JOMO KENYATTA UNIVERSITY OF AGRICULTURE & TECHNOLOGY

SCHOOL OF OPEN, DISTANCE AND eLEARNING

MODULE NOTES FOR DISTANCE LEARNERS

BIT 2210 BUSINESS INTELLIGENCE

Dr. J. Okello
(masenooj@gmail.com)

P.O. Box 62000, 00200
Nairobi, Kenya

BIT 2210 BUSINESS INTELLIGENCE

Course description

The essential of business intelligence; Origins and drivers of business intelligence; Major characteristics of business intelligence; Toward competitive intelligence and advantage; Structure and components of business intelligence. Data warehouse as basis of business WI-2-58-4-3(2/40/2/1) B. Sc. in Information Technology Intelligence; Data warehouse definitions and concepts; Data warehouse architecture; Data integration and the extraction, transformation and load process. Business analytics and data visualization; Online analytical processing; Reports and queries; Multidimensionality; Advanced business analytics; Data visualization; Geographic information systems and business analytics. Data, text and web mining; Data mining concepts and applications; Data mining techniques and tools; Text mining; Web mining. intelligence; Data warehouse definitions and concepts; Data warehouse architecture; Data integration and the extraction, transformation and load process. Business analytics and data visualization; Online analytical processing; Reports and queries; Multidimensionality; Advanced business analytics; Data visualization; Geographic information systems and business analytics. Data, text and web mining; Data mining concepts and applications; Data mining techniques and tools; Text mining; Web mining.

Prerequisite: ICS 2405Knowledge Based Systems

Course aims

Students will learn the theoretical background of business intelligence and will practice their knowledge using business intelligence tools in the lab.

Learning outcomes

Upon completion of this course you should be able to;

1. Discuss the origin and drivers of Business Intelligence, and evaluate its major characteristics. .

2. Discuss the data warehouse architecture, its role and the importance for business intelligence

3. Design and develop applications based on the major business analytics methods and tools to support competitive intelligence

4. Create data analysis using online analytical processing, data visualization and multidimensionality to improve companies

5. Discuss different applications of data mining, and produce data analysis based on different methods of data mining

Instruction methodology

- Lectures, practical and tutorial sessions in Computer Laboratory, individual and group assignments, exercises and project work

Instructional Materials/Equipment

Overhead projector and computer, handouts, white boards, Textbooks, appropriate software.

Course Text Books

1. Ramez Elmasri, Shamkant B. Navathe (2006). Fundamentals of Database Systems (5th ed.). Addison Wesley. ISBN: 0321369572.

2. Thomas M. Connolly, Carolyn E. Begg (2004). DataBase Systems: A Practical Approach to Design, Implementation and Management (4th ed.). Addison Wesley. ISBN: 0321294017

3. Margaret H. Dunham (2003). Data Mining: Introductory and Advanced Topics. Prentice Hall. ISBN-10: 0130888923

Reference Textbooks

1. Patricia Ward, George Dafoulas (2006). Database Management Systems. Thomson Course Technology. ISBN: 1844804526

2. Philip Lewis, Arthur Bernstein, and Michael Kifer (ISBN: 0-201-70872-8).

3. Databases and Transaction Processing: An Application-Oriented Approach. Addison- Wesley. Connolly, Begg, and Strachan (1998). Database Systems (2nd ed.). Addison Wesley ISBN 679-974689

Course Journals

1. International journal for information and communication technology research ISSN 6790456
2. Computer science journal ISSN 0348-70067
3. Acta Informatica ISSN 0001-5903
4. Advances in Computational Mathematics ISSN 1019-7168 5. Advances in data Analysis and Classification ISSN1 1862-5347 6. Annals Of software Engineering ISSN 1022-7091

Reference Journals

1. Journal of computer science and technology ISSN- 0916-534276
2. Journal of Computing ISSN-1002-784536
3. Directory of Open Access (DOAJ): Computer Science, ISSN 0034-567352 WI-2-58-4-3(2/40/2/1) B. Sc. in Information Technology
4. Journal of Science and Technology ISSN 1860-4749

Assessment information

The module will be assessed as follows;

- 30% Continuous Assessment (Tests 10%, Assignment 10%, Practical 10%) 70%
- End of Semester Examination.

Contents

LESSON 1

Business intelligence: an integrated approach

- What is business intelligence?
- Origins and drivers of business intelligence
- Major characteristics of business intelligence
- Why is business intelligence important
- What are the goals of business intelligence
- BI Framework

While the business world is rapidly changing and the business processes are becoming more and more complex making it more difficult for managers to have comprehensive understanding of business environment. The factors of globalization, deregulation, mergers and acquisitions, competition and technological innovation, have forced companies to re-think their business strategies and many large companies have resorted to Business Intelligence (BI) techniques to help them understand and control business processes to gain competitive advantage. BI is primarily used to improve the timeliness and quality of information, and enable managers better understand the position of their firm as in comparison to competitors. BI applications and technologies help companies to analyze changing trends in market share; changes in customer behavior and spending patterns; customers' preferences; company capabilities; and market conditions.

It is used to help analysts and managers determine which adjustments are most likely to respond to changing trends. It has emerged as a concept for analyzing collected data with the purpose to help decision making units get a better comprehensive knowledge of an organization's operations, and thereby make better business decisions.

BI is an area of Decision Support System (DSS) that which is an information system that can be used to support complex decision making, and solving complex, semi-structured, or ill-structured problems. The first reference to BI was made by Lunh (1958), which has replaced other terms such as Executive Information Systems and

Management Information Systems (Negash, 2004; Turban et al., 2008; Thomsen, 2003).

Although BI is a type of DSS, but it often has a broader meaning. It is the process of gathering high-quality and meaningful information about the subject matter being researched that will help the individual(s) to analyze the information, draw conclusions or make assumptions (Jonathan, 2000).

BI is the process of taking large amounts of data, analyzing that data, and presenting a high-level set of reports that condense the essence of that data into the basis of business actions, enabling management to make fundamental daily business decisions. BI is the way and method of improving business performance by providing powerful assistance to executive decision maker which enables them to have actionable information at hand. BI tools are viewed as technology that enhances the efficiency of business operation by providing an increased value to the enterprise information and hence the way this information is utilized.

It is the process of collection, treatment and diffusion of information that has an objective, the reduction of uncertainty in the making of all strategic decisions. It is a "business management term used to describe applications and technologies which are used to gather, provide access to analyzed data and information about an enterprise, in order to help them make better informed business decisions."

1.1. Goals of business intelligence

1. **Timeliness:** the data within system should be synchronized with all other applications.

2. **Accuracy:** the data should encompasses every data from any other application;

3. **Acceptance**: the users convinced of timeliness and accuracy of data should be able to actively use the system as support for decision making.

The rapidly changing business factors such as globalization, deregulation, mergers and acquisitions and technological innovation, have forced companies to re-think their business strategy. In this competitive environment, BI plays an important role in supporting of the decision making process to augment competitiveness, marking an efficient link between business strategies and IT. BI technology has been continuously expanding and improving to answer more and more complex business. The most widely applied BI enabling technologies, that has emerged include data warehousing (DW), on-line analytical processing (OLAP), and data mining (DM). BI technology aims to help people make “better" business decisions by making accurate, current, and relevant information available to them when they need it. Competitive organizations accumulate BI in order to assess environment to gain sustainable competitive advantage, and may regard such intelligence as a valuable core competence in some instances.

1.2. BI Framework

Although BI is adapted by organizations as per their requirements, history, environment etc to make informed, valuable customer oriented decisions. The main approaches are:

1. The traditional approach to BI is concerned with, data aggregation, business analytics and data visualization According to this approach, BI explores several technological tools, producing reports and forecasts, in order to improve the efficiency of the decision making. Such tools include Data Warehouse (DW), Extract-Transform and Load (ETL), On-Line Analytical Processing (OLAP), Data Mining (DM), Text Mining, Web Mining, Data Visualization, Geographic Information Systems (GIS), and Web Portals.

2. On the next level there is a concern with the integration of business processes on BI. According to this approach BI is a mechanism to bridge the gap between the business process Management to the business strategy. In addition to all the tools in traditional BI, tools such as Business Performance Management (BPM), Business Activity Monitoring (BAM), Service-Oriented Architecture (SOA), Automatic Decision Systems (ADS), and dashboards, are included.

3. Adaptive BI is concerned with self-learning adaptive systems, that can recommend the best actions, and that can learn with previous decisions, in order to improve continuously. Intelligence is incorporated on BI systems in this manner. However, the general framework for understanding and guidance of practitioners, academicians and researchers is presented here. The concept of BI can be decomposed into three parts:

(a) Data Capture/Acquisition,

(b) Data Storage and.

(c) Data Access & Analysis.

Data is collected from internal as well external sources. Internal sources of data are organizations operational database and data warehouse. External data sources include the data from customers, suppliers, government agencies, competitors, internet etc. The collected heterogeneous data is stored in a data warehouse after extract, transform and load (ETL) processes. . Finally the data stored in the data warehouse analyzed for decision making.

- Data Capture/ Acquisition

The acquisition component is the back end of the data warehousing system and consists of systems that have interface with the operational systems to load data into the data warehouse. Data is first entered or processed by a daily business process that is based on Online Transaction Processing(OLTP) environment and stored in operational database, which may consist of databases such as Oracle, DB2, Informix, SQL Server, SAP R/3, etc. Before data is loaded from operational database and external sources into the data warehouse, it needs to be processed through following stages:

1. Extraction & Cleanse: During data extraction data is acquired from multiple sources including the operational systems. The selected data is consolidated and filtered out from various forms of pollution. Data cleansing validates and cleans up the extracted data to correct inconsistent, missing, or invalid values. This step applies triggers, error reports and corrective processes.

2. Transform. Data transformation integrates data into standard formats and applies business rules that map data to the warehouse schema. Aggregates (e.g., summary table data) and imputed characteristics are generated.

3. Load. Data loading loads the cleansed data into the data warehouse.

- Data Storage

After ETL data is stored in data warehouse or data marts for future analysis. Data Warehousing: Data warehouse is a copy of transaction data specifically structured for query and analysis and is informational, analysis and decision support oriented, not operational or transaction processing oriented views data warehouse as a collection of corporate information derived directly from operational systems and some external data sources. Its specific purpose is to support business decisions, not business operations. Inmon who coined the term “data warehouse” in 1990, argues that a data warehouse is a subject oriented, integrated, time-variant, non-volatile collection of data that is used primarily in organizational decision making (Inmon,1996). Data warehouses, targeted for decision support, are maintained separately from the operational databases. The architecture of data warehouse can take a variety of forms in practice. But before designing a data warehouse, the requirements and resources of the organization should be taken into consideration. However, some of the options of architecture from which organizations may choose under different circumstances may include: Data Mart; Central Data Warehouse; Distributed Data Warehouse; Virtual Data Warehouse.

Data Marts:

Data marts or localized data warehouses are small sized data warehouses, typically created by individual departments or divisions to facilitate their own decision support activities. For example, a data mart can be created for specific products or functions, like customer management, marketing, finance etc. One of the purposes to build a data mart is to get prototype as soon as possible without waiting for a larger corporate data warehouse, because it’s small and easy to develop. But after having several data marts, organizations face operational difficulties in using them in an overall corporate data warehouse strategy, because individual data marts are not consistent with each other.

Metadata:

To understand and locate data in the data warehouse users need information about the data warehousing system and its content. This information known as metadata, data about data, includes format, encoding/decoding algorithms, domain constraints, and definitions of the data. It also includes business definitions, data quality alerts, organizational changes, business rules and assumptions, as well as other items of business interest. Metadata help the business user to understand what is available, how to access it, what it means, which data to use, when to use them, etc.

Metadata browsers provide an easy to understand view of the data warehouse.

- Data Access and Analysis

The access component of the BI is referred to as the front end. It consists of access tools and techniques that provide a business user with direct, interactive, or batch access to data, while hiding the technical complexity of data retrieval. The interface provides an intuitive, business-like presentation of information, friendly enough for use by a no technical person. This is accomplished by use of BI tools, a suite of software tools that presents a graphical user interface (GUI) with rich reporting and business analysis features. A variety of tools are typically used in an integrated fashion to serve the needs of different groups of users.

Revision Questions

Example ✏. ....

Solution: .... □

EXERCISE 1. ✍ ....

LESSON 2

Data warehouse

Data warehouse as basis of business

- Intelligence; Data warehouse definitions and concepts
- Data warehouse architecture
- Data integration and the extraction, transformation and load process.

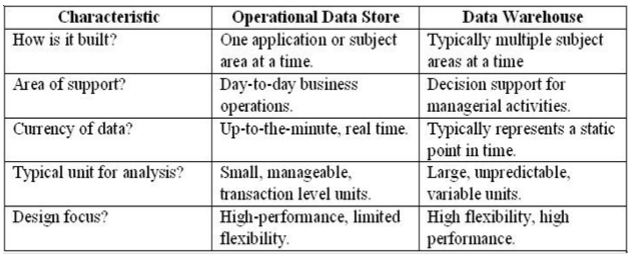
2.1. What is Data Warehouse?

A data warehouse is a collection of integrated databases designed to support a DSS. According to Inmon's (father of data warehousing) definition(Inmon,1992a,p.5): It is a collection of integrated, subject-oriented databases designed to support the DSS function, where each unit of data is non-volatile and relevant to some moment in time. It is used for evaluating future strategy. It needs a successful technician: Flexible, Team player, Good balance of business and technical understanding.
The ultimate use of data warehouse is Mass Customization. For example, it increased Capital One's customers from 1 million to approximately 9 millions in 8 years. Just like a muscle: DW increases in strength with active use. With each new test and product, valuable information is added to the DW, allowing the analyst to learn from the success and failure of the past. The key to survival: Is the ability to analyze, plan, and react to changing business conditions in a much more rapid fashion. In order for data to be effective, DW must be: Consistent, Well integrated, well defined and time stamped.

2.1.1. DW environment:

The data store, data mart & the metadata.
What is Data Store?
It is operational data store (ODS) used for storing data for a specific application. It feeds the data warehouse a stream of desired raw data. Is the most common component of DW environment. Data store is generally subject oriented, volatile, current commonly focused on customers, products, orders, policies, claims, etc...

EXERCISE 2. ✍ What is the difference between Data Store & Data Warehouse



Characteristic	Operational Data Store	Data Warehouse
How is it built?	One application or subject area at a time.	Typically multiple subject areas at a time
Area of support?	Day-to-day business operations.	Decision support for managerial activities.
Currency of data?	Up-to-the-minute, real time.	Typically represents a static point in time.
Typical unit for analysis?	Small, manageable, transaction level units.	Large, unpredictable, variable units.
Design focus?	High-performance, limited flexibility.	High flexibility, high performance.

Data Store's day-to-day function is to store the data for a single specific set of operational application and to feed the data warehouse data for the purpose of analysis.

EXERCISE 3. ✍ What is Data Mart?

It is lower-cost, scaled down version of the DW. It offer a targeted and less costly method of gaining the advantages associated with data warehousing and can be scaled up to a full DW environment over time. It's the last component of DW environments. It is information that is kept about the warehouse rather than information kept within the warehouse. Legacy systems generally don't keep a record of characteristics of the data (such as what pieces of data exist and where they are located).

EXERCISE 4. ✍ What is metadata? It's simply data about data.

EXERCISE 5. ✍ Explain NINE Characteristics of Data Warehouse

i. Subject oriented. Data are organized based on how the users refer to them.

ii. Integrated. All inconsistencies regarding naming convention and value representations are removed.

iii. Nonvolatile. Data are stored in read-only format and do not change over time.

iv. Time variant. Data are not current but normally time series.

v. Summarized Operational data are mapped into a decision-usable format

vi. Large volume. Time series data sets are normally quite large.

vii. Not normalized. DW data can be, and often are, redundant.

viii. Metadata. Data about data are stored.

ix. Data sources. Data come from internal and external unintegrated operational systems.

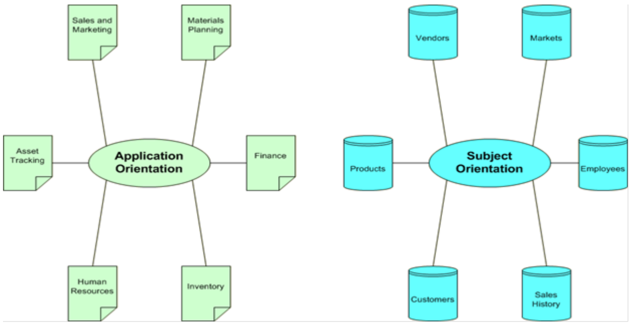


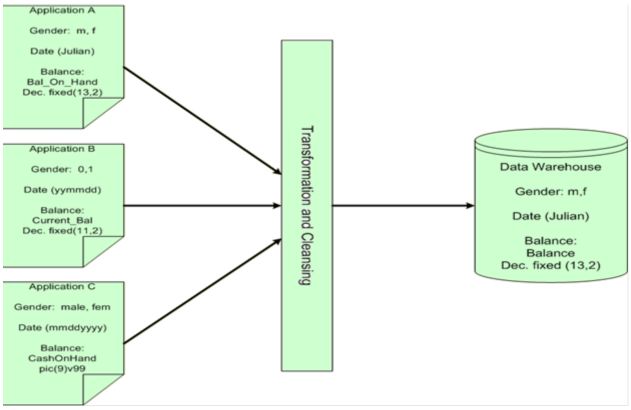
Figure 2.1: a datawarehouse is subject oriented

2.1.2. Subject Orientation

Application Environment	Data warehouse Environment
Design activities must be equally focused on both process and database design	DW world is primarily void of process design and tends to focus exclusively on issues of data modeling and database design

2.1.3. Data Integrated

1. Integration –consistency naming conventions and measurement attributers, accuracy, and common aggregation.
2. Establishment of a common unit of measure for all synonymous data elements from dissimilar database.
3. The data must be stored in the DW in an integrated, globally acceptable manner



2.1.4. Time Variant

1. In an operational application system, the expectation is that all data within the database are accurate as of the moment of access. In the DW data are simply assumed to be accurate as of some moment in time and not necessarily right now.

2. One of the places where DW data display time variance is in the structure of the record key. Every primary key contained within the DW must contain, either implicitly or explicitly an element of time(day, week, month, etc)

3. Every piece of data contained within the warehouse must be associated with a particular point in time if any useful analysis is to be conducted with it.

4. Another aspect of time variance in DW data is that, once recorded, data within the warehouse cannot be updated or changed.

2.1.5. Nonvolatility

1. Typical activities such as deletes, inserts, and changes that are performed in an operational application environment are completely nonexistent in a DW environment.

2. Only two data operations are ever performed in the DW: data loading and data access

Application	DW
The design issues must focus on data integrity and update anomalies. Complex processes must be coded to ensure that the data update processes allow for high integrity of the final product.	Such issues are no concern to in a DW environment because data update is never performed.
Data is placed in normalized form to ensure a minimal redundancy (totals that could be calculated would never be stored)	Designers find it useful to store many of such calculations or summarizations.
The technologies necessary to support issues of transaction and data recovery, roll back, and detection and remedy of deadlock are quite complex.	Relative simplicity in technology

EXERCISE 6. ✍ Explain issues of Data Redundancy between DW and operational environments

i. The lack of relevancy of issues such as data normalization in the DW environment may suggest that existence of massive data redundancy within the data warehouse and between the operational and DW environments.

ii. The data being loaded into the DW are filtered and "cleansed" as they pass from the operational database to the warehouse. Because of this cleansing numerous data that exists in the operational environment never pass to the data warehouse. Only the data necessary for processing by the DSS or EIS are ever actually loaded into the DW.

iii. The time horizons for warehouse and operational data elements are unique. Data in the operational environment are fresh, whereas warehouse data are generally much older.(so there is minimal opportunity of the data to overlap between two environments)

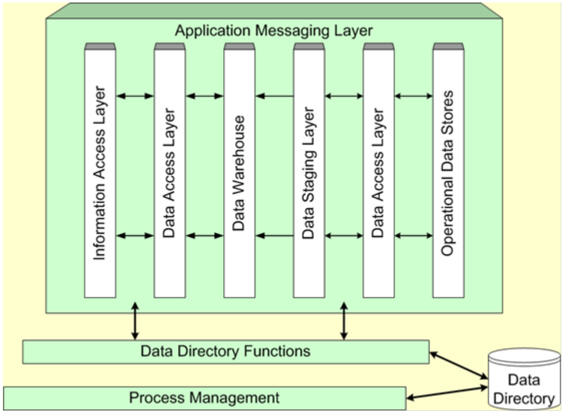
iv. The data loaded into the DW often undergo a radical transformation as they pass from operational to the DW environment. So data in DW are not the same.

- **The Data Warehouse Architecture**

The architecture consists of various interconnected elements:

1. Operational and external database layer – the source data for the DW
2. Information access layer – the tools the end user access to extract and analyze the data
3. Data access layer – the interface between the operational and information access layers iv. Metadata layer – the data directory or repository of metadata information

Components of the Data Warehouse Architecture



The Data Warehouse Architecture Additional layers includes:

- Process management layer – the scheduler or job controller
- Application messaging layer – the “middleware” that transports information around the firm
- Physical data warehouse layer – where the actual data used in the DSS are located
- Data staging layer – all of the processes necessary to select, edit, summarize and load warehouse data from the operational and external data bases

2.1.6. Data Warehousing Typology

- **The virtual data warehouse –** the end users have direct access to the data stores, using tools enabled at the data access layer
- **The central data warehouse** – a single physical database contains all of the data for a specific functional area
- **The distributed data warehouse** – the components are distributed across several physical databases

EXERCISE 7. ✍ Explain the Components of the Metadata
i. Transformation maps – records that show what transformations were applied
ii. Extraction & relationship history – records that show what data was analyzed
iii. Algorithms for summarization – methods available for aggregating and summarizing
iv. Data ownership – records that show origin
v. Patterns of access – records that show what data are accessed and how often

2.2. Typical Mapping Metadata

Transformation mapping records include:

1. Identification of original source
2. Attribute conversions
3. Physical characteristic conversions
4. Encoding/reference table conversions
5. Naming changes vi. Key changes
6. Values of default attributes
7. Logic to choose from multiple sources
8. Algorithmic changes

2.3. Implementing the Data Warehouse

Kozar list of “seven deadly sins” of data warehouse implementation:

1. “If you build it, they will come” – the DW needs to be designed to meet people’s needs
2. Omission of an architectural framework – you need to consider the number of users, volume of data, update cycle, etc.
3. Underestimating the importance of documenting assumptions – the assumptions and potential conflicts must be included in the framework
4. Failure to use the right tool – a DW project needs different tools than those used to develop an application v. Life cycle abuse – in a DW, the life cycle really never ends
5. Ignorance about data conflicts – resolving these takes a lot more effort than most people realize
6. Failure to learn from mistakes – since one DW project tends to beget another, learning from the early mistakes will yield higher quality later

Data Warehouse Technologies

1. No one currently offers an end-to-end DW solution. Organizations buy bits and pieces from a number of vendors and hopefully make them work together.
2. SAS, IBM, Software AG, Information Builders and Platinum offer solutions that are at least fairly comprehensive.
3. The market is very competitive.

The Future of Data Warehousing
As the DW becomes a standard part of an organization, there will be efforts to find new ways to use the data. This will likely bring with it several new challenges:

1. Regulatory constraints may limit the ability to combine sources of disparate data.
2. These disparate sources are likely to contain unstructured data, which is hard to store.

3. The Internet makes it possible to access data from virtually “anywhere”. Of course, this just increases the disparity.

2.3.1. Objective of Data Warehouse

1. Interesting Facts

   Harrah’s Entertainment’s Data Warehouse holds 30 terabytes, or 30 trillion bytes of data, roughly three times the number of printed characters in the Library of Congress Casinos, retailers, airlines, and banks are piling up data so vast, it would have been unthinkable years ago; result from the curse of cheap storage Storage Shipments as of 2004: 22 exabytes or 22 million trillion bytes of hard disk space, double the amount in 2002. Equivalent to 4x’s the space needed to store every word ever spoken by every human being who has ever lived. Should double again in 2006

2. Data Can be Used To

   Quantify the volume impact of vehicles across the marketing matrix Account for decay and saturation factors in the determination of investment choices and returns Execute “what-if” simulations of pricing or promotional scenarios before a proposed action is taken Provide a continuous planning, measurement, analysis and optimization cycle supported by a software structure Deliver robust data feeds into other systems supporting supply chain, sales, and financial reporting and endeavors

3. Robust Infrastructure

   Data Identification and Acquisition Data Cleansing, Mapping, and Transformation Production System Loading and Ongoing Update

4. Success of Data Warehouse Projects

   Over half of Data Warehouse projects are Doomed Fail due to lack of attention to Data Quality Issues More than half only have limited acceptance Consistency and Accuracy of Data Most businesses fail to use business intelligence (BI) strategically IT organizations build data warehouses with little to no business involvement Most challenging type of deployment for an enterprise Large scale and complex system configurations Sophisticated data modeling and analysis tools High visibility in broad range of important business functions within company Adoption of Linux-Based Platform

5. Implementing Data Warehouse

   Challenges: Identifying new processes Assuring there were of real use Implementing and ensuring cultural shifts Managing content and New communities towards a common benefit Linear models Standards, Governance, Controls, Valuation

6. Real Time Alerts & Integration

   Teradata 8.0 Version released in Oct 2004 Improves real-time alerts and integration Businesses can analyze operational info against historical info to identify events in near real-time using the new table design Used by: Continental Airlines in the US: reroute passengers on delayed flights, reissuing tickets, reserving a room in a hotel booking system Southwest Airlines- savings between $1.2-$1.4 Million

7. Identity Theft

   Government Regulation of Personal Data is Needed (National Consumer Protection Standards) ChoicePoint Folly Georgia-based data-collection company Founded in 1997 to analyze insurance claims information, but now provides data to customers including finance companies, law enforcement, and government Obtain personal information by perusing public records, or purchasing the information from other companies Duped by scammers who set 150 phony accounts to access personal data of as many as 145,000 people nationwide Scammers set user accounts by faxing in phony business licenses, undetected for one year 750 people had their identities stolen Theft would have gone unnoticed without California Identity theft law SB 1386 MSN Event Data Warehouse Information Gathering Over the Phone Interviews Trash Can Hunting Gathered from Doctors, Internet Transactions, Telephone Operators (Overseas or Prisoners)

Revision Questions

Example ✎. ....

Solution: .... □

EXERCISE 8. ✍ ....

LESSON 3

Business Analytics and Data Visualization

Learning Objectives:

1. List and briefly describe the major BA methods and tools
2. Describe how online analytical processing (OLAP), data visualization, and multidimensionality can improve decision making
3. Describe geographical information systems (GIS) and their support to decision making
4. Describe real-time BA v. Explain how the Web relates to BA
5. Describe Web intelligence and Web analytics and their importance to organizations
6. Describe implementation issues related to BA and success factors for BA

3.1. Lexmark Improves Operations with BI

1. Identify the challenges Lexmark faced regarding information flow
2. How were the information flows provided before and after implementation of the system?
3. Identify the decisions supported by the new system.
4. How can the new system improve customer service?

3.1.1. The Business Analytics (BA) Field: An Overview

Business Analytics

Is the use of analytical methods, either manually or automatically, to derive relationships from data remember that we defined business analytics (BA) to include the access, reporting, and analysis of data supported by software to drive business performance and decision making

The figure below shows Business Analytics categories

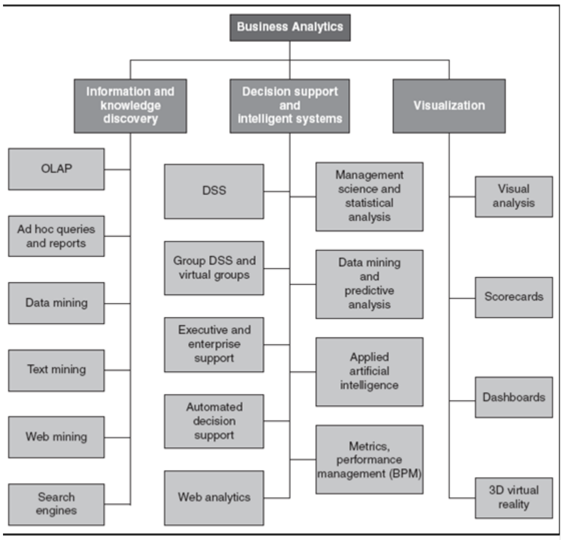

Business Analytics
Information and knowledge discovery
Decision support and intelligent systems
Visualization
OLAP
Ad hoc queries and reports
Data mining
Text mining
Web mining
Search engines
DSS
Group DSS and virtual groups
Executive and enterprise support
Automated decision support
Web analytics
Management science and statistical analysis
Data mining and predictive analysis
Applied artificial intelligence
Metrics, performance management (BPM)
Visual analysis
Scorecards
Dashboards
3D virtual reality

FIGURE 3.1 Categories of Business Analytics

3.1.2. MicroStrategy's classification of BA tools:

The micro strategy classification includes five styles of BI

1. Enterprise reporting
2. Cube analysis
3. Ad hoc querying and analysis
4. Statistical analysis and data mining
5. Report delivery and alerting

3.2. Executive information systems (EIS)

It provides rapid access to timely and relevant information aiding in monitoring an organization's performance. It also provides analysis support, communications, office automation, and intelligence support for the top management

EXERCISE 9. ✍ Explain how the Drill-down and online analytical processing can assist the manager make decision

Solution

Drill-down feature: The investigation of information in detail (e.g., finding not only total sales but also sales by region, by product, or by salesperson). Finding the detailed sources

Online analytical processing (OLAP): An information system that enables the user, while at a PC, to query the system, conduct an analysis, and so on. The result is generated in seconds

Exercise 10. ✍ Explain the difference between OLAP and OLTP
OLTP concentrates on processing repetitive transactions in large quantities and conducting simple manipulations
OLAP involves examining many data items complex relationships
OLAP may analyze relationships and look for patterns, trends, and exceptions
OLAP is a direct decision support method

Exercise 11. ✍ What is the different between Reports and Queries
a) Reports

- Routine reports
- Ad hoc (or on-demand) reports
- Multilingual support
- Scorecards and dashboards
- Report delivery and alerting
- Report distribution through any touch-point
- Self-subscription as well as administrator-based distribution
- Delivery on-demand, on-schedule, or on-event
- Automatic content personalization

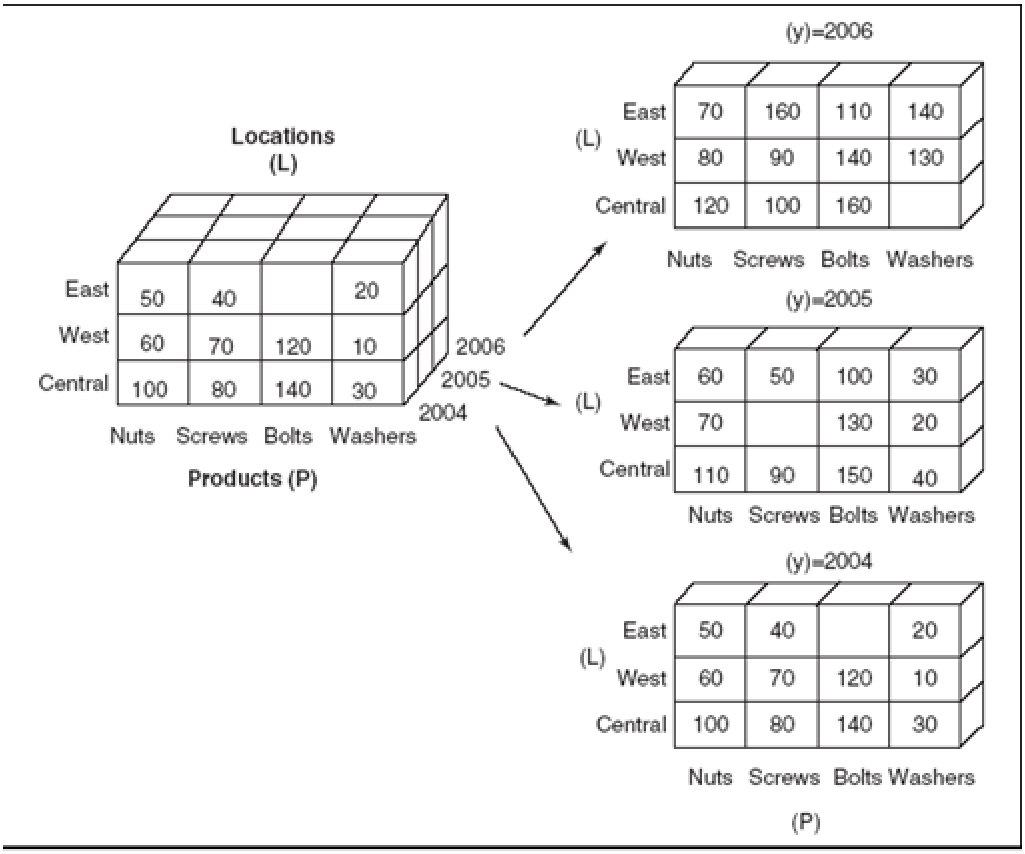
b) Ad hoc query -A query that cannot be determined prior to the moment the query is issued

Exercise 12. ✍ What is Structured Query Language (SQL) A data definition and management language for relational databases. SQL front ends most relational DBMS

3.3. What is Multidimensionality

The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions)
Multidimensional presentation

- Dimensions
- Measures



	Nuts	Screws	Bolts	Washers
East	50	40		20
West	60	70	120	10
Central	100	80	140	30

(y)=2006

(L)	Nuts	Screws	Bolts	Washers
East	70	160	110	140
West	80	90	140	130
Central	120	100	160	

(y)=2005

(L)	Nuts	Screws	Bolts	Washers
East	60	50	100	30
West	70		130	20
Central	110	90	150	40

(y)=2004

(L)	Nuts	Screws	Bolts	Washers
East	50	40		20
West	60	70	120	10
Central	100	80	140	30

FIGURE 3.3 Cube Analysis and Views

Figure 3.1: multidimensionality

- Time

3.3.1. Multidimensional database

A database in which the data are organized specifically to support easy and quick multidimensional analysis

- Data cube: A two-dimensional, three-dimensional, or higher-dimensional object in which each dimension of the data represents a measure of interest

- Cube : A subset of highly interrelated data that is organized to allow users to combine any attributes in a cube (e.g., stores, products, customers, suppliers) with any metrics in the cube (e.g., sales, profit, units, age) to create various two-dimensional views, or slices, that can be displayed on a computer screen

3.3.2. Multidimensional tools and vendors

Tools with multidimensional capabilities often work in conjunction with database query systems and other OLAP tools
Limitations of dimensionality

- The multidimensional database can take up significantly more computer storage room than a summarized relational database
- Multidimensional products cost significantly more than standard relational products
- Database loading consumes significant system resources and time, depending on data volume and the number of dimensions
- Interfaces and maintenance are more complex in multidimensional databases than in relational databases

3.3.3. Advanced Business Analytics

Data mining and predictive analysis

- Data mining
- Predictive analysis

Use of tools that help determine the probable future outcome for an event or the likelihood of a situation occurring. These tools also identify relationships and patterns

3.4. What is Data visualization

A graphical, animation, or video presentation of data and the results of data analysis.

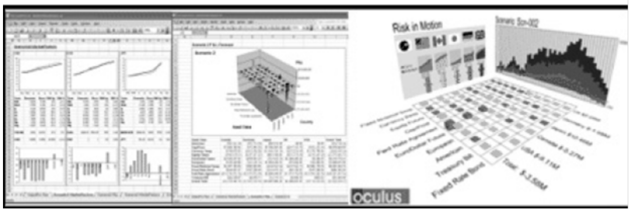
- The ability to quickly identify important trends in corporate and market data can provide competitive advantage
- Check their magnitude of trends by using predictive models that provide significant business advantages in applications that drive content, transactions, or processes

3.4.1. New directions in data visualization

In the 1990s data visualization has moved into:

- Mainstream computing, where it is integrated with decision support tools and applications.
- Intelligent visualization, which includes data (information) interpretation
- Dashboards and scorecards
- Visual analysis influence diagrams
- Financial data visualization

FIGURE 3.5 Visual Spreadsheet of Risk Analysis



3.5. Geographical information system (GIS)

An information system that uses spatial data, such as digitized maps. A GIS is a combination of text, graphics, icons, and symbols on maps As GIS tools become increasingly sophisticated and affordable, they help more companies and governments understand:

- Precisely where their trucks, workers, and resources are located
- Where they need to go to service a customer
- The best way to get from here to there

3.5.1. GIS and decision making

GIS applications are used to improve decision making in the public and private sectors including:

- Dispatch of emergency vehicles
- Transit management

- Facility site selection
- Drought risk management
- Wildlife management
- Local governments use GIS applications for used mapping and other decision-making applications

GIS combined with GPS

- Global positioning systems (GPS)
- Wireless devices that use satellites to enable users to detect the position on earth of items (e.g., cars or people) the devices are attached to, with reasonable precision

GIS and the Internet/intranets

- Most major GIS software vendors provide Web access that hooks directly to their software
- GIS can help the manager of a retail operation determine where to locate retail outlets
- Some firms are deploying GIS on the Internet for internal use or for use by their customers (locate the closest store location)

Real-Time BI

- The trend toward BI software producing real-time data updates for real-time analysis and real-time decision making is growing rapidly
- Part of this push involves getting the right information to operational and tactical personnel so that they can use new BA tools and up-to-the-minute results to make decisions

Concerns about real-time systems

- An important issue in real-time computing is that not all data should be updated continuously

- when reports are generated in real-time because one person's results may not match another person's causing confusion

- Real-time data are necessary in many cases for the creation of ADS systems

BA and the Web: Web Intelligence and Web Analytics

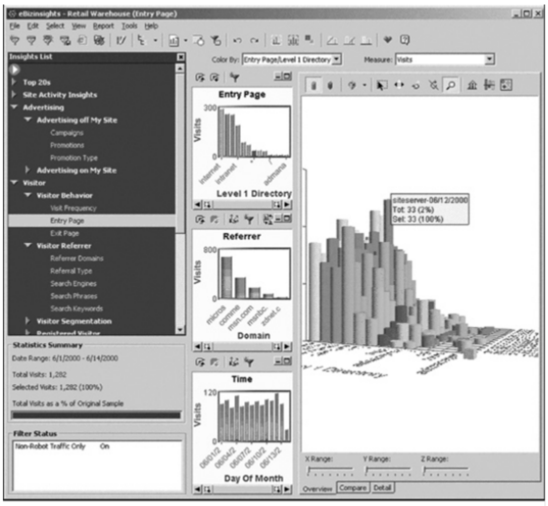
- Using the Web in BA

- Web analytics

The application of business analytics activities to Web-based processes, including e-commerce

- Click stream analysis: The analysis of data that occur in the Web environment.

- Click stream data : Data that provide a trail of the user's activities and show the user's browsing patterns (e.g., which sites are visited, which pages, how long)



Color By: Entry Page/Level 1 Directory

Measure: Visits

Statistics Summary

Date Range: 6/1/2000 - 6/14/2000

Total Visits: 1,282

Selected Visits: 1,282 (100%)

Total Visits as a % of Original Sample

Filter Status

Non-Robot Traffic Only On

FIGURE 3.7 Screen Shot from the eBizInsights Visual Portal Analysis of Web Performance

3.5.2. Usage, Benefits, and Success of BA

Usage of BA

- Almost all managers and executives can use some BA systems, but some find the tools too complicated to use or they are not trained properly.

- Most businesses want a greater percentage of the enterprise to leverage analytics; most of the challenges related to technology adoption involve culture, people, and processes
- Performance management systems (PMS) are BI tools that provide scorecards and other relevant information that decision makers use to determine their level of success in reaching their goals

Explain why BI/BA projects usually fails

- Failure to recognize BI projects as cross-organizational business initiatives and to understand that, as such, they differ from typical standalone solutions
- Unengaged or weak business sponsors
- Unavailable or unwilling business representatives from the functional areas
- Lack of skilled (or available) staff, or suboptimal staff utilization
- No software release concept (i.e., no iterative development method)
- No work breakdown structure (i.e., no methodology)
- No business analysis or standardization activities
- No appreciation of the negative impact of "dirty data" on business profitability
- No understanding of the necessity for and the use of metadata
- Too much reliance on disparate methods and tools

Revision Questions

Example ✎. ....

Solution: .... □

Exercise 13. ✍ ....

LESSON 4

Data Mining

Objectives

- What is Data Mining?
- Knowledge resource
- Knowledge types and/or knowledge datasets
- Data mining tasks
- Data mining techniques and applications used in knowledge management

In information era, knowledge is becoming a crucial organizational resource that provides competitive advantage and giving rise to knowledge management (KM) initiatives. Many organizations have collected and stored vast amount of data. However, they are unable to discover valuable information hidden in the data by transforming these data into valuable and useful knowledge. Managing knowledge resources can be a challenge. Many organizations are employing information technology in knowledge management to aid creation, sharing, integration, and distribution of knowledge. Knowledge management is a process of data usage. The basis of data mining is a process of using tools to extract useful knowledge from large datasets; data mining is an essential part of knowledge management. Data mining is one of the most important steps of the knowledge discovery in databases process and is considered as significant subfield in knowledge management. Research in data mining continues growing in business and in learning organization over coming decades.

4.1. Definition of Data Mining

Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data. The terms of KDD and data mining are different. KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge
Based on the figure above, KDD process consists of iterative sequence methods as follows

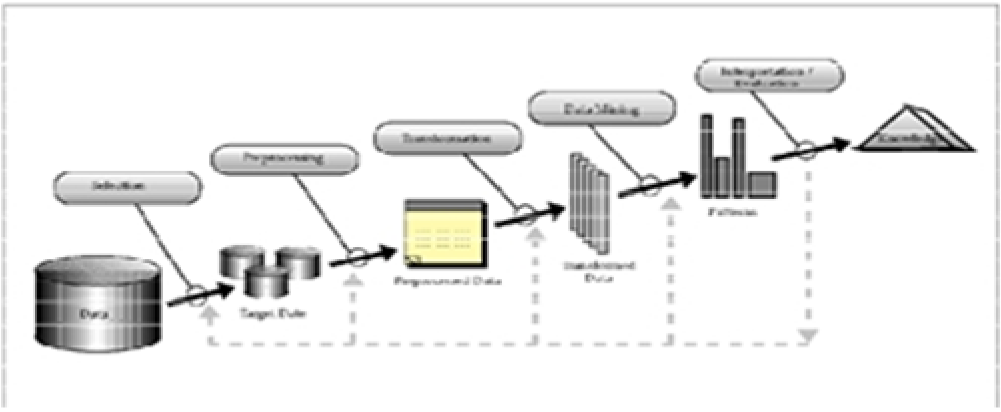


Figure 4.1: The figure above shows Data Mining and the KDD Process

1. Selection: Selecting data relevant to the analysis task from the database
2. Preprocessing: Removing noise and inconsistent data; combining multiple data sources
3. Transformation: Transforming data into appropriate forms to perform data mining
4. Data mining: Choosing a data mining algorithm which is appropriate to pattern in the data; Extracting data patterns
5. Interpretation/Evaluation : Interpreting the patterns into knowledge by removing
6. Redundant or irrelevant patterns; Translating the useful patterns into terms that human understandable

4.2. Data Mining Tasks

Define six main functions of data mining

1. Classification is finding models that analyze and classify a data item into several predefined classes
2. Regression is mapping a data item to a real-valued prediction variable
3. Clustering is identifying a finite set of categories or clusters to describe the data

4. Dependency Modeling (Association Rule Learning) is finding a model which describes significant dependencies between variables

5. Deviation Detection (Anomaly Detection) is discovering the most significant changes in the data vi. Summarization is finding a compact description for a subset of data

Data mining has two primary objectives of prediction and description. Prediction involves using some variables in data sets in order to predict unknown values of other relevant variables (e.g. classification, regression, and anomaly detection) Description involves finding human-understandable patterns and trends in the data (e.g. clustering, association rule learning, and summarization)

4.2.1. Definition of Knowledge Management

Knowledge management (KM) is an effort to increase useful knowledge within the organization. Ways to do this include encouraging communication, offering opportunities to learn, and promoting the sharing of appropriate knowledge artifacts. Knowledge management process focuses on knowledge flows and the process of creation, sharing, and distributing knowledge.
Each of knowledge units of capture and creation, sharing and dissemination, and acquisition and application can be facilitated by information technology. As technologies play an important role in KM, technologies stand to be a necessary tool for KM usage. Thus, KM requires technologies to facilitate communication, collaboration, and content for better knowledge capture, sharing, dissemination, and application.
Knowledge Management: Capture and Creation Tools Liao (2003) classifies KM technologies using seven categories:

1. KM Framework

2. Knowledge-Based Systems (KBS)

3. Data Mining

4. Information and Communication Technology

5. Artificial Intelligence (AI)/Expert Systems (ES)

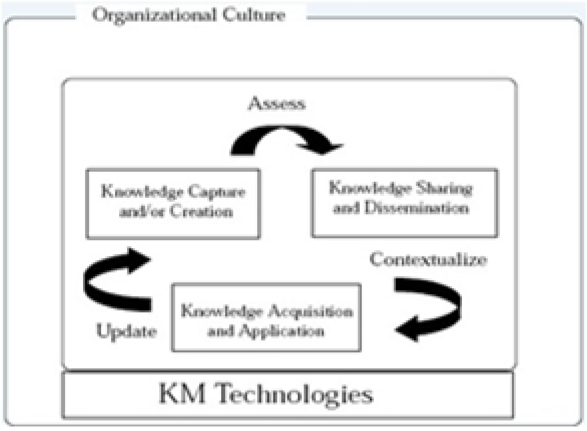
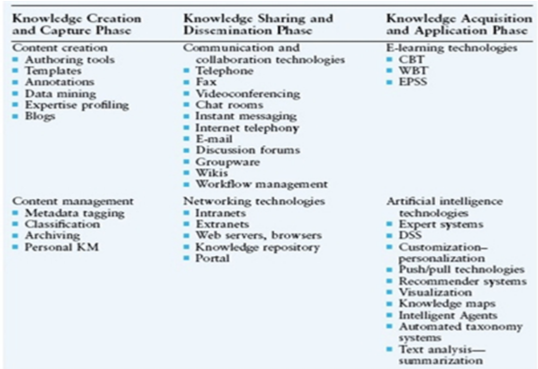

Organizational Culture
Assess
Knowledge Capture and/or Creation
Knowledge Sharing and Dissemination
Contextualize
Update
Knowledge Acquisition and Application
KM Technologies

Figure 4.2: Figure above shows KM Technologies Integrated KM Cycle (Source from Dalkir, K.,2005).

6. Database Technology (DT)

7. Modeling

Ruggles et.al. (1997) classify KM technologies as tools that generate knowledge (e.g. data mining), code knowledge, and transfer knowledge. Dalkir (2005) classifies KM tools according to the phase of the KM cycle (Figure below). We can see that data mining involves in the part of knowledge creation and capture phase.



Knowledge Creation and Capture Phase	Knowledge Sharing and Dissemination Phase	Knowledge Acquisition and Application Phase
Content creation ▪ Authoring tools ▪ Templates ▪ Annotations ▪ Data mining ▪ Expertise profiling ▪ Blogs	Communication and collaboration technologies ▪ Telephone ▪ Fax ▪ Videoconferencing ▪ Chat rooms ▪ Instant messaging ▪ Internet telephony ▪ E-mail ▪ Discussion forums ▪ Groupware ▪ Wikis ▪ Workflow management	E-learning technologies ▪ CBT ▪ WBT ▪ EPSS
Content management ▪ Metadata tagging ▪ Classification ▪ Archiving ▪ Personal KM	Networking technologies ▪ Intranets ▪ Extranets ▪ Web servers, browsers ▪ Knowledge repository ▪ Portal	Artificial intelligence technologies ▪ Expert systems ▪ DSS ▪ Customization–personalization ▪ Push/pull technologies ▪ Recommender systems ▪ Visualization ▪ Knowledge maps ▪ Intelligent Agents ▪ Automated taxonomy systems ▪ Text analysis—summarization

4.3. The applications of data mining in knowledge management

Knowledge management assists organization for effective capturing, storing and retrieving, and transferring knowledge. Areas of application

1. Knowledge resource;

2. Knowledge types and/or knowledge datasets;

3. Data mining tasks; and

4. Data mining techniques and applications used in KM.

4.3.1. Knowledge Resources

We divide knowledge resources into eight groups as that which knowledge object to be stored and manipulated in KM and how data mining aids.

1. **Health Care Organization**: this domain is a used in disease knowledge management system (KMS) in the hospital. Data mining tool is used to explore diseases, operations, and tumors relationships. This tool is used to build KMS to support clinical medicine in order to improve treatment quality.

2. **Retailing:** this is customer knowledge from household customers for product line and brand extension issues. Data mining can help and propose suggestions and solutions to the firm for product line and brand extensions. This is done by extracting market knowledge of customers, brands, products, and purchase data to fulfill the customers' demands behavior.

3. **Financial/Banking:** the domain knowledge covers financial and economic data; data mining can assist banking institutions making decision support and knowledge sharing processes to an enterprise bond classification. Small and Middle Businesses (food company and food supply chain): there are two methods and processes to obtain knowledge resources: knowledge seeding-the relative knowledge to the problems; knowledge cultivating-the process to find the key knowledge from knowledge seeding. Data mining and knowledge management integrated can help making better decisions.

4. **Entrepreneurial Science:** the knowledge resource is research assets in a knowledge institution .There are three types of the research assets: research products, intellectual capital, and research programs. Data mining facilitate for knowledge extraction and helped guiding managers in determining strategies on knowledge-oriented organization competition.

5. **Business**: data collected from questionnaire, and intensive literature review, and discussions with four KM experts. Data mining can discover hidden patterns between KM and its performance for better KM implementations.

6. **Collaboration and Teamwork:** Worker's log and documents are analyzed and each worker's referencing behavior and construct worker's knowledge flow. Data mining techniques can mine and construct group-based knowledge flows (GKFs) prototype for task-based groups.

7. **Construction Industry:** a large part of this enterprise information is available in the form of textual data formats. This leads to the influence of text mining techniques to handle textual information source for industrial knowledge discovery and management solutions.

8. **Small and Middle Businesses (food company and food supply chain):** there are two methods and processes to obtain knowledge resources: knowledge seeding-the relative knowledge to the problems; knowledge cultivating-the process to find the key knowledge from knowledge seeding. Data mining and knowledge management integrated can help making better decisions. Knowledge Base is an important part of EW&PC systems. It contained data analysis by managers and organizes in an appropriate way for other managers. Data mining methods is helpful for the EW&PC systems.

4.3.2. Knowledge Types

Here we describe knowledge types in 8 organization domains for data mining collaboration process in the knowledge creation.

1. **Health-care System domain, the dataset composed of three databases: the health-care providers' database;** the out-patient health-care statistics database; and the medical status database .Another data source was from hospital inpatient medical records.

2. **Construction Industry** domain, a sample data set is in the form of Post Project Reviews (PPRs) as defining good or bad information. Multiple Key Term Phrasal Knowledge sequences (MKTPKS) formation is generated through applications of text mining and is used an essential part of the text analysis in the text documents classification.

3. **Retailing domain:** customer data and the products purchased is collected and stored in databases to mine whether the customers' purchase habits and behavior affect the product line and brand extensions or not.

4. **Financial domain:** There are two datasets posed in financial domain: i. To identify bond ratings, knowledge sets contained strings of data, models, parameters and reports for each analytical study; and ii. To predict rating changes of bonds, cluster data of bond features as well as the model parameters were stored, classified, and applied to rating predictions.

5. **Small and Middle Businesses (SMBs) domain:** Knowledge types in small and middle Businesses in case of Food Company are related to the corporate conditions or goals of the problem among all departments to develop a decision system platform and then formed the knowledge tree to find relations by human-computer interaction method and optimize the process of decision making. To solve food supply chain networks problems, Li et al. (2010) developed EW&PC prototype which composed of major components of:

   (a) knowledge base,

   (b) task classifier and template approaches,

   (c) DM methods library with expert system for method selection

   (d) Explorer and predictor, and

   (e) User interface. This system built decision support models and helped managers to accomplish decision-making.

6. **Research Assets domain:** In Cantu & Cellbos (2010) focused on managing knowledge assets by applied acknowledge and information network (KIN) approach. This platform contained three components types of research products, human resources or intellectual capital, and research programs. The various types of research assets were handled on domain and databases.

7. **Business domain:** there are two types of knowledge attributes conducted: condition attributes and decision attribute. Condition attributes includes four independent attributes of the KM purpose, the explicit-oriented degree, the tacit-oriented degree, and the success factor. Decision attribute included one dependent attribute of the KM performance.

8. **Collaboration and Teamwork domain:** a dataset used from a research laboratory in a research institute. It contained 14 knowledge workers, 424 research documents, and a workers' log as that recorded the time of document

accessed and the documents of workers' needed . For the workers' log, it was generated to 2 levels of codified-level knowledge flow and topic-level knowledge flow. The two types of knowledge flow were determined to describe a worker's needs. To collect the knowledge flow, documents in the dataset are categorized into eight clusters by data mining clustering approach.

Revision Questions

Example ✐. ....

Solution: .... □

EXERCISE 14. ✍ ....

LESSON 5

Text Mining

Learning Objectives

- Describe text mining and understand the need for text mining
- Understand the different application areas for text mining
- Know the process of carrying out a text mining project
- Understand the different methods to introduce structure to text-based data
- How do we do patent analysis (PA)
- What are the challenges of patent analysis

Text Mining Concepts: 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text) Unstructured corporate data is doubling in size every 18 months. Tapping into these information sources is not an option, but a need to stay competitive. The solution is text mining.

Text mining is a semi-automated process of extracting knowledge from unstructured data sources knowledge discovery in textual databases.

Compare and contrast between Data Mining and Text Mining

1. Both seek for novel and useful patterns
2. Both are semi-automated processes

Difference is the nature of the data:

1. Structured versus unstructured data
2. Structured data: in databases
3. Unstructured data: Word documents, PDF files, text excerpts, XML files, and so on

Text mining – first, impose structure to the data, then mine the structured data

5.0.3. Text Mining Concepts

Benefits of text mining are obvious especially in text-rich data environments : e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc. Electronic communization records (e.g., Email)

- Spam filtering
- Email prioritization and categorization
- Automatic response generation

5.1. Text Mining for Patent Analysis

5.1.1. What is a patent?

It's a exclusive rights granted by a country to an inventor for a limited period of time in exchange for a disclosure of an invention.

How do we do patent analysis (PA)?

Patent analysis (PA) is to analyze the data of patent documents such as abstract and the number of issued patents. PA plays a major role in R&D policy because we can forecast the future aspect of a technology by the result of PA. So, most companies make efforts to perform PA for improving their competitiveness. One of the proposed model for PA is technology forecasting (TF). Here we combine the clustering and predictive results for TF. Using the retrieved patent documents from the United State Patent and Trademark Office, we make experiment to verify the performance.

5.1.2. What is Technology forecasting (TF)

Technology forecasting (TF) is to foresight the technological aspect in future. A considerable portion of the R&D plan has been depended on the TF results. Also, patent analysis (PA) plays an important role in the TF process. We use the document clustering and time series analysis and combine these results for constructing efficient TF model. Many researches published in PA fields are based on one analytical approach such as clustering, classification, and citation analyses.

But, they had some limitations to forecast the future state of a technology because they are depended on only one result of a TF method. So, to reduce this problem, we consider combining two analytical approaches which are patent document clustering and time series model. We use K-means clustering algorithm as a patent clustering method and time series regression (TSR) as a time series model.

5.1.3. Challenges of patent analysis

- Incompetence of the patent experts as pertained to the information presented in patents and the patent databases. Patent analysts with different levels of expertise require patent analysis tools with versatile capabilities

- Issues related to patent information are more complex and critical from the perspectives of searching the patent databases and retrieving information.

- The task of searching the patent databases to find relevant patents is supported by various data and text mining tools. Text mining tools with capabilities of mining text from structured and unstructured data have been developed. Mining the structured information from patent documents is relatively easier as compared to unstructured data because of its textual nature. Therefore, the task of parsing the unstructured data requires extraction tools having capabilities of segmentation of textual data into meaningful structures

- The visual output of the structured patent data is represented in the form of graphs and networks whereas the results from the unstructured patent data are represented as patent maps.

- Organizations should investigate the potential infringement risks before investing in new products because patent litigation may possibly result in huge financial bearings.

EXERCISE 15. ✍ How does text mining help in patent analysis?

EXERCISE 16. ✍ What are the benefits of patent analysis?

EXERCISE 17. ✍ With the aid of examples, explain the following text mining application areas i. Information extraction ii. Topic tracking iii. Summarization iv. Categorization v. Clustering vi. Concept linking vii. Question answering

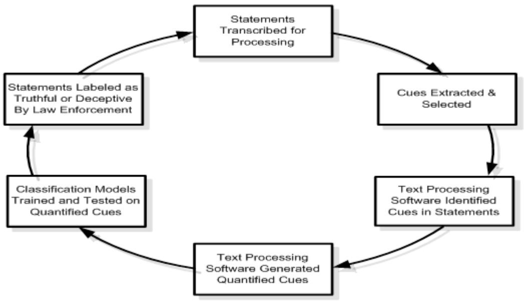
Exercise 18. ✍ With the aid of examples explain the following terminologies as used in text mining
i. Unstructured or semistructured data
ii. Corpus (and corpora)
iii. Terms
iv. Concepts
v. Stemming
vi. Stop words (and include words)
vii. Synonyms (and polysemes)
viii. Tokenizing ix. Term dictionary
x. Word frequency
xi. Part-of-speech tagging
xii. Morphology

What is Natural Language Processing (NLP)?

It's Structuring a collection of text; Old approach: bag-of-words,New approach: natural language processing. Its a very important concept in text mining, a subfield of artificial intelligence and computational linguistics What are the challenges in NLP

1. Part-of-speech tagging
2. Text segmentation
3. Word sense disambiguation
4. Syntax ambiguity
5. Imperfect or irregular input
6. Speech acts

The dream of AI community is to have algorithms that are capable of automatically reading and obtaining knowledge from text. WordNet; A laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets Sentiment Analysis :A technique used to detect favorable and unfavorable opinions toward specific products and services
Application Case : Mining for Lies


Statements Transcribed for Processing
Cues Extracted & Selected
Text Processing Software Identified Cues in Statements
Text Processing Software Generated Quantified Cues
Classification Models Trained and Tested on Quantified Cues
Statements Labeled as Truthful or Deceptive By Law Enforcement

Category	Example Cues
Quantity	Verb count, noun-phrase count, ...
Complexity	Avg. no of clauses, sentence length, …
Uncertainty	Modifiers, modal verbs, ...
Nonimmediacy	Passive voice, objectification, ...
Expressivity	Emotiveness
Diversity	Lexical diversity, redundancy, ...
Informality	Typographical error ratio
Specificity	Spatiotemporal, perceptual information ..
Affect	Positive affect, negative affect, etc.

Revision Questions

Example ✎. ....

Solution: .... □

EXERCISE 19. ✍ ....

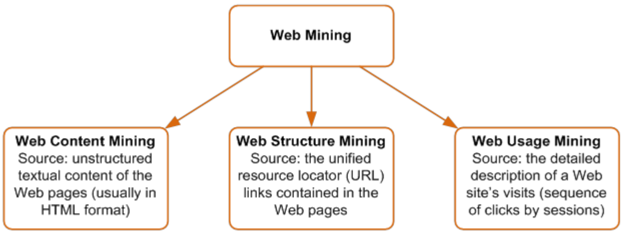
LESSON 6
Web Mining

Learning Objectives

1. Describe Web mining, its objectives, and its benefits
2. Understand the three different branches of Web mining
   (a) Web content mining
   (b) Web structure mining
   (c) Web usage mining
3. Describe the usage of Web mining
4. Differentiate between text mining, Web mining and data mining

6.1. Web Mining

Web mining (or Web data mining) is the process of discovering intrinsic relationships from Web data (textual, linkage, or usage)


Web Mining
Web Content Mining
Source: unstructured textual content of the Web pages (usually in HTML format)
Web Structure Mining
Source: the unified resource locator (URL) links contained in the Web pages
Web Usage Mining
Source: the detailed description of a Web site's visits (sequence of clicks by sessions)

6.1.1. Web Content/Structure Mining

Mining of the textual content on the Web Data collection via Web crawlers

- Web pages include hyperlinks
- Authoritative pages
- Hubs
- hyperlink-induced topic search (HITS)

6.1.2. Web Usage Mining

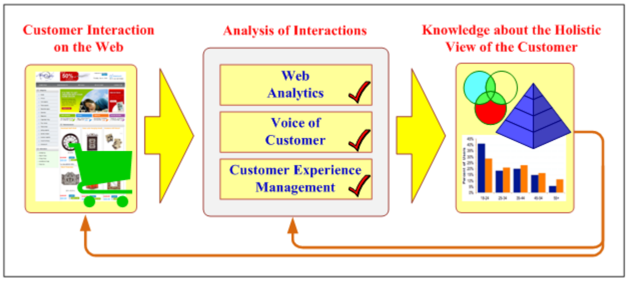
Extraction of information from data generated through Web page visits and transactions...

- data stored in server access logs, referrer logs, agent logs, and client-side cookies
- user characteristics and usage profiles
- metadata, such as page attributes, content attributes, and usage data
- Clickstream data
- Clickstream analysis

Web usage mining applications

- Determine the lifetime value of clients
- Design cross-marketing strategies across products.
- Evaluate promotional campaigns
- Target electronic ads and coupons at user groups based on user access patterns
- Predict user behavior based on previously learned rules and users' profiles
- Present dynamic information to users based on their interests and profiles...

Web Usage Mining (clickstream analysis)


Customer Interaction on the Web
Analysis of Interactions
Web Analytics
Voice of Customer
Customer Experience Management
Knowledge about the Holistic View of the Customer

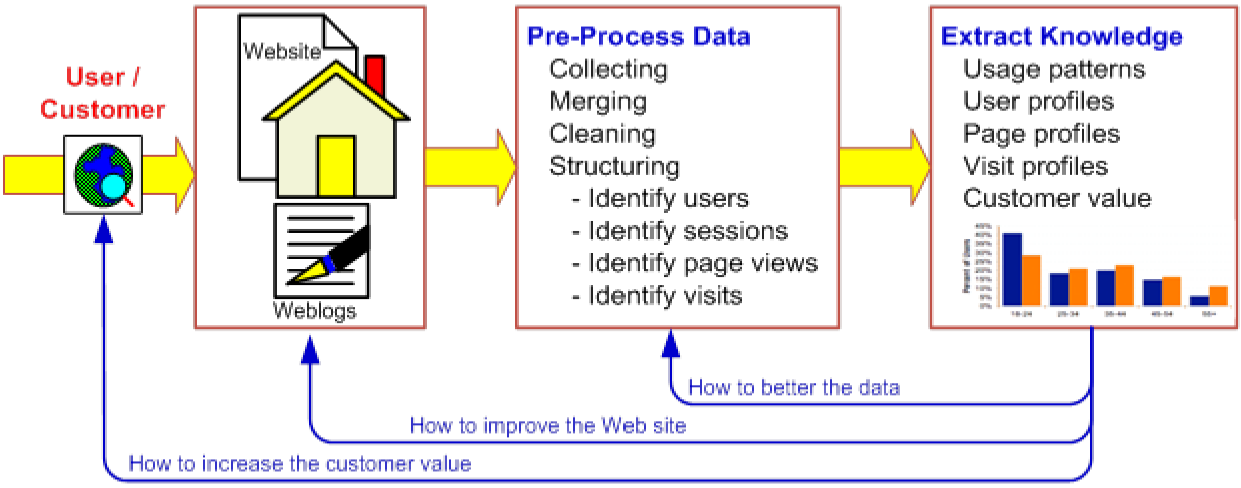


Figure 6.1: Web Mining Success Stories

Web Mining Tools

Product Name	URL
Angoss Knowledge WebMiner	angoss.com
ClickTracks	clicktracks.com
LiveStats from DeepMetrix	deepmetrix.com
Megaputer WebAnalyst	megaputer.com
MicroStrategy Web Traffic Analysis	microstrategy.com
SAS Web Analytics	sas.com
SPSS Web Mining for Clementine	spss.com
WebTrends	webtrends.com
XML Miner	scientio.com

Revision Questions

Example ✐. ....

Solution: .... □

EXERCISE 20. ✍ ....

LESSON 7

Role of IP in competitive Intelligence gathering and analysis

- What is Competitive Intelligence (CI)?
- Why is CI important? – Where to start from? – When to use competitive intelligence? ;Common goals.
- Why is CI required? – Porter Five Force Model
- CI versus Market Research
- Role of IP; information from patent
- References & Acknowledgements

7.1. Competitive Intelligence (CI)

There is a Chinese saying:

- **Know thy-self, know thy competition, and get it right almost every time.**
- **Know thy-self, not know thy competition, and get it right about half the time.**
- **Not know thy-self, not know thy competition, and get it wrong almost every time.**
- Competitive Intelligence is a process which gives insights into what might happen in the near future. This process involves stages from data to information to intelligence. It differs from data and information since it requires some form of analysis. The purpose of this analysis is to derive some meaning from the gathered data and information. This analysis and filtering of the data & information helps one to act on it and understand the options, giving an opportunity to make way forward decisions.
- When we present “intelligence” to people, they can draw a conclusion and make an important decision quickly. We need not collect and analyze all information for an exact picture, but get enough information so that we can

conclude what's going on. An outcome of CI is that it puts conclusions and recommendations upfront.

- CI is the analytical process that transforms disaggregated competitor intelligence into relevant, accurate and usable strategic knowledge about competitors, position, performance, capabilities and intentions

- The objective of competitor intelligence is not to steal a competitor‘s trade secrets or other proprietary property, but rather to gather in a systematic, legal manner a wide range of information that when collated and analyzed provides a fuller understanding of a competitor firm's structure, culture, behavior, capabilities and weaknesses.

7.1.1. The Importance of Competitive Intelligence

- It identifies the source for best practices – the only real way to isolate and find "best practices" is to engage in some form of CI; otherwise one may end up relying on crude and generic type benchmarking data.

- Helps formulate strategy through an understanding of the industry, oneself, and the competitors. CI is the essence of strategic business analysis.

- Helps identify areas for improvement as well as risks and opportunities.

- Isolates performance gaps in relation to the competition.

- Besides other critical things "strategic decision" is also based on certain assumptions. CI provides continuous flow of new intelligence which is essential keeping in mind the dynamics of requirement and market.

- CI helps test and validate these assumptions & it also fills in gaps, covering areas that were not considered in our assumptions.

7.1.2. Where to start from?

Fundamental information can be gathered from:

- Information about your industry (such as monitoring trade journals)

- Financial transactions (such as credit ratings)

- Publicly released information (such as regulatory filings by a competitor)
- Non Confidential Information between buyers and sellers to identify who the competition is.
- Screening Pat & Non Pat Publications of Competitor.
- Press releases, analyst reports, trade journals, transcripts of speeches, and other published sources of information.

When to use competitive intelligence?

- When competition increases from firms outside ones industry's traditional boundaries.
- • When consumers and customers become increasingly sophisticated and knowledgeable, demanding more and openly comparing products, services, and sources.
- When changes occur continuously in the nature and variety of the products and services one must offer to continue to compete.
- When significant changes occur in the ownership or senior management of firms in the industry, which may bring in new operating or marketing philosophies.

Some common goals of competitive intelligence:

- Detecting competitive threats
- Eliminating or lessening surprises
- Enhancing competitive advantage by lessening reaction time
- Finding new opportunities

7.1.3. Why is it required?

- Helps in uncovering the latest trends driving the pharmaceutical and biotech industry

- Identifying increasing trends such as mergers and acquisitions and how they impact on the industry
- Addressing the threat of expiring patents, dry pipelines and a difficult economic climate and strategies that manufacturers are adopting to overcome these hurdles
- Assessing current and future CI challenges and forecasting trends for CI in the pharmaceutical industry
- Utilizing intelligence information as a strategy determined to success in this competitive market
- Helps companies anticipate external threats and opportunities in a timely manner so companies can respond strategically
- R&D needs this information to stay abreast of competitors' developing products and services.
- Marketing needs this to monitor competitors' offerings, customer lists and price points.
- Sources like R & D Focus Drug News, Scrip World Pharmaceutical News, Drug Industry Daily and Pharma Business Week are used to filter the relevant information and thus create pharmaceutical competitive intelligence topics for the entire company to track, as well as topics for specific departments or individuals within the organization
- It helps in providing insight that drives new customers, helps retail current ones, streamlines operation, reduce costing.

Helps in market research by following ways:

- In answering what markets should we be getting into.
- Upside potential of a drug
- What drugs should be further segmented by form or strength.
- Who are primary competitors.

7.2. Porter Five Force Model

According to this there are five forces impacting an industry:

1. Customers – They have the power to bargain for lower prices and force certain kinds of change within your industry.
2. Suppliers – Suppliers may have certain power to dictate prices and options upon a company.
3. Substitutions – Your market share is subject to change from substitute products or a new use of an existing product.
4. New Competition – New entrants into the marketplace are not uncommon in this global world and they can be difficult to identify.
5. Existing Competition – Your organization is currently competing for market share, trying to position itself as a leader.

Knowing one‘s industry is an important building block of CI.

- An understanding of which strategy one‘s organization follows is critically important to understanding what one needs to monitor in the competitive universe.
- According to Porter, all companies must ultimately fall back on one of three types of strategies:
  - Low cost provider -very difficult to maintain, only a few companies can execute on this (such as Wal-Mart).
  - Differentiation – provides a unique type product or service, such as Swatch vs. Rolex when it comes to wrist watches – one is very fashionable and reasonably priced and the other is high quality and much more expensive.
  - Highly Focused – very geographic or market niche oriented. This is probably the more common type strategy for most companies. Examples include banks, software companies, media companies and airlines.

7.3. CI vs. Market Research

- Market-research teams typically focus on supporting a specific franchise or therapeutic area by identifying market gaps for their products to fill. Competitive-intelligence teams, on the other hand, use similar data—along with countless other resources—to predict market changes, preempt competitor activities, and develop strategic contingency plans.

- There are large areas of overlap; market research can also be considered as a subset of competitive intelligence

- The pharmaceutical industry in particular has been integrating competitive-intelligence and market-research efforts to engender greater strategic impact from their market-research functions

- Trend changed; Outsourcing became a quiet solution to data collection and cost savings. Most pharma companies outsourced all their market-research work during the 1980s

- The 1990s, however, saw a shift back to building internal market-research teams, which has lasted to the present. Those teams were built to support the blockbuster generation: Find a market gap, use the data to develop a first-to-market product, and brace yourself for the sales upswing. As scores of blockbuster products approach patent expiration and companies begin looking to lifecycle management teams to expand their franchises, market-research resources are settling to a more comfortable size

7.3.1. Role of IP

- When analyzing technological development in a field, the first concrete evidence of a new product, drug or industrial process may be a published patent document, often a patent application. Patents and other forms of Intellectual Property are often good indicators of what competitors are involved in.

- Intellectual Property usually has three broad functions within a corporation:

  - As a tool to protect price and market share by excluding others from a specific marketplace (patents) and as a guarantee of channels to market and goodwill (trademarks).

  - As insurance against legal action by other patent holders, operating to mitigate risk of infringement.
  - As a financial asset in the high-stake game of strategic alliances, in which technology is licensed, swapped, assigned, mortgaged, or held as a blocking strategy.

- Analysis of IP may reveal a great deal about a competing corporation‘s technology strategies. Patent and other IP data is available publicly from a number of sources, and is standardized to a high degree. Several companies offer software packages to collate, map and chart patent holdings to indicate patent filings over time, density/frequency in specific technologies, international equivalents, citation history, and activity of particular individuals, companies or groups.

7.3.2. Information from patents

Identification of competitors or collaborators
Assessment of human capital by analyzing inventor records for competing companies (Patent records may reveal jointly held patents with Universities or other research collaborations, indicating strategic use of human resources)
Assessment of competitors’ R&D effort and direction:

- Graphic mapping of the density and frequency of patent filings across all technologies for a competitor reveals the focus and intensity of their research efforts. Gaps in their IP portfolio may be discerned, and offer evidence of need to license or partner. Patent family searches (and also trademark searches) indicate the segments of the international market the competitor is targeting. Temporal profiles for patent filings may show a competitor is abandoning a particular field. Citation searches may reveal competitors "patenting around" a patent portfolio, filing improvements to a rival’s product line.

- Discover market trends, birth of new technologies

- Find new employees, consultants, and experts

Locate licensees:

- A thorough search of a company's area of patented technology may reveal newcomers to the field who should obtain a license to practice their patents.
- A glance at the front page of a patent can reveal a great deal about the quality of the document.

Reputable corporate or institutional source as assignee?

- If the patent is held by a private inventor, there is a probability the invention was not the result of a well funded R&D program. If the inventor drafted the patent (indicated by lack of a legal rep.) then there is a strong chance it is not of much worth

Patent Cooperation Treaty Filing?

- If the priority filing was a PCT app., this would indicate considerable funds has been expended for international filings, indicating some confidence in the technology.

Prior art cited? Literature cited?

- Studies indicate that patents issued to universities and research institutions providing generous citations reflect quality research.

Several inventors?

- Many inventors indicates well financed research team.

Continuations in part?

- Indicates ongoing serious research.

Prosecuted by solid law firm?
Certain patent firms specialize in particular industries and are not cheap. In theory a patent produced by such a firm would be well drafted.
Vigorous patent prosecution or lack thereof may indicate the priority placed on a technology in development. If a PCT application lies inactive for years, with little prosecution activity apparent, one can assume the invention will finally be abandoned.
Reviewing the legal history of a patent portfolio, as well as the infringement, opposition, re-examination, and other IP and trade-related litigation pursued by a company, indicates the company's degree of enforcement.

7.3.3. Explain the roles of IP

- Review Pipelines; therapeutic segments; access databases
- Look for product & other critical patent expiries & maintenance status.
- Patents Evaluation; Exclusivity Status; API Status; Generic Formulations
- Other patents of concern, if any?
- Continuous monitoring of newly published patent applications.
- IP fencing strategies.
- Keep track of IP situation so as to look for piggybacking opportunities
- Look for generic filings, Explore the generic market situation, checking activity in Non regulated markets
- Scrutinizing blocking patents for invalidation, purchase, license, or reverse engineering.
- • Analyze competitors patent
- Look for descriptive clues regarding a company‘s technological trajectory and the markets it is concentrating on.
- Commercial opportunities presented by expiration of patents held by rivals.
- Identify those involving a potential to manufacture under license (and other forms of joint venture) with other firms
- Patents citing their own early patents will tend to be of pioneers, while those citing other companies patents are usually “imitators”
- To identify those individuals or groups of inventors who are valuable as employees
- Sometimes there are difficulty in unearthing a company‘s patent portfolio.
- Sometimes firm operate under a group structure purposely try to baffle researchers by submitting patents under the names of numerous subsidiaries rather than that of parent company

- Difficult to determine which ones are commercially important to them
- Taking a trouble to renew a patent or filing it in other countries are probable indications that the owner places a greater commercial value to it.
- Considerable patent filings taking place in a certain field without this receiving much emphasis in a firms public statements

Caution

- No hasty assumptions
- Look for aberrations in a firms patenting behavior
- Regulatory filings
- Quotes from management and media reports

Indicators – to focus on

- Patents that appear to represent the cumulation of earlier research activity
- Completely new patents
- Filing large block of patents together – perhaps seeking to protect the spin-offs from the basic invention by simultaneously patenting many of them
- Where all research team members are named in a patent, if in others only some are recorded
- Patents are filed worldwide within a short time of each other
- E.g. Australia/ easy to apply/ filing there as matter of course
- Widely cited patents/ assembly list of competitors in a particular technical zone

Indicators – to focus on

- A highly cited patent represents core technology that other inventors have attempted to improve upon. When an important patent turns up in a search, the patents (and patent applications) citing it should be reviewed,

- Helping counter CI by screening speeches, publications, and presentations by employees to check for any inadvertent disclosure of vital information which might give clues

- Filing provisional patents around the larger firm's core technology can pave the way for an equitable joint venture or cross-licensing agreement.

- IP helps in positioning like identifying & buying the patent or the company, licensing or cross-licensing the IP, creating better technology and patenting around, suing for infringement or invalidating the patent by legal action or re-examination, and opposing pending applications.

Revision Questions

Example ✎. ....

Solution: .... □

EXERCISE 21. ✍ ....

LESSON 8

Developing a Comprehensive Competitive Business Intelligence

Introduction

- Defining Competitive Intelligence Programmes
- Common elements of designing competitive intelligence programs
- What are the objective of competitive intelligence programs
- What is the role the management accountant in competitive intelligence
- What is the competitive intelligence Process
- What are the tools and techniques for developing competitive intelligence

8.1. Defining Competitive Intelligence Programmes

Competitive intelligence programs are the foundation on which organizational objectives, strategies and tactics are built, assessed and modified. They permit organizations to assess both their industry life cycle and the capabilities of current and potential competitors in order to maintain or develop a competitive advantage. Competitive intelligence programs provide input for such decisions as which products, markets and business lines to invest in and develop, which to acquire or develop joint ventures around, and which to divest themselves of or exit. While there are many different ways of designing and implementing competitive intelligence programs, all have common elements:

1. Competitive intelligence programs focus on industries and on creating competitor profiles, particularly identifying organizational and performance implications of industry changes and of competitors' actions and reactions;
2. Gathered data (many unorganized, disconnected and unevaluated bits of input) become competitive intelligence (data that are organized and evaluated so that a firm gains new, different insights about its competition);
3. While individuals and/or units are formally charged with intelligence responsibilities, every organizational member is an intelligence antenna,

4. Competitive intelligence programs evolve to address changing critical issues and to permit organizational renewal; and
5. Competitive intelligence programs are not industrial espionage. Rather, they are the process of gathering, analyzing and using publicly available data. Obtaining confidential competitive information by nefarious means, and acting in clearly unethical or even illegal ways, is not competitive intelligence.

What are the objective of competitive intelligence programs

1. To provide an early warning of opportunities and threats, such as new acquisitions or alliances and future competitive products and services;
2. To ensure greater management awareness of changes among competitors, making the organization better able to adapt and respond appropriately;
3. To ensure that the strategic planning decisions are based on relevant and timely competitive intelligence; and
4. To provide a systematic audit of the organization's competitiveness that gives the CEO an unfiltered and unbiased assessment of the firms relative position.

8.1.1. Role the management accountant in competitive intelligence

Competitive intelligence is a process of gathering data, creating information and making decisions. Management accountants are trained to gather data, assimilate data into information and make decisions based upon information, frequently with their management counterparts. Competitive intelligence may also be viewed as a competitiveness audit, a concept that management accountants are familiar with. Management accountants' training and experience make them well-suited to the requirements of the competitive intelligence process. Management accountants may be actively involved in introducing a competitive intelligence process in several ways:

1. Identifying the need for a new or improved competitive intelligence process;
2. Educating top management and other senior managers about that need;

3. Developing a plan along with cross-functional team members for designing, developing and implementing the new, improved competitive intelligence practice, including its underlying architectures;
4. Identifying the appropriate tools and techniques for conducting competitor analysis;
5. Providing financial input, analysis and expertise to the competitive intelligence effort;
6. Contributing to and using competitive intelligence in target costing;
7. Ensuring that the competitive intelligence efforts are tied to the firm's goals, strategies, objectives and internal processes, as appropriate; and,
8. Continually assessing the new, improved competitive intelligence process and its implications for the organization, and continually improving the process.

8.1.2. What is the competitive intelligence Process

An effective competitive intelligence process allows the appropriate members of a firm to actively and systematically collect, process, analyze, disseminate and assimilate competitor information so that they can respond appropriately. There are many approaches to creating competitive intelligence. Corporate experience suggests that several elements are critical to an effective intelligence process. These include:

1. Define the business issue(s);
2. Determine the sources of competitive data;
3. Gather and organize the data;
4. Produce actionable intelligence;
5. Communicate results and findings;
6. Provide input into the strategic planning process; and
7. Provide feedback and re-evaluate.

Revision Questions

Example ✎. With the aid of examples explain the above mentioned process

Solution:

Tools and techniques for developing competitive intelligence Just as competition has increased for most firms during the past 50 years, so there has been an evolution of thought, practice, and tools and techniques that support competitive intelligence efforts. These tools and techniques can be categorized as strategic, product-oriented, customer-oriented, financial and behavioral.

a) Strategic Analysis Techniques Companies typically make superior profits either by entering a profitable industry or by establishing a competitive advantage over their rivals. Their strategy is usually defined by the answers to two basic questions: “Which business should we be in?” and “How should we compete?” The answer to the first question defines corporate strategy, which addresses issues such as diversification, vertical integration, entry and exit, and the allocation of resources within a diversified corporation. It emphasizes an in-depth understanding of the market, particularly of competitors and customers. The goal is not only to gain insight into current conditions but to anticipate changes that have strategic implications. The answer to the second question defines business strategy, or how the firm will compete within a specific industry or market. If the firm is to win, or even survive, it must adapt a strategy that establishes a sustainable competitive advantage. Strategic analysis has evolved significantly during the past 30-40 years, in many respects as competition has strengthened and become more global. Although the strategic analysis tools and techniques may have originally been developed to be used within a firm, many are equally applicable for competitive analysis. Tools and techniques that formerly were primarily internally focused have been turned around to focus more explicitly on the external environment and to analyze competitors in the same way that a firm would analyze and evaluate itself. Organizations use several strategic analysis techniques in developing competitive intelligence for their corporate and business strategies. Besides helping select the correct strategy, these techniques provide a framework for rational discussion of alternative ideas and the means to communicate the strategy throughout the organization. Some of these techniques are: 1) Industry classification analysis; 2) Core competencies and capabilities analysis; 3) Resource analysis; and 4) Future analysis.

b) Product-Oriented Analysis Techniques The pursuit of a sustainable advantage is typically the focus of corporate strategy. Protecting advantages has become increasingly difficult. Once the advantage is copied or overcome, it is no longer an advantage. It is now a cost of doing business. Ultimately, innovators are only able to exploit their advantage for a limited time before competitors launch a counterattack. Over time, organizations are forced to shift their cost (and price) and quality positions. Industries readjust their minimum acceptable level of quality and maximum acceptable price required to be a player in the marketplace. Revolutions in quality raise standards and then new revolutions shatter those standards. Innovations in product or process technology drive dramatic improvements in quality or reductions in cost. Since these cycles of change are growing progressively shorter, it is important for firms to regularly and systematically monitor competitors' products. One product-oriented intelligence technique that some companies have used for years and that more companies are emulating is reverse engineering/teardown analysis.

c) Customer-Oriented Analysis Techniques An important success factor for firms is the ability to deliver better customer value than the competition. Customer value can usually be achieved only when product quality, service quality, and value-based prices are in harmony and exceed customer expectations. Failing to meet customer expectations in any of the three areas leaves organizations in a situation of not having delivered good customer value. For example, if an organization offers poor-quality products or poor-quality service, then the price should fall. If an organization sets a price too high for a given level of product and service quality, sales should suffer. Providing great product quality and poor service quality will not maximize customer value. Organizations use several competitive intelligence techniques to help them determine how they are delivering customer value relative to their competitors. Some of these techniques are: 1) Customer value analysis; 2) Value chain analysis; and 3) Competitive benchmarking.

d) Financial Analysis Tools Financial strength obviously affects a company's strategic weaponry and the role that each product line or division plays in its portfolio. Thus, few competitor assessments would be complete without an in-depth financial analysis. Several financial analysis techniques can be utilized within competitive intelligence. Although these techniques have their limitations, thoughtful digging and analysis in the absence of hard data can help a firm understand the economic and

financial characteristics, capabilities and potential direction of competitors. These techniques include:
1) Traditional ratio analysis;
2) Sustainable growth rate analysis;
3) Disaggregated financial ratio analysis; and
4) Competitive cost analysis.
Tasks 1) With the aid of examples, explain the above mentioned financial analysis tools techniques □

Example ✏. ....

Solution: .... □

EXERCISE 22. ✍ ....

LESSON 9
OLAP Achitecture

Introduction

OLAP System Components
OLAP System Components Functions
OLAP Basic Features

9.1. Introduction

A successful company today has many decisions to make. The better those decisions are made, the more successful, and profitable, the company is. To many chief decision makers, the ability to analyze faster and better than the competition means better decisions, higher profitability, and more success. The optimization of the relational database (RDB) has enabled companies to efficiently collect data about transactions, giving decision makers more information to use. However, there is an upper limit to the amount of data that one can have in an RDB and still perform an efficient analysis on. On-Line Analytical Processing (OLAP) allows users to perform quick and effective analysis on large amounts of data. The data are stored in a multi-dimensional fashion that more closely models real business data. OLAP also allows users to access summary data faster and easier. They can then drill down into the summary figures to get more detailed data, if need be.

9.2. OLAP System Components

An OLAP system is comprised of multiple components. A top-level view of the system includes a data source, an OLAP server, and a client. The data source is the source of data to be analyzed. Data from the source are transferred or copied into the OLAP server, where it is organized and prepared to provide short query times. The client is the user interface to the OLAP server.

9.2.1. OLAP System Components Functions

- **Sources**

The source in an OLAP system is the server that supplies the data to be analyzed. Depending on the use of the OLAP product, the source could be an data warehouse, a legacy database housing corporate data, a collection of spreadsheets that holds financial data, or a combination of any of the above. The ability of an OLAP product to work with data from a number of sources is very important. Requiring that all source data is stored in a particular format or in a certain database is inconvenient for database administrators. It also reduces the power and flexibility of the OLAP product. Administrators and users find that OLAP products that allow data extraction from not only a wide variety of sources, but multiple sources, are more flexible and useful than those that have greater requirements.

- **Server**

The back-end of an OLAP system is the OLAP server. This is what does all of the work (depending on the model of the system), and where data that is actively accessed is stored. Different philosophies govern the architecture of the server. In particular, a major feature of an OLAP product is whether the server uses a multidimensional database (MDDB) to store the data, or a relational database (RDB). This section describes pros and cons to each approach.

- **MOLAP**

MOLAP stands for Multidimensional On-Line Analytical Processing. This means that the server uses an MDDB to store data. Because most OLAP products are based on an MDDB, the term OLAP usually refers to MOLAP as well. The purpose for using an MDDB is fairly straightforward. It can efficiently store data that are by nature multidimensional, providing a means of fast querying of the database. Data are transferred from a data source (as described above) into the multidimensional database, and then the database is aggregated. This pre calculation is what allows OLAP queries to be faster, since the calculation of summary data is already done. The query time becomes a function solely of the time required to access one piece of data, as opposed to the time to access many pieces of data and performing the calculation. The approach also supports the philosophy of doing the work once, and using the results over and over. Multidimensional databases are a relatively new technology. The use of MDDBs carries the same drawbacks that most new technologies do. Namely, they are not as robust as RDBs, and are not as optimized to the same extent. Another drawback is that most multidimensional databases are unable to be used while aggregating data, so it usually takes time for new information to become available for analysis.

- **Database Explosion**

Database explosion is a phenomenon of multidimensional databases. Though it is a problem that is experienced, it is difficult to explain how and why it happens. It appears to be related to the sparsity of the database and pre-aggregating the data. If a multidimensional database contains a small number of data points compared to the number of aggregation levels it performs, each piece of data will have a greater contribution to all data derived from it. When the database "explodes", the size of the database becomes magnitudes greater than it should be. It is difficult to determine conditions for database explosion, or to predict whether a particular configuration will explode. One approach that does seem to help the problem is dynamic sparse data handling. Dynamic sparse data handling allows a database to analyze it own storage patterns and optimize them to prevent database explosion.

- **ROLAP**

ROLAP is Relational On-Line Analytical Processing. The term ROLAP specifies that the OLAP server is based on a relational database. The source data are entered into a relational database, generally in a star or snowflake schema, which aids in fast retrieval times. The server provides a multidimensional model of the data, via optimize d SQL queries. There are a number of reasons to choose a relational database for storage as opposed to a multidimensional database. RDBs are a well-established technology that has had plenty of opportunities for optimization. Real world use has led to a more robust product. Additionally, RDBs support larger amounts of data than MDDBs do. They are designed for large amounts of data. A major argument against RDBs is that querying a large database with SQL to obtain summary data usually resulted in complex queries. An unskilled SQL programmer could easily tie up valuable system resources attempting to perform a query that is very simple in a MDDB.

- **Aggregated/ Pre-Aggregated Data:**

Fast query times are imperative for OLAP. This is one of the basic tenets for OLAP the ability to intuitively manipulate data requires quick retrieval of information. Generally, the more calculation that needs to be done in order to produce a piece of information, the slower the response time. Thus, in order to keep query times small, pieces of information that are typically accessed frequently, but need to be calculated, are pre-aggregated. That is, they are calculated and then stored in the database as new data. An example of a type of data that may be precalculated is summary data, such as sales numbers for months, quarters, or years, when the data that was actually entered was daily sales numbers.

Different vendors have different approaches as to what values need to be pre-aggregated, and how many values to precalculate. Approaches to aggregation affects both the size of the database and the response time of queries. If more values are precalculated, a user is more likely to request a value that has already been calculated, thus the response time will be faster because the value does not need to be requested. However, if all possible values are precalculated, not only will the size of the database be unmanageable, but the time it takes to aggregate will be long. Additionally, when values are added to the database, or perhaps changed, that information again needs to be propagated through the precalculated values that depend on the new data. Thus, updating the database can be time-consuming if too many values are pre-calculated for the database. Since it is common for a database to be off-line while aggregating, it is desired to keep aggregation times small.

- **Client**

The client is what is used to view and manipulate data in the database. A client can be as simple as a spreadsheet that incorporates OLAP features such as pivoting and drilling, it can be a specialized yet still simple report viewer, or it can be as complicated as a custom-built application designed for more complicated data manipulation. The World Wide Web is the newest form of client. This also bears the mark of a new technology; many web solutions vary greatly in the features included and the functionality of the solution as an OLAP solution. While the server is what provides the backbone of an OLAP solution, the front end is also very important. The server may provide a solid foundation for easy data manipulation, but if the front end is complicated or lacking in features, the user will not benefit to the same extent. The client is so important that many vendors focus on building only the front-end. What is included in these applications is a standard look and feel to the interface, pre-defined functions and structure, and a quick solution to more or less standard situations. For instance, financial packages are popular. A pre-built financial application will allow many to use common financial tools without having to design the structure of the database or common forms and reports.

- **Query Tool/Report Writer**

A query tool or report writer provides a simple tool to access OLAP data. They have an easy-to-use graphical interface, and allow users to create reports by dragging and dropping objects into the report. While a traditional report writer will enable a user to quickly produce formatted reports, report writers that support OLAP produce live reports. The end product is a report that allows viewers to drill down to detail, pivot reports, support hierarchies, etc.

- **Spreadsheet Add-In**

Many businesses currently do various forms of analysis on corporate data using a spreadsheet. In some ways, this is an ideal means of creating reports and viewing data. Analysts can create macros that treat the data in the desired way, a template can be designed so that when data is input, formulas calculate the correct values, and simple calculations don't have to be entered over and over. However, this provides a "flat" report, meaning that once a report is created, it is difficult to view it from different aspects. For instance, if a graph shows information for a time period, say a month. If someone wants to see the values per day (as opposed to the sum for the month), then an entire new graph needs to be created. New data sets need to be defined, new labels have to be added to the graph, and many simple yet time-consuming changes have to be made. There are also many areas where mistakes can be made, making it much more prone to errors. When OLAP functionality is added to the spreadsheet, one graph can be created, then it can be manipulated so that the user can see the information he or she needs, without the hassle of creating all possible different views.

- **The World-Wide Web as a Client:**

The newest member in the family of OLAP clients is the Web. There are a great number of advantages to deploying OLAP reports via the Web. Themost significant is that no special software needs to be installed for someone to access the information. This saves lots of time and money for any organization. Every Web product is different. Some make it easier to create Web pages, but are less flexible. Others allow you to create data views, then save them as static HTML files. These will allow someone to view the data on the Web, but they aren't able to actively manipulate the data. Others are interactive and dynamic, which allow them to be fully functional. Users can drill down, pivot, limit dimensions, etc. Before choosing a means of Web deployment, it is important to decide what functionality one needs from a Web solution, and then determine which product will fulfill that functionality best.

- **Applications**

Applications are a type of client that use OLAP databases. They are similar to the query tool or report writers discussed above, but they also incorporate greater functionality into the product. An application is usually more robust than a query tool.

- **Development**

Typically, OLAP vendors provide a development environment for users to create their own customized applications. The development environment is generally a graphical interface that supports object oriented development of applications. In addition, most vendors provide an API that can be used to integrate an OLAP database with other applications.

9.3. OLAP Basic Features

The basic features are those that are the core of the theory behind OLAP and multidimensional analysis. There are two key concepts to keep in mind when learning about OLAP: First, an OLAP implementation should allow many users access to the same data in whatever way they want to. Also, OLAP is providing a user a way to get the most information out of the data, and OLAP uses the idea that a human is going to be able to get most out of the data if they are able to access it any way they want. The user should be able to follow their thoughts when looking at the data, and the OLAP product should be able to follow along.

9.3.1. Multidimensional Conceptual View

Multidimensionality is key to OLAP. Some may claim that an OLAP product is simply a multidimensional database. That's not quite true. OLAP products may be based on a multidimensional database, but the important part is the conceptual view. ROLAP products are able to store data in a relational database while providing a multidimensional view of the data. Multidimensionality is important because of the multidimensional nature of business data. Common dimensions are time, geographical region, and product. An example of a data measure that can have this dimensionality it the number of products sold; another is the amount of money that is brought in (either during a time period, in a particular location, or for a product). Analysts may be interested in different numbers or reports. Thus, it makes sense to define the measures similarly, and redundancy in the database can be reduced. A multidimensional view is intuitive for human users. As the number of dimensions grows, it is more and more difficult for humans to visualize the model of the database. However, the visualization of a three-dimensional "cube" can be generalized to more dimensions.

9.3.2. Intuitive

Data Manipulation While the multidimensional view aides in the storage of data and query times, intuitive data manipulation is intended to reinforce the ability of humans to perform successful analysis. It is very common for a person to look at an initial table or graph that is a summary of some data, see a value that might not look quite right, and then want to learn more about what goes into that value. When using a static reporting tool (either on paper or on-line), it is necessary to generate an entirely new report in order to see more detail. This entails determining where the data comes from, defining an area, and then putting it into a report format. Depending on how automated a system is, it might not be too difficult. With an OLAP tool, it can be as easy as a double-mouse click, and that value is to be exploded into more detailed data. Another example is that sometimes understanding a table can be easier when it is rotated for some reason; an OLAP tool can easily change the axis of a table or graph to make them easier for the user to understand and gain knowledge from. This rule is to disallow complicated procedures to change a view. The flow of analysis is interrupted when, to change a view, a user must select the correct item form a menu, enter some information into a box, and then wait for the view to come up. Intuitive implies that there is a graphical interface and communication tools such as mouse clicks and drag/drop routines.

9.3.3. Accessibility

In practice, an analyst will need data from a variety of sources, including an OLTP database, a legacy system, or perhaps spreadsheet files. The OLAP server should be able to access these multiple sources seamlessly when queried. The OLAP database will contain much of the data, however most systems also have a legacy database (an old database) that contains historical data that has not been stored in the production database (the newer, in this case OLAP, database).

9.3.4. Multi-User Support

A robust OLAP product will allow multiple users access to data concurrently. Multi-user support also includes the ability of the server to allow individual users to use different analysis models.

9.3.5. Reporting Features

The reporting features are integral to a good data analysis tool. Some personal preference goes into the style of the reports that are created, but there are a couple of rules that reporting tools should follow in order to comply with full OLAP functionality.

- Flexible Reporting: The ad-hoc nature of analysis required for OLAP products dictates that a reporting tool should be flexible. A user should be free to show a display of the cube in any way; that is, there should be no constraints on the arrangement of the dimensions on the axes of a table, and cubes, when displayed, should be able to be rotated and sliced and diced.

- Uniform Reporting Performance: Again, to prevent a disruption of flow when performing analyses, the time it takes a tool to present a report should be consistent. This means that the queries should be run at a consistent pace, and also that the report is generated in consisted times as well. As far as the query is concerned, when the database is fully pre-calculated, there shouldn't be any problem with uniform query times due to the fact that there is just a retrieval time, and no calculation time.

9.3.6. Dimension Control

- **Generic Dimensionality**

This rule is intended to stipulate that all dimensions are equivalent, and that an operation performed on one dimension will similarly perform on another dimension. A dimension can be given special rules or formulae, but those rules should be extensible to all other dimensions. This is a controversial rule; many vendors want to make dimensions "smart" for ease of use by analysts. "Smart" dimensions usually include time or perhaps special accounts. Time dimensions will understand the structure of a year into months, quarters, etc., and will treat values dimensioned by time accordingly. Another example of smart dimensions is account dimensions, such as expense accounts. The extent to which a product adheres to this rule depends on the needs of the user. For some users, it is better that extra rules are already accounted for by the database. For users who need more flexibility, or have little use for the dimensions that are specially defined, a more generic approach is favorable.

- **Unlimited Dimensions & Aggregation Levels**

Ideally, an OLAP product should allow users to define as many dimensions as needed to fit their business model. Within each dimension, users should be able to determine the number of aggregations levels as well. While unlimited dimensionality is theoretically required, technically it is impossible. Statistically, it is determined that few business models will exceed 15 dimensions, 20 dimensions is a better number to target.

Revision Questions

Example ✎. ....

Solution: .... □

EXERCISE 23. ✍ ....

LESSON 10

The Business Value of Intelligence

Introduction

- How organizations can take advantage of an expanding BI toolset.
- Delivering Business Intelligence
- Today's BI
- Expanded Tools: what are the BI tools?
- The Growth of Self-service BI
- Determining Business Drivers
- What are the Elements of a BI solution?

10.1. How organizations can take advantage of an expanding BI toolset.

Business intelligence has been a technology mainstay of many organizations for decades. Today, it is playing an ever larger role in the success of businesses of all sizes. From small operations to global enterprises, more organizations are recognizing BI as a powerful tool that can help them survive and thrive in a challenging and competitive business environment. The need for BI or the expansion of BI for those that already use it in some form creates additional demands on IT teams, who frequently take a leadership role in BI initiatives. IT departments have faced numerous challenges in recent years. Budgets have grown tighter than ever, even as the IT group has been asked to deal with the security and governance challenges of bring-your-own-device (BYOD) programs and cloud computing. This white paper provides an overview of BI building blocks and deployments from classic data warehouses to Big Data and real-time analytics, as well as the demand for mobile solutions that can provide actionable intelligence from anywhere.

10.2. Delivering Business Intelligence

The good news is that BI has a well-deserved reputation for providing insights based on the inherent power of bringing data together (often from disparate sources) into a central repository, or data warehouse, from which it can be queried, analyzed and explored to guide decision-making. Organizations can gain substantial benefits and clarity from using well-defined reporting and analytic tools that provide all decision-makers with deep insights that they previously would have been unable to find. Expanding the power of BI to mobile, incorporating non-structured data from the vast realm of Big Data or drawing upon cloud-based resources are all simply extensions of the core BI story. The IT team was there from the birth of BI, and it is the group best prepared to help an enterprise evaluate and deploy whatever BI toolsets and data sources it needs.

10.2.1. Today's BI

To appreciate where BI is today and where it is headed it is helpful to look back to its origins, which can be traced to the previous century and a German émigré named Hans Peter Luhn, who joined IBM as a research scientist in 1941. He is credited with creating the term business intelligence in a 1958 article he wrote for the IBM Journal of Research and Development. In defining what he meant by business intelligence, Luhn simply cited the definition of intelligence straight from Webster's Dictionary: "the ability to apprehend the interrelationships of presented facts in such a way as to guide action toward a desired goal." That search for "the interrelationships of presented facts" to "guide actions toward a desired goal" is, more than 50 years later, still the goal of all BI deployments whether these facts are pulled from an organization's data warehouse or churned out from crunching terabytes of Big Data.

10.2.2. Expanded Tools

For the past 50 years, the quest for BI has been one of the drivers behind computer science, and BI has benefitted from a long series of breakthroughs in both hardware and software. Today, with the ability to purchase terabyte hard drives at commodity prices and the ubiquity of multi-core processing, much of the focus is on the software that can be brought to bear on vast data stores. In its most simple description, BI could be looked upon as a two-step process: Step 1, gather a lot of data. Step 2; pull actionable insights from that data. The basic tools of BI, the reporting, analytical and predictive tools that make its actionable insights so valuable to an organization, include the following:

1. Reporting tools:

   Reporting tools are the bread and butter of BI. This functionality enables an IT department to define and run recurring reports against data imported from across an enterprise. Whether referred to as an organizational data store, a data warehouse or any number of other terms, the power of classic BI is seen in the gathering of data and then running reports against it.

2. ii. Analytic tools:

   Analytic tools support the exploration of data. They go beyond fixed reports to seek out unusual insights that might otherwise go unseen. For example, a retailer might find that a hunting jacket popular in Maine isn't as popular in Texas and adjust its sales strategies in those states accordingly.

3. Predictive tools:

   In a way, predictive tools can be seen as a third step along the path of BI from reporting to analytics to prediction. Of course, plenty of predictive value can be derived from solid analytics, but the emphasis on predictive tools is generally on real-time or near-real-time results.

   The wealth of data and promise of BI have inspired thinkers for centuries. Here are some classic thoughts on the matter:"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay!" "Without Big Data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway.""Data is a precious thing and will last longer than the systems themselves." "Numbers have an important story to tell. They rely on you to give them a clear and convincing voice." "Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world." Predictive analytics is based on feeding data into an algorithm to generate what's predicted to be the most effective response. To an extent, the need for predictive analytics has been driven by the world's move to mobile devices, as well as the demands of the mobile ad space. Vendors have invested heavily in finding out who will respond best to which ad or, given what is known about certain visitors, which offers should be presented upon their landing on a website.

4. Contextual tools:

Contextual tools are focused on the interaction between data and context. The context can be as simple as matching ZIP code demographics to sales, or as subtle as evaluating the brand of smartphone a person is using, the time of day and even his or her location (from the smart phone's GPS) to determine how to analyze data points or how to fine-tune real-time predictions.

5. Cognitive tools:

   Cognitive tools incorporate models that attempt to replicate the weighing of facts and context that goes into human decision-making. The quest for cognitive tools goes back to the birth of BI, and later, the emergence of artificial intelligence. A cognitive element is seen in predictive analytic tools that incorporate machine learning. Machine learning simply means that once a system makes a prediction (such as whether a user will click on a proffered link) the results are monitored and fed back into the system, with the goal of enabling it to make smarter choices in the future.

6. Social media tools:

   Social media tools are used to import and analyze non-structured data from the vast realm of social media. For a business, this could involve gaining a feel for customer sentiment by analyzing how the company is perceived through automated searches of blogs, Twitter feeds, Face-book posts and other social media. The same information can be analyzed to predict trends and patterns that could better inform an organization's BI efforts.

10.2.3. The Growth of Self-service BI

For as long as BI has existed, enterprises have held a strong desire for self-service BI. With classic BI, the IT team works with business groups to design and run a set of recurring reports, while specialized analysts use complex tools to run their own searches and ad hoc queries against the same data store. Of course, it wasn't long before knowledge workers started making requests for custom reports. Traditionally, such requests were submitted to an IT group or analyst group, where the report would be created. Such requests demanded resources from the IT and analyst groups, especially when the person requesting the report perhaps after having waited days or longer to get it later added other variables to the request. Now available user-friendly tools can create ad hoc reports and let regular business users design, test and modify their own queries. These tools have become very popular and have boosted bottom-line productivity. Tools designed with a user-friendly dashboard interface allow users to explore data and drill down into exacting detail whenever needed. Today, regular users can create their own reports and define their own key performance indicators without having to go to night school to learn a sequential query language.

- **Mobile Trends**

The adoption of smartphones and other mobile devices has been so strong that researchers at Stanford University reported that most people are more likely to leave the house without their wallet than forget their smartphone. Similar studies have found that the majority of people take their smartphones to bed with them, and that we come to view them as extensions of ourselves. Against that backdrop, it isn't surprising that decision makers throughout an organization want to be able to use mobile devices to access the same kinds of BI tools that they have come to depend on when using desktop or notebook computers. Mobile applications also help workers for whom mobile work is a requirement field technicians, delivery drivers, plant workers and a world of others who can do their job better by tapping into BI and analytics (whether to determine an optimized delivery route or to project mean time to failure for a piece of equipment) from wherever they may be working. The mobile trend extends the path on which business analytics has been growing. A generation of desktop BI and analytics tools has freed BI from the back office and pushed it to the edges of the enterprise so that everyone can make better decisions. Mobile is a powerful tool for extending this practice. The challenge for IT departments is to find the best mobile self-service tools. Trends such as the consumerization of IT and BYOD have sometimes forced the IT department into a reactive position of accommodating multiple apps that have been downloaded onto users' mobile devices. Some organizations have taken a pre-emptive position by creating their own internal app stores where employees can download apps that have been vetted and integrated into back-end BI systems. Most new BI solutions support the mobile workforces that organizations employ. Many of these solutions let mobile workers use a smartphone or tablet to make queries, extract data, run analyses and conduct other BI-related tasks.

- **Cloud Trends**

Cloud computing represents an interesting challenge for IT groups. The cloud can look enticing: It offers all of the computing and storage power an organization may need, without the capital expenditures of deploying its own servers. The cloud certainly makes for easier capacity planning too. An enterprise can spin up instances as needed, and spin them down (and stop paying for the resources) when not needed. But IT teams must maintain careful oversight and control of cloud resources. Storing an organization's data on devices that are controlled and owned by another entity, perhaps located in another country, raises a number of questions with regard to security and governance. Many organizations are taking a cautious approach to the cloud. But the basic value proposition is strong enough that small enterprises wishing to deploy or expand BI platforms, or groups seeking to capture and explore immense volumes of Big Data, may be drawn toward cloud-based resources. This is an area where the IT department can provide guidance in assessing security and governance concerns. IT may also be able to help with the architectural design of hybrid solutions that incorporate cloud resources while maintaining local control of data to meet industry compliance or internal governance restrictions. The consumerization of IT, again, can be a complicating factor. In the same way that workers want BYOD, many are also launching their own clouds, which can be accomplished in minutes. Launching cloud-based resources in such an ad-hoc manner can make it difficult for the IT staff to maintain control of enterprise information. When considering cloud resources, an organization is well served to remember that its BI platform, to a great extent, constitutes its crown jewels. Wherever the infrastructure is deployed whether in the back office or in the cloud it must be carefully secured and protected, which provides good arguments for IT involvement.

10.3. Determining Business Drivers

Determining the business drivers for a BI implementation should be a first step for all projects whether planning to introduce BI for the first time, or expanding an existing BI solution to serve new groups or purposes. This means walking stakeholders through a discussion of what they need and why they need it. Analysts and someone with systems integration experience would add value to this process. The purpose of this discussion is to define needs and propose solutions — while identifying a set of metrics that can be measured both before and after deployment to gauge the success of the project. If an organization already has a data warehouse, the project could be as simple as defining a set of reports for a specific group and providing it with easy-to-use tools to enable self-service ad hoc reporting and analytics. For groups with heavier analytical needs, the IT team can provide a separate data mart (a subset of the entire data warehouse) with data structured (using a dimensional model) to more directly meet their needs. Because a data mart is essentially structured to answer common questions, a sales department might be best served by one data mart, finance by another, customer service or manufacturing by another.

The IT group should ensure that all data marts are plugged into the same central repository, or data warehouse, and that a common set of data structures, or dimensions, is established so that every data mart uses the same definition of time, product, customer, supplier, branch, etc. It can be surprisingly difficult to get different groups within the same enterprise to agree on what constitutes basic metrics. But without uniform definitions, the organization can't get to a single view of the truth, which should be a common goal of all BI deployments. When launching a new BI platform, it is perhaps even more important to identify the business drivers and the metrics that will determine success. To maximize the chance for success, a BI project should begin small and then grow as users and IT staff gain experience and expertise. Rather than rolling out a BI solution for the entire organization, the IT team should find the group that could most benefit from BI and work with it to create a pilot project with a set of pre-deployment metrics that can be used to measure success. Success breeds attention. Once a successful BI platform is provided to one group, others will want the same. So whatever is deployed should have a scalable infrastructure, one that can be built out on an as-needed basis.

10.4. Delivering Information that is relevant and actionable

Once the IT department has a good sense of what a group needs, the next step is determining the best solution to meet these needs. In many cases, a line of business can be served by an organization's existing BI infrastructure, by creating a set of custom reports and giving the LOB access to self-service reporting and analytic tools. If an enterprise can't use a solution that's already in-house, it should consider third-party products specifically designed to meet its needs. The discussion of which solution an enterprise decides to implement should be guided by the goal of providing relevant, actionable information.

10.4.1. The Elements of a BI Solution

An effective BI solution includes the following elements:

1. Software: The software platform that facilitates gathering, analyzing and reporting on the data is an essential key to a successful BI solution.
2. Hardware: Servers and storage are among the hardware components that provide the engine for disseminating information to clients.
3. Appliances: Appliances are preloaded, all-in-one solutions that package hardware and software components together.
4. Services: A BI solution's holistic components hardware and software must meet the organization's reporting requirements with implementation and configuration services.

Revision Questions

Example ✐. ....

Solution: .... □

EXERCISE 24. ✍ ....