UNIT VI

6 Big Data Technologies Application and Impact

Syllabus

Social media analytics, *Text mining, Mobile analytics, Data analytics life cycle of case studies, Organizational impact, understanding decision theory, creating big data strategy, big data value creation drivers, Michael Porter's valuation creation models, Big data user experience ramifications, Identifying big data use cases, Big Data Analytics Challenges and Research directions.*

Contents

6.1 Introduction to Social Network Analysis

SPPU : May-18, Dec.-18, 19

- Social Network Analysis [SNA] is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The term "social network" has been introduced by Barnes in 1954.
- SNA is the study of social relations among a set of actors. The methods of data collection in network analysis are aimed at collecting relational data in a reliable manner. Data collection is typically carried out using standard questionnaires and observation techniques that aim to ensure the correctness and completeness of network data.
- Social network analysis is based on an assumption of the importance of relationships among interacting units. The social network perspective encompasses theories, models, and applications that are expressed in terms of relational concepts or processes.
- The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships.
- **Features of social network analysis :** Structural intuition, systematic relational data, graphic representation and mathematical or computational models.
- A social network is a group of collaborating, and/or competing individuals or entities that are related to each other. It may be presented as a graph, or a multi-graph; each participant in the collaboration or competition is called an actor and depicted as a node in the graph theory.
- Valued relations between actors are depicted as links, or ties, either directed or undirected, between the corresponding nodes.
- Actors can be persons, organizations, or groups - any set of related entities. As such, SNA may be used on different levels, ranging from individuals, web pages, families, small groups, to large organizations, parties and even to nations.
- In general, a social network consists of actors (e.g., persons, organizations) and some form of relation among them. The network structure is usually modeled as a graph, in which vertices represent actors and edges represent ties, i.e., the existence of a relation between two actors.
- The vocabulary, models and methods of network analysis also expand continuously through applications that require to handle ever more complex data sets.

- The purpose of social network analysis is to identify important actors, crucial links, roles, dense groups, and so on, in order to answer substantive questions about structure.
- Analysis methods available in vison are divided into four main categories according to the level or subject of interest : vertex, dyad, group, and network level.
- Available analysis methods include actor-level centrality indices, e.g. closeness, betweenness, and page rank, cohesive subgroups like cliques, k-cliques, and k-clans, centrality and connectedness
- These levels break further down into measures of the same objective, e.g., connectedness or cohesiveness. Analysis methods are accessible using the analysis tab in the control area

Key concepts and measures in network analysis

- Social network analysis has developed a set of concepts and methods specific to the analysis of social networks.
- Several analytic tendencies distinguish social network analysis :
  1. There is no assumption that groups are the building blocks of society: the approach is open to studying less-bounded social systems, from nonlocal communities to links among websites.
  2. Rather than treating individuals (persons, organizations, states) as discrete units of analysis, it focuses on how the structure of ties affects individuals and their relationships.
  3. In contrast to analyses that assume that socialization into norms determines behavior, network analysis looks to see the extent to which the structure and composition of ties affect norms.

Application of Social Network Analysis

- SNA is an important and valuable tool for knowledge extraction from massive and un-structured data. Social network provides a powerful abstraction of the structure and dynamics of diverse kinds of inter-personal connection and interaction.
- Facebook is a social networking service and website that connects people with other people, and share data between people. A user can create a personal profile, add other users as friends, exchange data, create and join common interest communities.
- Twitter is a social net-working and microblogging service. The users of Twitter can exchange text-based posts called tweets. A tweet is a maximum 140 characters long but can be augmented by pictures or audio recording. The main concept of

Twitter was to build a social network formed by friends and followers. Friends are people who you follow, followers are those who follow you.

- The role of social networks in labor markets deserves attention for at least two reasons : first, because of the central role networks play in disseminating information about job openings they place a critical role in determining whether labor markets function efficiently; and second, because network structure ends up having implications for things like human capital investment as well as inequality.
- Social Network Analysis (SNA) primarily focuses on applying analytic techniques to the relationships between individuals and groups, and investigating how those relationships can be used to infer additional information about the individuals and groups.
- SNA is used in a variety of domains. For example, business consultants use SNA to identify the effective relationships between workers that enable work to get done; these relationships often differ from connections seen in an organizational chart.

6.1.1 Social Media Analytics

- Social media analytics deals with development and evaluation of tools and frameworks to collect, monitor, analyze, summarize and visualize social media data.
- Buyer's perspective about various brands and businesses can be statistically analyzed to extract various insights required for decision making from a very large amount of unstructured and semi structured social media data.
- In social media, the two sources of information are the content (images, audios, customer feedbacks, product reviews, videos, bookmarks, sentiments, etc.) generated by users and the relationships between the entities of network (people, organizations, products, etc.).
- The social media analytics can be categorized into two parts : Content-based analytics and Structure-based analytics.
- In content-based analytics, analytics is performed on the content posted by the users on the social media platforms. Such content is of high volume, unstructured, noise and dynamic nature.
- To extract insights from such data, the text, audio and video analytics techniques can be applied. The data processing challenges are addressed by the big data technologies.
- In structure-based analytics, the focus is on the structural attributes of the social network. Insights are extracted from the relationships of the entities.

Process of social media data analytics :

- Fig. 6.1.1 shows process of social media data analytics.

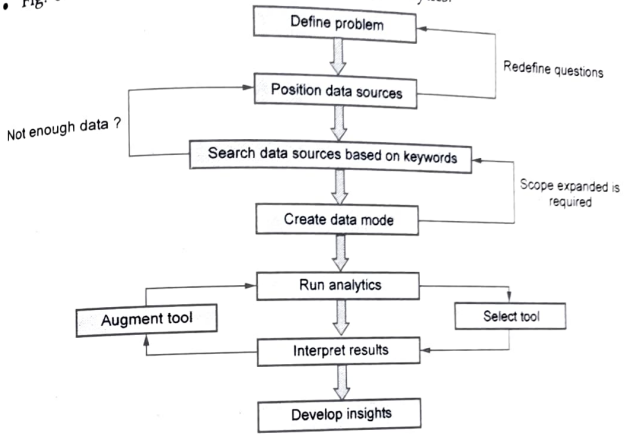


Fig. 6.1.1

- Social media analytics as a part of social analytics is the process of gathering data from stakeholder conversations on digital media and processing into structured insights leading to more information-driven business decisions and increased customer centrality for brands and businesses.
- Data analysis is the set of activities that assist in transforming raw data into insight, which in turn leads to a new base of knowledge and business value.
- In other words, data analysis is the phase that takes filtered data as input and transforms that into information of value to the analysis.
- Many different types of analysis can be performed with social media data. The data analysis step begins once we know that problem we want to solve and know that we have sufficient data that is enough to generate a meaningful result.

Applications :

- **Retail companies :** To harness their brand awareness, service improvement, advertising / marketing strategies, identifying influencers.
- **Finance :** To determine market sentiment, news data for trading.
- **Government and public officials :** Monitoring public perception on political candidates, election campaigns and announcements. Prediction at national level of happiness, unemployment, etc.

- **Public health and sociology** : Given that two people have been in approximately the same geographic locate, at same time, on multiple occasions, how likely they know each other ?

Review Questions

1. *What is social media analytics ? Explain it's need with sample case study.* **SPPU : May-18 (End Sem), Marks 9**
2. *Explain the process of social media data analytics with example.* **SPPU : Dec.-18 (End Sem), Marks 8**
3. *How social media analytics helps in value creation ? Explain with suitable example.* **SPPU : Dec.-19 (End Sem), Marks 9**

6.2 Text Mining

SPPU : May-18, Dec.-18, 19

- Text Analysis (TA) aims to extract machine-readable information from unstructured text in order to enable data-driven approaches towards managing content. The purpose of text analysis is to create structured data out of free text content.
- Text analysis, also known as text mining, is the process of automatically classifying and extracting meaningful information from unstructured text.
- Text mining is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from textual document repositories.
- Text mining can be visualized as consisting of two phases : Text refining that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form.
- Fig 6.2.1 shows high level text mining functional architecture.

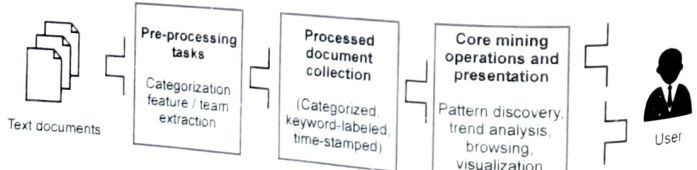


Fig. 6.2.1 High level text mining functional architecture

- A process of text mining involves a series of activities to be performed to mine the information. These activities are :

1. **Text Pre-processing :** It involves a series of steps as shown in below :
   i. **Text Clean-up :** Text Clean-up means removing any unnecessary or unwanted information. Such as remove ads from web pages, normalize text converted from binary formats.
   ii. **Tokenization :** Tokenizing is simply achieved by splitting the text into white spaces.
   iii. **Part of speech tagging :** Part-of-Speech (POS) tagging means word class assignment to each token. Its input is given by the tokenized text. Taggers have to cope with unknown words (OOV problem) and ambiguous word-tag mappings.
2. **Text transformation (Attribute generation) :** A text document is represented by the words it contains and their occurrences. Two main approaches to document representation are : Bag of words and Vector Space
3. **Feature selection (Attribute selection) :** Feature selection also is known as variable selection. It is the process of selecting a subset of important features for use in model creation. Redundant features are the one which provides no extra information. Irrelevant features provide no useful or relevant information in any context.
4. **Data mining :** At this point, the text mining process merges with the traditional process. Classic data mining techniques are used in the structured database. Also, it resulted from the previous stages.
5. **Evaluate :** Evaluate the result, after evaluation, the result can be discarded.

- The five fundamental steps involved in text mining are :

1. Gathering unstructured data from multiple data sources like plain text, web pages, pdf files, emails, and blogs, to name a few.
2. Detect and remove anomalies from data by conducting pre-processing and cleansing operations. Data cleansing allows user to extract and retain the valuable information hidden within the data and to help identify the roots of specific words.
3. Convert all the relevant information extracted from unstructured data into structured formats.
4. Analyse the patterns within the data via Management Information System (MIS).
5. Store all the valuable information into a secure database to drive trend analysis and enhance the decision-making process of the organization.

6.2.1 Use of a Text Mining Tool

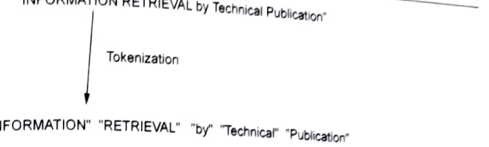
1. **Text analytics :** Involves extracting useful information and patterns from text. Most tools provide this feature.
2. **Text processing :** Involves transforming and manipulating unstructured text so that analysis methods can be applied to it.
3. **Classification/Categorization :** Many tools are used for classification and categorization of text/documents.
4. **Sentiment Analysis :** Is used to identify subjective information from text. Many tools provide for sentiment analysis also called as opinion mining.
5. **Knowledge discovery :** Deals with identification of useful information from huge amount of text. Most tools provide for knowledge discovery and information retrieval features

6.2.2 Text Pre-processing

- Text pre-processing is required to transform the text into an understandable format so that machine learning algorithms can be applied to it.
- As we know Machine Learning needs data in the numeric form. We basically used encoding technique to encode text into numeric vector. But before encoding we first need to clean the text data and this process to prepare or clean text data before encoding is called text pre-processing.
- The various text pre-processing steps are : Tokenization, Lower casing, Stop words removal, Stemming and Lemmatization.

6.2.2.1 Tokenization

- Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. The simplest form of analysis is to reduce different word forms into tokens.
- Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords.
- Tokenization can be broadly classified into 3 types : Word, character, and subword (n-gram characters) tokenization.
- These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.



- Tokenization can be done to either separate words or sentences. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.
- Tokens are the building blocks of Natural Language, the most common way of processing the raw text happens at the token level.
- The major question of the tokenization phase is what are the correct tokens to use ? You chop on whitespace and throw away punctuation characters.
- Types of tokenization are white space, dictionary based, rule based, penn tree, spacy, subword etc.

6.2.2.2 Stemming

- Stemming is the process for reducing inflected words to their stem, base or root form, generally a written word form. Stemming allows a query term such as "orienteering" to match an occurrence of "orienteers", or "runs" to match "running".
- For example, "orienteering" and "orienteers" might reduce to the root form "orienteer"; "runs" and "running" might reduce to "run".
- In an IR system a stemmer may be applied at both indexing time and query time. During indexing each token is passed through the stemmer and the resulting root form is indexed.
- At query time, the query terms are passed through the same stemmer and matched against the index terms. Thus the query term "runs" would match an occurrence of "running" by way of their common root form "run".
- Stemming is the particular case of tokenization which reduces inflected forms to a single base form or stem. Stemming algorithms are basic string - handling algorithms, which depend on rules which identify affixes that can be stripped
- The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance :

  am, are, is $\Rightarrow$ be

  car, cars, car's, cars' $\Rightarrow$ car
- The result of this mapping of text will be something like : The boy's cars are different colors the boy car be differ color.

6.2.2.3 Stop Words

- The removal of high frequency words, 'stop' words or 'fluff' words is one way of implementing Luhn's upper cut-off. This is normally done by comparing the input text with a 'stop list' of words which are to be removed.
- Sometimes, some extremely common words that would appear to be of little value in helping select documents matching a user need are excluded from stop words the vocabulary entirely. These words are called stop words.
- Function words are words that have no well - defined meanings in and of themselves. Function words are usually among the most frequently occurring words in any language. At query time these stopwords are stripped from the query, and retrieval takes place on the basis of the remaining terms alone.
- Examples of a few stop words in English are "the", "a", "an", "so", "what".
- Removal of stop words definitely reduces the dataset size and thus reduces the training time due to the fewer number of tokens involved in the training.
- The phrase query "Capital of the India," which contains two stop words, is more precise than Capital AND "India".
- Stop word list is as follows :

  a an and are as at be by for from has he in is it its of on that the to was were will with
- With stopwords present in the index, the IR system can make a decision on a query - by - query basis. Ranking methods in a modern commercial search engine will incorporate many ranking features, including features based on term frequency and proximity.
- For features that do not consider proximity between query terms, stopwords may be elminated. For features that do consider proximity between query terms, particularly to match their occurrence in phrases, it may be appropriate to retain stopwords.
- Good compression techniques means the space for including stopwords in a system is very small.
- Good query optimization techniques mean you pay little at query time for including stop words.

6.2.2.4 Lemmatization

- Many complex or technical concepts and many organization and product names are multiword compounds or phrases. An information retrieval system uses phrases to index, retrieve, organize and describe documents.

- Phrases are identified that predict the presence of other phrases in documents. Documents are the indexed according to their included phrases.
- Most recent search engines support a double quotes syntax ("technical publications") for phrase queries, which has proven to be very easily understood and successfully used by users.
- One approach to handling phrases is to consider every pair of consecutive terms in a document as a phrase. For example the text ***Friends, Romans, Countrymen*** would generate the biwords :

  friends romans

  romans countrymen
- Each of these biwords is now a dictionary term. Two - word phrase query - processing is now immediate.
- Biword indexes are not the standard solution but can be part of a compound strategy.

6.2.3 Bag of Words

- Bag of words model helps convert the text into numerical representation such that the same can be used to train models using machine learning algorithms.
- The bag-of-words model is commonly used in methods of document classification where the occurrence of each word is used as a feature for training a classifier.
- Here are the key steps of fitting a bag-of-words model :
  1. Create a vocabulary index of words or tokens from the entire set of documents. The vocabulary indices can be created in alphabetical order.
  2. Construct the numerical feature vector for each document that represents how frequent each word appears in different documents. The feature vector representing each will be sparse in nature as the words in each document will represent only a small subset of words out of all words present in entire set of documents.
- Fig. 6.2.2 shows turning raw text into a bag of words representation.
- Bag of words simply refers to a matrix in which the rows are documents and the columns are words. The values matching a document with a word in the matrix, could be a count of word occurrences within the document or use tf-idf.
- Classifiers are used to train the bag of words and a special kind of algorithm used to break words down into categories.

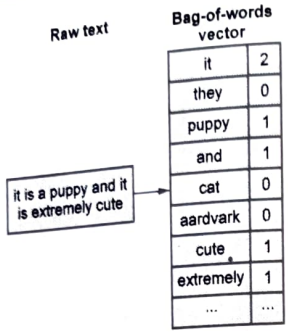


Fig 6.2.2 Turning raw text into a bag of words representation

- Traditionally, text documents are represented in NLP as a bag-of-words. This means that each document is represented as a fixed-length vector with length equal to the vocabulary size.
- Each dimension of this vector corresponds to the count or occurrence of a word in a document. Being able to reduce variable-length documents to fixed-length vectors makes them more amenable for use with a large variety of Machine Learning (ML) models and tasks.

6.2.4 TF-IDF

- Term Frequency (TF) : Frequency of occurrence of query keyword in document
- Inverse Document Frequency (IDF) : How many documents the query keyword occurs in.
- Inverse Document Frequency (IDF) is a popular measure of a word's importance. It's defined as the logarithm of the ratio of number of documents in a collection to the number of documents containing the given word. This means rare words have high IDF and common words have low IDF.
- Term frequency is a measure of the importance of terms *i* in document *j*.
- Inverse document frequency is a measure of the *general* importance of the term.
- High term frequency for "apple" means that apple is an important word in a specific document. But high document frequency (low inverse document frequency) for "apple", given a particular set of documents, means that apple is not all that important overall, since it is in all of the documents.

- The weight increases as the number of documents in which the term appears decreases. High value indicates that the word occurs more often in this document than average.
- The term frequency $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d.
- A document with tf = 10 occurrences of the term is more relevant than a document with tf = 1 occurrence of the term. But not 10 times more relevant. Relevance does not increase proportionally with term frequency.
- The document frequency is the number of documents in the collection that the term occurs in. We define the idf weight of term t as follows :

$$\text{idf weight } (idf_t) = \log 10 \frac{N}{df_t}$$

here N is the number of documents in the collection

- The tf-idf weight of a term is the product of its tf weight and its idf weight

$$W_{t,d} = (1 + \log tf_{t,d}) \cdot \log \frac{N}{df_t}$$

Stop lists and stemming :

- **Stoplists :** This is a list of words that we should ignore when processing documents, since they give no useful information about content.
- **Stemming :** This is the process of treating a set of words like "fights, fighting, fighter, ..." as all instances of the same term - in this case the stem is "fight".

Review Questions

1. *What is text mining ? Draw and explain text mining architecture and explain its need.* **SPPU : May-18 (End Sem), Marks 9**
2. *Draw and explain details architecture of text mining and explain why it is required ?* **SPPU : Dec.-18 (End Sem), Marks 8**
3. *What is text mining ? Draw and explain text mining architecture and its use.* **SPPU : Dec.-19 (End Sem), Marks 10**

6.3 Mobile Analytics

SPPU : May-18, Dec.-18

- Analytics is the practice of measuring and analyzing data of users in order to create an understanding of user behavior as well as website or application's performance. If this practice is done on mobile apps and app users, it is called "mobile analytics".

- Mobile analytics is the practice of collecting user behavior data, determining intent from those metrics and taking action to drive retention, engagement and conversion.
- Mobile analytics is similar to web analytics where identification of the unique customer and recording their usages.
- With mobile analytics data, you can improve your cross-channel marketing initiatives, optimize the mobile experience for you customers and grow mobile user engagement and retention.
- Analytics usually comes in the form of a software that integrates into companie's existing websites and apps to capture, store and analyze the data.
- It is always very important for businesses to measure their critical KPIs (Key Performance Indicators), as the old rule is always valid : "If you can't measure it, you can't improve it".
- To be more specific, if a business find out 75 % of their users exit in the shipment screen of their sales funnel, probably there is something wrong with that screen in terms of its design, user interface (UI) or user experience (UX) or there is a technical problem preventing users from completing the process.

Working of Mobile Analytics :

- Most of the analytics tools need a library (an SDK) to be embedded into the mobile app's project code and at minimum an initialization code in order to track the users and screens.
- SDKs differ by platform so a different SDK is required for each platform such as iOS, Android, Windows Phone etc. On top of that, additional code is required for custom event traking.
- With the help of this code, analytics tools track and count each user, app launch, tap, event, app crash or any additional information that the user has, such as device, operating system, version IP address (and probable location).
- Unlike web analytics, mobile analytics tools don't depend on cookies to identify unique users since mobile analytics SDKs can generate a persistent and unique identifier for each device.
- The tracking technology varies between websites, which use either JavaScript or cookies and apps, which use a Software Development Kit (SDF).
- Each time a website or app visitor takes an action, the application fires off data which is recorded in the mobile analytics platform.

6.3.1 Difference between Mobile Analytics and Web Analytics

Sr. No.	Mobile analytics	Web analytics
1.	When web site is using, then mobile user called as USER.	When web site is using, then user called as VISITER.
2.	Interaction with site is called as SESSION.	Interaction with site is called as VISTIS.
3.	On mobile, users have less screen real estate (4 to 7 iches) and interact by toucing, swiping and holding.	On a desktop, users have larger screes (10 to 17 inches) and interact by clicking, double-clicking and using key commands.
4.	Session timeout may be as short as 30 seconds.	Session will end after 30 minutes of inactivity for websites.
5.	Unique users are identified via user IDs.	Cookies are used to identifies user.

Review Questions

1. *How mobile analytics is different than social media analytics. Explain with suitable example.*
   SPPU : May-18 (End Sem), Marks 9
2. *What is mobile analytics ? What is the importance of mobile analytics ?*
   SPPU : Dec.-18 (End Sem), Marks 8

6.4 Data Analytics Life Cycle of Case Studies

- Each big data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis. An evaluation of a Big Data analytics business case helps decision-makers understand the business resources that will need to be utilized and which business challenges the analysis will tackle.
- Data identification stage is dedicated to identifying the datasets required for the analysis project and their sources. Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations.
- During the data acquisition and filtering stage, the data is gathered from all of the data sources that were identified during the previous stage. The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.
- Data extraction : Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution. The need to address disparate types of data is more likely with data from external sources. For example,

extracting the required fields from delimited textual data, such as with webserver log files, may not be necessary if the underlying Big Data solution can already directly process those files.

- Data validation and cleansing : Invalid data can skew and falsify analysis results. Unlike traditional enterprise data, where the data structure is pre-defined and data is pre-validated, data input into Big Data analyses can be unstructured without any indication of validity. Its complexity can further make it difficult to arrive at a set of suitable validation constraints.
- The data validation and cleansing stage is dedicated to establishing often complex validation rules and removing any known invalid data.
- Data aggregation and representation : Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID. In other cases, the same data fields may appear in multiple datasets, such as date of birth. Either way, a method of data reconciliation is required or the dataset representing the correct value needs to be determined.
- The data analysis stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered.
- Data visualization : The ability to analyze massive amounts of data and find useful insights carries little value if the only ones that can interpret the results are the analysts.
- Fig 6.4.1 shows data analytics life cycle.

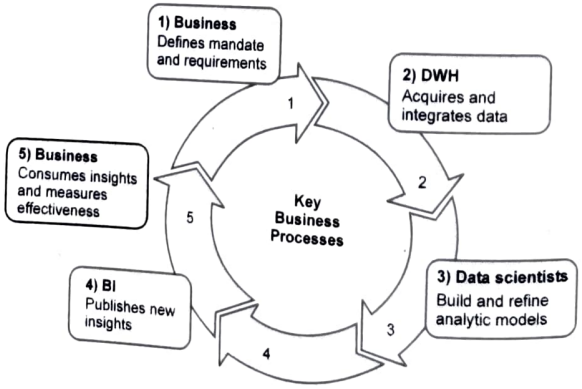


Fig. 6.4.1

6.5 Organizational Impact

- Business Intelligence and data science including wide statistics-helps to describe a possible future event information, data engineering, programming, and data seeing have very varied aspects and require varied skills and approaches.
- Big data has reverted the details for defining and putting into numbers terms such as valuable, important, and successful. It is these details that fuel the that are the source of competitive wholesomeness.
- New big data sources and new wide capabilities of deep learning, supports higher loyalty answers to these questions.
- Most organizations now understand that if they capture all the data that streams into their businesses, they can deploy big data analytics to get significant leverage in understanding their customers, forecasting business trends, reducing operational costs, and realizing more profits. Regardless of industry, data analysis and visualization become accessible and impact business in critical ways.
- Big data analytics helps to improve quality in industries where inconsistencies are hard to reduce.
- Big Data Analytics is beneficial for organizations like Travel and hospitality, Healthcare, Government, Retail and more, that relies on agile and quick decisions to stay competitive.
- The high-performance analytics retrieved from Big Data analytics resources helps organization do things they have never thought before.
- They can get quick results in few seconds rather than days which can in turn accelerate quick reactions to key business challenges and questions.
- Big Data and Analytics are becoming closely intertwined and work together on the unstructured data to get precise answers for hard-to-solve problems and uncover new growth opportunities.
- Optimize the allotment of sufficient sales resources in view of best sales opportunities by the sales department. Identification of great potential and important business account is an equally important task done by the sales team.
- To identify and confirm suppliers who are cost-effective and supply good quality products in a timely manner.
- To measure the device performance and process variance are the main indicators of manufacturing, processing or quality problems.

6.6 Understanding Decision Theory

- Decision theory is the analysis of the behavior of an individual facing non-strategic uncertainty, that is, uncertainty that is due to what we term "Nature" or, if other individuals are involved, their behavior is treated as a statistical distribution known to the decision maker.
- Decision theory depends on probability theory. Decision theory is based on various decisions taken at different stages of analytical life cycle.
- Statistical problems can be interpreted using decision theory. In this view, problems are considered solved when an optimal "decision rule" is chosen from a set of allowed rules.
- The optimal choice is usually given in terms of an optimization problem involving a risk/loss/objective function to be minimized.
- Most decisions are not made in short time due to complicated processes. So it is therefore natural to divide them into phases or stages.
- One interesting specialty of big data is it is challenging to the ordinary thinking. The reason is the suitable large amount of data is generated which need to model by sampling datasets.
- Business Intelligence (BI) is the broad category of skills sets, processes, technologies, applications and practices used to support better decision making.
- BI can also be defined as the ability to access the right data needed at the right time to make informed strategic decisions.
- Significant training and handcrafting were needed or demanded that help to merchant users.
- Business users, who are by their nature are not experts and have struggled to learn and need to combine varied things together so they work as one unit. Also should be able to convert into expertness for their daily merchantry processes.
- Merchantry intelligence tools did not help merchantry users to make the transformation from one thing to another. This comes from deep thinking by understandings of deep things and optimization considering that tools were not unbearable in helping users understanding why something happened.

6.6.1 Business Process

- Decision processes : Most decisions are not made in short time due to complicated processes. So it is therefore natural to divide them into phases or stages. One interesting specialty of big data is it is challenging to the ordinary thinking. The reason is the suitable large amount of data is generated which need to model by sampling datasets.

- Big data professionals don't necessarily have to exercise on huge data in diverse ways. As doing of these things to huge data sets may be the pioneer of some obstacles to understanding why un-revokable behaviors happen.
- Specific area of data analysis where users are trying to apply the statistical algorithm to their data in order to measure cause and effects. This helps to identify the correlation between certain activities and results of it. Users hopes are if they can quantify cause
- The BI vendors added statistical and analytic capabilities to their products, all in the hope of moving business users beyond retrospective reporting into the area of predictive analytics.

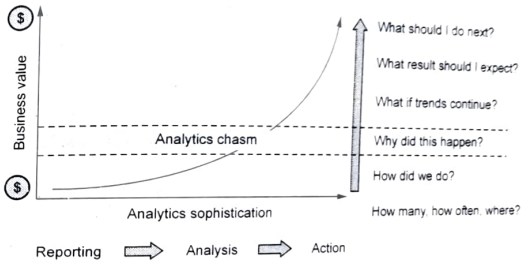


Fig. 6.6.1 Analytics Chasm

- Unfortunately, the BI tools failed to help business users make the transition because the tools were totally inadequate in helping users understand.
- Something happened these tools required business users to quantify cause-and-effect in order to build the models necessary to predict what to do next, and that was beyond their training and interest. As a result, the users' transition to a forward-looking view of their business fell into the "analytics chasm."
- Trying to turn the average business user into a statistical specialist failed, in the early 2000's and it continues to fail today.
- The average business user's career aspiration is not to become a statistical expert. They are in the retail or medical or telecommunications or banking industries because they like that industry, not because they want to master statistics or manipulate large data sets. The tools today are way too hard to make that process trivial.

6.7 Creating Big Data Strategy

- Big data is like traditional data in many ways : It must be captured, stored, organized, and analyzed, and the results of the analysis need to be integrated into established processes and influence how the business operates.
- The importance of big data and analytics has seen it rise to the top of the decision making tree, becoming a C-level decision to address what should be done.
- Five key points for building a data strategy that will help your business be a big data success :
  1. Identify the business problem
  2. People before technology : Support for a big data strategy needs to come from the top
  3. Define policies : Policies are a vital piece of the puzzle and key areas should be thought about such as governance, data ownership and areas of responsibility.
  4. Design information feedback loops
  5. Plan for the future
- Document for Big Data Strategy :
  1. Business Strategy : The targeted business strategy always clearly defines the scopes where the big data initiative will be the focus. This includes the title of the document.
  2. Business Initiatives : This section breaks business plan into various new supporting subsets. This subset is allotted time duration from total approximately 9 months to one-year project duration.
  3. Outcomes and Critical Success Factors (CSF) : This section helps to generate the result and through successful execution of the organizations merchantry approach to finding something new.
  4. Tasks : This is the next level which has various writing task that needs to be done for successful completion of merchantry attempts.
  5. Data Source : Data documents focus on the key data source required and support merchantry plan to achieve their goals through various merchantry attempts

6.8 Big Data Value Creation Drivers

- Understanding value creation process : As some organization has their fixed understanding or imagination about how solutions on big data help to achieve their key merchantry attempt.

- Visualization exercise help among the merchantry user to identify particular areas so can have an idea wherever big data affect their organization business.
- All business users recognize a different type of questions they are trying to answer in view of their key business processes. Fig. 6.8.1 shows Big data drives value creation processes.

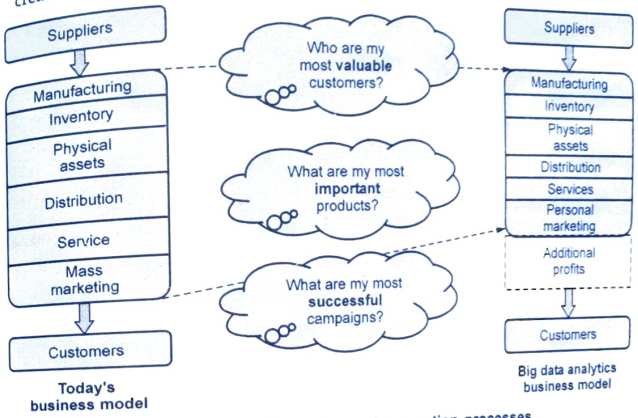


Fig. 6.8.1 Big data drives value creation processes

- Understanding big data value creation process : It is important to understand big data "merchantry" drivers and the values of "megacosm" process. Different "merchantry" drivers are shown below :
  1. Structured Data : Mostly business transactional data. It helps to get Enable more coarse as well as exhaustive decisions.
  2. Unstructured Data : Data from variety of sources mostly in text documents. It helps to get complete and accurate business decisions.
  3. Velocity : Real time data with high speed and very low data latency. It helps in real time business decisions frequently
  4. Predictions : Predictor, Experiment, cause, instrumentations. It help to predict actionable business decisions.

6.9 Michael Porter's Valuation Creation Models

- Five Forces Analysis provides inputs, perspective on an organization's competitive as industry wide and outside-in data as a driver. Following are the five forces :

1. **Competitive jealousy** : This mainly focuses on the total number and volume of other organizations competing to the organization. It also include by and large organizational size and its directions.
2. **Power of supplier** : This includes a factor those are related to the supplier as the reputation of the brand, an area covered by the supplier, various products and provided services. Also gives an idea about the capacity to bid on different products and services available
3. **Power of buyer** : This force helps to "merchantry" as it include buyer choice and preferences, a volume of a buyer and switching frequency and tendency.
4. **Development of Products and Technologies** : This is about quality and price of various products and services offered. Also, exposure to marketplace allocation, market trends, and compliance risk, legislative and government actions is also included under product and technology development.
5. **New market entrants** : This gives an idea about the barriers to the newcomer or the new entry. Different geographical and cultural factors those affects on "merchantry" are also discovered under this title.

6.9.1 Porter's Value Chain Analysis

- The Porter's value chain concept says that there is a chain of events which occur in a company right from the procurement of raw materials to the delivery of goods as well as the post sales service
- Value chain analysis is an analytical framework that assists in identifying business activities that can create value and competitive advantage to the business. Fig. 6.9.1 shows the essence of apple value chain analysis.

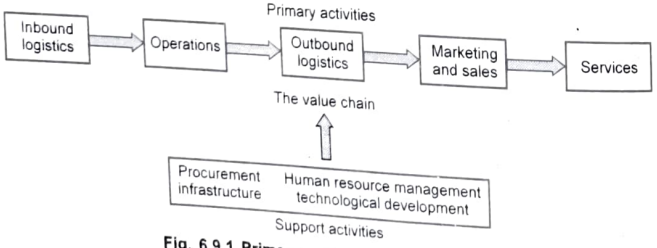


Fig. 6.9.1 Primary and secondary activities

- Most organizations engage in hundreds, even thousands, of activities in the process of converting inputs to outputs. These activities can be classified generally as either primary or support activities that all businesses must undertake in some form.

- According to Porter, the primary activities are :
  1. Inbound Logistics - Involve relationships with suppliers and include all the activities required to receive, store, and disseminate inputs.
  2. Operations - Are all the activities required to transform inputs into outputs (products and services).
  3. Outbound Logistics - Include all the activities required to collect, store, and distribute the output.
  4. Marketing and Sales - Activities inform buyers about products and services, induce buyers to purchase them, and facilitate their purchase.
  5. Service - Includes all the activities required to keep the product or service working effectively for the buyer after it is sold and delivered.
- Secondary activities are :
  1. Procurement - Is the acquisition of inputs, or resources, for the firm.
  2. Human Resource management - Consists of all activities involved in recruiting, hiring, training, developing, compensating and (if necessary) dismissing or laying off personnel.
  3. Technological Development - Pertains to the equipment, hardware, software, procedures and technical knowledge brought to bear in the firm's transformation of inputs into outputs.
  4. Infrastructure - Serves the company's needs and ties its various parts together, it consists of functions or departments such as accounting, legal, finance, planning, public affairs, government relations, quality assurance and general management.

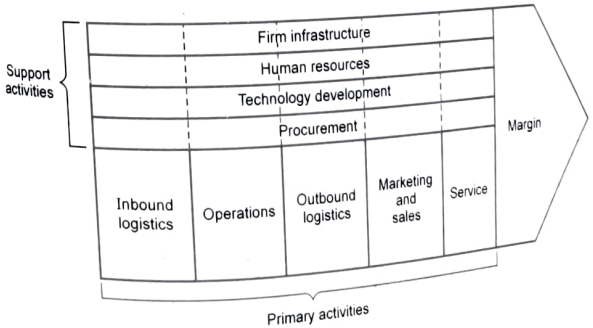


Fig. 6.9.2 Porter's value chain analysis

6.10 Big Data User Experience Ramifications

- This use case provides a good example of the process that one can employ in order to identify the most relevant questions that need to be answered in order to support an organization's key business decisions.
- And it all starts by understanding your organization's key business initiatives.

Step 1 : Understand your organization's key business initiatives.

Step 2 : Capture the decisions that an organization needs to make in order to support the organization's key business initiatives.

Step 3 : Identify those questions that need to be answered in order to facilitate making the decisions.

- Understanding the relationship between your customer's objectives, decisions, and questions that need to be answered is key to creating a user experience that provides the right information to the right customer to make the right decisions at the right time.
- It defines Big Data Model Maturity Index so that products insights extracted from big data have a very powerful impact on consumer user experience.
- There are various opportunities based on customer behavior graph which can be used to find useful and relevant insights for the customer. With these results building the profitable and actionable relationship with the customer can be possible.

6.11 Identifying Big Data Use Cases

SPPU : Dec.-19

- The term "use case" describes an application of data and analytics to improve business performance. To identify potential use cases in a business, the first step is to understand what data is available and what modeling approaches fit the business challenge to solve with the information at hand.
- An effective use case is one in which the application of analytics to a business challenge provides benefits sufficient to justify the investments required to acquire, prepare, analyze, and act on the data. Even as the cost of data capture and storage and analytical processing power fall, this ROI threshold is a high bar for potential use cases to exceed.
- What are the characteristics of potential use cases likely to meet the required ROI threshold? They often solve problems in which even small performance improvements can yield large returns. An example of such a use case is customer price segmentation, often called yield management. Reducing the loss associated

with pricing actions, even if by relatively small amounts, can yield large revenue gains.

- Other use cases with high ROIs are those that avoid substantial costs, such as unexpected down time in a production operation, for example, predicting the failure of important parts where unexpected failure can cause substantial delays.
- A key attribute for successful use case outcomes is that the incremental revenue or reduced cost is substantial and measurable.
- An approach to operationalizing data and analytics that maximizes the potential ROI is to define the application and possible approaches before starting down the path of data capture. Most data will come from the systems in place, so finding out what data they keep and how to access the data will help frame the available modelling options.
- How do merchantry and IT work together to identify the right merchantry opportunity upon which to focus the big data effort to doing something, and then diamond the right use of big data money-making opportunities ?
- How do you make sure of the successful use service of these new big data skills given the upper rate of failure for the adoption of new technologies ?

Cases of use of Big Data in Factories 4.0

- The amount of information produced by IoT and today's manufacturing systems must be translated into actionable ideas. That's why Big Data classifies the information collected and draws relevant conclusions that help improve companies' operations in the following ways :
  a) Improving warehouse processes : Using sensors and portable devices, companies can improve operational efficiency by detecting human errors, performing quality controls and showing optimal production or assembly routes.
  b) Elimination of bottlenecks : Big Data identifies variables that can affect performance, at no extra cost, guiding manufacturers in identifying the problem.
  c) Predictive demand : More accurate and meaningful predictions thanks to the visualization of activity through internal analysis (customer preferences) and external analysis beyond historical data. This allows the company to modify/optimise its product portfolio.
  d) Predictive maintenance : Data fed sensors identify possible failures in the operation of machinery before it becomes a breakdown, by identifying breakdowns in patterns. The system sends an alert to the equipment so that it can react in time.

- Finding the right use case is essential to the success of data science project because it enables :
  a) Understand your problem from an end-user perspective.
  b) Find the right data-driven solution.
  c) Define how to measure the project's success which ensures that your solution adds business value.
  d) Go beyond data science in a technical sense and view your use case from a business perspective.

Review Question

1. *Explain four big data use cases.* **SPPU : Dec.-19 (End Sem), Marks 8**

6.12 Big Data Analytics Challenges and Research Directions

- In recent years, applications of big data and AI in education have made significant headways. This highlights a novel trend in leading-edge educational research. The convenience and embeddedness of data collection within educational technologies, paired with computational techniques have made the analyses of big data a reality.
- We are moving beyond proof-of-concept demonstrations and applications of techniques, and are beginning to see substantial adoption in many areas of education. The key research trends in the domains of big data and AI are associated with assessment, individualized learning, and precision education.
- Model-driven data analytics approaches will grow quickly to guide the development, interpretation, and validation of the algorithms. However, conclusions from educational analytics should, of course, be applied with caution.
- At the education policy level, the government should be devoted to supporting lifelong learning, offering teacher education programs, and protecting personal data.
- With regard to the education industry, reciprocal and mutually beneficial relationships should be developed in order to enhance academia-industry collaboration.
- Furthermore, it is important to make sure that technologies are guided by relevant theoretical frameworks and are empirically tested.
- Intelligent educational systems employing big data techniques are capable of collecting accurate and rich personal data. Data analytics can reveal students' learning patterns and identify their specific needs. Hence, big data and AI have the potential to realize individualized learning to achieve precision education.

6.13 Multiple Choice Questions

Q.1 Which of the following is not considered a primary activity in the value chain framework developed by Michael Porter ?

a Sales and service
b Inbound logistics
c Operations
d Procurement

Q.2 Which of the following is not considered a secondary activity in the value chain framework developed by Michael Porter ?

a Human resource management
b Firm infrastructure management
c Sales and service
d Procurement

Q.3 Which of the following criterion is not applicable to decision-making under risk ?

a Maximize expected return
b Maximize return
c Minimize expect regret
d Knowledge of likelihood occurrence of each state of nature

Q.4 _____ is a list of words that we should ignore when processing documents, since they give no useful information about content.

a Stemming
b Term frequency
c Stoplists
d None

Q.5 Text Mining is _____.

a conceptual
b theoretical
c empirical
d all of the these

Q.6 Using TF - IDF values for features in a uni - gram bag - of - words model should have an effect most similar to which of the following ?

a Lowercasing the data
b Dropout regularization
c Removing stop words
d Increasing the learning rate

Q.7 **A scheme where a weight is assigned to a term based upon the number of occurrences of the term within a document is called :**

a Bag of words

b Document frequency

c Term frequency

d Optimal weight

Answer Keys for Multiple Choice Questions :

Q.1	d	**Q.2**	c
Q.3	b	**Q.4**	c
Q.5	c	**Q.6**	c
Q.7	c		

❑❑❑