Big Data Processing

Big data primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many entries (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.[2] Though used sometimes loosely partly because of a lack of formal definition, the interpretation that seems to best describe big data is the one associated with a large body of information that we could not comprehend when used only in smaller amounts.[3]Big data analysis challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: volume, variety, and velocity.[4] The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Thus a fourth concept, veracity, refers to the quality or insightfulness of the data. Without sufficient investment in expertise for big data veracity, then the volume and variety of data can produce costs and risks that exceed an organization's capacity to create and capture value from big data.[5]Current usage of the term big data tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from big data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."[6] Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on".[7] Scientists, business executives, medical practitioners, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet searches, fintech, healthcare analytics, geographic information systems, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics,[8] connectomics, complex physics simulations, biology, and environmental research.[9]The size and number of available data sets have grown rapidly as data is collected by devices such as mobile devices, cheap and numerous information-sensing Internet of things devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.[10][11] The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[12] as of 2012, every day 2.5 exabytes (2.5×260 bytes) of data are generated.[13] Based on an IDC report prediction, the global data volume was predicted to grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020. By 2025, IDC predicts there will be 163 zettabytes of data.[14] According to IDC, global spending on big data and business analytics (BDA) solutions is estimated to reach $215.7 billion in 2021.[15][16] While Statista report, the global big data market is forecasted to grow to $103 billion by 2027.[17] In 2011 McKinsey & Company reported, if US healthcare were to use big data creatively and effectively to drive efficiency and quality, the sector could create more than $300 billion in value every year.[18] In the developed economies of Europe, government administrators could save more than €100 billion ($149 billion) in operational efficiency improvements alone by using big data.[18] And users of services enabled by personal-location data could capture $600 billion in consumer surplus.[18]

One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.[19]Relational database management systems and desktop statistical software packages used to visualize data often have difficulty processing and analyzing big data. The processing and analysis of big data may require "massively parallel software running on tens, hundreds, or even thousands of servers".[20] What qualifies as "big data" varies depending on the capabilities of those analyzing it and their tools. Furthermore, expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."[21]Definition[edit]The term big data has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term.[22][23] Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.[24] Big data philosophy encompasses unstructured, semi-structured and structured data; however, the main focus is on unstructured data.[25] Big data "size" is a constantly moving target; as of 2012 ranging from a few dozen terabytes to many zettabytes of data.[26] Big data requires a set of techniques and technologies with new forms of integration to reveal insights from data-sets that are diverse, complex, and of a massive scale.[27]"Variety", "veracity", and various other "Vs" are added by some organizations to describe it, a revision challenged by some industry authorities.[28] The Vs of big data were often referred to as the "three Vs", "four Vs", and "five Vs". They represented the qualities of big data in volume, variety, velocity, veracity, and value.[4] Variability is often included as an additional quality of big data.A 2018 definition states "Big data is where parallel computing tools are needed to handle data", and notes, "This represents a distinct and clearly defined change in the computer science used, via parallel programming theories, and losses of some of the guarantees and capabilities made by Codd's relational model."[29]In a comparative study of big datasets, Kitchin and McArdle found that none of the commonly considered characteristics of big data appear consistently across all of the analyzed cases.[30] For this reason, other studies identified the redefinition of power dynamics in knowledge discovery as the defining trait.[31] Instead of focusing on the intrinsic characteristics of big data, this alternative perspective pushes forward a relational understanding of the object claiming that what matters is the way in which data is collected, stored, made available and analyzed.Big data vs. business intelligence[edit]The growing maturity of the concept more starkly delineates the difference between "big data" and "business intelligence":[32]Business intelligence uses applied mathematics tools and descriptive statistics with data with high information density to measure things, detect trends, etc.Big data uses mathematical analysis, optimization, inductive statistics, and concepts from nonlinear system identification[33] to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density[34] to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.[33][35] [promotional source?]Characteristics[edit]Shows the growth of big data's primary characteristics of volume, velocity, and variety.Big data can be described by the

following characteristics:VolumeThe quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. The size of big data is usually larger than terabytes and petabytes.[36]VarietyThe type and nature of the data. Earlier technologies like RDBMSs were capable to handle structured data efficiently and effectively. However, the change in type and nature from structured to semi-structured or unstructured challenged the existing tools and technologies. Big data technologies evolved with the prime intention to capture, store, and process the semi-structured and unstructured (variety) data generated with high speed (velocity), and huge in size (volume). Later, these tools and technologies were explored and used for handling structured data also but preferable for storage. Eventually, the processing of structured data was still kept as optional, either using big data or traditional RDBMSs. This helps in analyzing data towards effective usage of the hidden insights exposed from the data collected via social media, log files, sensors, etc. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.VelocityThe speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time. Compared to small data, big data is produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.[37]VeracityThe truthfulness or reliability of the data, which refers to the data quality and the data value.[38] Big data must not only be large in size, but also must be reliable in order to achieve value in the analysis of it. The data quality of captured data can vary greatly, affecting an accurate analysis.[39]ValueThe worth in information that can be achieved by the processing and analysis of large datasets. Value also can be measured by an assessment of the other qualities of big data.[40] Value may also represent the profitability of information that is retrieved from the analysis of big data.VariabilityThe characteristic of the changing formats, structure, or sources of big data. Big data can include structured, unstructured, or combinations of structured and unstructured data. Big data analysis may integrate raw data from multiple sources. The processing of raw data may also involve transformations of unstructured data to structured data.Other possible characteristics of big data are:[41]ExhaustiveWhether the entire system (i.e., =all) is captured or recorded or not. Big data may or may not include all the available data from sources.Fine-grained and uniquely lexicalRespectively, the proportion of specific data of each element per element collected and if the element and its characteristics are properly indexed or identified.RelationalIf the data collected contains common fields that would enable a conjoining, or meta-analysis, of different data sets.ExtensionalIf new fields in each element of the data collected can be added or changed easily.ScalabilityIf the size of the big data storage system can expand rapidly.Architecture[edit]Big data repositories have existed in many forms, often built by corporations with a special need. Commercial vendors historically offered parallel database management systems for big data beginning in the 1990s. For many years, WinterCorp published the largest database report.[42][promotional source?]Teradata Corporation in 1984 marketed the parallel processing DBC 1012 system. Teradata systems were the first to store and analyze 1 terabyte of data in 1992. Hard disk drives were 2.5 GB in 1991 so the definition of big data continuously evolves. Teradata installed the first petabyte class RDBMS based system in 2007. As of 2017, there are a

few dozen petabyte class Teradata relational databases installed, the largest of which exceeds 50 PB. Systems up until 2008 were 100% structured relational data. Since then, Teradata has added unstructured data types including XML, JSON, and Avro.In 2000, Seisint Inc. (now LexisNexis Risk Solutions) developed a C++-based distributed platform for data processing and querying known as the HPCC Systems platform. This system automatically partitions, distributes, stores and delivers structured, semi-structured, and unstructured data across multiple commodity servers. Users can write data processing pipelines and queries in a declarative dataflow programming language called ECL. Data analysts working in ECL are not required to define data schemas upfront and can rather focus on the particular problem at hand, reshaping data in the best possible manner as they develop the solution. In 2004, LexisNexis acquired Seisint Inc.[43] and their high-speed parallel processing platform and successfully used this platform to integrate the data systems of Choicepoint Inc. when they acquired that company in 2008.[44] In 2011, the HPCC systems platform was open-sourced under the Apache v2.0 License.CERN and other physics experiments have collected big data sets for many decades, usually analyzed via high-throughput computing rather than the map-reduce architectures usually meant by the current "big data" movement.In 2004, Google published a paper on a process called MapReduce that uses a similar architecture. The MapReduce concept provides a parallel processing model, and an associated implementation was released to process huge amounts of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the "map" step). The results are then gathered and delivered (the "reduce" step). The framework was very successful,[45] so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open-source project named "Hadoop".[46] Apache Spark was developed in 2012 in response to limitations in the MapReduce paradigm, as it adds in-memory processing and the ability to set up many operations (not just map followed by reducing).MIKE2.0 is an open approach to information management that acknowledges the need for revisions due to big data implications identified in an article titled "Big Data Solution Offering".[47] The methodology addresses handling big data in terms of useful permutations of data sources, complexity in interrelationships, and difficulty in deleting (or modifying) individual records.[48]Studies in 2012 showed that a multiple-layer architecture was one option to address the issues that big data presents. A distributed parallel architecture distributes data across multiple servers; these parallel execution environments can dramatically improve data processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of MapReduce and Hadoop frameworks. This type of framework looks to make the processing power transparent to the end-user by using a front-end application server.[49]The data lake allows an organization to shift its focus from centralized control to a shared model to respond to the changing dynamics of information management. This enables quick segregation of data into the data lake, thereby reducing the overhead time.[50][51]Technologies[edit]A 2011 McKinsey Global Institute report characterizes the main components and ecosystem of big data as follows:[52]Techniques for analyzing data, such as A/B testing, machine learning, and natural language processingBig data technologies, like business intelligence, cloud computing, and

databasesVisualization, such as charts, graphs, and other displays of the dataMultidimensional big data can also be represented as OLAP data cubes or, mathematically, tensors. Array database systems have set out to provide storage and high-level query support on this data type. Additional technologies being applied to big data include efficient tensor-based computation,[53] such as multilinear subspace learning,[54] massively parallel-processing (MPP) databases, search-based applications, data mining,[55] distributed file systems, distributed cache (e.g., burst buffer and Memcached), distributed databases, cloud and HPC-based infrastructure (applications, storage and computing resources),[56] and the Internet.[citation needed] Although, many approaches and technologies have been developed, it still remains difficult to carry out machine learning with big data.[57]Some MPP relational databases have the ability to store and manage petabytes of data. Implicit is the ability to load, monitor, back up, and optimize the use of the large data tables in the RDBMS.[58] [promotional source?]DARPA's Topological Data Analysis program seeks the fundamental structure of massive data sets and in 2008 the technology went public with the launch of a company called "Ayasdi".[59][third-party source needed]The practitioners of big data analytics processes are generally hostile to slower shared storage,[60] preferring direct-attached storage (DAS) in its various forms from solid state drive (SSD) to high capacity SATA disk buried inside parallel processing nodes. The perception of shared storage architectures—storage area network (SAN) and network-attached storage (NAS)— is that they are relatively slow, complex, and expensive. These qualities are not consistent with big data analytics systems that thrive on system performance, commodity infrastructure, and low cost.Real or near-real-time information delivery is one of the defining characteristics of big data analytics. Latency is therefore avoided whenever and wherever possible. Data in direct-attached memory or disk is good—data on memory or disk at the other end of an FC SAN connection is not. The cost of an SAN at the scale needed for analytics applications is much higher than other storage techniques.