Lecture Notes

Introduction to Big Data

Course

BE Computer Engineering

Semester – VIII

Subject

Big Data Analytics

Code: 2771607

Prepared by,

Prof. Yogesh M. Kapuriya

Assistant Professor,

Computer Engineering

C. K. Pithawalla College of Engineering and Technology,

Surat

Index

Motivation

Data Generation

- Day to day habits changed due to technological innovation
- More technology means more data generation
- Sources of data
  - Internet, mobile devices, industry equipment, environmental sensors, Internet of Things and many more
- This led us to deal with huge amount of data.
- Size only doesn't matters
  - It is not just about size of data, we are supposed to deal with rate at which it is generated, courtesy social networking, wireless sensors, biological records etc. It is about variety of form it takes; text, images, videos etc. It is about diversity of sources from where data is collected.
- Being a human being we always thought beyond data generation. Or we can say that we always wanted to figure out how available data can be utilized to generate meaningful information to gather important knowledge which can help us making wise decision for future actions.

Data Leads to Actions

There are lots of hurdles when you travel on a path from data to action. These hurdles includes collection, data integration, pre-processing, analysis, visualization, security etc. So there is a need of strong tools and technologies which can overcome this.

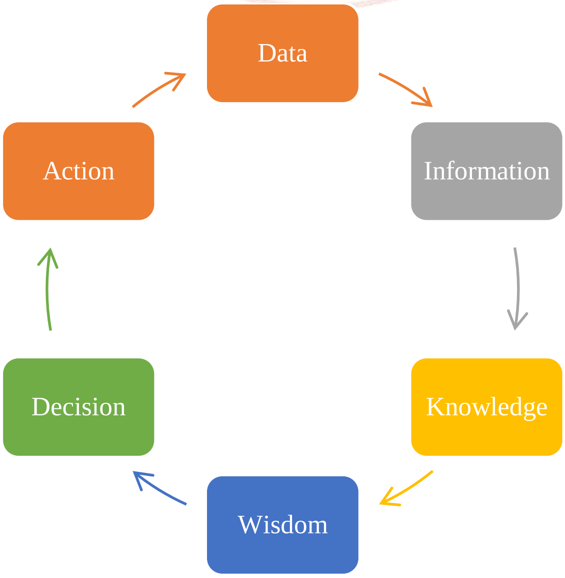

Figure 1. Big Data Process

Big Data Analytics

When Albert Einstein said, *"We can't solve problems by using the same kind of thinking we used when we created them"*, he may well have said the same about trying to manage the data explosion with the same technology that causes it. Conventional tools and methodologies are not empowered to deal with various big data requirements such as storage, transfer, transform, process and visualization etc.

Figure 2. Big Data Analytics

[Source: https://developers.googleblog.com/2016/02/introducing-autotrack-for-analyticsjs.html]

Thus we need new technological paradigm to deal with big data. Big data analytic is introduced as an answer to that. Even though the main purpose of big data analytics is analysis, but it also tries to find out solution to other associated issues also. Various domains like Health, Agriculture, Environment, Industrial, Social etc. almost all can make use of Big Data Analytic.

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyse your data and get answers from it almost immediately – an effort that's slower and less efficient with more traditional business intelligence solutions. This leads to the fact that big data affects organizations across practically every domain.

Due to its great value, big data has been essentially changing and transforming the way we live, work, and think. Jin et al. discusses importance of big data as,

- Significance to national development,
- Significance to industrial upgrades,
- Significance to scientific research,
- Significance to emerging interdisciplinary research,
- Significance to helping people better perceive the present
- Significance to helping people better predict the future.

Case Studies

McKinsey & Company observed how big data created values after in-depth research on the U.S. healthcare, the EU public sector administration, the U.S. retail, the global manufacturing, and the global personal location data.

During the 2009 flu pandemic, Google obtained timely information by analysing big data, which even provided more valuable information than that provided by disease prevention centre.

Microsoft purchased Farecast, a sci-tech venture company in the U.S. in 2008. Farecast has an airline ticket forecast system that predicts the trends and rising/dropping ranges of airline ticket price.

Consumer analytics is at the epicentre of a big data revolution. Technology helps capture rich and plentiful data on consumer phenomena in real time. Sunil Erevelles et al. proposed a conceptual framework to better understand the impact of big data on various marketing activities, enabling firms to better exploit its benefits.

Findings of Y. Wang can help healthcare organizations understand the big data analytics capabilities and potential benefits and support them seeking to formulate more effective data-driven analytics strategies.

Big Data

Definition

As such there is no universally accepted definition of Big Data. Wikipedia[1] defines it as "Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate."

SAS[2] defines it as "Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis."

Gartner[3] defines it as, "Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

So in general we can say that, big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.

Features of Big Data – 'V's

For proper understanding of term big data, it is often described as using five Vs: Volume, Velocity, Variety, Veracity and Value

- **Volume** refers to the vast amount of data generated every second. If we take all the data generated in the world between the beginning of time and the year 2000, it is the same amount generated every minute now. This increasingly makes data sets too large to store and analyse using traditional database technology.
- **Velocity** refers to the speed at which new data is generated and speed at which new data moves from one location to another. Imagine how social media messages going viral in minutes! This makes us to look for technologies which can analyse data even before it gets stored in database.
- **Variety** refers to the different type of data we can use now. Earlier we were dealing with structured data, which can be stored and processed in relational tables easily. But now most of the data are unstructured or at the most semi-structured, therefore cannot be put easily into the database. We are dealing with messages, social media conversations, photos, sensor data, and video or voice recordings etc. virtually on the single platform.

[1] https://en.wikipedia.org/wiki

[2] http://www.sas.com/en_in/home.html

[3] http://www.gartner.com/technology/home.jsp

- **Veracity** refers to the messiness or trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable. The volumes often make up for the lack of quality or accuracy.
- **Value** refers to our ability to turn our data into value. It is important that businesses make a case for any attempt to collect and leverage big data. This is why value is the one V of big data that matters the most.

Challenges of Big Data

The inherent behaviour or more precisely 4Vs (Volume, Velocity, Variety, Veracity) of big data raises many challenges for 5th V i.e. Value.

Conventional DBMS technology is not capable to handle big data. Volume requires scalability and parallelism that are beyond capability of DBMS. Variety of big data make it unfit to the restriction of closed processing architecture of current DBMS. Velocity asks for appropriate real-time efficiency which is beyond scope of DBMS. So the fundamental problem is to deal with features of big data such as heterogeneity; scale; speed; accuracy, trust and provenance; privacy crisis, interactivity and garbage mining etc.

Brief summary of challenges associated with handling big data is shown in Table 1.

Table 1. Summary of Challenges handling Big Data

Challenges	Sub-challenges/Tasks
Storage and Retrieval	Data Distribution, Data Transportation, Indexing etc.
Data Pre-processing	Data collection, Data filtering, Data quality, Data cleaning, Data reduction, Data representation, Data transformation etc.
Analytical	Data querying, Various analysis like Descriptive, Estimative, Predictive, Prescriptive etc., Scalability of algorithms
Security and Privacy	Secure computations, Secure data storage and transaction logs, End point input validation/filtering, Real time security/Compliance monitoring, Privacy preserving data mining and analytics, Access control and secure communication, Granular access control, Honeypot detection etc.
Other	Skill requirements, Statistical challenges, Spatial challenges, Interactiveness, Visualization, Garbage mining etc.

Big Data and Science: Myths and Reality

H. V. Jagadish et al. explored few myths about big data and discovers underlying truths as following;

- Only size matters – Variety and Veracity are far more challenging than Volume and Velocity.
- Challenge is to invent new computing architecture and algorithms – Facilitating human interaction to big data space is more challenging.
- Analytics is the central problem with big data – Many other steps involved in analysis pipeline e.g. Data acquisition, information extraction and cleaning, data integration, aggregation and representation, modelling and analysis, interpretation and visualization etc.
- Data reuse is low hanging fruit – Timeliness of data poses many challenges for data reuse.
- Data science is same as big data – However there exist a perspective difference between two; big data begins with data characteristics (and works up from there), Data science begins with data use (and works down from there).
- Big data is all hype – May be hyped, but there is more than enough substance there for it to deserve attentions.

Drivers for Big Data

There are five key drivers that catalysed the growth of Big Data.

Data Explosion

We have more data now than we ever had in past and even more is expected in the future. Various reports are available documenting about how the world's annual collection of data moved from about 1-2 Exabytes in 2000 to 2700 Exabytes in 2012 with a projection of 40,000 Exabytes in 2020. This is not at all surprising as we continue to add data from our operations, partners, third parties, public departments, civic bodies and customers.

Increased Computing Power

We have more computing power and better and economical ways to harness it. Due to the technology evolution compute power has grown substantially, the fastest supercomputer is more than 10.000 times faster today than it was in 2000, and the average desktop power has increased more than 100 times in the same period. Further due to cloud economics a massive amount of compute power is available on demand. Cloud computing enables access to an extreme level of compute power on pay for what you use model, a completely different economic dynamic than in the past. With cloud, companies can scale IT infrastructure without having to allocate cash for up-front capital and then amortize equipment over 5 years. With more individuals and organization adopting the cloud this trend will continue for long and so is the case of more powerful computing hardware.

Technological Innovation to Handle Variety of Data

There are a lot of innovations and developments in software for unstructured data. The majority of new data is unstructured—social media posts, video, audio, pictures, research articles, etc. The new tools, like Hadoop and NoSQL, are giving companies data workhorses that can crunch any amount of data thrown at them. Also, most of these tools are open source and are also available on Software as Service model, having a profound impact.

Improved Intelligence

Analytical algorithms are advancing in several areas, most importantly with machine learning. This has led to lot of activities in futuristic technology space including areas like robotics, artificial intelligence, virtual reality, autonomous vehicles, speech recognition, facial recognition, fraud detection, and medical diagnoses to name a few. Advancement in analytical algorithms has also disrupted various traditional domains like financial markets, retail, and weather forecasting to name a few with better capability to predict what is going to happen.

Internet of Things

The uptake and evolution of the Internet of Things (IoT) has made a substantial impact on everything from consumer electronics to military equipment. IoT has enabled most physical objects like handheld devices, machines, vehicles, buildings and other items to collect and exchange data. The substantial improvement in technology driving IoT includes embedded electronics, software, and network connectivity. IoT is also at the heart of technologies such as smart grids, smart homes, intelligent transportation and smart cities that are future of our society.

In conclusion all these factors are both drivers and are driven by the developments in the Big Data ecosystem, hence we have exciting times ahead if we want to leverage true potential of Big Data.

Application Domain of Big Data

Industry

Process and manufacturing industry

In most industrial plants and systems, the whole production line is equipped with sensors, which feed into monitoring and control systems. A single plant may have on the order of tens of thousands of sensors, often sampled at millisecond rates. Furthermore, enterprises today consist of not a single plant but several, distributed at different locations, often worldwide. The enterprises are in turn connected in supply chains, where each part depends on the others. There is a demand to use analytics of the collected data to get an overview of the production situation in the whole chain; to detect deviations and problems in time; to predict the production outcome; and to plan and dynamically adjust the production in response to both the internal situations and external demands.

Telecom and internet

Obvious enablers for the current rapid development in ICT, including Big Data, mobility, cloud services, and Internet of Things, are the telecommunication industry and the availability of internet everywhere. However, to monitor, maintain and upgrade this huge infrastructure, and to protect it from various threats on all levels, requires Big Data Analytics of the data flows, to find patterns, trends, and unexpected events in it.

Transportation

Modern vehicles contain numerous sensors and electronic control units that generate large amounts of data. This data can be very useful for diagnostics, traffic safety, product development, etc. At the same time, the distributed and mobile nature of those systems makes them challenging to analyse.

However, the benefits that can be obtained, for example by increasing fuel efficiency or by reducing the number of dangerous situations, make it an important area for both research and innovation.

Finance

The role of Big Data Analytics in financial applications has gone through a generational change in which traditional analysts without a specialisation in data analytics have been urged towards modern analytics that involve network- and transactional data. The extreme constraints on financial data flows, where information value drops by the millisecond, means that any Big Data Analytics tools must rely on sampling of real-time flows. The maturity of Big Data Analytics tools and methods is key to keeping algorithmic trading safe and efficient.

Streaming media

There has been a rapidly growing business around streaming media. Streaming audio services has been around for a while, and there are already several competing services for movies. Other examples include YouTube, and that e.g. SVT offers all their programs on the web. A significant part of all web traffic today consists of streaming material.

Characteristic for a provider of streaming media is the huge volumes of data to distribute, the requirement to minimise delays, and typically a very large number of customers. Together this calls for analytics both to monitor the distribution itself to detect possible problems, and to analyse user behaviour and trends, to be able to predict user demand or provide recommendations. Sometimes it is also used as a basis for caching of the distributed material.

Academics

Physic/eScience

There are ever increasing sources of non-structured data from high throughput experiments in biology, large experimental facilities in Physics, energy grids, large climate studies and simulations, etc. New parallel, distributed, heterogeneous high performance computing architectures like clouds, multicores, clusters, FPGAs, as well as a new generation of algorithmic and statistical techniques is being developed to address this.

Life science

Genomics research has high value to both society and industry. It is used by biomedicine researchers, hospital diagnostics, food industries, agronomy, and pharmaceutical industries. A quantum shift is happening in the area of human genomics. A huge wave of big data is approaching, driven by the decreasing cost of sequencing genomic data, which has been halving every 5 months since 2004.

These improvements in both the cost and throughput of DNA sequencing machines have caused a mismatch between the increasing rate at which they can generate genomic data and the ability of our existing tools and computational infrastructure to both store and analyse this data. The scale of the storage requirements for genomic data is huge – a single human genome amounts to coping with the analysis of three billion base pairs. In addition to the storage of genomic data, its analysis will require both massive parallel computing infrastructure and data-intensive computing tools and services to perform analyses in reasonable time.

Society

Health

With the changing age profile of the society, it is becoming more and more important to provide systems that can support people in their lives in an unobtrusive but efficient way. Those systems need to “understand” humans and seamlessly adapt to their habits. An important concept is that of “aging together”, where the focus shifts from comfort and convenience for young and healthy people, towards safety and protection for elderly or sick ones. With the ubiquity of cheap sensors available today this is possible in theory, but new developments in data analysis techniques are needed in order to implement it in practice.

Transports and smart cities

A large number of data is being gathered every hour in today's cities, but there is surprisingly little global analysis that is being done on it. While combining data from multiple sources needs to be done in a careful way to preserve privacy, the benefits of being able to detect abnormal situations or discover surprising relations between events definitely make it worthwhile. This area is a prime example of the need for combining very diverse types of information, and for presenting results in a flexible way.

Urban and physical planning

Data for urban and physical planning is collected and produced by local, regional and national authorities, but is not generally shared and used in an efficient manner. To this data from all available sources can be added and used. An important part of this is to create work processes from the early data collection stage to the visualisation and presentation stage in order to optimise well-grounded political and/or business decisions.

An example is location of preschools. An efficient preschool planning require historical and present data in order to make prognosis of future demographical development.

Socio-economic planning

For socio-economic planning there exist data from Statistics Sweden for small areas called NYKO (similar to census districts), and the planning system itself is organised in each municipality. Today, it has become possible to build data systems on Internet that could be used by municipalities, private companies, media, and citizens. This would give a more Open Data situation. Such systems can incorporate functions like database handling, visualisation, change detection, small area mapping, forecasts, and comparisons with other regions or even countries.

Spatialized big data

Big data analysis of large text volumes in Internet can be combined with geography. For example “which topics are discussed in certain areas?” Here it can be needed to aggregate geo-positions given by addresses to towns or districts, a typical geospatial task. One can also combine with background geo statistics to relate to population properties in different regions/districts, e.g. districts with large immigration or large emigration (national or international).

Integrated data initiatives

Detailed geodata have been available for at least 10 years, however because of high costs most users have been restricted to very few data sources, e.g. registers from one authority or municipality, remote sensing data, or international data in outdated formats. This is now changing rapidly through Open source technologies, Open data initiatives and geodata cooperation.

To some extent, it is the same type of data that was available also 5-10 years ago, but the extended access has dramatically increased the possibility to combine them for new applications.

Big Data Analytics

Fundamental of Analysis

Even though term analysis and analytics often used interchangeably, there exist basic difference between two. Analysis is more related to functions and processes. It is 'a process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making'.

On the other hand Analytics refers to 'the practice of deriving data-driven insights to drive business planning and future performance of an organization through statistics, predictive modelling, data mining and operations analysis, and the communication of these results to appropriate audiences'.

Analytics is a systematic examination and evaluation of data or information, by breaking it into its component parts to uncover their interrelationships.

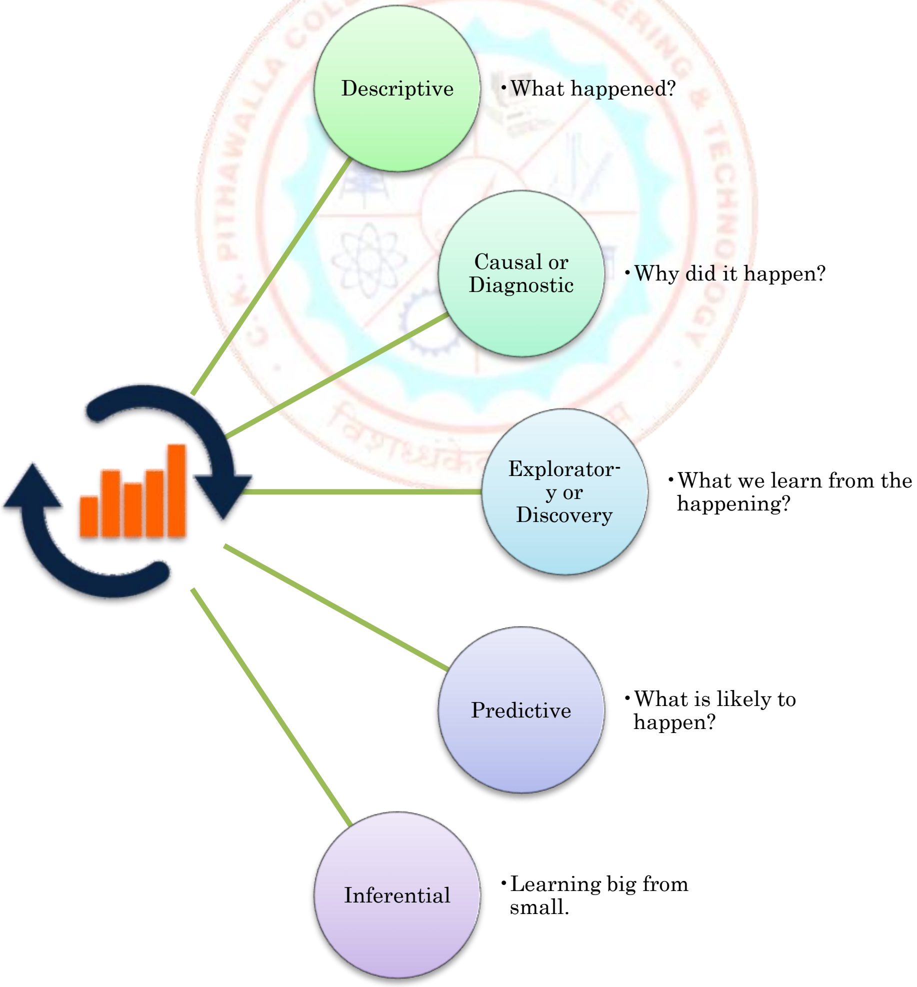
Descriptive
•What happened?
Causal or Diagnostic
•Why did it happen?
Explorator-y or Discovery
•What we learn from the happening?
Predictive
•What is likely to happen?
Inferential
•Learning big from small.

Figure 3. Types of Analysis

Analysis from Big data perspective

Nowadays, the data that need to be analysed are not just large, but they are composed of various data types, and even including streaming data. Since big data has the unique features of "massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous," which may change the statistical and data analysis approaches. Although it seems that big data makes it possible for us to collect more data to find more useful information, the truth is that more data do not necessarily mean more useful information. It may contain more ambiguous or abnormal data. As a result, the whole data analytics has to be re-examined from the following perspectives:

- From the volume perspective, the overflow of input data is the very first thing that we need to face because it may paralyze the data analytics.
- In addition, from the velocity perspective, real-time or streaming data bring up the problem of large quantity of data coming into the data analytics within a short duration but the device and system may not be able to handle these input data.
- From the variety perspective, because the incoming data may use different types or have incomplete data, how to handle them also bring up another issue for the input operators of data analytics.

The overall process of extracting insights from big data can be broken down into five stages as mentioned in Figure 4.

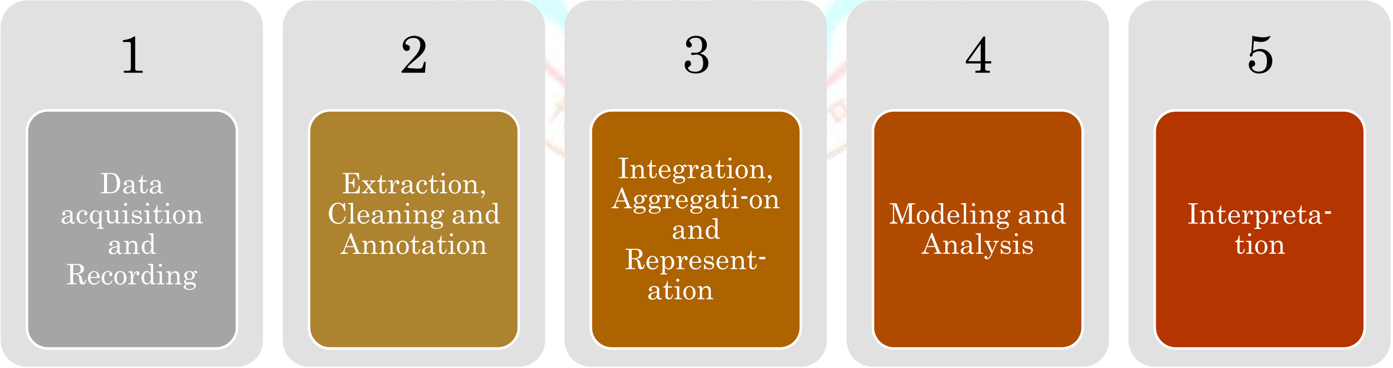
1
Data acquisition and Recording
2
Extraction, Cleaning and Annotation
3
Integration, Aggregati-on and Represent-ation
4
Modeling and Analysis
5
Interpreta-tion

Figure 4. Process of extracting insights from big data

These five stages form the two main sub-processes: data management and analytics.

- Data management involves processes and supporting technologies to acquire and store data and to prepare and retrieve it for analysis, i.e. step 1-3 in Figure 4
- Analytics, on the other hand, refers to techniques used to analyse and acquire intelligence from big data, i.e. step 4-5 in Figure 4.

Thus, big data analytics can be viewed as a sub-process in the overall process of 'insight extraction' from big data.

Types of Data

Data can be classified in major two categories from big data perspective.

Structured data

Data that resides in a fixed field within a record or file is called *structured data*. This includes data contained in relational databases and spreadsheets. Structured data has the advantage of being easily entered, stored, queried and analysed.

Examples – Data stored in database tables or spreadsheet.

Semi-structured data

Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyse. With some process you can store them in relation database (it could be very hard for some one kind of semi structured data), but the semi structure exist to ease space, clarity or compute.

Examples – XML and JSON documents, NoSQL databases

Unstructured data

The phrase *unstructured data* usually refers to information that doesn't reside in a traditional row-column database. Unstructured data files often include text and multimedia content. Digging through unstructured data can be burdensome and costly.

Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents.

Email is a good example of unstructured data. It's indexed by date, time, sender, recipient, and subject, but the body of an email remains unstructured. Other examples of unstructured data include books, documents, medical records, and social media posts.

Significance of unstructured data

Experts estimate that 80 to 90 percent of the data in any organization is unstructured. Also the amount of unstructured data in enterprises is growing significantly - often many times faster than structured databases are growing. So it is but obvious that in order to derive knowledge from text data, most of the time we are dealing with unstructured data.

Types of Analytics

Four major classes of big data analytics are, Text analytics, Audio analytics, Video analytics and Social media analytics

Text Analytics

Text analytics refers to techniques that extract information from textual data. Social network feeds, emails, blogs, online forums, survey responses, corporate documents, news, and call centre logs are examples of textual data held by organizations. Text analytics involve statistical analysis, computational linguistics, and machine learning. Text analytics enable businesses to convert large volumes of human generated text into meaningful summaries, which support evidence-based decision-making. For instance, text analytics can be used to predict stock market based on information extracted from financial news. Various techniques for Text analytics is shown in Figure 5.

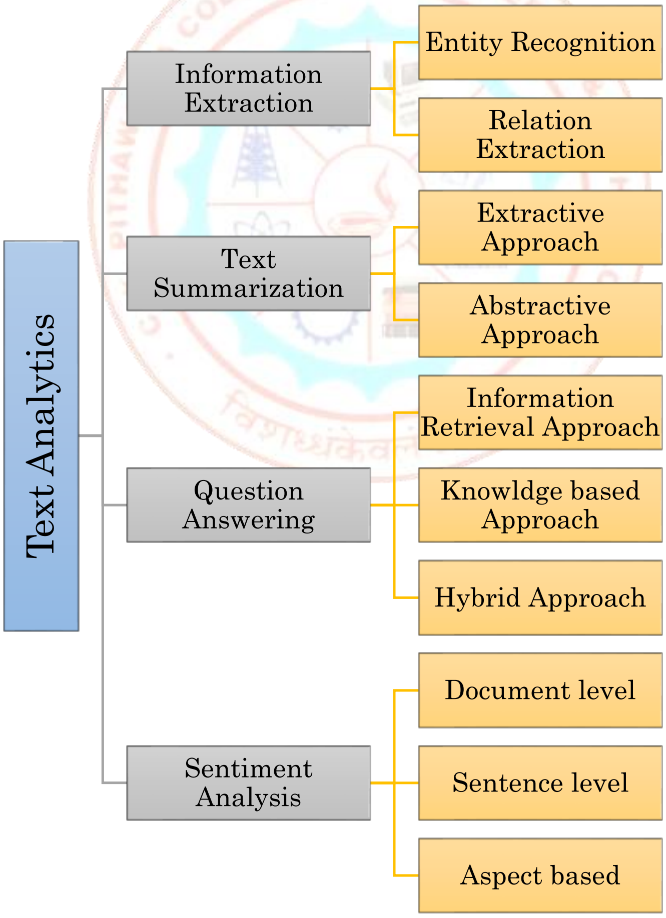

Figure 5. Classification of Text Analytics

1. Information Extraction

- Information extraction extracts structured data from unstructured text. For example, IE algorithms can extract structured information such as drug name, dosage, and frequency from medical prescriptions.
- Two sub-tasks in IE are Entity Recognition (ER) and Relation Extraction (RE).
  - ER finds names in text and classifies them into predefined categories such as person, date, location, and organization.
  - RE finds and extracts semantic relationships between entities (e.g., persons, organizations, drugs, genes, etc.) in the text.
- For example, given the sentence "Steve Jobs co-founded Apple Inc. in 1976", an RE system can extract relations such as FounderOf [Steve Jobs, Apple Inc.] or FoundedIn [Apple Inc., 1976].

2. Text Summarization

- Text summarization automatically produce a concise summary of a single or multiple documents.
- Two approaches;
  - Extractive approach – Resultant summary is a subset of original document. It does not require an understanding of the text.
  - Abstractive approach – On the other hand abstractive approach involves extracting semantic information from the text.
- Extractive systems are easier to adopt, especially for big data.

3. Question Answering

- Question answering techniques provide answers to questions posed in natural language.
- Apple's Siri and IBM's Watson are examples of commercial QA systems.
- QA systems rely on complex natural language processing techniques.
- QA techniques can be further classified into three sub-categories;
  - Information retrieval (IR) based approach - IR based QA systems involves question processing, document processing and answer processing.
  - Knowledge based approach – Knowledge-based QA systems generate a semantic description of the question, which is then used to query structured resources.
  - Hybrid approach – In hybrid QA systems, like IBM's Watson, while the question is semantically analysed, candidate answers are generated using the IR methods.

4. Sentiment analysis

- Also known as opinion mining techniques.
- It analyse opinionated text, which contains people's opinions toward entities such as products, organizations, individuals, and events.

- Sentiment analysis techniques are further divided into three subgroups.
  - Document-level techniques – determine whether the whole document expresses a negative or a positive sentiment.
  - Sentence level techniques – attempt to determine the polarity of a single sentiment about a known entity expressed in a single sentence.
  - Aspect-based techniques – recognize all sentiments within a document and identify the aspects of the entity to which each sentiment refers. For instance, customer product reviews usually contain opinions about different aspects (or features) of a product.

Audio Analytics

- Audio analytics analyse and extract information from unstructured audio data.
- Also referred to as speech analytics, when applied to human spoken language. Since these techniques have mostly been applied to spoken audio, the terms audio analytics and speech analytics are often used interchangeably.
- Currently, customer call centers and healthcare are the primary application areas of audio analytics.
  - Call centers use audio analytics for efficient analysis of thousands or even millions of hours of recorded calls. These techniques help improve customer experience, evaluate agent performance, enhance sales turnover rates, gain insight into customer behaviour, and identify product or service issues, among many other tasks.
  - In healthcare, audio analytics support diagnosis and treatment of certain medical conditions that affect the patient's communication patterns.

As shown in Figure 6 audio analytics follows two common technological approaches: the transcript-based approach (widely known as large-vocabulary continuous speech recognition, LVCSR) and the phonetic-based approach.

1. Transcript based approach

- This methods follows two phase approach indexing and searching.
- Phase-1
  - Attempt to transcribe the speech content of the audio.
  - This is performed using automatic speech recognition (ASR) algorithms that match sounds to words.
  - The words are identified based on a predefined dictionary.
  - If the system fails to find the exact word in the dictionary, it returns the most similar one.
  - The output of the system is a searchable index file that contains information about the sequence of the words spoken in the speech.

- Phase-II
    - Standard text-based methods are used to find the search term in the index file.

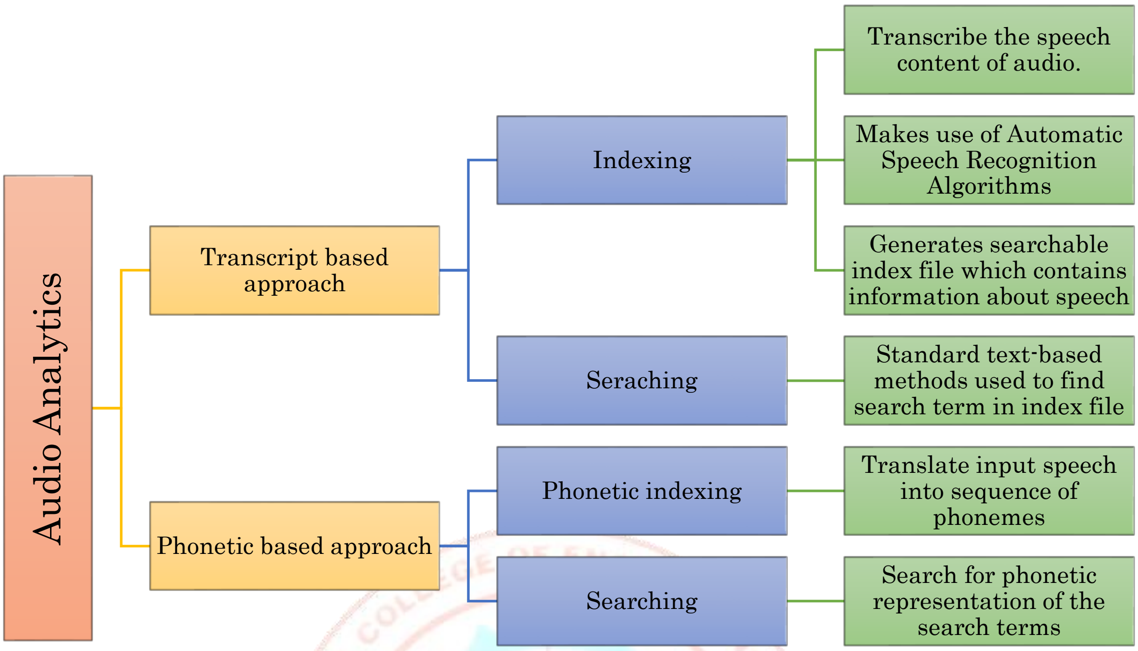
Audio Analytics
Transcript based approach
Phonetic based approach
Indexing
Seraching
Phonetic indexing
Searching
Transcribe the speech content of audio.
Makes use of Automatic Speech Recognition Algorithms
Generates searchable index file which contains information about speech
Standard text-based methods used to find search term in index file
Translate input speech into sequence of phonemes
Search for phonetic representation of the search terms

Figure 6. Summary of Audio Analytics

2. Phonetic based approach

- This approach work with sounds or phonemes.
- Phonemes are the perceptually distinct units of sound in a specified language that distinguish one word from another (e.g., the phonemes ‘s’ and ‘r’ differentiate the meanings of “sun” and “run”).
- Phonetic-based systems also consist of two phases: phonetic indexing and searching.
- Phase-I
    - The system translates the input speech into a sequence of phonemes. This is in contrast to LVCSR systems where the speech is converted into a sequence of words.
- Phase-II
    - The system searches the output of the first phase for the phonetic representation of the search terms.

Video Analytics

- Video analytics, also known as video content analysis (VCA), involves a variety of techniques to monitor, analyse, and extract meaningful information from video streams.
- Various techniques have already been developed for processing real-time as well as pre-recorded videos.

- The increasing prevalence of closed-circuit television (CCTV) cameras and the booming popularity of video-sharing websites are the two leading contributors to the growth of computerized video analysis.
- A key challenge, however, is the sheer size of video data. Because one second of a high-definition video, in terms of size, is equivalent to over 2000 pages of text.
- Big data technologies turn this challenge into opportunity. It can be leveraged to automatically sift through and draw intelligence from thousands of hours of video.
- Primary applications
    - Automated security and surveillance systems.
        - The data generated by CCTV cameras in retail outlets can be extracted for business intelligence.
    - Automatic video indexing and retrieval
        - The indexing of a video can be performed based on different levels of information available in a video including the metadata, the soundtrack, the transcripts, and the visual content of the video.
        - Audio analytics and text analytics techniques can be applied to index a video based on the associated soundtracks and transcripts, respectively.

In terms of the system architecture, there exist two approaches to video analytics, namely server-based and edge-based:

1. Server-based architecture,

- The video captured through each camera is routed back to a centralized and dedicated server that performs the video analytics.
- Due to bandwidth limits, the video generated by the source is usually compressed by reducing the frame rates and/or the image resolution.
- The resulting loss of information can affect the accuracy of the analysis.
- However, the server-based approach provides economies of scale and facilitates easier maintenance.

2. Edge-based architecture

- Analytics are applied at the 'edge' of the system.
- That is, the video analytics is performed locally and on the raw data captured by the camera. As a result, the entire content of the video stream is available for the analysis, enabling a more effective content analysis.
- Edge-based systems, however, are more costly to maintain and have a lower processing power compared to the server-based systems.

Social Media Analytics

- Social media analytics refer to the analysis of structured and unstructured data from social media channels.
- Social media can be categorized into the following types:
  - Social networks (e.g., Facebook and LinkedIn),
  - Blogs (e.g., Blogger and WordPress),
  - Microblogs (e.g., Twitter and Tumblr),
  - Social news (e.g., Digg and Reddit),
  - Social bookmarking (e.g., Delicious and StumbleUpon),
  - Media sharing (e.g., Instagram and YouTube),
  - Wikis (e.g., Wikipedia and Wikihow),
  - Question-and-answer sites (e.g., Yahoo! Answers and Ask.com)
  - Review sites (e.g., Yelp, TripAdvisor)
- The application of social media analytics spans across several disciplines, including psychology, sociology, anthropology, computer science, mathematics, physics, and economics.
  - Marketing has been the primary application of social media analytics in recent years.
- Sources of information
  1. User-generated content (e.g., sentiments, images, videos, and bookmarks)
  2. The relationships and interactions between the network entities (e.g., people, organizations, and products) are the two sources of information in social media.

Based on this categorization, the social media analytics can be classified into two groups as shown in Figure 7.

1. Content based analytics

- Focuses on the data posted by users on social media platforms, such as customer feedback, product reviews, images, and videos.
- Such content on social media is often voluminous, unstructured, noisy, and dynamic.
- Various text, audio, and video analytics can be applied to derive insight from such data.
- Big data technologies can be adopted to address the data processing challenges.

2. Structure based analytics

- Also referred to as social network analytics
- This type of analytics are concerned with synthesizing the structural attributes of a social network and extracting intelligence from the relationships among the participating entities.
- The structure of a social network is modelled through a set of nodes and edges, representing participants and relationships, respectively.

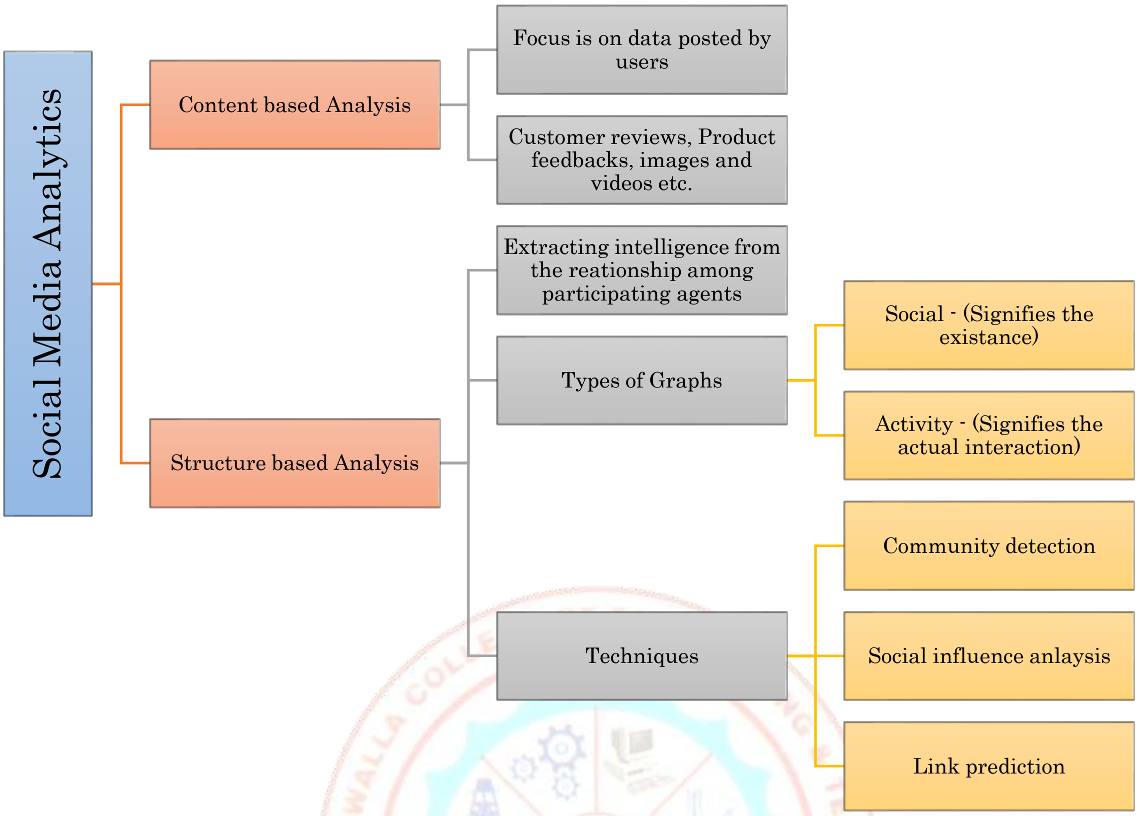
Social Media Analytics
Content based Analysis
Focus is on data posted by users
Customer reviews, Product feedbacks, images and videos etc.
Structure based Analysis
Extracting intelligence from the reationship among participating agents
Types of Graphs
Social - (Signifies the existance)
Activity - (Signifies the actual interaction)
Techniques
Community detection
Social influence anlaysis
Link prediction

Figure 7. Classification of Social Media Analytics

- The model can be visualized as a graph composed of the nodes and the edges.
- Types of graphs :
    - Social graphs
        - An edge between a pair of nodes only signifies the existence of a link (e.g., friendship) between the corresponding entities.
        - Used for identifying communities or determine hubs.
    - Activity graphs.
        - The edges represent actual inter-actions between any pair of nodes.
        - The interactions involve exchanges of information (e.g., likes and comments).
        - Preferable to social graphs, because an active relationship is more relevant to analysis than a mere connection.
- Applications
    - Community detection
    - Social influence analysis
    - Link prediction

Distributed File Systems

Considering various features of big data such as volume etc., big data applications often required distributed file system architecture. Distributed file system has support for remote information sharing, user mobility, availability etc. These properties are essential for big data processing.

Services provided by DFS

- Storage services –
  - Allocation and management of space on a secondary storage device thus providing a logical view of the storage system.
- True file services –
  - Includes file-sharing semantics, file-caching mechanism, file replication mechanism, concurrency control, multiple copy update protocol etc.
- Name/Directory services –
  - Responsible for directory related activities such as creation and deletion of directories, adding a new file to a directory, deleting a file from a directory, changing the name of a file, moving a file from one directory to another etc.

Desirable features of DFS

1. Transparency
   - Structure transparency – Clients should not know the number or locations of file servers and the storage devices. Note: multiple file servers provided for performance, scalability, and reliability.
   - Access transparency – Both local and remote files should be accessible in the same way. The file system should automatically locate an accessed file and transport it to the client's site.
   - Naming transparency – The name of the file should give no hint as to the location of the file. The name of the file must not be changed when moving from one node to another.
   - Replication transparency – If a file is replicated on multiple nodes, both the existence of multiple copies and their locations should be hidden from the clients.
2. User mobility
   - Automatically bring the user's environment (e.g. user's home directory) to the node where the user logs in.
3. Performance
   - Performance is measured as the average amount of time needed to satisfy client requests.
   - This time includes CPU time + time for accessing secondary storage + network access time.
   - It is desirable that the performance of a distributed file system be comparable to that of a centralized file system.

4. Simplicity and ease of use
   - User interface to the file system be simple and number of commands should be as small as possible.
5. Scalability
   - Growth of nodes and users should not seriously disrupt service.
6. High availability
   - A distributed file system should continue to function in the face of partial failures such as a link failure, a node failure, or a storage device crash.
   - A highly reliable and scalable distributed file system should have multiple and independent file servers controlling multiple and independent storage devices.
7. High reliability
   - Probability of loss of stored data should be minimized. System should automatically generate backup copies of critical files.
8. Data integrity
   - Concurrent access requests from multiple users who are competing to access the file must be properly synchronized by the use of some form of concurrency control mechanism. Atomic transactions can also be provided.
9. Security
   - Users should be confident of the privacy of their data.
10. Heterogeneity
    - There should be easy access to shared data on diverse platforms (e.g. UNIX workstation, Wintel platform etc.).

Big Data – File Systems

- File systems are the foundation of the applications at upper levels.
- Google's GFS
  - An expandable distributed file system to support large-scale, distributed, data-intensive applications.
  - GFS uses cheap commodity servers to achieve fault-tolerance and provides customers with high performance services.
  - GFS supports large-scale file applications with more frequent reading than writing.
  - Limitations
    - Single point of failure
    - Poor performances for small files.
  - Such limitations have been overcome by Colossus, the successor of GFS.
- Other solutions

  - Companies and researchers also have their solutions to meet the different demands for storage of big data.
    - HDFS - Derivatives of open source codes of GFS.
    - Microsoft developed Cosmos to support its search and advertisement business.
    - Facebook utilizes Haystack to store the large amount of small-sized photos.
- Distributed file systems have been relatively mature after years of development and business operations.

Programming Models

- Big data are generally stored in hundreds and even thousands of commercial servers. Thus, the traditional parallel models, such as Message Passing Interface (MPI) and Open Multi-Processing (OpenMP), may not be adequate to support such large-scale parallel programs.
- Recent developments in parallel programming models effectively improve the performance of NoSQL and reduce the performance gap to relational databases.
- Therefore, these models have become the foundation for the analysis of massive data. Examples of this programming model includes MapReduce, Drayad, All-pairs and Pregel etc.

MapReduce

MapReduce is a simple but powerful programming model for large-scale computing using a large number of clusters of commercial PCs to achieve automatic parallel processing and distribution.

In MapReduce, computing model only has two functions, i.e., Map and Reduce, both of which are programmed by users.

- The Map function processes input key-value pairs and generates intermediate key-value pairs.
- MapReduce will combine all the intermediate values related to the same key and transmit them to the Reduce function.
- Reduce function merges together these values to form a possibly smaller set of values. Typically just zero or one output value is produced per Reduce invocation.
- The intermediate values are supplied to the user's reduce function via an iterator. This allows to handle lists of values that are too large to fit in memory.

MapReduce has the advantage that it avoids the complicated steps for developing parallel applications, e.g., data scheduling, fault-tolerance, and inter-node communications. The user only needs to program the two functions to develop a parallel application.

Execution overview:

- The Map invocations are distributed across multiple machines by automatically partitioning the input data into a set of M splits.
- The input splits can be processed in parallel by different machines.
- Reduce invocations are distributed by partitioning the intermediate key space into R pieces using a partitioning function (e.g. $hash(key)\ mod\ R$).
- The number of partitions (R) and the partitioning function are specified by the user.

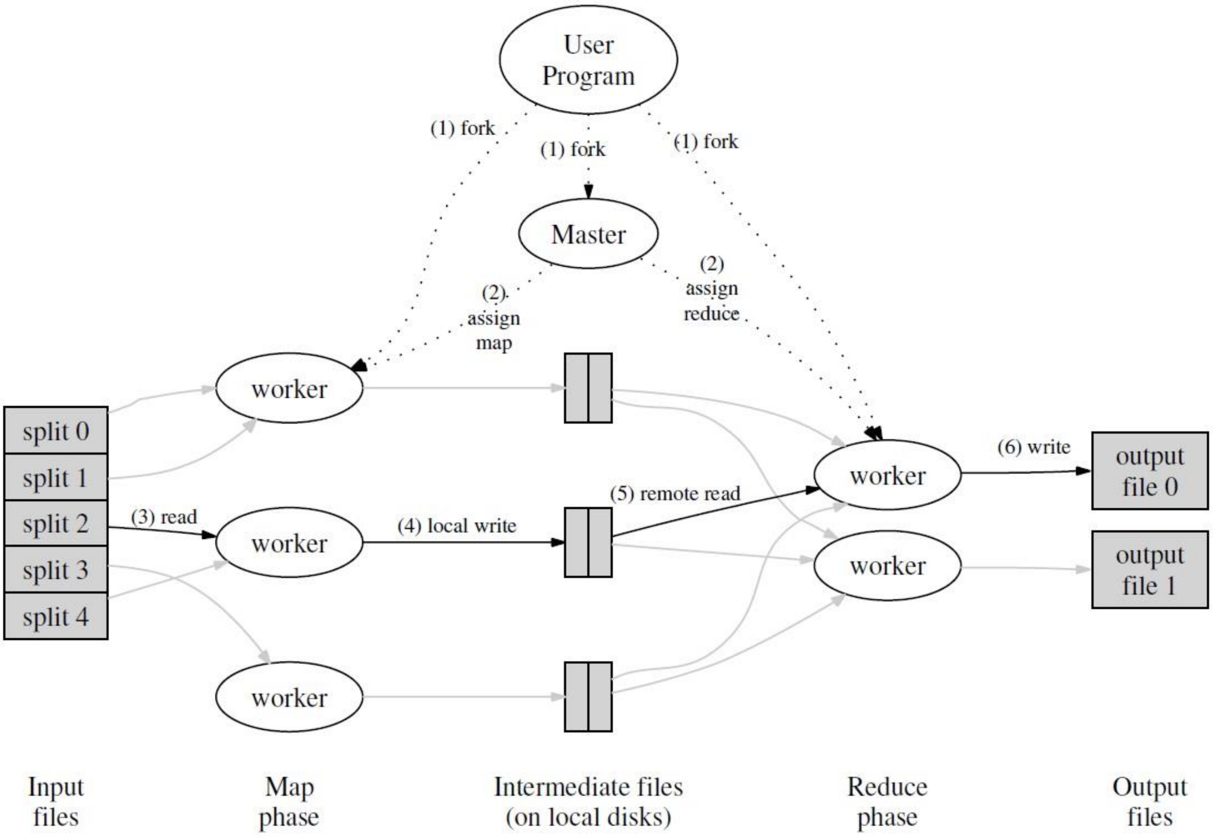

Figure 8. MapReduce Programming Model

Overall flow of a MapReduce operation is shown in Figure 8. When the user program calls the MapReduce function, the following sequence of actions occurs:

1) Input files is been split in to *M* pieces of typically 16 MB to 64 MB per piece. It then starts up many copies of the program on a cluster of machines.
2) The master - a special copy of a program, will assign work to the rest i.e. workers. The master picks idle workers and assigns each one a map task or a reduce task. Total there are *M* map task and *R* reduce task to be assigned.
3) A worker who is assigned a map task read the content of input split, parses the key/value pair and passes each pair to user-defined *Map* function. Generated intermediate key/value pair are buffered in memory.
4) On timely bases buffered pairs will be written to local disk and partitioned into *R* regions by the partitioning function. The location of these pairs on local disk are passed back to the master, who is responsible for forwarding these locations to the reduce workers.
5) Upon receiving notification from master, reduce worker reads buffered data from the local disk of the map workers using remote procedure calls. It then sorts intermediate key in order to group them together. In case of large intermediate data external sort is used.

6) The reduce worker iterates over sorted keys and passes the key and corresponding set of intermediate values to the user-defined *Reduce* function. The output of *Reduce* function is appended to a final output file for this reduce partition.
7) Once all map and reduce tasks have been accomplished, the master wakes up user program and MapReduce call in the user program returns back to the user code.

After successful completion, the output of the MapReduce execution is available in the R output files. (One per reduce task, with file names as specified by the user). Typically, users do not need to combine these R output files into one file. They often pass these files as an input to another MapReduce call, or use them from another application.

Example – Word Count

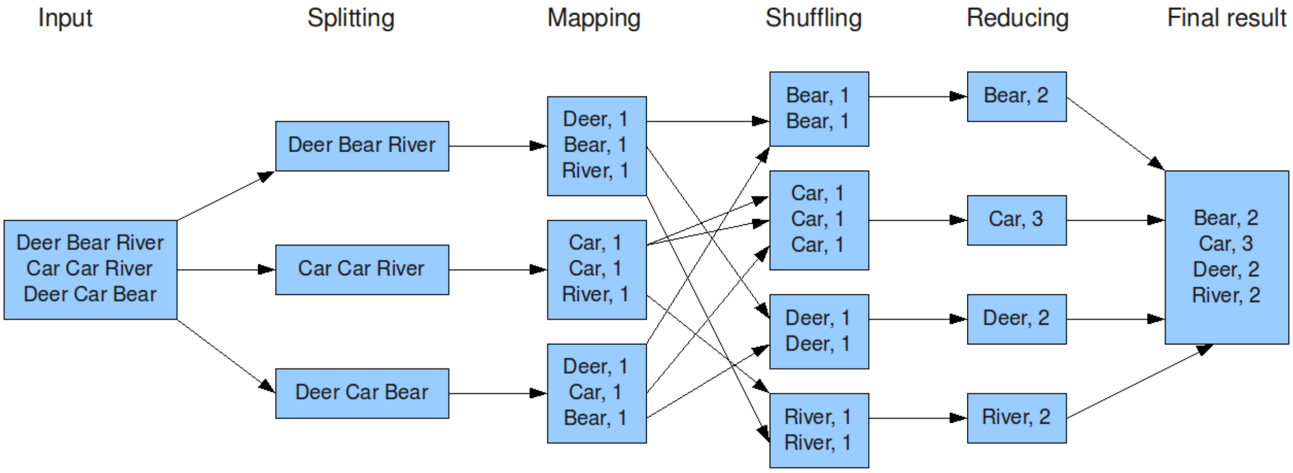

Figure 9. Example of MapReduce

Consider the problem of counting the number of occurrences of each word in a large collection of documents. Detailed computation of word count using MapReduce model is shown in Figure 9.

Over the past decades, programmers are familiar with the advanced declarative language of SQL, often used in a relational database, for task description and dataset analysis. However, the concise MapReduce framework only provides two non-transparent functions, which cannot cover all the common operations. Therefore, programmers have to spend time on programming the basic functions, which are typically hard to be maintained and reused. In order to improve the programming efficiency, some advanced language systems have been proposed, e.g., Sawzall of Google, Pig Latin of Yahoo, Hive of Facebook, and Scope of Microsoft.

Processing Mechanisms

Processing methods are techniques used to process different types of data. There are number of processing methods depending upon the hardware/software capabilities and the type of data need to be processed by the organization. Commonly used processing methods are batch processing, stream (online) processing, and real time processing.

Batch Processing

- Batch processing is very efficient in processing high volume data.
- Data is collected, entered to the system, processed and then results are produced in batches. Here time taken for the processing is not an issue.
- Batch jobs are configured to run without manual intervention, trained against entire dataset at scale in order to produce output in the form of computational analyses and data files.
- Depending on the size of the data being processed and the computational power of the system, output can be delayed significantly.
- Batch processing requires separate programs for input, process and output. As there are 3Vs of Big Data (Volume, Velocity, Variety). If only concern is Volume of the data, then batch processing is a way to go.
- MapReduce is an example of the batch processing system of Hadoop ecosystem.

Stream Processing

- In contrast, stream processing involves a continual input, process and output of data.
- Data must be processed in a small time period (or near real-time).
- Radar systems, customer services and bank ATMs are examples.
- Streaming analytics has the potential to accelerate "time to insight" from the massive amounts of data originating from market data, sensors, mobile phones, the Internet of Things, Web clickstreams, and transactions.
- Apache Strom is a stream processing framework that also does micro-batching.
- Sometimes streaming used as a sort of synonym for real-time. Real-time stuff usually takes the form of needing to respond to an event in milliseconds, as in a synchronous API which is not streaming according to

While most organizations use batch data processing, sometimes an organization needs real time data processing. Real time data processing and analytics allows an organization the ability to take immediate action for those times when acting within seconds or minutes is significant. The goal is to obtain the insight required to act prudently at the right time which increasingly means immediately. Differences of batch and stream processing summarised in Table 2.

Table 2. Batch Processing vs. Stream Processing

Batch Processing	**Stream Processing**
• Has access to all data • Might compute something big and complex • Is generally more concerned with throughput than latency of individual components of the computation • Has latency measured in minutes or more	• Computes a function of one data element, or a smallish window of recent data • Computes something relatively simple • Needs to complete each computation in near-real-time — probably seconds at most • Computations are generally independent • Asynchronous – source of data doesn't interact with the stream processing directly, like by waiting for an answer