

WHITE PAPER

BIG DATA & ANALYTICS

Transforming the business and driving revenue through big data and analytics

Collection, storage and extraction of business value from data generated from a variety of sources are top priorities for enterprises. The path of transforming raw data into actionable insights necessitates a number of business and technical decisions, many of which can be challenging.

This paper highlights the key trends contributing to the explosion of big data and analytics, such as the widespread adoption and enhancements in sensors and online technology, rapid advances in storage and distributed computing and the increased awareness and application of large-scale machine-learning techniques. The paper also discusses important strategic and tactical considerations for business and IT leaders who are developing internal standards and processes for big data and analytics.

*Big Data Analytics Group
Verizon Corporate Technology, Palo Alto*

Introduction: Big Data and Value Generation

Big data is typically defined as data having one or more of the “4 Vs” – volume, velocity, variety and veracity. With an estimated 2.5 exabytes of data generated each day¹ from a variety of sources and often including a variety of formats, it is essential that enterprises understand and monitor trends in key areas of big data and analytics, particularly the explosion in online, mobile and sensor data, the development in storage and computing technology and advances in large-scale machine learning techniques.

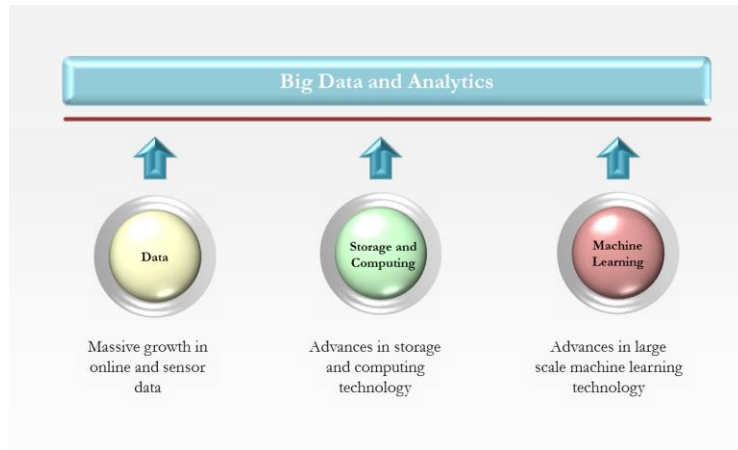


Figure 1. Three key trends in Big Data and Analytics

Need for Big Data and Analytics

Enterprises are rapidly moving toward leveraging data assets in new ways and pushing the limits on analytics to generate new revenue streams, reduce operational costs and enhance business decision-making.

Growth in Online, Mobile and Sensor Data

The past decade has seen an explosion of data generation, primarily from two main channels – online/mobile and data derived from sensors. Online and mobile sources comprise data such as ecommerce transactions, browsing behavior, search logs, clickstreams, location and social networks.

With more than 50 billion connected devices projected by 2020², new sources of machine data are quickly emerging in applications such as healthcare, utilities, transportation and city planning. The need for storage and computing for such massive amounts of data has led to many developments in infrastructure systems.

¹ <http://hbr.org/2012/10/big-data-the-management-revolution>

² http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

Advances in Storage and Computing Technology

The field of computing infrastructure and storage has rapidly evolved from centralized, structured, computing-as-a-product paradigm to distributed, structure-independent, computing as a service paradigm (Table 1).

Traditional Databases	Big Data Infrastructure
<ul style="list-style-type: none">• Shared systems• Centralized data storage• Vertical scaling• Computing as a product• Schema-based systems	<ul style="list-style-type: none">• Shared nothing systems• Distributed data storage• Massive horizontal scaling• Computing as a Service• Schema-less systems

Table 1. Traditional versus Big Data infrastructure

Distributed systems allow for inherently distributed computation, increased reliability, availability and fault-tolerance. Shared-nothing systems, a popular architecture for distributed computing, allows scaling as the nodes in the distributed system do not share memory or disk usage. Moreover, each node is often implemented with inexpensive commodity hardware, enabling cost-effective scaling. Such advances in distributed computing are being leveraged by technologies such as Hadoop to enable development of massively scalable and cost-efficient big data solutions. The Apache Hadoop project, for example, develops open-source software for reliable, scalable, distributed computing³. Hadoop is a software framework for distributed computing and includes modules such as a high-throughput distributed files system (HDFS) and a programming paradigm for parallel processing of large data sets (MapReduce). Hadoop delivers a highly available service that uses software, rather than relying on special hardware for high throughput and high availability.

The combination of distributed computing technology and cloud infrastructure has enabled the deployment of solutions with “pay-as-you-go” models, thereby greatly enhancing the capability of enterprises to access efficient solutions in a cost-effective manner. Business users and data scientists are no longer constrained to querying and drawing conclusions from a small subset of data.

Distributed Computing

Distributed computing has become increasingly common because of advances that have made both machines and networks cheaper and faster.

³ <http://hadoop.apache.org/>

Machine Learning

While reporting on big data can help gain new insights primarily due to access to a larger volume of data than before, machine learning focuses on building models that learn the properties of data and use that learning to predict future events.

Analytics and Large Scale Machine Learning

The big data and analytics team at Verizon is developing advancements in machine learning that take advantage of recent developments in distributed systems theory.

Reporting is one of the most important operational activities for any enterprise. Reporting usually emphasizes selecting and counting events of interest or testing hypothesis. Reporting from large volumes of data (“big data reporting”) can present insights on historical events that may have been previously unknown or unfeasible to detect from smaller data.

Analytics goes beyond traditional reporting to discover novel and hidden patterns in data, and to learn from large amounts of data for describing existing phenomena or predicting future events. Machine learning, an advanced area of analytics, is concerned with the development of systems that automatically learn from data to improve the system’s performance in a given set of tasks -- usually to predict new events based on learning from data. Advances in machine learning, such as deep learning, are now capable of learning information at multiple levels and modeling complex real-world phenomenon. This results in a level of analytics that was not possible in the past, due to either the lack of necessary volumes of data or limitations in algorithms.

Table 2 summarizes the difference in reporting and analytics. In this paper, we focus the discussion on large scale machine learning techniques that add insights and value beyond that provided by tools in the reporting and data mining domain.

Reporting	Analytics
<ul style="list-style-type: none">• Emphasis on counting events• Performed by business intelligence tools and queries• Present information based on events that have occurred in the past	<ul style="list-style-type: none">• Emphasis on discovering new patterns and trends• Performed by predictive or descriptive algorithms• Build models that can forecast/predict future events or classify events

Table 2. Reporting versus Analytics

Big Data to Insights: Strategic and Technical Considerations

The McKinsey Global Institute⁴ reports that big data can add value in several key areas – increasing transparency, improving efficiency, automating human tasks and developing new products and business models. However, when enterprises find themselves with increasing volumes of data, it is enticing to invest in large systems to store and catalog the data, prior to a thorough analysis that maps the analytics to real business problems.

There are several key considerations along the path from data to actionable insight – formulating the business need, identifying and gathering the relevant data, designing the system and delivering the right solutions. Figure 2 outlines a lifecycle for developing big data and analytics solutions. A big data initiative is best positioned for success when it starts with an analysis of the impact on business goals and clearly defines the use cases enabled by such an activity.

Designing a Big Data Solution

Designing a system to solve big data problems can be a complex process involving several business and technical considerations.

Designing the right system and picking from the myriad of technologies available is often the most complex step in the process. We focus this section on the three key aspects of designing a big data and analytics system: 1) Data collection, storage and processing, 2) Machine learning, and 3) Visualization.

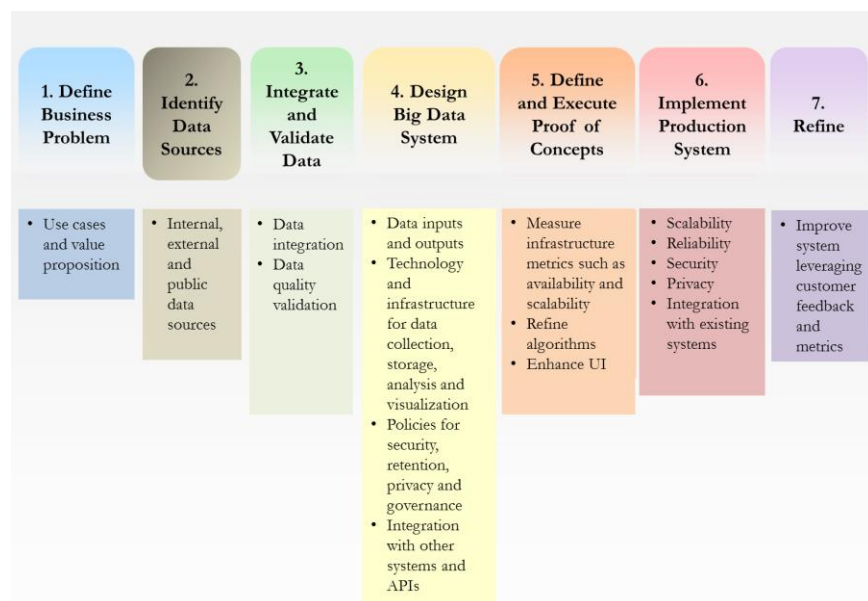


Figure 2. Big Data and Analytics initiatives -- life cycle

⁴ Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011

Technologies for Big Data Collection, Storage and Processing

A big data solution may often comprise more than one big data technology. It is important to carefully evaluate the strengths and shortcomings of any technology before integrating it with existing systems

Data Collection, Storage and Access

Data can exist in many disparate data stores and across several domains and may need special processing to collect and store. The variety of data types -- structured, unstructured and semi-structured -- can add an additional layer of complexity.

With several database and business intelligence players and a highly active open-source community developing big data technologies, this space is evolving very rapidly. In order to make the right technology choices, it is important to carefully evaluate the strengths and shortcomings of any technology before integrating it with existing systems. Furthermore, a big data solution may often comprise more than one big data technology, such as an inexpensive file system for batch storage, a key-value store for analytics and an in-memory database for fast and efficient visualization of the insights. Below are two key considerations when building a big data platform.

Data Collection: Data may exist in data stores or may be generated by real-time sources. Connectors, collectors and messaging tools may be needed to gather data from a variety of sources.

Data Storage and Access: Advances in distributed storage technologies, such as the Hadoop ecosystem, have greatly enhanced the available choices for both batch and near real-time storage and access to data. Table 3 highlights some of the key open source Apache projects that provide tools for data collection, storage and access in the Hadoop ecosystem.

Analytical data stores built on columnar or in-memory technology are optimized for fast data-access and may be better suited for data services and interactive applications. In cases where data can be stored as “key-value” pairs, NoSql technologies are gaining popularity.

Technology	Description
Pig	High-level data-flow language and execution framework
Hive	Data warehouse infrastructure
HBase	Scalable, distributed database for structure data
Cassandra	Scalable multi-master database
Flume	Distributed, service for efficiently collecting, aggregating and moving large amounts of log data
Sqoop	Efficient transfer of bulk data between Apache Hadoop and structured data stores such as relational databases

Table 3. Examples of Hadoop-related data collection, storage and access projects at Apache

Analytics and Large Scale Machine Learning

Identifying the right approach and techniques to solve an analytics problem is central to the field of data science. In addition, domain-specific knowledge is often required to develop the right analytics system.

Analytics and Large Scale Machine Learning

Analytics and Large Scale Machine learning aims to learn from data with the goal of predicting new events based on the learned patterns. These techniques create a dynamic system where algorithms can continually learn and adapt to new observations. Machine learning techniques often involve higher dimensional analysis and complex optimization techniques to answer business questions that cannot be answered by reporting tools. Large scale machine learning techniques take machine learning to the next level, often dealing with challenges imposed by high volume and dimensionality, complex data types, speed of data input and real-time need for output. Advances in distributed computing technology have played a key role supporting the increased use of large scale machine learning to solve business problems. The Hadoop ecosystem, for example, includes the Apache Mahout project, which is a library of scalable machine learning algorithms. Mahout includes capabilities to develop classification, clustering and recommendation systems.

Identifying the right approach and techniques to solve an analytics problem is central to the field of data science. There are often many considerations such as the scale and quality of data, computational complexity of the algorithms under consideration, batch versus real-time nature of analysis, tradeoff between specificity and sensitivity of the output and applicability and performance of the chosen techniques in a production environment. In addition, domain-specific knowledge may be required to develop the right analytics system.

Visualization

Creating visualizations that are actually used

The world is moving towards a DRIP mode – Data Rich, Information Poor. Enterprises face the key challenge of generating and sharing insights through analytics that are meaningful across its various end-users; otherwise, these analytics are wasteful in time and money. The answer to this challenge lies in understanding these three key points:

- Who is my audience and how will they read and interpret the information?
- What are their expectations and what type of information is most useful to them?
- How can the viewer take meaningful action from the information?

Compelling visualizations provide your viewers with actionable insights for better decision-making and a desire for more details to pursue.

Subpar visualizations make your viewers repeatedly ask “So what?”. When executed poorly, more data (i.e., big data) does not lead to more value.

Big Data Visualization

Big data has posed new challenges to visualization, particularly to human perceptual capabilities and interactive exploration. As a result, conventional approaches have evolved to define new ways to visualize the large volumes of data.

Visualization in the big data and analytics life cycle

Visualization has traditionally been used for designing dashboards intended to present information. However, big data and analytics projects can use visualization for other important purposes.

Data Quality Validation: visualizations to explore raw data sources to validate data quality and to determine pre-processing for parsing needs

Data Discovery: visualizations to explore the data, to uncover the grains of truth and to search for patterns to create features into machine learning

Storytelling: visualizations, from both raw data and machine learning output data, to develop a compelling story for a business audience

Delivery via a Product: visualizations to present insights to your audience in the optimal format, whether it is through reports, dashboards, alerts or tools

Challenges in big data visualization

Big data has posed new challenges to visualization, particularly to human perceptual capabilities and interactive exploration. Given the resolution of conventional displays (~1-3 million pixels), visualizing every data point can lead to over-plotting and will likely overwhelm users' perceptual and cognitive capacities. On the other hand, reducing the data through sampling or filtering can omit interesting structures or outliers. Similarly, querying large data stores can incur slow response times, disrupting a fluid user experience. As a result, conventional approaches have evolved to define new ways to visualize the large volumes of data.

New ways of visualizing big data?

Visual Querying: manipulating visual portions of graphs or charts to query and re-query data without writing code, e.g., drill across to another dimension, drill down and up hierarchies, drill through to another source or to another dashboard, filter, group, pivot, rank and sort.

Visualization via Animation: if a particular dimension, such as time, has thousands of values (such as daily values over multiple years), manually clicking through every day is not practical. Launching an automated, animated display of the dimension is a more practical approach.

Sparklines and Microcharts: both are tiny graphics that fit within a row of a table or even in a sentence. Sparklines communicate trends via a time-series graph while Microcharts communicate status via gauges or meters such as bulb/LED lights, clocks, dials, funnels, pyramids, street lights and temperature gauges.

Geospatial Visualization: overlaying data on maps makes information more consumable and easier to comprehend. Maps also support analysis of geospatial relationships (e.g., location, area, distance, etc.)

The Road Ahead: Conclusions and Future Outlook

Big data and analytics enable enterprises to generate greater insights and improve efficiency or develop new products, services or business models. However, as discussed, there are many considerations on the path from data to actionable insights. Successful big data and analytics initiatives involve a variety of decisions, from mapping business strategy to use cases to making the right technical choices to implement a cost-effective system. Organizations also need to consider aspects such as data governance, privacy, organizational change and talent management.

As big data technologies evolve at a rapid pace and analytics make it possible to extract greater insights, big data and analytics are set to become a mainstay in enterprises.

About Verizon Big Data and Analytics

Powered by investments in security, data centers, 4GLTE, cloud computing and our immense global IP network, our portfolio of solutions effortlessly meets the demands and challenges that shape technology and business today.

Verizon is a leader in the world of big data solutions, using innovation in infrastructure, algorithms and analytics solutions to benefit our customers. Our Big Data Lab is located in Palo Alto and focuses on cutting-edge research and product development in big data solutions for marketing analytics, network optimization, machine-to-machine (M2M) and cyber security. To find out more about our solutions for big data or to discuss your needs, contact us at www.verizonenterprise.com.

© 2014 Verizon. All Rights Reserved.

The Verizon names and logos and all other names, logos and slogans identifying Verizon's products and services are trademarks and service marks or registered trademarks and service marks of Verizon Trademark Services LLC or its affiliates in the United States and/or other countries. All other trademarks and service marks are the property of their respective owners.