**UNIT I BUSINESS INTELLIGENCE**

**Effective and timely decisions – Data, information and knowledge – Role of mathematical models – Business intelligence architectures: Cycle of a business intelligence analysis – Enabling factors in business intelligence projects – Development of a business intelligence system – Ethics and business intelligence.**

## PART-A

1. **Define Business Intelligence.(APRIL/MAY 2017)**

   Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes.

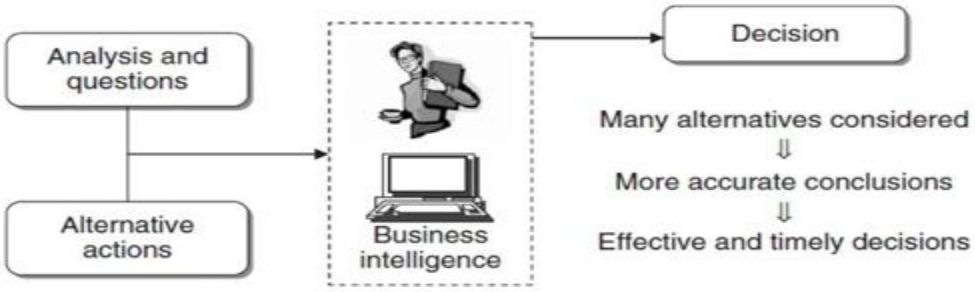
2. **What are effective decisions?**

   The application of rigorous analytical methods allows decision makers to rely on information and knowledge which are more dependable. As a result, they are able to make better decisions and devise action plans that allow their objectives to be reached in a more effective way.

3. **What are timely decisions?**

   Enterprises operate in economic environments characterized by growing levels of competition and high dynamism. As a consequence, the ability to rapidly react to the actions of competitors and to new market conditions is a critical factor in the success or even the survival of a company.

4. **What are the benefits of a business intelligence system? (APRIL/MAY 2017)**




   The decision makers ask themselves a series of questions and develop the corresponding analysis. Hence, they examine and compare several options, selecting among them the best decision, given the conditions at hand. If decision makers can rely on a business intelligence system facilitating their activity, we can expect that the overall quality of the decision-making process will be greatly improved. With the help of mathematical models and algorithms, it is actually possible to analyze a larger number of alternative actions, achieve more accurate conclusions and reach effective and timely decisions.

5. **Define data**.

   Data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities. For example, for a retailer data refer to primary entities such as customers, points of sale and items. They need to be processed by means of appropriate extraction tools and analytical methods capable of transforming them into information and knowledge that can be subsequently used by decision makers.

6. **Difference between information and knowledge.**

| Information | Knowledge |
|---|---|
| Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain | Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. |
| For example:<br>To the sales manager of a retail company, the number of customers holding a loyalty card who have reduced by more than 50% the monthly amount spent in the last three months, represent meaningful pieces of information that can be extracted from raw stored data. | For example:<br>For a retail company, a sales analysis may detect that a group of customers, living in an area where a competitor has recently opened a new point of sale, have reduced their usual amount of business. |

**7. What is attrition ?**

Low customer loyalty, also known as customer attrition or churn, is a critical factor for many companies operating in service industries.

**8. Who are knowledge workers?**

In complex organizations, public or private, decisions are made on a continual basis. Such decisions may be more or less critical, have long- or short-term effects and involve people and roles at various hierarchical levels. The ability of these knowledge workers to make decisions, both as individuals and as a community, is one of the primary factors that influence the performance and competitive strength of a given organization.

**9. What do you mean by knowledge management?**

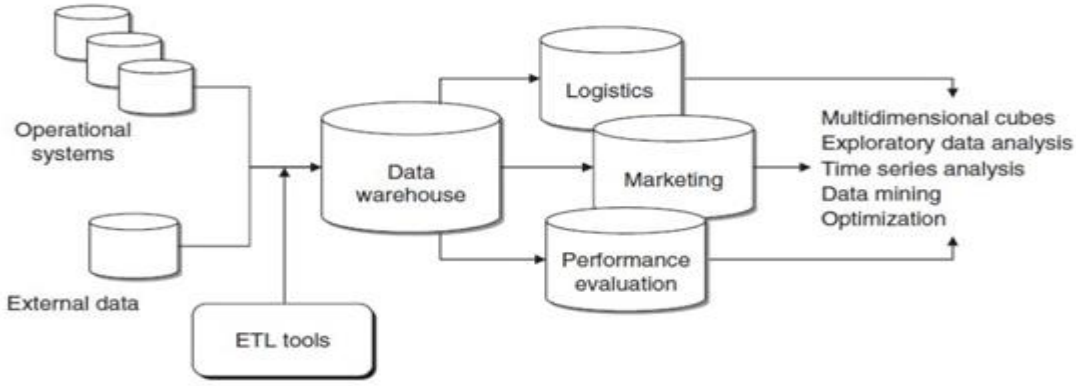
Several public and private enterprises and organizations have developed in recent years formal and systematic mechanisms to gather store and share their wealth of knowledge, which is now perceived as an invaluable intangible asset. The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as knowledge management.

**10. How to summarize a typical business intelligence analysis?**

- First, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.
- Mathematical models are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.
- Finally, what-if analyses are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameters.

**11. What are the three major components of business intelligence architecture?**

(i) Data Sources
(ii) Data warehouses and data marts
(iii) Business intelligence methodologies




**12. What are data sources and data warehouses?**

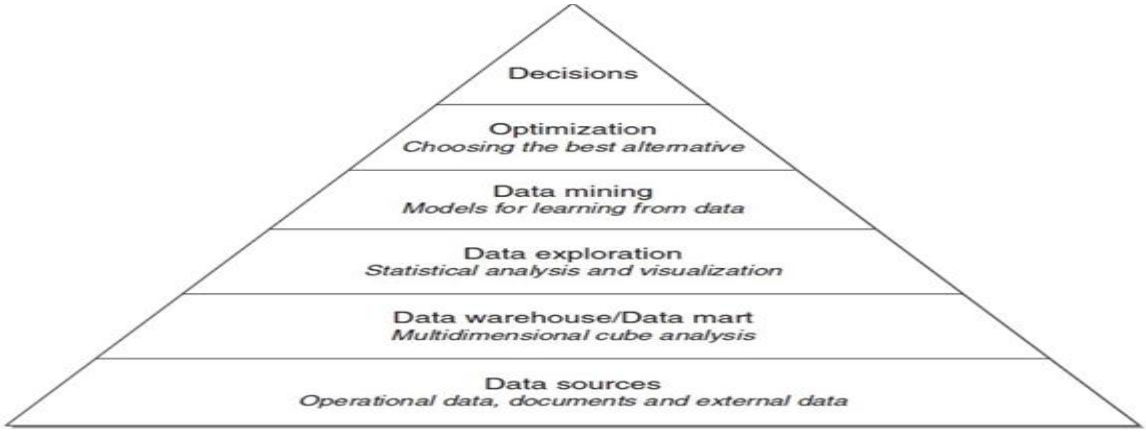
- **Data sources-** In a first stage, it is necessary to gather and integrate the data stored in the various primary and secondary sources, which are heterogeneous in origin and type. The sources consist for the most part of data belonging to operational systems, but may also include unstructured documents, such as emails and data received from external providers. Generally speaking, a major effort is required to unify and integrate the different data sources.
- **Data warehouses -** Using extraction and transformation tools known as extract, transform and load (ETL), the data originating from the different sources are stored in databases intended to support business intelligence analyses. These databases are usually referred to as data warehouses.

**13. What are business intelligence methodologies?**

Data are finally extracted and used to feed mathematical models and analysis methodologies intended to support decision makers. In a business intelligence system, several decision support applications may be implemented, such as:

• multidimensional cube analysis.
• exploratory data analysis.
• time series analysis.
• inductive learning models for data mining.
• optimization models.

**14. What are the main components of business intelligence system?**




.

**15. What is data exploration?**

At this level of pyramid, find the tools for performing a *passive* business intelligence analysis, which consist of query and reporting systems, as well as statistical methods. These are referred to as passive methodologies because decision makers are requested to generate prior hypotheses or define data extraction criteria, and then use the analysis tools to find answers and confirm their original insight

**16. What is data mining?**

At this level, it includes active business intelligence methodologies, whose purpose is the extraction of information and knowledge from data. These include mathematical models for pattern recognition, machine learning and data mining techniques. Unlike the tools described at the previous level of the pyramid, the models of an active kind do not require decision makers to formulate any prior hypothesis to be later verified. Their purpose is instead to expand the decision makers' knowledge.

**17. What are decisions?**

A decision corresponds to the choice and the actual adoption of a specific decision and in some way represents the natural conclusion of the decision-making process. Even when business intelligence methodologies are available and successfully adopted, the choice of a decision pertains to the decision makers, who may also take advantage of informal and unstructured information available to adapt and modify the recommendations and the conclusions achieved through the use of mathematical models.

**18. Who are all responsible for functioning of a good business intelligence system?**

The required competencies are provided for the most part by the information systems specialists within the organization, usually referred to as database administrators. Analysts and experts in mathematical and statistical models are responsible for the intermediate phases. Finally, the activities of decision makers responsible for the application domain appear dominant at the top.

**19. What are the phases involve in the cycle of business intelligence analysis?**

(i) Analysis
(ii) Insight
(iii) Decision
(iv) Evaluation

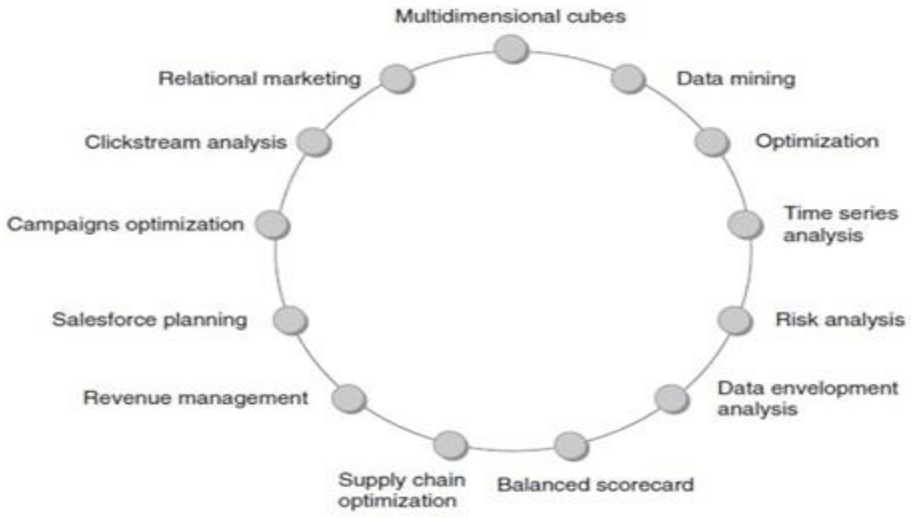
**20. What is analysis phase in the cycle of Business intelligence analysis?**

During the analysis phase, it is necessary to recognize and accurately spell out the problem at hand. Decision makers must then create a mental representation of the phenomenon being analyzed, by identifying the critical factors that are perceived as the most relevant. The availability of business intelligence methodologies may help already in this stage, by permitting decision makers to rapidly develop various paths of investigation. Thus, the first phase in the business intelligence cycle leads decision makers to ask several questions and to obtain quick responses in an interactive way.

**21. What are the phases of the development of a business intelligence system?**

(i) Analysis
(ii) Design
(iii) Planning
(iv) Implementation and control.

**22. What are the methodologies available in business intelligence system?**




**23. What is the need for decision support system?( NOV/DEC 2017)**

Decision support system is a computer solution using statistical data to help managers and operation planners overcome strategic deficiencies in order to implement streamlined, efficient solutions. DSS help those in positions of authority compile enough information to make an informed decision about changes in policy, implementation and so forth.

**24. What are structured, unstructured and semi-structured decisions NOV/DEC 2017)**

- Structured decisions are routine are repetitive decisions often taken at operational management level.
- Semi-structured decisions involve combination of standard procedures and unstructured elements taken at middle management level.
- Unstructured decisions are fuzzy , strategic and complex taken at top management level.

## PART-B

**1) Explain the need and benefits of business intelligence system**.

Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes. Let us describe in general terms the problems entailed in business intelligence, highlighting the interconnections with other disciplines and identifying the primary components typical of a business intelligence environment.

In complex organizations, public or private, decisions are made on a continual basis. Such decisions may be more or less critical, have long- or short-term effects and involve people and roles at various hierarchical levels. The ability of these knowledge workers to make decisions, both as individuals and as a community, is one of the primary factors that influence the performance and competitive strength of a given organization. Most knowledge workers reach their decisions primarily using easy and intuitive methodologies, which take into account specific elements such as experience, knowledge of the application domain and the available information. This approach leads to a stagnant decision-making style which is inappropriate for the unstable conditions determined by frequent and rapid changes in the economic environment. Indeed, decision-making processes within today's organizations are often too complex and dynamic to be effectively dealt with through an intuitive approach, and require instead a more rigorous attitude based on analytical methodologies and mathematical models. The importance and strategic value of analytics in determining competitive advantage for enterprises has been recently pointed out by several authors. Examples 1 and 2 illustrate two highly complex decision-making processes in rapidly changing conditions.

**Example 1:**

The marketing manager of a mobile phone company realizes that a large number of customers are discontinuing their service, leaving her company in favor of some competing provider. As can be imagined, low customer loyalty, also known as customer attrition or churn, is a critical factor for many companies operating in service industries.

Suppose that the marketing manager can rely on a budget adequate to pursue a customer retention campaign aimed at 2000 individuals out of a total customer base of 2 million people. Hence, the question naturally arises of how she should go about choosing those customers to be contacted so as to optimize the effectiveness of the campaign. In other words, how can the probability that each single customer will discontinue the service be estimated so as to target the best group of customers and thus reduce churning and maximize customer retention? By knowing these probabilities, the target group can be chosen as the 2000 people having the highest churn likelihood among the customers of high business value. Without the support of advanced mathematical models and data mining techniques, , it would be arduous to derive a reliable estimate of the churn probability and to determine the best recipients of a specific marketing campaign.

**Example 2 :**

The logistics manager of a manufacturing company wishes to develop a medium-term logistic-production plan. This is a decision-making process of high complexity which includes, among other choices, the allocation of the demand originating from different market areas to the production sites, the procurement of raw materials and purchased parts from suppliers, the production planning of the plants and the distribution of end products to market areas. In a typical manufacturing company this could well entail tens of facilities, hundreds of suppliers, and thousands of finished goods and components, over a time span of one year divided into weeks. The magnitude and complexity of the problem suggest that advanced optimization models are required to devise the best logistic plan. Optimization models allow highly complex and large-scale problems to be tackled successfully within a business intelligence framework. The main purpose of business intelligence systems is to provide knowledge workers with tools and methodologies that allow them to make effectively and timely decisions.

**(i)Effective decisions:** The application of rigorous analytical methods allows decision makers to rely on information and knowledge which are more dependable. As a result, they are able to make better decisions and devise action plans that allow their objectives to be reached in a more effective way. Indeed, turning to formal analytical methods forces decision makers to explicitly describe both the criteria for evaluating alternative choices and the mechanisms regulating the problem under investigation. Furthermore, the ensuing in-depth examination and thought lead to a deeper awareness and comprehension of the underlying logic of the decision-making process.

**(ii)Timely decisions:** Enterprises operate in economic environments characterized by growing levels of competition and high dynamism. As a consequence, the ability to rapidly react to the actions of competitors and to new market conditions is a critical factor in the success or even the survival of a company. Figure 1.1 illustrates the major benefits that a given organization may draw from the adoption of a business intelligence system. When facing problems such as those described in Examples 1.1 and 1.2 above, decision makers ask themselves a series of questions and develop the corresponding analysis. Hence, they examine and compare several options, selecting among them the best decision, given the conditions at hand. If decision makers can rely on a business intelligence system facilitating their activity, we can expect that the overall quality of the decision-making process will be greatly improved. With the help of mathematical models and algorithms, it is actually possible to analyze a larger number of alternative actions, achieve more accurate conclusions and reach effective and timely decisions.

As observed above, a vast amount of data has been accumulated within the information systems of public and private organizations. These data originate partly from internal transactions of an administrative, logistical and commercial nature and partly from external sources. However, even if they have been gathered and stored in a systematic and structured way, these data cannot be used directly for decision-making purposes. They need to be processed by means of appropriate extraction tools and analytical methods capable of transforming them into information and knowledge that can be subsequently used by decision makers. The difference between data, information and knowledge can be better understood through the following remarks

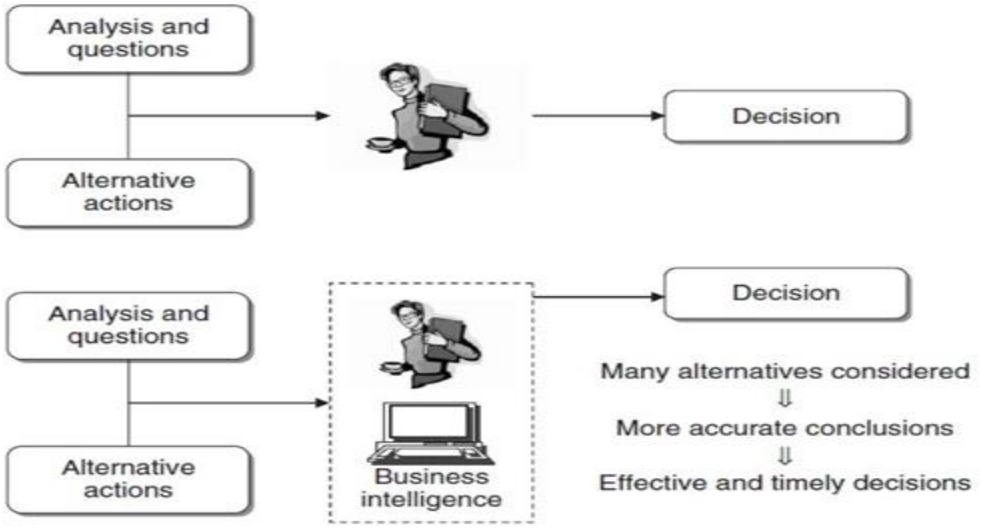



*Figure 1.1 Benefits of a business intelligence system*

**(i)Data.** Generally, data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities. For example, for a retailer data refer to primary entities such as customers, points of sale and items, while sales receipts represent the commercial transactions.

**(ii)Information.** Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain. For example, to the sales manager of a retail company, the proportion of sales receipts in the amount of over 100 per week, or the number of customers holding a loyalty card who have reduced by more than 50% the monthly amount spent in the last three months, represent meaningful pieces of information that can be extracted from raw stored data.

**(iii)Knowledge.** Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. Therefore, we can think of knowledge as consisting of information put to work into a specific domain, enhanced by the experience and competence of decision makers in tackling and solving complex problems. For a retail company, a sales analysis may detect that a group of customers, living in an area where a competitor has recently opened a new point of sale, have reduced their usual amount of business.

The knowledge extracted in this way will eventually lead to actions aimed at solving the problem detected, for example by introducing a new free home delivery service for the customers residing in that specific area. We wish to point out that knowledge can be extracted from data both in a passive way, through the analysis criteria suggested by the decision makers, or through the active application of mathematical models, in the form of inductive learning or optimization, as described in the following chapters. Several public and private enterprises and organizations have developed in recent years formal and systematic mechanisms to gather, store and share their wealth of knowledge, which is now perceived as an invaluable intangible asset. The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as knowledge management.

It is apparent that business intelligence and knowledge management share some degree of similarity in their objectives. The main purpose of both disciplines is to develop environments that can support knowledge workers in decision-making processes and complex problem-solving activities. To draw a boundary between the two approaches, we may observe that knowledge management methodologies primarily focus on the treatment of information that is usually unstructured, at times implicit, contained mostly in documents, conversations and past experience. Conversely, business intelligence systems are based on structured information, most often of a quantitative nature and usually organized in a database. However, this distinction is a somewhat fuzzy one: for example, the ability to analyze emails and web pages through text mining methods progressively induces business intelligence systems to deal with unstructured information.

**2) Explain business intelligence architecture in detail (or) Explain the framework of business intelligence in detail.(APRIL/MAY 2017)(NOV/DEC 2017)**

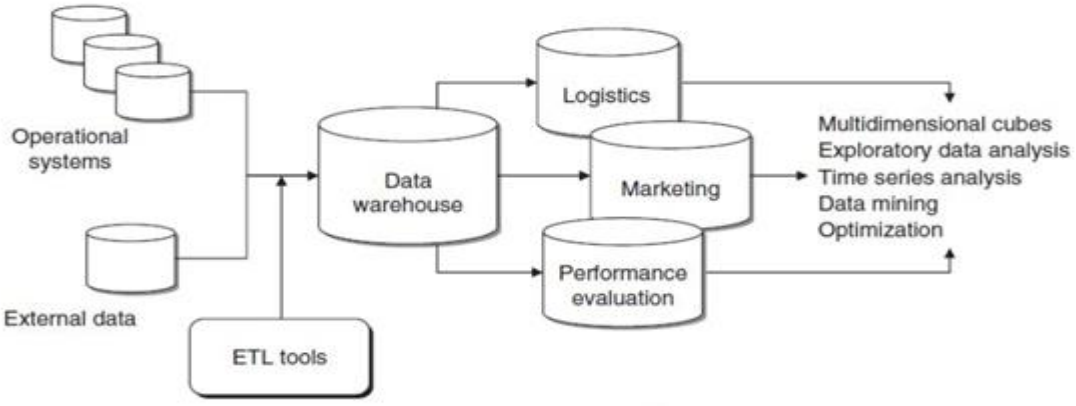
The architecture of a business intelligence system, depicted below, includes three major components.

**(i)Data sources-** In a first stage, it is necessary to gather and integrate the data stored in the various primary and secondary sources, which are heterogeneous in origin and type. The sources consist for the most part of data belonging to operational systems, but may also include unstructured documents, such as emails and data received from external providers. Generally speaking, a major effort is required to unify and integrate the different data sources.

**(ii)Data warehouses and data marts-** Using extraction and transformation tools known as *extract, transform, load* (ETL), the data originating from the different sources are stored in databases intended to support business intelligence analyses. These databases are usually referred to as *data warehouses* and *data marts*.

**(iii)Business intelligence methodologies-** Data are finally extracted and used to feed mathematical models and analysis methodologies intended to support decision makers. In a business intelligence system, several decision support applications may be implemented, such as:

- multidimensional cube analysis.
- exploratory data analysis.
- time series analysis.
- inductive learning models for data mining.
- optimization models.




The pyramid shows the building blocks of a business intelligence system. So far, we have seen the components of the first two levels when discussing Figure 1.2. We now turn to the description of the upper tiers.

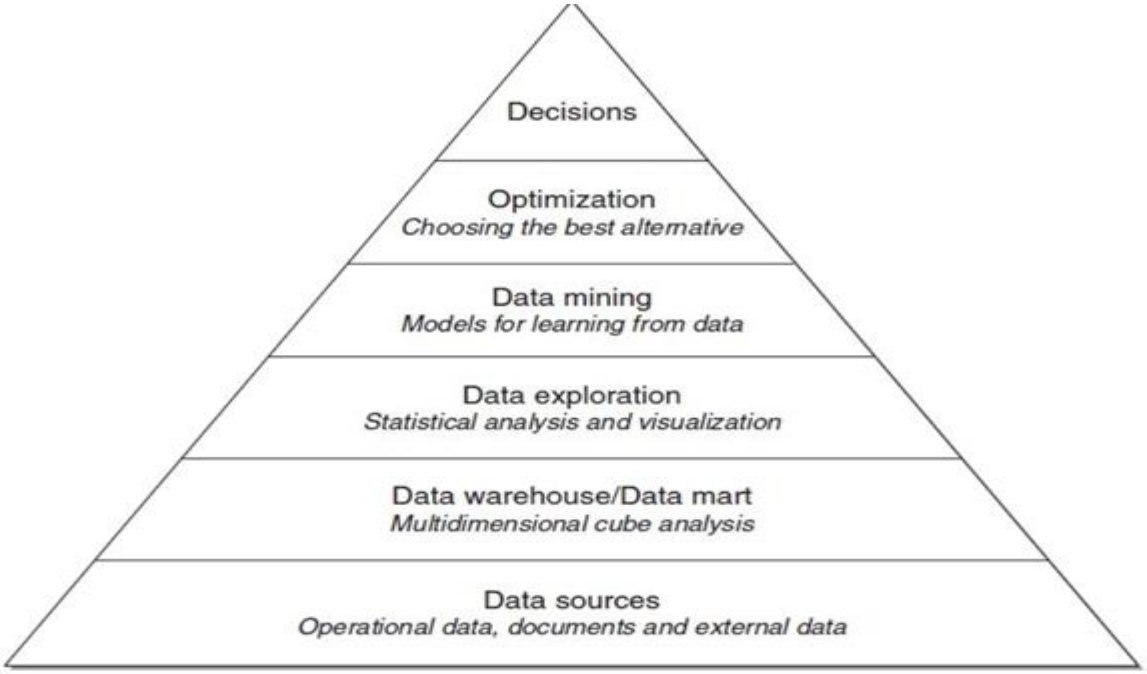
**(i)Data exploration.** At the third level of the pyramid we find the tools for performing a *passive* business intelligence analysis, which consist of query and reporting systems, as well as statistical methods. These are referred to as passive methodologies because decision makers are requested to generate prior hypotheses or define data extraction criteria, and then use the analysis tools to find answers and confirm their original insight.For instance, consider the sales manager of a company who notices that revenues in a given geographic area have dropped for a specific group of customers. Hence, she might want to bear out her hypothesis by using extraction and visualization tools, and then apply a statistical test to verify that her conclusions are adequately supported by data.

**(ii)Data mining.** The fourth level includes *active* business intelligence methodologies, whose purpose is the extraction of information and knowledge from data. These include mathematical models for pattern recognition, machine learning and data mining techniques. Unlike the tools described at the previous level of the pyramid, the models of an active kind do not require decision makers to formulate any prior hypothesis to be later verified. Their purpose is instead to expand the decision makers' knowledge.

**(iii)Optimization.** By moving up one level in the pyramid we find optimization models that allow us to determine the best solution out of a set of alternative actions, which is usually fairly extensive and sometimes even infinite. Example 1.2 shows a typical field of application of optimization models.

**(iv)Decisions.** Finally, the top of the pyramid corresponds to the choice and the actual adoption of a specific decision, and in some way represents the natural conclusion of the decision-making process. Even when business intelligence methodologies are available and successfully adopted, the choice of a decision pertains to the decision makers, who may also take advantage of informal and unstructured information available to adapt and modify the recommendations and the conclusions achieved through the use of mathematical models.

As we progress from the bottom to the top of the pyramid, business intelligence systems offer increasingly more advanced support tools of an active type. Even roles and competencies change. At the bottom, the required competencies are provided for the most part by the




information systems specialists within the organization, usually referred to as *database administrators*. Analysts and experts in mathematical and statistical models are responsible for the intermediate phases. Finally, the activities of decision makers responsible for the application domain appear dominant at the top.

As described above, business intelligence systems address the needs of different types of complex organizations, including agencies of public administration and associations. However, if we restrict our attention to enterprises, business intelligence methodologies can be found mainly within three departments of a company, as depicted : marketing and sales; logistics and production; accounting and control.

**3) Explain in detail about the role of mathematical models.**

A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms. In some instances, this activity may reduce to calculations of totals and percentages, graphically represented by simple histograms, whereas more elaborate analyses require the development of advanced optimization and learning models. In general terms, the adoption of a business intelligence system tends to promote a scientific and rational approach to the management of enterprises and complex organizations. Even the use of a spreadsheet to estimate the effects on the budget of fluctuations in interest rates, despite its simplicity, forces decision makers to generate a mental representation of the financial flows process. Classical scientific disciplines, such as physics, have always resorted to mathematical models for the abstract representation of real systems. Other disciplines, such as operations research, have instead exploited the application of scientific methods and mathematical models to the study of artificial systems, for example public and private organizations. The rational approach typical of a business intelligence analysis can be summarized schematically in the following main characteristics.

• First, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.
• Mathematical models are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.
• Finally, what –if analyses are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameters.
Although their primary objective is to enhance the effectiveness of the decision making process, the adoption of mathematical models also affords other advantages, which can be appreciated particularly in the long term.

First, the development of an abstract model forces decision makers to focus on the main features of the analyzed domain, thus inducing a deeper understanding of the phenomenon under investigation. Furthermore, the knowledge about the domain acquired when building a mathematical model can be more easily transferred in the long run to other individuals within the same organization, thus allowing a sharper preservation of knowledge in comparison to empirical decision-making processes. Finally, a mathematical model developed for a specific decision-making task is so general and flexible that in most cases it can be applied to other ensuing situations to solve problems of similar type.

**4) Explain all the phases in the cycle of business intelligence analysis.**

Each business intelligence analysis follows its own path according to the application domain, the personal attitude of the decision makers and the available analytical methodologies. However, it is possible to identify an ideal cyclical path characterizing the evolution of a typical business intelligence analysis, even though differences still exist based upon the peculiarity of each specific context.

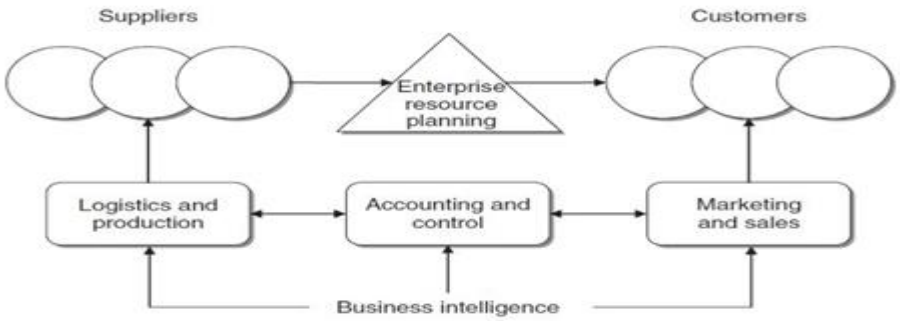
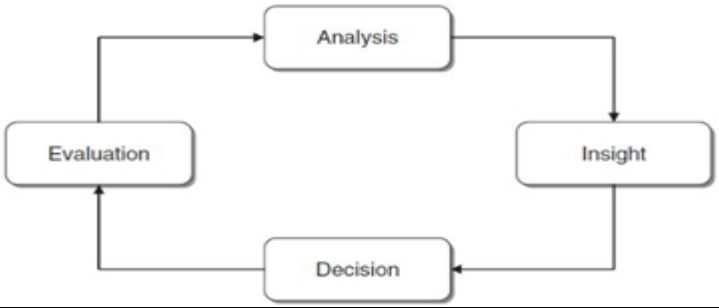



*Figure 1.4 Departments of an enterprise concerned with business intelligence systems*

**(i)Analysis.** During the analysis phase, it is necessary to recognize and accurately spell out the problem at hand. Decision makers must then create a mental representation of the phenomenon being analyzed, by identifying the critical factors that are perceived as the most relevant. The availability of business intelligence methodologies may help already in this stage, by permitting decision makers to rapidly develop various paths of investigation. For instance, the exploration of data cubes in a multidimensional analysis, according to different logical views, allows decision makers to modify their hypotheses flexibly and rapidly, until they reach an interpretation scheme that they deem satisfactory. Thus, the first phase in the business intelligence cycle leads decision makers to ask several questions and to obtain quick responses in an interactive way.



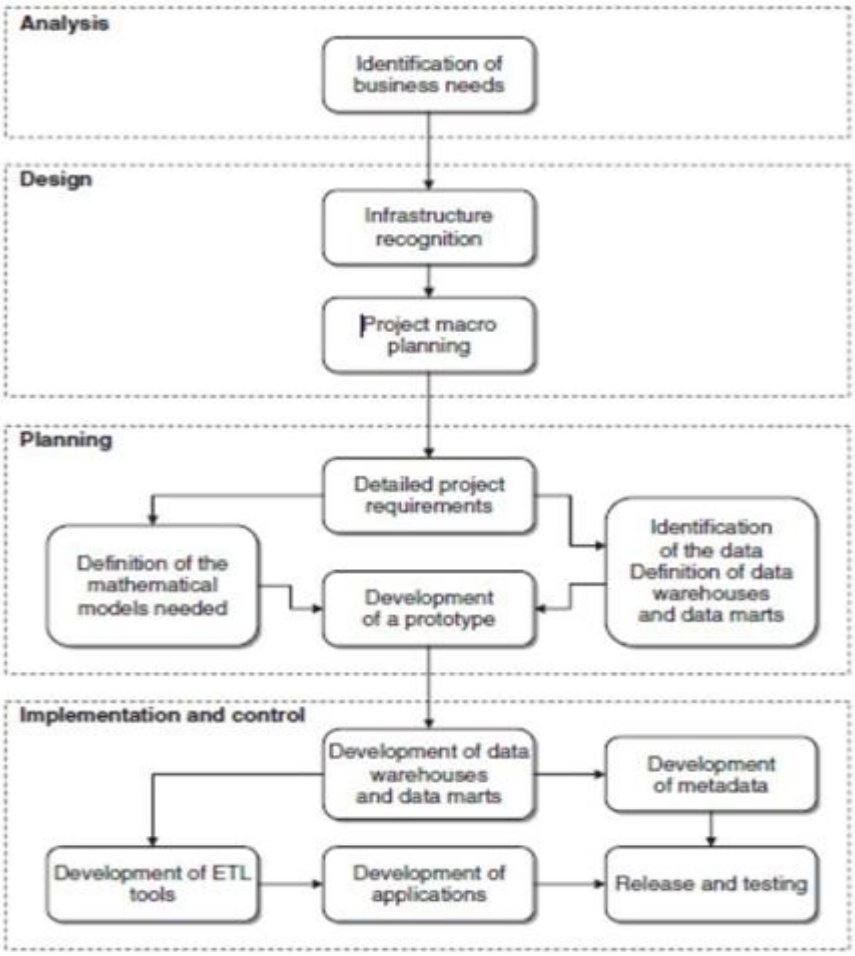
**(ii) Insight.** The second phase allows decision makers to better and more deeply understand the problem at hand, often at a causal level. For instance, if the analysis carried out in the first phase shows that a large number of customers are discontinuing an insurance policy upon yearly expiration, in the second phase it will be necessary to identify the profile and characteristics shared by such customers. The information obtained through the analysis phase is then transformed into knowledge during the insight phase. On the one hand, the extraction of knowledge may occur due to the intuition of the decision makers and therefore be based on their experience and possibly on unstructured information available to them. On the other hand, inductive learning models may also prove very useful during this stage of analysis, particularly when applied to structured data.

**(iii) Decision.** During the third phase, knowledge obtained as a result of the insight phase is converted into decisions and subsequently into actions The availability of business intelligence methodologies allows the analysis and insight phases to be executed more rapidly so that more effective and timely decisions can be made that better suit the strategic priorities of a given organization. This leads to an overall reduction in the execution time of the *analysis-decision-action- revision* cycle, and thus to a decision-making process of better quality.

**(iv) Evaluation.** Finally, the fourth phase of the business intelligence cycle involves performance measurement and evaluation. Extensive metrics should then be devised that are not exclusively limited to the financial aspects but also take into account the major performance indicators defined for the different company departments. DEA is the powerful analytical methodologies for performance evaluation.

**5) Explain all the phases in the development of business intelligence system in detail.(APRIL/MAY 2017)**

The development of a business intelligence system can be assimilated to a project, with a specific final objective, expected development times and costs, and the usage and coordination of the resources needed to perform planned activities. Figure 1.6 shows the typical development cycle of a business intelligence architecture. Obviously, the specific path followed by each organization might differ from that outlined in the figure. For instance, if the basic information structures, including the data warehouse and the data marts, are already in place,

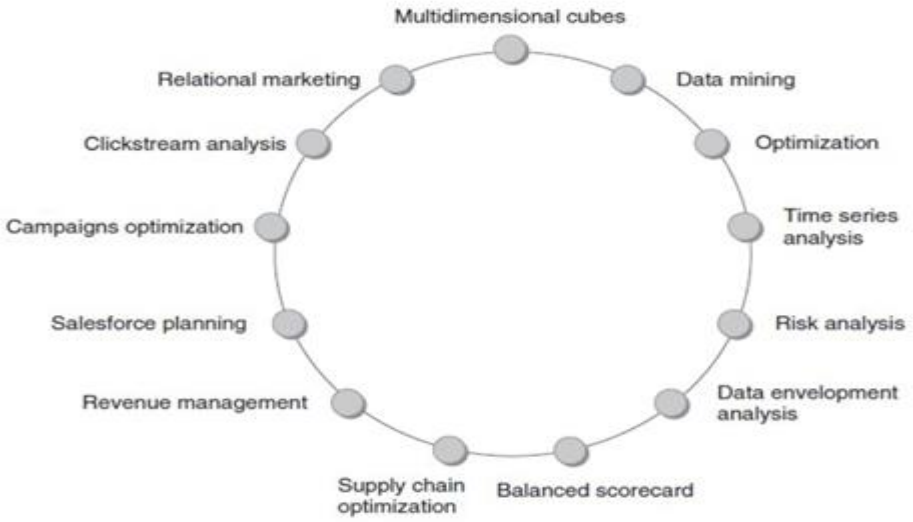


**(i) Analysis:** During the first phase, the needs of the organization relative to the development of a business intelligence system should be carefully identified. This preliminary phase is generally conducted through a series of interviews of knowledge workers performing different roles and activities within the organization. It is necessary to clearly describe the general objectives and priorities of the project, as well as to set out the costs and benefits deriving from the development of the business intelligence system design**.** The second phase includes two sub-phases and is aimed at deriving a provisional plan of the overall architecture, taking into account any development in the near future and the evolution of the system in the midterm.

First, it is necessary to make an assessment of the existing information infrastructures. Moreover, the main decision-making processes that are to be supported by the business intelligence system should be examined, in order to adequately determine the information requirements. Later on, using classical project management methodologies, the project plan will be laid down, identifying development phases, priorities, expected execution times and costs, together with the required roles and resources.

**(ii)Planning:** The planning stage includes a sub-phase where the functions of the business intelligence system are defined and described in greater detail. Subsequently, existing data as well as other data that might be retrieved externally are assessed. This allows the information structures of the business intelligence architecture, which consist of a central data warehouse and possibly some satellite data marts, to be designed. Simultaneously with the recognition of the available data, the mathematical models to be adopted should be defined, ensuring the availability of the data required to feed each model and verifying that the efficiency of the algorithms to be utilized will be adequate for the magnitude of the resulting problems. Finally, it is appropriate to create a system prototype, at low cost and with limited capabilities, in order to uncover beforehand any discrepancy between actual needs and project specifications.

**(iii)Implementation and control:** The last phase consists of five main sub-phases. First, the data warehouse and each specific data mart are developed. These represent the information infrastructures that will feed the business intelligence system. In order to explain the meaning of the data contained in the data warehouse and the transformations applied in advance to the primary data, a *metadata* archive should be created. Moreover, ETL procedures are set out to extract and transform the data existing in the primary sources, loading them into the data warehouse and the data marts. The next step is aimed at developing the core business intelligence applications that allow the planned analyses to be carried out. Finally, the system is released for test and usage.




The above figure provides an overview of the main methodologies that may be included in a business intelligence system, most of which will be described in the following chapters. Some of them have a methodological nature and can be used across different application domains, while others can only be applied to specific tasks.

**6) Discuss about ethics to be followed in business intelligence.**

The adoption of business intelligence methodologies, data mining methods and decision support systems raises some ethical problems that should not be overlooked. Indeed, the progress toward the information and knowledge society opens up countless opportunities, but may also generate distortions and risks which should be prevented and avoided by using adequate control rules and mechanisms. Usage of data by public and private organizations that is improper and does not respect the individuals' right to privacy should not be tolerated.

More generally, we must guard against the excessive growth of the political and economic power of enterprises allowing the transformation processes outlined above to exclusively and unilaterally benefit such enterprises themselves, at the expense of consumers, workers and inhabitants of the Earth ecosystem. However, even failing specific regulations that would prevent the abuse of data gathering and invasive investigations, it is essential that business intelligence analysts and decision makers abide by the ethical principle of respect for the personal rights of the individuals. The risk of overstepping the boundary between correct and intrusive use of information is particularly high within the relational marketing and web mining fields, are important. For example, even if disguised under apparently inoffensive names such as 'data enrichment', private information on individuals and households does circulate, but that does not mean that it is ethical for decision makers and enterprises to use it. Respect for the right to privacy is not the only ethical issue concerning the use of business intelligence systems. There has been much discussion in recent years of the social responsibilities of enterprises, leading to the introduction of the new concept of stakeholders. This term refers to anyone with any interest in the activities of a given enterprise, such as investors, employees, labor unions and civil society as a whole.

There is a diversity of opinion on whether a company should pursue the short-term maximization of profits, acting exclusively in the interest of shareholders, or should instead adopt an approach that takes into account the social consequences of its decisions. As this is not the right place to discuss a problem of such magnitude, we will confine ourselves to pointing out that analyses based on business intelligence systems are affected by this issue and therefore run the risk of being used to maximize profits even when different considerations should prevail related to the social consequences of the decisions made, according to a logic that we believe should be rejected.

**7) Discuss briefly about the user interface in business intelligence system. (APRIL/MAY 2017)**

**B**usiness intelligence (BI) tools are the window to the data that helps business users identify new opportunities and make fact-based decisions. Despite the critical role that BI tools provide, adoption remains low and BI tools are considered difficult to use, with largely unappealing interfaces.

Assessing a BI tool's ease of use is difficult because it is influenced by subjective factors. However, the importance of ease of use rated even higher than specific tool capabilities and analytic power makes it a critical aspect to address in making BI more pervasive and with bigger business impact. Each group of potential BI users whether power users who work with data for most of the day, or casual users who only have a short time to glance at information brings their own set of experiences and expectations to a BI tool. Customers and vendors a like need to consider these differences so they can improve ease of use, leveraging whichever innovation is appropriate for a particular type of application and user segment.

Currently, BI tools are rated as some of the most difficult to use relative to a variety of technologies including Google, email, and Excel. Power users rate BI tools as more difficult than casual users. Companies that described their predominant BI tool as very easy to use had a much higher BI adoption rate than those who described their BI tool as only somewhat easy or difficult.

An appealing interface provides a powerful first impression, particularly when trying to embrace new users. The way a product looks and works influences how pleasant and effective it is to continue using a particular BI tool. With BI, a cumbersome interface and workflow competes with other ways of working, whether gut feel decision making or asking the expert. If someone doesn't have to absolutely get to the data to do his or her job, an unappealing interface can be a barrier to continued use. In considering various interfaces within a BI platform, the authoring interface is rated the least appealing. Dashboard interfaces rated the most appealing, but still leave room for improvement.

In trying to address ease of use and interface appeal, vendors have introduced a number of innovations such as BI's integration with Microsoft Office, in particular Excel, Adobe Flash animation, and BI search. Of these innovations, Microsoft Office integration is the most widely adopted. Lackluster adoption for other innovations is mainly because customers have not purchased the option. Despite claims that BI tools should be easy enough to use so that training is not required, most people want some training, at least half a day. Even if the BI tool is easy to use, understanding the data requires time and explanation.

**8) Explain the scope for business intelligence and business value in the following applications (NOV/DEC 2017)**

**(i) Customer segmentation**

For each customer segment, what are various measures like sales revenues, sales profitability, average customer tenure etc. This analysis helps you to validate the assumptions you have made on a customer segment behavior. Over months and quarters, you can re-validate these assumptions and fine-tune your strategy. For example, if you find that a given segment is buying more than expected, you may heighten the advertising budget for that segment. You may also like to go back and check your methods which you deployed in the previous customer segment to behavior analysis.

**Customer segment movement over time**: You can analyze on how the customer moments happening across the segments over time. For example, over a year- %age distribution of your customers in various segments could change. This information will tell you, if your customer acquisition is as per strategy. Taking this example further- If you find that % of low-income, low-usage customers have gone up over last one year, it may not be matching with your strategy to go for high usage customers. Another example is – Tracking on how are we converting high risk customers to lower risk, and medium value customers to higher value? For example- one of the measures of success of a 'customer value' initiative could be on how many customers have moved from low value to medium value. The way to achieve it is to have a time-stamp as the attribute to the customer dimension. For example, if you are having a way to assign the Customer segment class every month, a new record will be added with a time-stamp.

**New Customer segment growth:** This analysis talks on the profile on new customers which are acquired. For example- one may note that while existing customer portfolio is fairly moderate risk, the customers acquired in last three months are primarily high risk. This becomes important because in a normal analysis, last three months acquisitions may not move the averages, but it serves the shareholders to know that sales function is suddenly getting over-aggressive in acquisition.

**(ii) Customer profitability**

It is important to measure profitability over the duration of a cycle that is inherent in the business (e.g. lifetime of a car lease, a growing season for a farmer, a project lifetime for a building contractor, a redemption cycle for a loyalty program customer). There are a wide variety of measurements that may be used to provide insight into the profitability of your customers' business. Choosing what to measure and how to measure will profoundly affect the usefulness of the information you produce.Sales effectiveness is best monitored through a combination of activity measurements and results measurements. Activity measurements are needed to promote productivity and identify actions that can be taken to improve individual performance. As management guru Tom Peters would say, this is about "doing the thing right."

Results measurement is needed to ensure that our sales people are doing "the right thing." Historically, we have spent lots of time, effort and money acquiring business that provides low or negative value to our companies. The classic statistic first uncovered in retail banking is that 80% of their client base is at or below the zero profit mark. As measurement of customer profitability has spread, we hear similar findings echoed in a number of different industries.It is quite clear that acquiring profitable customers is a key to managing the margin and the bottom line of your business. Without a disciplined analysis of the profitability of your existing client base it is very difficult to tell which types of customers you should be identifying for acquisition. All of the other target marketing information you presently use remains valid and useful: the difference is you learn which customers you want to acquire.

**(iii) Fraud detection**

a) Reduce operational risk and fraud losses by detecting fraud and other malicious activity in real-time.

b) Deter potential fraudulent users just by knowing that all their actions are recorded and evaluated in real-time.

c) Upgrade the efficiency of internal audits by alerting on detection of suspicious behavior and providing internal auditors with full visibility to all actions of each and every suspicious end-user.

d) Enforce corporate security policies by detecting security breaches and exceptions.

e) Improve compliance with government regulations by creating a full audit trail for all end-user activity, including queries that typically do not leave traces in most systems.

**UNIT II KNOWLEDGE DELIVERY**

**The business intelligence user types, Standard reports, Interactive Analysis and Ad Hoc Querying, Parameterized Reports and Self-Service Reporting, dimensional analysis, Alerts/Notifications, Visualization: Charts, Graphs, Widgets, Scorecards and Dashboards, Geographic Visualization, Integrated Analytics, Considerations: Optimizing the Presentation for the Right Message.**

## PART-A

1. **What is Knowledge Delivery .**

   Knowledge Delivery has evolved into data visualization, executive reporting, and real-time dashboards. The mainstream of Knowledge Delivery has adopted some strict rules, or grammar, associated with the presentation of information in dashboards, reports, etc.

2. **What are the Business Intelligence User types.**

   Power users, Business users, Casual users, Data aggregators or Information Providers, Operational analytics users, extended enterprise users, IT users.

3. **What are the different modes of presentation that are relevant to different user types.**
   - Standard reports
   - Interactive Analysis and Ad Hoc Querying
   - Parameterized Reports and Self-Service Reporting

4. **What are the groups of items being measured in standard reports.**
   - The number of owner-occupied housing units with a mortgage
   - Value of the houses
   - Mortgage status
   - Household income for the previous 12 months

5. **What are standard reports in BI.**

   **Standard reports** usually have a fixed format, are parameter-driven and, in their simplest form, are pre-run. Standard reports provide a core set of information about what's going on in a particular business area. These reports are the backbone of BI applications.

6. **What is adhoc querying.**

   An Ad-Hoc Query is a query that cannot be determined prior to the moment the query is issued. It is created in order to get information when need arises. Ad hoc queries are used intensively in the internet. Search engines process millions of queries every single second from different data sources. Any keywords typed the internet user are dynamically generated with an ad hoc query against virtually any database back end. As the basic structure of an SQL statement consist of SELECT keyword FROM table WHERE conditions, an ad hoc query dynamically supplies the keyword, data source and the conditions without the user knowing it.

7. **Write some of the caveats allowing users to formulate and execute adhoc queries.**
   - **Performance**. Writing efficient queries is a skill, and many queries involve joins across multiple tables that can bring a system's performance to its knees.
   - **Semantic consistency**. Allowing users to write their own queries implies they know and understand the meanings of the data elements they have selected to include in their result sets.
   - **Repeatability.** The ad hoc process involves a sequence consisting of multiple iterations of the two-phased query and review of the result set process.

**8. What is parameterized report. (NOV/DEC 2017)**

The parameterized approach is particularly beneficial in operational scenarios in which similar queries and drill-downs are done over and over again. Parameterized reports bridges the gap between static, canned reports and free flowing ad hoc queries.

**9. What is self service business intelligence.**

Self-service business intelligence (SSBI) is an approach to data analytics that enables business users to access and work with corporate data even though they do not have a background in statistical analysis, business intelligence (BI) or data mining.

**10. What is Canned reports.**

prebuilt or "canned" report. These are typically distributed across a wide audience with polished formatting. Canned reports are short and sweet, batched and scripted data for often monthly or weekly metrics. Canned reports are the Pre defined reports ,They are reports that run daily, weekly, monthly and don't change.

**11. What are the different types of visualization modes for data.**

Bar chart, line chart, Pie chart, Scatter plot, Bubble chart, Gauge, Directional indicators (arrows up or down), Heat map, Spider or radar chart and Sparkline.

**12. What is adhoc analysis and adhoc reports.**

Ad hoc analysis may be used to create a report that does not already exist, or drill deeper into a static report to get details about accounts, transactions, or records. The process may be also used to get more current data for the existing areas covered by a static report. OLAP dashboards are specifically designed to facilitate ad hoc analysis by providing quick, easy access to data from the original report. The idea of ad hoc reporting is used to talk about 'one-off' or one-time reports that are done in a customized way, to provide results for one specific question or objective.

**13. Define OLAP.**

OLAP (online analytical processing) is computer processing that enables a user to easily and selectively extract and view data from different points of view. OLAP data is stored in a multidimensional database. It can be used for datamining or the discovery of previously undiscerned relationships between data items. OLAP products are typically designed for multiple-user environments.

**14. What is OLAP cube.**

A cube can be considered a multi-dimensional generalization of a two- or three-dimensional spreadsheet. For example, a company might wish to summarize financial data by product, by time-period, and by city to compare actual and budget expenses. Product, time, city and scenario (actual and budget) are the data's dimensions.

**Slice** is the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension.

**Dice:** The dice operation produces a subcube by allowing the analyst to pick specific values of multiple dimensions

**15. What is Gauge.**

A gauge is an indicator of magnitude in the context of critical value ranges. A gauge is good for conveying relative status of critical variables and points that should trigger some action. A traditional example is an automobile's fuel gauge, which indicates the relative fullness of the tank, as well as an area close to the "empty" measure marked as red to indicate the need for refueling.

**16. What is scorecards and dashboards.**

**Score cards**: A scorecard is a data visualization tool that helps organizations, individuals, or groups of individuals to reach goals by displaying progress toward objectives against the objectives themselves. A report that gives information about the status, condition, or success of someone or something.

**Dashboards**: **Dashboards** often provide at-a-glance views of KPIs (key performance indicators) relevant to a particular objective or business process (e.g. sales, marketing, human resources, or production). In real-world terms, "dashboard" is another name for "progress report" or "report."

**17. List Some characteristics of business processes that are nicely suited to integrated analytics.**

- The business process has distinct performance objectives.
- The business process involves decision points by one or more actors.
- The process's performance can be impaired by absence of information.
- The process's performance can be impaired by ill-informed decisions.
- The process can be improved with well-informed decision-making.
- Participants do not need to be "tech-savvy" to be informed.

**18. Who are Operational analytics users?**

Who indirectly rely on the results of analytics embedded within operational applications. Examples include call center representatives whose scripts are adjusted interactively in relation to customer profiles, predicted behavioural predispositions, and real-time customer responses, web site offers and ad placement, or users of retail shelf management systems that adjust stock levels based on demand across multiple regions.

**19. What is sparkline.**

Sparklines are small line graphs without axes or coordinates. Many sparklines can be used in relative comparison regarding trends. As an example, the trends of different stock price histories for similar companies can be compared to determine of there are industry trends relating to stock price.

**20. What is integrated analysis? (APRIL/MAY 2017)**

. It implies the need for real-time integration of data from multiple sources of both analytics results and operational data. In turn, the delivery of the actionable knowledge must be seamlessly presented in a way that is best suited to business operations and is seamlessly integrated with the common suites of desktop productivity tools.

**21. Write the major issues in modeling. (APRIL/MAY 2017)**

Some models are more opaque than others; that is, it's hard to understand the logic the model used to identify relevant patterns and relationships in the data. The problem with these models is that business people often have a hard time trusting them until they see quantitative results, such as reduced costs or higher revenues. Getting business users to understand and trust the output of analytical models is perhaps the biggest challenge in data mining.

**22. What is interactive analysis? (NOV/DEC 2017)**

Interactive analysis often manifested as a pivot table. These pivot tables enable broader flexibility in grouping data within ordered hierarchies, development of static graphs and charts, or just perusing the data from different angles.

**23. Write some guidelines for laying out a BI dashboard.**

- Choose the right visualization graphic
- Manage your ‘‘real estate.’’
- Maintain context
- Be consistent
- Keep it simple

## PART B

**1. Explain in detail on business intelligence user types.**

- **Power users**: who constitute a community of experienced, sophisticated analysts who want to use complex tools and techniques to analyze data and whose results will inform decision-making processes;
- **Business users,** who rely on domain-specific reporting and analyses prepared by power users, but who also rely on their own ad hoc queries and desire access to raw data for drilling down, direct interaction with analytics servers, extraction, and then further manipulation, perhaps using desktop utility tools;
- **Casual users**, who may represent more than one area of the business, and rely on rolled-up metrics from across functions or operational areas summarized from predesigned reports presented via scorecards or dashboards;
- **Data aggregators or Information Providers**, which are businesses that collect industry- or society-wide data and enhance and reorganize that data as a way of providing value-added services to customers and subscribers. Some examples include database marketing services, financial and credit information services, real estate business information services, audience measurement services, market research providers, and national statistical agencies, among others;
- **Operational analytics users,** who indirectly rely on the results of analytics embedded within operational applications. Examples include call center representatives whose scripts are adjusted interactively in relation to customer profiles, predicted behavioral predispositions, and real-time customer responses, web site offers and ad placement, or users of retail shelf management systems that adjust stock levels based on demand across multiple regions;

- **Extended enterprise users**, comprising external parties, customers, regulators, external business analysts, partners, suppliers, or anyone with a need for reported information for tactical decision-making; and
- **IT users**, mostly involved in the development aspects of BI, and whose use of BI is more for supporting the needs of other information consumers.

**2. Explain standard report with an example. (NOV/DEC 2017)**

The most ‘‘generic’’ approach to presentation of information reflects a relatively basic, two-dimensional alignment of information, characterized within a grid of rows and columns. Standard, static reports derived from user specifications provide a consistent view of particular aspects of the business, generated in batch and typically delivered on a scheduled basis through a standard (web)interface. The columns typically articulate the item or characteristic being measured, while the rows will generally correspond to the division and hierarchies for which those measures are provided. The intersection of each row and column provides the specific measure for the column’s characteristic for the row’s item.

In this example, there are two measures (in the columns) the estimate of owner occupied housing units with a mortgage, and the margin of error associated with the measure. There are four groups of items being measured:

- The number of owner-occupied housing units with a mortgage
- Value of the houses
- Mortgage status
- Household income for the previous 12 months

Within some of these groups, there are further hierarchical breakdowns, such as the dollar groupings for value, or the categories for mortgage status. These are relatively generic categories/hierarchies, and this is reflected in the fact that these are indeed ‘‘canned’’ (or static) reports that have been already prepared for presentation.

The presumption is that the static nature of standard reports will drive the need for alternative methods for additional insight. In other words, standard reports present analytical results, but may not provide enough information for analysts seeking actionable insight unless any of the reported numbers are perceived to be beyond the bounds of expectations. And in either case, the standard report only provides a view into what was intended to be shared, but is limited in providing answers to specific business questions.

**3. Explain Interactive Analysis and Ad Hoc Querying with an example. (APRIL/MAY 2017) (NOV/DEC 2017)**

BI users looking for additional details regarding information delivered in standard reports may opt to drill into the data, either with broader visibility into the existing data or with a finer level of granularity. Both are intended to go beyond the relatively strict format of the standard report, even if they open up different views into the data.

The result sets are also suitable for loading into desktop tools for further organization and analysis, as well as forming the basis for static charts and graphs. However, there are some caveats when allowing users to formulate and execute adhoc queries; here are some:

- **Performance.** Writing efficient queries is a skill, and many queries involve joins across multiple tables that can bring a system’s performance to its knees. The users would be expected to be highly trained before letting many loose in writing their own queries.
- **Semantic consistency.** Allowing users to write their own queries implies they know and understand the meanings of the data elements they have selected to include in their result sets. However, without comprehensive, standardized business term glossaries and metadata repositories, users may see data element names and impute their definitions, potentially assigning meanings that are different than what was intended by the data creators. These discrepancies may impact believability of the results.
- **Repeatability.** The ad hoc process involves a sequence consisting of multiple iterations of the two-phased query and review of the result set process. The operational process allows the analyst to effectively follow a thread or a train of thought, but without a means for capturing the thought processes driving the sequence, it is difficult to capture the intuition that drives the ultimate result.

In other words, the sequence may yield some results, but it may be difficult to replicate that process a second or third time. Standard reports can provide knowledge to a broad spectrum of consumers, even if those consumers must have contextual knowledge to identify the key indicators and take action. Ad hoc queries enable greater drill-down and potential for insight.owever, given the growth of data into the petabytes coupled with the complexity and performance impacts of ad hoc queries, standard reporting is rapidly yielding to more organized methods for delivering results, through parameterized reporting, dimensional analysis, and notification, alerts, and exception reporting.

**4. Explain in detail on Parameterized Reports and Self-Service Reporting. (APRIL/MAY 2017) (NOV/DEC 2017)**

After monitoring the types of ad hoc queries performed, it became apparent in many scenarios that users within similar categories were executing very similar queries. The problem was that despite the similarity of the queries, each was being executed in isolation, with each contributing to degradation of overall performance. Parameterized reports provide one approach to self-service business intelligence, or "self-service BI."In a self-service BI framework, tools simplify the different aspects of generating results and reports, including simplifying. The data discovery process is done by presenting a palette of data sets that the user can access and use. The data access methods by masking or virtualizing access to the data to be queried. The documentation of the ''make-up'' of the report via collaborative means so that the results, and more importantly, the process for generating the results, can be shared with other analysts, The development of the presentation layer, whether that is simple row/column reports, or using more sophisticated visualization techniques.

Another benefit of self-service BI is that it is intended to reduce or eliminate the IT bottleneck. In many environments, the IT department is responsible for developing reports, and as the BI program gains more acceptances, there will be greater demand for IT resources for report development. This becomes a bottleneck when the time for responding to a request exceeds the window of opportunity for exploiting the actionable knowledge. For example, one group might want to evaluate how a product performs within its first week of release so that adjustments and tweaks can be made; if it takes three weeks for the report to be readied, it is already too late to take action.

**5. Explain Alerts/Notification method.**

When you think about the ways that individuals usually review the data in the standard report's layout, you will recognize that in many cases, the individual's attention is only focused on one or two key pieces of information. In these situations, the individual's goal is examining some specific variable's value, and either verifying that the value is within an expected range, or determining that the value is outside the expected range and then taking some action.For example, a national call center manager might review average hold times by regional call center.

As long as the average hold time is between 30 and 60 seconds, the averages remain within the acceptable level of service. However, once an average hold time for any region exceeds 60 seconds, the call center manager will need to reach out to the regional call center manager to investigate why the hold times are longer than expected.Of course, you can envision many similar scenarios in which the action needs to be triggered only when certain variables hit specific values. And in each of these cases, reviewing the entire report is overkill the business user only needs to know the specific variable's value, and only when that value would need to trigger an action; otherwise, the variable's value can be ignored. This realization means that instead of presenting an entire report, alerts or notifications can be an alternative method for delivering actionable knowledge.

This method is nicely suited to operational environments in which notifications can be delivered via different methods. Some examples include email, instant messages, direct messages delivered through (potentially internal) social networking sites, smartphones, other mobile devices, radio transmissions, or even visual cues (such as scrolling message boards, light banks, or visual consoles). In these situations, the notification method can embody the context; for example, a flashing amber light provides the medium for notification as well as the message. This approach not only simplifies the delivery of the critical piece of information, it reduces the effort for inspecting the critical value and thereby enables actions to be taken in a rapid manner.

**6. Explain some of the visualization modes for data.**

There are many different types of visualization modes for data, and while this is not intended to provide a comprehensive overview of visualization techniques, it is meant to provide an overview of a handful of ways to present actionable knowledge:

- **Line chart.** A line chart maps points on a grid connected by line segments. A line chart can be used to show a series of connected values, such as a time series. An example would be mapping the rise and fall of gas prices per gallon using the price of a gallon of gas on the first day of each month for the previous 36 months.
- **Bar chart**. A bar chart maps values using rectangles whose lengths correspond to the charted values. Bar charts are good fo for comparing different values of the same variable across different contexts. An example would be a chart of the average life expectancy in years across different countries.
- **Pie chart.** A pie chart is conveyed as a circle that is broken out into sectors representing some percentage of a whole. A pie chart is good for showing distributions of values across a single domain. An example is showing the relative percentages of owner-occupied homes by ethnicity within a Zip code area. The total of all the components always will add up to 100%, and each slice of the pie represents a percentage of the whole.
- **Scatter plot**. A scatter plot graphs points showing a relationship between two variables. Typically one variable is fixed (the dependent variable) and the other is not (the independent variable). In a two-dimensional scatter plot, the x axis represents the independent variable value and the y axis represents the dependent variable. A scatter plot is used to look for correlation between the dependent and independent variable. An example graphs an individual's age (the dependent variable) and the individual's observed weight (the independent variable).
- **Bubble chart.** A bubble chart is a variation on a scatter plot in which a third variable can be represented using the size of the item in the chart. An example would graph the dollar sales volume by the number of items sold, and the bubbles could represent the percentage of the overall market share.

**7. (i) Explain Scorecards and Dashboards with example. (APRIL/MAY 2017) (NOV/DEC 2017)**

Scorecards and dashboards are two different approaches for consolidating the presentation of reported results to a particular user type. A scorecard usually presents the values of key performance indicators as well as indicators reflecting whether those KPI values are acceptable or not. The scorecard presentation may also be enhanced with historical trends and indications if the KPIs have been improving or not over time. Scorecards are often updated on a periodic basis

Dashboards provide some degree of flexibility to the user in crafting the presentation that is most relevant to the way he/she operates. Given an inventory of presentation graphics (such as those described in the previous section), an analyst and business user can work together in selecting the most appropriate methods of presentation. Dashboards can connect to real-time sources, and allow the business data consumer to customize an up-to-date presentation of summarized performance metrics, allowing continuous monitoring throughout the day.Pervasive delivery mechanisms can push dashboards to a large variety of channels, ranging from the traditional browser-based format to handheld mobile devices. Through the interactive nature of the dashboard, the knowledge worker can drill down through the key indicators regarding any emerging opportunities, as well as take action through integrated process-flow and communication engines.

Another approach to dashboards is the concept of a mashup, which allows the knowledge consumers themselves the ability to identify their own combination of analytics and reports with external data streams, news feeds, social networks, and otherWeb 2.0 resources in a visualization framework that specifically suits their own business needs and objectives. The mashup framework provides the ''glue'' for integrating data streams and BI with interactive business applications.

**(ii)Explain Geographic visualization in detail. (APRIL/MAY 2017) (NOV/DEC 2017)**

Aspects of location intelligence and spatial analytics, and the results of that type of analysis can be presented within the context of a map. Instead of using the standard graphical widgets described in a previous section, aggregate values and totals can be attributed to a visual representation of a map.

For example, population statistics for each country in the European Union can be superimposed on top of a map of Europe. These maps can satisfy the desire to drill down; interactive selection or clicking on one segment of the map can zoom in from a geographic standpoint. In addition, spatial analysis results can be layered within the mapping interface. For example, in an insurance management application, hazard zones can be superimposed on top of regions potentially affected by weather events to help guide determination of heightened risk areas that will command additional insurance premiums in a way that balances risk across the company's customer base. Another example employs the heat map concept to geographic regions using sizes and colors to present a collection of variable values.

Often dashboards will link more than one visualization component to others, and this can be easily applied to geographic visualization. For example, a dimensional analysis presentation (such as a pivot table) for a geographic hierarchy can be presented in one frame while the aggregated values are displayed within a map. Realigning the dimensions in the grid will automatically update the map, and drilling through regions on the map will trigger a reload of the drilled-through data within the grid.

**8. Explain in detail about Integrated Analytics.**

**S**pecific analytic values trigger specific actions within well-defined business processes, employing a combination of alerts and notifications with visualization tools reduces the need for the end users to have deep training in the use of the BI tool set. In other words, when the analytical results are fed directly into operational activities, the end-user may be a consumer of BI and may not even be aware of it.

Some characteristics of business processes that are nicely suited to integrated analytics include:

- The business process has distinct performance objectives.
- The business process involves decision points by one or more actors.
- The process's performance can be impaired by absence of information.
- The process's performance can be impaired by ill-informed decisions.
- The process can be improved with well-informed decision-making.
- Participants do not need to be ''tech-savvy'' to be informed.

Yet in order to make integrated analytics work, the implementers must make sure that all the information necessary can be delivered to the appropriate person with the right time frame to facilitate the best decisions. That might imply the need for real time integration of data from multiple sources of both analytics results and operational data.

In turn, the delivery of the actionable knowledge must be seamlessly presented in a way that is best suited to business operations and is seamlessly integrated with the common suites of desktop productivity tools. This type of event driven notification allows analytics to be directly embedded within operational processes and supporting applications across multiple channels. As this approach gains popularity, more widespread adoption of BI services coupled with lowered barriers to deployment will open new opportunities for integrating BI results.

**9. How to optimize the presentation of the right message?**

Here are some quick guidelines to keep in mind when laying out a BI dashboard:

- **Choose the right visualization graphic.** Don't let the shiny graphics fool you into using a visual component that does not properly convey the intended result. For example, line charts are good for depicting historical trends of the same variable over time, but bar charts may not be as good a choice.
- **Manage your ''real estate.''** The available screen space limits what can be displayed at one time, and this is what is referred to as screen ''real estate.'' Different delivery channels allow different amounts of real estate. A regular desktop screen affords more presentation area than the screen on a laptop, while both trump the screen of a portable tablet and especially a smartphone. When considering the channel and the consumer, employ the right visualization components that fit within the available space yet still deliver the actionable knowledge.

- **Maintain context.** You must recognize that the presentation of a value is subject to variant interpretations when there is no external context defining its meaning. For example, presenting a value on a dial-gauge conveys the variable's magnitude, but not whether that value is good, bad, or indifferent. Adjusting the dial gauge with a red zone (to indicate a bad value) and a green zone (to indicate a good value) provides the context of the displayed magnitude.
- **Be consistent**. When the power of self-service dashboard development is placed in the hands of many data consumers, their own biases will lead to an explosion of variant ways of representing the same or similar ideas. The result is that what makes sense to one data consumer has less meaning when presented to another, and the confusion only grows with wider dissemination. Consistent representations and presentations (and corresponding selection of standard visualization graphics) will help to ensure consistent interpretations.
- **Keep it simple**. Don't inundate the presentation with fancy-looking graphics that don't add to the decision-making process. Often the simpler the presentation, the more easily the content is conveyed.
- **Engage**. Engage the user community and agree on standards, practices, and a guide book for developing visualization parameters for delivery and presentation.

**10. Explain OLAP with slice and dice example.**

OLAP is an acronym for Online Analytical Processing. OLAP performs multidimensional analysis of business data and provides the capability for complex calculations, trend analysis, and sophisticated data modeling. Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data. Here is the list of OLAP operations:

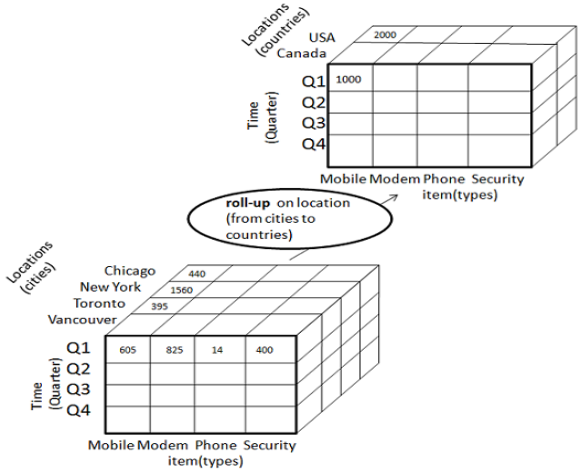
- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

**Roll-up**

Roll-up performs aggregation on a data cube in any of the following ways:

- By climbing up a concept hierarchy for a dimension
- By dimension reduction

The following diagram illustrates how roll-up works.



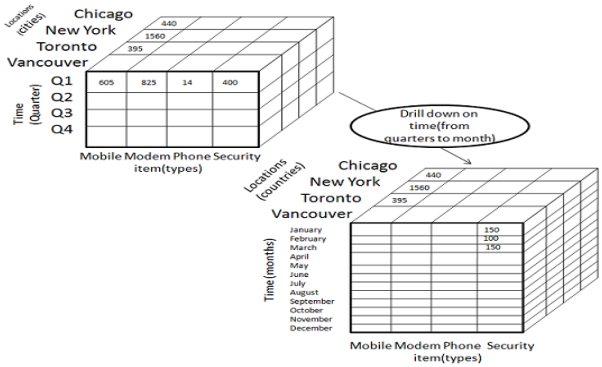

- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

**Drill-down**

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways:

- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.

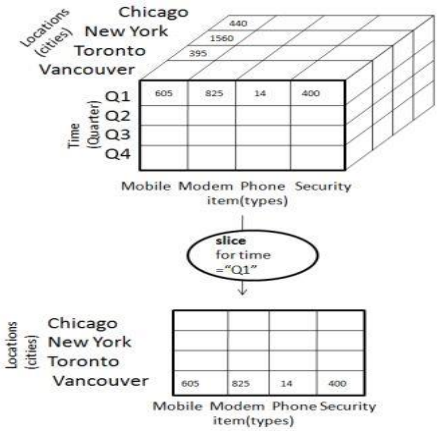
The following diagram illustrates how drill-down works:




- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

**Slice**

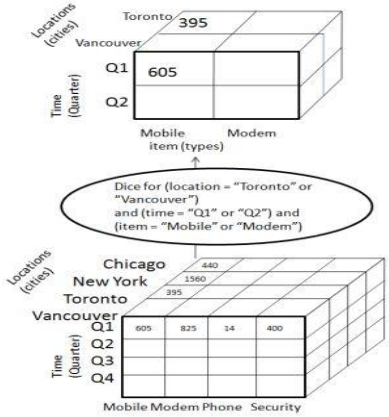
The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.




- Here Slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

**Dice**

Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.




The dice operation on the cube based on the following selection criteria involves three dimensions.

(location = "Toronto" or "Vancouver") (time = "Q1" or "Q2") (item =" Mobile" or "Modem")

**Pivot**

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation. In this the item and location axes in 2-D slice are rotated.

**UNIT III EFFICIENCY**

**Efficiency measures – The CCR model: Definition of target objectives- Peer groups – Identification of good operating practices; cross efficiency analysis – virtual inputs and outputs – Other models. Pattern matching – cluster analysis, outlier analysis.**

## PART-A

**1. What is data envelopment analysis?**

Data envelopment analysis is to compare the operating performance of a set of units such as companies, university departments, hospitals, bank branch offices, production plants, or transportation systems. In order for the comparison to be meaningful, the units being investigated must be homogeneous.

**2. What do you mean by productivity indicator?**

Data envelopment analysis relies on a productivity indicator that provides a measure of the efficiency that characterizes the operating activity of the units being compared. This measure is based on the results obtained by each unit, which will be referred to as outputs, and on the resources utilized to achieve these results, which will be generically designated as inputs or production factors.

**3. What are decision making units (DMU)?**

In DEA the units being compared are called decision making units (DMUs),as they enjoy a certain decisional autonomy.

**4. How to find the efficiency of a decision making unit?**

Lets evaluate the efficiency of $n$ units, where $N = \{1, 2, \ldots, n\}$ denote the set of units being compared.

If the units produce a single output using a single input only, the *efficiency* of the $j$ th decision making unit DMU$j$ , $j \in N$, is defined as $\theta j = yj/xj$ , in which $yj$ is the output value produced by DMU$j$ and $xj$ the input value used.

**5. How to find the efficiency of a DMU with multiple outputs?**

If the units produce multiple outputs using various input factors, the efficiency of DMU$j$ is defined as the ratio between a weighted sum of the outputs and a weighted sum of the inputs. Denote by $H = \{1, 2, \ldots, s\}$ the set of production factors and by $K = \{1, 2, \ldots, m\}$ the corresponding set of outputs. If $xij$ , $i \in H$, denotes the quantity of input $i$ used by DMU$j$ and $yrj$ , $r \in K$, the quantity of output $r$ obtained,

The efficiency of DMU$j$ is defined as $\theta j = (u1y1j + u2y2j + \ldots + umymj)/(v1x1j + v2x2j + \ldots + vsxsj) = (\Sigma r \in K\ uryrj)/(\Sigma i \in H\ vixij)$, for weights $u1, u2, \ldots, um$ associated with the outputs and $v1, v2, \ldots, vs$ assigned to the inputs.

**6. What is efficient frontier?**

The efficient frontier, also known as production function, expresses the relationship between the inputs utilized and the outputs produced. It indicates the maximum quantity of outputs that can be obtained from a given combination of inputs. At the same time, it also expresses the minimum quantity of inputs that must be used to achieve a given output level. Hence, the efficient frontier corresponds to technically efficient operating methods

**7. What are called the production possibility sets?**

Slope of the line connecting each point to the origin represents the efficiency value associated with the corresponding branch. The line with the maximum slope, represented by a solid line, is the efficient frontier for all branches being analyzed. The branches that are on this line correspond to efficient units, while the branches that are below the efficient frontier are inefficient units. The area between the efficient frontier and the positive horizontal semi-axis is called the production possibility set.

**8. What is input oriented efficiency and output oriented efficiency.**

Efficient frontier provides some indications for improving the performance of inefficient units. Indeed, it identifies for each input level the output level that can be achieved in conditions of efficiency. By the same token, it identifies for each output level the minimum level of input that should be used in conditions of efficiency. (i.e)for each DMU$j$ , $j \in N$,

(i) Input oriented efficiency $\theta_{jI}$ can be defined as the ratio between the ideal input quantity $x*$ that should be used by the unit if it were efficient and the actually used quantity $xj$ : $\theta Ij = x*/x_j$

(ii) Similarly, the output oriented efficiency $\theta^{Oj}$ is defined as the ratio between the quantity of output $yj$ actually produced by the unit and the ideal quantity $y*$ that it should produce in conditions of efficiency: $\theta Oj = y*/ y_j$

**9. How to convert inefficient unit to efficient?**

Inefficient unit can be brought close to the efficient frontier. In this case, the inefficiency of a given unit is evaluated by the length of the segment connecting the unit to the efficient frontier along the line passing through the origin of the axes.

$$\theta_A = \frac{\overline{OP}}{\overline{OA}},$$

The inefficient unit may be made efficient by a displacement along segment *OA* that moves it onto the efficient frontier. Such displacement is tantamount to progressively decreasing the quantity of both inputs while keeping unchanged the quantity of output.

**10. What is CCR model ?**

Using data envelopment analysis, the choice of the optimal system of weights for a generic DMU*j* involves solving a mathematical optimization model whose decision variables are represented by the weights $ur$ , $r \in K$, and $vi$ , $i \in H$, associated with each output and input. Various formulations have been proposed, the best-known of which is probably the Charnes-Cooper-Rhodes (CCR) model. The CCR model formulated for DMU*j* takes the form

$$\begin{aligned} \max \quad & \vartheta = \sum_{r \in \mathcal{K}} u_r y_{rj}, \\ \text{s.to} \quad & \sum_{i \in \mathcal{H}} v_i x_{ij} = 1, \\ & \sum_{r \in \mathcal{K}} u_r y_{rj} - \sum_{i \in \mathcal{H}} v_i x_{ij} \le 0, \qquad j \in \mathcal{N}, \end{aligned}$$

**11. What is the formulation taken by the CCR model for decision making units?**

(i) The objective function involves the maximization of the efficiency measure for DMU*j*

(ii)Constraints require that the efficiency values of all the units, calculated by means of the weights system for the unit being examined, be lower than one.

(iii)Conditions guarantee that the weights associated with the inputs and the outputs are non-negative.

**12. What are target objectives?**

Based on the efficiency values, data envelopment analysis therefore gives a measure for each unit being compared of the savings in inputs or the increases in outputs required for the unit to become efficient. To determine the target values, it is possible to follow an input- or output oriented strategy.

**13. How to determine the target values?**

Target values can be determined in two cases.

(i) In the first case, the improvement objectives primarily concern the resources used, and the target values for inputs and outputs are given by

$$\begin{aligned} x_{ij}^{\text{target}} &= \vartheta^* x_{ij} - s_i^{-*}, \quad i \in \mathcal{H}, \\ y_{rj}^{\text{target}} &= y_{rj} + s_r^{+*}, \quad r \in \mathcal{K}. \end{aligned}$$

(ii) In the second case, target values for inputs and outputs are given by

$$\begin{aligned} x_{ij}^{\text{target}} &= x_{ij} - \frac{s_i^{-*}}{\vartheta^*}, \quad i \in \mathcal{H}, \\ y_{rj}^{\text{target}} &= \frac{y_{rj} + s_r^{+*}}{\vartheta^*}, \quad r \in \mathcal{K}. \end{aligned}$$

**14. What are the other performance improvement strategies?**

The target values for the inputs are set in such a way as to minimize the quantity used of the resources to which the highest priority has been assigned, without allowing variations in the level of other inputs or in the outputs produced; The target values for the outputs are set in such a way as to maximize the quantity produced of the outputs to which highest priority has been assigned, without allowing variations in the level of other outputs or inputs used; Preferences expressed by the decision makers with respect to a decrease in some inputs or an increase in specific outputs.

**15. What are peer groups in DEA?**

DEA identifies for each inefficient unit a set of excellent units, called a peer group, which includes those units that are efficient if evaluated with the optimal system of weights of an inefficient unit. The peer group, made up of DMUs which are characterized by operating methods similar to the inefficient unit being examined, is a realistic term of comparison which the unit should aim to imitate in order to improve its performance.

**16. Define outlier analysis?**

The data objects which do not comply the general behavior or model of the data, then such data are grossly different or inconsistent with the remaining set of data called outliers. Thus the outlier detection is a data mining task known as outlier analysis.

**17. What is the use of weight restrictions?**

The units that are really efficient are separated from those whose efficiency score largely depends on the selected weights system, we may impose some restrictions on the value of the weights to be associated with inputs and outputs. In general, these restrictions translate into the definition of maximum thresholds for the weight of specific outputs or minimum thresholds for the weight of specific inputs.

**18. What is cross efficiency analysis? (NOV/DEC 2017)**

The analysis of cross efficiency is based on the definition of the efficiency matrix which provides information on the nature of the weights system adopted by the units for their own efficiency evaluation. The square efficiency matrix contains as many rows and columns as there are units being compared. The generic element $\theta ij$ of the matrix represents the efficiency of DMU$j$ evaluated through the optimal weights structure for DMU$i$ , while the element $\theta jj$ provides the efficiency of DMU$j$ calculated using its own optimal weights.

**19. Define cluster analysis?**

Cluster analysis groups data objects based only on information found in the data that describes the object ts and their relationships. The goal is that the objects within a group be similar(or related) to one another and different from the objects in other groups. The greater the similarity within a group and greater the difference between the groups makes the better, or more distinct the clustering.

**20. What are virtual inputs and outputs in DEA? (NOV/DEC 2017)**

Virtual inputs and virtual outputs provide information on the relative importance that each unit attributes to each individual input and output, for the purpose of maximizing its own efficiency score. The virtual inputs of a DMU are defined as the product of the inputs used by the unit and the corresponding optimal weights. Similarly, virtual outputs are given by the product of the outputs of the unit and the associated optimal weights

**21. Define linear regression. (APRIL/MAY 2017)**

Linear regression is a statistical technique for analyzing data in order to obtain a measure of correlation between two variables were the relationship between the variables is expected to be linear. Regression is mostly used in predictive analysis.

**22. What is link analysis? (APRIL/MAY 2017)**

Link analysis is based on a branch of mathematics called graph theory, which represents relationships between different objects as edges in a graph. Link analysis is not a specific modeling technique, so it can be used for both directed and undirected data mining. It is often used for creating new derived variables for use by other modeling techniques. It can also be used for undirected data mining, by exploring the properties of the graphs themselves.

## PART-B

**1. Explain in detail about Efficiency measures and Efficient frontier.**

The purpose of data envelopment analysis (DEA) is to compare the operating performance of a set of units such as companies, university departments, hospitals, bank branch offices, production plants, or transportation systems. In order for the comparison to be meaningful, the units being investigated must be homogeneous. The performance of a unit can be measured on several dimensions. For example, to evaluate the activity of a production plant one may use quality indicators, which estimate the rate of rejects resulting from manufacturing a set of products, and also flexibility indicators, which measure the ability of a system to react to changes in the requirements with quick response times and low costs. Data envelopment analysis relies on a productivity indicator that provides a measure of the efficiency that characterizes the operating activity of the units being compared.

This measure is based on the results obtained by each unit, which will be referred to as outputs, and on the resources utilized to achieve these results, which will be generically designated as inputs or production factors. If the units represent bank branches, the outputs may consist of the number of active bank accounts, checks cashed or loans raised; the inputs may be the number of cashiers, managers or rooms used at each branch. If the units are university departments, it is possible to consider as outputs the number of active teaching courses and scientific publications produced by the members of each department; the inputs may include the amount of financing received by each department, the cost of teaching, the administrative staff and the availability of offices and laboratories.

**EFFICIENCY MEASURES**

In data envelopment analysis the units being compared are called *decision making units* (DMUs), since they enjoy a certain decisional autonomy. Assuming that we wish to evaluate the efficiency of $n$ units, let $N = \{1, 2, \ldots, n\}$ denote the set of units being compared.

(i) If the units produce a single output using a single input only, the efficiency of the $j$ th decision making unit DMU$j$ , $j \in N$, is defined in which $yj$ is the output value produced by DMU$j$ and $xj$ the input value

$$\theta j = yj/xj$$

(ii) If the units produce multiple outputs using various input factors, the efficiency of DMU$j$ is defined as the ratio between a weighted sum of the outputs and a weighted sum of the inputs. Denote by $H = \{1, 2, \ldots, s\}$ the set of production factors and by $K = \{1, 2, \ldots, m\}$ the corresponding set of outputs. If $xij$ , $i \in H$, denotes the quantity of input $i$ used by DMU$j$ and $yrj$ , $r \in K$, the quantity of output $r$ obtained, the efficiency of DMU$j$ is defined as

$$\theta j = (u1y1j + u2y2j + \ldots + umymj)/(v1x1j + v2x2j + \ldots + vsxsj) = (\Sigma r \in K\ uryrj)/(\Sigma i \in H\ vixij)$$

for weights $u1, u2, \ldots, um$ associated with the outputs and $v1, v2, \ldots, vs$ assigned to the inputs. In this second case, the efficiency of DMU$j$ depends strongly on the system of weights introduced. At different weights, the efficiency value may undergo relevant variations and it becomes difficult to fix a single structure of weights that might be shared and accepted by all the evaluated units. In order to avoid possible objections raised by the units to a preset system of weights, which may privilege certain DMUs rather than others, data envelopment analysis evaluates the efficiency of each unit through the weights system that is best for the DMU itself - that is, the system that allows its efficiency value to be maximized. Subsequently, by means of additional analyses, the purpose of data envelopment analysis is to identify the units that are efficient in absolute terms and those whose efficiency value depends largely on the system of weights adopted.

**EFFICIENT FRONTIER**

The efficient frontier, also known a production function, expresses the relationship between the inputs utilized and the outputs produced. It indicates the maximum quantity of outputs that can be obtained from a given combination of inputs. At the same time, it also expresses the minimum quantity of inputs that must be used to achieve a given output level. Hence, the efficient frontier corresponds to technically efficient, operating methods. The efficient frontier may be empirically obtained based on a set of observations that express the output level obtained by applying a specific combination of input production factors. In the context of data envelopment analysis, the observations correspond to the units being evaluated.

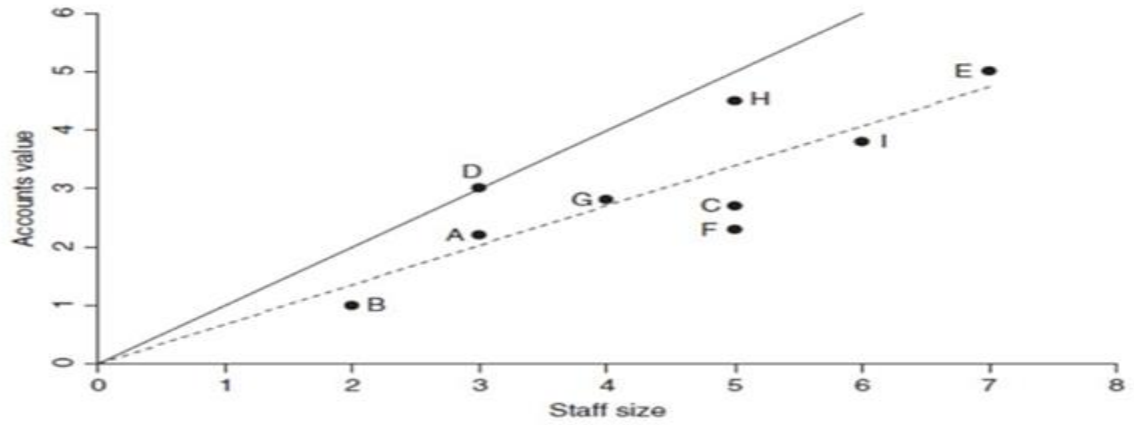
Most statistical methods of parametric nature, which are based for instance on the calculation of a regression curve, formulate some prior hypotheses on the shape of the production function. Data envelopment analysis, on the other hand, forgoes any assumptions on the functional form of the efficient frontier, and is therefore nonparametric in character. It only requires that the units being compared are not placed above the production function, depending on their efficiency value. To further clarify the notion of efficient frontier considers Example 1.

**Example 1 - Evaluation of the efficiency of bank branches.** A bank wishes to compare the operational efficiency of its nine branches, in terms of staff size and total value of savings in active accounts. Table 1 shows for each branch the total value of accounts, expressed in hundreds of thousands of euros, and the number of staff employed, with the corresponding efficiency values calculated based on definition. The graph shown in Figure 1 shows for each branch the number of employees on the horizontal axis and the value of accounts on the vertical axis. The slope of the line connecting each point to the origin represents the efficiency value associated with the corresponding branch.

The line with the maximum slope, represented in Figure 1 by a solid line, is the efficient frontier for all branches being analyzed. The branches that are on this line correspond to efficient units, while the branches that are below the efficient frontier are inefficient units. The area between the efficient frontier and the positive horizontal semi-axis is called the production possibility set.A possible alternative to the efficient frontier is the regression line that can be obtained based on the available observations, indicated in Figure 1 by a dashed line.

In this case, the units that fall above the regression line may be deemed excellent, and the degree of excellence of each unit could be expressed by its distance from the line. However, it is appropriate to underline the difference that exists between the prediction line obtained using a regression model and the efficient frontier obtained using data envelopment analysis. The Table 1 Input and output values for the bank branches in Example 1 bank branch staff size accounts value efficiency regression line reflects the average behavior of the units being compared,

| bank branch | staff size | accounts value | efficiency |
|---|---|---|---|
| A | 3 | 2.5 | 0.733 |
| B | 2 | 1.0 | 0.500 |
| C | 5 | 2.7 | 0.540 |
| D | 3 | 3.0 | 1.000 |
| E | 7 | 5.0 | 0.714 |
| F | 5 | 2.3 | 0.460 |
| G | 4 | 3.2 | 0.700 |
| H | 5 | 4.5 | 0.900 |
| I | 6 | 4.5 | 0.633 |




while the efficient frontier identifies the best behavior, and measures the inefficiency of a unit based on the distance from the frontier itself. Notice also that the efficient frontier provides some indications for improving the performance of inefficient units. Indeed, it identifies for each input level the output level that can be achieved in conditions of efficiency. By the same token, it identifies for each output level the minimum level of input that should be used in conditions of efficiency. In particular, for each DMU$j$ , $j \in N$, the *input-oriented* efficiency $\theta_{jI}$ can be defined as the ratio between the ideal input quantity $x*$ that should be used by the unit if it were efficient and the actually used quantity $xj$ :

$$\theta Ij = x*/x_j$$

Similarly, the *output-oriented* efficiency $\theta^{Oj}$ is defined as the ratio between the quantity of output $yj$ actually produced by the unit and the ideal quantity $y*$ that it should produce in conditions of efficiency:

$$\theta Oj = y*/ y_j$$

The problem of making an inefficient unit efficient is then turned into one of devising a way by which the inefficient unit can be brought close to the efficient frontier. If the unit produces a single output only by using two inputs, the efficient frontier assumes the shape shown in Figure 2. In this case, the inefficiency of a given unit is evaluated by the length of the segment connecting the unit to the efficient frontier along the line passing through the origin of the axes. For the example illustrated in Figure 2, the efficiency value of DMU$A$ is given by

$$\theta_A = \frac{\overline{OP}}{\overline{OA}},$$

where *OP* and *OA* represent the lengths of segments *OP* and *OA*, respectively. The inefficient unit may be made efficient by a displacement along segment *OA* that moves it onto the efficient frontier. Such displacement is tantamount to progressively decreasing the quantity of both inputs while keeping unchanged the quantity of output. In this case, the production possibility set is defined as the region delimited by the efficient frontier where the observed units being compared are found.

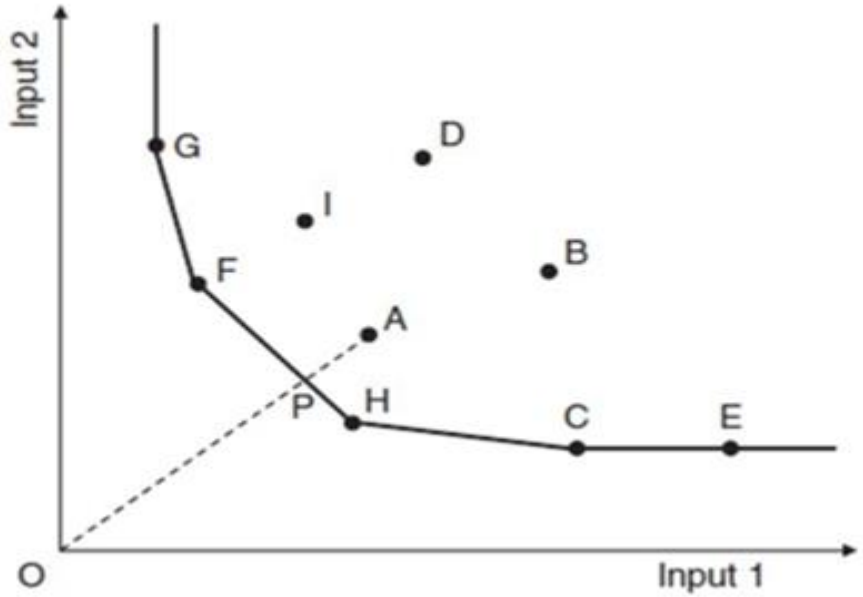



**Figure 2: Efficient frontier with two inputs and one output**

$$\max \quad \vartheta = \frac{\sum_{r \in \mathcal{K}} u_r y_{rj}}{\sum_{i \in \mathcal{H}} v_i x_{ij}},$$

$$\text{s.to} \quad \frac{\sum_{r \in \mathcal{K}} u_r y_{rj}}{\sum_{i \in \mathcal{H}} v_i x_{ij}} \leq 1, \qquad j \in \mathcal{N},$$

$$u_r, v_i \geq 0, \qquad r \in \mathcal{K}, i \in \mathcal{H}.$$

**2. Explain the CCR model in detail. (APRIL/MAY 2017)**

Using data envelopment analysis, the choice of the optimal system of weights for a generic DMU*j* involves solving a mathematical optimization model whose decision variables are represented by the weights *ur* , $r \in K$, and *vi* , $i \in H$, associated with each output and input. Various formulations have been proposed, the best-known of which is probably the Charnes-Cooper-Rhodes (CCR) model. The CCR model formulated for DMU*j* takes the form

(i) The objective function involves the maximization of the efficiency measure for DMU*j*

(ii) Constraints require that the efficiency values of all the units, calculated by means of the weights system for the unit being examined, be lower than one.

(iii) Finally, conditions guarantee that the weights associated with the inputs and the outputs are non-negative.

In place of these conditions, sometimes the constraints $ur, vi \geq \delta$, $r \in K$, $i \in H$ may be applied, where $\delta > 0$, preventing the unit from assigning a null weight to an input or output. Model can be linearized by requiring the weighted sum of the inputs to take a constant value, for example 1. This condition leads to an alternative optimization problem, the input oriented CCR model, where the objective function consists of the maximization of the weighted sum of the outputs

$$\max \quad \vartheta = \sum_{r \in \mathcal{K}} u_r y_{rj},$$

$$\text{s.to} \quad \sum_{i \in \mathcal{H}} v_i x_{ij} = 1,$$

$$\sum_{r \in \mathcal{K}} u_r y_{rj} - \sum_{i \in \mathcal{H}} v_i x_{ij} \leq 0, \qquad j \in \mathcal{N},$$

$ur, vi \geq 0, r \in K, i \in H$

Let $\vartheta*$ be the optimum value of the objective function corresponding to the optimal solution $(\mathbf{v}*, \mathbf{u}*)$ of problem. DMU$j$ is said to be efficient if $\vartheta* = 1$ and if there exists at least one optimal solution $(\mathbf{v}*, \mathbf{u}*)$ such that $\mathbf{v}* > \mathbf{0}$ and $\mathbf{u}* > \mathbf{0}$. By solving a similar optimization model for each of the $n$ units being compared, one obtains systems of weights. The flexibility enjoyed by the units in choosing the weights represents an undisputed advantage, in that if a unit turns out to be inefficient based on the most favorable system of weights, its inefficiency cannot be traced back to an inappropriate evaluation process. However, given a unit that scores $\vartheta* = 1$, it is important to determine whether its efficiency value should be attributed to an actual high-level performance or simply to an optimal selection of the weights structure. For the input-oriented CCR model, the following dual problem, which lends itself to an interesting interpretation, can be formulated:

$$\begin{aligned}
\min \quad & \vartheta, \\
\text{s.to} \quad & \sum_{j \in \mathcal{N}} \lambda_j x_{ij} - \vartheta x_{ij} \leq 0, \qquad i \in \mathcal{H}, \\
& \sum_{j \in \mathcal{N}} \lambda_j y_{rj} - y_{rj} \geq 0, \qquad r \in \mathcal{K}, \\
& \lambda_j \geq 0, \qquad j \in \mathcal{N}.
\end{aligned}$$

Based on the optimum value of the variables $\lambda*_j, j \in N,$

(i) The aim of model is to identify an ideal unit that lies on the efficient frontier and represents a term of comparison for DMU$j$ .

(ii) Constraints and of the model require that this unit produces an output at least equal to the output produced by DMU$j$ , and uses a quantity of inputs equal to a fraction of the quantity used by the unit examined.

(iii) The ratio between the input used by the ideal unit and the input absorbed by DMU$j$ is defined as the optimum value $\vartheta*$ of the dual variable $\vartheta$. If $\vartheta* < 1$, DMU$j$ lies below the efficient frontier. In order to be efficient, this unit should employ $\vartheta* xij$ , $i \in H$, of each input.

The quantity of inputs utilized by the ideal unit and the level of outputs to be produced are expressed as a linear combination of the inputs and outputs associated with the $n$ units being evaluated:

$$x_i^{\text{ideal}} = \sum_{j \in \mathcal{N}} \lambda_j^* x_{ij}, \qquad i \in \mathcal{H},$$

$$y_r^{\text{ideal}} = \sum_{j \in \mathcal{N}} \lambda_j^* y_{rj}, \qquad r \in \mathcal{K}.$$

For each feasible solution $(\vartheta, \lambda)$ to problem, the slack variables $si$ ,$i \in H$, and $s+r$ , $r \in K$, can be defined, which represent respectively the quantity of input $i$ used in excess by DMU$j$ and the quantity of output $r$ produced in shortage by the DMU$j$ with respect to the ideal unit:

$$s_i^- = \vartheta x_{ij} - \sum_{j \in \mathcal{N}} \lambda_j x_{ij}, \quad i \in \mathcal{H},$$

$$s_r^+ = \sum_{j \in \mathcal{N}} \lambda_j y_{rj} - y_{rj}, \quad r \in \mathcal{K}.$$

As with the primal problem, it is possible also for the dual problem to provide a definition of efficiency. DMU$j$ is efficient if $\vartheta* = 1$ and if the optimum value of the slack variables is equal to zero $s-*i = 0$, $i \in H$, and $s+*r = 0$, $r \in K$.

In other words, DMU$j$ is efficient if it is not possible to improve the level of an input used or the level of an output produced without deterioration in the level of another input or of another output. If $\vartheta* < 1$, DMU$j$ is said to be technically inefficient, in the sense that, in order to obtain the same output, the input quantities used could be simultaneously reduced in the same proportion. The maximum reduction allowed by the efficient frontier is defined by the value $1 - \vartheta*$. If $\vartheta* = 1$, but some slack variables are different from zero, DMU$j$ presents a mix inefficiency since, keeping the same output level, it could reduce the use of a few inputs without causing an increase in the quantity of other production factors used.

**3. Discuss in detail about the methods of identifying good operating practices.**

By identifying and sharing, good operating practices one may hope to achieve an improvement in the performance of all units being compared. The units that appear efficient according to data envelopment analysis certainly represent terms of comparison and examples to be imitated for the other units. However, among efficient units some more than others may represent a target to be reached in improving the efficiency.

The need to identify the efficient units, for the purpose of defining the best operating practices, stems from the principle itself on which data envelopment analysis is grounded, since it allows each unit to evaluate its own degree of efficiency by choosing the most advantageous structure of weights for inputs and outputs. In this way, a unit might appear efficient by purposely attributing a non-negligible weight only to a limited subset of inputs and outputs. Furthermore, those inputs and outputs that receive greater weights may be less critical than other factors more intimately connected to the primary activity performed by the units being analyzed. In order to identify good operating practices, it is therefore expedient to detect the units that are really efficient, that is, those units whose efficiency score does not primarily depend on the system of weights selected. To differentiate these units, we may resort to a combination of different methods: cross efficiency analysis, evaluation of virtual inputs and virtual outputs and weight restrictions.

**CROSS-EFFICIENCY ANALYSIS**

The analysis of cross efficiency is based on the definition of the efficiency matrix which provides information on the nature of the weights system adopted by the units for their own efficiency evaluation. The square efficiency matrix contains as many rows and columns as there are units being compared. The generic element $\theta ij$ of the matrix represents the efficiency of DMU$j$ evaluated through the optimal weights structure for DMU$i$ , while the element $\theta jj$ provides the efficiency of DMU$j$ calculated using its own optimal weights. If DMU$j$ is efficient (i.e. if $\theta jj = 1$), although it exhibits a behavior specialized along a given dimension with respect to the other units, the efficiency values in the column corresponding to DMU$j$ will be less than 1. Two quantities of interest can be derived from the efficiency matrix. The first represents the average efficiency of a unit with respect to the optimal weights systems for the different units, obtained as the average of the values in the $j$ th column.

The second is the average efficiency of a unit measured applying its optimal system of weights to the other units. The latter is obtained by averaging the values in the row associated with the unit being examined. The difference between the efficiency score $\theta jj$ of DMU$j$ and the efficiency obtained as the average of the values in the $j$ th column provides an indication of how much the unit relies on a system of weights conforming with the one used by the other units in the evaluation process. If the difference between the two terms is significant, DMU$j$ may have chosen a structure of weights that is not shared by the other DMUs in order to privilege the dimensions of analysis on which it appears particularly efficient.

**VIRTUAL INPUTS AND VIRTUAL OUTPUTS**

Virtual inputs and virtual outputs provide information on the relative importance that each unit attributes to each individual input and output, for the purpose of maximizing its own efficiency score. Thus, they allow the specific competencies of each unit to be identified, highlighting at the same time its weaknesses. The virtual inputs of a DMU are defined as the product of the inputs used by the unit and the corresponding optimal weights. Similarly, virtual outputs are given by the product of the outputs of the unit and the associated optimal weights.

Inputs and outputs for which the unit shows high virtual scores provide an indication of the activities in which the unit being analyzed appears particularly efficient. Notice that model admits in general multiple optimal solutions, corresponding to which it is possible to obtain different combinations of virtual inputs and virtual outputs. Two efficient units may yield high virtual values corresponding to different combinations of inputs and outputs, showing good operating practices in different contexts. In this case, it might be convenient for each unit to follow the principles and operating methods shown by the other, aiming at improving its own efficiency on a specific dimension.

**WEIGHT RESTRICTIONS**

To separate the units that are really efficient from those whose efficiency score largely depends on the selected weights system, we may impose some restrictions on the value of the weights to be associated with inputs and outputs. In general, these restrictions translate into the definition of maximum thresholds for the weight of specific outputs or minimum thresholds for the weight of specific inputs. Notice that, despite possible restrictions on the weights, the units still enjoy a certain flexibility in the choice of multiplicative factors for inputs and outputs. For this reason it may be useful to resort to the evaluation of virtual inputs and virtual outputs in order to identify the units with the most efficient operating practices with respect to the usage of a specific input resource or to the production of a given output.

**4. Explain the ways of converting inefficient units to efficient unit?**

Data envelopment analysis identifies for each inefficient unit a set of excellent units, called a *peer group*, which includes those units that are efficient if evaluated with the optimal system of weights of an inefficient unit. The peer group, made up of DMUs which are characterized by operating methods similar to the inefficient unit being examined, is a realistic term of comparison which the unit should aim to imitate in order to improve its performance. The units included in the peer group of a given unit DMU$j$ may be identified by the solution to model. Indeed, these correspond to the DMUs for which the first and the second member of constraints are equal:

$$E_j = \left\{ j : \sum_{r \in \mathcal{K}} u_r^* y_{rj} = \sum_{i \in \mathcal{H}} v_i^* x_{ij} \right\}$$

Alternatively, with respect to formulation ,the peer group consists of those units whose variable $\lambda j$in the optimal solution is strictly positive:

$$E_j = \{ j : \lambda_j^* > 0 \}$$

Notice that within a peer group a few excellent units more than others may represent a reasonable term of comparison. The relative importance of a unit belonging to a peer group depends on the value of the corresponding variable $\lambda j$in the optimal solution of the dual model. The analysis of peer groups allows one to differentiate between really efficient units and apparently efficient units for which the choice of an optimal system of weights conceals some abnormal behavior. In order to draw this distinction, it is necessary to consider the efficient units and to evaluate how often each belongs to a peer group. One may reasonably expect that an efficient unit often included in the peer groups uses for the evaluation of its own efficiency a robust weights structure. Conversely, if an efficient unit rarely represents a term of comparison, its own system of optimal weights may appear distorted, in the sense that it may implicitly reflect the specialization of the unit along a particular dimension of analysis.

**5. Explain cluster analysis in detail. (APRIL/MAY 2017)**

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression.

Although classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labelling of a large set of training tuples or patterns, which the classifier uses to model each group. It is often more desirable to proceed in the reverse direction: First partition the set of data into groups based on data similarity (e.g., using clustering), and then assign labels to the relatively small number of groups. Additional advantages of such a clustering-based process are that it is adaptable to changes and helps single out useful features that distinguish different group. Clustering is a challenging field of research in which its potential applications pose their own special requirements.

The following are typical requirements of clustering in data mining:

- **Scalability**: Many clustering algorithms work well on small data sets containing fewer than several hundred data objects; however, a large database may contain millions of objects. Clustering on a sample of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed. Ability to deal with different types of attributes: Many algorithms are designed to cluster interval-based (numerical) data. However, applications may require clustering other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types. Discovery of clusters with arbitrary shape: Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. It is important to develop algorithms that can detect clusters of arbitrary shape. Minimal requirements for domain knowledge to determine input parameters: Many clustering algorithms require users to input certain parameters in cluster analysis (such as the number of desired clusters). The clustering results can be quite sensitive to input parameters. Parameters are often difficult to determine, especially for data sets containing high-dimensional objects. This not only burdens users, but it also makes the quality of clustering difficult to control.
- **Ability to deal with noisy data:** Most real-world databases contain outliers or missing, unknown, or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality. Incremental clustering and insensitivity to the order of input records: Some clustering algorithms cannot incorporate newly inserted data (i.e., database updates) into existing clustering structures and, instead, must determine a new clustering from scratch. Some clustering algorithms are sensitive to the order of input data.That is, given a set of data objects, such an algorithm may return dramatically different clusterings depending on the order of presentation of the input objects. It is important to develop incremental clustering algorithms and algorithms that are insensitive to the order of input
- **High dimensionality:** A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling low-dimensional data, involving only two to three dimensions. Human eyes are good at judging the quality of clustering for up to three dimensions. Finding clusters of data objects in high dimensional space is challenging, especially considering that such data can be sparse and highly skewed.

- **Constraint-based clustering:** Real-world applications may need to perform clustering under various kinds of constraints. Suppose that your job is to choose the locations for a given number of new automatic banking machines (ATMs) in a city. To decide upon this, you may cluster households while considering constraints such as the city's rivers and highway networks, and the type and number of customers per cluster. A challenging task is to find groups of data with good clustering behavior that satisfy specified constraints.
- **Interpretability and usability**: Users expect clustering results to be interpretable, comprehensible, and usable. That is, clustering may need to be tied to specific semantic interpretations and applications. It is important to study how an application goal may influence the selection of clustering features and methods.

**Categorization of Major Clustering Methods**

Many clustering algorithms exist in the literature. It is difficult to provide a crisp categorization of clustering methods because these categories may overlap, so that a method may have features from several categories. Nevertheless, it is useful to present a relatively organized picture of the different clustering methods\

- **Partitioning methods**
- **Hierarchical methods**
- **Density-based methods**
- **Grid-based methods**
- **Model-based methods**

**6. Explain outlier analysis in detail. (APRIL/MAY 2017)**

"What is an outlier?" Very often, there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers.

Outliers can be caused by measurement or execution error. For example, the display of a person's age as 999 could be caused by a program default setting of an unrecorded age. Alternatively, outliers may be the result of inherent data variability. The salary of the chief executive officer of a company, for instance, could naturally stand out as an outlier among the salaries of the other employees in the firm. Many data mining algorithms try to minimize the influence of outliers or eliminate them all together. This, however, could result in the loss of important hidden information because one person's noise could be another person's signal. In other words, the outliers may be of particular interest, such as in the case of fraud detection, where outliers may indicate fraudulent activity.

Thus, outlier detection and analysis is an interesting data mining task, referred to as outlier mining. Outlier mining has wide applications. As mentioned previously, it can be used in fraud detection, for example, by detecting unusual usage of credit cards or telecommunication services. In addition, it is useful in customized marketing for identifying the spending behavior of customers with extremely low or extremely high incomes, or in medical analysis for finding unusual responses to various medical treatments. Outlier mining can be described as follows: Given a set of n data points or objects and k, the expected number of outliers, find the top k objects that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data. The outlier mining problem can be viewed as two sub problems: (1) define what data can be considered as inconsistent in a given data set, and (2) find an efficient method to mine the outliers so defined

**1) Statistical Distribution-Based Outlier Detection**

The statistical distribution-based approach to outlier detection assumes a distribution or probability model for the given data set (e.g., a normal or Poisson distribution) and then identifies outliers with respect to the model using a discordancy test. Application of the test requires knowledge of the data set parameters (such as the assumed data distribution), knowledge of distribution parameters (such as the mean and variance), and the expected number of outliers "How does the discordancy testing work?" A statistical discordancy test examines two hypotheses: a working hypothesis and an alternative hypothesis. A working hypothesis, H, is a statement that the entire data set of n objects comes from an initial distribution model, F, that is, The hypothesis is retained if there is no statistically significant evidence supporting its rejection. A discordancy test verifies whether an object, oi, is significantly large (or small) in relation to the distribution F.

**2) Distance-Based Outlier Detection**

The notion of distance-based outliers was introduced to counter the main limitations imposed by statistical methods.

An object, o, in a data set, D, is a distance-based (DB) outlier with parameters pct and dmin,11 that is, a DB(pct;dmin)-outlier, if at least a fraction, pct, of the objects in D lie at a distance greater than dmin from o. In other words, rather than relying on statistical tests, we can think of distance-based outliers as those objects that do not have "enough" neighbors, where neighbors are defined based on distance from the given object. In comparison with statistical-based methods, distance based outlier detection generalizes the ideas behind discordancy testing for various standard distributions.

Distance-based outlier detection avoids the excessive computation that can be associated with fitting the observed distribution into some standard distribution and in selecting discordancy tests. For many discordancy tests, it can be shown that if an object, o, is an outlier according to the given test, then o is also a DB(pct, dmin)-outlier for some suitably defined pct and dmin. For example, if objects that lie three or more standard deviations from the mean are considered to be outliers, assuming a normal distribution, then this definition can be generalized by a DB(0:9988, 0:13σ) outlier.

Several efficient algorithms for mining distance-based outliers have been developed. These are outlined as follows.

- Index-based algorithm: Given a data set, the index-based algorithm uses multidimensional indexing structures, such as R-trees or k-d trees, to search for neighbors of each object o within radius dmin around that object. Let M be the maximum number of objects within the dmin-neighborhood of an outlier. Therefore, onceM+1 neighbors of object o are found, it is clear that o is not an outlier. This algorithm has a worst-case complexity of O(n2k), where n is the number of objects in the data set and k is the dimensionality. The index-based algorithm scales well as k increases. However, this complexity evaluation takes only the search time into account, even though the task of building an index in itself can be computationally intensive.

- Nested-loop algorithm: The nested-loop algorithm has the same computational complexity as the indexbased algorithm but avoids index structure construction and tries to minimize the number of I/Os. It divides the memory buffer space into two halves and the data set into several logical blocks. By carefully choosing the order in which blocks are loaded into each half, I/O efficiency can be achieved.

- Cell-based algorithm: To avoidO(n2) computational complexity, a cell-based algorithm was developed for memory-resident data sets. Its complexity is O(ck +n), where c is a constant depending on the number of cells and k is the dimensionality. In this method, the data space is partitioned into cells with a side length equal to dmin 2pk . Each cell has two layers surrounding it. The first layer is one cell thick, while the second is d2pk _ 1e cells thick, rounded up to the closest integer. The algorithm counts outliers on a cell-by-cell rather than an object-by-object basis. For agiven cell, it accumulates three counts—the number of objects in the cell, in the cell and the first layer together, and in the cell and both layers together. Let's refer to these counts as cell count, cell + 1 layer count, and cell + 2 layers count, respectively.

• An object, o, in the current cell is considered an outlier only if cell + 1 layer count is less than or equal to M. If this condition does not hold, then all of the objects in the cell can be removed from further investigation as they cannot be outliers.

• If cell + 2 layers count is less than or equal to M, then all of the objects in the cell are considered outliers. Otherwise, if this number is more than M, then it is possible that some of the objects in the cell may be outliers. To detect these outliers, object-by-object processing is used where, for each object, o, in the cell, objects in the second layer of o are examined. For objects in the cell, only those objects having no more than M points in their dmin-neighborhoods are outliers. The dmin-neighborhood of an object consists of the object's cell, all of its first layer, and some of its second layer.

**3) Density-Based Local Outlier Detection**

Statistical and distance-based outlier detection both depend on the overall or "global" distribution of the given set of data points, D. However, data are usually not uniformly distributed.

**4) Deviation-Based Outlier Detection**

Deviation-based outlier detection does not use statistical tests or distance-based measures to identify exceptional objects. Instead, it identifies outliers by examining the main characteristics of objects in a group. Objects that "deviate" from this description are considered outliers. Hence, in this approach the term deviation is typically used to refer to outliers. In this section, we study two techniques for deviation-based outlier detection. The first sequentially compares objects in a set, while the second employs an OLAP data cube approach.

**Sequential Exception Technique**

The sequential exception technique simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects. It uses implicit redundancy of the data. Given a data set, D, of n objects, it builds a sequence of subsets, {D1, D2, : : : , Dm}, of these object
with 2<= m <=n such that Dissimilarities are assessed between subsets in the sequence. The technique introduces the following key terms.

- Exception set: This is the set of deviations or outliers. It is defined as the smallest subset of objects whose removal results in the greatest reduction of dissimilarity in the residual set.
- Dissimilarity function: This function does not require a metric distance between the objects. It is any function that, if given a set of objects, returns a low value if the objects are similar to one another. The greater the dissimilarity among the objects, the higher the value returned by the function. The dissimilarity of a subset is incrementally computed based on the subset prior to it in the sequence.

**OLAP Data Cube Technique**

An OLAP approach to deviation detection uses data cubes to identify regions of anomalies in large multidimensional data. For added efficiency, the deviation detection process is overlapped with cube computation. The approach is a form of discovery-driven exploration, in which precomputed measures indicating data exceptions are used to guide the user in data analysis, at all levels of aggregation. A cell value in the cube is considered an exception if it is significantly different from the expected value, based on a statistical model. The method uses visual cues such as background color to reflect the degree of exception of each cell. The user can choose to drill down on cells that are flagged as exceptions. The measure value of a cell may reflect exceptions occurring at more detailed or lower levels of the cube, where these exceptions are not visible from the current level.

The model considers variations and patterns in the measure value across all of the dimensions to which a cell belongs. For example, suppose that you have a data cube for sales data and are viewing the sales summarized per month. With the help of the visual cues, you notice an increase in sales in December in comparison to all other months. This may seem like an exception in the time dimension. However, by drilling down on the month of December to reveal the sales per item in that month, you note that there is a similar increase in sales for other items during December. Therefore, an increase in total sales in December is not an exception if the item dimension is considered. The model considers exceptions hidden at all aggregated group-by's of a data cube. Manual detection
of such exceptions is difficult because the search space is typically very large, particularly when there are many dimensions involving concept hierarchies with several levels.

**7. Write short notes on cross efficiency analysis**. (**APRIL/MAY 2017)**

The analysis of cross efficiency is based on the definition of the efficiency matrix which provides information on the nature of the weights system adopted by the units for their own efficiency evaluation. The square efficiency matrix contains as many rows and columns as there are units being compared. The generic element $\theta ij$ of the matrix represents the efficiency of DMU$j$ evaluated through the optimal weights structure for DMU$i$ , while the element $\theta jj$ provides the efficiency of DMU$j$ calculated using its own optimal weights.

If DMU*j* is efficient (i.e. if $\theta jj = 1$), although it exhibits a behavior specialized along a given dimension with respect to the other units, the efficiency values in the column corresponding to DMU*j* will be less than 1. Two quantities of interest can be derived from the efficiency matrix. The first represents the average efficiency of a unit with respect to the optimal weights systems for the different units, obtained as the average of the values in the *j* th column. The second is the average efficiency of a unit measured applying its optimal system of weights to the other units. The latter is obtained by averaging the values in the row associated with the unit being examined. The difference between the efficiency score $\theta jj$ of DMU*j* and the efficiency obtained as the average of the values in the *j*th column provides an indication of how much the unit relies on a system of weights conforming with the one used by the other units in the evaluation process. If the difference between the two terms is significant, DMU*j* may have chosen a structure of weights that is not shared by the other DMUs in order to privilege the dimensions of analysis on which it appears particularly efficient.

**8. Explain about hierarchical clustering analysis in detail with an example. (NOV/DEC 2017)**

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds. The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds.

**9. Explain about peer group arrangement and identification of good operating practices for Case-Based Reasoning (CBR). (NOV/DEC 2017)**

**Refer Question no 3 and 4.**

**UNIT IV BUSINESS INTELLIGENCE APPLICATIONS**

**Marketing models – Logistic and Production models – Case studies.**

## PART-A

**1. Write short notes on Marketing decision processes?**

Marketing decision processes are characterized by a high level of complexity due to the simultaneous presence of multiple objectives and countless alternative actions resulting from the combination of the major choice options available to decision makers

**2. What is relational marketing?**

The aim of a relational marketing strategy is to initiate, strengthen, intensify and preserve over time the relationships between a company and its stakeholders, represented primarily by its customers, and involves the analysis, planning, execution and evaluation of the activities carried out to pursue these objectives.

**3. What is customer relationship management (CRM)? (NOV/DEC 2017)**

Customer relationship management (CRM) is a term that refers to practices, strategies and technologies that companies use to manage and analyze customer interactions and data throughout the customer lifecycle, with the goal of improving business relationships with customers, assisting in customer retention and driving sales growth.

**4. Write something about sales force automation (SFA)?**

Sales force automation (SFA) is an integrated application of customizable customer relationship management (CRM) tools that automate and streamline sales inventory, leads, forecasting, performance and analysis. SFA tools include Web-based (hosted CRM) and in-house systems. SFA is also known as sales force management system.

**5. Why do we have relational marketing analysis?**

Relational marketing analyses can be used to identify customers who are more likely to take up the offer of additional services and products (cross-selling), or of alternative services and products of a higher level and with a greater profitability for the enterprise (up-selling).

6. **How the Acquisition can be done, why?**
   Although retention plays a prominent role in relational marketing strategies, for many companies the acquisition of new customers also represents a critical factor for growth. The acquisition process requires the identification of new prospects, as they are potential customers who may be totally or partially unaware of the products and services offered by the company.
7. **What is Cross-selling?**
   The term cross-selling refers to the attempt to sell an additional product or service to an active customer, already involved in a long-lasting commercial relationship with the enterprise. By means of classification models, it is possible to identify the customers characterized by a high probability of accepting a cross-selling offer, starting from the information contained in the available attributes.
8. **Write brief notes on Market basket analysis?**
   The purpose of market basket analysis is to gain insight from the purchases made by customers in order to extract useful knowledge to plan marketing actions. It is mostly used to analyze purchases in the retail industry and in e-commerce activities, and is generally amenable to unsupervised learning problems. It may also be applied in other domains to analyze the purchases made using credit cards, the complementary services activated by mobile or fixed telephone customers, the policies or the checking accounts acquired by a same household. Each transaction consists of a list of purchased items. This list is called a basket.
9. **What is web mining?**
   The web is a critical channel for the communication and promotion of a company's image. Moreover, e-commerce sites are important sales channels. Hence, it is natural to use web mining methods in order to analyze data on the activities carried out by the visitors to a website.
   Web mining methods are mostly used for three main purposes,
   (i) Content mining (ii) Structure mining (iii) Usage mining.
10. **What does meant by salesforce?**
    The term salesforce is generally taken to mean the whole set of people and roles that are involved, with different tasks and responsibilities, in the sales process
11. **Write short notes on content mining?**
    Content mining involves the analysis of the content of web pages to extract useful information. Search engines primarily perform content mining activities to provide the links deemed interesting in relation to keywords supplied by users. Content mining methods can be traced back to data mining problems for the analysis of texts, both in free format or HTML and XML formats, images and multimedia content.
12. **Write short notes on structure mining?**
    The aim of this type of analysis is to explore and understand the topological structure of the web. Using the links presented in the various pages, it is possible to create graphs where the nodes correspond to the web pages and the oriented arcs are associated with links to other pages. Results and algorithms from graph theory are used to characterize the structure of the web.
13. **Write short notes on usage mining?**
    Analyses aimed at usage mining are certainly the most relevant from a relational marketing standpoint, since they explore the paths followed by navigators and their behaviors during a visit to a company website. Methods for the extraction of association rules are useful in obtaining correlations between the different pages visited during a session.
14. **Write down the taxonomy of salesforces?**
    A preliminary taxonomy of salesforces is based on the type of activity carried out, as indicated below.
    • **Residential**. Residential sales activities take place at one or more sites managed by a company supplying some products or services, where customers go to make their purchases. This category includes sales at retail outlets as well as wholesale trading centers and cash-and-carry shops.
    • **Mobile.** In mobile sales, agents of the supplying company go to the customers' homes or offices to promote their products and services and collect orders. Sales in this category occur mostly within B2B relationships, even though they can also be found in B2C contexts.
    • **Telephone**. Telephone sales are carried out through a series of contacts by telephone with prospective customer.
15. **List the decision-making processes relative to salesforce management?**
    The decision-making processes relative to salesforce management can be grouped into three categories: (i) Design. (ii) Planning. (iii)Assessment.

**16. What are Response functions?**

Response functions play a key role in the formulation of models for designing and planning a sales network. In general terms, a response function describes the elasticity of sales in terms of the intensity of the sales actions, and is a formal method to describe the complex relationship existing between sales actions and market reactions.

**17. Write about Sales territory design?**

Sales territory design involves allocating sales coverage units to individual agents so as to minimize a weighted sum of two terms, representing respectively the total distance between areas belonging to the same territory and the imbalance of sales opportunities for the agents.

**18. Define Supply Chain?**

A supply chain may be defined as a network of connected and interdependent organizational units that operate in a coordinated way to manage, control and improve the flow of materials and information originating from the suppliers and reaching the end customers, after going through the procurement, processing and distribution subsystems of a company.

**19. What is Backlogging?**

The term backlog refers to the possibility that a portion of the demand due in a given period may be satisfied in a subsequent period, incurring an additional penalty cost. Backlogs are a feature of production systems more likely to occur in B2B or make-to-order manufacturing contexts. In B2C industries, such as mass production consumer goods, on the other hand, one is more likely to find a variant of the backlog, known as lost sales, in which unfulfilled demand in a period cannot be transferred to a subsequent period and is lost.

**20. State minimum lot conditions?**

A further feature often appearing in manufacturing systems is represented by minimum lot conditions: for technical or scale economy reasons, it is sometimes necessary that the production volume for one or more products be either equal to 0 (i.e. the product is not manufactured in a specific period) or not less than a given threshold value, the minimum lot.

**21. What is Revenue management?**

Revenue management is a managerial policy whose purpose is to maximize profits through an optimal balance between demand and supply. It is mainly intended for marketing as well as logistic activities and has found growing interest in the service industry, particularly in the air transportation, tourism and hotel sectors. More recently these methods have also begun to spread within the manufacturing and distribution industries.

**22. State the use of data mining and business intelligence in health care department. (APRIL/MAY 2017)**

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Some experts believe the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. Data mining involves uncovering patterns from vast data stores and using that information to build predictive models.

**23. Define mental models. (APRIL/MAY 2017)**

Mental model is a explanation of someone's thought process about how something works in the real world. It is a representation of the relationships between its various parts and a person's intuitive perception about his or her own acts and their consequences.

**24. Comment on information broadcasting tools for business intelligence. (NOV/DEC 2017)**

Information broadcasting allows you to make objects with business intelligence content available to a wide spectrum of users, according to your own requirements. It also offers functions to optimize performance and for exception reporting. BEx Information Broadcasting is the Business Explorer component you use to distribute BI objects that were generated with the various BEx tools. With the BEx Broadcaster, you can precalculate queries, query views, Web templates, reports and workbooks and broadcast them by e-mail or to the portal.

## PART-B

**1. Explain in detail about Relational Marketing components and decision making option with suitable examples.**

In order to fully understand the reasons why enterprises develop relational marketing initiatives, consider the following three examples: an insurance company that wishes to select the most promising market segment to target for a new type of policy; a mobile phone provider that wishes to identify those customers with the highest probability of churning, that is, of discontinuing their service and taking out a new contract with a competitor, in order to develop targeted retention initiatives; a bank issuing credit cards that needs to identify a group of customers to whom a new savings management service should be offered. These situations share some common features: a company owning a massive database which describes the purchasing behavior of its customers and the way they make use of services, wishes to extract from these data useful and accurate knowledge so as to develop targeted and effective marketing campaigns.

The aim of a relational marketing strategy is to initiate, strengthen, intensify and preserve over time the relationships between a company and its stakeholders, represented primarily by its customers, and involves the analysis, planning, execution and evaluation of the activities carried out to pursue these objectives.

Relational marketing became popular during the late 1990s as an approach to increasing customer satisfaction in order to achieve a sustainable competitive advantage. So far, most enterprises have taken at least the first steps in this direction, through a process of cultural change which directs greater attention toward customers, considering them as a formidable asset and one of the main sources of competitive advantage. A relational marketing approach has been followed in a first stage by service companies in the financial and telecommunications industries, and has later influenced industries such as consumer goods, finally reaching also manufacturing companies, from automotive and commercial vehicles to agricultural equipments, traditionally more prone to a vision characterized by the centrality of products with respect to customers.

**Motivations and objectives**

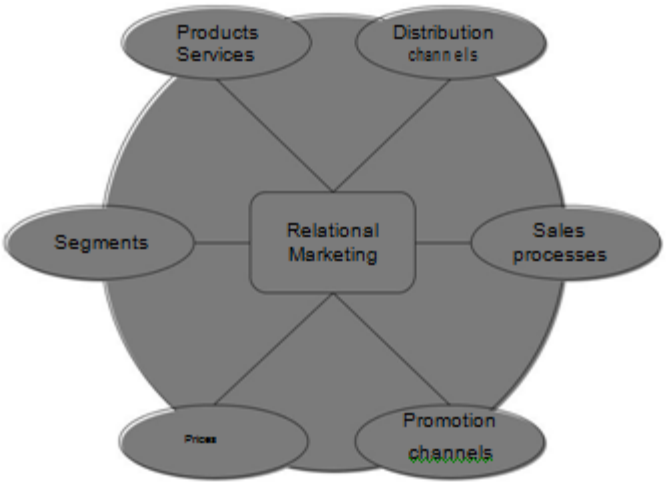
The reasons for the spread of relational marketing strategies are complex and interconnected. Some of them are listed below, although for additional information the reader is referred to the suggested references at the end of the chapter.

- The increasing concentration of companies in large enterprises and the resulting growth in the number of customers have led to greater complexity in the markets.
- Since the 1980s, the innovation – production – obsolescence cycle has progressively shortened, causing a growth in the number of customized options on the part of customers, and an acceleration of marketing activities by enterprises.
- The increased flow of information and the introduction of e-commerce have enabled global comparisons. Customers can use the Internet to compare features, prices and opinions on products and services offered by the various competitors.
- Customer loyalty has become more uncertain, primarily in the service industries, where often filling out an on-line form is all one has to do to change service provider.
- In many industries a progressive commoditization of products and ser-vices is taking place, since their quality is perceived by consumers as equivalent, so that differentiation is mainly due to levels of service.
- The systematic gathering of sales transactions, largely automated in most businesses, has made available large amounts of data that can be transformed into knowledge and then into effective and targeted marketing actions.
- The number of competitors using advanced techniques for the analysis of marketing data has increased.

Relational marketing strategies revolve around the choices shown , which can be effectively summarized as formulating for each segment, ideally for each customer, the appropriate offer through the most suitable channel, at the right time and at the best price.

The ability to effectively exploit the information gathered on customers' behavior represents today a powerful competitive weapon for an enterprise. A company capable of gathering, storing, analyzing and understanding the huge amount of data on its customers can base its marketing actions on the knowledge extracted and achieve sustainable competitive advantages. Enterprises may profitably adopt relational marketing strategies to transform occasional contacts with their customers into highly customized long-term relationships. In this way, it is possible to achieve increased customer satisfaction and at the same time increased profits for the company, attaining a win – win relationship.

To obtain the desired advantages, a company should turn to relational marketing strategies by following a correct and careful approach. In particular, it is advisable to stress the distinction between a relational marketing vision and the software tools usually referred to as customer relationship management (CRM). Relational marketing is not merely a collection of software applications, but rather a coherent project where the various company departments are called upon to cooperate and integrate the managerial culture and human resources, with a high impact on the organizational structures. It is then necessary to create within a company a true data culture, with the awareness that customer-related information should be enhanced through the adoption of business intelligence and data mining analytical tools.

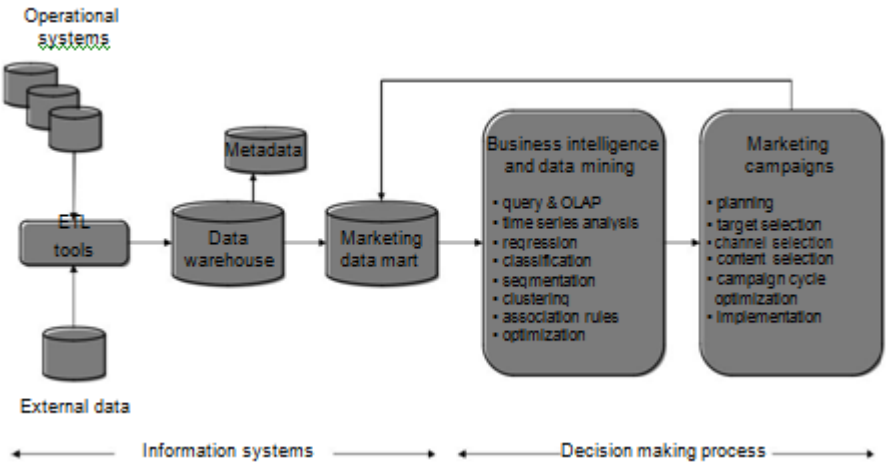



*Decision-making options for a relational marketing strategy*

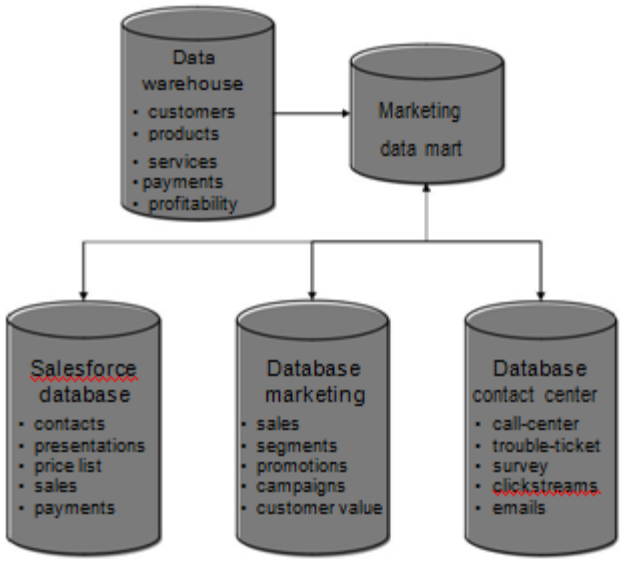
**2. Explain in detail about an environment for relational marketing analysis.**

The Figure shows the main elements that make up an environment for relational marketing analysis. Information infrastructures include the company's data warehouse, obtained from the integration of the various internal and external data sources, and a marketing data mart that feeds business intelligence and data mining analyses for profiling potential and actual customers. Using pattern recognition and machine learning models as described in previous chapters, it is possible to derive different segmentations of the customer base, which are then used to design targeted and optimized marketing actions. A classification model can be used, for example, to generate a scoring system for customers according to their propensity to buy a service offered by a company, and to direct a cross-selling offer only toward those customers for whom a high probability of acceptance is predicted by the model, thus maximizing the overall redemption of the marketing actions.

Effective management of frequent marketing campaign cycles is certainly a complex task that requires planning, for each segment of customers, the content of the actions and the communication channels, using the available




human and financial resources. The corresponding decision-making process can be formally expressed by appropriate optimization models. The cycle of marketing activities terminates with the execution of the planned campaign, with the subsequent gathering of information on the results and the redemption among the recipients. The data collected are then fed into the marketing data mart for use in future data mining analyses. During the execution of each campaign, it is important to set up procedures for controlling and analyzing the results obtained. In order to assess the overall effectiveness of a campaign, it would be advisable to select a control group of customers, with characteristics similar to those of the campaign recipients, toward whom no action should be undertaken.




*Types of data feeding a data mart for relational marketing analysis*

The above describes the main types of data stored in a data mart for relational marketing analyses. A company data warehouse provides demographic and administrative information on each customer and the transactions carried out for purchasing products and using services. The marketing database contains data on initiatives carried out in the past, including previous campaigns and their results, promotions and advertising, and analyses of customer value. A further possible data source is the salesforce database, which provides information on established contacts, calls and applicable sales conditions. Finally, the contact center database provides access to data on customers' contacts with the call center, problems reported, sometimes called trouble tickets, and related outcomes, website navigation paths and forms filled out on-line, and emails exchanged between customers and the support center.

**3. Describe in brief given below,**

**I. Acquisition**

Although retention plays a prominent role in relational marketing strategies, for many companies the acquisition of new customers also represents a critical factor for growth. The acquisition process requires the identification of new prospects, as they are potential customers who may be totally or partially unaware of the products and services offered by the company, or did not possess in the past the characteristics to become customers, or were customers of competitors. It may also happen that some of the prospects were former customers who switched their custom to competitors, in which case much more information is usually available on them.

Once prospects have been identified, the enterprise should address acquisition campaigns to segments with a high potential profitability and a high probability of acquisition, in order to optimize the marketing resources. Traditional marketing techniques identify interesting segments using predefined pro-filing criteria, based on market polls and socio-demographic analyses, according to a top-down perspective. This approach can be successfully integrated or even replaced by a top-down segmentation logic which analyzes the data available in the data mart (demographic information, contacts with prospects, use of products and services of competitors), and derives classification rules that characterize the most promising profiles for acquisition purposes. Also in this case, we are faced with a binary classification problem, which can be analyzed with the techniques

**II. Retention**

The maturity stage reached by most products and services and the subsequent saturation of their markets have caused more severe competitive conditions. As a consequence, the expansion of the customer base of an enterprise consists more and more of switch mechanisms – the acquisition of customers at the expense of other companies. This phenomenon is particularly apparent in service industries, such as telecommunications, banking, savings management and insurance, although it also occurs in manufacturing, both for consumer goods and industrial products. For this reason, many companies invest significant amounts of resources in analyzing and characterizing the phenomenon of attrition, whereby customers switch from their company to a competitor. There are also economic reasons for devoting substantial efforts to customer retention: indeed, it has been empirically observed that the cost of acquiring a new customer, or winning back a lost customer, is usually much higher – of the order of 5 to 9 times higher – than the cost of the marketing actions aimed at retaining customers considered at risk of churning. Furthermore, an action to win back a lost customer runs the risk of being too late and not achieving the desired result. In many instances, winning back a customer requires investments that do not generate a return.

One of the main difficulties in loyalty analysis is actually recognizing a churn. For subscription services there are unmistakable signals, such as a formal notice of withdrawal, while in other cases it is necessary to define adequate indicators that are correlated, a few periods in advance, with the actual churning. A customer who reduces by more than a given percentage her purchases at a selected point of sale using a loyalty card, or a customer who reduces below a given threshold the amount held in her checking account and the number of transactions, represent two examples of disaffection indicators. They also highlight the difficulties involved in correctly defining the appropriate threshold values. To optimize the marketing resources addressed to retention, it is therefore necessary to target efforts only toward high-value customers considered at risk of churning. To obtain a scoring system corresponding to the probability of churning for each customer, it is necessary to derive a segmentation based on the data on past instances of churning. Predicting the risk of churning requires analysis of records of transactions for each customer and identifying the attributes that are most relevant to accurately explaining the target variable. Again, we are faced with a binary classification problem. Once the customers with the highest risk of churning have been identified, a retention action can be directed toward them. The more accurately such action is targeted, the cheaper it is likely to be.

**III. Cross-selling and up-selling**

Data mining models can also be used to support a relational marketing analysis aimed at identifying market segments with a higher propensity to purchase additional services or other products of a company. For example, a bank also offering insurance services may identify among its customers segments interested in purchasing a life insurance policy. In this case, demographic information on customers and their past transactions recorded in a data mart can be used as explanatory attributes to derive a classification model for predicting the target class, consisting in this example of a binary variable that indicates whether the customer accepted the offer or not.

The term cross-selling refers to the attempt to sell an additional product or service to an active customer, already involved in a long-lasting commercial relationship with the enterprise. By means of classification models, it is possible to identify the customers characterized by a high probability of accepting a cross-selling offer, starting from the information contained in the available attributes.

In other instances, it is possible to develop an up-selling initiative, by persuading a customer to purchase an higher-level product or service, richer in functions for the user and more profitable for the company, and therefore able to increase the lifetime value curve of a customer. For example, a bank issuing credit cards may offer customers holding a standard card an upgrade to a gold card, which is more profitable for the company, but also able to offer a series of complementary services and advantages to interested customers. In this case too, we are dealing with a binary classification problem, which requires construction of a model based on the training data of customers' demographic and operational attributes. The purpose of the model is to identify the most interesting segments, corresponding to customers who have taken up the gold service in the past, and who appear therefore more appreciative of the additional services offered by the gold card. The segments identified in this way represent the target of up-selling actions.

**IV. Market basket analysis.**

The purpose of market basket analysis is to gain insight from the purchases made by customers in order to extract useful knowledge to plan marketing actions. It is mostly used to analyze purchases in the retail industry and in e-commerce activities, and is generally amenable to unsupervised learning problems. It may also be applied in other domains to analyze the purchases made using credit cards, the complementary services activated by mobile or fixed telephone customers, the policies or the checking accounts acquired by a same household.

The data used for this purpose mostly refer to purchase transactions, and can be associated with the time dimension if the purchaser can be tracked through a loyalty card or the issue of an invoice. Each transaction consists of a list of purchased items. This list is called a basket , just like the baskets available at retail points of sale.

**4. Explain in detail the Taxonomy of web mining analysis.**

The web is a critical channel for the communication and promotion of a company's image. Moreover, e-commerce sites are important sales channels. Hence, it is natural to use web mining methods in order to analyze data on the activities carried out by the visitors to a website. Web mining methods are mostly used for three main purposes, as shown in the below Figure

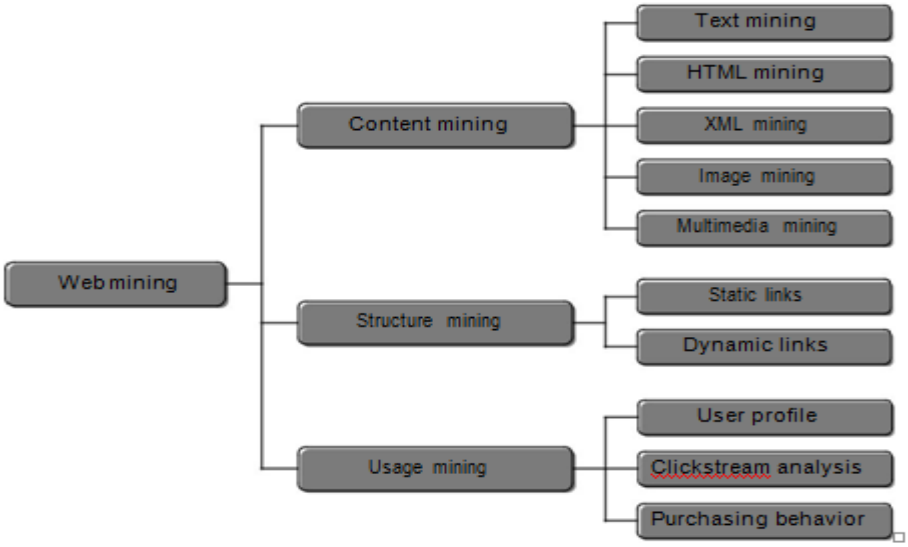
**(i) Content mining.** Content mining involves the analysis of the content of web pages to extract useful information. Search engines primarily perform content mining activities to provide the links deemed interesting in relation to keywords supplied by users. Content mining methods can be traced back to data mining problems for the analysis of texts, both in free format or HTML and XML formats, images and multimedia content. Each of these problems is in turn dealt with using the learning models described in previous chapters.

For example, text mining analyses are usually handled as multicategory classification problems, where the target variable is the subject category to which the text refers, while explanatory variables correspond to the meaningful words contained in the text. Once it has been converted into a classification problem, text mining can be approached using the methods described in Chapter 10. Text mining techniques are also useful for analyzing the emails received by a support center. Notice that the input data for content mining analyses are easily retrievable, at least in principle, since they consist of all the pages that can be visited on the Internet.

**(ii) Structure mining.** The aim of this type of analysis is to explore and understand the topological structure of the web. Using the links presented in the various pages, it is possible to create graphs where the nodes correspond to the web pages and the oriented arcs are associated with links to other pages. Results and algorithms from graph theory are used to characterize the structure of the web, that is, to identify areas with a higher density of connections, areas disconnected from others and maximal cliques, which are groups of pages with reciprocal links.

In this way, it is possible to pinpoint the most popular sites, or to measure the distance between two sites, expressed in terms of the lowest number of arcs along the paths that connect them in the links graph. Besides analyses aimed at exploring the global structure of the web, it is also possible to carry out local investigations to study how a single website is articulated. In some investigations, the local structure of websites is associated with the time spent by the users on each page, to verify if the organization of the site suffers from inconsistencies that jeopardize its effectiveness. For example, a page whose purpose is to direct navigation on the site should be viewed by each user only briefly. Should this not be the case, the page has a problem due to a possible ambiguity in the articulation of the links offered.

**(iii) Usage mining.** Analyses aimed at usage mining are certainly the most relevant from a relational marketing standpoint, since they explore the paths followed by navigators and their behaviors during a visit to a company website. Methods for the extraction of association rules are useful in obtaining correlations between the different pages visited during a session. In some instances, it is possible to identify a visitor and recognize her during subsequent sessions. This happens if an identification key is required to access a web page, or if a cookie-enabling mechanism is used to keep track of the sequence of visits. Sequential association rules or time series models can be used to analyze the data on the use of a site according to a temporal dynamic. Usage mining analysis is mostly concerned with clickstreams – the sequences of pages visited during a given session. For e-commerce sites, information on the purchase behavior of a visitor is also available.




*Taxonomy of web mining analyses*

**5. Describe salesforce management. Explain in detail the decision processes, models in salesforce management.**

Most companies have a sales network and therefore rely on a substantial number of people employed in sales activities, who play a critical role in the profitability of the enterprise and in the implementation of a relational marketing strategy. The term salesforce is generally taken to mean the whole set of people and roles that are involved, with different tasks and responsibilities, in the sales process. A preliminary taxonomy of salesforces is based on the type of activity carried out, as indicated below.

**(i) Residential:** sales activities take place at one or more sites managed by a company supplying some products or services, where customers go to make their purchases. This category includes sales at retail outlets as well as wholesale trading centers and *cash-and-carry* shops.

**(ii) Mobile:** agents of the supplying company go to the customers' homes or offices to promote their products and services and collect orders. Sales in this category occur mostly within B2B relationships, even though they can also be found in B2C contexts.

**(iii)Telephone:** sales are carried out through a series of contacts by tele-phone with prospective customers.

There are various problems connected with managing a mobile salesforce management, which will be the main focus of this section. They can be subdivided into a few main categories:

- designing the sales network
- planning the agents' activities
- contact management
- sales opportunity management
- customer management
- activity management
- order management
- area and territory management
- support for the configuration of products and services;
- knowledge management with regard to products and services.

Designing the sales network and planning the agents' activities involve decision-making tasks that may take advantage of the use of optimization models, such as those that will be described in the next sections. The remaining activities are operational in nature and may benefit from the use of software tools for salesforce automation (SFA), today widely implemented.

**Decision processes in salesforce management**

The design and management of a salesforce raise several decision-making problems.When successfully solved, they confer multiple advantages: maximization of profitability, increased effectiveness of sales actions, increased efficiency in the use of resources, and greater professional rewards for sales agents.

The decision processes described in Figure 13.15 should take into account the strategic objectives of the company, with respect to other components of the marketing mix, and conform to the role assigned to the salesforce within the broader framework of a relational marketing strategy. The two-way connections indicated in the figure suggest that the different components of the decision-making process interact with each other and with the general objectives of the marketing department. In particular, the decision-making processes relative to salesforce management can be grouped into three categories: design, planning and assessment .


*Decision processes in salesforce management*

**(i) Design:** Salesforce design is dealt with during the start-up phase of a commercial activity or during subsequent restructuring phases, for example following the merger or acquisition of a group of companies. As shown the design phase is usually preceded by the creation of market segments through the application of data mining methods and by the articulation of the offer of products and services, which are in turn subdivided into homogeneous classes. Salesforce design includes three types of decisions.

**(ii) Organizational structure.** The organizational structure may take different forms, corresponding to hierarchical agglomerations of the agents by group of products, brand or geographical area. In some situations the structure may also be differentiated by markets. In order to determine the organizational structure, it is necessary to analyze the complexity of customers, products and sales activities, and to decide whether and to what extent the agents should be specialized.

**(iii) Sizing.** Sales network sizing is a matter of working out the optimal number of agents that should operate within the selected structure, and depends on several factors, such as the number of customers and prospects, the desired level of sales area coverage, the estimated time for each call and the agents' traveling time.

One should bear in mind that a reduction in costs due to a decrease in the salesforce size is often followed by a reduction in sales and revenues. A better allocation of the existing salesforce, devised during the planning phase by means of optimization models, is usually more effective than a variation in size.

**(iv) Sales territories.** Designing a sales territory means grouping together the geo-graphical areas into which a given region has been divided and assigning each territory to an agent. The design and assignment of sales territories should take into account several factors, such as the sales potential of each geographical area, the time required to travel from one area to another and the total time each agent has available. The purpose of the assignment consists of determining a balanced situation between sales opportunities embedded in each territory, in order to avoid disparities among agents. The assignment of the geographical areas should be periodically reviewed since the sales potential balance in the various territories tends to vary over time.

**(v) Planning:** Decision-making processes for *planning* purposes involve the assignment of sales resources, structured and sized during the design phase, to market entities. Resources may correspond to the work time of agents or to the budget, while market entities consist of products, market segments, distribution channels and customers. Allocation takes into account the time spent pitching the sale to each customer, the travel time and cost, and the effectiveness of the action for each product, service or market segment. It is also possible to consider further ancillary activities carried out at the customers' sites, such as making suggestions that are conducive to future sales or explaining the technical and functional features of products and services. Salesforce planning can greatly benefit from the use of optimization models, as explained below.

**(vi) Assessment**

The purpose of assessment and control activities is to measure the effectiveness and efficiency of individuals employed in the sales network, in order to design appropriate remuneration and incentive schemes. To measure the efficiency of the sales agents it is necessary to define adequate evaluation criteria that take into account the actual personal contribution of each agent, having removed effects due to the characteristics that may make an area or product more or less advantageous than others.

**Models for salesforce management**

Before proceeding, it is useful to introduce some notions common to the different models that will be described. Assume that a region is divided into $J$ geographical sales areas, also called *sales coverage units*, and let $J = \{1, 2, \ldots, J\}$. Areas must be aggregated into disjoint clusters, called *territories*, so that each area belongs to one single territory and is also connected to all the areas belonging to the same territory. The connection property implies that from each area it is possible to reach any other area of the same territory. The time span is divided into $T$ intervals of equal length, which usually correspond to weeks or months, indicated by the index $t \in T = \{1, 2, \ldots, T\}$.

Each territory is associated with a sales agent, located in one of the areas belonging to the territory, henceforth considered as her area of residence. The choice of the area of residence determines the time and cost of traveling to any other area in the same territory. Let $I$ be the number of territories and therefore the number of agents that form the sales network, and let $I = \{1, 2, \ldots, I\}$.

In each area there are customers or prospects who can be visited by the agents as part of their promotions and sales activities. In some of the models that will be presented, customers or prospects are aggregated into segments, which are considered homogeneous with respect to the area of residence and possibly to other characteristics, such as value, potential for development and purchasing behaviors. Let $H$ be the number of market entities, which in different models may represent either single customers or segments, and let $H = \{1, 2, \ldots, H\}$. Let $D_j$ be the set of customers, or segments of customers where necessary, located in area $j$ .Finally, assume that a given agent can promote and sell $K$ products and services during the calls she makes on customers or prospects, and let $K = \{1, 2, \ldots, K\}$.

**6. Explain in detail calls and product presentations planning with optimization model?**

Optimization models for calls and product presentations planning are intended to derive for each agent the optimal sales activity plan.

**Calls planning**

The aim of the first model described is to identify the optimal number of calls to each customer or prospect (taken together as *market entities* in what follows) located in the territory assigned to a specific agent. The objective function expresses the difference between revenues and transfer costs. The decision variables are defined as

$X_h$ =number of calls to market entity $h$ $W_j$ =number of trips to market area $j$,

while the parameters have the following meanings:

$a_h$ = strategic relevance of market entity $h$,

$c_j$ = transfer cost to area $j$,

$v_j$ = transfer time to area $j$,

$t_h$ = time spent with market entity $h$ in each call,

$l_h$ = minimum number of calls to market entity $h$,

$u_h$ = maximum number of calls to market entity $h$,

$b$ = total time available to the sales agent.

The corresponding optimization problem can be formulated as

$$\max \sum_{h \in H} a_h r(X_h) - \sum_{j \in J} c_j W_j,$$

$$\text{s.to} \sum_{h \in H} t_h X_h + \sum_{j \in J} v_j W_j \le b,$$

$$X_h \le u_h, \quad X_h \ge l_h, \qquad h \in H,$$

$$W_j \ge X_h, \qquad j, h \in D,$$

$$X_h, W_j \ge 0 \text{ and integer}, \qquad h \in H, j \in J.$$

**Product presentations planning**

The aim of this model is to determine for each period in the planning horizon the optimal number of mentions for each product belonging to the sales portfolio of a given agent. Through an index called *relative exposure* the model also incorporates the dynamic effects determined by the mentions of each product made in past periods.

while the parameters have the following meanings:

$a_h$ = strategic relevance of market entity $h$,

$c_j$ = transfer cost to area $j$,

$v_j$ = transfer time to area $j$,

$t_h$ = time spent with market entity $h$ in each call,

$l_h$ = minimum number of calls to market entity $h$,

$u_h$ = maximum number of calls to market entity $h$,

$b$ = total time available to the sales agent.

The decision variables of the model are consequently defined as

$X_{kt}$ = number of calls for product $k$ in period $t$,

$Z_{kt}$ = cumulated exposure level for product $k$ in period $t$.

The parameters are

$d_{kt}$ = number of units of product $k$ available in period $t$,

$p$ = maximum number of mentions for each product,

$\lambda$ = memoryless parameter.

The quantity $\sigma (X_{kt})$ expresses the relative exposure of product $k$ as a func-tion of the number of times $k$ has been mentioned in period $t$ . The relative exposure formalizes the relationship between the level of cumulative exposure and the number of mentions made in period $t$ through constraints.

The resulting optimization model is formulated as

$$\max \quad \sum_{t \in T} \sum_{k \in K} d_{kt} r_k (Z_{kt}),$$

$$\text{s.to} \quad \sum_{k \in K} X_{kt} \leq Kp, \qquad t \in T,$$

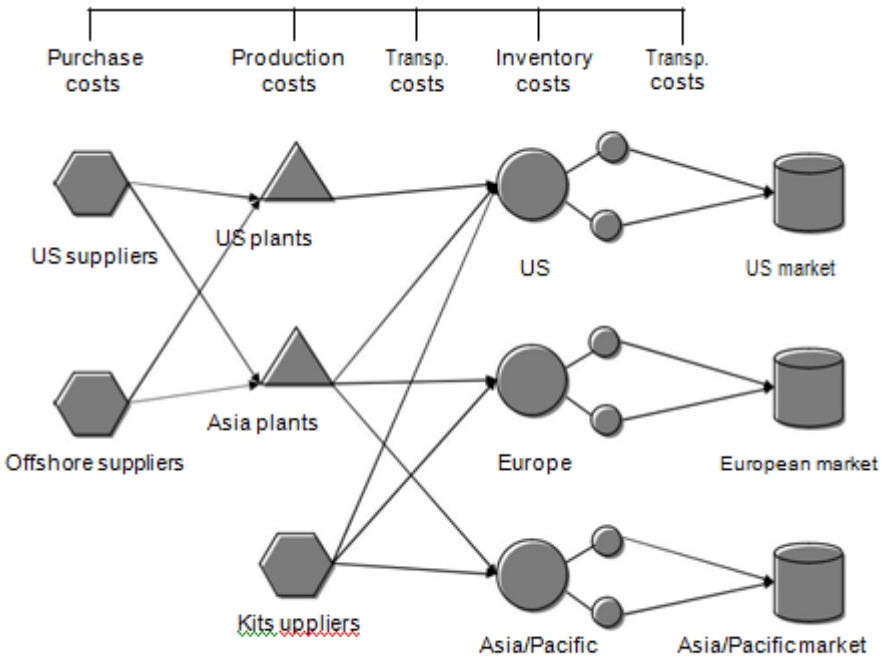
$$Z_{kt} = \lambda \sigma (X_{kt}) + (1 - \lambda)\, \sigma (X_{kt-1}), \qquad k \in K, t \in T,$$

$$X_{kt}, Z_{kt} \geq 0 \text{ and integer}, \qquad k \in K, t \in T.$$

**7. Describe supply chain optimization. Explain in detail the Optimization models for logistics planning.**

A supply chain may be defined as a network of connected and interdependent organizational units that operate in a coordinated way to manage, control and improve the flow of materials and information originating from the suppliers and reaching the end customers, after going through the procurement, processing and distribution subsystems of a company**.** The aim of the integrated planning and operations of the supply chain is to combine and evaluate from a systemic perspective the decisions made and the actions undertaken within the various subprocesses that compose the logistic system of a company. Many manufacturing companies, such as those operating in the consumer goods industry, have concentrated their efforts on the integrated operations of the supply chain, even to the point of incorporating parts of the logistic chain that are outside the company, both upstream and downstream

The major purpose of an integrated logistic process is to minimize a function expressing the total cost, which comprises processing costs, transportation costs for procurement and distribution, inventory costs and equipment costs. Note that




*An example of global supply chain*

The need to optimize the logistic chain, and therefore to have models and computerized tools for medium-term planning and for capacity analysis, is particularly critical in the face of the high complexity of current logistic systems, which operate in a dynamic and truly competitive environment. We are referring here to manufacturing companies that produce a vast array of products and that usually rely on a multicentric logistic system, distributed over several plants and markets, characterized by large investments in highly auto-mated technology, by an intensive usage of the available production capacity and by short-order processing cycles. The features of the logistic system we have described reflect the profile of many enterprises operating in the consumer goods industry. In the perspective outlined above, the aim of a medium-term planning process is therefore to devise an optimal logistic production plan, that is, a plan that is able to minimize the total cost, understood as the sum of procurement, processing, storage, distribution costs and the penalty costs associated with the failure to achieve the predefined service level. However, to be implemented in practice, an optimal logistic production plan should also be feasible, that is, it should be able to meet the physical and logical constraints imposed by limits on the available production capacity, specific technological conditions, the structure of the bill of materials, the configuration of the logistic network, minimum production lots, as well as any other condition imposed by the decision makers in charge of the planning process.

Optimization models represent a powerful and versatile conceptual paradigm for analyzing and solving problems arising within integrated supply chain planning, and for developing the necessary software. Due to the complex interactions occurring between the different components of a logistic production system, other methods and tools intended to support the planning activity seem today inadequate, such as electronic spreadsheets, simulation systems and planning modules at infinite capacity included in enterprise resource planning software. Conversely, optimization models enable the development of realistic mathematical representations of a logistic production system, able to describe with reasonable accuracy the complex relationships among critical components of the logistic system, such as capacity, resources, plans, inventory, batch sizes, lead times and logistic flows, taking into account the various costs. Moreover, the evolution of information technologies and the latest developments in optimization algorithms mean that decision support systems based on optimization models for logistics planning can be efficiently developed.

**(i) Tactical planning**

In its simplest form, the aim of tactical planning is to determine the production volumes for each product over the $T$ periods included in the medium-term planning horizon in such a way as to satisfy the given demand and capacity limits for a single resource, and also to minimize the total cost, defined as the sum of manufacturing production costs and inventory costs.

We therefore consider the decision variables

$P_{it}$ = units of product $i$ to be manufactured in period $t$

$I_{it}$ = units of product $i$ in inventory at the end of period $t$,

and the parameters

$d_{it}$ = demand for product $i$ in period $t$,

$c_{it}$ = unit manufacturing cost for product $i$ in period $t$,

$h_{it}$ = unit inventory cost for product $i$ in period $t$,

$e_i$ = capacity absorption to manufacture a unit of product $i$,

$b_t$ = capacity available in period $t$.

The resulting optimization model is,

$$\min \sum_{t \in T} \sum_{i \in I} (c_{it} P_{it} + h_{it} I_{it})$$

$$\text{s.to} \quad P_{it} + I_{i,t-1} - I_{it} = d_{it}, \qquad i \in I, t \in T,$$

$$\sum_{i \in I} e_i P_{it} \le b_t, \qquad t \in T,$$

$$P_{it}, I_{it} \ge 0, \qquad i \in I, t \in T.$$

**(ii) Extra capacity**

A first extension of the basic model (14.1) deals with the possibility of resorting to *extra capacity*, perhaps in the form of overtime, part-time or third-party capacity. In addition to the decision variables already included in model, we define the variables $O_t$ =extra capacity used in period $t$, and the parameters $q_t$ =unit cost of extra capacity in period $t$.

The optimization problem now becomes

$$\min \sum_{t\in T}\sum_{i\in I}(c_{it}P_{it} + h_{it}I_{it}) + \sum_{t\in T} q_t O_t$$

$$\text{s.to } P_{it} + I_{i,t-1} - I_{it} = d_{it}, \quad i \in I, t \in T,$$

$$\sum_{i\in I} e_i P_{it} \le b_t + O_t, \quad t \in T,$$

$$P_{it}, I_{it}, O_t \ge 0, \quad i \in I, t \in T.$$

**(iii) Multiple resources**

If the manufacturing system requires $R$ critical resources, a further extension of model can be devised by considering multiple capacity constraints. The decision variables already included in model remain unchanged, though it is necessary to consider the additional parameter

$b_{rt}$ =quantity of resource $r$ available in period $t$

$e_{ir}$ = quantity of resource $r$ absorbed to manufacture one unit of product $i$.

The resulting optimization problem is given by

$$\min \sum_{t\in T}\sum_{i\in I}(c_{it}P_{it} + h_{it}I_{it})$$

$$\text{s.to } P_{it} + I_{i,t-1} - I_{it} = d_{it}, \quad i \in I, t \in T,$$

$$\sum_{i\in I} e_{ir} P_{it} \le b_{rt}, \quad r \in R, t \in T,$$

$$P_{it}, I_{it} \ge 0, \quad i \in I, t \in T.$$

**(iv)Backlogging**

Another feature that needs to be modeled in some logistic systems is *backlog-ging*. The term *backlog* refers to the possibility that a portion of the demand due in a given period may be satisfied in a subsequent period, incurring an additional penalty cost. Backlogs are a feature of production systems more likely to occur in B2B or make-to-order manufacturing contexts. In B2C industries, such as mass production consumer goods, on the other hand, one is more likely to find a variant of the backlog, known as *lost sales*, in which unfulfilled demand in a period cannot be transferred to a subsequent period and is lost. To model backlogging, it is necessary to introduce new decision variables

$B_{it}$ =units of demand for product $i$ delayed in period $t$,

and the parameters

$g_{it}$ = unit cost of delaying the demand for product $i$ in period $t$.

The resulting optimization problem is

$$\min \sum_{t\in T}\sum_{i\in I}(c_{it}P_{it} + h_{it}I_{it} + g_{it}B_{it})$$

$$\text{s.to } P_{it} + I_{i,t-1} - I_{it} + B_{it} - B_{i,t-1} = d_{it}, \quad i \in I, t \in T,$$

$$\sum_{i\in I} e_i P_{it} \le b_t, \quad t \in T,$$

$$P_{it}, I_{it}, B_{it} \ge 0, \quad i \in I, t \in T.$$

**(v) Minimum lots and fixed costs**

A further feature often appearing in manufacturing systems is represented by *minimum lot* conditions: for technical or scale economy reasons, it is sometimes necessary that the production volume for one or more products be either equal to 0 (i.e. the product is not manufactured in a specific period) or not less than a given threshold value, the minimum lot. To incorporate minimum lot conditions into the model, we define the binary decision variables

$$Y_{it} = \begin{cases} 1 & \text{if } P_{it} > 0, \\ 0 & \text{otherwise,} \end{cases}$$

and the parameters

$l_i$ = minimum lot for product ***i,***

$\gamma$ = constant value larger than any producible volume for ***i.***

The optimization problem is now

$$\min \sum_{t \in T} \sum_{i \in I} (c_{it} P_{it} + h_{it} I_{it})$$

$$\text{s.to} \quad P_{it} + I_{i,t-1} - I_{it} = d_{it}, \qquad i \in I, t \in T,$$

$$e_i P_{it} \le b_t, \qquad t \in T,$$

$$\sum_{i \in I} P_{it} \ge l_i Y_{it}, \qquad i \in I, t \in T,$$

$$P_{it} \le \gamma Y_{it}, \qquad i \in I, t \in T,$$

$$P_{it}, I_{it} \ge 0, \; Y_{it} \in \{0, 1\}, \qquad i \in I, t \in T.$$

**8. What is Revenue management system? Explain decision processes in revenue management.**

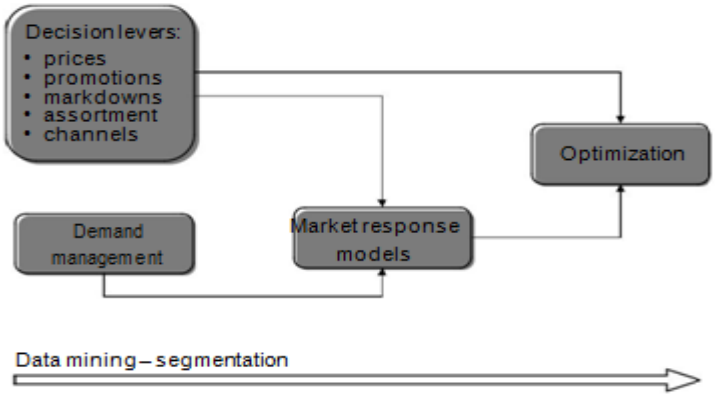
Revenue management is a managerial policy whose purpose is to maximize profits through an optimal balance between demand and supply. It is mainly intended for marketing as well as logistic activities and has found growing interest in the service industry, particularly in the air transportation, tourism and hotel sectors. More recently these methods have also begun to spread within the manufacturing and distribution industries.

**Decision processes in revenue management**

Revenue management involves the application of mathematical models to predict the behavior of customers at a micro-segmentation level and to optimize the availability and price of products in order to maximize profits. In this respect, we can use the same definition introduced in Chapter 13 to summarize relational marketing objectives: to formulate for each segment, ideally for each customer, the appropriate offer through the most suitable channel, at the right time and at the best price. The purpose of revenue management is therefore to maximize profits, aligning the offer of products and services to the expected demand, using both the major levers of the marketing mix (e.g. prices, promotions, assortment) and the levers of logistics (e.g. efficiency and timeliness). Specific and innovative features of revenue management strategies are a closer focus on demand than supply and a greater emphasis on costs than revenues; such features are often absent from the managerial policies adopted by most enterprises.

As already observed, in recent years revenue management has been applied with more and more success by many companies operating in the service industry. Among the pioneers in this field are airlines, hotel chains, automobile rental companies, theme parks, theaters and other entertainment-related enterprises.

The common characteristics of these fields are well apparent: a highly perish-able product, a fairly low marginal sales cost and the possibility of applying dynamic pricing policies and exploiting multiple sales channels. Revenue management affects some highly complex decision-making processes of strategic relevance, as shown in Figure:




*Decision processes in revenue management*

- market segmentation, by product, distribution channel, consumer type and geographic area, performed using data mining models;
- prediction of future demand, using time series and regression models;
- Identification of the optimal assortment, i.e. the mix of products to be allocated to each point of sale;
- definition of the market response function, obtained by identifying models and rules that explain the demand based on company actions, the initiatives of competitors and other exogenous contextual events;
- management of activities aimed at determining the price of each product (*pricing*) as well as the timing and the amount of markdowns;
- planning, management and monitoring of sales promotions, and assessment of their effectiveness;
- sales analysis and control, and use of the information gathered to evaluate market trends;
- material procurement and stock management policies, such as control policy, frequency of issued orders, reorder quantities;
- integrated management of the different sales and distribution channels.

Revenue management relies on the following basic principles:

- To address sales to micro-segments: segmentation carried out by means of business intelligence and data mining models is critical to achieve an adequate knowledge of the market.
- To exploit the product value cycle: to generate the highest revenues, it is required to grasp the value cycle of products and services, in order to optimally synchronize their availability over time and to determine the price for each market micro-segment. Notice that the value cycle also depends on the sensitivity of micro-segments to price variations.
- To have a price-oriented rather than cost-oriented approach in balancing supply and demand: when supply and demand are out of balance, most enterprises tend to react by increasing or decreasing capacity. In many instances it might, however, be more convenient to adopt price variations, avoiding repeated variations in capacity.
- To make informed and knowledge-based decisions: a consistent use of prediction models tends to mean that decisions rest on a more robust knowledge basis. In particular, a correct prediction of consumer pur-chasing behaviors is essential to evaluate elasticity and reactions to price variations.
- To regularly examine new opportunities to increase revenues and profits: the possibility of timely access to the available information, combined with the possibility of considering alternative scenarios, strengthens the competencies of marketing analysts and increases the effectiveness of their activity

**10. Explain about marketing models. (NOV/DEC 2017)**
**Refer Question no 3,4,5 and 6**

**11. Explain in detail about Logistics and production model for business intelligence. (APRIL/MAY 2017)(NOV/DEC 2017)**
**Refer Question no 7 and 8.**

**UNIT V FUTURE OF BUSINESS INTELLIGENCE**

**Future of business intelligence – Emerging Technologies, Machine Learning, Predicting the Future, BI Search & Text Analytics – Advanced Visualization – Rich Report, Future beyond Technology.**

## PART-A

**1. Write about workload automation?**

Sophisticated automation tools are at the forefront of technological advances that are paving the way for the simplification of BI and play a considerable role in sustaining hybrid or centralized BI architecture. Formerly, lengthy querying and reporting jobs that involved time-consuming scripts can now be expedited due to self-service and script less automation job scheduling. With the proper IT configuration, end users are able to issue on-demand queries in close to real time with some of the most viable applications for BI such as Big Data and data from conventional sources. These queries greatly simplify the integration of BI with other platforms.

**2. What is Data Discovery?**

Discovery tools have lessened the need for conventionally lengthy (and time-consuming) BI reports, and significantly augmented them in cases in which they're essential. Dashboards and interactive visualizations graphically represent data and results from BI in ways in which trends are readily discernible, data smashups and in-memory analytics enable users to quickly query a variety of disparate sources, and search tools offer text-derived analysis of either structured or unstructured quantitative and qualitative data.

**3. What are BI technology evaluation?**

- Embrace
- Adopt Where Appropriate
- Evaluate and Test
- Monitor and Understand

**4. What are the Emerging Trends in Business Intelligence?**

- Data Discovery Accelerates Self-Service BI and Analytics
- Unified Access and Analysis of All Types of Information Improves User Productivity
- Big Data Generated by Social Media Drives Innovation in Customer Analytics
- Text Analytics Enables Organizations to Interpret Social Media Sentiment Trends and Commentary
- Decision Management Enables Organizations to be Predictive and Proactive in Real Time

**5. What is the difference between BI Analytics and Predictive Analytics?**

| BI Analytics | Predictive Analytics |
|---|---|
| Business intelligence allows you to answer questions about the demographics or characteristics of your customers, products, stores, etc., and an-swer questions about the performance of your business across a number of different dimensions | Predictive analytics allow organizations to go beyond the answers generated by BI by providing more predictive answers and recommendations to many of the same questions |
| Example:<br>➢ when did customer X last visit the store.<br>➢ How much revenue did store X generate last Christmas? | Example:<br>➢ How many customers are likely to visit the store next week.<br>➢ How much revenue will be generated by store X next Christmas? |

**6. List the Top Business Intelligence Software Products?**

- Sisense
- Corporater EPM Suite
- SAP BusinessObjects
- Hyperion
- PowerCenter
- Cognos
- Active Intelligence
- Teradata Database

**7. Define BI search?**

A BI Search interface promises to change the way users' access information. Picture a Google interface to BI. Without any training in a BI tool, users can enter a phrase such as "Recent sales for customer A" and then be presented with either a list of predefined reports or, in some cases, a newly generated query.

**8. What are the Potential Benefits of BI Search?**

- Self service for report consumers is the best reason for BI search.
- BI search is mostly about finding pre-built reports, not creating new ones.
- BI search unearths more facts for decision making.
- Users associate ease of use with BI search

**9. What is text analytics?**

Text analytics is closely related to search in that unstructured information or text can be transformed into quantitative data. For example, it allows for searching of information in a comment field to see how many times a customer praised a particular product. Text analytics is the numerical analysis of textual information.

**10. What is the difference between structured and unstructured data?**

| Structured data (quantitative) | Unstructured data(textual) |
|---|---|
| Structured data refers to the numerical values typically captured in the operational systems and subsequently stored in a data warehouse | Unstructured content refers to information stored in textual comment fields, documents, annual reports, websites, and so on. |
| Example:<br>➢ RDMS data | Example:<br>➢ Textual data<br>➢ Social network data |

**11. Write notes on advanced visualization?**

Advanced visualization goes beyond a simple chart such as a bar or line chart to includes:

- Spark lines, a highly condensed trend line the size of a word.
- Bullet graphs, a construct by Stephen Few that includes a target indicator within the bar chart.
- Small multiples, which are series of small, similar graphics or charts.
- Heat maps that display two variables as different intensifying colors.
- Decomposition trees, a visualization that displays each drill-down akin to an ever-expanding organization chart.
- Geographic maps that display things such as sales figures in a map form, using color to highlight sales performance. By mousing over a particular country, region, or state, you can see the individual data values.

**12. What is Rich reportlets?**

- Rich reportlets are powered by Web 2.0 technologies to create rich Internet applications (RIA). When BI suites were first re-architected for the Web, report consumers could only view a static page.
- With rich reportlets, someone accesses a report over the Web but in a much more interactive and appealing way.
- At a simple click, data can be re-sorted, filtered, or graphed, without having to launch a complicated report editor.

**13. What are the Future Trends of Business Intelligence?**

Looking beyond the obvious trends (Social, Mobile, Cloud), would like to focus on three trends brewing at the tectonic levels of the BI industry.

Business intelligence tools need to be, (i) Simplified (ii) Specialized (iii) Personalized.

**14. Define the term Decision management?**

Decision management is the term industry experts and vendors use to describe the integration of analytics with business rules and process management systems to achieve a predictive and proactive posture in a real-time world.

**15. Write the Scope of BI Search?**

The list of files, documents, reports, and systems indexed by a BI search implementation constitute its scope. The scope of the implementation determines many things, including what's visible through the search index (all else is invisible) and what elements of the indexed source are included in the index.

**16. What are the machine learning strategies?**

Machine learning can be done by applying specific learning strategies, such as:

- A supervised strategy to map the data inputs and model them against desired outputs.
- An unsupervised strategy, to map the inputs and model them to find new trends.

**17. What is Machine Learning?**

In simple terms, machine learning is a branch of the larger discipline of Artificial Intelligence, which involves the design and construction of computer applications or systems that are able to learn based on their data inputs and/or outputs. Basically, a machine learning system learns by experience; that is, based on specific training, the system will be able to make generalizations based on its exposition to a number of cases and then be able to perform actions after new or unforeseen events.

**18. What are the three components of learning algorithms?**

Learning = Representation + Evaluation + Optimization

- **Representation** means the use of a classifier element represented in a formal language that a computer can handle and interpret;
- **Evaluation** consists of a function needed to distinguish or evaluate the good and bad classifiers;
- **Optimization** represents the method used to search among these classifiers within the language to find the highest scoring ones.

**19. Difference between artificial and business intelligence?**

| Artificial intelligence | Business intelligence |
|---|---|
| This subject deals with Logic, Reasoning,Graph raversing/Mining etc. It deals with automatic ways of reasoning and reaching to a conclusion by computers. | This field is even more business goal focused than Data Mining. |
| Some of the algorithms of this domain are BFS, DFS, A*, Djikstra, Best First, Backtracking etc. | There is no algorithm for this rather it is a skill of converting raw data to useful information for the business. |
| Search and Optimization are two big use cases of AI. | Data sources could be internet search, internal business sources, web click data, customer feedback data etc. |
| Robot Navigation, Automatic Clinical Decision System, Knowledge representation | Business dashboards, Reports, SWOT Analysis |

**20. What is Event-Streaming?**

Event-streamed data is the term for Big Data generated from sensors, instruments, and other monitoring systems that constantly produce data. Such data is frequently related to weather, global positioning, video monitoring, etc. Although sensor data is already used in industry specific applications such as oil mining or in retail stores via smartphones, the analytics and BI industry is preparing to make tools for such data more commonplace.

**21. List the application areas of text mining. (APRIL/MAY 2017)**

- Cybercrime prevention
- Customer care service
- Fraud detection
- Contextual Advertising

**22. What are emerging technologies in BI?**

- data visualizations,
- BI on clouds or SaaS,
- Hadoop,
- Apache Spark and Shark,
- mobile BI,
- social media,
- Internet of things,

**23. Define discriminant analysis. (APRIL/MAY 2017)**

Discriminant analysis is a statistical tool with an objective to assess the adequacy of a classification. It is often used to complement the findings of cluster analysis and principal component analysis**.**

**24. Define information extraction. (NOV/DEC 2017)**

Extraction of information from a text in the form of text strings and processed text strings which are placed into slots labeled to indicate the kind of information that can fill them. Convert the benefits of powerful query tools such as SQL. This method of getting meaning from text is called Information Extraction.

**25. What is morphological analysis? (NOV/DEC 2017)**

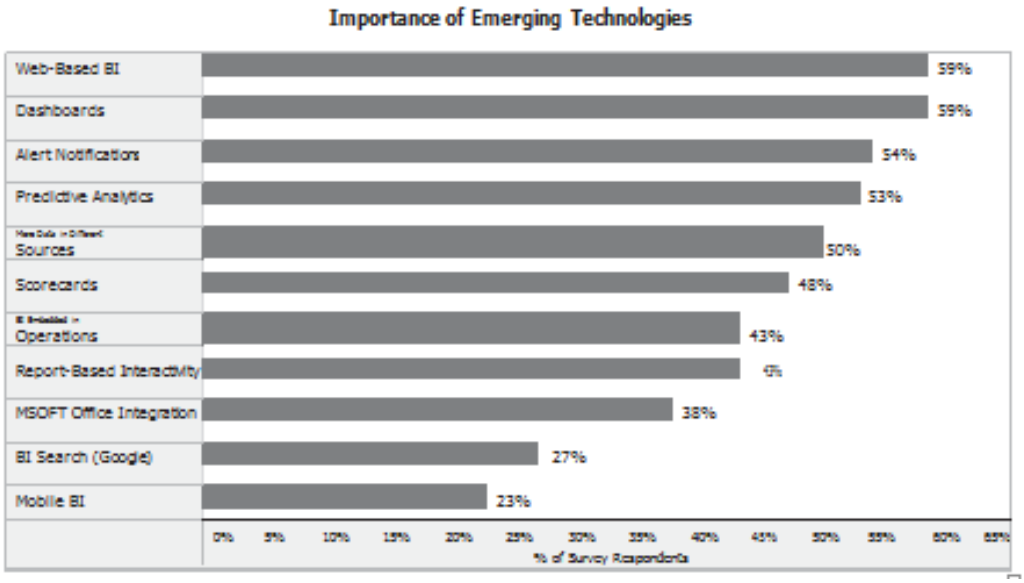
Morphological analysis is a method developed by Fritz Zwicky for exploring all the possible solutions to a multi-dimensional, non-quantified complex problem. Consider a complex, real-world problem, like those of marketing or making policies for a nation, where there are many governing factors, and most of them cannot be expressed as numerical time series data, as one would like to have for building mathematical models

## PART-B

**1. Explain in detail the emerging technologies in future of business intelligence.**

As part of the Successful BI Survey, respondents were asked to choose items from a list of emerging technologies that they believe will help their companies achieve greater success.

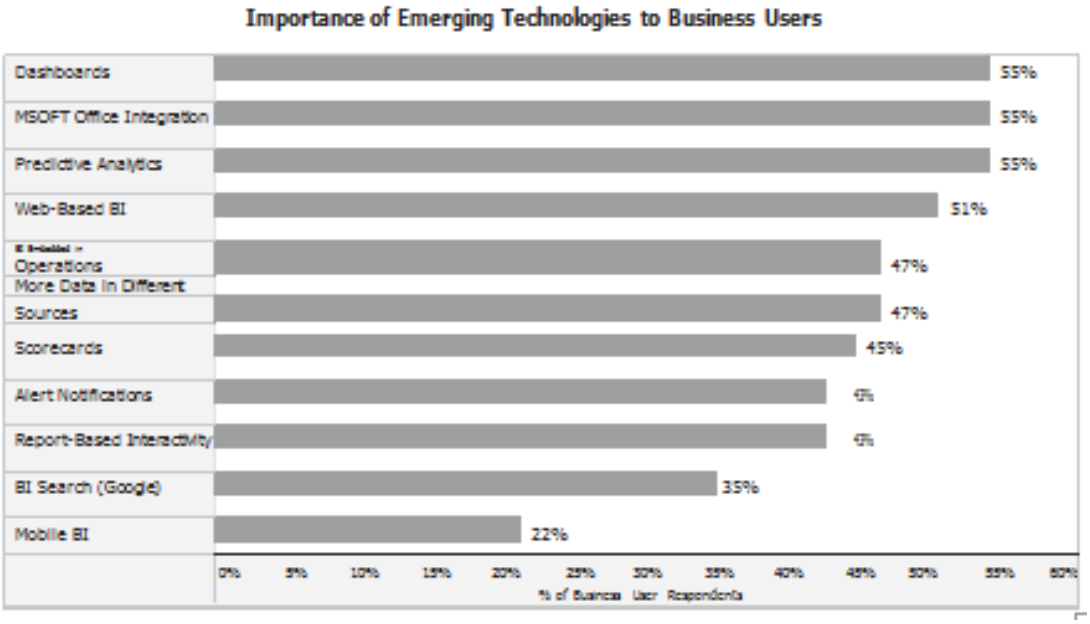
Figure shows which items are considered most important in helping companies achieve greater success. Web-based business intelligence and dashboards were rated the highest, with predictive analytics and alerting also at the top. Surprising to me, Microsoft Office Integration, BI Search, and Mobile BI were selected by only a small percentage of survey respondents. The view according to business users, however, is slightly different, as shown in Figure. Business users account for only 10% of the survey respondents. Those who describe themselves as hybrid business-IT personnel account for 23% of respondents. When viewing responses only for business users, the importance of Microsoft Office integration moves to the top of the list, while alerting moves down.



Some of these differences can be explained by gaps in understanding of the feature benefits, but also by a respondent's point of view. For example, IT professionals have been burned in the past by the thousands of disconnected spreadsheets and the ensuing data chaos. As Microsoft Office integration with BI has improved dramatically in 2007, IT professionals may not realize that spreadsheet-based analysis can now be "safely" enabled and can be something to be embraced for knowledge workers familiar not just with Microsoft Excel, but also with Word, Outlook, and PowerPoint. In a similar fashion, if you are a BlackBerry user, you may rate Mobile BI high.

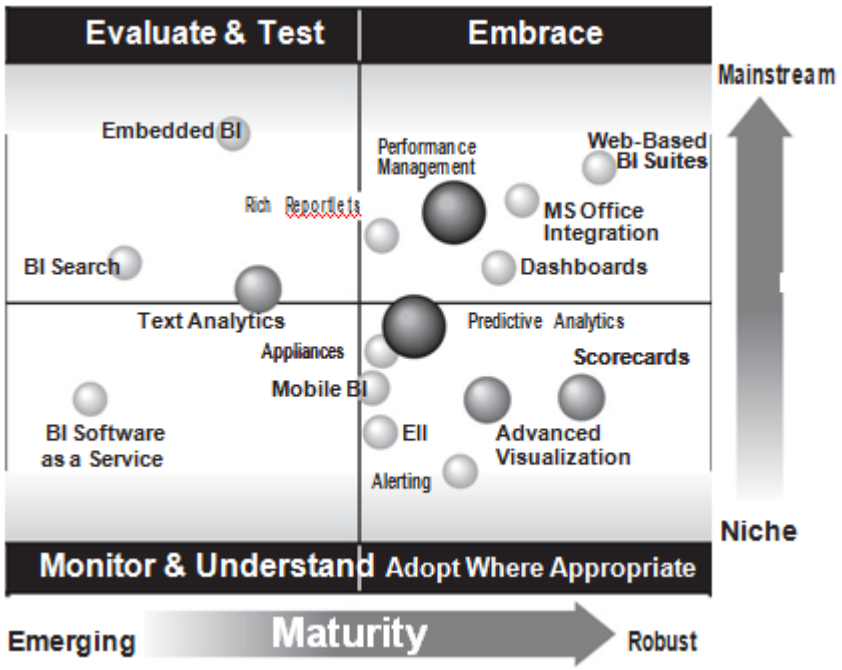
While web-based business intelligence may have been introduced in the late 1990s, these solutions only reached the rich functionality of their desktop predecessors in the 2005 time frame. A number of companies are not yet on the latest releases, though, and still use client/server BI deployments. Depending upon where a survey respondent is in their web-based BI deployment will influence how this capability was rated.

At The Data Warehousing Institute's (TDWI) Executive Summit in February 2007, I participated in a panel on the role of emerging technologies in extending the reach and impact of business intelligence. Attendance was restricted to BI directors and executive sponsors who influence their company's BI strategy. Attendees could vote on a limited number of items that they thought would have the biggest impact in the next few years. The most highly ranked item: performance management and predictive analytics. The things that got few to no votes were BI search, dashboards, and rich Internet applications, contrary to what I believe will have the biggest impact. As we delved into what these technologies mean, and in some cases, demonstrated them, the perceptions changed significantly. In this way, we sometimes don't know the impact any of these capabilities will have until the technology has become more mature and the industry understands it better. If you think about the way breakthroughs like the iPod and YouTube have revolutionized their markets, when they first were introduced, they were met with a mixture of fascination and confusion, without a clear understanding of where they would lead. Recent BI innovations must go through a similar process of the industry first understanding their potential, accepting or rejecting them, and then either embracing or adopting the innovations in a limited fashion. Another fundamental difference in evaluating impact is whether the impact is measured by the value of one user or decision or aggregated by multiple users and decisions. Thousands of individual decisions can have as big, if not bigger, an impact on a company than a single decision. There seems to be a natural tendency to rate capabilities that have a single big impact as being more important or more likely to help achieve greater BI success.



**2. Explain in detail the framework of BI technology evaluation.**

The below Figure provides a framework for evaluating changes in BI technology to determine which new and emerging capabilities will prove most valuable to your company, how mature they are, and when to monitor them or when to embrace and actively deploy them (adapted from TDWI's Technology Evaluation Framework). The X axis provides an indication of how mature the technology is, and the Y axis gives an indication of which technology will make BI pervasive. Recall from the section on BI users as a percentage of employees in Chapter 4 that the average usage of BI within a company is currently at 25%, and even if budget were available and the deployment were wildly successful, survey respondents felt the use rate would extend only to 54% of employees.




The Y axis, then, indicates the degree to which an enabling technology will take BI's reach closer to 100% of employees. Business impact and BI prevalence are not linearly correlated, however. One enabling technology, such as predictive analytics, may yield a big value for a single decision, say, a $4 million savings by better marketing campaign management. Another enabling technology such as BI embedded in operational processes may affect thousands of users, each of whom makes dozens of decisions on a daily basis; the monetary value of these individual decisions may be small when measured in isolation, but enormous when taken in aggregate. The size and shading of the bubbles in the above figure give an indication of which items have a bigger single value. For each innovation, consider both the technical maturity and the business impact to decide how to proceed:

- **Embrace** Items in the upper-right quadrant show innovations that are mature and that should be embraced as they will help speed user adoption across multiple user segments.
- **Adopt Where Appropriate** Items in the lower-right quadrant show innovations that are mature but that may serve only specific segments of users. Mobile BI is an example of this; the technology is more mature than BI search, for example, but benefits only those users who have smartphones such as a BlackBerry.
- **Evaluate and Test** Items in the upper-left quadrant are relatively new but will have a profound impact on user adoption. BI Search is a good example of this. The technology is very new and not well under-stood. A number of usability and performance issues still need to be worked out, but the potential impact on user adoption is enormous.
- **Monitor and Understand** Items in the lower-left quadrant are so new that they may be riskier investments. Items here are less proven and have less market adoption.

**3. Explain in detail predicting the future using data mining algorithms with business intelligence technology.**

Data mining, statistical analysis, and predictive analytics are nothing new. These technologies are well established and are used in a number of different applications such as fraud detection, customer scoring, risk analysis, and campaign management. What's changed is how they have become integrated in the BI platform. Traditionally, predictive analytics has been a backroom task performed by a limited few statisticians who would take a snapshot of the data (either from a data warehouse or from a purpose-built extract from the source system), build a model, test a model, finalize it, and then somehow disseminate the results.

While the expertise to build such models remains a unique skill set, the industry recognizes that the results of the analysis should be more broadly shared, not as a stand-alone application, but rather, as an integral part of the BI solution. This does not mean that predictive analytics software will become "mainstream," but rather that the results of such analyses can be readily incorporated into everyday reports and decision making. The analysis, then, is what needs to become mainstream.

Predictive analytic tools from different vendors do continue to differ significantly in how they work and in what information is stored in the database versus calculated and presented in a report or incorporated into an operational process. At Corporate Express, for example, predictive analytics are being used to improve customers' online shopping experience. Market basket analysis helps retailers understand which products sell together and provide product recommendations. In the past, Corporate Express provided these recommendations by logical product pairings. So if a customer ordered a stapler, the online store would recommend a staple remover as the marketing team had marked this as a complementary product.

In analyzing the data, though, it turned out that what was most often purchased with a stapler was not a staple remover, but rather a ruler, tape dispenser, and a wastepaper basket—items that indicate a purchase for a new employee. With the manually associated product recommendations, there was no significant impact on sales. Leveraging MicroStrategy and SPSS, Corporate Express tested a new market basket option. The model analyzes past shopping carts and produces recommendations to ensure the greatest lift. As a result, the average order size for market basket pairings doubled (versus those orders with no pairings), and the market basket application is expected to generate an incremental gross profit of more than $2 million in 2007. Dow Chemical also has begun extending the reach of predictive analytics with SAS's JMP product (pronounced "jump"), a solution that combines visual analysis with statistics. As discussed in Chapter 12, Dow uses BusinessObjects and Cognos PowerPlay as enterprise report-ing and analysis standards. Through these tools and the data in the data warehouse, Dow began looking at the high cost of railroad shipments: $400 million annually across North America.

A team of statistical experts studied the variables that most affected these costs and pulled data from the data warehouse and external data sources into SAS JMP. By benchmarking current payments versus industry norms, the analysis showed Dow was overpaying by 20%, or $80 million. In entering new contracts, the purchasing department now uses the software to predict appropriate rates, enabling them to negotiate more aggressively. For both Corporate Express and Dow Chemical, the move to predictive analytics has been evolutionary. The underlying information architecture and a culture of fact-based decision making had to first reach a level of maturity and data quality before predictive analytics could be embraced.

While both companies have been doing statistical analysis for decades, the degree to which predictive analytics has now been incorporated into daily processes (online store at Corporate Express and purchasing negotiations at Dow Chemical) reflects the degree to which predictive analytics has shifted from the backroom to the front line, with the most casual of users deriving value from such analytics.

**4. Discuss in detail BI search and Text Analytics. (APRIL/MAY 2017)( NOV/ DEC '17)**

BI Search offers a number of promising benefits to business intelligence:

- Simple user interface.
- A more complete set of information to support decision making, with the integration of structured (quantitative) and unstructured content (textual). Structured data refers to the numerical values typically captured in the operational systems and subsequently stored in a data warehouse. Unstructured content refers to information stored in textual comment fields, documents, annual reports, websites, and so on
- Users can find what they need through search, rather than through navigating a long list of reports.

Text analytics is closely related to search in that unstructured information or text can be transformed into quantitative data. For example, it allows for searching of information in a comment field to see how many times a customer praised a particular product. Text analytics is the numerical analysis of textual information. Despite all the improvements in data warehousing and BI front-end tools, users continue to feel overwhelmed with reports yet undersatisfied with meaningful information. They don't know what's available or where. Similar reports are created over and over because users don't know which reports already exist or how, for example, the report "Product Sales" differs from "Product Sales YTD." Some of the most valuable information is hidden in textual data.

A BI Search interface promises to change the way users access information. Picture a Google interface to BI. Without any training in a BI tool, users can enter a phrase such as "Recent sales for customer A" and then be presented with either a list of predefined reports or, in some cases, a newly generated query. The added benefit is that in addition to displaying reports coming from the BI server, the search engine will also list textual information that may be relevant—a customer letter, sales call notes, headline news. When search capabilities are combined with text analytics, a report may include numerical data that scans the comment field to indicate number of complaints with number of positive comments. Never before has such unstructured data been so nicely accessible with structured or quantitative data.

If the integration of search and BI is successful, it is yet another innovation that will make BI accessible and usable by every employee in an organization. According to Tony Byrne, founder/president of CMS Watch, a technology evaluation firm focusing on enterprise search and content management systems, search as a technology has existed for more than 50 years. Consumer search (Google and Yahoo, for example) as a technology emerged with the Internet in the mid-1990s. In many respects, the success of consumer search has helped spur hype around enterprise search, in which companies deploy search technology internally to search myriad document repositories. Text analytics has existed for 25 years but with usage in limited sectors, particularly, the government. The convergence of search with business intelligence first emerged in 2006. Google is not the only enterprise search solution that BI vendors support but it is one that has the most consumer recognition and thus has helped business users to understand the possibilities. To illustrate the point, note that BI search was selected by only 27% of the Successful BI Survey respondents as a capability that would help foster greater success.

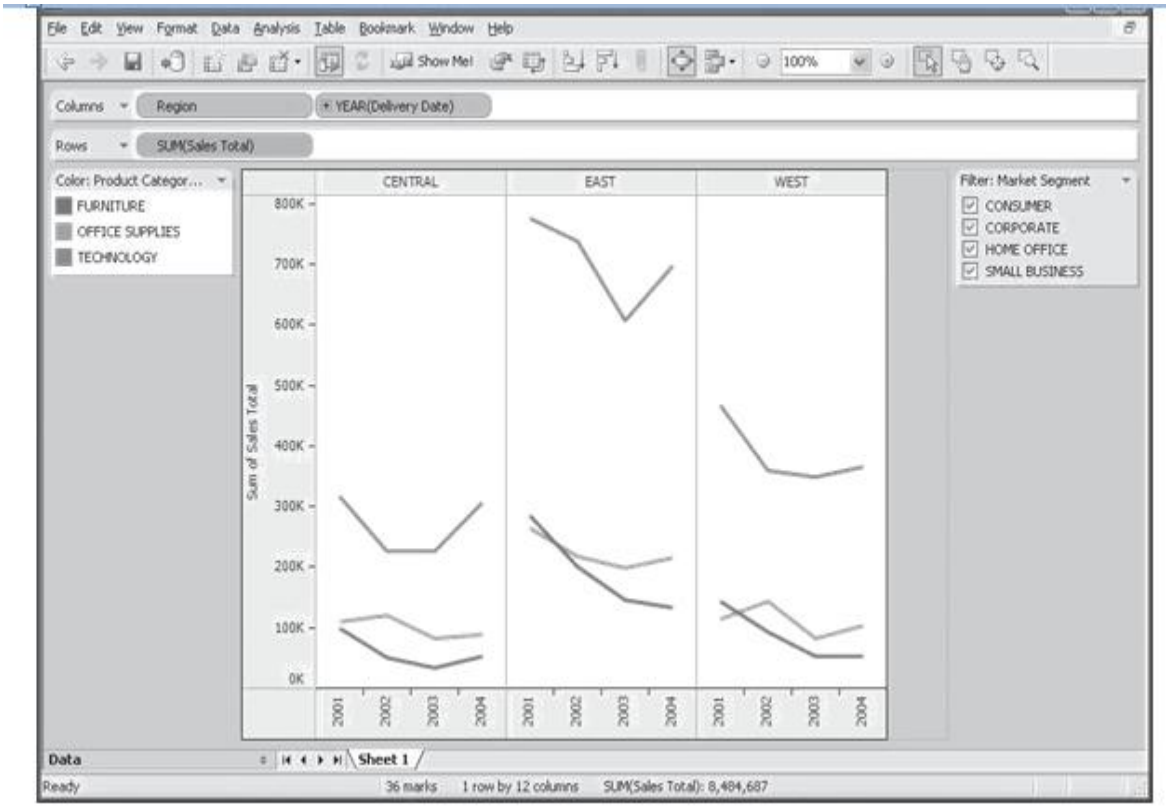
The number of customers taking advantage of the BI Search and text analytics integration is only a handful. BlueCross BlueShield (BCBS) of Tennessee (TN) is an early adopter of these capabilities. BCBS of TN is a not-for-profit provider of health insurance. In 2006, it paid $17 billion in benefits for its 2 million commercial members. Managing claims and negotiating rates with providers is critical in ensuring BCBS can meet its obligations to the members it insures. While the insurer has had a mature business intelligence deployment for ten years, Frank Brooks, the senior manager of data resource management and chief data architect, recognized that there was value in bringing the text data stored in comment fields from call center notes together with information in the data warehouse.

**5. Explain in detail Advanced Visualization with suitable examples. (APRIL/MAY 2017)**

All leading BI tools have basic visualization capabilities: you can take tabular data and turn it into a bar chart, trend line, and so on. They also support some kind of conditional formatting of data: display positive numbers in green, display negative numbers in red, and enlarge those with the worst variances. Advanced visualization goes beyond a simple chart such as a bar or line chart to include things as:

(i) Spark lines, a highly condensed trend line the size of a word.
(ii) Bullet graphs, a construct by Stephen Few that includes a target indicator within the bar chart.
(iii) Small multiples, which are series of small, similar graphics or charts. As they use the same scale and are positioned side-by-side, they facilitate visual discovery by letting users make comparisons at a glance.
(iv) Heat maps that display two variables as different intensifying colors.
(v) Decomposition trees, a visualization that displays each drill-down akin to an ever-expanding organization chart.
(vi) Geographic maps that display things such as sales figures in a map form, using color to highlight sales performance. By mousing over a particular country, region, or state, you can see the individual data values.

Advanced visualization software and capabilities also help you apply best practices in data visualization, even for basic visualizations. As an example, many reports today are designed as a dense page of numbers. The dense page of numbers may not help facilitate insight, but they are what users are accustomed to. With visualization software, a dense set of numbers can quickly be converted to a more meaningful display. Figure shows several charts created in Tableau Software. By displaying multiple graphs side-by-side, as "small multiples," you can more easily see which product category technology in this case is the top-selling product



**6. (i) Discuss in detail on Rich Reportlets.**

Report-based interactivity is a ho-hum term that warrants a better name. "Active reports," "on-report formatting," and "navigable reports" are similar terms that also don't fully capture the value of this capability. I suspect poor terminology and lack of awareness also explains why survey respondents ranked this capability on the low end of importance for emerging technologies. So, after much thought and brainstorming, will refer to this capability as "rich reportlets." The difference in power and appeal with rich report-lets versus, say, green-bar paper reports and much of what is currently deployed over the Web, is comparable to the difference between a black Ford Model T and a red Mercedes sports coupe.

Rich reportlets are powered by Web 2.0 technologies to create rich Internet applications (RIA). When BI suites were first re-architected for the Web, report consumers could only view a static page. Given how static a display this was, more sophisticated users would export the data to Excel for analysis. Less sophisticated users would submit requests to IT or to the BI team to modify the report design. The Web in this case is only a delivery vehicle for data; it does not facilitate user adoption and insight. With rich reportlets, someone accesses a report over the Web but in a much more interactive and appealing way. At a simple click, data can be re-sorted, filtered, or graphed, without having to launch a complicated report editor. With the use of either Adobe Flex or Macromedia Flash, these reports come to life in ways that make business intelligence fun. I have seen, for example, a bubble chart that displays bubbles dancing across the screen as the time axis marches onward. Such animation makes BI appealing as well as insightful as users see the trend in action. In this regard, the term "report" doesn't do justice to the capability that is more akin to a mini application.

This type of interactivity affects all BI users, whether casual or power users. The appeal makes BI more engaging, and while some technologists may scoff at the importance of this, when other barriers to adoption exist, appeal matters. A lot! The ability to interact with the data in a simple and intuitive way facilitates greater insight at the hands of the decision maker. The report consumer is not forced to delay this insight until a power user can modify the report. Lastly, the cost of ownership is lowered because a single reportlet can be "tweaked" to that decision maker's needs, without IT having to maintain thousands of individualized reports.

**(ii) Discuss the future Beyond Technology.**

Technical innovation is only one aspect that will help increase BI's prevalence. In discussing future plans with many of the case study companies, much of their concern was not about technology, but rather, in finding new ways to use BI to address common business problems. For the more large-scale deployments, some expressed concern about man-aging the risk of making any kind of major change to such a business critical, complex application. With success, of course, comes greater demands on the systems and the people. Ensuring an effective way of prioritizing competing requests warrants constant attention. One business leader expressed frustration at his department's inability to make wise investments, while witnessing other departments, working in more unison and getting more value from business intelligence. Yet he remains optimistic that his business will get there and that BI will be the first thing people look at, even before email. "To have one screen I can get to with a single click, that shows sales, margin, price, opportunities in graphical form, with drill down—that would be magic!" His comments remind me that the technology is sometimes the easy part; getting the organization aligned is harder. Even the most successful BI companies, then, continue to have their battles.

**7. Discuss the pros and cons of business intelligence technologies**

**Refer Question no 2,3 ,4 and 5**

**8. How artificial intelligence makes business more productively in the varying trends of business intelligence.**

Enterprise seems to be entering a new era ruled by data. What was once the realm of science fiction, AI in business intelligence is evolving into everyday business as we know it. Companies can now use machines algorithms to identify trends and insights in vast reams of data and make faster decisions that potentially position them to be competitive in real-time. It's not a simple process for companies to incorporate machine learning into their existing business intelligence systems, though Skymind CEO and past TechEmergence podcast guest Chris Nicholson advises that it doesn't have to be daunting. "AI is just a box," he says. "Math and code. If this, then that. That is the simplest way to describe it." Organizing data collection and testing an algorithm with this data for accuracy over the first few months are is where many businesses get stuck.

But as AI has gained momentum, prominent application providers have gone beyond creating traditional software to developing more holistic platforms and solutions that better automate business intelligence and analytics processes. Major vendors— including General Electric, SAP, and Siemens—offer such software suites and operating systems, but there are a growing number of emerging providers in the market as well.

## DOMO – AI for Business Dashboards

It is not just the giants like SAP developing machine learning platforms for business. Domo, a fast-growing business management software company that's raised over $500 million in funding, has created a dashboard that gathers information to help companies make decisions. The cloud-based dashboard can scale with the size of the company, so it can be used by teams as few as 50 or by much larger enterprises. There are more than 400 native software connectors that let Domo collect data from third-party apps, which can be used to offer insights and give context to business intelligence.

This gives companies using Domo a way to pull data from Salesforce, Square, Facebook, Shopify, and many other applications that they use to gain insight on their customers, sales, or product inventory. For instance, Domo users who are merchants can extract data from their Shopify point-of-sale and e-commerce software, which is used to manage online stores. The extracted information can be used to generate reports and spot trends in real-time, such as in product performance, which can be shared to any device used by the company.

In March, Domo announced Mr. Roboto, a set of new features for the platform that draw upon AI, machine learning, and predictive analytics. The expectation is for Mr. Roboto to offer recommendations and insights to decision makers at companies. Once these features are rolled out, expected in late spring 2017, the platform is supposed to issue new alerts and notifications for significant changes, such as the detection of anomalies or new patterns in data (similar to approaches used in cyber security already).

Detecting these changes and patterns is expected to fuel the predictive analytics side of Mr. Roboto and help companies predict the return on investment for marketing in real-time, customer churn, and sales forecasts.

## Apptus – AI in Sales Enablement

There are numerous ways for machine learning to enhance applications, including those from Apptus, which offer recommendations on actions that companies can take to boost their sales channels. Apptus says it specializes in the connection between a customer's intent to buy and the realization of revenue by a company. The Apptus eSales solution is designed to, among other features, automate merchandising based on a predictive understanding of consumers. The software combines big data and machine learning to determine which products might appeal to a potential customer as they search online or get recommendations.

For example, when a customer visits an online store that uses Apptus eSales and starts to type in search terms to look up products, the machine learning solution can predict and automatically display related search phrases. It can also display products associated with those search terms.

This is a potential threshold moment for business and industry, where machine learning might weave its way further into how operations are handled, the way decisions are made, and resources get managed. It will depend on whether or not businesses collectively find real value in AI; the investment in the technology must prove its worth.

Nicholson notes that though the accuracy and capabilities of deep learning have increased, the technology is still trickling out into the world among early adopters. The next phase, he says, will be about whether or not such resources will flow more freely and be embraced by the business community at large. "The future is already here," he says, citing cyberpunk Author William Gibson, "It's just not widely distributed."

**9. Explain Internet of Things (IOT) in terms of business intelligence**.

The Dresner analyst team delivers regarding the intersection of big data and the Internet of Things (IoT), big data adoption, analytics, and big data distributions. The report also provides an analysis of Cloud Business Intelligence (BI) feature requirements, architecture, and security insights. IoT adoption is thoroughly covered in the study, with a key finding being that large organizations or enterprises are the strongest catalyst of IoT adoption and use.

Mature BI programs are also strong advocates or adopters of IoT and as a result experience greater BI success. IoT advocates are defined as those respondents that rated IoT as either critical or very important to their initiatives and strategies. The combined rankings of IoT as critical and very important are highest for sales, strategic planning and the Business Intelligence (BI) Competency Centers. Sales ranking IoT so highly is indicative of how a wide spectrum of companies, from start-ups to large-scale enterprises, is attempting to launch business models and derive revenue from IoT. Strategic planning's prioritization of IoT is also driven by a long-term focus on how to capitalize on the technology's inherent strengths in providing greater contextual intelligence, insight, and potential data-as-a-service business models.

**10. Explain the four steps of case based reasoning(CBR) process and Write the various benefits of case based Reasoning.** (**APRIL/MAY 2017)( NOV/ DEC '17)**

Case–based reasoning (CBR) is a knowledge–based problem solving technique that is based on reuse of past experience. Today, many CBR systems are available and successfully running. Maintenance of case–based reasoning systems is a hot topic in CBR research and application. Many aspects of case–based reasoner maintenance have been explored mainly concentrating on case base maintenance. Software system maintenance, in general, has to deal with three types of maintenance : corrective, adaptive, and perfective maintenance. And — to add another level of difficulty — in knowledge–based systems we do not only have to deal with the maintenance of the case–based reasoning application itself we also have to deal with the maintenance of the system's knowledge. But in this paper we concentrate on the maintenance of the knowledge.

The two models are complementary and represent two views on case–based reasoning. The first is a dynamic model that identifies the main sub processes of a CBR cycle, their interdependencies and products. The second is a task–oriented view, where a task decomposition and related problem solving methods are described. Aamodt and Plaza identified and discussed important problem areas of CBR, and means of dealing with them. All task–decompositions are complete, i.e., the set of subtasks of a task are intended to be sufficient to accomplish the task. The four processes retrieve, reuse, revise, and retain describe the general tasks in a case–based reasoner. They provide a global external view to what is happening in the system.

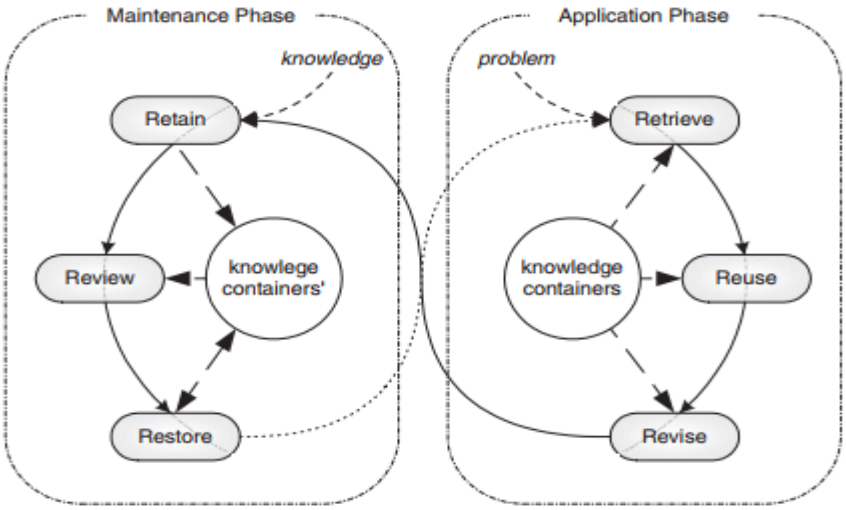
The four tasks are decomposed into a hierarchy of CBR tasks the system has to achieve. This task–oriented view is suitable for describing the detailed mechanisms from the perspective of the case–based reasoner itself.

Aamodt and Plaza call the top level task of CBR problem solving and learning from experience which directly matches our two phases application and maintenance.

**(i) Application Phase :** The application phase consists of the three steps retrieve, reuse, and revise. In the first step the case–based reasoner retrieves the most similar case or cases. Then, it reuses the information and knowledge in that case and proposes a solution. If the solution is rejected the CBR system revises the proposed solution. This far the systems stays unchanged. If the revised solution is stored for later use during the retain step the case–based reasoner enters the maintenance phase.

**(ii) Maintenance Phase:** The maintenance phase in the original CBR cycle consists of the retain step alone. During this phase the revised solution is stored in the case base.1 The indexing structures to the newly added case are adapted, too. If no new case was constructed existing indexing structures are modified to improve the similarity assessment.

The proposed six step process model consists of the original four steps retrieve, reuse, revise, and retain, and the two novel steps review and restore.




**(i) Retrieve** : The retrieve step provides information through its result, namely, how often a case is retrieved. Also, to improve performance the mean retrieval time may be stored. Maximum and minimum may also be of interest. Other usage information are the last update or last retrieval time. The queries may be logged here, too. In the maintenance phase the queries could be used to enhance the vocabulary, the similarity measures or the adaptation knowledge. They even may be used to seed new cases. Comparisons of retrieval results of different versions of the case–based reasoner regarding the stored queries may yield interesting insights into the behavior of the system over time.

**(ii) Reuse** : The reuse step adapts the retrieved cases to solve the given problem. An adaptation counter can count which of the retrieved cases could be adapted to the query and is therefore suggested as a possible solution.

**(iii) Revise**: If the suggested solution is rejected this may be counted as well as how often the case has been accepted as a solution. This information can help to decrease the case base size because it may be more favorable to remove an often rejected case from the case base than accepted cases.

**(iv)Retain**: This step adds adapted cases to the system. The user then has a choice between confirmed and unconfirmed cases. A second purpose of the retain step is to modify the similarity measures by modifying the indexing structures. Modifications like that should only be used by the case–based reasoner if there are possibilities to track the changes, or better measure the impact of those changes.