International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

A Literary Review on Big Data & Hadoop

Anudeepa Gon

Abstract: ***This report file is prepared on the topic Big Data Analytics and Hadoop; it has been tried to elucidate all the relevant details to the topic to be included in the report. In the beginning this report gives an overall view about this topic. 'Big Data' is the data but with a huge size. 'Big Data' is used to explain the collection of data which is huge in size and still growing exponentially with respect to time. Basically, this data is so large and complex as none of the traditional data management tools can be store it or process it efficiently and perfectly. Big data analytics gives permission to data scientists and various other users to calculate large volumes of transaction data and other sources of data that traditional business systems are unable to handle. Modern software programs that are used for big data analytics, while the unstructured data used in big data analytics may not be applicable to conventional data warehouses. Requirements which are high in processing associated with Big data may also make traditional data warehousing a poor fit. As a result, newer, bigger data analytics environments and technologies have emerged, including Hadoop, MapReduce and NoSQL databases. These technologies make up an open-source software framework that's used to process huge data sets over clustered systems. Hadoop is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems, formerly known as Apache Hadoop. The technology is developed as part of an open source project within the Apache Software Foundation (ASF). Big Data is nothing but a concept which describes how to handle large amount of datasets. Hadoop is just a single framework out of dozens of tools. Initially Hadoop is used for batch processing technology. The difference between big data and the open source software Hadoop is a distinct and fundamental one.***

Keywords: BigData, Concept, Definition, Hadoop, Structure, procedure

1. Introduction

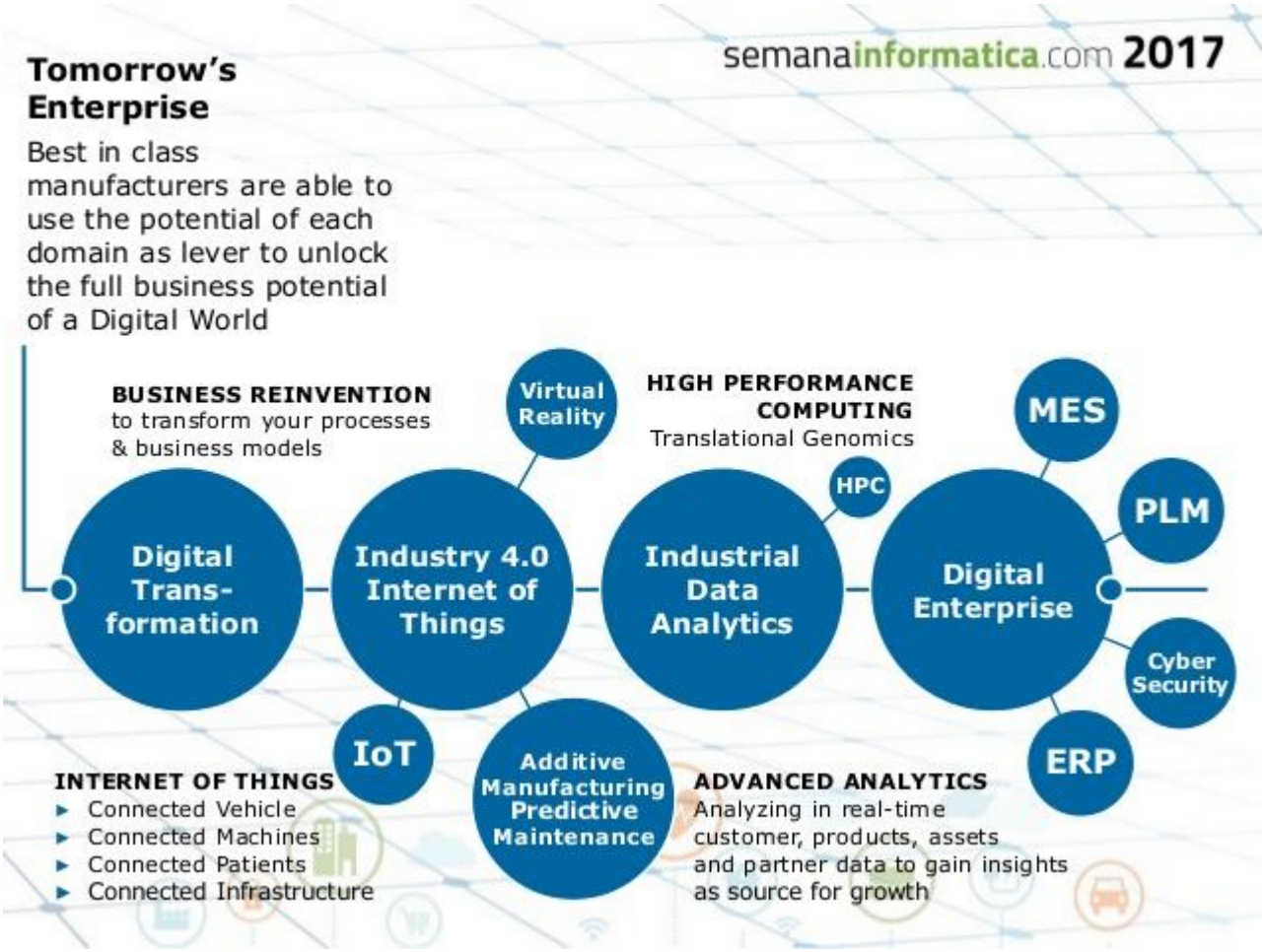
Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reductions and reduced risk.

Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, combat crime and so on." Scientists, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research."

Digital technologies have rapidly been assimilated into our everyday lives, for instance via digital payment systems and the widespread diffusion of broadband infrastructure. Increasing digitization throughout all aspects of life is creating significant challenges for the insurance industry. It has fundamentally changed the behaviors, needs and requirements of the customer. The Internet generation – which, by the way, is already in its 30s – has grown up with computers, mobile phones and the Internet and is accustomed to communicating and shopping using smartphones. If digital natives are interested in an insurance product, they inform themselves on the Web via consumer and advisory websites, have a look at relevant blogs and search for the most affordable offers on comparison websites. Ultimately they conclude the contract online, for example by using an insurance app which they have downloaded to their smartphone.
Digital transformation helps companies embrace a culture of change and remain competitive in a global environment, but, when companies decide to go digital, the process is a little bit like losing weight (’tis the season!). You can’t go on a diet for a few weeks and expect fast or lasting results. Losing weight has to be a lifestyle change, and so does incorporating big data into your business strategies.

Big data allows companies to make meaningful, strategic adjustments that minimize costs and maximize results. If you know what consumers and employees are doing currently, you can create projections for what they will do in the future, and start implementing changes to address their needs and your goals. A digital transformation isn’t complete unless a business adopts big data. Big data and analytics are topics firmly embedded in our business dialogue. The amount of data we’re now generating is astonishing. Cisco predicts that annual global IP traffic will reach 3.3 ZB per year by 2022 and that the number of devices connected to IP networks will be more than three times the global population by 2022, while Gartner predicts $2.5M per minute in IoT spending and 1M new IoT devices will be sold every hour by 2022. It’s testament to the speed with which digital connectivity is changing the lives of people all over the world. (1)

Data has also evolved dramatically in recent years, in type, volume, and velocity – with its rapid evolution attributed to the widespread digitisation of business processes globally. Data has become the new business currency and its further rapid increase will be key to the transformation and growth of enterprises globally, and the advancement of employees, ‘the digital natives’.



Digital Transformation

Internet of Things	Cloud Computing	Big Data Analytics	Augmented-/Virtual Reality	Artificial Intelligence & Machine Learning
Smart Factory, Smart Offices, Cyber-Physical-Systems (CPS)	All data are available anytime, anywhere	Data can be used to better understand the world and/or human behavior	Creation of an alternative reality resp. digitalization of the perception of reality	Machines think and learn like humans

Network Economy

Pressure to innovate and change	Ceasing organizational borders	Business Model disruption (Everything as a Service)	New work environment (Digital Workplace)	Competencies instead of knowledge
Shortened life span of products, price decline, decreasing customer loyalty	Crowd Sourcing, project based work, integration of customers, suppliers and competitors into the service offering	With increasing focus on the needs of customers even manufactures become service provider	Working independent from time and place: Because it gets easier through emerging technology and because work is increasingly digital.	Elaborate competencies are an imperative to add a value in a world where "ability and knowledge" of machines increase.

Agility as a response to the Digital Transformation

Agility as the ability to change efficient and effective in order to handle the increasingly complex environment. A complexity that is immanent to a digital and globally tightly connected network economy.

Big data means better business

Data is an enabler of future strategies and immediate change, thanks to the power of predictive analytics and advanced data science. Correctly harnessing data can help to achieve better, fact-based decision-making and improve the overall customer experience. By using new Big Data technologies, organisations can answer questions in seconds rather than days, and in days rather than months. This acceleration allows businesses to enable the type of quick reactions to key business questions and challenges that can build competitive advantage and improve performance, and provide answers for complex problems or questions that have resisted analysis.

Big Data and analytics are becoming closely intertwined and need to work together to deliver the promised results of Big Data. Traditionally, Data management and analytics have resided in different parts of the organisation. Breaking down organisational boundaries and creating better integration between the IT and business departments is a critical step on the road to successful transformation. (2)

There is also a widespread realisation of the need for better Business Analytics. Business Analytics are the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. The key is integrating Big Data with traditional Business Analytics to create a data ecosystem that allows the business to generate new insights while executing on what it already knows.

Definition of Big Data

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. Big data "size" is a constantly moving target, as of 2018 ranging from a few dozen terabytes to many petabytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

Big Data Analytics

On a broad scale, data analytics technologies and techniques provide a means to analyze data sets and draw conclusions about them to help organizations make informed business decisions. BI queries answer basic questions about business operations and performance. Big data analytics is a form of advanced analytics, which involves complex applications with elements such as predictive models, statistical algorithms and what-if analysis powered by high-performance analytics systems.

Meta Data Under Big Data Analytics

In a 2001 research report and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data. In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." Additionally, a new V "Veracity" is added by some organizations to describe it.

Concept of 4Vs

If Gartner's definition (the 4Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between big data and Business Intelligence, regarding data and their use:

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;
- Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships, dependencies and perform predictions of outcomes and behaviors.

A more recent, consensual definition states that "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value".(3)

Big data can be described by the following characteristics:

Volume – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

Variety - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of Big Data.

Velocity - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

Variability - This is a factor which can be a problem for those who analyse the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Veracity - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

Complexity - Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

Factory work and Cyber-physical systems may have a 6C system:

1. Connection (sensor and networks),
2. Cloud (computing and data on demand),
3. Cyber (model and memory),
4. Content/context (meaning and correlation),
5. Community (sharing and collaboration), and
6. Customization (personalization and value).

In this scenario and in order to provide useful insight to the factory management and gain correct content, data has to be processed with advanced tools (analytics and algorithms) to generate meaningful information. Considering the presence of visible and invisible issues in an industrial factory, the information generation algorithm has to be capable of detecting and addressing invisible issues such as machine degradation, component wear, etc. in the factory floor. (4)

The importance of big data analytics

Driven by specialized analytics systems and software, as well as high-powered computing systems, big data analytics offers various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals.

Big data analytics applications enable big data analysts, data scientists, predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs. That encompasses a mix of semi-structured and unstructured data -- for example, internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile phone records, and machine data captured by sensors connected to the internet of things. (5)

Categories Of 'Big Data'

Big data' could be found in three forms:

1. **Structured**
2. **Unstructured**
3. **Semi-structured**

Structured

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kinds of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when the size of such data grows to a huge extent, typical sizes being in the range of multiple zettabytes.

An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_ Name	Gender	Department	Salary_In_ lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, unstructured data poses multiple challenges in terms of its processing for deriving value out of it. Typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Nowadays organizations have a wealth of data available with them but unfortunately they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Semi-structured data can contain both the forms of data. We can see semi-structured data as a strcutured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in XML file.(6)

Examples of Semi-structured Data

Personal data stored in a XML file-(7)

<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>

- <rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>

<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>

Architectural demand for Big Data

Logical layers offer a way to organize your components. The layers simply provide an approach to organizing components that perform specific functions. The layers are merely logical; they do not imply that the functions that support each layer are run on separate machines or separate processes. A big data solution typically comprises these logical layers:

1. Big data sources

2. Data massaging and store layer
3. Analysis layer
4. Consumption layer

- **Collection point**: Where the data is collected, directly or through data providers, in real time or in batch mode. The data can come from a primary source, such as weather conditions, or it can come from a secondary source, such as a media-sponsored weather channel.

- **Location of data source**: Data sources can be inside the enterprise or external. Identify the data to which you have limited-access, since access to data affects the scope of data available for analysis.

- **Data massaging and store layer:** This layer is responsible for acquiring data from the data sources and, if necessary, converting it to a format.

- Data is to be analyzed. For example, an image might need to be converted so it can be stored in an Hadoop Distributed File System (HDFS) store or a Relational Database Management System (RDBMS) warehouse for further processing. Compliance regulations and governance policies dictate the appropriate storage for different types of data.

- **Analysis layer:** The analysis layer reads the data digested by the data massaging and store layer. In some cases, the analysis layer accesses the data directly from the data source. Designing the analysis layer requires careful forethought and planning. Decisions must be made with regard to how to manage the tasks to:

  - Produce the desired analytics
  - Derive insight from the data
  - Find the entities required
  - Locate the data sources that can provide data for these entities
  - Understand what algorithms and tools are required to perform the analytics.

- **Consumption layer:** This layer consumes the output provided by the analysis layer. The consumers can be visualization applications, human beings, business processes, or services. It can be challenging to visualize the outcome of the analysis layer. Sometimes it's helpful to look at what competitors in similar markets are doing.(8)

Each layer includes several types of components, as illustrated below.

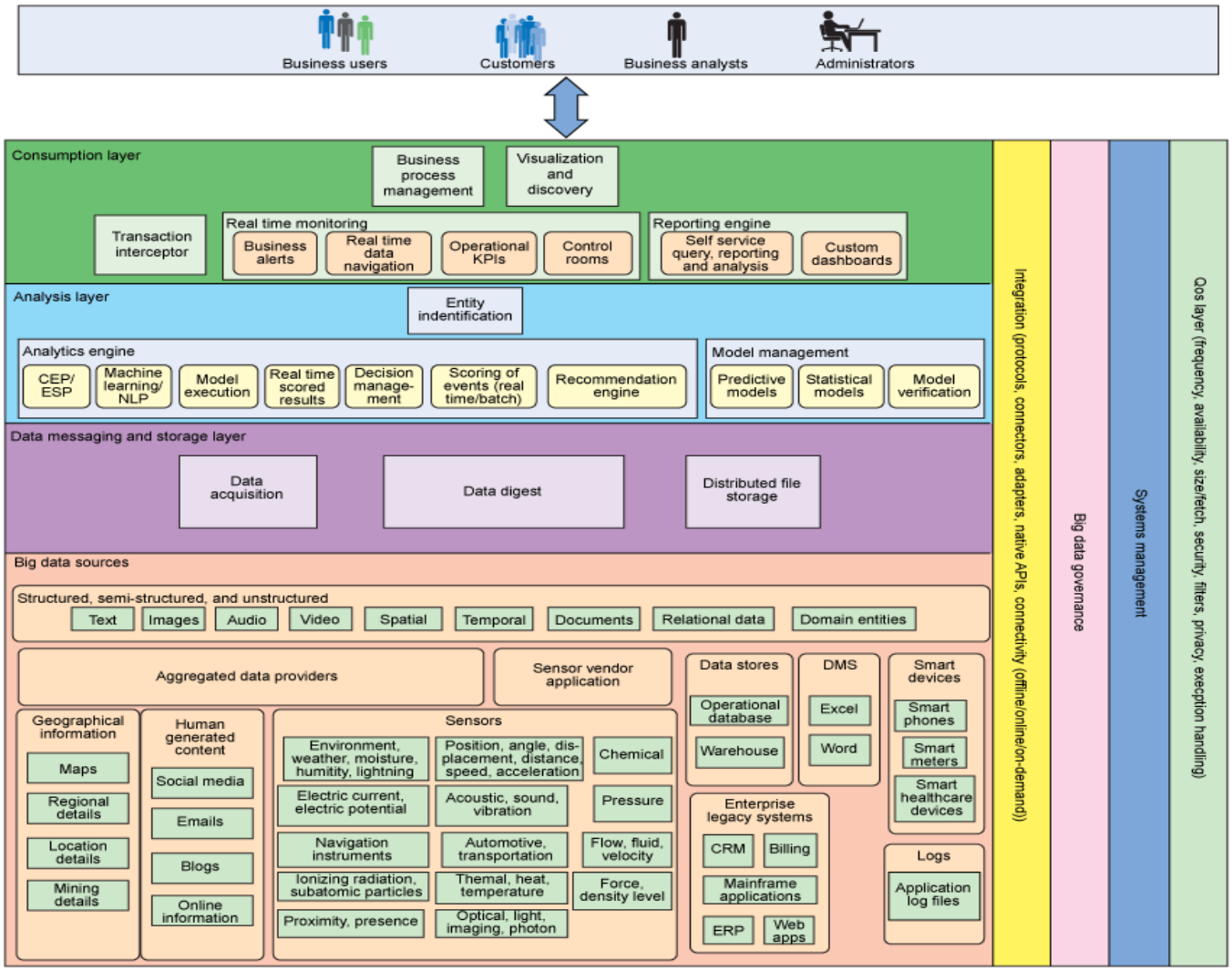


Figure 1: Components by logical and vertical layer

Big data sources

This layer includes all the data sources necessary to provide the insight required to solve the business problem. The data is structured, semi-structured, and unstructured, and it comes from many sources:

- **Enterprise legacy systems:** These are the enterprise applications that drive the analytics and insights required for business:
    - Customer relationship management systems
    - Billing operations
    - Mainframe applications
    - Enterprise resource planning
    - Web applications
- Web applications and other data sources augment the enterprise-owned data. Such applications can expose the data using custom protocols and mechanisms.
- **Data management systems (DMS)**— The data management systems store legal data, processes, policies, and various other kinds of documents:
    - Microsoft® Excel® spreadsheets
    - Microsoft Word documents

These documents can be converted into structured data that can be used for analytics. The document data can be exposed as domain entities or the data massaging and storage layer can transform it into the domain entities.

- **Data stores**: Data stores include enterprise data warehouses, operational databases, and transactional databases. This data is typically structured and can be consumed directly or transformed easily to suit requirements. Such data may or may not be stored in the distributed file system, depending on the context of the situation.
- **Smart devices**: Smart devices are capable of capturing, processing, and communicating information on most widely used protocols and formats. Examples include smartphones, meters, and healthcare devices. Such devices can be used to perform various kinds of analysis. For the most part, smart devices do real-time analytics, but the information stemming from smart devices can be analyzed in batch, as well.
- **Aggregated data providers**: These providers own or acquire the data and expose it in sophisticated formats, at required frequencies, and through specific filters. Huge volumes of data pour in, in a variety of formats, produced at different velocities, and made available by various data providers, sensors, and existing enterprises.
- **Additional data sources**: A wide range of data comes from automated sources:
- Geographical information:
    - Maps
    - Regional details
    - Location details
    - Mining details
- Human-generated content:
    - Social media
    - Email
    - Blogs
    - Online information

Sensor data:

- Environment: Weather, moisture, humidity, lightening
- Electricity: Current, energy potential, etc.
- Navigation instruments
- Ionizing radiation, subatomic particles, etc.
- Proximity, presence, and so on
- Position, angle, displacement, distance, speed, acceleration
- Acoustic, sound vibration, etc.
- Automotive, transportation, etc.
- Thermal, heat, temperature
- Optical, light, imaging, photon
- Chemical
- Pressure
- Flow, fluid, velocity

Big data Analytics Technologies

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. A 2011 McKinsey report suggests suitable technologies include A/B testing, crowdsourcing, data fusion and integration, genetic algorithms, machine learning, natural language processing, signal processing, simulation, time series analysis and visualisation. Multidimensional big data can also be represented as tensors, which can be more efficiently handled by tensor-based computation, such as multilinear subspace learning. Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data mining, distributed file systems, distributed databases, cloud based infrastructure (applications, storage and computing resources)(9) and the Internet.

The big data analytics technology is a combination of several techniques and processing methods. What makes them effective is their collective use by enterprises to obtain relevant results for strategic management and implementation.

In spite of the investment enthusiasm, and ambition to leverage the power of data to transform the enterprise, results vary in terms of success. Organizations still struggle to forge what would be consider a “data-driven” culture. Of the executives who report starting such a project, only 40.2% report having success. Big transformations take time, and while the vast majority of firms aspire to being “data-driven”, a much smaller percentage have realized this ambition. Cultural transformations seldom occur overnight.

In the evolution of big data, the challenges for most companies are not related to technology. The biggest

impediments to adoption relate to cultural challenges: organizational alignment, resistance or lack of understanding, and change management.
Here are some key technologies that enable Big Data for Businesses: (10)

1) Predictive Analytics

One of the prime tools for businesses to avoid risks in decision making, predictive analytics can help businesses. Predictive analytics hardware and software solutions can be utilised for discovery, evaluation and deployment of predictive scenarios by processing big data. Such data can help companies to be prepared for what is to come and help solve problems by analyzing and understanding them.

2) NoSQL Databases

These databases are utilised for reliable and efficient data management across a scalable number of storage nodes. NoSQL databases store data as relational database tables, JSON docs or key-value pairings.

3) Knowledge Discovery Tools

These are tools that allow businesses to mine big data (structured and unstructured) which is stored on multiple sources. These sources can be different file systems, APIs, DBMS or similar platforms. With search and knowledge discovery tools, businesses can isolate and utilise the information to their benefit.

4) Stream Analytics

Sometimes the data an organisation needs to process can be stored on multiple platforms and in multiple formats. Stream analytics software is highly useful for filtering, aggregation, and analysis of such big data. Stream analytics also allows connection to external data sources and their integration into the application flow.

5) In-memory Data Fabric

This technology helps in distribution of large quantities of data across system resources such as Dynamic RAM, Flash Storage or Solid State Storage Drives. Which in turn enables low latency access and processing of big data on the connected nodes.

6) Distributed Storage

A way to counter independent node failures and loss or corruption of big data sources, distributed file stores contain replicated data. Sometimes the data is also replicated for low latency quick access on large computer networks. These are generally non-relational databases.

7) Data Virtualization

It enables applications to retrieve data without implementing technical restrictions such as data formats, the physical location of data, etc. Used by Apache Hadoop and other distributed data stores for real-time or near real-time access to data stored on various platforms, data virtualization is one of the most used big data technologies.

8) Data Integration

A key operational challenge for most organizations handling big data is to process terabytes (or petabytes) of data in a way that can be useful for customer deliverables. Data integration tools allow businesses to streamline data across a number of big data solutions such as Amazon EMR, Apache Hive, Apache Pig, Apache Spark, Hadoop, MapReduce, MongoDB and Couchbase.

9) Data Preprocessing

These software solutions are used for manipulation of data into a format that is consistent and can be used for further analysis. The data preparation tools accelerate the data sharing process by formatting and cleansing unstructured data sets. A limitation of data preprocessing is that all its tasks cannot be automated and require human oversight, which can be tedious and time-consuming.

10) Data Quality

An important parameter for big data processing is the data quality. The data quality software can conduct cleansing and enrichment of large data sets by utilising parallel processing. These softwares are widely used for getting consistent and reliable outputs from big data processing.

Advantages & disadvantages of Big Data Analytics

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. A 2011 McKinsey report suggests suitable technologies include A/B testing, crowdsourcing, data fusion and integration.

Unstructured and semi-structured data types typically don't fit well in traditional data warehouses that are based on relational databases oriented to structured data sets. Further, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently -- or even continually, as in the case of real-time data on stock trading, the online activities of website visitors or the performance of mobile applications.

As a result, many of the organizations that collect, process and analyze big data turn to **NoSQL** databases, as well as **Hadoop** and its companion tools, including:

- **YARN**: A cluster management technology and one of the key features in second-generation Hadoop.
- **MapReduce**: A software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.

- **Spark**: An open source, parallel processing framework that enables users to run large-scale data analytics applications across clustered systems.
- **HBase**: A column-oriented key/value data store built to run on top of the Hadoop Distributed File System (HDFS).
- **Hive**: An open source data warehouse system for querying and analyzing large data sets stored in Hadoop files.
- **Kafka**: A distributed publish/subscribe messaging system designed to replace traditional message brokers.
- **Pig**: An open source technology that offers a high-level mechanism for the parallel programming of MapReduce jobs executed on Hadoop clusters.
- **Mongodb:** MongoDB is an open source database management system (DBMS) that uses a document-oriented database model which supports various forms of data. It is one of numerous nonrelational database technologies which arose in the mid-2000s under the NoSQL banner for use in big data applications and other processing jobs involving data that doesn't fit well in a rigid relational model.

Working procedure of Big Data Analytics

In some cases, Hadoop clusters and NoSQL systems are used primarily as landing pads and staging areas for data before it gets loaded into a data warehouse or analytical database for analysis -- usually in a summarized form that is more conducive to relational structures.

More frequently, however, big data analytics users are adopting the concept of a Hadoop data lake that serves as the primary repository for incoming streams of raw data. In such architectures, data can be analyzed directly in a Hadoop cluster or run through a processing engine like Spark. As in data warehousing, sound data management is a crucial first step in the big data analytics process. Data being stored in the Hadoop Distributed File System must be organized, configured and partitioned properly to get good performance out of both extract, transform and load (ETL) integration jobs and analytical queries. (11)

Applications Of Big Data Analytics

Big data has increased the demand of information management specialists in that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. In 2010, this industry was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.

Many vendors offer off-the-shelf solutions for Big Data, experts recommend the development of in-house solutions custom-tailored to solve the company's problem at hand if the company has sufficient technical capabilities.

Big Data has been playing a role of big game changer for most of the industries over the last few years. In this Big Data Applications blog, I will take you through various industry domains, where I will be explaining how Big Data is revolutionizing them. (12)

Big Data Applications

The primary goal of Big Data applications is to help companies make more informative business decisions by analyzing large volumes of data. It could include web server logs, Internet click stream data, social media content and activity reports, text from customer emails, mobile phone call details and machine data captured by multiple sensors.

Organisations from different domains are investing in Big Data applications, for examining large data sets to uncover all hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. In this blog we will we be covering:

- Big Data Applications in Healthcare
- Big Data Applications in Manufacturing
- Big Data Applications in Media & Entertainment
- Big Data Applications in IoT
- Big Data Applications in Government

Big Data applications are playing a major role in different domains.

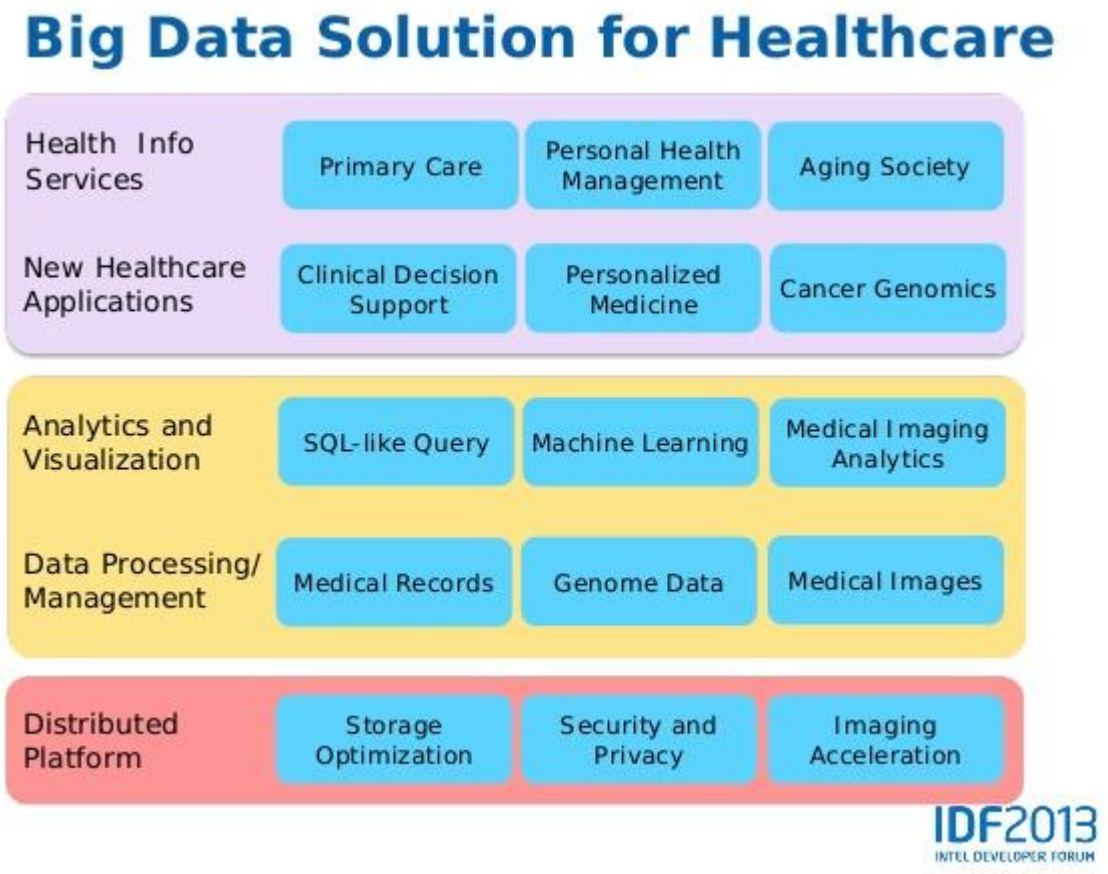
Big Data Applications: Healthcare

The level of data generated within healthcare systems is not trivial. Traditionally, the health care industry lagged in using Big Data, because of limited ability to standardize and consolidate data.

But now Big data analytics have improved healthcare by providing personalized medicine and prescriptive analytics. Researchers are mining the data to see what treatments are more effective for particular conditions, identify patterns related to drug side effects, and gains other important information that can help patients and reduce costs.

With the added adoption of mHealth, eHealth and wearable technologies the volume of data is increasing at an exponential rate. This includes electronic health record data, imaging data, patient generated data, sensor data, and other forms of data.

By mapping healthcare data with geographical data sets, it's possible to predict disease that will escalate in specific areas. Based of predictions, it's easier to strategize diagnostics and plan for stocking serums and vacancies.



Big Data Applications: Manufacturing

Predictive manufacturing provides near-zero downtime and transparency. It requires an enormous amount of data and advanced prediction tools for a systematic process of data into useful information.

Major benefits of using Big Data applications in manufacturing industry are:

- Product quality and defects tracking
- Supply planning
- Manufacturing process defect tracking
- Output forecasting
- Increasing energy efficiency
- Testing and simulation of new manufacturing processes
- Support for mass-customization of manufacturing

Big Data Applications: Media & Entertainment

Various companies in the media and entertainment industry are facing new business models, for the way they – create, market and distribute their content. This is happening because of current consumer's search and the requirement of accessing content anywhere, any time, on any device.

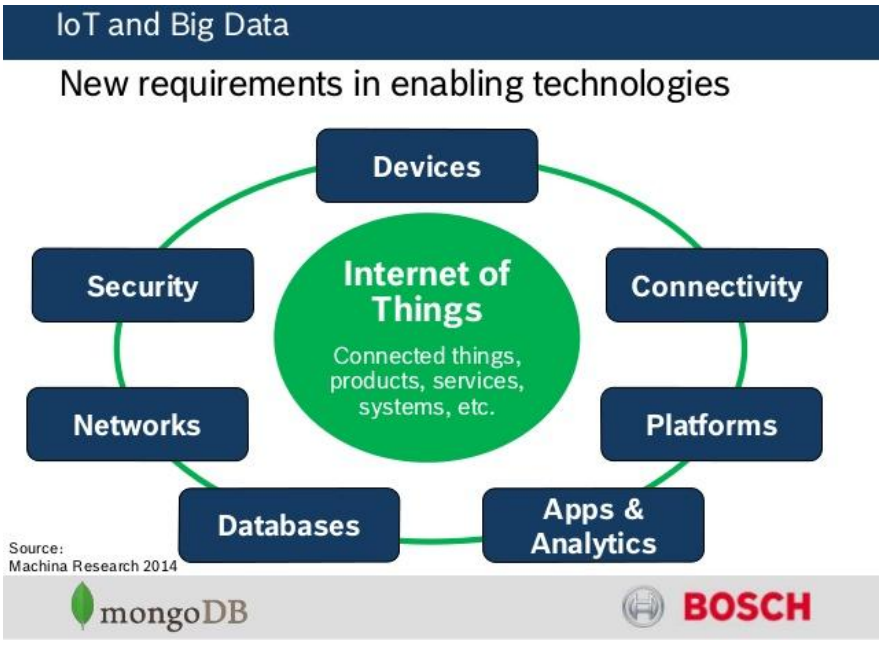
Big Data provides actionable points of information about millions of individuals. Now, publishing environments are tailoring advertisements and content to appeal to consumers. These insights are gathered through various data-mining activities. Big Data applications benefits media and entertainment industry by: (13)

- Predicting what the audience wants
- Scheduling optimization
- Increasing acquisition and retention
- Ad targeting
- Content monetization and new product development

Technical Requirement	Technology	Research Question
Semantic Data Enrichment	Common and open ontologies	Relation extraction
	Graph databases	Scalability of non-relational databases
	Crowdsourced curation platform	Blended algorithmic & manual curation at scale
Data Quality	Open data platforms	Data standardisation & interoperability
	Unstructured data processing	Natural language processing at scale
	Heterogeneous data storage	Data-agnostic storage architectures
Data-driven Innovation	Machine learning (ML)	Integrating ML approaches into databases
	Networks, sensors, wearable tech	Commercialisation of auto-generated data
	Customer recommendation tools	Improving algorithmic recommendations
Data Analysis	Descriptive analytics	Data mining for subjective factors e.g. sentiment
	Data visualisation solutions	Business-user friendly applications
	Customer relationship platforms	Understand contexts to enhance data delivery

Big Data Applications: Internet of Things (IoT)

Data extracted from **IoT** devices provides a mapping of device inter-connectivity. Such mappings have been used by various companies and governments to increase efficiency. IoT is also increasingly adopted as a means of gathering sensory data, and this sensory data is used in medical and manufacturing contexts.



Big Data Applications: Government

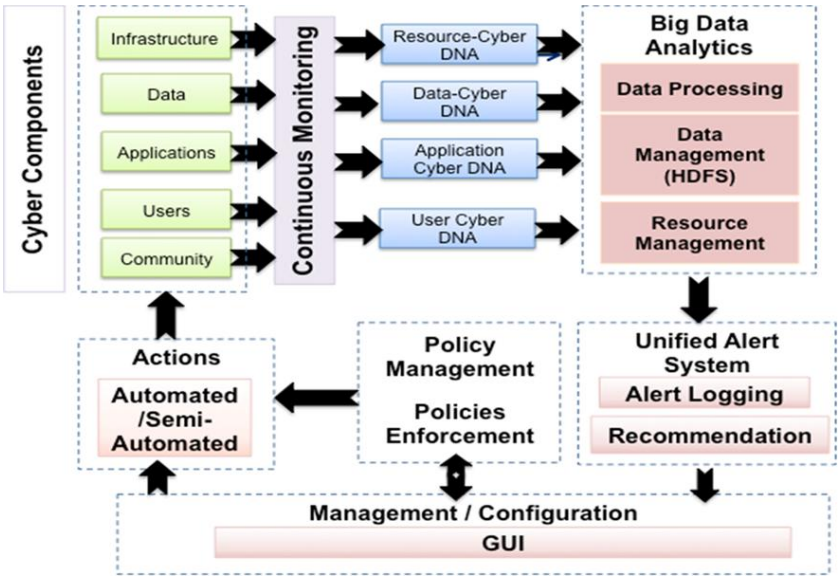
The use and adoption of Big Data within governmental processes allows efficiencies in terms of cost, productivity, and innovation. In government use cases, the same data sets are often applied across multiple applications & it requires multiple departments to work in collaboration.

Since Government majorly acts in all the domains, thus it plays an important role in innovating Big Data applications in each and every domain. Let me address some of the major areas:

Cyber security & Intelligence

The federal government launched a cyber security research and development plan that relies on the ability to analyze large data sets in order to improve the security of U.S. computer networks.

The National Geospatial-Intelligence Agency is creating a "Map of the World" that can gather and analyze data from a wide variety of sources such as satellite and social media data. It contains a variety of data from classified, unclassified, and top-secret networks.

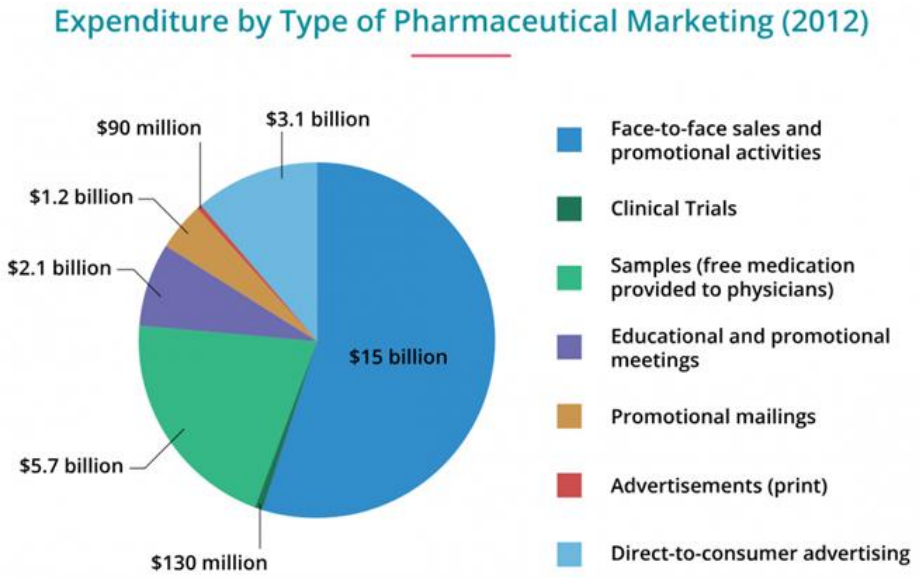


Crime Prediction and Prevention

Police departments can leverage advanced, real-time analytics to provide actionable intelligence that can be used to understand criminal behaviour, identify crime/incident patterns, and uncover location-based threats.

Pharmaceutical Drug Evaluation

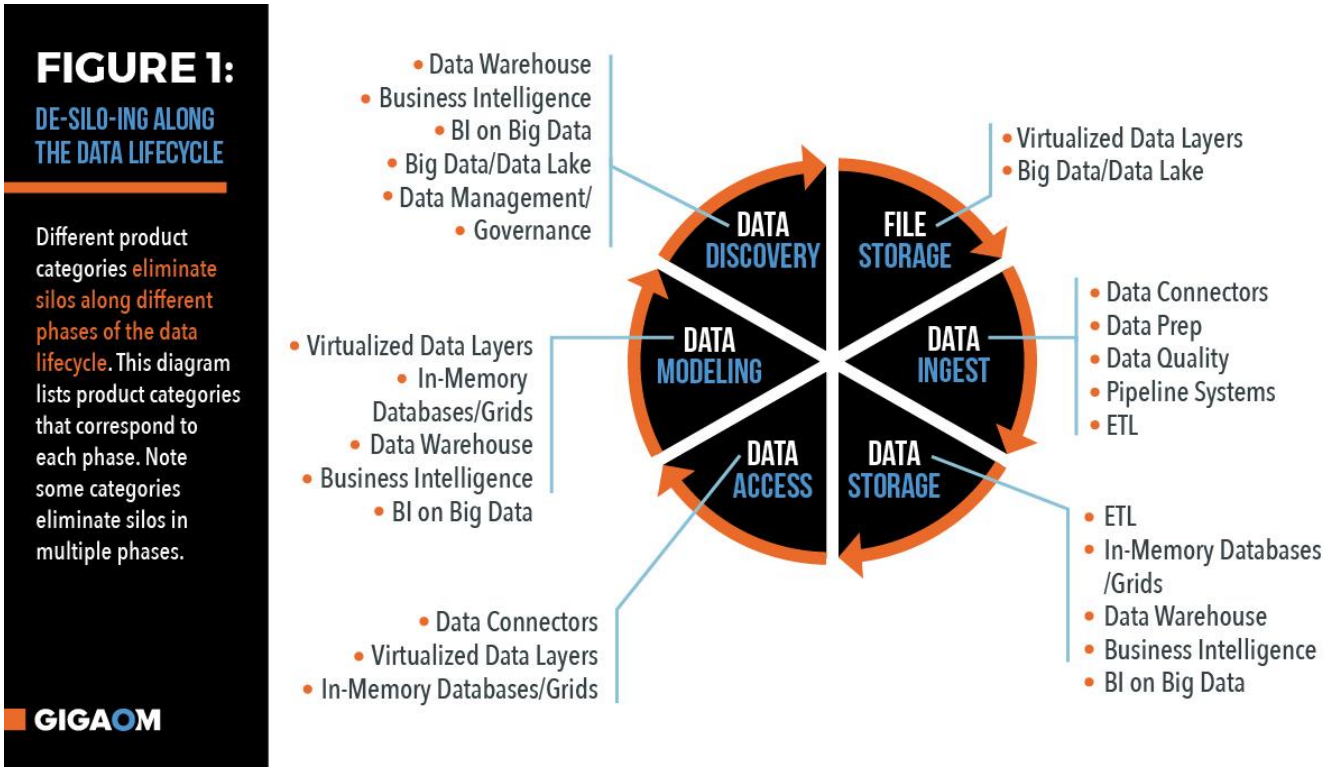
According to a McKinsey report, Big Data technologies could reduce research and development costs for pharmaceutical makers by $40 billion to $70 billion. The FDA and NIH use Big Data technologies to access large amounts of data to evaluate drugs and treatment.



Scientific Research

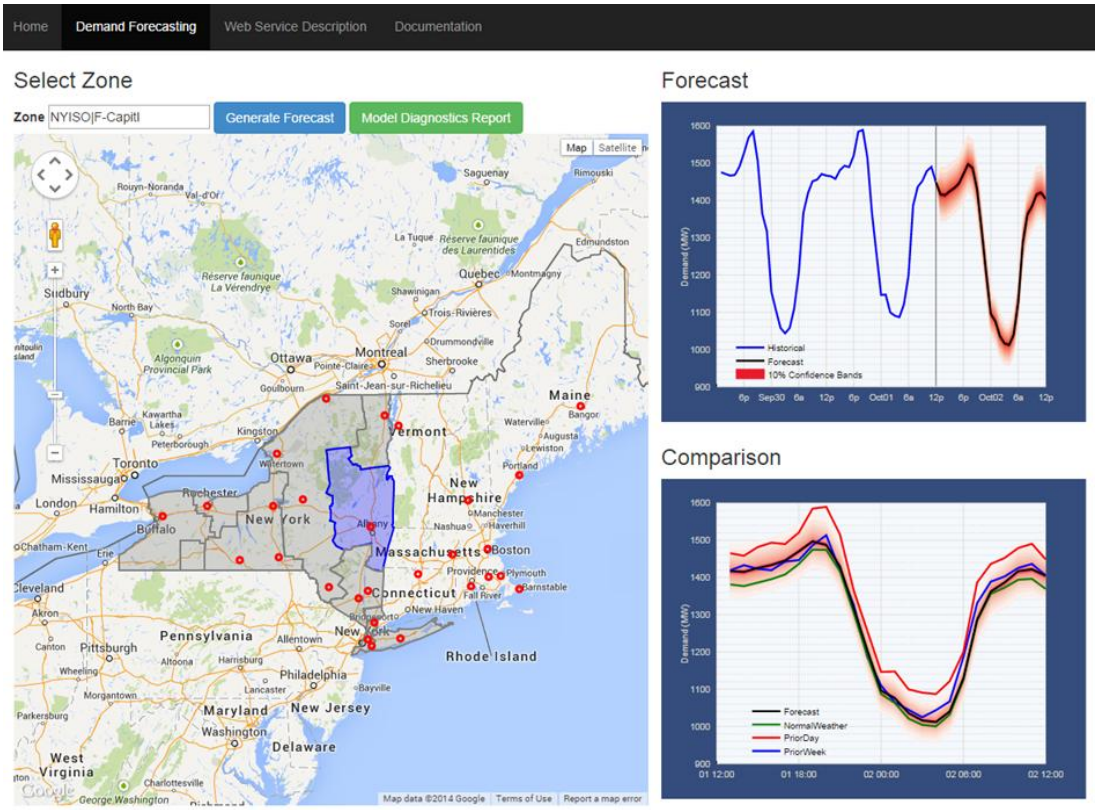
The National Science Foundation has initiated a long-term plan to:

- Implement new methods for deriving knowledge from data
- Develop new approaches to education
- Create a new infrastructure to "manage, curate, and serve data to communities".



Weather Forecasting

The NOAA (National Oceanic and Atmospheric Administration) gathers data every minute of every day from land, sea, and space-based sensors. Daily NOAA uses Big Data to analyze and extract value from over 20 terabytes of data.



Tax Compliance

Big Data Applications can be used by tax organizations to analyze both unstructured and structured data from a variety of sources in order to identify suspicious behavior and multiple identities. This would help in tax fraud identification. (14)

Traffic Optimization

Big Data helps in aggregating real-time traffic data gathered from road sensors, GPS devices and video cameras. The potential traffic problems in dense areas can be prevented by adjusting public transportation routes in real time.

United States of America

In 2012, the Obama administration announced the Big Data Research and Development Initiative, to explore how big data could be used to address important problems faced by the government. The initiative is composed of 84 different big data programs spread across six departments.

- Big data analysis played a large role in Barack Obama's successful 2012 re-election campaign.
- The United States Federal Government owns six of the ten most powerful supercomputers in the world.
- The Utah Data Center is a data center currently being constructed by the United States National Security Agency. When finished, the facility will be able to handle a large amount of information collected by the NSA over the Internet.

Storage space is unknown, but more recent sources claim it will be on the order of a few exabytes.

India

- Big data analysis was, in parts, responsible for the BJP and its allies to win a highly successful Indian General Election 2014.
- The Indian Government utilises numerous techniques to ascertain how the Indian electorate is responding to government action, as well as ideas for policy augmentation

United Kingdom

Examples of uses of big data in public services:

- Data on prescription drugs: by connecting origin, location and the time of each prescription, a research unit was able to exemplify the considerable delay between the release of any given drug, and a UK-wide adaptation of the National Institute for Health and Care Excellence guidelines. This suggests that new/most up-to-date drugs take some time to filter through to the general patient.
- Joining up data: a local authority blended data about services, such as road gritting rotas, with services for people at risk, such as 'meals on wheels'. The connection of data allowed the local authority to avoid any weather related delay.

International development

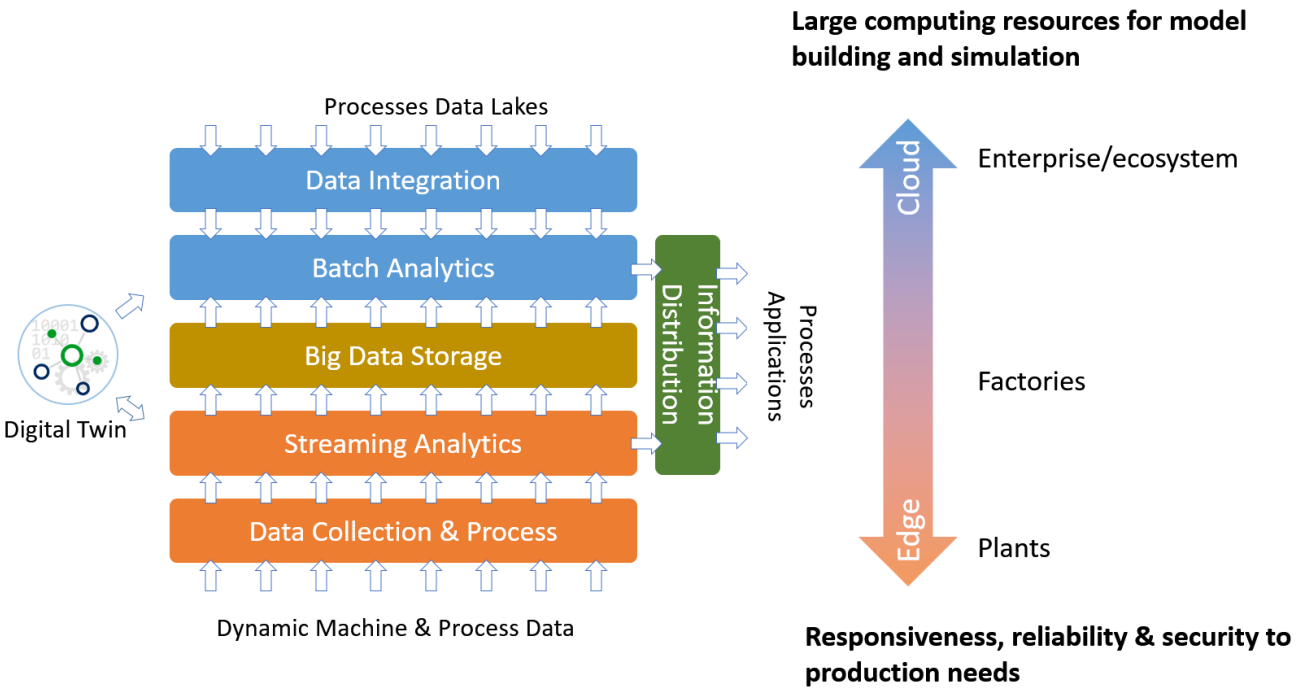
Research on the effective usage of information and communication technologies for development (also known as ICT4D) suggests that big data technology can make important contributions but also present unique challenges to International development. Advancements in big data analysis offer cost-effective opportunities to improve decision-making in critical development areas such as health care, employment, economic productivity, crime, security, and natural disaster and resource management. However, longstanding challenges for developing regions such as inadequate technological infrastructure and economic and human resource scarcity exacerbate existing concerns with big data such as privacy, imperfect methodology, and interoperability issues.

Media

Internet of Things (IoT)

To understand how the media utilises Big Data, it is first necessary to provide some context into the mechanism used for the media process.(15) It has been suggested by Nick Couldry and Joseph Turow that practitioners in Media and Advertising approach big data as many actionable points of information about millions of individuals. The industry appears to be moving away from the traditional approach of using specific media environments such as newspapers, magazines, or television shows and instead tap into consumers with technologies that reach targeted people at optimal times in optimal locations. The ultimate aim is to serve, or convey, a message or content that is (statistically speaking) in line with the consumers mindset. For example, publishing environments are increasingly tailoring messages (advertisements) and content (articles) to appeal to consumers that have been exclusively gleaned through various data-mining activities.

- Targeting of consumers (for advertising by marketers)
- Data-capture



Big Data and the IoT work in conjunction. From a media perspective, data is the key derivatives of device inter connectivity and allows accurate targeting. The Internet of Things, with the help of big data, therefore transforms the media industry, companies and even governments, opening up a new era of economic growth and competitiveness. The intersections of people, data and intelligent algorithms have far-reaching impacts on media efficiency. The wealth of data generated allows an elaborate layer on the present targeting mechanisms of the industry. (16)

Technology

- EBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay's 90PB data warehouse
- Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.
- Facebook handles 50 billion photos from its user base.
- As of August 2012, Google was handling roughly 100 billion searches per month.

Private sector

Retail

- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.

Retail Banking

- FICO Card Detection System protects accounts world-wide.
- The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.

Real Estate

- Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

Stock market prediction

It is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. Others disagree and those with this viewpoint possess myriad methods and technologies which purportedly allow them to gain future price information.

Prediction methods

Prediction methodologies fall into three broad categories which can (and often do) overlap. They are fundamental analysis, technical analysis (charting) and technological methods.

Fundamental analysis

Fundamental Analysts are concerned with the company that underlies the stock itself. They evaluate a company's

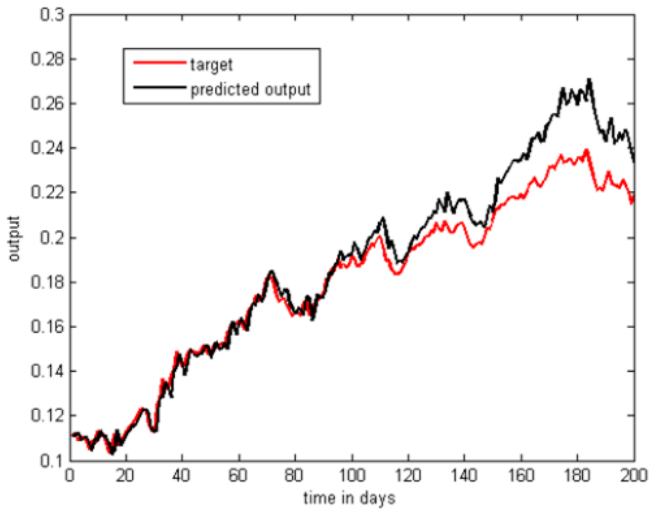
past performance as well as the credibility of its accounts. Many performance ratios are created that aid the fundamental analyst with assessing the validity of a stock, such as the P/E ratio. Warren Buffett is perhaps the most famous of all Fundamental Analysts.

Fundamental analysis is built on the belief that human society needs capital to make progress and if a company operates well, it should be rewarded with additional capital and result in a surge in stock price. Fundamental analysis is widely used by fund managers as it is the most reasonable, objective and made from publicly available information like financial statement analysis.

Another meaning of fundamental analysis is beyond bottom-up company analysis, it refers to top-down analysis from first analyzing the global economy, followed by country analysis and then sector analysis, and finally the company level analysis.

Technical analysis

Technical analysts or chartists are not concerned with any of the company's fundamentals. They seek to determine the future price of a stock based solely on the trends of the past price (a form of time series analysis). Numerous patterns are employed such as the head and shoulders or cup and saucer. Alongside the patterns, techniques are used such as the exponential moving average (EMA). Candle stick patterns, believed to have been first developed by Japanese rice merchants, are nowadays widely used by technical analysts.



Science

The Large Hadron Collider experiments represent about 150 million sensors delivering data 40 million times per second. There are nearly 600 million collisions per second. After filtering and refraining from recording more than 99.99995% of these streams, there are 100 collisions of interest per second.

- As a result, only working with less than 0.001% of the sensor stream data, the data flow from all four LHC experiments represents 25 petabytes annual rate before replication (as of 2012). This becomes nearly 200 petabytes after replication.
- If all sensor data were to be recorded in LHC, the data flow would be extremely hard to work with. The data flow would exceed 150 million petabytes annual rate, or nearly 500 exabytes per day, before replication. To put the number in perspective, this is equivalent to 500 quintillion (5×1020) bytes per day, almost 200 times more than all the other sources combined in the world.

The Square Kilometre Array is a telescope which consists of millions of antennas and is expected to be operational by 2024. Collectively, these antennas are expected to gather 14 exabytes and store one petabyte per day. It is considered to be one of the most ambitious scientific projects ever undertaken.

Science and Research

- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.
- When the Sloan Digital Sky Survey (SDSS) began collecting astronomical data in 2000, it amassed more in its first few weeks than all data collected in the history of astronomy. Continuing at a rate of about 200 GB per night, SDSS has amassed more than 140 terabytes of information. When the Large Synoptic Survey Telescope, successor to SDSS, comes online in 2016 it is anticipated to acquire that amount of data every five days.
- Decoding the human genome originally took 10 years to process, now it can be achieved in less than a day: the DNA sequencers have divided the sequencing cost by 10, 000 in the last ten years, which is 100 times cheaper than the reduction in cost predicted by Moore's Law.

Hadoop

Hadoop is sub-project of Lucene (a collection of industrial-strength search tools), under the umbrella of the Apache Software Foundation. Hadoop parallelizes data processing across many nodes (computers) in a compute cluster, speeding up large computations and hiding I/O latency through increased concurrency. Hadoop is especially well-suited to large data processing tasks (like searching and indexing) because it can leverage its distributed file system to cheaply and reliably replicate chunks of data to nodes in the cluster, making data available locally on the machine that is processing it. Hadoop is written in Java. Hadoop programs can be written using a small API in Java or Python. Hadoop can also run binaries and shell scripts on nodes in the cluster provided that they conform to a particular convention for string input/output. Hadoop provides to the application programmer the abstraction of map and reduce (which may be familiar to those with functional programming experience). Map and reduce are available in many languages, such as Lisp and Python.

Hadoop Applications

Making Hadoop Applications More Widely Accessible

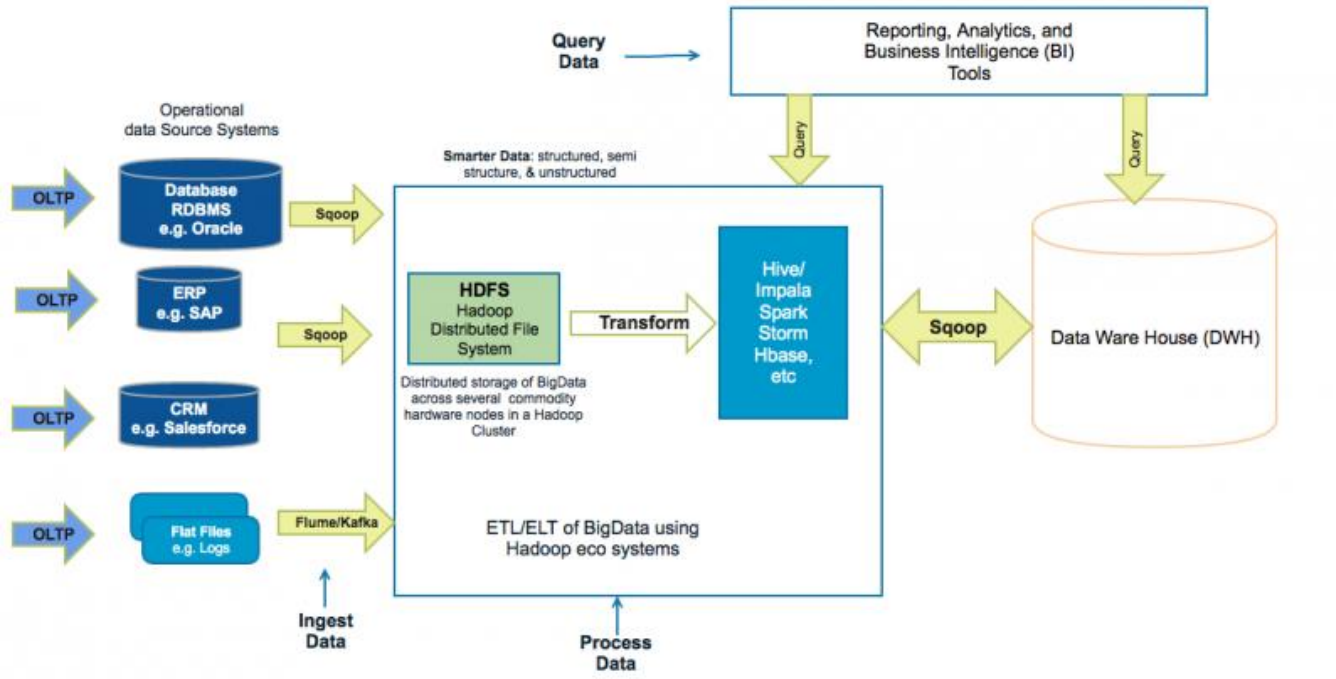
Apache Hadoop, the open source MapReduce framework, has dramatically lowered the cost barriers to processing and analyzing big data. Technical barriers remain, however, since Hadoop applications and technologies are highly complex and still foreign to most developers and data analysts. Talend, the open source integration company, makes the massive computing power of Hadoop truly accessible by making it easy to work with Hadoop applications and to incorporate Hadoop (17)into enterprise data flows.

A Graphical Abstraction Layer on Top of Hadoop Applications

In keeping with our history as an innovator and leader in open source data integration, Talend is the first provider to offer a pure open source solution to enable big data integration. Talend Open Studio for Big Data, by layering an easy to use graphical development environment on top of powerful Hadoop applications, makes big data management accessible to more companies and more developers than ever before. With its Eclipse-based graphical workspace, Talend Open Studio for Big Data enables the developer and data scientist to leverage Hadoop loading and processing technologies like HDFS, HBase, Hive, and Pig without having to write Hadoop application code. By simply selecting graphical components from a palette, arranging and configuring them, you can create.

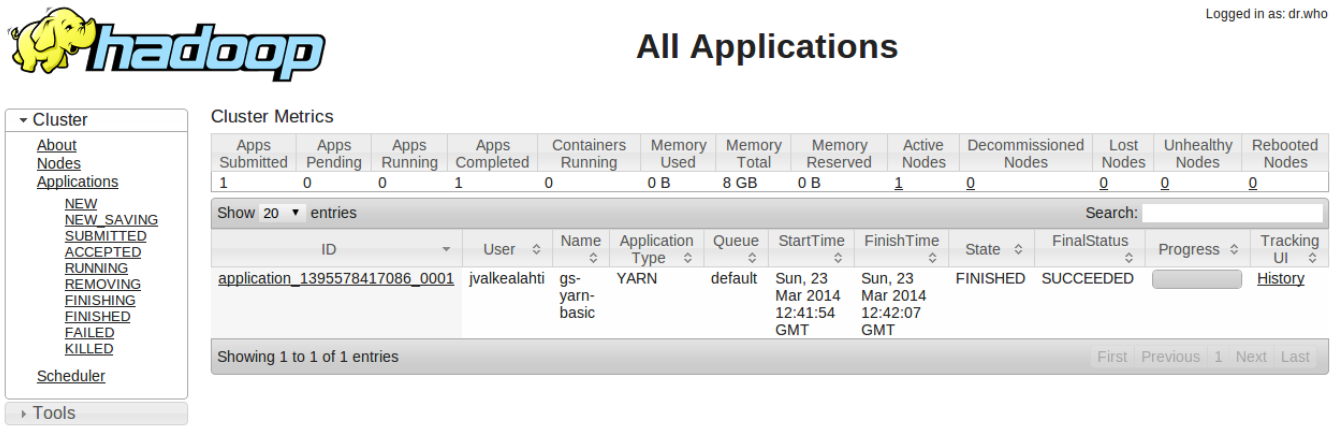
Hadoop jobs that, for example:

- Load data into HDFS (Hadoop Distributed File System)
- Use Hadoop Pig to transform data in HDFS
- Load data into a Hadoop Hive based data warehouse
- Perform ELT (extract, load, transform) aggregations in Hive
- Leverage Sqoop to integrate relational databases and Hadoop



Hadoop Applications, Seamlessly Integrated

For Hadoop applications to be truly accessible to your organization, they need to be smoothly integrated into your overall data flows. Talend Open Studio for Big Data is the ideal tool for integrating Hadoop applications into your broader data architecture. Talend provides more built-in connector components than any other data integration solution available, with more than 800 connectors that make it easy to read from or write to any major file format, database, or packaged enterprise application. For example, in Talend Open Studio for Big Data, you can use drag 'n drop configurable components to create data integration flows that move data from delimited log files into Hadoop Hive, perform operations in Hive, and extract data from Hive into a MySQL database (or Oracle, Sybase, SQL Server, and so on). Big Data is open source software, free to download and use under an Apache license.



Hadoop Architecture

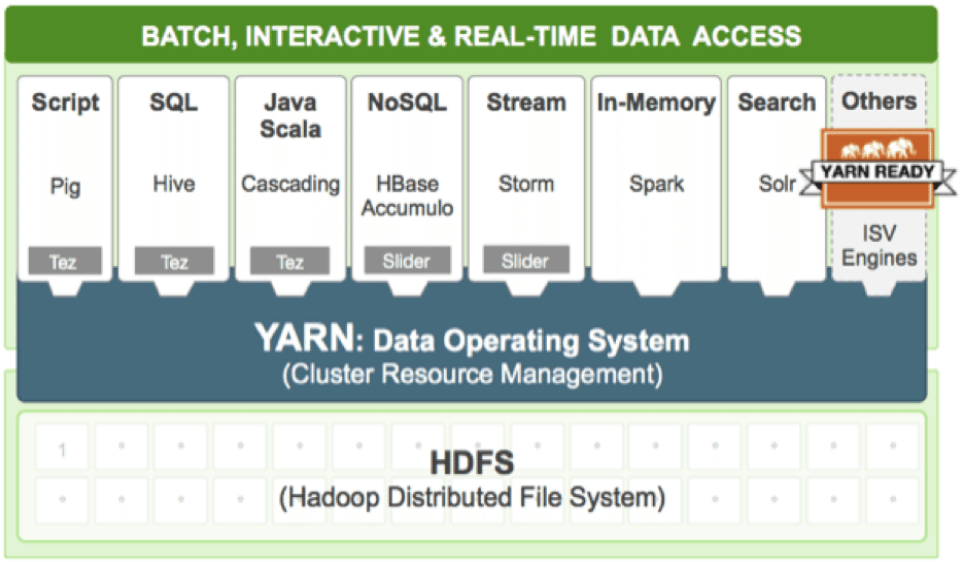
Hadoop framework includes following four modules: (18)

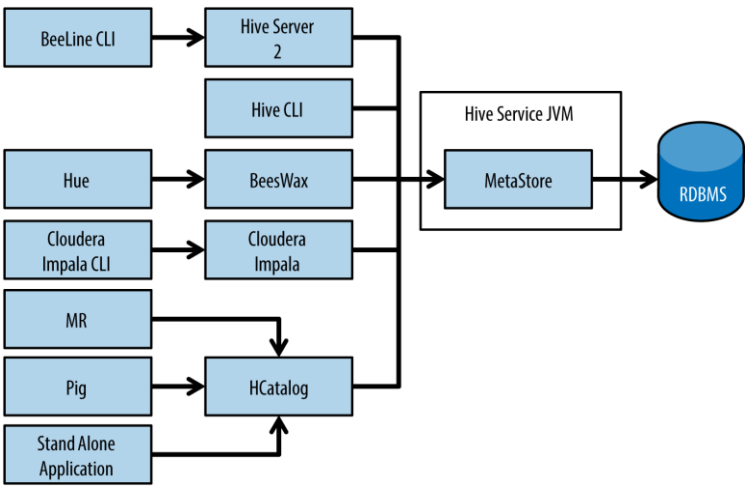
Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

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 a YARN-based system for parallel processing of large data sets. We can use the following diagram to depict these four components available in Hadoop framework.



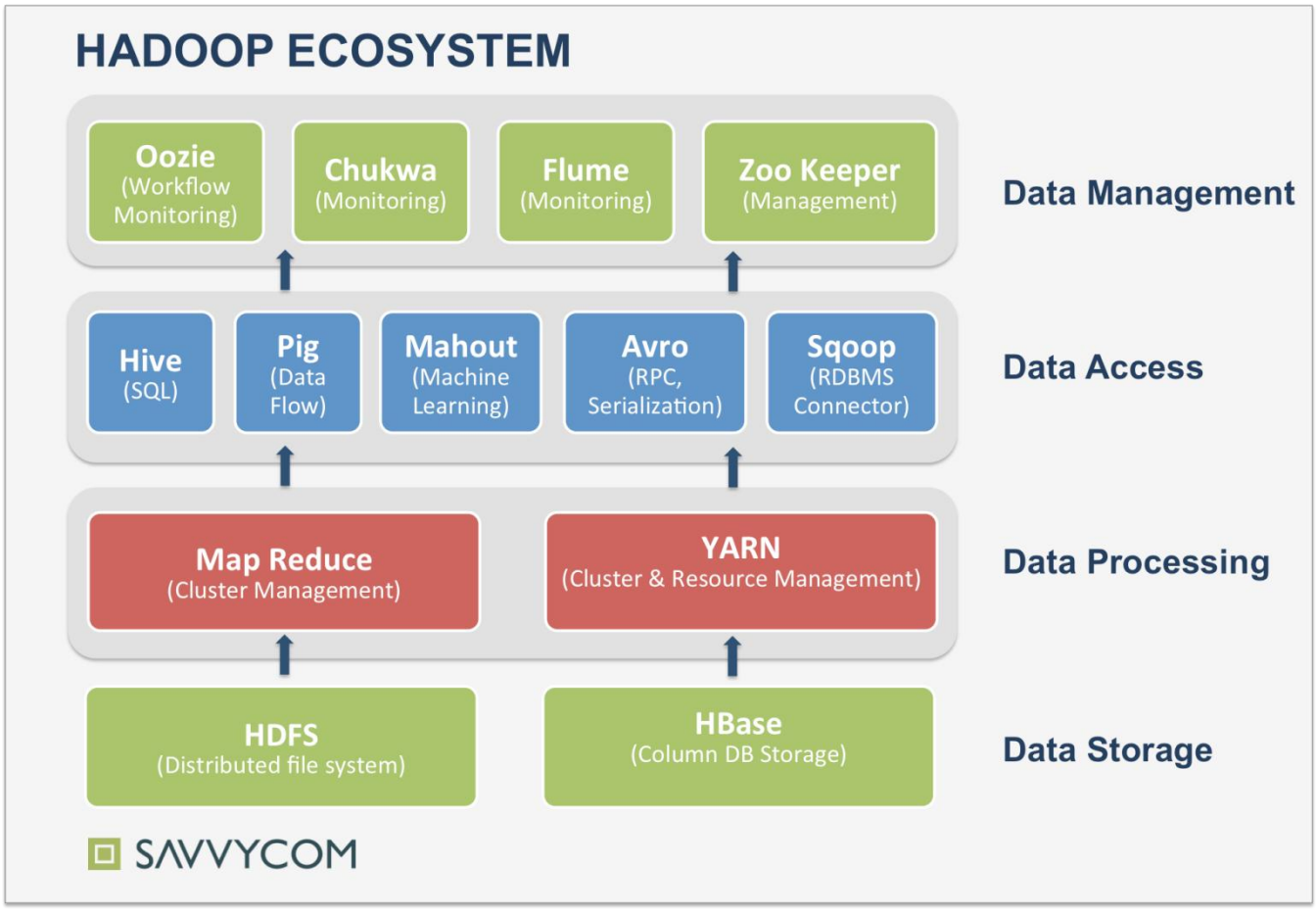


Since 2012, the term "Hadoop" often refers not just to the base modules mentioned above but also to the collection of additional software packages that can be installed on top of or alongside Hadoop, such as Apache Pig, Apache Hive, Apache HBase, Apache Spark etc.

Hadoop Ecosystem Overview (19)

Remember that Hadoop is a framework. If Hadoop was a house, it wouldn't be a very comfortable place to live. It would provide walls, windows, doors, pipes, and wires. The Hadoop ecosystem provides the furnishings that turn the framework into a comfortable home for big data activity that reflects your specific needs and tastes.

The Hadoop ecosystem includes both official Apache open source projects and a wide range of commercial tools and solutions. Some of the best-known open source examples include Spark, Hive, Pig, Oozie and Sqoop. Commercial Hadoop offerings are even more diverse and include platforms and packaged distributions from vendors such as Cloudera, Hortonworks, and MapR, plus a variety of tools for specific Hadoop development, production, and maintenance tasks.



Most of the solutions available in the Hadoop ecosystem are intended to supplement one or two of Hadoop's four core elements (HDFS, MapReduce, YARN, and Common). However, the commercially available framework solutions provide more comprehensive functionality. The sections below provide a closer look at some of the more prominent components of the Hadoop ecosystem, starting with the Apache projects

Apache open source Hadoop ecosystem elements

The Apache Hadoop project actively supports multiple projects intended to extend Hadoop's capabilities and make it easier to use. There are several top-level projects to create development tools as well as for managing Hadoop data flow and processing. Many commercial third-party solutions build on the technologies developed within the Apache Hadoop ecosystem.

Spark, Pig, and Hive are three of the best-known Apache Hadoop projects. Each is used to create applications to process Hadoop data. While there are a lot of articles and discussions about whether Spark, Hive or Pig is better, in practice many organizations do not only use a single one because each is optimized for specific functions.(20)

Spark

Spark is both a programming model and a computing model. It provides a gateway to in-memory computing for Hadoop, which is a big reason for its popularity and wide adoption. Spark provides an alternative to MapReduce that enables workloads to execute in memory, instead of on disk. Spark accesses data from HDFS but bypasses the MapReduce processing framework, and thus eliminates the resource-intensive disk operations that MapReduce requires. By using in-memory computing, Spark

workloads typically run between 10 and 100 times faster compared to disk execution.

Spark can be used independently of Hadoop. However, it is used most commonly with Hadoop as an alternative to MapReduce for data processing. Spark can easily coexist with MapReduce and with other ecosystem components that perform other tasks.

Spark is also popular because it supports SQL, which helps overcome a shortcoming in core Hadoop technology. The Spark programming environment works interactively with Scala, Python, and R shells. It has been used for data extract/transform/load (ETL) operations, stream processing, machine learning development and with the Apache GraphX API for graph computation and display. Spark can run on a variety of Hadoop and non-Hadoop clusters, including Amazon S3.

Hive

Hive is data warehousing software that addresses how data is structured and queried in distributed Hadoop clusters. Hive is also a popular development environment that is used to write queries for data in the Hadoop environment. It provides tools for ETL operations and brings some SQL-like capabilities to the environment. Hive is a declarative language that is used to develop applications for the Hadoop environment; however it does not support real-time queries.

Hive has several components, including:

- **HCatalog** – Helps data processing tools read and write data on the grid. It supports MapReduce and Pig.
- **WebHCat** – Lets you use an HTTP/REST interface to run MapReduce, Yarn, Pig, and Hive jobs.
- **HiveQL** – Hive's query language intended as a way for SQL developers to easily work in Hadoop. It is similar to SQL and helps both structure and query data in distributed Hadoop clusters.

Hive queries can run from the Hive shell, JDBC, or ODBC. MapReduce (or an alternative) breaks down HiveQL statements for execution across the cluster.

Hive also allows MapReduce-compatible mapping and reduction software to perform more sophisticated functions. However, Hive does not allow row-level updates or support for real-time queries, and it is not intended for OLTP workloads. Many consider Hive to be much more effective for processing structured data than unstructured data, for which Pig is considered advantageous.

Pig

Pig is a procedural language for developing parallel processing applications for large data sets in the Hadoop environment. Pig is an alternative to Java programming for MapReduce, and automatically generates MapReduce functions. Pig includes Pig Latin, which is a scripting language. Pig translates Pig Latin scripts into MapReduce, which can then run on YARN and process data in the HDFS cluster. Pig is popular because it automates some of the complexity in MapReduce development.

Pig is commonly used for complex use cases that require multiple data operations. It is more of a processing language than a query language. Pig helps develop applications that aggregate and sort data and supports multiple inputs and exports. It is highly customizable, because users can write their own functions using their preferred scripting language. Ruby, Python and even Java are all supported. Thus, Pig has been a popular option for developers that are familiar with those languages but not with MapReduce. However, SQL developers may find Hive easier to learn.

HBase

HBase is a scalable, distributed, NoSQL database that sits atop the HFDS. It was designed to store structured data in tables that could have billions of rows and millions of columns. It has been deployed to power historical searches through large data sets, especially when the desired data is contained within a large amount of unimportant or irrelevant data (also known as sparse data sets). It is also an underlying technology behind several large messaging applications, including Facebook's.

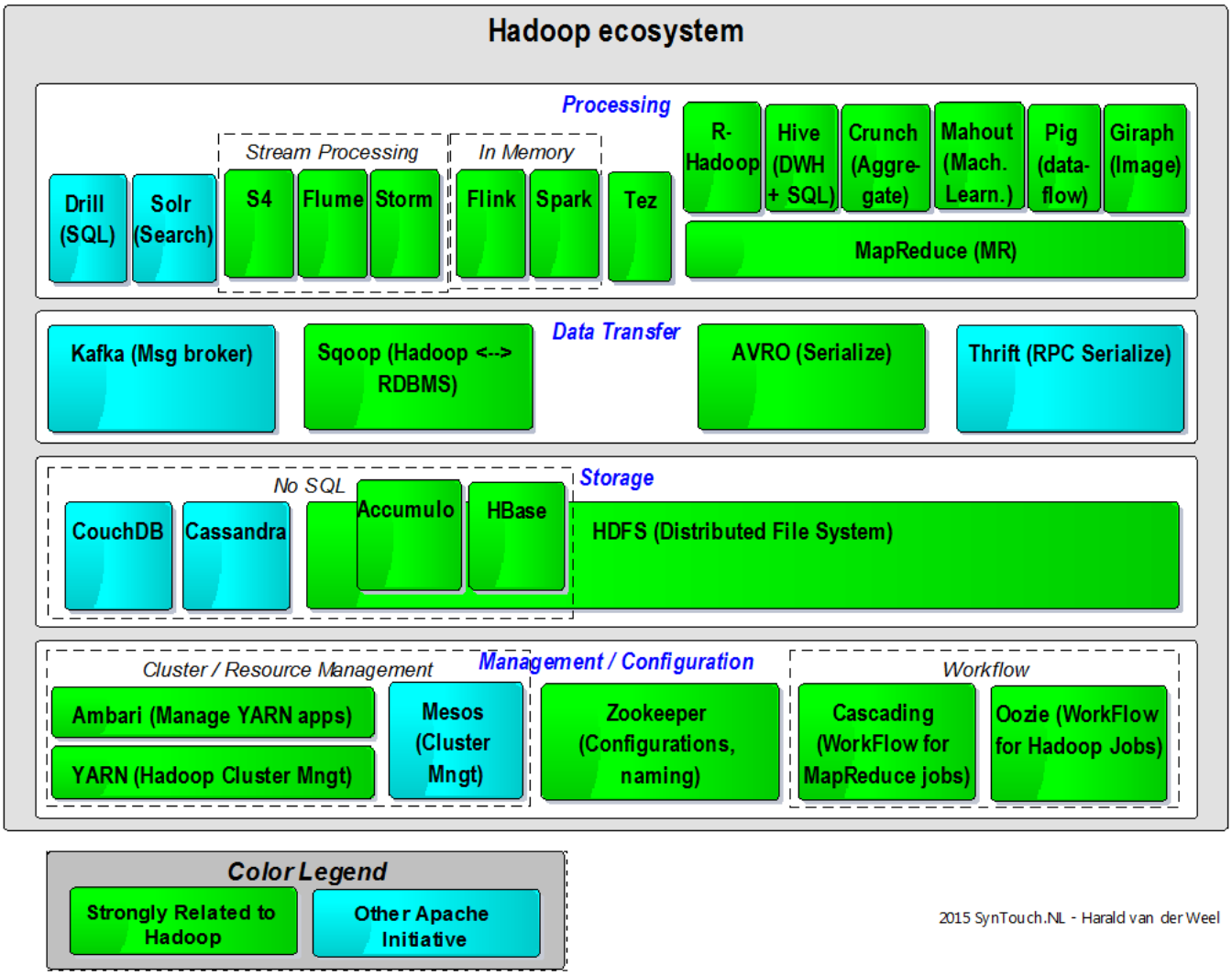
HBase is not a relational database and wasn't designed to support transactional and other real-time applications. It is accessible through a Java API and has ODBC and JDBC drivers. HBase does not support SQL queries, however there are several SQL support tools available from the Apache project and from software vendors. For example, Hive can be used to run SQL-like queries in HBase.

Oozie

Oozie is the workflow scheduler that was developed as part of the Apache Hadoop project. It manages how workflows start and execute, and also controls the execution path. Oozie is a server-based Java web application that uses workflow definitions written in hPDL, which is an XML Process Definition Language similar to JBOSS JBPM jPDL. Oozie only supports specific workflow types, so other workload schedulers are commonly used instead of or in addition to Oozie in Hadoop environments.

Sqoop

Think of Sqoop as a front-end loader for big data. Sqoop is a command-line interface that facilitates moving bulk data from Hadoop into relational databases and other structured data stores. Using Sqoop replaces the need to develop scripts to export and import data. One common use case is to move data from an enterprise data warehouse to a Hadoop cluster for ETL processing. Performing ETL on the commodity Hadoop cluster is resource efficient, while Sqoop provides a practical transfer method.



Other Apache Hadoop-related open source projects

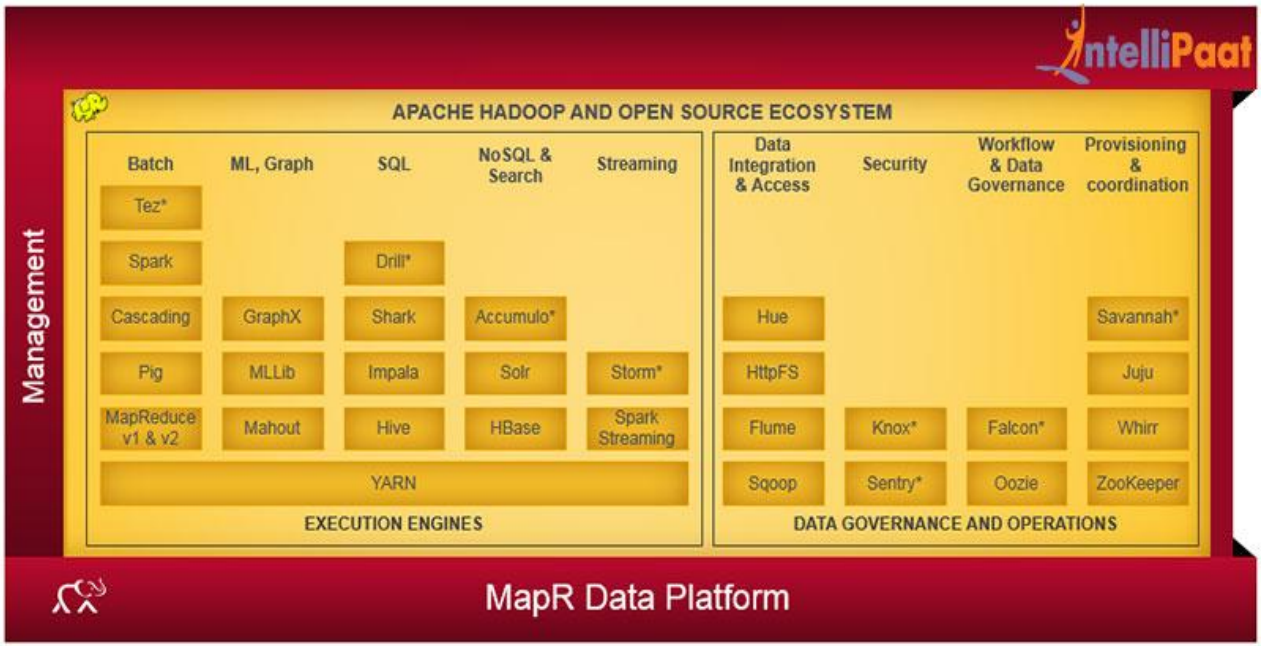
Here is how the Apache organization describes some of the other components in its Hadoop ecosystem.

- **Ambari** – A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig, and Sqoop.
- **Avro** – A data serialization system.
- **Cassandra** – A scalable multi-master database with no single points of failure.
- **Chukwa** – A data collection system for managing large distributed systems.
- **Impala** – The open source, native analytic database for Apache Hadoop. Impala is shipped by Cloudera, MapR, Oracle, and Amazon.
- **Flume** – A distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming event data.
- **Kafka** – A messaging broker that is often used in place of traditional brokers in the Hadoop environment because it is designed for higher throughput and provides replication and greater fault tolerance.
- **Mahout** – A scalable machine learning and data mining library.
- **Tajo** – A robust big data relational and distributed data warehouse system for Apache Hadoop. Tajo is designed for low-latency and scalable ad-hoc queries, online aggregation, and ETL on large-data sets stored on HDFS and other data sources. By supporting SQL standards and leveraging advanced database techniques, Tajo allows direct control of distributed execution and data flow across a variety of query evaluation strategies and optimization opportunities.
- **Tez** – A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive, Pig and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace MapReduce as the underlying execution engine.
- **Zookeper** – A high-performance coordination service for distributed applications.

The ecosystem elements described above are all open source Apache Hadoop projects. There are numerous commercial solutions that use or support the open source Hadoop projects. Some of the more prominent ones are described in the following sections.

Commercial Hadoop distributions

Hadoop can be downloaded from www.hadoop.apache.org and used for free, which thousands of organizations have done. There are also commercial distributions that combine core Hadoop technology with additional features, functionality and documentation. The leading commercial distribution Hadoop vendors include Cloudera, Hortonworks, and MapR. There are also many less comprehensive, more task-specific tools for the Hadoop environment, such as developer tools and job schedulers.



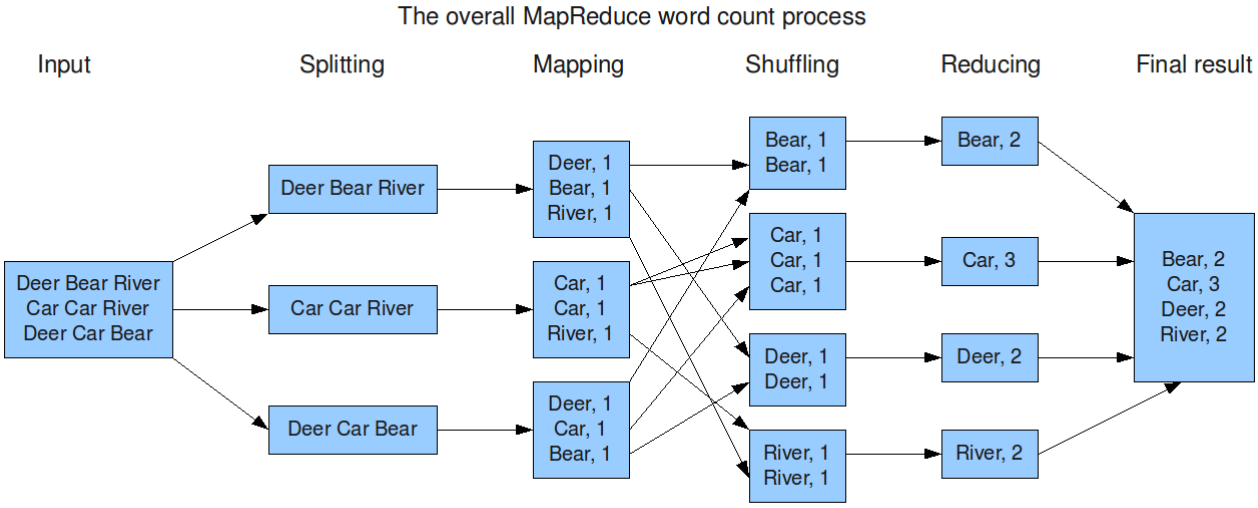
MapReduce

Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The term MapReduce actually refers to the following two different tasks that Hadoop programs perform: (21)

The Map Task: This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples (key/value pairs).

The Reduce Task: This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task. Typically both the input and the output are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks. The MapReduce framework consists of a

single master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves TaskTracker execute the tasks as directed by the master and provide task-status information to the master periodically. The JobTracker is a single point of failure for the Hadoop MapReduce service which means if JobTracker goes down, all running jobs are halted.



Importance of Hadoop

Ability to store and process huge amounts of any kind of data, quickly. With data volumes and varieties constantly increasing, especially from social media and the Internet of Things (IoT), that's a key consideration.

Computing power
Hadoop's distributed computing model processes big data fast. The more computing nodes you use, the more processing power you have.

Fault tolerance

Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. Multiple copies of all data are stored automatically.

Flexibility

Unlike traditional relational databases, you don’t have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.

Low cost

The open-source framework is free and uses commodity hardware to store large quantities of data.

Scalability

You can easily grow your system to handle more data simply by adding nodes. Little administration is required.

Hadoop Distributed File System

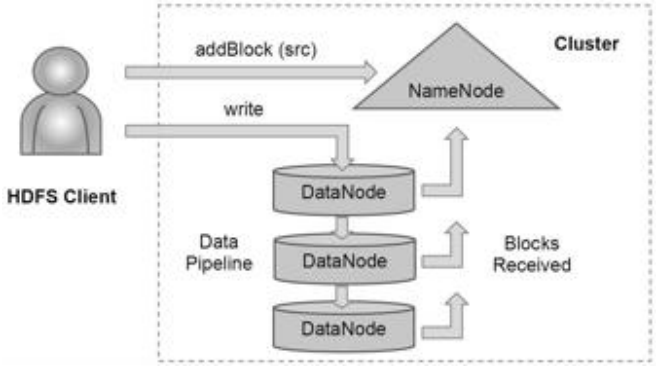
Hadoop can work directly with any mountable distributed file system such as Local FS, HFTP FS, S3 FS, and others, but the most common file system used by Hadoop is the Hadoop Distributed File System (HDFS).

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner.

HDFS uses a master/slave architecture where master consists of a single NameNode that manages the file system metadata and one or more slave DataNodes that store the actual data.

A file in an HDFS namespace is split into several blocks and those blocks are stored in a set of DataNodes. The NameNode determines the mapping of blocks to the DataNodes. The DataNodes takes care of read and write operation with the file system. They also take care of block creation, deletion and replication based on instruction given by NameNode.

HDFS provides a shell like any other file system and a list of commands are available to interact with the file system. These shell commands will be covered in a separate chapter along with appropriate examples.



Working procedure of Hadoop

Stage 1

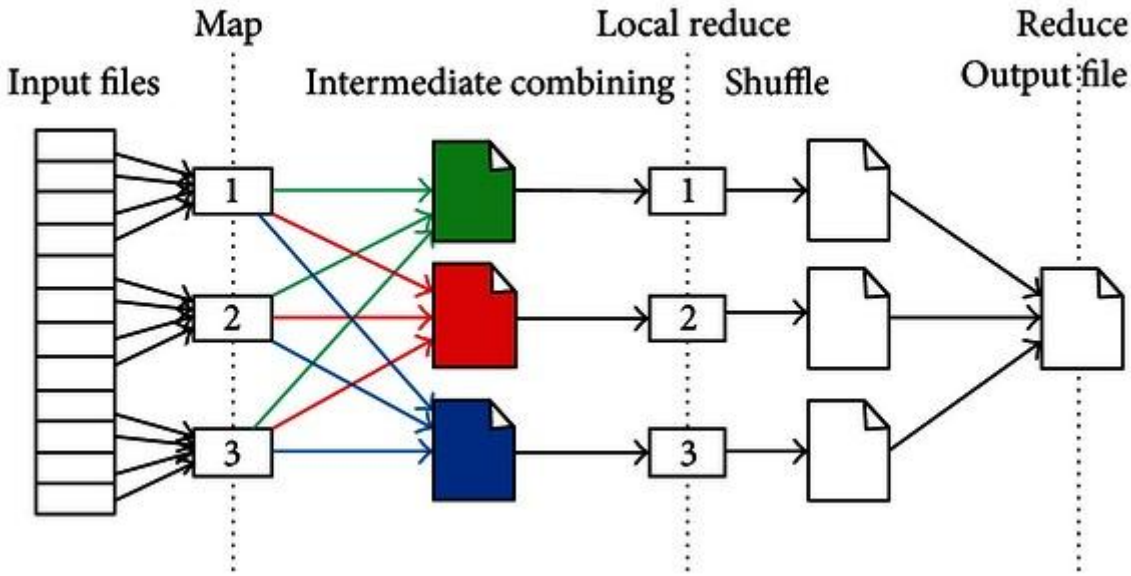
A user/application can submit a job to Hadoop (a Hadoop job client) for the required process by specifying the following items: 1. The location of the input and output files in the distributed file system. 2. The java classes are in the form of a jar file containing the implementation of map and reduce functions. 3. The job configuration by setting different parameters specific to the job.

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.

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.



Advantages of Hadoop (22)

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 many thousands of terabytes of data.

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.

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. Hadoop can be used for a wide variety of purposes, such as log processing,

recommendation systems, data warehousing, market campaign analysis and fraud detection.

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.

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.

Disadvantages of Hadoop:

As the backbone of so many implementations, Hadoop is almost synomous with big data.

1. Security Concerns

Just managing complex applications such as Hadoop can be challenging. A simple example can be seen in the Hadoop security model, which is disabled by default due to sheer complexity. If whoever managing the platform lacks how to enable it, your data could be at huge risk. Hadoop is also missing encryption at the storage and network levels, which is a major selling point for government agencies and others that prefer to keep their data under wraps.

2. Vulnerable By Nature

Speaking of security, the very makeup of Hadoop makes running it a risky proposition. The framework is written almost entirely in Java, one of the most widely used yet controversial programming languages in existence. Java has been heavily exploited by cybercriminals and as a result, implicated in numerous security breaches.

3. Not Fit for Small Data

While big data is not exclusively made for big businesses, not all big data platforms are suited for small data needs. Unfortunately, Hadoop happens to be one of them. Due to its high capacity design, the Hadoop Distributed File System, lacks the ability to efficiently support the random reading of small files. As a result, it is not recommended for organizations with small quantities of data.

4. Potential Stability

Issues Like all open source software, Hadoop has had its fair share of stability issues. To avoid these issues, organizations are strongly recommended to make sure they are running the latest stable version, or run it under a third-party vendor equipped to handle such problems.

5. General Limitations

The article introduces Apache Flume, MillWheel, and Google's own Cloud Dataflow as possible solutions. What each of these platforms has in common is the ability to improve the efficiency and reliability of data collection, aggregation, and integration. The main point the article stresses is that companies could be missing out on big benefits by using Hadoop alone.

Usages of Hadoop

Low-cost storage and data archive

The modest cost of commodity hardware makes Hadoop useful for storing and combining data such as transactional, social media, sensor, machine, scientific, click streams, etc. The low-cost storage lets you keep information that is not deemed currently critical but that you might want to analyze later.

Sandbox for discovery and analysis

Because Hadoop was designed to deal with volumes of data in a variety of shapes and forms, it can run analytical algorithms. Big data analytics on Hadoop can help your organization operate more efficiently uncover new opportunities and derive next-level competitive advantage. The sandbox approach provides an opportunity to innovate with minimal investment.

Data lake

Data lakes support storing data in its original or exact format. The goal is to offer a raw or unrefined view of data to data scientists and analysts for discovery and analytics. It helps them ask new or difficult questions without constraints.

Data lakes are not a replacement for data warehouses. In fact, how to secure and govern data lakes is a huge topic for IT. They may rely on data federation techniques to create a logical data structures.

Complement your data warehouse

We're now seeing Hadoop beginning to sit beside data warehouse environments, as well as certain data sets being offloaded from the data warehouse into Hadoop or new types of data going directly to Hadoop. The end goal for every organization is to have a right platform for storing and processing data of different schema, formats, etc. to support different use cases that can be integrated at different levels.

IoT and Hadoop

Things in the IoT need to know what to communicate and when to act. At the core of the IoT is a streaming, always on torrent of data. Hadoop is often used as the data store

for millions or billions of transactions. Massive storage and processing capabilities also allow you to use Hadoop as a sandbox for discovery and definition of patterns to be monitored for prescriptive instruction. You can then continuously improve these instructions, because Hadoop is constantly being updated with new data that doesn't match previously defined patterns.

2. Conclusion

The availability of Big Data, low-cost commodity hardware, and new information management and analytic software have produced a unique moment in the history of data analysis. The convergence of these trends means that we have the capabilities required to analyze astonishing data sets quickly and cost-effectively for the first time in history. These capabilities are neither theoretical nor trivial. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue, and profitability.

The Age of Big Data is here, and these are truly revolutionary times if both business and technology professionals continue to work together and deliver on the promise.

Reference

[1] www.google.com
[2] www.wikipedia.com
[3] Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–19 (2010)
[4] Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)
[5] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)
[6] Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)
[7] Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, pp. 1–24 (2012)
[8] http://www.ey.com/CA/en/Newsroom/News-releases/2014-Global-forensic-data-analytics-survey
[9] http://www.ikanow.com/how-can-i-use-big-data-analytics-for-fraud-detection/
[10] http://www.pactera.com/resources/blog/how-big-data-is-revolutionizing-fraud-detection-in-financial-services/
[11] Ruchi Verma, Sathyan Ramakrishna Mani, Using Analyrtics for Insurance Fraud Detection‖; FINsights, Infosys, Issue 10
[12] http://www-01.ibm.com/software/analytics/solutions/customer-analytics/social-media-analytics/
[13] http://www.news-sap.com/sentiment-analysis-with-big-data/
[14] https://datafloq.com/read/big-datas-impact-food-industry/96
[15] http://venturebeat.com/2013/03/01/ibm-brings-big-data-tech-to-food-to-prevent-the-next-horse-meat-scandal/
[16] http://www.forbes.com/sites/daniellegould/2012/09/24/food-industry-understand-trends-big-data-tools/
[17] Puneet Singh Duggal, Sanchita Paul, ― Big Data Analysis: Challenges and Solutions‖, International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV
[18] Marcin Jedyk, MAKING BIG DATA, SMALL, Using distributed systems for processing, analysing and managing large huge data sets, Software Professional's Network, Cheshire Data systems Ltd.
[19] S. Ghemawat, H. Gobioff, and S. Leung, The Google File System.‖ in ACM Symposium on Operating Systems Principles, Lake George, NY, Oct 2003, pp. 29 – 43.
[20] Jefry Dean and Sanjay Ghemwat, MapReduce: A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1, January 2010, pp 72-77.
[21] PIG Tutorial, YahooInc., http://developer.yahoo.com/hadoop/tutorial/pigtutorial.html
[22] IBM-What.is.Jaql, www.ibm.com/software/data/infosphere/hadoop/jaql/