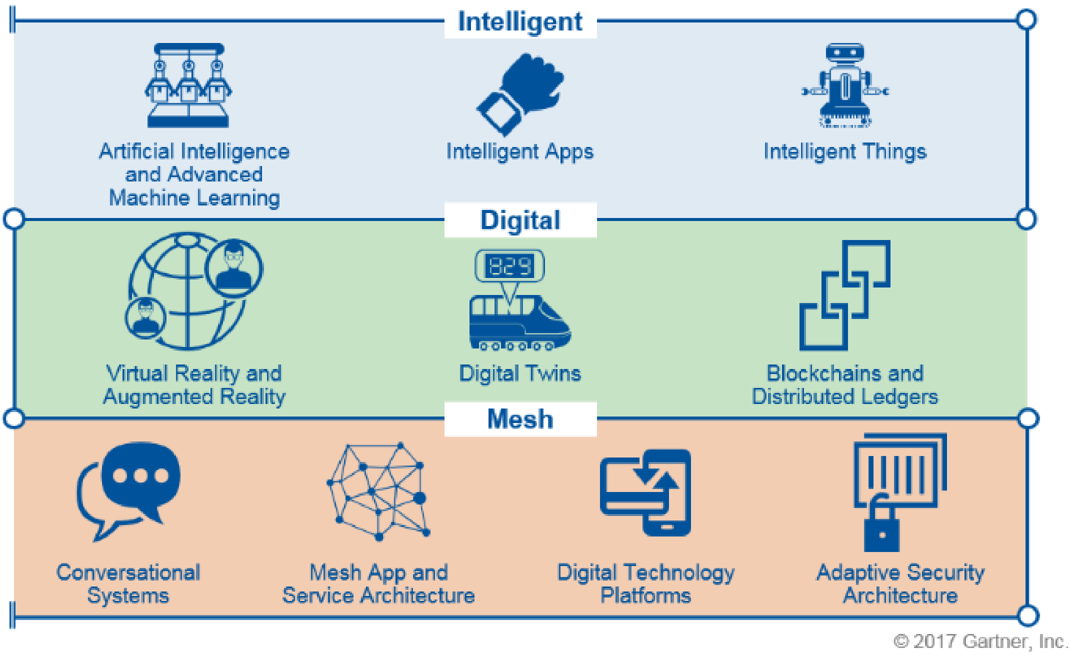

Intelligent
Artificial Intelligence and Advanced Machine Learning
Intelligent Apps
Intelligent Things
Digital
Virtual Reality and Augmented Reality
Digital Twins
Blockchains and Distributed Ledgers
Mesh
Conversational Systems
Mesh App and Service Architecture
Digital Technology Platforms
Adaptive Security Architecture
© 2017 Gartner, Inc.

Figure 22: Top 10 Strategic Technology Trends for 2017

4.2.1. Advanced Security Analytics

Advanced analytics has been incorporated into fraud detection applications for decades. More recently, it has been introduced into the overall security picture for behavioral and anomaly detection. In the past, security methods were largely rule-based to detect attacks and protect assets and information inside an organization's network. Advanced analytics, especially for detection, has been lacking in enterprise security. Then, a few years ago, end users realized their security systems were struggling, if not failing, to stop the onslaught of breaches. In addition, detection and monitoring products are challenged with separating out false positives while maintaining an adequate level of visibility. Analytics applies logic and mathematics to data to uncover insights to make better decisions in real time and to help address the inadequacies of historic detection and monitoring tools. New developments in processing power and software intelligence are finally making intrusion detection a reality. Advanced security analytics is defined as applying attributes that answer questions in the four categories of:

- ✓Descriptive — What is happening?
- ✓Diagnostic — Why did it happen?
- ✓Predictive — What will happen?
- ✓Prescriptive — What should I do about it?

Advanced security analytics enable smarter and more automated outcomes of how data is handled and protected. Including advanced security analytics in existing product sets will address the current lack of more automated responses to detect and mitigate potential data theft.

4.2.2. Artificial Intelligence and Machine Learning

Frequently used buzzwords, AI and ML are often used interchangeably when discussing big data and analytics, but they are different. Machine learning is an application of artificial intelligence, where AI can be viewed as the broad concept of machines acting "smart" and having humanlike capabilities. AI comes in two ways, general and applied.

- Applied AI has been around for years. It is commonly found in systems that make autonomy work, for example, in autonomous vehicles, such as the Google car.
- General AI is more advanced and takes AI a step further. It takes data, and based on characteristics and patterns of the data, the code and algorithms can make automated decisions based on what the system has learned about the data.

These technologies are in the very early stage when applied in the data security realm. Technology strategic planners must be careful to not just add a technology for the sake of adding technology to products. Instead, they must be able to communicate the business value of these technologies to their customers. Providers should be able to attribute specific benefits of the capability (see "Tech Go-To-Market: Creating the Path From Features to Value"). Often, the features and benefits are easy to communicate, but the differentiation and value are not brought forward. While AI and ML are disruptive technologies in the data security market, end users need to understand how these technologies will improve insights and create actionable analyses in considerably shorter periods of time (in minutes versus hours). An early example is seen with telecom operator BT's announcement in September 2017. BT is testing the use of AI and ML to determine patterns that require an action and automating that action in a time frame that surpasses human capability.

An example of a vendor utilizing ML in a cloud storage environment to protect data is Amazon Web Services' (AWS') Amazon Macie . Macie is a security service that uses ML to help protect stored data by automatically discovering and classifying sensitive data, which is useful for rogue data that may go unnoticed. Once proper classification is applied, sensitive data can be appropriately protected with either encryption, masking or redaction, depending on the policy and rules. The Amazon Macie service claims to continuously monitor data access activity, generating alerts should anomalies be detected. It is currently only available with Amazon's S3 cloud storage service.

4.2.3. Multiparty Computing

Multiparty computing (MPC) is a method of cryptography that enables entities (applications, individuals or devices) to work with data while keeping that data in a protected, confidential and private state. Specifically, MPC allows for multiple parties (that is, users) to share data while maintaining confidentiality without exposing encryption keys. As such, MPC provides advantages over traditional approaches to encryption. An early use-case example of MPC is in supply chain collaboration where the parties involved do not want to share data because of security concerns. The use of MPC in supply chain management for data security is evidenced by studies conducted by Florian Kerschbaum and others from 2008 through 2011.

A number of MPC-based security solutions have emerged in the market to help end-user organizations increase security and privacy in a more effective and efficient manner. Due to the advantages that MPC brings over traditional methods of encryption, technology strategic planners should consider expanding their capabilities to include MPC by increasing partnerships .

4.2.4. Blockchain

Labeled "disruptive," blockchain is a promising technology. A core concept of blockchain technology is a "distributed ledger." Data, regardless of type, typically resides in ledgers (for example, databases), therefore falling squarely within the concept of the technology. Blockchain-enabled data security applications offer alternative methods to establish trust and resiliency, and to track digital assets (data types, identifiers, encryption keys, transactions, device attributes) with little reliance on centralized arbiters. Vendors such as Schedule1 (DataPassports), Post-Quantum and Acronis are utilizing blockchain in different ways to improve their data security applications. For example, Post-Quantum uses blockchain to secure each block through a quantum key rather than a digital key. Acronis uses blockchain to verify and protect data using timestamps and certificates of authority.

4.2.5. Differential Privacy

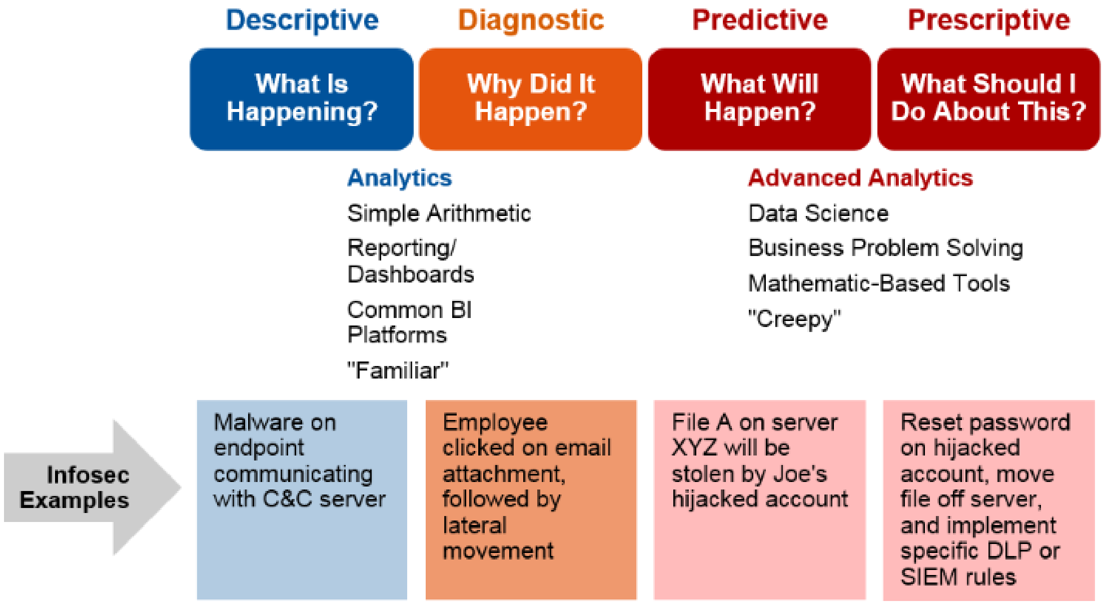
Essentially, differential privacy is a technique that was developed in 2006 and is generally applied to "privacy" use cases. However, once differential privacy is applied, it becomes nearly impossible to identify a single individual based on their usage patterns. In 2016, Apple announced it will adopt differential privacy as a means to collect data on individuals, mainly to improve the user experience, which will anonymize the data. Differential privacy is used to discover usage patterns for a large number of users, then it adds "mathematical noise" to a small sample of users within the group. As more users share the same pattern, patterns become "general." Hence, the difficulty in identifying a single individual/entity increases. While this technique is still unproven, it is potentially promising for use in data protection to improve the privacy of individuals and their data.

4.3. Security Wakes Up to Advanced Analytics

Advanced analytics have been incorporated into fraud detection applications since at least 1993, when credit card systems started using neural networks to detect fraud. In contrast, security methods used to detect attacks and protect assets and information inside an organization's network have been largely rule-based. Advanced analytics using machine learning, especially for detection, have been lacking in the world of enterprise security until the past few years, when organizations realized their security systems were struggling, if not failing, to stop the onslaught of breaches.

It is important to note that applying advanced analytics to security use cases is more difficult than it is with fraud use cases. That's because there are many more variables and possible relationships to look at in security than there are in fraud. For example, with payment card fraud, the relevant variables are relatively finite and limited to account holder, account attributes, account history, purchase amount, purchase type, purchase location and merchant. With security use cases, such as insider threats, an employee can engage in numerous activities from numerous devices that access multiple files on multiple servers, each of which has its own set of variables to inform the analysis.

Beginning in 2010,[15] new security vendors started introducing products with advanced analytics, and user, entity and peer group profiling. The results are generally better detection capabilities with lower false-positive rates and higher staff productivity than traditional rule- and signature-based security detection products provide. The pace of change and innovation in advanced analytics for security continues to accelerate, so that by 2018, 25% of security products used for detection will have some form of machine learning built into them.

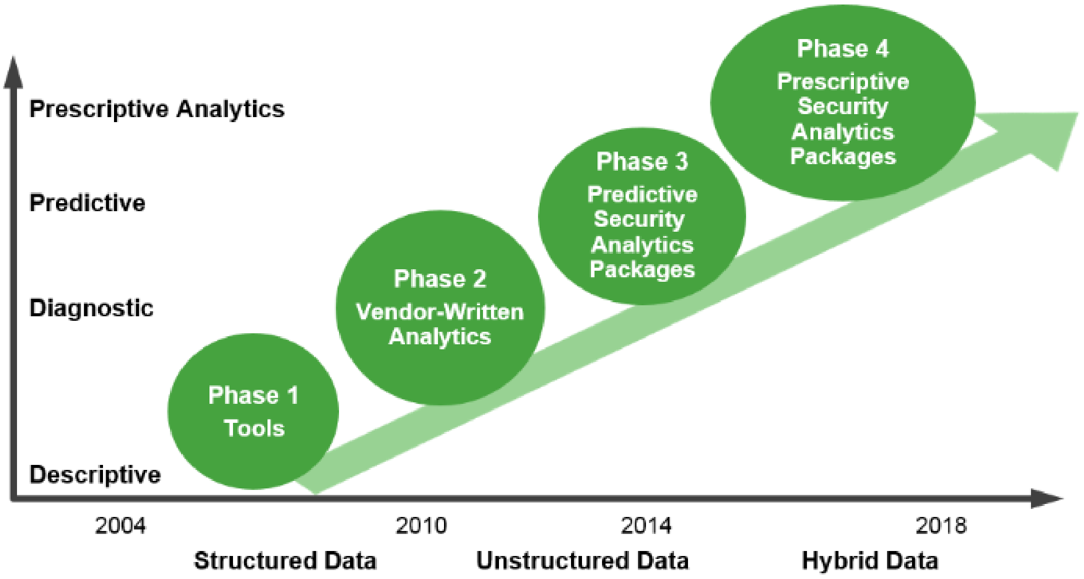

Descriptive
Diagnostic
Predictive
Prescriptive
What Is Happening?
Why Did It Happen?
What Will Happen?
What Should I Do About This?
Analytics
Simple Arithmetic
Reporting/ Dashboards
Common BI Platforms
"Familiar"
Advanced Analytics
Data Science
Business Problem Solving
Mathematic-Based Tools
"Creepy"
Infosec Examples
Malware on endpoint communicating with C&C server
Employee clicked on email attachment, followed by lateral movement
File A on server XYZ will be stolen by Joe's hijacked account
Reset password on hijacked account, move file off server, and implement specific DLP or SIEM rules

Source: Gartner (April 2016)

Figure 23: Advanced Analytics Defined through Infosec Examples

As noted in above schema, analytics capabilities become more advanced as products and functions move from the left of this chart to the right. Today, most security analytics products have descriptive and diagnostic analytic capabilities. Several user and entity behavior analytics (UEBA) products support predictive analytics by indicating, through a high risk score, that a specific event is likely to lead to an unauthorized or malicious activity. Gartner has not yet seen packaged prescriptive analytics (that is, prescribing what to do about it) deployed, but expects them to be introduced by 2018.

At the same time, the types of data ingested by analytic packages are becoming more complex, evolving from structured to hybrid data containing text, objects and things. Below schema, on evolution of advanced security analytics, shows how the market is evolving from being only a set of tools and analytics that vendors wrote on a customized basis for their clients, to packaged applications employing predictive and prescription analytics.



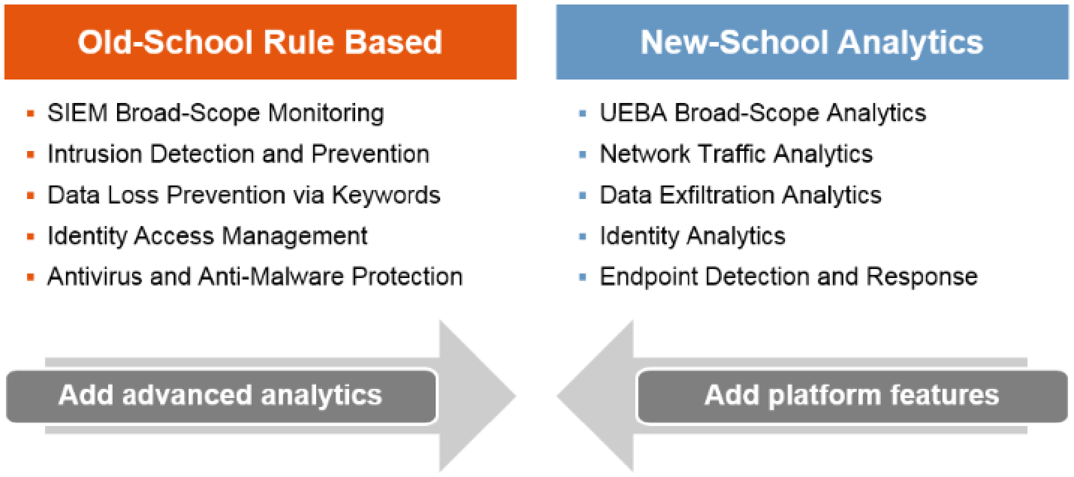
Source: Gartner (April 2016)

Figure 24: Emerge of Advanced Security Analytics

4.4. Old-School Versus New-School Security Products

For at least the past two years, Gartner has witnessed many new vendors with advanced analytics appear in several security market segments. One area that has spurred a lot of innovation is UEBA, which enables broad-scope security analytics, much like security information and event management (SIEM) enables broad scope security monitoring. UEBA provides analytics around user behavior, but also around other entities such as endpoints, networks and applications. The correlation of the analyses across various entities makes the analytics' results more accurate and threat detection more effective, just as it does with SIEM.

There also has been substantial analytics product development and innovation in other discrete segments of the security market (Scema 25).



Old-School Rule Based

- SIEM Broad-Scope Monitoring
- Intrusion Detection and Prevention
- Data Loss Prevention via Keywords
- Identity Access Management
- Antivirus and Anti-Malware Protection

New-School Analytics

- UEBA Broad-Scope Analytics
- Network Traffic Analytics
- Data Exfiltration Analytics
- Identity Analytics
- Endpoint Detection and Response

Add advanced analytics

Add platform features

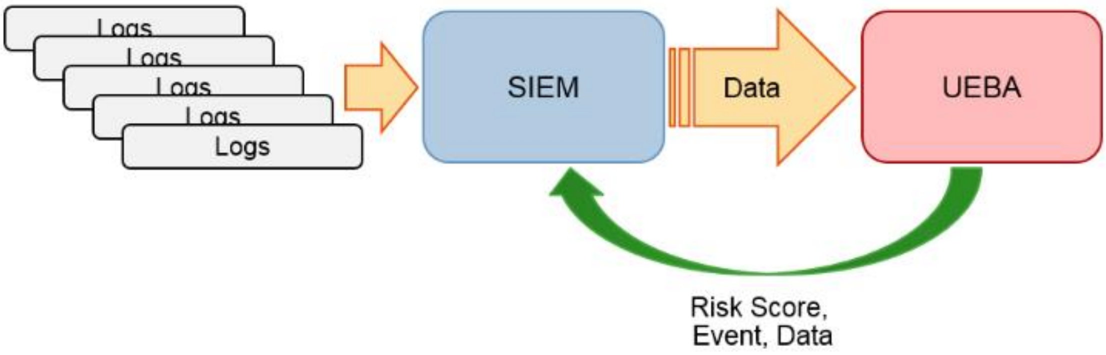
Source: Gartner (April 2016)

Figure 25: Old School vs New school Security Products

Some vendors in the "old school," largely rule-based, markets are adding advanced analytics approaches into their products, either through native development, acquisitions or OEM relationships. Conversely, some of the "new school" analytics vendors are adding platform features found in the old-school products, and competing with them directly.

SIEM and UEBA are examples of where these old-school and new-school capabilities are presently being utilized in combination as a "force multiplier" to better detect threats — such as advanced external attackers and threats from insiders — while reducing the daily operational overhead (for example, monitoring analysts and reducing the daily operational overhead incurred from staff who monitor security operations or tune rules). UEBA products typically need a data source, and SIEM products are commonly the central aggregation point for security logs for an organization. SIEM and UEBA tools complement each other in several other ways. UEBA can profile a user over a long period of time (and SIEM is not optimized for that). However, SIEM is broader in scope and stores the detail that is needed for incident investigation.

Another way in which these functionalities are complementary is that SIEM products are commonly the central monitoring and alerting tool for security analysis, such that a UEBA (or other behavioral analysis) tool-generated alert can be sent into the SIEM system for enrichment with events from other sources, and then the alert can be presented to an analyst for triage (Schema 26).



Source: Gartner (April 2016)

Figure 26: Interrelationship of SIEM and UEBA

4.4.1. Intrusion Prevention System (IPS) and Network Traffic Analytics (NTA)

Intrusion prevention is an established market in security that has undergone considerable disruption in the past few years. Year over year, more buyers are moving away from stand-alone IPSs to integrated platforms (such as a next-generation firewall [NGFW]). Many of the major providers have been acquired or sold off to other providers, and the market growth is moving toward decline. However, IPS still remains a budgeted line item technology for approximately 40% of security buyers, and a strategic tool for many customer security operations centers (SOCs). Analyzing network traffic to discover threats isn't a new concept; however, it's an area that most IPS providers haven't fully realized given the advancements in analytic sciences. It's likely that leading IPS providers will incorporate new-school network traffic analysis to further differentiate their technologies. NTA providers may expand their scope to meet IPS market definitions, and pursue buyers with existing IPS budgets; although, this is unlikely since IPS often focuses on the network perimeter, while NTA focuses on insider and egress traffic.

Indeed, many NTA vendors focus their efforts at where today's threats hide — inside the perimeter — and not at the Internet-exposed assets. However, NTA tools' use of advanced analytics can potentially make them into "a better intrusion detection system (IDS)" for threats inside and across the perimeter.

4.4.2. Endpoint Protection Platform (EPP) and Endpoint Detection and Response (EDR)

The EDR market is good example of a market taking advantage of detection gaps in established technologies (like EPP) to gather buyer attention. Although the scope and approach vary through the field of EDR vendors, many EDR tools use behavioral analytics to investigate hosts or applications (even though others remain merely search tools for threats).

The advantage to such analytical EDR systems is they can propose to find threats otherwise missed by the field of EPP tools and network security controls. EDR tools also enable better and faster security incident response. As attractive as this new detection capability is, the problem remains that the EPP itself is a platform including other feature beyond threat detection, such as whole-disk encryption, personal firewall, vulnerability assessment and data loss prevention (DLP) functionality. This platform gap makes it such that an EDR solution does not replace all the functions of an EPP. Gartner anticipates that many of the leading EPP providers will incorporate EDR functionality into their platforms as a way to further differentiate in the EPP market.

4.4.3. Data Loss Prevention (DLP) and Data Exfiltration Analytics

DLP is an established market in security, filled with long-standing providers that offer platform solutions to help buyers detect, identify, monitor and/or control the flow of their sensitive data . Despite the DLP market's maturity, there remains some fundamental gaps in capabilities — mainly the ability to identify and classify data that misses canned "detection filters." For example, certain patterned or keyword-based data, such as medical records, is relatively easy to identify; however, what about a personally written memo from the head of R&D about the five-year future of product innovation?

Data exfiltration providers seek to use analytic sciences to bridge that identification gap. Instead of trying to identify the document via patterns or filters, they can instead use analytics to watch other indicators, like data movement, activity, popularity and the like. Other UEBA providers incorporate even more watch factors about user profiles, such that they can be linked to data (documents) in order to better identify, or set, risk thresholds. UEBA deployment "on top of" traditional DLP is also not uncommon; in this scenario, the UEBA tool is used as an "add-on brain" for DLP tools and practices.

Given the range of functionality of leading DLP platforms, Gartner anticipates that many new-school data and/or exfiltration providers will become attractive partners for DLP providers.

4.4.4. Identity and Access Management (IAM) and Identity Analytics

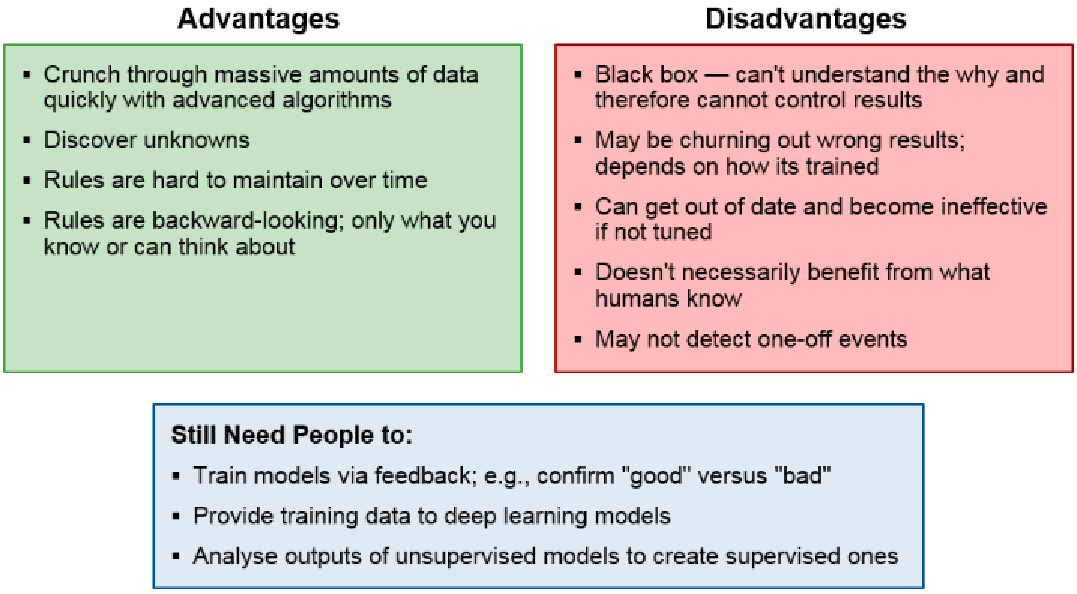
Identity analytics is the discipline that applies logic and science to identity and access data to provide insights for making better IAM decisions. Identity analytics delivers intelligent interactive and actionable analytics, so that IAM administrators can quickly identify risks as well as the origin of the risk, and modify existing access policies as appropriate. Identity analytics closes the loop between administrative controls and runtime activities.

UEBA tools have been used for identity analytics projects, but unlike threat detection scenarios, their algorithms were aimed at optimizing the identity management infrastructure or refining the access rights and privileges given to the users.

4.5. What's in Store for the Future?

Advanced analytics are not a panacea. Simply speaking, they raise the bar for detection systems by reducing the signal-to-noise ratio. Certainly, it will be harder for a malicious actor or process to escape detection when advanced analytics systems are being used instead of only rules. But if the past is any indicator of the future, criminals and other "bad actors" will study how advanced analytics and profiling systems work, and figure out how to beat them as well. Generally, that will involve a lot of surveillance and testing on their part. And if they can't beat the system detection, they will figure out how to socially engineer employees, contractors, partners, customers and other individuals associated with their target organization to perpetrate their crimes.

But certainly, advanced analytics will make an organization's detection capabilities much more effective, and will improve staff productivity and reduce the time to investigate and respond to a security incident. It's important to remember that machine learning has its pros and cons, and people will always be needed to manage security operations and systems. The nature of their work will change, however, as security systems are increasingly automated. See Schema 27 for a list of pros and cons of machine learning, and areas where humans will still be needed. [15]



Advantages

- Crunch through massive amounts of data quickly with advanced algorithms
- Discover unknowns
- Rules are hard to maintain over time
- Rules are backward-looking; only what you know or can think about

Disadvantages

- Black box — can't understand the why and therefore cannot control results
- May be churning out wrong results; depends on how its trained
- Can get out of date and become ineffective if not tuned
- Doesn't necessarily benefit from what humans know
- May not detect one-off events

Still Need People to:

- Train models via feedback; e.g., confirm "good" versus "bad"
- Provide training data to deep learning models
- Analyse outputs of unsupervised models to create supervised ones

Source: Gartner (April 2016)

Figure 27: Advantages & Disadvantages of Machine Learning

Notwithstanding the information in Figure 5, it is clear that the security market and its various domains and products will be adopting more and more advanced analytics capabilities in the coming years.

4.5.1. Analysis & Predictions

By 2018, at least 50% of major SIEM vendors will incorporate UEBA functionality into their products through acquisitions or native development. In 2015, we have already seen in Splunk acquire UEBA vendor Caspida and HP ArcSight acting as an OEM for UEBA products from Securonix.[15]

By 2018, 25% of security products used for detection will have some form of machine learning built into them. We are already seeing SIEM vendors, like LogRhythm and RSA, The Security Division of EMC, integrate advanced analytics into their portfolios.[15]

By 2018, prescriptive analytics will be deployed in at least 10% of UEBA products to automate incidence response, up from zero today. We are already seeing UEBA and SIEM vendors integrate with incident response systems such as IBM's Resilient Systems, FireEye Invotas and Hexadite, and this integration will be much tighter and more fully automated when UEBA products adopt prescriptive analytics.[15]

By 2020, sophisticated criminals will be able to beat 80% of the organizations who have deployed advanced analytic systems. We have seen incidents in the fraud domain, where advanced analytics have been used for years, for example, in online banking. In these domains, we see the criminals migrating their attack tactics so that they are increasingly socially engineering customers, partners and other members of a bank's ecosystem to perpetrate their crimes. This, no doubt, will happen in the security domain as well, so that criminals evade the analytics systems altogether by operating outside their purview.[15]

5. MARKET

Big data can offer accountants and finance professionals the possibility of reinvention, the chance to take a more strategic, ‘future-facing’ role in organizations. The transition, however, will not be easy. The accountants and finance professionals who differentiate themselves will be those who develop new skills and new ways of thinking, and who form new collaborations and partnerships. The vast amount of data continually collected, stored and transferred by technologies is changing the priorities of businesses and posing important questions for their leaders. How can diverse, disparate and often amorphous datasets be managed profitably and responsibly? Get big data right and it will facilitate ways of improving performance and productivity, and creating new wealth for shareholders and stakeholders. Get it wrong and the result will be poor decisions, breaches to data security and privacy codes, damage to organizational reputation and brand, and destroyed value. Data management is becoming a business-critical function as leaders seek ways to use the resource of big data strategically and unlock the insights that transform companies – without threatening their relationships with customers or exposing themselves to unacceptable risks. The market for big data analytics is, unsurprisingly, growing rapidly, forecast to reach US$23bn by 2016 (IDC 2013).

The open-source software movement and the software industry have developed solutions such as new programming models and new suites of data tools.1 Combined with increases in processing power, these solutions make it possible to synthesize vast amounts of information with previously unimaginable speed and accuracy, but they are only part of the answer. Managing big data effectively requires the right people. This has far-reaching implications for the accountancy and finance professions.

5.1. Opportunities & Challenges

The paradox of new technology is that it offers the chance to replace the value lost as it commoditizes traditional skills. Advances in automation, such as self-service data retrieval, are freeing accountants and the finance function from the more routine aspects of internal reporting and compliance work – and creating opportunities for them to alter their profile in business radically. Trained to gather, analyses and benchmark information and to use data in modeling and forecasting, accountants and finance professionals can provide a new and critical service: making big data smaller, ‘distilling’ vast amounts of information into actionable insights. Responsible for the ‘integrity’ of reports and accounts, they can help act as custodians of non-financial datasets and set quality and ethical standards for the information used in making strategic decisions and for that sold to third parties. This role will become increasingly important as more companies look for ways of developing new products and services from data they generate and own, particularly within the context of growing concerns around privacy and ethical data usage.

Big data offers the finance professional the possibility of moving into a more strategic, proactive role in business. It is important, however, to understand the realities of what it means: the opportunities are matched by the challenges. To differentiate themselves in the marketplace in the next 5 to 10 years and turn big data to their advantage, accountants and finance professionals will need to do three things.

- Develop methods and services for the valuation of data – and extend their role in compliance and internal control to the ethical and effective stewardship of data assets.

- Use big data to offer more specialized decision-making support – often in real time – and decide when data can most usefully be shared with internal and external stakeholders or ‘monetised’ as new products.
- Use big data and its associated tools not only to identify risks in real time and improve forensic accounting but also to evaluate the risks and rewards of long-term investment in new products and new markets.

5.1.1. A New Professional Agenda

The opportunities and challenges suggest three imperatives in the next 10 years, ie those of:

- developing new metrics
- learning new analytical skills
- creating a visual language of data ‘art’

Combined, these imperatives make up a new professional agenda. Accountants and finance professionals must find ways not only to measure big data as an organizational asset but also to use it as a measure of organizational performance. The trend towards integrated reporting (IR) and the inclusion of non-financial ‘capitals’ in company reports and accounts makes adopting this approach all the more urgent. It will increasingly be necessary to combine ‘hard’ financial data with ‘softer’ and non-financial datasets to provide the bigger picture of performance.

Meanwhile, there will also be requirements to extract value from big data through advanced analytics – and to interpret the meaning of big data in ‘visual language’ that can be used in company dashboards, decision-making ‘cockpits’ and information ‘hubs’. The accountants and finance professionals who succeed in the future will form a bridge between data science and data art, combining analytical skills and sophisticated models developed by mathematicians and statisticians with the skills of data art and data ‘storytelling’. They will collaborate closely with the IT and information management departments in cross-functional and multidisciplinary teams: the future could see the emergence of a new professional ‘hybrid’, the chief financial technology officer (CFTO) or chief financial information officer (CFIO). Most importantly, they will form partnerships with senior leaders in the development of strategy and the management of risk – and provide a service critical to the future of business.

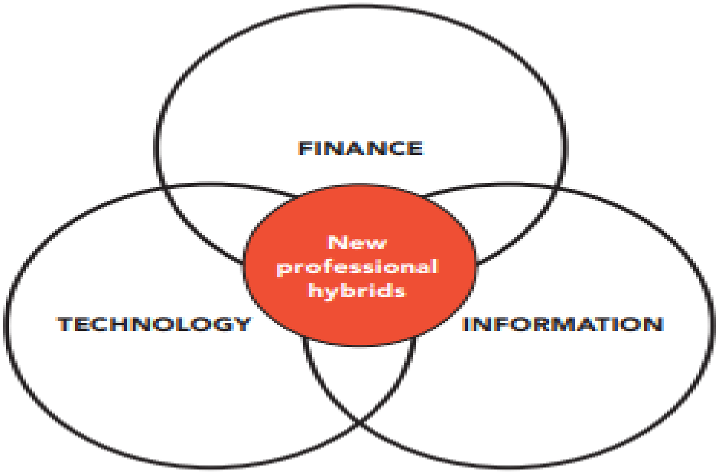
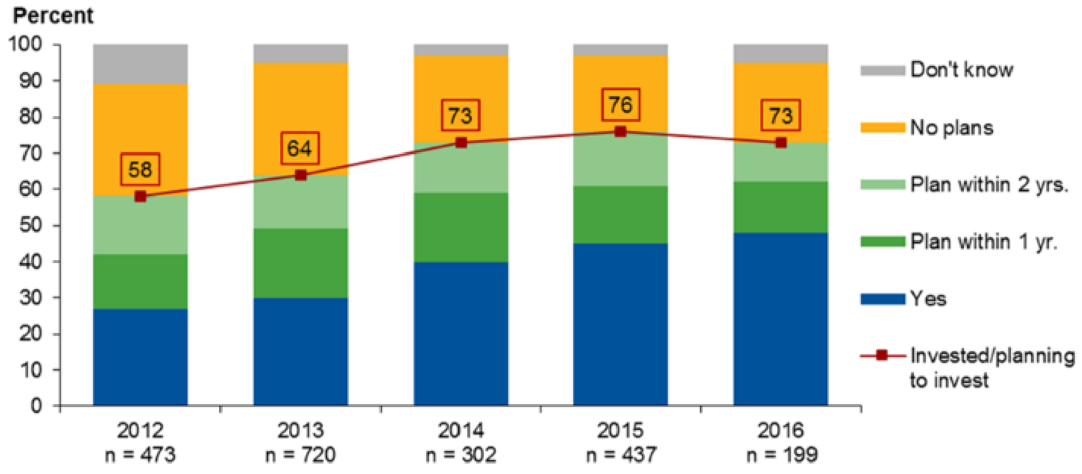


Figure 28: New Accounting & Finance Professional Hybrids

5.2. Big Data Investments Begun Tapering in 2016

The survey data shows the continued shift of focus from big data as a topic unto itself toward standard practices and business impact. The big issue is not so much big data itself, but rather, how it is used. Organizations have advanced beyond treating new data sources and technologies as unfamiliar. They continue to shift focus from ad hoc technology investments to infrastructure-level deployments that deliver on strategic business needs and force enterprise architectures to evolve. There is also an understanding that big data is not just about a specific technology. Rather, big data is a collection of different data management technologies and practices to support multiple analytics use cases.

Big data investment continues, but is showing signs of tapering. Overall, 73% of respondents say their organizations have invested, or have plans to invest, in big data. This number is down slightly from 2015, when the number was 76% (see Figure 1). Nearly half (48%) of respondents say their organization has already invested in big data, a three percentage point increase over 2015. In addition, 25% have plans to invest within two years, down from 31% in 2015. Those with no plans to invest stayed steady at 22% (21% in 2015). [17]



Base: Total respondents by survey year

Q. Which of the following best describes your organization's stage of big data adoption?

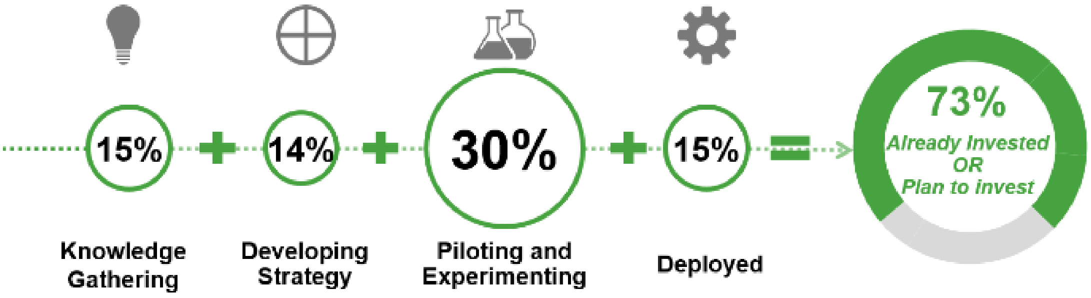
Source: Gartner (September 2016)

Figure 29: Big Data Investments Trending

There are a few reasons why big data investment is tapering. First, calling any project a "big data" project is no longer helpful. Companies are moving from vague notions of data and analytics to specific business problems that can be addressed with data. Second is the difficulty in getting big data projects to production. While 73% of respondents say their organization has invested or is planning to invest in big data, many remain stuck at the pilot stage (see Figure 2). Only 15% of businesses report deploying their big data project to production, effectively unchanged from last year (14%). Only 15% of respondents report their organization has deployed their big data project to production, effectively unchanged from last year (14%)[17]. Based on inquiries with Gartner clients, there are several reasons why big data efforts fail to make it to production:

- Big data projects receive less spending priority than competing IT initiatives (see Schema 31).
- There is a lack of effective business leadership or involvement in data initiatives. This challenge is also reflected in this year's survey data (see Schema 32). Pilots and experiments are built with ad hoc technologies and infrastructure and not

created with production-level reliability in mind, or there is a gap in how Mode 2 exploration translates to Mode 1 production



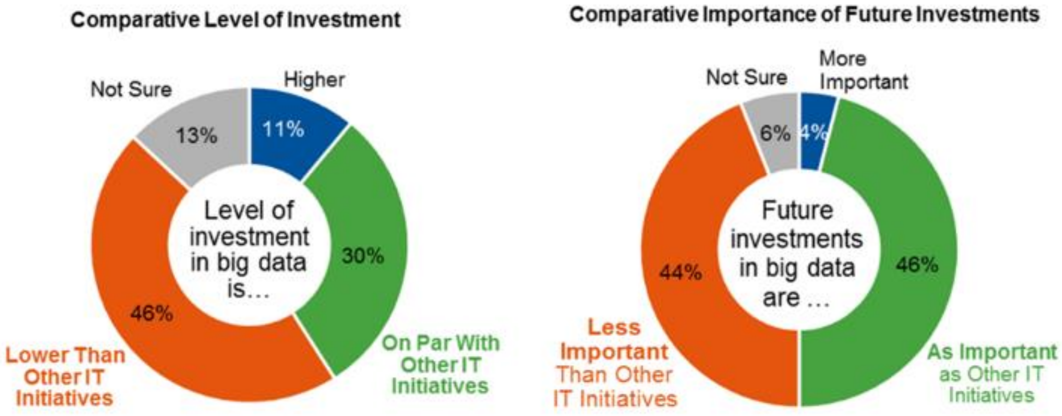
Q. Which of the following best describes your organization's stage of big data adoption?

Total respondents 2016 (n = 199)

Source: Gartner (September 2016)

Figure 30: Big Data Deployment Stage

Survey data indicates that organizations may be prioritizing other IT investments over big data initiatives. One reason is likely due to many big data projects not having a tangible ROI that can be determined upfront. Other reasons could be that the big data initiative is a part of a larger funded initiative. This will be more common as the term "big data" fades away and dealing with larger datasets and multiple data types continues to be the norm.



Base: Already invested in big data (n = 96)

Base: Planning to invest in big data (n = 50)

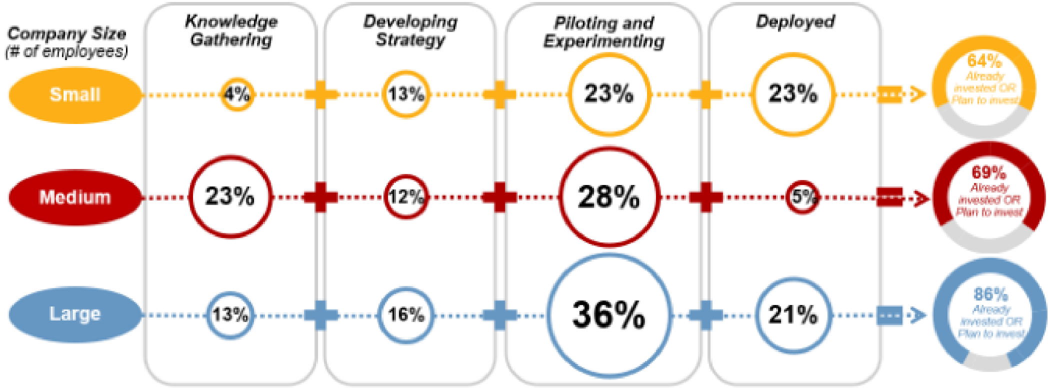
Q. Is your organization's business investment in big data lower or higher than investment in other IT initiatives?

Q. Is future investment in big data seen by your organization to be more important or less important than other IT initiatives?

Source: Gartner (September 2016)

Figure 31: Comparative Level of Investment

This year, we also compared big data deployments by size of organization. Small organizations are those with up to 1,000 employees, midsize organizations have between 1,000 and 9,999 employees, and large organizations are those with over 10,000 employees. While respondents working within large organizations indicate their organizations have been the pathfinders in investing in big data (86%), responses also indicate that both medium and small organizations are not that far behind in their investments (69% and 64%, respectively). This highlights that organizations of all sizes see value in trying to improve insights using more and varied data. Open-source and cloud options are also making it easier for smaller organizations to get started and to experiment with big data.


Company Size
(# of employees)
Knowledge Gathering
Developing Strategy
Piloting and Experimenting
Deployed
Small
4%
13%
23%
23%
64% Already invested OR Plan to invest
Medium
23%
12%
28%
5%
69% Already invested OR Plan to invest
Large
13%
16%
36%
21%
86% Already invested OR Plan to invest

Base: Total respondents: Small organization less than 1,000 employees (n = 47); Medium organization 1,000-9,999 employees (n = 74); Large organization 10,000-plus employees (n = 70)

Q. Which of the following best describes your organization's stage of big data adoption?

Q. Has your organization already invested in technology specifically designed to address the big data challenge?

Note: Values may not add up to 100% due to rounding.

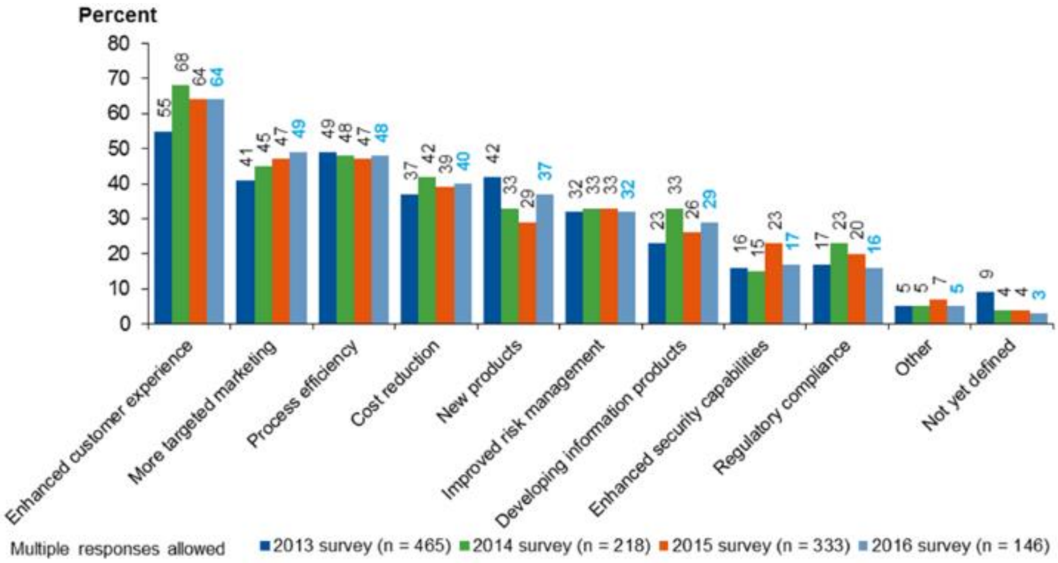
Source: Gartner (September 2016)

Figure 32: Big Data Deployment Stage by Company Size

Interestingly, both small and large companies have more success deploying their big data efforts than midsize companies. Relative to small and large organizations, a larger percentage of respondents from midsize enterprises indicate their organizations are still knowledge gathering and piloting while their peers are moving forward with deployment. Midsize organizations might be lagging because the benefits of big data may be less apparent. Another reason could be that other IT projects simply have higher priority over big data initiatives.

The good news is that the barrier to entry to deploy big data is lower than ever thanks to the availability of robust open-source technologies. Small and midsize organizations now have access to a powerful array of the best big data infrastructure that money doesn't need to buy.

With minor variation, the reasons organizations invest in big data haven't changed in the four years we have conducted this survey. Enhancing customer experience remains the top reason for investing at 64%, followed by more targeted marketing (49%) and improving process efficiency (48%).



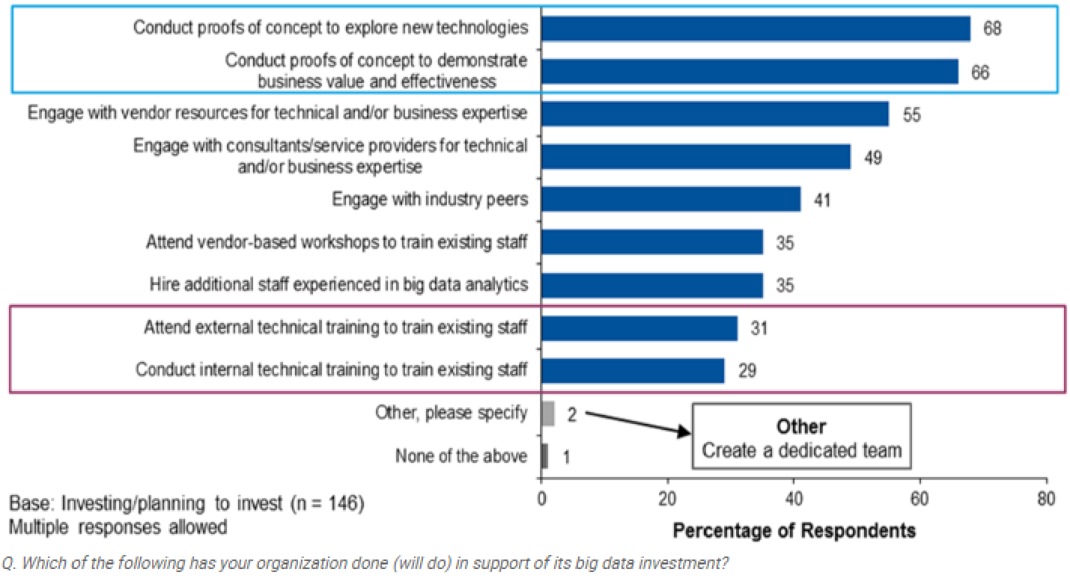
Source: Gartner (September 2016)

Figure 33: Big Data Opportunities

However, over the last four years, Gartner inquiries for these types of data and analytics efforts have focused on how organizations can do these things themselves. For example, many organizations are trying to build their own customer 360 hubs or build cybersecurity solutions on new infrastructure instead of simply engaging with vendors that are already offering purpose-built applications and infrastructure. Based on inquiries, many organizations are still at the crafting stage. Industrialization, and the performance and stability guarantees that come with it, have yet to penetrate big data thinking.

5.2.1. Skills Continue to Challenge Investing Organizations

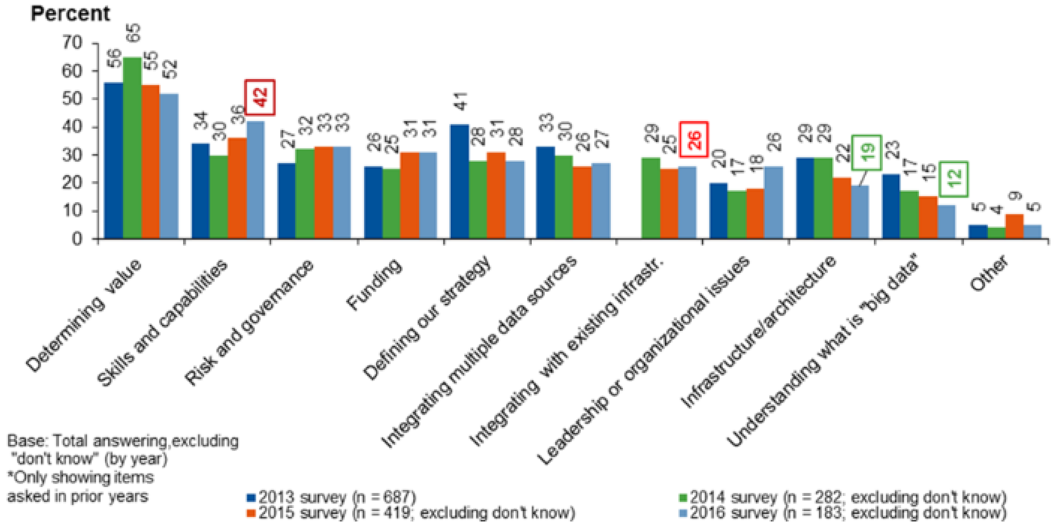
Regardless of company size, finding and developing skills continues to be a top challenge for companies either making or planning to make investments in big data. This year, 42% of respondents reported skills and capabilities as a key challenge for big data adoption (Schema 35). This is a six percentage point increase over last year.



Source: Gartner (September 2016)

Figure 34: Big Data Investment Support

Enterprises aren't clear on how best to cultivate big data skills. When we asked what organizations were doing to support big data investment, most responded that they were conducting proofs of concept (POCs) for new technologies or to demonstrate business value and effectiveness (68% and 66%, respectively; Schema 35). This is understandable given the top challenge for big data remains simply determining its value (52% agreeing; see Schema 36). However, performing a POC in a low-skill environment isn't guaranteed to provide results. And based on survey results, companies aren't planning on training existing staff. For those planning to or already investing in big data, only 31% plan to or are using external training and only 29% look to internal training (Schema 35).



Q. What are your organization's top 3 hurdles or challenges with big data?

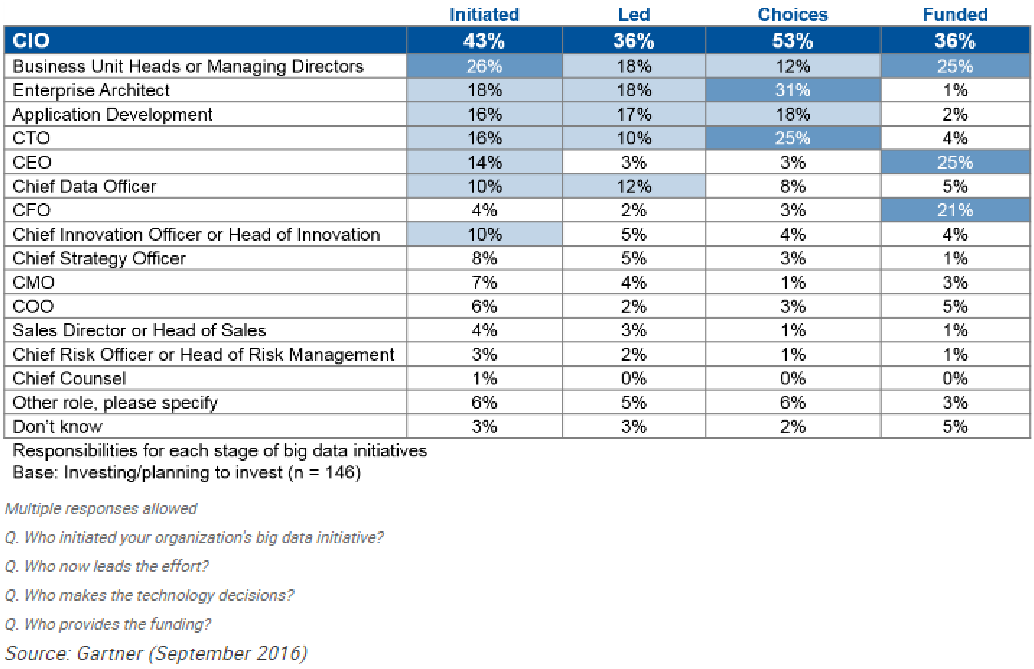
Source: Gartner (September 2016)

Figure 35: Big Data Challenges, Trending

What about hiring the necessary skills? This is understandably difficult given the shortage of data and analytics talent, but 35% report their organization is either hiring or planning to hire (Schema 35). Additionally, about half of respondents report that their organization engages with vendors or service providers for the required expertise. However, this can be a risky strategy. As more businesses shift to the digital era, companies must rely on internal expertise and business understanding. Over the long term, outsourcing big data skills, which will translate to general data and analytics skills, is effectively outsourcing the future of your business.

Over that last year, leadership and organizational issues were cited as an increasing challenge. This is up eight percentage points over 2015, from 18% to 26% (Schema 36). While still low compared with perennial challenges like risk, governance and funding, this increase may indicate a lack of business involvement or leadership for big data projects. It may also indicate conflict over who controls the budget for big data initiatives. Big data projects frequently reside within business units, but with support from IT. One party may control the budget and another the infrastructure and governance, resulting in friction between disparate groups (Schema 37). The CIO is the primary initiator and leader of big data efforts, and also makes the technology decisions, but funding is more distributed. The CIO funds 36% of big data projects, while business units and the CEO each fund projects at 25% and the CFO funds 21% of projects.

As other challenges decrease, like infrastructure architecture, it is likely that classic IT challenges around funding, governance and leadership will continue to increase.



	Initiated	Led	Choices	Funded
CIO	**43%**	**36%**	**53%**	**36%**
Business Unit Heads or Managing Directors	26%	18%	12%	25%
Enterprise Architect	18%	18%	31%	1%
Application Development	16%	17%	18%	2%
CTO	16%	10%	25%	4%
CEO	14%	3%	3%	25%
Chief Data Officer	10%	12%	8%	5%
CFO	4%	2%	3%	21%
Chief Innovation Officer or Head of Innovation	10%	5%	4%	4%
Chief Strategy Officer	8%	5%	3%	1%
CMO	7%	4%	1%	3%
COO	6%	2%	3%	5%
Sales Director or Head of Sales	4%	3%	1%	1%
Chief Risk Officer or Head of Risk Management	3%	2%	1%	1%
Chief Counsel	1%	0%	0%	0%
Other role, please specify	6%	5%	6%	3%
Don't know	3%	3%	2%	5%

Responsibilities for each stage of big data initiatives
Base: Investing/planning to invest (n = 146)

Multiple responses allowed

Q. Who initiated your organization's big data initiative?

Q. Who now leads the effort?

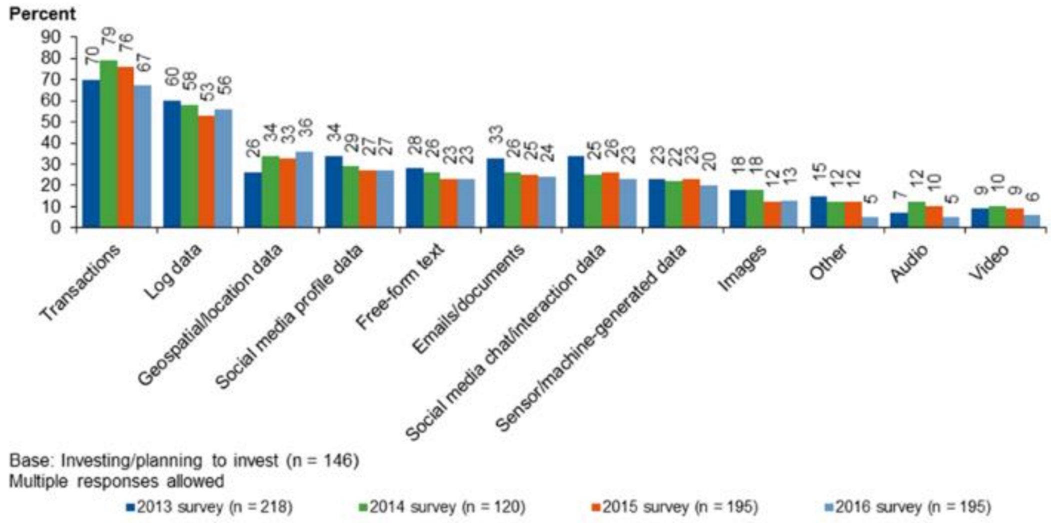
Q. Who makes the technology decisions?

Q. Who provides the funding?

Source: Gartner (September 2016)

Figure 36: Big Data Leadership

Transactions and log data continue to be the top two data types that are currently analyzed at 67% and 56%, respectively (Schema 38). But interest in analyzing geospatial/location data has increased over the past four years from 26% to 36%. In many ways, this makes sense given that enhancing customer experience and more targeted marketing are the top business problems addressed with big data (Schema 34). The smartphone is the ubiquitous personal device that can reveal our past and current location, along with our preferences. Geospatial/location data can be used to build customer profiles and notify organizations when customers are ready to buy. Log data can provide insights about the path the customers took on your website or e-commerce site, and transactional data provides insight if a marketing campaign results in a sale or about what the customer purchased in the past.



Base: Investing/planning to invest (n = 146)
Multiple responses allowed

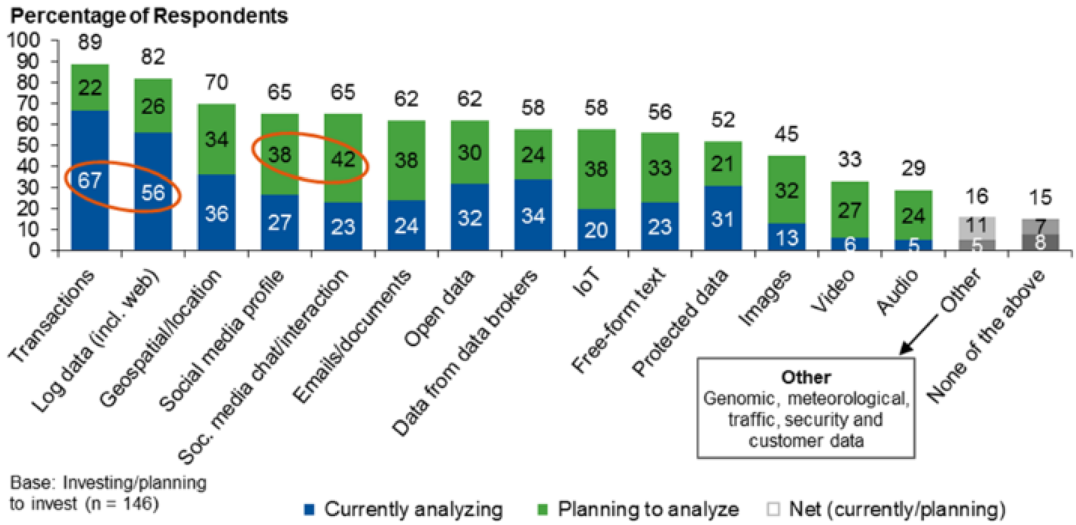
Q. Which types of big data does your organization currently analyze, and which do you plan to add to your analytics in the future?

Source: Gartner (September 2016)

Figure 37: Data Types

Interestingly, social media profile and interaction data, along with the Internet of Things (IoT), rank in the top for organizations that have not yet deployed big data (Schema 39).

This is likely due to the overall hype about "social business" continuing to subside and social analytics becoming entrenched in the practice of "business as usual." Social media engagement for customer service or marketing purposes is becoming critical to organizations as they look to progress their customer experience strategies to meet customers in their preferred communication channels. And, interest in IoT is growing by organizations planning to use big data as they experiment with deploying connected endpoint devices, each continuously generating data that potentially can help improve process efficiency (third-ranked business use), especially in asset-intensive industries.



Q. Which types of big data does your organization currently analyze, and which do you plan to add to your analytics in the future?

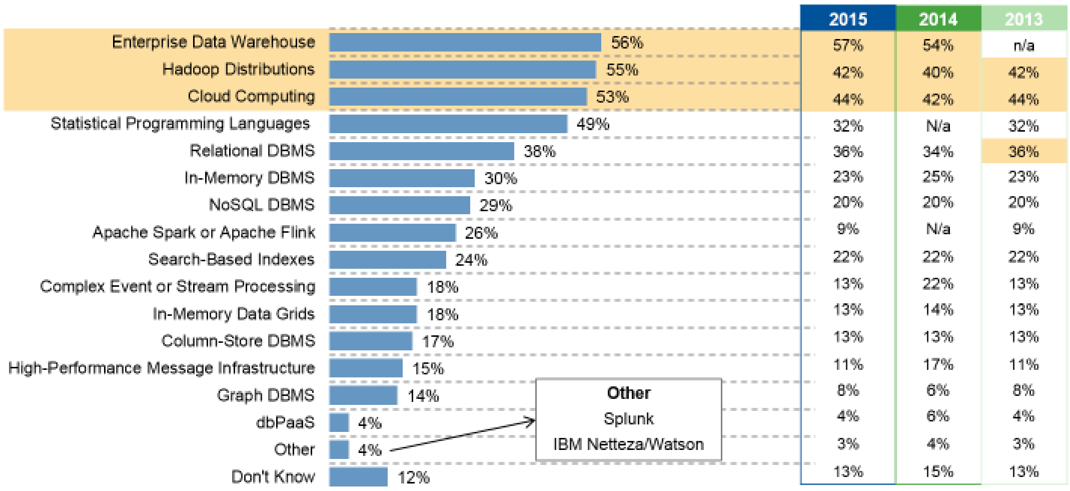
Source: Gartner (September 2016)

Figure 38: Current and Future Data Types for Analysis

More than 50% [17] of organizations use or plan to use an enterprise data warehouse (EDW), Hadoop distribution and/or cloud computing to derive value from big data (Schema 40). The EDW has consistently been the top technology used to store, analyze and manage big data for the past three years. But Hadoop and cloud computing are now equal in terms of technologies used for big data. This is likely the result of organizations finally realizing that they need a portfolio of capabilities to manage the different types and sizes of data rather than just one data management technology. Even though Hadoop is an open-source technology, it has become a mainstream capability for building data lakes to store and manage any type of data. And, the availability of high-performance compute power via public clouds like AWS and Azure makes it easier, faster and more affordable to explore and perform big data analytics, without having to procure and manage clusters of hardware inside the organization.

Interest in using statistical programming languages has grown year over year from 32% to 49% [17] by both organizations that have already invested as well as those planning to invest in big data. This is likely the result of organizations moving beyond just trying to store big data. Now, they are extracting insights by using data science languages like open-source R, Python or proprietary vendor tools to build models that can analyze the stored data and uncover patterns or trends.

Other DBMS technologies such as relational, in-memory and NoSQL are increasingly being used for big data projects. Both in-memory and NoSQL databases jumped nearly 10 percentage points because organizations understand the value of these technologies to solve specific big data analytics challenges, such as the need for faster performance. Apache Spark and Flink jumped from single digits to 26% because more organizations see value in using memory-centric data processing frameworks to more effectively run distributed analytic workloads (both batch and streaming) to support a variety of use cases.



		2015	2014	2013
Enterprise Data Warehouse	56%	57%	54%	n/a
Hadoop Distributions	55%	42%	40%	42%
Cloud Computing	53%	44%	42%	44%
Statistical Programming Languages	49%	32%	N/a	32%
Relational DBMS	38%	36%	34%	36%
In-Memory DBMS	30%	23%	25%	23%
NoSQL DBMS	29%	20%	20%	20%
Apache Spark or Apache Flink	26%	9%	N/a	9%
Search-Based Indexes	24%	22%	22%	22%
Complex Event or Stream Processing	18%	13%	22%	13%
In-Memory Data Grids	18%	13%	14%	13%
Column-Store DBMS	17%	13%	13%	13%
High-Performance Message Infrastructure	15%	11%	17%	11%
Graph DBMS	14%	8%	6%	8%
dbPaaS	4%	4%	6%	4%
Other	4%	3%	4%	3%
Don't Know	12%	13%	15%	13%

Base: Investing/planning to invest 2016 (n = 146); 2015 (n = 333); 2014 (n = 218); 2013 (n = 465)

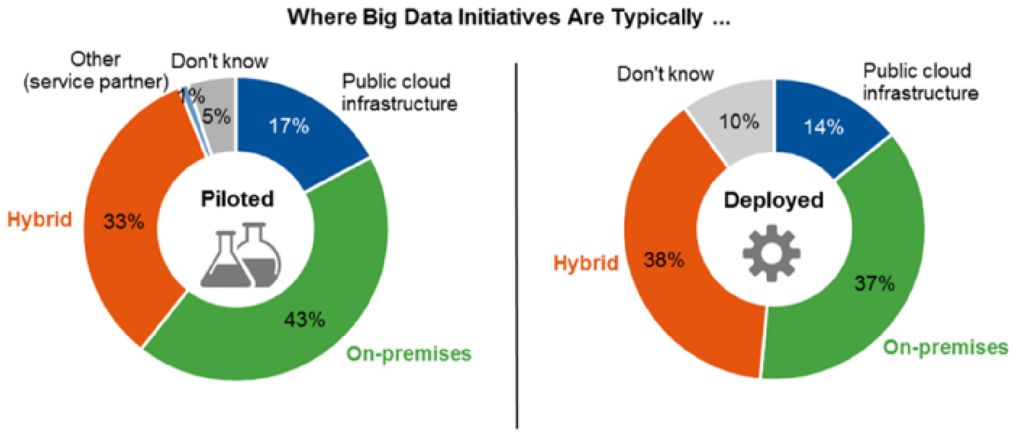
Multiple responses allowed

Q. What technologies is your organization using or planning to use to derive value from big data?

Source: Gartner (September 2016)

Figure 39: Technology Choices

When it comes to cloud vs. on-premises, the answer is primarily "on-premises." Of those that have deployed their big data project, 37% have deployed on-premises exclusively, with another 14% deploying exclusively on public cloud infrastructure. More interesting is that 38% of deployed organizations have deployed in a hybrid (cloud IaaS and on-premises) environment (Schema 41). [17] Because hybrid includes on-premises, the survey results show that about 75% pilot and/or deploy at least partially on-premises. Conversely, only about 50% use cloud at least partially. We expect the trend for hybrid deployments to increase as tooling improves.



Base: Investing/planning to invest (n = 146)

Q. Where does your organization typically pilot big data initiatives?

Q. Where does your organization typically deploy big data initiatives?

Note: Values may not add up to 100% due to rounding.

Source: Gartner (September 2016)

Figure 40: Deployment Environment Selection

6. BEYOND BIG DATA

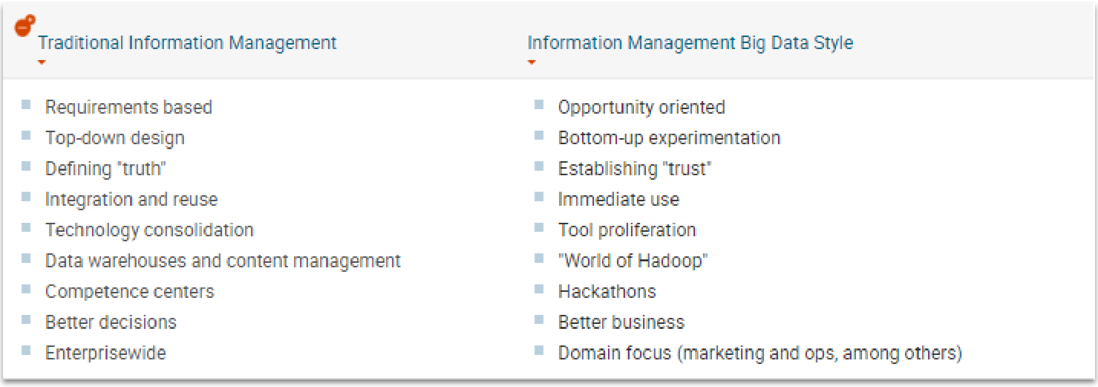
Information management, dominated with conceptual terminology, such as "metadata," "data warehousing" and "governance," has not been the hottest business topic. It's no surprise that information managers have not had the senior management attention they asked for and deserved. Instead, the discipline has often been relegated to the "IT basement" within the organization.

However, big data and the Nexus of Forces are making information more relevant to the consumer. In turn, this is significantly affecting how enterprises manage information. As a result, decision makers within organizations are paying attention to big data trends and instead of being relegated to the basement, information management is now even discussed in the boardroom in many organizations.

Information management has often been labeled as "infrastructure," in the sense of being part of "everything below the level I care about," but it has fast become core business for some organizations. This new-found attention requires information architects and leaders to reconsider their communication style, topics and methodologies and even address human emotion. For this reason, this analysis showcases a collection of research organized around conflict, excitement and fear.

6.1. Conflict: View Diverging Approaches as Complimentary

In many organizations, two schools of thought now offer competing approaches to information management for analytics



Traditional Information Management	Information Management Big Data Style
Requirements based	Opportunity oriented
Top-down design	Bottom-up experimentation
Defining "truth"	Establishing "trust"
Integration and reuse	Immediate use
Technology consolidation	Tool proliferation
Data warehouses and content management	"World of Hadoop"
Competence centers	Hackathons
Better decisions	Better business
Enterprisewide	Domain focus (marketing and ops, among others)

Source: Gartner (March 2014)

Table 1: Comparison of Traditional Information Management & the New Big Data Style

Traditional information managers have taken a strategic approach to managing information (with varying degrees of success) for the past 20 years. They understand the organization's corporate goals and translate them into performance indicators and relevant business questions. They know how to build data models to acquire and house all relevant information. They've also taken a deductive approach by creating data warehouses, going through two rounds of consolidating systems and working toward achieving one version of the truth.

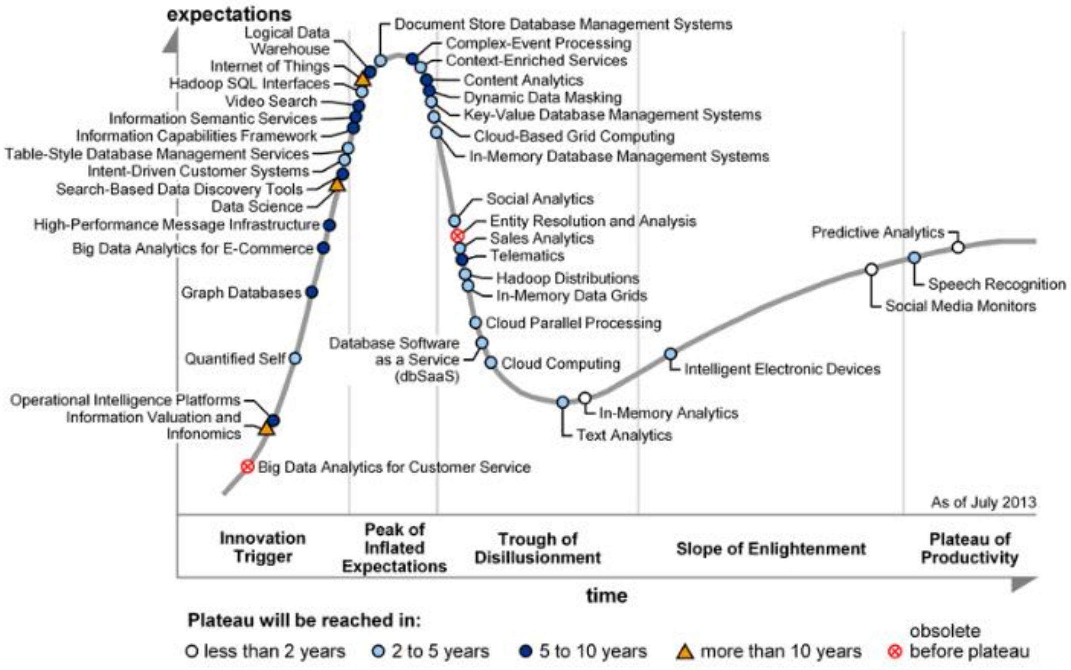
The newer big data-focused professionals do not believe in the idea of corporate strategy or a concrete business question as the starting point; they favor a bottom-up approach. They don't agree with the deductive approach of traditional information managers; instead, they favor an inductive approach. In other words, they ask: "What would the data tell us if it could talk?" With little concern for the so-called "single version of truth" endemic to business applications and data warehousing, the newer big data professionals are more interested in experimentation. They believe that the innovation is in the data itself and that exploring this will lead to new business strategies, new business opportunities and new business questions.

Traditional information managers fear this emerging approach because its focus on big data will engender fragmentation and escalated data ownership costs. As a result, they want to institute restrictive guardrails or halt its progress outright. At the same time, some proponents of the "new big data way" claim that their new technologies will replace the "old generation."

Both are wrong in their resolute and extreme approach. Success is in reconciling both approaches within the same strategic framework. Gartner is seeing more organizations well advanced in enterprise information management doing just that. New leadership roles are emerging. The number of large enterprises with chief data officers has doubled in the last year and 17% of CEOs say they will have established this role by 2014.

6.2. Excitement: Consider Multiple Dimensions and Sources of Big Data

Big data is one of the most hyped technology terms of 2013 and 2014 doesn't look much better — but there is substance behind the hype.



Source: Gartner (July 2013)

Figure 41: Hype Cycle for Big Data 2013

Clients ask Gartner how much of this hype the vendor is marketing versus the real uptake and investment by organizations in their industry. Investment numbers across industries are impressive and point to big data planning and investment activities beyond the hype.

Big data will grow past its hype toward 2016. Today's big data will become "just data" again once these data sources become commonplace, technologies mature, skills become more prevalent and organizations implement enterprise-class big data solutions.

However, big data isn't the only stream of innovation in information management. Businesses are feeling the impact and can take advantage of what Gartner calls the

Nexus of Forces — the convergence of mobile, social, cloud and information to create exciting new business opportunities. Managing information within the Nexus of Forces presents opportunities as well as challenges.

In addition, Gartner contends that information is an asset in its own right, as it has value and substance. Our research on the emerging concept of what we term "infonomics" shows how information can and should be managed, leveraged and quantifiably valued with the same discipline as actual balance sheet assets. In fact, Gartner predicts that by 2016, 30% of businesses will have begun directly or indirectly monetizing their information assets via bartering or selling them outright.

This is already happening in various industries, such as in commerce and the public sector and in large and small enterprises. For example:

- GE started an analytical software business unit that uses sensors and contextual information to help with fuel efficiency, more efficient operations and predictive maintenance. [18]
- John West (a manufacturer of tinned fish products), connects sensors to every batch of fish that it catches, so that consumers can track where the fish came from. [19]
- The Climate Corp (recently acquired by Monsanto), collects granular weather and soil data from around the world to create crop insurance products for small farmers in developing countries. [20]
- In the public sector, governments allow businesses and researchers to use public data to gain more insight into local markets.

6.3. Fear: Consider Privacy and Ethical Issues Around the Use of Data

In search of competitive advantage, aggressive innovation sometimes leads to unintended consequences. Companies may misuse information either accidentally or intentionally and this may involve a breach of privacy, leading to a consumer backlash, brand damage or legal action. Ignorance can never be a good excuse. Through 2016, 25% of organizations using consumer data will face reputation damage due to inadequate understanding of information trust issues.

Organizations should be aware of the limitations in the use of information and put the proper information governance in place. Information governance is moving from a control function to limit access to data, to a form of prioritization to maximize the asset value of information.

Information architects and leaders who don't want to be part of the 25% of organizations facing reputation damage can use the following research examples from industry to guide their decisions around information innovation:

- "Privacy and Ethical Concerns Can Make Big Data Analytics a Big Risk Too" uses real-world examples to explore the delicate balance between the benefits that big data analytics bring and the ethical and privacy risks they pose. Information leaders should use this research to initiate an internal debate on the limitations of big data analytics.
- "Big Data Analytics Requires An Ethical Code of Conduct" covers how ethical guidelines will help business analytics professionals avoid the unintended

consequences of powerful analytics. Information architects and leaders can follow these rules to do the right thing.

6.4. What Enterprise accomplish with Big Data

Despite all the hype surrounding big data, information leaders and analytics leaders are struggling to turn proofs of concept into demonstrable business value. Time is ticking and trust will disappear once big data hits the Trough of Disillusionment. Big data requires new skills and new technologies, both of which often meet resistance in organizations. Gartner has had several inquiries already in which big data teams complain that none of their new insights have really been implemented.

6.4.1. Dialogue with consumers

Today’s consumers are a tough nut to crack. They look around a lot before they buy, talk to their entire social network about their purchases, demand to be treated as unique and want to be sincerely thanked for buying your products. Big Data allows you to profile these increasingly vocal and fickle little ‘tyrants’ in a far-reaching manner so that you can engage in an almost one-on-one, real-time conversation with them. This is not actually a luxury. If you don’t treat them like they want to, they will leave you in the blink of an eye.

Just a small example: when any customer enters a bank, Big Data tools allow the clerk to check his/her profile in real-time and learn which relevant products or services (s)he might advise. Big Data will also have a key role to play in uniting the digital and physical shopping spheres: a retailer could suggest an offer on a mobile carrier, on the basis of a consumer indicating a certain need in the social media.

6.4.2. Re-develop your products

Big Data can also help you understand how others perceive your products so that you can adapt them, or your marketing, if need be. Analysis of unstructured social media text allows you to uncover the sentiments of your customers and even segment those in different geographical locations or among different demographic groups.

On top of that, Big Data lets you test thousands of different variations of computer-aided designs in the blink of an eye so that you can check how minor changes in, for instance, material affect costs, lead times and performance. You can then raise the efficiency of the production process accordingly.

6.4.3. Perform risk analysis

Success not only depends on how you run your company. Social and economic factors are crucial for your accomplishments as well. Predictive analytics, fueled by Big Data allows you to scan and analyze newspaper reports or social media feeds so that you permanently keep up to speed on the latest developments in your industry and its environment. Detailed health-tests on your suppliers and customers are another goodie that comes with Big Data. This will allow you to take action when one of them is in risk of defaulting.

6.4.4. Keeping your data safe

You can map the entire data landscape across your company with Big Data tools, thus allowing you to analyze the threats that you face internally. You will be able to detect potentially sensitive information that is not protected in an appropriate manner and make sure it is stored according to regulatory requirements. With real-time Big Data analytics you can, for example, flag up any situation where 16 digit numbers – potentially credit card data - are stored or emailed out and investigate accordingly.

6.4.5. Create new revenue streams

The insights that you gain from analyzing your market and its consumers with Big Data are not just valuable to you. You could sell them as non-personalized trend data to large industry players operating in the same segment as you and create a whole new revenue stream.

One of the more impressive examples comes from Shazam, the song identification application. It helps record labels find out where music sub-cultures are arising by monitoring the use of its service, including the location data that mobile devices so conveniently provide. The record labels can then find and sign up promising new artists or remarket their existing ones accordingly.

6.4.6. Customize your website in real time

Big Data analytics allows you to personalize the content or look and feel of your website in real time to suit each consumer entering your website, depending on, for instance, their sex, nationality or from where they ended up on your site. The best-known example is probably offering tailored recommendations: Amazon’s use of real-time, item-based, collaborative filtering (IBCF) to fuel its ‘Frequently bought together’ and ‘Customers who bought this item also bought’ features or LinkedIn suggesting ‘People you may know’ or ‘Companies you may want to follow’. And the approach works: Amazon generates about 20% more revenue via this method.

6.4.7. Reducing maintenance costs

Traditionally, factories estimate that a certain type of equipment is likely to wear out after so many years. Consequently, they replace every piece of that technology within that many years, even devices that have much more useful life left in them. Big Data tools do away with such unpractical and costly averages. The massive amounts of data that they access and use and their unequalled speed can spot failing grid devices and predict when they will give out. The result: a much more cost-effective replacement strategy for the utility and less downtime, as faulty devices are tracked a lot faster.

6.4.8. Offering tailored healthcare

We are living in a hyper-personalized world, but healthcare seems to be one of the last sectors still using generalized approaches. When someone is diagnosed with cancer they usually undergo one therapy, and if that doesn’t work, the doctors try another, etc. But what if a cancer patient could receive medication that is tailored to his individual genes? This would result in a better outcome, less cost, less frustration and less fear.

With human genome mapping and Big Data tools, it will soon be commonplace for everyone to have their genes mapped as part of their medical record. This brings medicine closer than ever to finding the genetic determinants that cause a disease and developing drugs expressly tailored to treat those causes — in other words, personalized medicine.

6.4.9. Offering enterprise-wide insights

Previously, if business users needed to analyze large amounts of varied data, they had to ask their IT colleagues for help as they themselves lacked the technical skills for doing so. Often, by the time they received the requested information, it was no longer useful or even correct. With Big Data tools, the technical teams can do the groundwork and then build repeatability into algorithms for faster searches. In other words, they can develop systems and install interactive and dynamic visualization tools that allow business users to analyze, view and benefit from the data.

References

[1] https://www.ericsson.com/assets/local/publications/white-papers/wp-big-data.pdf
[2] https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/big-data-meets-big-data-analytics-105777.pdf
[3] http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics
[4] https://www.ibm.com/analytics/us/en/technology/hadoop/big-data-analytics/
[5] https://en.wikipedia.org/wiki/Big_data
[6] https://www.sas.com/en_us/insights/analytics/big-data-analytics.html
[7] http://dbmanagement.info/Books/MIX/Bigdata_presentation_for_BIM_community_Big_Data.pptx
[8] http://www.oracle.com/us/products/oracle-big-data-executive-brief-2415577.pdf
[9] https://www.vertica.com/product/vertica-for-sql-on-hadoop/
[10] https://www.oracle.com/big-data/index.html
[11] http://bigdata.teradata.com/
[12] http://www.teradata.com/products-and-services/integrated-big-data-platform
[13] http://www.teradata.com/products-and-services/teradata-analytics-platform
[14] https://www.statista.com/statistics/254266/global-big-data-market-forecast/
[15] https://www.gartner.com/document/3274217/meter/charge
[16] https://www.gartner.com/document/3815368?ref=solrAll&refval=193495747&qid=5cd7747b7aa5e1ed080de539dc77e2f0
[17] https://www.gartner.com/document/3446724/meter/charge
[18] https://www.informationweek.com/it-leadership/ge-ceo-jeff-immelts-analytics-lessons-learned/d/d-id/1111893
[19] https://www.foodnavigator.com/Article/2012/11/27/John-West-rolls-out-can-tracking-scheme
[20] http://www.takepart.com/notfound.html