Date 25.7.2017
Page

Data warehousing and data mining are the foundation of business intelligence.

Syllabus

→ Introduction
→ Data Warehousing
→ Business Analytics and Data Visualisation

- Under introduction :

BI impacts :
↳ improving operational performance, quality, speed, dependability, quantity, flexibility, cost
↳ improving customer service ; Feedback from customers to find loopholes and make amendments accordingly
↳ Identifying new opportunities ; track the product

Capabilities of BI :
↳ Gathering of data (from various data sources)
↳ store
↳ analyze
↳ present the result in the form of results and graphs

Theories behind BI :
↳ OLAP (data cube) and OLTP (databases)

Characteristics :
↳ easy way of handling the system

Scanned by CamScanner

(R) 3 copy

↳ should monitor the indicators (profit, loss)
↳ reporting facilities (graphs or any visual representation of result or the analysis)
↳ consistency, quality

Implementation:
↳ Business Analysts (Power users)
↳ Managers, Executives
↳ IT team (Engineers)
↳ Processes involved in making strategies and planning computerized.

· Under business analytics and Data Visualization.
↳ Automated Decision Support (ADS): tools that covers business rules + predictive model for the use of business to take decision on ~~sp~~ repetitive problems.
↳ Web Intelligence: tools used for visual interaction report

→ Data, Text and Web Mining
→ Technologies Enabling Business Intelligence
→ Management and future of Business Intelligence

Scanned by CamScanner

What is Business Intelligence?

BI is an umbrella-term which comprises of strategies, processes, data application, techniques, technology, architecture which are used by an enterprise to gather data, understand data, perform processing, obtain results and present analysis of the result.

The key output of a BI is decision (information, knowledge also)

Input is data (internal or external)
↳ Eg. business ranking of organization

- BI must be able to handle structured as well as unstructured data.
- Results presented via scorecards, dashboards etc

BI solution supports BI product
↳ makes use of BI tool.

BI tools are provided by BI vendors and BI solution is deployed in an organization.

Eg. A grocery chain collects detailed data of customers on purchase by using customer reward card and uses BI sol^n to categorize customers and to develop knowledge on how offers should be customized.

Scanned by CamScanner

Real-time BI : real-time input to decision makers, whenever needed so that business processes are not shown.

Operational BI : focuses on operations such as quantity cost profit loss rather than planning or generating insights

* Factors driving BI
→ Large amount of data
→ Complicated decision
→ Quick decisions
→ Technological process

(1) Exploring Data Volume
→ Organizations have accumulated large data volume about aspects such as sales, customers, complaints etc.
→ BI solutions provide managers the ability to effectively utilize this larger data volume

(2) Increasingly complicated decision.

With increase in competition across countries. Availability of too much

Scanned by CamScanner

information, unstructured data from various sources such as website, search, media etc.
Decision making has become complicated which has been made easy by BI solution.

(3) Need for Quick Reflexes.

Market and environmental influences may affect company capitalisation for which managers need to take decisions before window of opportunity closes.

BI solution helps to overcome delays in
→ converting data from various sources
→ integrating information across sources
→ making results available to decision maker.

(4) Technological process is made possible by development in decision support system, resource planning system, data mining and text mining.

Scanned by CamScanner

Date 28/7/2017
Page

Business Intelligence and related technologies

* Knowledge Management (KM)

It is about doing what is needed to get the most out of the knowledge resource. It focuses on creating, sharing and applying knowledge.

Knowledge that can be imparted easily or verbally → implicit
Knowledge that cannot be imparted easily → explicit (tacit)

KM system examples:
→ Database of information that are retrieved from people who had already tackled a problem and found a solution.
→ Contact details of a person who has expertise in a domain
→ Feedback from customers and employees.

NOTE: BI focuses on internal and external data whereas data warehousing and data mining focuses ~~internally~~ primarily on internal data and structured data.

Table 1.1 in the book.

Scanned by CamScanner

* Business Intelligence in Contemporary Organization :-

(i) Contributions of BI :

→ spreading of real time information in a user-friendly format by providing users a single point of access to important information.

→ Creation of new knowledge based on past. Eg. arriving at new insights based on customer behaviour pattern.

→ responsive and anticipative decision; meaning decisions that have to be taken to compete with current market trends. Eg. what kind of products an organization should focus on.

→ improving planning for the future. Eg. recruitment of manpower or investment in R and D.

* Obstacles to BI

→ Huge initial and ongoing investment on BI.

→ Business events not being consistently defined throughout the organization.

Technological obstacles

(i) BI tools are difficult to use.

(ii) BI training is expensive.

Scanned by CamScanner

Organizational drawbacks :-

(i) ~~(iii)~~ Profits that are received is difficult to justify.

Technological

~~(ii)~~ (iii). Difficult to customize as per user needs.

Organizational.

(ii) Finding single solution for all BI needs for an organization is difficult.

Scanned by CamScanner

Chapter 2

Business Intelligence Capabilities

(a) Organizational memory
(b) Information Integration
(c) Insight Creation
(d) Presentation Capabilities

(a) It is a storage of information and knowledge in such a way that it can be accessed later.

(b) Ability to link structured and unstructured data from variety of sources

(c) Ability to develop new insight and use them in short term & long term to make better decisions.

(d) Ability to use appropriate reporting and balanced scorecard tools and thereby make ~~the~~ BI more valuable to users.

* Features of BI :-

(a) Benchmarking relative to competition and industry needs

Scanned by CamScanner

(b) Intelligence in terms of ability to search and utilise data across various sources.

(c) Convenience in terms of customization and connectivity for presentation.

* Factors leading to organizational memory capability :-

(a) Technological progress and cheap data storage have increased data volume in an organization and these data have to be stored for a longer period of time.

(b) Availability and use of unstructured data and explicit knowledge.

* Organizational memory (OM) capability

→ focuses on structured and unstructured data.
Structured data or information can be stored in database records.
Unstructured explicit knowledge can be stored in knowledge repositories.

→ focuses on intellectual resources of an organization.

Scanned by CamScanner

→ data consistency and accuracy are the two important necessities of OM.

* Technologies enabling OM capabilities.

→ Any transactional systems (database systems)
→ Enterprise resource planning systems. (ERP systems)
→ ERP systems — Softwares which help business collect information from all core areas of an organization for decision-making.
→ technologies that support the capture of unstructured information.

Factors necessating structured data and info may exist in a number of transactional databases that are mutually disconnected and often incompatible with each other. There are considerable amounts of data/information on websites, media etc.

* Capabilities of Information Integration

→ OM focuses on past data. BI focuses on latest data, information & knowledge.

Scanned by CamScanner

→ Context from various sources needs to be synthesised (sort, cluster, classify etc) before it can be used for producing insights.

→ Therefore structured and unstructured data from a variety of sources with real-time data is information integration.

→ It is important to integrate the following:
1. Structured information from ERP
2. External information
3. External written sources & knowledge from text mining.

→ Factors leading to insight creation
- Need for quick reflexes
- Increase domain complexity

Insight Creation Capability:
If we have the description of what had happened, we can find new trends or patterns ~~in~~ prior to events and actions.

If we have an understanding of what had happened, models could

Scanned by CamScanner

be developed. A prediction of future behaviour, which could depend on a model, that is developed based on understanding of the past, tested and refined using additional information over time.

* Technologies enabling insight creation.
→ Data Mining / Knowledge Discovery
→ Business Analytics
→ Realtime Decision Support (use of models of DM or DA to support operational decisions in a real-time fashion.
Eg. Airlines use real-time decision support along with presentational capability to make decisions aimed at flight arrivals & departure on schedule.

* Factors necessitating capability
(1.) Content : level of aggregating information
(2.) Format : Use of tables / charts & videos.

The key factors that affect content & format are :
(1.) Role of intended user.
(2.) Task of intended user
(3.) Preference

The only capability of presentational model should be user friendly based on role and task.

Scanned by CamScanner

Technologies enabling presentation

(1) OLAP ~~Cu~~ Cubes : OLAP is a multi-dimensional data model which gives different views of data.
Three different OLAP architectures are - MOLAP, ROLAP and HOLAP.

(2) Digital Dashboards : It is a GUI that organises and presents information in a way that is easy to read and provides at a glance inside to what is actually happening in an organisation.
Dashboards enable top level executives to gain a view of business and identify strengths and weaknesses in real-time freeing them from traditional bottleneck of manual reporting.

(3) Scorecard : Strategic planning and management tool used by organisation to align their business activities to their organisation's vision and strategy.
They measure periodic results against a pre-defined goal.

Scanned by CamScanner

Chapter 3

Major Theories and Characteristics of Business Intelligence.

Changing Business Environment and Computerized Business Support.

The present business environment and need for computerised support for managerial decision-making leads to the model, the business pressure-responses-support.

Remember fig 1.1

- Business Strategy is long term business planning.

- Globalisation → The process by which business or other organisation develops international influence or start operating on international scale

* Decision Support System

- Decision Making Process

Scanned by CamScanner

1. Defining the Problem
2. Constructing a Model that describes the real-world problem.
3. Identifying possible solutions to the modeled problem and evaluating the solution.
4. Comparing, choosing and recommending a potential solution to the problem.

* Need for Computerised Decision Support for System.

1. Improved Communication and Collaboration
2. Increased Productivity of Group Members.
3. Improved Data Management.
4. Managing Giant Data Warehouses becomes easy.
5. Overcoming cognitive limits in processing and storing information.

* Decision Support Matrix / Framework.

Remember Fig 1.2

* Types of Decision ~~Structure~~

1. Structured Decision : These are taken on processes that are routined and

typically repetitive for which standard solution exists.

(2) Unstructured Decisions are those that are taken on fuzzy processes and complex problems for which there is no clear-cut solution.

(3) Semi-Structured : Decisions taken on processes that have structured and unstructured problems.

* Types of Control

(1) Strategic Planning : It is a process to develop and deliver and support strategies, plans and tactics.

(2) Managerial Control : The means by which an organisation resources are directed, monitored and measured. (physical resources and intellectual properties)

(3) Operational Control : It is effective and efficient execution of specific tasks.

Scanned by CamScanner

ADS → Automated Decision Support.

Date 5/8/2017

Phases of Decision Making

1. Intelligence Gathering : Searching for conditions in environment that calls for decision.
   This step includes finding a problem, defining the problem, gathering data to solve the problems, constraints identification w.r.t the problem.

2. Design phase : also known as solution discovery is about, finding out possible solutions.

3. Choice : includes selection of best proposed solution

4. Implementation : solving the real problem or finding out if the solution is working or can a better solution be proposed.

Conventional Management Information Systems (MIS) and Management Science (MS) tools are insufficient for semi-structured and unstructured problems, hence decision support systems (DSS) are used.

Scanned by CamScanner

MS and ADS is used for structured decisions.

* MIS : These tools are computerized reporting tools for manufacturing, marketing, transportation, human resources and accounting organized & programmed in such a way that it produces reports on operations for every level of management in a company.

Features of MIS :

1. Produces reports on performance such as weekly, monthly and annual sales.
2. Reports can be viewed from web browsers and can be securely shared with others.
3. Reports can be exported in any format.
4. Suitable for any kind of business.

* Management Science

These tools are based on mathematical modelling or algebraic expressions that describe the problem.

Scanned by CamScanner

Date
Page

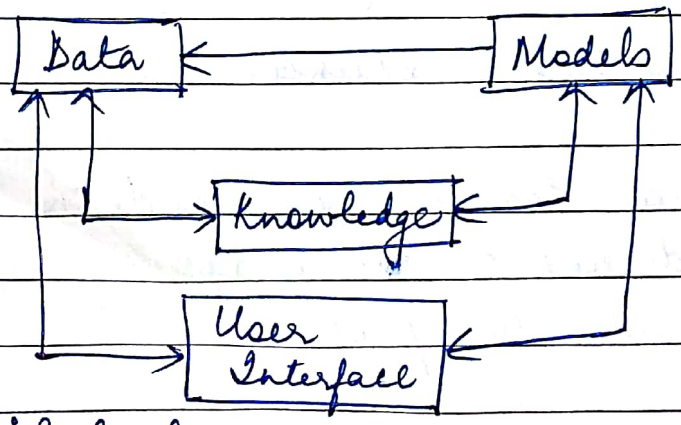
* Automated Decision Making : It is a rule-based system that provides a solution in one functional area to a specific repetitive managerial problem

* Computer Support for Unstructured Decision : Decision for unstructured problem can be partially supported by standard computerized quantitative methods. It is usually necessary to develop customized solutions.

* Computer Support for Semi-Structured Problems : This involves a combination of standard solution procedures and human judgement.

Remember fig 1.3
(for phases of decision making)

*



High-level architecture of a Decision Support ~~Structure~~ System

Scanned by CamScanner

DSS is framework for the process of supporting managerial decisions-making by modeling and employing the models for solution analysis.

EXPLANATION OF FIG. 1.5 (DSS)

To solve any problem or to achieve a particular strategy laid out, data is important.

Models being the second component uses the data for manipulation.

Knowledge denotes any intelligent system that is required to provide decision for unstructured or semi-structured problem.

User can interact with the DSS using user interface.

(Remember fig 1.6) → Evolution of BI.

* High-level architecture of BI (fig 1.7)

Data Warehouse → includes historical data along with real-time data, organized and summarized so that end users can easily view or manipulate data.

Business analytics tools :-

→ Reports and Queries : includes both

static and dynamic reporting, querying, multidimensional view etc.

→ Tools for advanced analytics includes statistical, financial and mathematical models.

→ Tools for data mining, web mining and text mining.

→ Business process management, also known as corporate performance management are the set of application and methodologies ~~that contains~~ used for monitoring and measuring performance.

User Interface: Dashboard, scorecards, digital scorecards.

Intelligent systems are customized purely on the business needs.

8/8/2017

Successful Business Intelligence

BI Maturity Level: Maturity level of an organization refers to ability of a particular organization for continuous improvement in a particular discipline.

Scanned by CamScanner

Four broad categories of maturity level:

1. Descriptive → reports and dashboards used to describe the measurements of any aspect of business which enables users to report on goals and actual values of any ~~methods~~ metrics ~~networks~~

2. Diagnostic → uses more advanced and critical capabilities such as interactive visualisation to enable users to drill ~~data~~ down more easily into data.

3. Predictive → discovering new insights or forecasting about future.

4. Prescriptive → insights from the predictive models are integrated into the business process to take corrective actions.

Successful Business Implementation

* Roadmap for BI at Organizational Level :-

1. Understand organisation's goals, strategies, decision making process and performance metrics to help define

Scanned by CamScanner

BI priorities.

(2.) Understand organisation's data (quantity, quality, availability)

(3.) Review the current BI maturity level

(4.) Review the existing IT infrastructure.

(5.) Review the current BI tool sets.

(6.) Work with stakeholders of organisation about goals and metrics.

(7.) Drive performance management from top-down and bottom-up.

(1.) BI should not only serve to hold the operational staff accountable but, also put the same emphasis on accountability and responsibility on senior executives as well.

* Use an agile modular approach.

→ Focus on breaking down the performance management needs by

Scanned by CamScanner

organisational priority and go about implementing BI using a highly iterative approach and shorter release cycles.

* Design an easy way to expand BI solution framework

→ The overall design of the BI solution framework should be easily expandable to accomodate features such as predictive and prescriptive analysis.

* Leverage existing IT infrastructure and BI tools.

* Focus on right metrics with each release

→ Implementation of BI should be in such a way that the identified metrics of an organisation should be measured w.r.t the target.

* Pay greater attention to data quality.
1. Source of authentication
2. Cleaning up of data
3. Data transformation rules must be incorporated.

Scanned by CamScanner

(4.) Implementing multiple levels of data validation in the loading process.

* Keep the ~~sy~~ presentation of data as simple as possible.

* Launch early and often thereafter (same as agile)

* Establish a centralized governance structure

* Encourage feedback and proactively manage change

9/8/2017

Data Mining Architecture

(1.) Enterprise Data Warehousing

* Pros: Single environment for data warehousing.

* Cons: Complex workload, querying, reporting, real-time access and loads will have to be done on a single data warehouse

Scanned by CamScanner

data silos : a separate database or set of data files that are not part of an organization's ~~enterprise-wide data administration~~.

Personal Requirement :
Complexity in managing the entire single environment needs more personal employees.
Slow to adapt to technology.
Monolithic architecture does not take advantage of rapid advances in server hardware.
Eg. parallel performance.

Expensive Maintenance : Multiprocessing based data warehouses makes maintenance cost increasingly difficult.

Scaling becomes difficult.
Limited environment agility.

(2) Distributed Independent Mart.
* Pros : Highly relevant to a particular department.
Appropriately sized, based on departmental needs.

* Cons : Data silos → implementing cross-organizational view of data silos requires significant time and expense.

Might result in data duplication.

Scanned by CamScanner

In case of data duplication, consistency of data has to be always taken care of.

Proliferation of technologies : Multiple hardware and software technologies are deployed across departments.

Data management costs across different departments.

Advantages of Hub and Spoke Architecture

1. Data mart tables can be generated on a hub without impacting data mart users and then be published to one or many spokes.

2. Clear-cut separation of management and user workload.

3. Ability to size a database as per the specific needs.

4. Provides scalability

Scanned by CamScanner

Date 11/8/2017

Remember → fig 5.5a – 5.5e
to be referred from slides.

* Selection of Data Warehouse Architecture

It depends on the following concerns:

1. Information interdependence between organizational units.
2. Top-level management's information needs.
3. Urgency of need for a data warehouse
4. Nature of end users tasks
5. Constraints on IT resources.
6. Strategic view of data warehouse prior to implementation.
7. Compatibility with existing systems.

Integration, Extraction, Transformation and Loading.

1. Enterprise Application Integration
2. Enterprise Information Integration
3. Extraction, Transformation and Loading.

1. Enterprise Application Integration.

→ provides solution to load data from

different source systems into a different warehouse.

→ It also involves integrating applications functionalities and is focused on sharing functionalities across systems, thereby enabling flexibility and reuse.

→ EAI is accomplished using SOA (Service-Oriented Architecture) which is a style of S/w design where service is provided by application's components over a network.

(2.) Enterprise Information Integration.

→ It is a tool which performs real-time data integration from a variety of sources.

→ It displays data in a uniform representation on a data-monitoring interface, meaning integrated data appear relational to end users.

→ EII also uses XML which is structured and may be extended at a later time.

(3.) Extraction Transformation and Loading

Scanned by CamScanner

→ Uses of ETL tools :

1. performs ETL process
2. maintains organizational data warehouse.
3. helps in detecting and removing any anomalies in real-time data.
4. should support documentation of how metadata changes as data moves between source and target.

→ During ETL process

1. all input files are written to a staging area which are designed to facilitate the load process.
2. defining business rules such as how the data will be used (cardinality constraints)
3. summarization rules containing dimension algorithms, i.e what dimension to include, what level within each dimension to include, data granularity.
4. standardization of encoded attributes:
5. consistency across columns or table names
5. data quality issues must be dealt.
6. business rules must be stored in central metadata repository.

Scanned by CamScanner

→ Criteria in selecting ETL tools:
→ ability to read and write from large number of data source.
→ automating, capturing and delivery of metadata.
→ a history of conforming to open standard.
→ an easy to use interface for all ~~there are~~ types of users.

Remember fig 5.6

DATA WAREHOUSE DEVELOPMENT

* Direct benefits from a data warehouse

1. Consolidated view of corporate data
2. Allows users to perform extensive analysis.
3. Timely information with better representation
4. Simplified access.

* Indirect benefits of data warehouse

1. Enhances business knowledge
2. Competitive advantage with other industries

Scanned by CamScanner

(3) Enhanced customer service and satisfaction
(4) Facilitates decision making
(5) Reforms business process.

* Crucial factors for the development of data warehouse.

(1) Support from the management end users
(2) Time frame and budget of the project.
(3) Complexities of business objectives, goals and strategies.

* Guidelines that need to be considered when developing a vendor list.

(1) Financial strength of the organisation
(2) Qualified consultants
(3) ERP linkages
(4) Market share
(5) Industry experiences
(6) Established partnerships.

* Data Warehouse Development Approaches:

(1) Enterprise data warehouse: Top-down development and follows a spiral approach.

Scanned by CamScanner

(2) Data mart approach: Development starts with the data marts of various departments followed by development of a centralized data warehouse at the organizational level if needed on a long term.

Remember table - 5.2.

* Tasks performed during the development of data warehouse.

(1) Determine the goals, objectives & business processes for an organization
(2) Design a functional data warehouse dimensional model
(3) Identify the data sources.
(4) Identifying ETL best practices for development & maintenance.
(5) Identify the need of ODS (operational data sources)
(6) Identify the appropriate technologies for ETL, metadata and data warehouse
(7) Launch a tiered deployment (DBA support, DB refresh, cloning, cleansing, security etc)
(8) Perform backups.

Scanned by CamScanner

Best Practices for Implementing a Data Warehouse

1. The project must fit with the corporate strategy and business objectives.
2. Follow incremental development of data warehouse.
3. Adaptability to any environment.
4. Develop a business supplier as well as IT relationship.
5. They should be complete by end of the product.

Risks involved in data warehousing projects

(i). No mission or objectives identified clearly.
(ii) Quality of data unknown
(iii) Skills not in place
(iv) Inadequate budget
(v) Lack of supporting softwares
(vi) Source of data not identified properly.
(vii) Too many changing requirements
(viii) Architectural and Design risks
(ix) Platform incompatibility
(x) Key IT people leaving the project.
(xi) Geographically distributed environment.

Scanned by CamScanner

Issues w.r.t a successful data warehouse

(i) Starting with wrong sponsorship chain.
(a) Need for an executive sponsor (senior member of the project, i.e a chairperson who has influence over the necessary resources)
(b) Need of an executive project driver, i.e someone who has earned the respect of other executives.
(c) Need of an IT manager to head the project.

(ii) Loading the warehouse with information just because it is available.

(iii) Believing that data warehousing design and database design are same.

(iv) Setting up expectations that you cannot meet.

(v) Choosing a data warehouse manager who is technology-oriented rather than a user-oriented (it becomes an issue)

Scanned by CamScanner

(vi) Data should be made accessible to the right people at the right time in the right format.

(vii) Believing that the problems are over when data warehouse is up and running.

(viii) Data warehousing is a continuous process as BI keeps evolving.

(ix). Focussing on ad hoc, periodic reporting instead of alerts.
→ Alert systems: These are BI tools which send notifications to the user that a certain event or condition has happened. Eg. when daily sales figure falls under a certain percentage of the target

(x) Delivering data with overlapping or confusing definitions

(xi) User participation in data warehouse development

(i) Data modelling (tasks performed)
→ define business rules
→ decide on aggregations

Scanned by CamScanner

(ii) Axis modelling (how data are to be retrieved from a data warehouse)

(iii) Indexing (for faster access)

19/8/2017

Massive data warehouses and scalability.

→ ~~#~~ Size of a data warehouse grows when:

→ amount of data in warehouse increases

→ complexity of user queries increases

→ as business expands

→ as number of concurrent users increases.

→ ~~scalo~~

Scalability of a data warehouse is said to be good when queries and data access function grow linearly with size of the data warehouse and data warehouse grows horizontally and vertically.

Scanned by CamScanner

Real-time data warehousing.

This is also called active data warehousing, zero latency data warehousing.

Need for real-time data warehousing:
→ increased data volume and accelerating update speeds.
→ to make fast and consistent decisions.
→ traditional data warehouses are not business critical as data is updated on a weekly basis.

Definition of a real-time data warehouse:
It provides integrated information repository to drive strategic and tactical decision support, within an organization
Eg. flight management dashboard application.

* Comparison between traditional and real-time data warehouses.

Traditional	Active
→ strategic decisions only.	→ strategic and tactical decisions
→ results sometimes hard to measure	→ results measured with operations.

Scanned by CamScanner

→ daily, weekly, monthly updation → data updation is every minute.

→ moderate user concurrency → high number of users accessing and querying the system simultaneously.

→ highly restrictive reporting; often uses pre-developed summary tables. → flexible adhoc reporting.

→ power users and internal users → operational staffs, call centers, external users.

Critical issues in real-time data warehousing

1. Not all data should be updated continuously
2. Real-time reports are different when produced at slightly different times.

Data warehousing administration and security issues.

1. Knowledge on high-performance software and hardware, network technologies, business insights, decision-making process, data warehouse design and

Scanned by CamScanner

good communication skill

* Security issues of a data warehouse

→ Establishing effective corporate security policies which should be communicated among all levels in an organization.
Eg. non-disclosure policies

→ Implementing logical security techniques to restrict access.
Eg. user authentication, encryption.

→ Limiting physical access to data center environment

→ Establishing an effective internal control review process with an emphasis on security and privacy

22/8/2017

Business Analytics Field and Overview

BA is a broad category of applications and techniques for gathering, storing, analyzing and providing access to data to help enterprise ~~to~~ users make better

Scanned by CamScanner

business and strategic decisions.
BA is also termed as analytical processing, BI tool, BI application. Eg. an analytic application used for credit scoring for a loan applicant might
(i) automatically accept or deny the loan application.
(ii) select a loan limit or
(iii) to select which other type of loan to approve.

Tools and techniques of BA.

Broad category of BA.
1. Information and Knowledge Discovery (OLAP querying, data mining, text mining, web mining).
2. Decision Support and Intelligence System (ADS, ESS, EIS, DIS, Web Analytics)
3. Visualisation (dashboard, scorecard)

* Microstrategy's classification of BA tools.

Enterprise reporting → These are pixel perfect report format.

Scanned by CamScanner

Cube Analysis → used to provide simple OLAP, multidimensional slice and dice analytical capabilities.

Ad hoc queries and analysis → perform on ROLAP to query a database for analysis. Statistical analysis and data mining: used to perform predictive analysis or to discover correlation between two metrics.

Alert system: to notify users whenever an event occurs.

* SAP's classification of BA tools:
  1. Operational
  2. Managerial
  3. Strategical

Executive Information and Support System

EIS:

(i) Computer-based system that provides timely and relevant information to the top executives to monitor organization's performance.

(ii) Provides exception reporting

(iii) Provides drill down operations

Scanned by CamScanner

(iv) ESE: goes beyond EIS to include analysis support communications, office automation and intelligent support.

* Capabilities of EIS and ESS
→ drill down
→ helps to measure critical success factors
Eg. customer satisfaction, quality of food w.r.t restaurant.

→ measures key performance indicators: these are measures which help know whether a business is moving in the direction of its set goals and objectives. Eg. profit, sale, loss.

→ helps produce status reports: A report on current ~~affairs~~ or the latest data available on key performance indicator (KPI).

→ helps do ad hoc analysis, i.e analysis done at any time.
→ helps provide exception reporting
→ provides slicing and dicing

Scanned by CamScanner

classmate
Date 23/8/2017
Page

Online Analytical Processing

These are activities performed by end users in online systems for generating reports, graphs, statistical analysis and query.

* Difference between OLTP and OLAP.

OLTP	OLAP
→ Operational Workers	→ Top-level executives
→ Repetitive transactions	→ Activity can be ad hoc.
→ Simple queries	→ Complex queries.
→ Answers user queries	→ Pattern analysis, trend analysis etc.
→ Nature of design is application-oriented.	→ Nature of design is subject-oriented
→ Answers close ended questions	→ Answers open ended questions.
→ Structured query language	→ DMQL
→ Nature of data is relational	→ Nature of data is historical, summarized and multidimensional.

Scanned by CamScanner

* Types of OLAP

MOLAP, ROLAP, Desktop OLAP,
Web OLAP → OLAP data accessible from web browser.
Desktop OLAP → Multidimensional analysis is done in a local machine.

* Characteristics of OLAP tools

1. Multidimensional conceptual view for formulating queries.
2. Transparency to the user
   - Underlying data repository
   - Source of data (nature)
   - Computing architecture

   } should be made known to user.
3. Easy accessibility
   - Batch and online access should be made available.
   Eg. ETL batch processing.
4. Consistent reporting performance.
   - Performance to run a report should not degrade as size of the database increases.
5. Client-server architecture

Scanned by CamScanner

(6.) Generic dimensionality
- Treat each dimension equally on the basis of operation capabilities.

(7.) Dynamic sparse matrix handling.
- OLAP tools must be able to handle missing values in the incoming data in case of real-time data warehouses.

(8.) Multi-user support
(9.) Unrestricted cross dimensional operation
- Should allow for drill down, roll-up and any other data manipulations across dimensions

(10.) Intuitive data manipulating
- Avoid menu or multiple trips to UI
- Enable consolidated path.

(11.) Flexible reporting
- Provide capabilities to the business users to arrange columns or rows or any cells in a manner that facilitates easy analysis and generation of information.

(12.) Appropriate number of dimensions and aggregation.
- Accomodate 15 to 20 dimensions within an analytical model and should have suitable levels.

Scanned by CamScanner

* 4 types of ~~E~~ Analytical Processing ~~Cates~~ Performed by Analysts

1. Categorical : Static analysis based on historical data. Eg. disease prediction
2. Exegetical : Drill down operation performed on historical data. Eg. drill down operation performed on revenue of entire business will give revenue per business unit ~~or~~ in a quarter.
3. Contemplative : Analysis done ~~on~~ by changing a single value to determine its impact.
4. Formulaic : Allows for changes to multiple variables.

26/8/2017

\# Reports and Queries

1. Runtime Reports : Reports which are automatically and periodically distributed to subscribers. Eg. weekly sales report sent to managers.
2. Ad hoc / On-demand reports : Reports generated whenever needed and data in the reports depend upon user demand.

Scanned by CamScanner

attrition : leaving

* Multilingual Support in which reports generated may be in certain language and can be translated to any other language.

(3) Representative types of report.

(i) Finance / Accounting : Eg. accounts receivable and payable reporting, balance sheet reporting, profit and loss income statement analysis.

(ii) Human Resource : Eg. career path reporting and analysis, employee expense reporting, time and labor management reporting

(iii) Sales forecasting and marketing : Eg. channel analysis, sales pipeline reporting and analysis.

(iv) Customer relationship management : Eg. customer acquisition retention and attrition. Eg. customer loyalty analysis.

(v) Product sales : Eg. discount and promotion analysis, inventory analysis.

Scanned by CamScanner

(vi) Operations and Supply Chain Management : Eg. call centre analysis, quality control analysis

(vii) Website Analysis : Web Navigation Analysis and web traffic analysis.

* Features that a report could have (additional features)

1. Report distribution to any touchpoint; Report content can be carried by emails, file servers, portals and in any format.

2. Self-subscription as well as administrator based distribution : An individual personally gets subscribed to some reports to have knowledge on other projects.

3. Delivery on-demand on-schedule or on-event

4. Automatic content personalisation : Report that provide only that information which is requested.

Scanned by CamScanner

Queries:

(1) Ad hoc queries : Any query that cannot be determined prior to the moment the query is issued. Eg. a user may ~~to~~ decide to place a query after he/she gets a report ; on-the-fly SQL query is generated.

Once reports are generated, managers would want to go for detailed investigation of the displayed data in the report. Such an analysis can be facilitated by multi-dimensionality

Multi-dimensional databases support multi-dimensionality in which data are organized specifically to support easy and quick access.

* Limitations of dimensionality

(1) MDB can significantly take up more computer storage than relational database

(2) Multidimensional products are costlier.

Scanned by CamScanner

(3) Loading MDB takes more time and resources.

(4) Maintenance space is costlier.

Advanced Business Analytics

(1) Data Mining
(2) Predictive Analysis

29/8/2017

* Data Visualization : Refers to tools and technologies that support visualisation and interpretation of data and information at several points along data processing chain.

Technologies used for data visualisation are

- any digital image
- geographical information systems
- data analytical Tools
- Virtual reality

* Spreadsheets for Visualisation :

Scanned by CamScanner

GIS is a system designed to capture, store, manipulate, analyze, manage and present spatial or geographic data

Features of spreadsheets

- Charts
- Sparkline
- Conditional formatting
- Pivot Table (summarize large datasets using 'n' dimensions)
- Power map (helps visualise geographic data on a map)

* Geographic Information System (GIS)

- Site selection to open up new outlet of business
- Use GIS to answer questions such as "How many" and "What if".
- GIS can help find demographic profile of a store's neighbourhood to determine the kind of products that can be kept in the store.
- Helps assess damage due to natural calamity.
- Helps identify clusters of call disconnects and dispatch technicians accordingly.

* Application of GIS in banks:

- Determine branch and ATM locations
- Analysing customer demographics such as age, education, income etc will help them come up with new products of marketing campaigns.

Scanned by CamScanner

- Helps managers analyse "what if" questions.
- Determining market potential for bank activities.

* Features of GIS

- Creation of maps based on geographic information.
- Toolkits such as clip tool, buffer tool, continuous and discrete raster.

* GIS functions

(1) Data capture : Data for GIS can come from many sources such as scanning for area photographs, paper maps (undergoes manual digitization, i.e conversion of paper maps into numerical digits that can be stored in computer in the form of points or lines), GPS.

(2) Data compilation
Involves cleaning up of errors and noisy data followed by assigning data to their respective attributes

Scanned by CamScanner

(3) Data Storage : Data storage is based on two modes : Vector and Raster.

(4) Data manipulation : Involves co-ordinate change, projections, edge detection, measuring area, perimeter.

(5) Analysis :

ADS. 30/8/2017

Automated Decision Support Systems :
These are rule-based ~~solutions~~ systems that provide solutions to repetitive managerial problems.
ADS is also called as enterprise decision management. ADS systems are closely related to BI and BA in the following ways :
(a) BA models operate on business tools.
(b) Business rules can be used to trigger automatic decisions that can be a part of BI application.

* Business Rules : These are embedded in a rule-driven work-flow engine.

Scanned by CamScanner

There are four types of business rules:

(a) Analysis rules → These are used to calculate performance metrics from business transaction data.

(b) Context Rules → These enable performance metrics to be tied to business goals and forecast.

(c) Exception Rule → Sending an alert to a business user when a metric exceeds a threshold defined in exception rule.

(d) Actions Rule → Manual decision making.

* Benefits / Characteristics of ABS :-

(1) Rapidly builds rule-based applications
(2) Injecting predictive analysis is also possible.
(3) Providing ABS to legacy systems is also possible.

Scanned by CamScanner

* Implementing ADS.

1. Rule Engine ~~It all~~
2. Mathematical and statistical algorithms
3. Industry specific packages
4. Workflow Applications : Once a rule engine makes a decision, the workflow system moves the business process through required steps.

* Competitive Intelligence :-
Monitoring the activities of the competitors in the market. CI can drive business performance by increasing market knowledge and also improve strategic planning.
Technologies used in ~~CI~~ CI are —
→ Optical Character Recognition (OCR): This helps identify printed characters from business magazines or articles. ~~Using~~ so that it can be stored in knowledge management system.
→ Radio Frequency Identification
→ Intelligent Agents are the softwares which helps business monitor competitor's website or even monitor the entire internet for a specific information.

Scanned by CamScanner

→ Internet : helps business know the feedback about their products or even offer competitive prizes for the products.

Another aspect of CI is industrial espionage (IE) who exist in industries look for confidential marketing plans or strategic plans of the competitors. IE is considered unethical and illegal

Sessional Portion till here

2/9/2017

Web Analytics

Application of business analytics is BI. BI is a method for decision making and web analytics is a solution for BI through which data can be collected for business growth.

Web Analytics is about gathering data from web and analyzing it.

Scanned by CamScanner

Web Analytics w.r.t BI is about determining market interest, segmentation of market, target market, analyze market trends and determine the behaviour of website visitor.

* Benefits of Web Analytics for Business Growth.

(1) Measure web traffic : Web Analytics helps measure :
(a) how many viewers are visiting their website, ~~or any other advertisements~~).
(b) source of visitors (emails, social media,
(c) location of visitors or demographic data.
(d) keywords used by visitors to search products and services

(2) Visitor count and number of visits by a visitor : helps know how effective a page is so that the content of the page can be improved.

(3) Track bounce rate : A bounce is when a visitor visits a page on the website and then leaves the same page without taking any action on the page.

Scanned by CamScanner

Bounce rate is total bounces divided by total visits on the website. High bounce rate is an indicator to marketers that visitor website journey should be improved.

(4) Identify exit pages : Exit is not the same thing as bounce. Exit means that visitor visits multiple pages on a website and on a particular page leaves the site.

(5) Optimise marketing campaigns : Marketers can use certain tools to measure the campaigns performance and which campaigns drive best visitors. This helps marketers to align their resources in right campaign or channel. Resources when aligned properly would result high ~~o~~ return on investment (ROI).

(6) Identify target market : helps reveal the ~~correct~~ current market trends and demands from various geographic location.

Scanned by CamScanner

* Clickstream Analysis : It is a way to implement web analytics. Clickstream data includes how and when a user has entered a website, navigation among the pages, time spent on every page, where exactly the user has left the website.
Clickstream Analysis is about gathering clickstream data and analyzing the user's behaviour either in a website or anywhere on the internet.

* Critical points to be considered while performing clickstream analysis.

1. Raw clickstream data must be obtained from multiple servers.
2. Identify the sessions belonging to a particular customer
3. Obtaining individual customer data which is buried in a mass of other data.
4. Obtaining all information w.r.t a particular page as a single page request can generate multiple entries in server logs.

Scanned by CamScanner

Web Crawlers and Web Scrapers

Web Crawlers and Web Scrapers are means of implementing business analytics.

→ helps in obtaining competitive information ; such as prizes offered by the competitors' new products or services.

→ to monitor the news and social media and industry forums to get information ~~from~~ about what is being said about you and your competitors.

→ lead generation crawler can be set up to identify customers who are in need of a business product or service.

6/9/2017

Usage of BI

Based on three perspectives :

(1.) Descriptive analysis → which use data aggregation and data mining algorithms to provide insight into past and answer "What has happened?"
Usually statistics that are applied are count, sum, average etc.

Scanned by CamScanner

DA is useful because they allow us to learn from past behaviour and understand how they might influence future outcomes. Eg. reports that provide historical insights regarding company's productions, finance, sales etc.

(2) Predictive Analysis → Use statistical models to understand the future and answer "What would happen?".
Eg. Predict the demand for inputs, predict credit score of a customer etc.

(3) Prescriptive analytics → uses optimization and simulation algorithm that answer the question "What should we do?"
These analytics are about advising. It recommends one and more possible courses of action and quantifies the effect of future decisions before the decision are actually made.

Scanned by CamScanner

Date 12/9/2017

Data, Text and Web Mining

Data mining is used to build BI solution by gathering data, processing it, analyzing, organising it in order to get a better understanding of their customer and their own operations.

Need for data mining in health insurance company :-

1. Identify cluster of people suffering from same type of disease so that an insurance plan can be targeted at them.
2. Identify what age group needs what kind of plan.
3. Predictive modeling might be able to predict which patient is likely to fall sick in future.

Difficulties in data mining project.

1. Costly and Complex 2. depends on efficiency of organization's data

Scanned by CamScanner

(3.) Investment ~~of~~ on data warehouse for its development and maintenance.
(4.) Integration of data warehouses when servers are situated in different geographical areas.
(5.) Format of data mining differ in different warehouses.
(6.) Consistency in data gathering effort.

* Data Mining is a process that uses statistical, mathematical formulae, AI, ML and NN to ~~abo~~ extract and identify useful information and knowledge and pattern from ~~a~~ large databases.

A pattern may be a rule, correlation, trend or predictions

* Reasons for sudden raise in data mining

(1.) General recognition of untapped values in large databases.
(2.) Reduction in cost of data storage processing.
(3.) Intense competition for a customer's

Scanned by CamScanner

attention in an increasingly saturated marketplace.

* Characteristics of Data Mining

(1.) Data mining techniques works on data collected in the past.

(2.) A data miner need not have knowledge about the algorithm that is applied.

(3.) Because of the large amount of data, it is better to go for parallel programming.

(4.) Data mining techniques works on unstructured data.

(5.) Data mining environment is usually a client-server architecture.

13/9/2017

* Working of Data Mining follows three patterns:

(i) Simple models includes SQL queries

Scanned by CamScanner

OLAP operations

(ii) Intermediate model includes classification, clustering, regression.

(iii) Complex models from neural networks

Broad categories of Data mining

1. Classification
2. Clustering
3. Association Rule Mining
4. Sequence Pattern Mining
5. Regression
6. Time Series Analysis

1. Classification : It is a supervised techniques. Model helps predict the classes of unclassified instances. The model is built using training data set and accuracy of the model is obtained by using the test data set.

Tools used for classification are neural networks, decision trees

(i) Neural Networks → Algorithms are

Scanned by CamScanner

Information Gain and Entropy are ~~too~~ inversely proportional

effective when number of variables are more in number and relationship among them is complex. Each node of the network can be computed using parallel processing. Time needed to train the model increases as data volume increases

(ii) Decision Trees → significantly faster that neural networks

(2) Clustering : The goal is to create groups so that members within each group have maximum similarity and members across groups have minimum similarity. Once clusters are identified, they can be used to classify new data. It is an unsupervised technique.

(3) Association Rule Mining : Finding frequent pattern correlations from the data set.

(4) Sequence Pattern Mining : It is the identifications of associations over time. Helps in identifying behaviour of a customer over time.
Also keeps track of elapsed time

Scanned by CamScanner

between associated events.

(5.) Regression : It is a statistical technique used to map data to a prediction value. Two types – linear and non-linear.

(6.) Time Series Analysis : Models created by working with time-based data to derive insights.

Applications of Data Mining Technique in Business :-

→ To predict customer churn : (classification)
~~We~~ Information such as billing, website visits, customer interaction etc are gathered to give each customer a probability score; Targets offers and incentives to customers who they think to be at a higher risk of churning.
~~(clustering)~~

→ Retailers segment customers into recency, frequency and monetary groups and target marketing and promotions to different groups.
(clustering)

Scanned by CamScanner

→ People who bought / viewed a product, also bought / viewed other related products. (Association)

→ Analyzing the basket in a single basket / purchase. (Association)

→ Cross-marketing (Association)
Eg. Vehicle dealers or manufacturers have cross-marketing campaigns with oil and gas companies.

→ Catalogue-design : selection of items in a business catalogue are often designed to complement each other so that purchase of one item will lead to purchase of another item. (Association)

→ Web server log can be used to discover user access trends and predict future-visit patterns; helps in pre-fetching documents from servers; recommend web-pages or placing advertisements in web pages (Sequence Pattern)

→ Predict sales next year (Regression + Time Series)

→ Stock Market Analysis (Regression + Time Series)

Scanned by CamScanner

Other Applications include (w.r.t marketing) :

→ Predicting which customers will respond to internet banners.

→ Predicting whether a customer will buy a product.

→ Segmenting customer based on demographics

Applications w.r.t banking :

→ Forecasting level of bad loans.

→ Forecasting which kind of customer will respond to a new service.

→ Credit card spending by new customers

Applications w.r.t retailing and sales:

→ Predicting inventory level based on sales

→ Predicting the distribution schedules among outlets.

Applications w.r.t marketing & production:

→ Predicting when to expect machinery failure

Application w.r.t brokerage:

→ Predicting when to buy a particular share.

Scanned by CamScanner

Application w.r.t computer H/W & S/W.
→ Prediction of disk failure.

Application w.r.t broadcasting
→ Predicting what programs are best shown during the prime time and how to maximise returns by knowing where to insert advertisements.

Two types of data mining :-

(1) Hypothesis-driven data mining begins with a proposition by the user who then seeks to validate the truthfulness of the proposition. Eg. Are DVD sales related to sales of TV sets?

(2) Discovery-driven data mining finds patterns, associations, relationships among the data.

* Case-Based Reasoning (one of DM techniques)

This is done for developing solutions to unsolved problems based on

Scanned by CamScanner

case-base : database of sol^n of all the cases.

existing solutions of similar nature (pretty much what ADS makes use of). Eg. Diagnostic assistants to help users diagnose problems in their computers. When user gives a description of their problem, the closest cases in the case-base are retrieved.

16/9/2017

* Other Application w.r.t Defence (applications continued)

→ Data mining helps in predict ~~or~~ potential attacks to a certain area.

* Statistical Methods (DM technique)

(1.) Linear Regression : It refers to a model that can show relationship between two variables (dependent and independent variables) and how one can impact the other. Eg. if a company's sales have increased steadily every month for the past few years by conducting analysis on the sales data with monthly sales, the company could forecast, sales in future months (linear analysis on sales data)

Scanned by CamScanner

(2) Non-linear Regression : It refers to modelling when the variables in the data set do not have a linear relationship. Eg. fund invested on marketing a particular product may not always yield a substantial amount of ROI.

(3) Point Estimation : The process of finding an approximate value of some population. Eg. deviation of sales in a year.

(4) Neural Computing (DM techniques)

Applications of neural computing are common in financial services such as daily currency exchange rate, stock prediction, pricing of a product, hedging (investment risk).

* Intelligent Agents (DM technique)

These are S/W programs that uses a built-in or pre-defined knowledge to carry out specific tasks for a user. Eg. chatbots, delete junk emails, schedule appointments,

Scanned by CamScanner

Sort emails, customize news reports, form-filler bots.
These are used to observe and report on an equipment (monitoring bots).
Intelligent bots that keep track of company's inventory level, observe competitor's price.

19/9/2017

* Genetic Algorithm (GA): It is a search-based optimisation technique based on the principle of genetics and natural selection. It is used to find optimal or near optimal solutions. In GA, we have a pool of possible solutions (population) for a given problem. Initially random solutions are identified, these solutions then undergo recombination and mutation producing new solutions and the process is repeated until a stopping criteria is reached. Each candidate solution is assigned a fitness value. GA is used in clustering and association rule mining.

Advantages:
(i) Provides a list of good solutions

Scanned by CamScanner

and not a single solution.
(ii) Useful when search space is large and large number of parameters are involved.

Disadvantages :
(i) If not implemented properly, GA may not cover all optimal solutions
(ii) There is no guarantee on quality of all the solutions

* Roughset Algorithm (used for classification : It is a non-statistical approach in data science for classification ; based on the concept of approximation of lower and upper spaces of a set. The lower subset is characterised by objects that will definitely form part of interest whereas, the upper set is characterised by objects that will possibly form part of an interest.
(Interest eg. fraud detection)
Every subset here is called a rough set.

Scanned by CamScanner

Entropy: level of uncertainty of a particular node.

Factors of Classification Algorithm (CA):

1. Accuracy: ability of a CA to correctly predict the class label of an unknown example / instance.

2. Speed: refers to the speed and computational cost involved in generating a model and using it for prediction or classification.

3. Robustness: ability of an algorithm to predict correctly even when the dataset has missing values or noisy data.

4. Scalability: ability of a CA to generate the model when the size of the dataset increases.

5. Interpretability: refers to the level of understanding the insights provided by the model.

* To Do: Problems on Clustering and Decision Tree

Scanned by CamScanner

#

	ID3	C4.5	C5.0	CART
Type of I/p	Categorical	Continuous & Categorical	Continuous, Categorical, Timestamp	Continuous & Nominal
Speed	Low	Faster than ID3	Highest	Average
Pruning	No pruning	Pre-pruning	Pre-pruning	Post-pruning
Missing Values	Cannot handle missing values.	Can handle missing values	Can handle missing values	Can handle missing values.
Branching Procedure	Information Gain & Entropy	Split information and gain ratio	Split information and gain ratio.	Gini Index

20.9.2017

Cluster Analysis for Data Mining

Cluster analysis is a statistical tool for identifying groups in a dataset. Cluster analysis results may be used to:
(i) identify a classification scheme
(ii) suggest statistical models to describe populations
Eg. discriminant analysis ~~a~~ is a statistical ~~analysis~~ models used to understand fitness of a cluster.

Scanned by CamScanner

(iii) indicate rules for assigning cases to classes

(iv) ~~rear~~ rare event detection
Eg. identifying outliers

(v) provides measures such as cluster density, similarity between the data points

Cluster Analysis Methods :-

(1) Statistical method : (i) Model-based
(ii) Partition
(iii) Density-based

All these are non-hierarchical.

Hierarchical — (i) Divisive (ii) Agglomerative

(2) Neural Network

(3) Fuzzy Logic : In fuzzy clustering, each data point can belong to multiple clusters.

(4) Genetic Algorithm

Critical Issues in Clustering

(1) Size of the cluster (number of clusters)

(2) Should all clusters have approximately

Scanned by CamScanner

same number of items?

(3) Dimensionality of the data set Vs. similarity measures

Data Mining Project Process

Data mining projects during its development follow a project management process.

(1) Cross-Industry standard process

Business Understanding ⟷ Data Understanding

↓

Data Preparation ⟷ Modelling

↓

Evaluation

↓

Deployment

The model includes sequence of steps that cannot be sequential always and may involve iterative work especially in understanding the data requirement from the business.

(2) ~~(ii)~~ Six-Sigma Model

Define → Measure → Analyze → Improve

↓

Control.

Scanned by CamScanner

Six-Sigma is a methodology used for eliminating defects and quality control problems used in manufacturing, service-delivery etc.
The basic idea of the steps is to have an idea of what to expect. This is accompanied with other steps like applying measurements at each step & analyze the measurements w.r.t the data, obtain the feedback in order to improve on the mechanisms.

22/9/2017

Six Sigma is a data-driven, systematic, problem solving approach which is divided into two process - DEMAIC (Define, Measure, Analyze, Improve, Control) and DMADV (Define, Measure, Analyze, Design, Verify)

DEMAIC process is an improvement system for existing process falling below a specification level and looking forward for incremental improvement.

DMADV is a process followed during the development of a new product. ~~Eg. for~~

Scanned by CamScanner

Example for DMAIC : Improving the heat efficiency of existing steel manufacturing process in a steel company.

* Key steps of DMAIC process :

(i) Define
→ defines the requirement of customers
→ defines goal and scope of project
→ map the defined requirements and goals to business flow
→ identify factors such as critical to quality, critical to process etc.

(ii) Measure :
→ measuring CTx (x can be quality, time, process)
→ assess performance specification such as :
ⓐ Project Effort (also called as Team effort or Work Effort) → defined as total project team time that is spent on project-related activities in the life cycle of the project.
ⓑ Efficiency / Yield / Productivity of the project

Scanned by CamScanner

(c) Project Duration → defined as project cycle time and time to market.
(d) Schedule Predictability → measure of how much the original project duration estimate differs from actual project duration.
(e) Requirement completion ratio → measures the extent to which planned functional requirements were satisfied in the final product implementation.
(f) Post-release Defects

(iii) Analyze:
→ analyzes the source of variability (Consider a fuel-fed system in which there was too much variation in the concentration of pet coke. Two cases may arise - ~~resulting in~~ too much pet coke resulting in violation of environmental parameters and too little pet coke may result in increased cost of operation)

(iv) Improve:
→ process is improved by eliminating wasteful causes (design, idea, features)
→ in this phase, various models are generated and evaluated for the data gathered

Scanned by CamScanner

Just for understanding (Not asked in exam)

in measure phase.
*(contd. below)

Data Mining Techniques used in Six Sigma phases

1. Measure : sampling, data partition etc
2. Analyze : exploratory analysis and modelling (regression etc)
3. Improve : interactive grouping node in an algorithm in SAS miner (it is a tool of Six Sigma which helps in optimization)

23/9/2017

*(V) Control :
(contd.) → audits or inspections are carried out.
→ observations are documented
→ procedures carried out and steps to be followed for the usage are documented

Data model just like cross-industry standard, six sigma

3. SEMMA (Sample, Explore, Modify, Model, Assess)

(1) Sample :
→ by extracting a portion of large data

set which is big enough to contain the significant information, yet small enough to manipulate quickly

(ii) Explore :

→ explore the data to search for some unanticipated trends in the sampled data

(iii) Modify :

→ data can be modified to add new variables, reduce number of variables, feature selection and transform the variables.

(iv) Model :

→ application of data mining algorithms

(v) Assess :

→ assess your model to estimate how well it performs.

Scanned by CamScanner

TEXT MINING

Definition: Text mining is an application of data mining to unstructured or less structured text files.
Source of data may be electronic text, server logs, corporate documents, application firms, research papers, emails etc.

* Applications of text mining:

→ helps to provide better services to their customers through their feedback and helps increase customer base.

→ helps maintain better relationship with partner companies and forge new relationships.

→ data about competitors can help them stay ahead of its competitors.

→ helps in HR management to monitor employee satisfaction level, storing and reading CVs for the selection of new personnel.

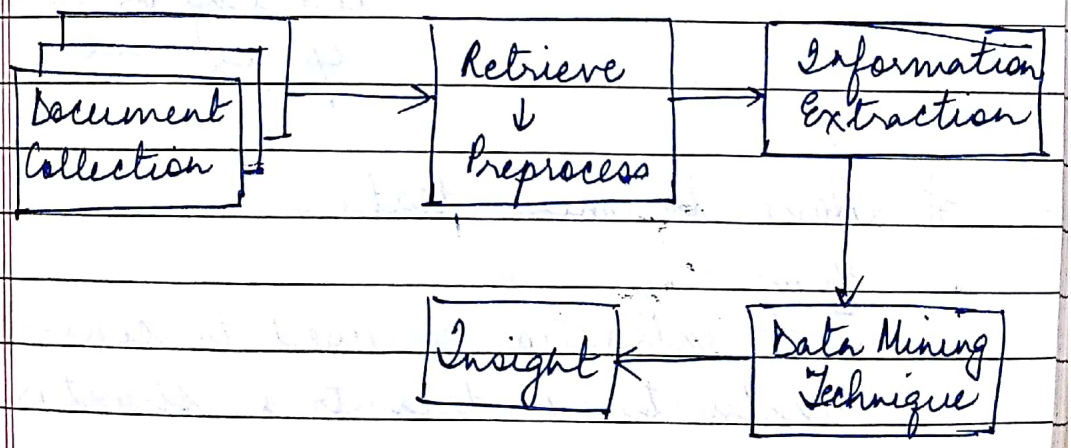
Scanned by CamScanner

→ identifying the diseases that are hit in a particular region (especially rural areas)

→ find complaints from call centres call - by converting audio to a transcript.

→ can be used to rate call centre performance.

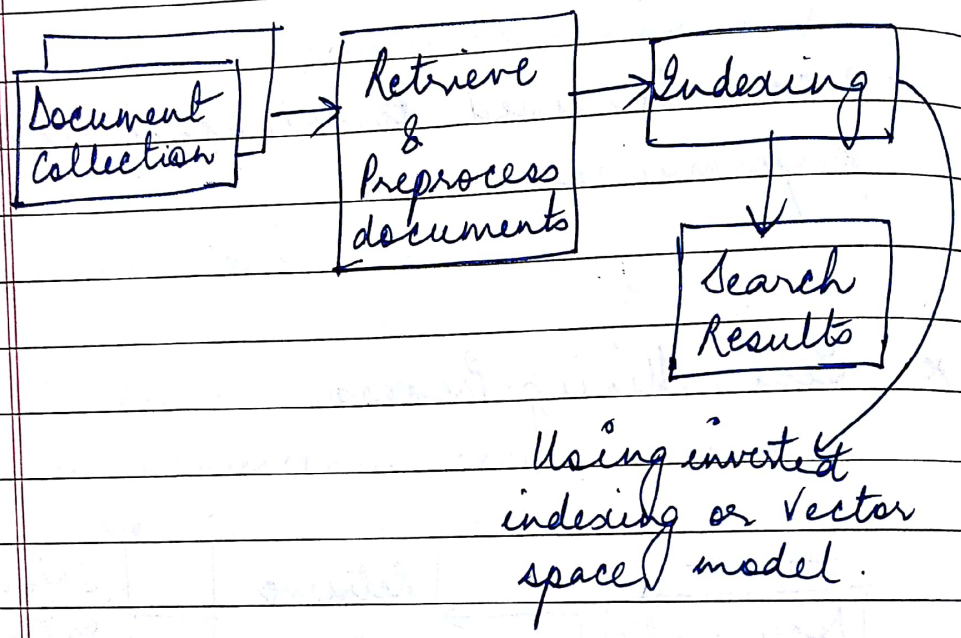
* Text Mining Process



During document collection, the file may be of any format.
In preprocess step, steps like stemming, tokenisation, Stop - word removal are applied.

Scanned by CamScanner

W.r.t information extraction, converting retrieved text into a format of a data set with attributes and class labels (if necessary) by applying techniques such as natural language processing (NLP).



How to mine text.

~~Text~~ Term extraction is used to convert unstructured data to a structured format.

Text is reduced to a list of terms and weights.

To do this, text mining involves the following steps:

Scanned by CamScanner

(i) Eliminate stop-words such as — and, the etc.
(ii) Replace words with their stems or roots. Eg. stemming becomes stem.
(iii) Consider synonyms and phrases.
(iv) Calculate the weights of the terms such as term frequency, inverse document frequency.

26/9/2017

Mistakes in data mining project

→ Selecting wrong problem for data mining.
→ Leaving insufficient time for data preparation
→ Being ignorant about keeping track of mining procedures and result
→ Ignoring suspicious findings such as outlier examples in the data set.
→ Running mining algorithms blindly.

Scanned by CamScanner

WEB MINING

It is the discovery and analysis of interesting and useful information from web by making use of web based tools and text mining.

Three categories under web mining
→ Web Content Mining
→ Web Structure Mining
→ Web Usage Mining.

1. Web Content Mining — webpage content mining and search result mining.

2. Web Usage Mining — general access pattern tracking and customised usage tracking.

I. WEB CONTENT MINING refers to extraction of useful information from webpages. The web content may be text, video - audio images etc.

Web crawlers and web scrapers are used to read the content of web site.

Scanned by CamScanner

Applications of Web Content Mining

→ To know the performance of competitors.

II WEB STRUCTURE MINING refers to developing useful information from the links included in the web document. It can be used to determine the popularity of a document using page rank algorithm.

Eg.

For the business to understand their usefulness & or how informative ~~the~~ the webpages are.

Links within a document can indicate the depth of coverage of a topic.

Hubs are the pages that point to many authorities and authorities are the pages which provide information about a topic and are pointed from other links.

A good understanding of the structure of a website is helpful in determining

Scanned by CamScanner

which pages should be added to a collection of websites.

Hyperlink analysis is a technique that uses a graph structure in order to determine the inter-relationships between links and importance of a webpage.

Applications of hyperlink analysis: user behaviour of shopping, user behaviour of searching any information from web.

* ~~Web Usage~~

III. WEB USAGE MINING is about discovering patterns in user's behaviour. Main source of data is web server log.

Applications → helps in identifying where to place the advertisements. Eg. customers who search for hotels are likely to search for flights. Hence advertisements for hotels and flights can be displayed together.

→ to understand the peak time of customer's shopping so that a better

support can be given.

27.9.2017

* Search Result Mining (comes under Web Content Mining) :
It is optimisation of search result based on query done by user. Search result of a user query would be too many to go through, hence terms from the query can be extracted and based on the similarity between user's earlier queries on a related topic and documents viewed earlier, clustering may be applied such that the most appropriate cluster of links or documents may be shown to the user.

* General Access Pattern Tracking (comes under Web Usage Mining) :
It analyzes the web logs to understand access patterns and trends w.r.t several users. It helps in understanding the resource providers.
It can be used to restructure sites.
It targets specific users for specific selling advertisements.

Scanned by CamScanner

* Customized Usage Tracking : analyzes individual customers' trends.
It helps in customisation of websites to users.
(same like General Access Package Tracking)

TECHNOLOGIES ENABLING BUSINESS INTELLIGENCE

Organisational memory provides business archives which is a unique asset which every company has.
It captures knowledge and experience of day-to-day's working of a company and is an integral part of business

Organisations would be at a risk of losing their memory if employees switch job without passing on the knowledge that they have learnt.

Organizational memory (OM) is related to aspects such as: corporate memory,

Scanned by CamScanner

knowledge repositories and institution memory.
Eg. for corporate memory that are stored in knowledge repository - company culture (products, employees, work ethics), product literature, websites, information about growth and recession, commercial activities, analysis of success and failures in R & D departments.

28.9.2017.

Two technologies involved in OM :
1. ERP Systems
2. Data warehouse

1. ERP Systems : It is a software for transactional processing that captures organizational memory related to all the business processes that an organisation is involved in.
Eg. order to cash, captures all the transactions involved in the organisation from the moment the customer places an order to the moment the organisation receives

Scanned by CamScanner

payment for the product.

ERP expectations: (i) Integration of all the functionalities of an organisation (finance, marketing, HR, production etc)

(ii) Integrating the systems which are geographically apart ~~wher~~ by providing one single data repository.

(iii) Transparency of information across departments

(iv) Modular, flexible, easy-to-add functionalities.

* Advantages of ERP:

→ Transparency ~~of~~ in the system.

→ One integrated solution for the entire organisation.

→ A single database without duplication of records.

→ Realtime access of data for operational staff

→ High ~~quality~~ security of business data

→ Improves customer service especially in case of order fulfilment process.

→ Replaces custom code or spaghetti code used to artificially interconnect

extant : still in existence, surviving

the functionalities through APIs.

* Configuration control in ERP system

Since ERP systems are developed based on the generic needs of an organisation, it has numerous functionalities. Enterprises are required to pick up the relevant components to meet their specific business requirements.
Configuration of ERP systems deals with handling of numerous usage controls which can be switched off or switched on so as to balance its functionalities extent needs.

Thousands of configuration tables are present in the middle ware of the ERP systems which define how the system should operate and how the signals or messages should appear, how a screen should look like etc.

* Success factors of ERP implementation

(1) Top management commitment includes support from managers; business

Scanned by CamScanner

plan and vision has to be set during the startup of a project.

(2) Team members' skill set :-

(i) Experience with programming environment.

(ii) Familiarity with modern programming practices.

(3) Strong Project Management includes business process redesign in case an organisation had legacy system

(4) Effective communication among various users ~~to~~ of ERP.

* Issues with ERP System

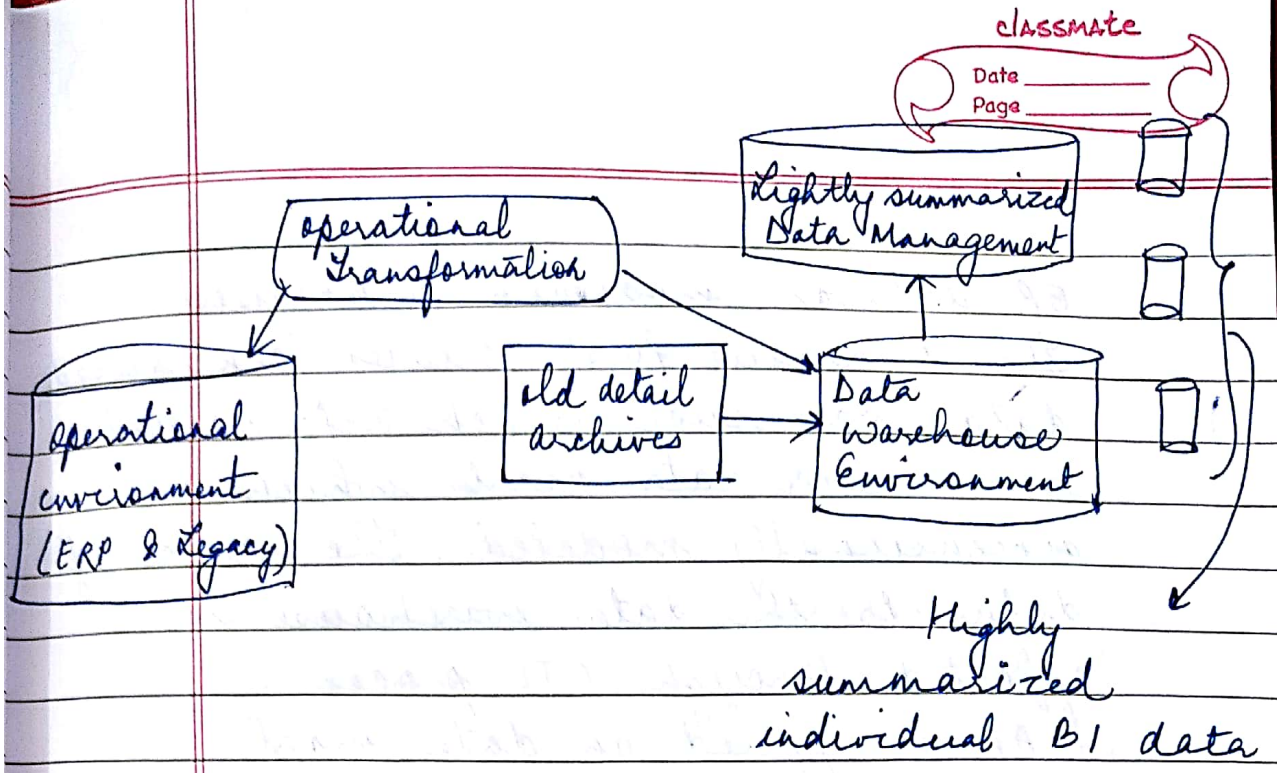
(1) Large financial investment

(2) Change in business process to align to ERP system

(3) Technical changes to integrate ERP with legacy system

II. Data Warehouse

Scanned by CamScanner



Enterprise data model has four levels:

(i) Operational level which consists of ERP systems and legacy system.
(ii) Data warehouse level
(iii) Department level / Data mart level
(iv) Individual data level

Modelling of Data Warehouses

Organizations choose appropriate modelling techniques for data warehouse based on their analytical needs. Two techniques:

(i) Entity Relational Modelling: wherein data warehouse is an integrated database modeled using traditional

ER database modeling techniques.
The technique first creates centralised data warehouse which acts as data source for data marts which are dimensionally modeled. The source of data to the data warehouse is affected through ETL process.
OLAP is applied on data marts.
ER modeling could take longer time but could provide a more powerful method by providing a data repository that could be used as a basis for extracts into the other data stores and acts as a backup.

(2) Dimensional Modeling : Data warehouse is a collection of dimensionally modeled data marts. The dimension tables which are common across the departments are modeled first and then fact tables are added. As a result, data warehouse is a collection of interlinked dimensionally modeled data marts

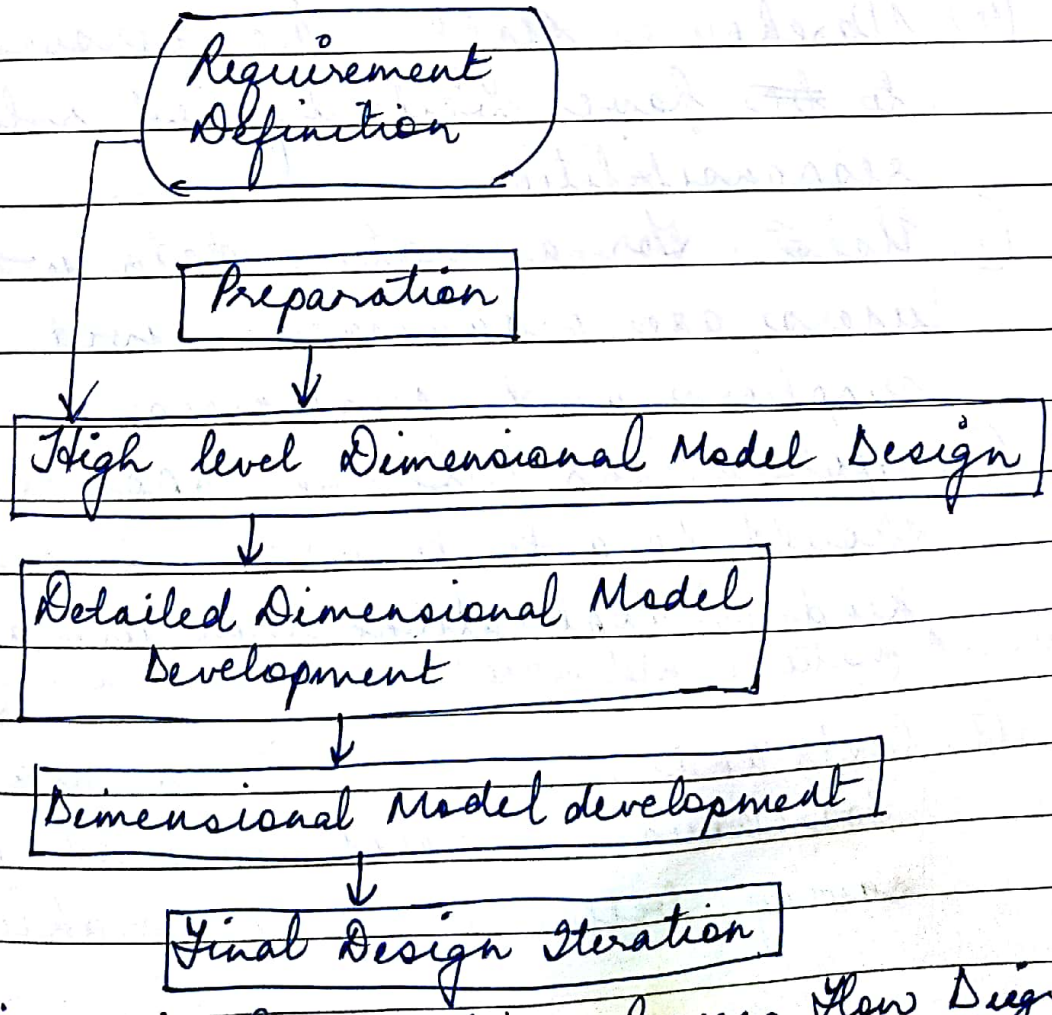
* Challenges in deploying a Data Warehouse.

Scanned by CamScanner

① Abundant technical options at each stage such as extract, storing, accessing analyzing.

② Changes in technologies and vendors. The tools and technologies available for a company are many and hence companies have to assess their requirements and figure out which data warehouse tool can effectively help them grow.

③ Integration requirements is required in order to integrate tools across vendors and across versions.



Dimensional Modeling Process Flow Diagram

Scanned by CamScanner

* Characteristics of a Mature Data Warehouse.

(1) Data reflects the enterprise (different functional areas of an organisation, well-integrated and evolved over multiple time periods)

(2) Architecture : For a mature data warehouse, there should be a clear structure with dependent data marts.

(3) Stability of the Production Environment : The procedures involved are routined, maintained and well-documented.

(4) Warehouse Staff : The personnel has to ~~be~~ have well-defined roles and responsibilities.

(5) Users : For a mature data warehouse, users are organisation-wide including suppliers and customers.

(6) Applications : Several applications should be able to access it. Refers to the kinds of applications that use a data warehouse. A mature data warehouse should include reports, queries, DSS etc.

(7) Cost and Benefit : For a mature data warehouse, benefits should be time-saving, better decision-making

Scanned by CamScanner

Design of Enterprise Architecture

It is a blue-print for creation of unified IT environment (standardized hardware and software systems) and its strategy across the organisation.
The goals are:

(i) Promote alignment → Keeping employees aligned to the main purpose inspite of its size and complexity.

(ii) Standardization → set of standards for employees and products.

(iii). Reusing of existing IT assets → common methods for project management and software development.

Examples: (i) Put a process in place to achieve sales. (ii) Create a mandatory procedure for staff for opening and closing of business outlets. (iii) Set a standard or policy for staff clothing at the customer service offices.

Firms define their underlying enterprise architecture by making two important choices about their business operations unit.

(a) how standardized their business processes

Scanned by CamScanner

should be across operational unit.

(6) how integrated their business process should be across those units.

* Standardization refers to defining exactly how a process will be executed regardless of who is performing the process or where it is completed.

For any process, when there is more than one person completing it, there is likely to be more than one method used. This has the potential to cause issues with the consistency of output. One way to overcome this issue is to work with process owners to standardize the process ensuring that each time a process is carried out, it is completed in the same way.

* Benefits of Standardization

(1) Consistent output from the process
(2) The method can be roled out to the new employees.
(3) Internal audits and outside regulatory bodies are more likely to be satisfied with the standard process.

Scanned by CamScanner

* Integration

This links the efforts of processes of organizational units through shared data. To enable end to end transaction processing so that a company can project a single phase to customers.

Integration relates to how different areas of a company coordinates their operations.

A highly integrated company has strong connection between departments with each section working under a cohesive set of rules and strategies.

Based on two choices of business process of standardization and integration will operate in one of the four possible operating models. (These operating models decide on how to implement your business processes and its infrastructure and is the first step towards designing of enterprise architecture.)

* Diversification Model (Low standardization and low integration)

Scanned by CamScanner

→ It is used by organisations that follow a decentralised organisational design.
→ Organizations may have same shared data but the business units operates in a unique manner.

* Coordination Model

Low standardization, high integration is used by firms that have high level of integration through sharing customers, products, suppliers and partners but operationally every business unit is unique. The autonomous business units have a high degree of control over business process designs, to adapt to its specific operations.

7/10/2017

→ Enterprise Architecture :-
→ Steps towards construction of an Enterprise Architecture
→ Standardization (↑)
→ Integration (↓)

Scanned by CamScanner

* Replication Model (High Standardization, Low Integration)
→ Used by firms that provides these units with the autonomy to run independently (no data sharing) following highly standardized business process.
→ For firms in this model, their units do not depend on each other's transactions or data while business process design is centrally managed.

* Unification Model (High standard, High Integration)
→ Used by firms that are highly integrated through supply chains, customers etc and have a centralized business process.

Examples:

(1) Diversification : GMAC is an independent business which has companies such as GMAC mortage, GMAC real state, GMAC
Each company has different customers and suppliers but the IT services are shared

Scanned by CamScanner

(2) Coordination : A firm USAA originally provided only insurance services, now they also offer banking, investment services and products. There are common customers across their business units.

(3) Replication : Each of the restaurants of McDonalds are independently managed and accountable for results but all outlets operates in a similar manner.

(4) Unification : Dow chemical is a company in which 60% of the different ~~co~~ departments, core chemical processes are standardized while supply chains to several departments are same.

Once a firm has identified an appropriate model, it is ready to define the enterprise architecture which has four levels.

Four levels of architecture which resides below enterprise architecture:

Scanned by CamScanner

→ The business process architecture which refers to activities that comprise the major business process identified by a process owner.
Eg. Invoice, generating bill, customer service.

→ The data or information architecture which includes shared data definitions. The customer data that comes to a billing section is passed on to a shipping section.

→ Application architecture which refers to the individual application systems architecture. Eg. multitier architecture, client tier architecture.

→ Technology architecture : Architecture of technology that is used to provide a functionality for an application built over it.

→ Enterprise architecture : depicted in one page picture called core diagram which depicts a high-level view of process data and technologies that comprises the organizations digitized platforms for execution of their business process. Companies that select a specific operating model will have similar core diagrams.

Scanned by CamScanner

Date 10/10/2017
Page

Knowledge Repository

→ Includes technologies that support the capture of unstructured information and knowledge (document management system) and archiving of audio and video files (digital content management system).

Related terms are:
→ Enterprise Content Management System
→ Web Content Management System.

For knowledge repositories to be effective, they must attract a critical volume of knowledge seekers and knowledge owners. Knowledge seekers should be able to explore all the documents pertaining to their context. Knowledge owners should share their knowledge with a trusted group and they should decide when to share and conditions for sharing.

* Characteristics of Document Management System:

Scanned by CamScanner

→ should reside in a primary location that affords multiple access points.
→ should be centralized or distributed
→ stores unstructured data
→ should be platform-independent and should provide web-interface.
→ should follow indexing, classification or hierarchial structure.

Knowledge Repositories (KR) can be classified according to organisations per person.

Types of KR:

① Incident Report Databases : used to store information about malfunctioning of an equipment or any other incident. Eg. bug reports, downtime of an equipment due to voltage sag. These reports include information such as what is the incident, incident investigation, screenshots if any, recommendations if any, etc.

② Alert Systems : spread information about a negative or a positive incident that is going to occur.

Scanned by CamScanner

Date
Page

Eg. product recall.

(3) Best Practices Database : describes successful efforts w.r.t business practices such as technical documentation, industry standards etc.

(4) Lessons Learnt Systems : captures and provides lessons that can benefit employees who encounter situations that closely resemble a previous experience in a similar situation.

(5) Expertise locator system : helps locate an expert who can be a source of information.

Remember Table 3.1 from Textbook

The difference among the knowledge repositories are based upon :

(i) Content origin :
Is information from lesson learnt or technical documentation or industry standard ?

(ii) Application :

Scanned by CamScanner

Does the information explain the complete process?

(iii) Results:
Does the information describe a success or a failure?

(iv) Orientation:
Does the information support an organization or an industry?

End of Second Sessional :)

Scanned by CamScanner

TECHNOLOGIES ENABLING INFORMATION INTEGRATION

This chapter is about integration of structured information from data warehouses and ERP system with unstructured information from other sources such as web to produce new insights.

Integration of data sources in a business intelligence application

Some organizations may have large data warehouse and some others may have spreadsheets. Therefore BI applications will rely on wide range of software tools to integrate and create insights as there is difference in data sources.

The decision of whether to use external data or not depends upon:
(i) Business need
(ii) Inhouse data availability

Scanned by CamScanner

(iii) Lack of inhouse expertise to generate data

(iv) BI application sophistication to improve the predictive ability.

(v) BI budget

Designing an organization's architecture for knowledge creation.

① Crisp understanding of business objective
Example: to understand the customer response rate to a marketing campaign, the foremost step is to target those customers from campaign database who are likely to buy.

② Understand the data

* Data collection
→ Define data sources
→ Find out data owner
→ Cost of using the data source
→ Size
→ Structure

* Data description : describes the content of each of the data source such as primary key, foreign key, number of columns, size of each table etc.

Scanned by CamScanner

* Data quality and verification : Data quality dimensions are —
(i) Validity of data (values fall within acceptable range)
(ii) Consistency — same values across multiple locations
(iii) Integrity —
(iv) Accuracy — data has to be correct
(v) Relevance — how appropriate the data is to business objective.

③ Environmental Scanning

It is scanning for information about events in a company's outside environment, the ~~the~~ knowledge of which would assist top management in its task of deciding the company's future course of action, both at tactical and strategic level.
One popular method of environment scanning is 'SWOT analysis in which strength and ~~opportun~~ weakness are factors within the company, opportunities and threats come from sources outside the organisation.
Eg. most scans include a thorough

Scanned by CamScanner

look at competition, economics, technology, legal issues, social or demographic factors.
Example on competition – a shop which was the only one in an area for 20 years but now a competitor crept into the market this year. The existing shop has to work upon customer feedback, services, pricing, competitor's services etc.
Example on economics – are people losing their jobs and cutting back on spendings.
Example on Technology – Business must reveal how much internet impacts sales and how to make web work for them. If a business does not want to sell on internet, they can use it to advertise or keep customers uptodate about their special offers and products.
Example on legal issues – what would happen if a restaurant is filthy or what would happen if a construction company has lots of injuries.

Environmental scanning is not always performed to achieve profit but to

Scanned by CamScanner

keep a business in the long run.

12/10/2017

Adaptive Organization

→ Organizations might have to adapt to change in economy, market-forces etc.

→ Four modes of ES are based on two concepts : organizational analyzability, organizational perceiveness

Adaptive organizations are the ones able to keep up with the rapid changes in its environment. They are able to find out external factors affecting the organizations and how they may affect the decision-making during the course of action. Eg. adaptive organization accept change in customer needs, change in technology, change in ~~customer needs~~ economy etc.

There are four modes of environmental scanning; these four modes are based on two dimensions

(1) Environment analyzability → It is about

Scanned by CamScanner

management belief about analyzability of external environment such as government policies, market forces, stakeholder attitude.

Environment is said to be analyzable when process are hard (fixed or routine), measurable and determinant. Example; banks have to follow government policies while sanctioning house loan.

Environment is said to be unanalyzable when ~~process are~~ ~~there~~ there is a need for reasonable interpretation to suggest upcoming steps.
Eg: companies selling air-conditioners, what offers should they offer, what climatic conditions should they consider for those offer.
→ Attitude of stock holder, there is no clearcut way of predicting that Eg. Fluctuations in product pricing dependending upon a climatic condition. Stockholders attitude towards a company.

(2) Organizational Intrusiveness : Extent to

Scanned by CamScanner

which organization allocates resources for information search or extent to which organizations which actively intrude into environment.

Active Intrusiveness : Organizations actively search for information on a continuous basis. Eg. build a plan to gather information.

Passive Intrusiveness : Organizations generally accept information given by any random source.

Four modes of environmental scanning:

1). Undirected Viewing (Environment is unanalyzable and intrusiveness is also passive. EA is unanalyzable and OI is passive).
Meaning satisfied with limited information and does not seek comprehensive data from external source, Eg. flipping through TV channels, magazines and surfing internet for an information.

Scanned by CamScanner

2). Conditional Viewing (EA is analyzable; OI is passive): Makes use of standard procedure to gather information such as browsing of particular sections of newspaper, periodicals or websites that report regularly on topics of interest

3). Searching: EA is analyzable, OI is active. Refers to deliberate effort following a pre-established plan or methodology to secure a specific topic. Eg: market forecast which predicts trend analysis.

4). Exacting (EA is unanalyzable, OI is active): Organization constructs their own experiments or unstructured effort to obtain a specific information

An organisation's environment could be ~~perceived~~ perceived as uncertain due to number of reasons.

1). Market
2). Technology
3). Regulatory
4). Economy
5). Social factors
6). Political Factors

Scanned by CamScanner

Date 21/10/2017

Text mining and web mining are the most important techniques used to identify and extract external business information for environment scanning.

* Text Mining : Techniques can be classified into four main layers.

(1) Natural Language Processing (NLP) or Linguistic Analysis : used to identify key concept descriptors which are embedded in textual documents. Descriptors of documents are any word that answers 'WH' questions like what, when, where etc. Examples of linguistic analysis are : stemming, stop-word removal, proximity.
Unit of analysis is a word.

(2) Statistical and co-occurrence analysis → Co-occurrence of two terms is frequent occurrence of two terms from a text corpus along side each other in a certain order.

Scanned by CamScanner

(3) Statistical and Neural Networks clustering and categorization:
These are the techniques used to find similarity between documents

(4) Visualization → used to interpret the initial dataset and the results from from the techniques applied.

Refer to figure 4.2 from textbook which describes the text mining process

* Web Mining:

(1) Web Structure Mining:
(a) Intrapage Structure Mining → study of arrangiment of various HTML and XML tags within a webpage.
(b) Interpage Structure Mining → Refers to study of hyperlinks connecting one page to another

(2) Web Usage Mining:
Three main tasks performed →
(a) Preprocessing: includes (i) collecting data from multiple websites.
(ii) cleansing the data

Scanned by CamScanner

(iii)
(a) Dataset preparation

(b) ~~Web~~ Pattern Analysis → involves noticing unusual patterns, identifying the relationship between variables

(c) Pattern Discovery → using data mining techniques, patterns are found by considering certain variation in the application. Eg. application of market basket analysis on shopping carts.

(3.) Web Content Mining
Web mining can be used to search for images on the internet. Techniques are:
(i) Keyword-based image retrieval techniques which seek to match keywords from the user query to those keywords used to annotate the images.
Limitation is: annotating images is a tedious job.

(ii) Content-based image retrieval: This method allows automatic extraction of features (colors, shapes, objects) from images and aims to find images

Scanned by CamScanner

via specific features of queried image. Limitation is that it may often find irrelevant images with similar features as those of queried image.

(iii) This method combines both visual and semantic feature vectors from the images and textual content of web pages where the images are found.

Customer Relationship Management (CRM)

→ consists of mechanisms and technologies used to manage the interactions between a company and its customers.

Application in CRM can be characterised as:

→ Operational CRM: Includes the support of call centers and sales force automation tracking customer packages.

→ Analytical CRM: Employ data mining techniques to create new insights about their customer in order to better understand and serve their clients.

* CRM process design

Scanned by CamScanner

Involves –

(1) Identification of customer market segments

(2) Integration of result of data mining study (market segmentation) into a predefined campaign management software (CMS).

CMS manages :

→ planning, execution, assessment and refinement of marketing campaign

→ monitors company's communication with its customers about campaigns such as email, direct point of sale etc

(3) Data mining prediction model is used to predict customer score which is a numeric value assigned to each customer database record that specifies the probability that customer in question will behave in a specific manner

Eg. customer churn is a measure of customer attrition to a particular service or product.

RFID (Radio Frequency Identification)

It is a technology to record the presence of an object using radio

Scanned by CamScanner

signals. Consists of two parts.

→ Antenna : To receive and transmit signals.

→ RFID tag : Integrated circuit which modulates and demodulates a radio frequency signal, processes and stores information.

* Uses of RFID in business :

1. Asset tracking
2. Supply chain management
3. Inventory management
4. Retailing

Scanned by CamScanner

Date 27/10/2017
Page

TECHNOLOGIES CREATING INSIGHTS AND DECISIONS

The process of creating new insights from information is known as business analytics

BA is a part of BI. Organizations relying on BA make use of:

1. Extensive use of data.
2. Statistical Analysis
3. Predictive Modelling
4. Decision Making

* Business Analytics Process

Since BA makes use of predictive modelling which is a technique in data mining, it become apparent to develop a standard process model.

Cross industry standard process for data mining (CRISP-DM) is one such process model used for

Scanned by CamScanner

data mining projects.

Steps in CRISP - DM process:

(1) Business understanding involves:
→ determining business objectives
→ business success criteria
→ inventory of resources, requirements
→ any assumptions, constraints, risks, costs and benefit.

(2) Data understanding involves:
→ data collection.
→ data description
→ data quality verification
→ data analysis

(3) Data Preparation
→ selection involves — sampling, rational for inclusion and exclusion of independent variables.
→ construction and transformation of variables: Construction involves computing domain of values for an attribute from the raw data. Transformation involves converting the data from one format to another format to support model construction.

Scanned by CamScanner

→ Data Integration : It involves data consolidation from multiple disjoint databases.

(4.) Model Building and Validation
The task involves trying out several algorithms on the same particular dataset until and the results are compared to see which model is the best accurate result.
Development of model is followed by validation. The most popular validation technique is enfold cross validation.

(5.) Model Evaluation and Interpretation
Evaluation of a model can be done by giving an input whose label is already known or by considering an example from the dataset itself.

Data Mining Model : Describing what happened

Techniques used :
1. Association
2. Clustering

Scanned by CamScanner

Data mining model describing what will happen

1. Neural Networks
2. Rule Induction or Decision-Trees
3. Statistical Methods

* Evaluation of results of data mining algorithm

Once the result of DM study have been obtained, it is important to carefully evaluate when the performance of the model is examined. Eg. In the case of a system used to predict cancer in a patient, a 'yes' when an answer is 'no' or in the vice-versa case, can be very costly in terms of seriousness of ~~a~~ the error.

28/10/2017

Effective Implementation of Business Analytics

Common mistakes that organisations seeking the deployment of data mining technology must avoid.

(1) User expectations are too high. - overly impressed by vendors claims for a small set of data. This mistake can be overcome by working with vendors by using actual company data to produce all functional demos

(2) Putting the right tools in wrong hands - Data mining tools will be highly beneficial if used by people at the managerial level rather than at the operational level.

(3) Dishing up data users need to figure out how to use - Data mining ~~helps~~ reports will point out performance measures but no help in how to improve the performance. This mistake can be overcome by appointing someone who can ensure that the mechanisms are in place to address those problems.

(4) Training business users only at the beginning of project.

(5) The organisations ~~fo~~ go for a Big Bang.

* Real-time Decision Support: Integrating Results with Actions

It refers to putting analytics into action. A necessary condition for real-time decision support is integration of analytics into existing systems.

Challenges to be addressed are:

(1) Scaling analysis to large databases: Bringing in the large dataset into computer memory to be manipulated for real-time decision making.

(2) Scaling to high dimensional data & models: It is increasingly difficult to formulate

models on a very large number of variables in a dataset.

(3) Automating the search – Selecting a sample dataset from large number of samples should also be automated.

(4) Finding patterns and models understandable and interesting to users – Earlier DM techniques focused on measures accuracy and utility (benefit of pattern). These days techniques are expected to incorporate reports with high user friendly visualization.

TECHNOLOGIES ENABLING PRESENTATION

Why visualization is important?

Trying to find a pattern from a large dataset manually would take too many days which can be overcome by visual analytics. Eg. values corresponding to attributes which depicts an outlier.

Visual analytics mean use of computer graphics to create visual representation of large collection of information.

Scanned by CamScanner

Date
Page

* Performance Dashboards :-

Characteristics of Performance Dashboards (Three 3's)

(1) Three (3) Applications : Integrates monitoring, analysis management applications - Monitoring, Analysis and Management.

(2) Three Layers : Integrates the 3 applications with the informational layer allowing the users drill down to root cause of the problem.

(3) Three types : It can support operational, tactical and strategic decisions support.

* Balanced Scorecards

Dashboards graphically display various sources of information whereas scorecards show performance with respect to the targets

→ Scorecard focuses on four organisational perspectives — financial, customer, internal business process, learning and growth

→ Balanced scorecards are used to link long term strategic objective with short term actions.

Scanned by CamScanner

1. Translating the vision – helps in translating the vision into set of objectives and measures.
2. Communicating and linking – Balanced scorecards let business leaders communicate their business strategies, measures, objectives at all levels of the organisation.
3. Business planning.
4. Feedback and learning : Feedback can be taken from any organisational person or customer to improve the process.

* IT GOVERNANCE

It refers to specifying the decisions right and a framework to encourage desirable behaviour in the use of IT. The goal of IT governance is to determine, who makes the IT decisions and how a decision is made and what decisions have to be taken.
An effective IT governance should answer the above questions and in answering these questions, five inter-related IT decisions must be addressed.

1. IT principles : is about how IT is used in business. In order to define IT principles organisations should shortlist this business strategy.
2. IT Architecture : refers to enterprise architecture which depends upon the level of

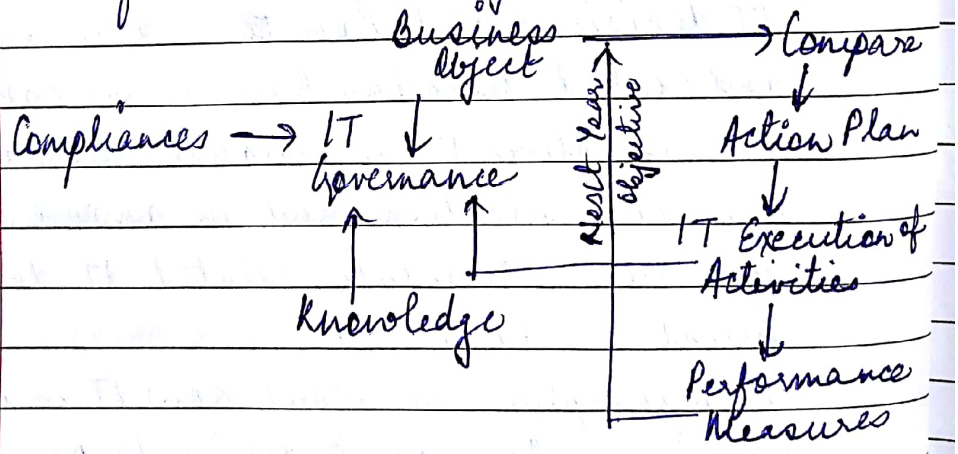
standardization and integration of data and services etc that is appropriate for an organisation.

(3) IT infrastructure: This includes technology components such as operating systems, database softwares, also include human infrastructure such as skills etc.

(4) Business Application needs: Refer to applications that facilitate a business process such as loan applications, complaint support systems. Two goals of business application are:
(i) Creativity (ii) Discipline

(5) IT investment & prioritization: Refer to how the organisation decides how much to spend on IT, on what to spend and ensure that investments are inline with strategic priorities.

* Role of Business Intelligence in IT Governance



* Impact of BI on Corporate Performance

Performance dashboards and balanced scorecards are helpful for organisations to

communicate their strategic objectives by providing the infrastructure to measure, monitor and manage key activities and processes that an organisation must perceive in order to meet its goals and KPIs.

Eg. of KPIs : operational cost, product schedules, customer loyalty and satisfaction.

→ The goal of using BI to develop KPIs is to provide decision makers with the following capabilities

(I) Collect data about operational processes
(II) Analyse data in real-time
(III) Prompt managers to take corrective actions
(IV) Detect events within a process. Eg. a fraudulent activity.

→ Ex. (step 1) : Collection of sales data w.r.t everyday so that a manager will know how many units have been sold out in a quarter/month, if at all the number of units sold out is not upto the target then analyse the reason (raw material, cost, quality) and take corrective actions.

Mistakes to be avoided while designing a dashboard :

Scanned by CamScanner

(1) Failing to apply the three 3's :

(a) Failing to deliver three applications.

→ Monitoring application – conveying data that is timely and relevant by ~~making~~ making use of all features of graphical interface.

→ Analysis Application – Allowing users to drill down into the details.

→ Management Applications – To allow alignment, collaboration, coordination amongst the tasks across processes.

(b) Failing to deliver three layers :

→ Not presenting three layers of information necessary to get root cause of a problem.

The three layers of information are:

(i) Transactional view of operational data – bottom most layers.

(ii) Multi-dimensional view of matrix – middle layer.

(iii) Graphical view of KPIs – top most layer.

(c) Failing to create right type of performance dashboard (operational, tactical, strategic)

Scanned by CamScanner

(2) A maximum of 7 KPIs can be presented at a time.
(3) Deciding an enterprise architecture
(4) Deciding on KPIs which are aligned with strategic objectives

Impact of BI (Dashboards & Scorecards) on Corporate Performance (KPIs):

(1) Communicate business strategy.
(2) Increase visibility on strategy goals
(3) Increase enterprise-wise coordination of activities.
(4) Better decision-making
(5) Consolidated and consistent view of business.

Scanned by CamScanner

Date 3/11/2017
Page

BUSINESS INTELLIGENCE TOOLS AND VENDORS

Figure 7.1 : BI tools in relation to capabilities.

BI tool is an application software which facilitates access, analysis of data and presentation of results.

BI tools put together enables four BI capabilities, which come together in a BI solution which are deployed in the organisation.

* BI tools classified into categories.

1. OLAP, data mining and reporting.
2. Based on cost, functionalities, BI complexity and number of users.
3. Based on the support for data, information, knowledge and decisions.
4. Based on BI capabilities.

* Customisation Vs. Standardisation of BI Tools:

Scanned by CamScanner

→ Customisation could lead to waste of time, money and effort. Standardization on the other hand will be overly restrictive

→ Customisation becomes more important as we go along organisational memory to presentation and standardization becomes more important and vice-versa

→ Customisation and standardisation are inversely proportional to each other.

* Three types of users of BI tools:

(1) Power Users: include analysts who have good understanding of the BI tools who can use advance features of the tools, developing their own reporting tools & who can obtain insights from BI tools.

(2) Business ~~Z~~ Users: are information users and lack BI expertise

(3) Middle-class Users: possesses some BI skills, occasionally design reports and queries or perform minor analysis.

* Recommendations for customisation and standardisation:

Scanned by CamScanner

Organisational memory should be standard, information integration tools should be customised w.r.t roles, insight creation tools should be customised w.r.t roles and tasks. Presentation tools should be customised as per roles, tasks and individuals.

7/11/2017

Future of BI

→ discusses about safeguard for corporate data, future directions of BI due to the increased use of social network, web 2.0 and enterprise 2.0

→ Certain recommended safeguards for corporate data/organisation data secure:
(i) appoint a high ranking corporate official in charge of information protection.
(ii) develop a comprehensive strategy for risk management.
(iii) develop policies, procedures and training protocols about the data standards.

Scanned by CamScanner

(iv) become fully PCI-compliant (pay card industry data security standard)

* Requirements of PCI:

→ maintain a secure network (firewall configuration, vendor-supplied-default passwords).
→ protect card-to-card holder data (encryption)
→ maintain a vulnerability management programme (anti-virus software secure system)
→ implement strong access control measures.

* WEB2.0 AND ENTERPRISE 2.0

Web2.0 is defined as world wide web technology that has dynamic web pages which enhanced creativity, information sharing, collaboration.

The use of web2.0 like applications within organisation is known as enterprise2.0 which includes
(i) search for information
(ii) authoring ~~for~~ such as wikis and blogs

Scanned by CamScanner

(iii) tags to categorise and share contents.

(iv) extensions are recommendations matched to user preferences.

(v) signals. Eg. RSS feed which is a data format. used to provide users to subscribe to a specific content

Scanned by CamScanner