TABLE OF CONTENTS

ICT394Ans1

Business intelligence

Business intelligence is composed of data itself, the analysis of the data and the presentation of the data in a format that is easily consumable by the users who requested it. Through methods, technologies and tools business intelligence seeks to help improve organisational decision making thereby resulting in positive organisational outcomes such as increased profit. While the impact of business intelligence can be positive there are downsides/cost in adopting business intelligence. For example- Loyalty cards for BI monitoring customer behaviour

https://www.netsuite.com/portal/resource/articles/business-strategy/business-intelligence-examples.shtml

Downsides of Business intelligence (impact)

Downsides of Business intelligence include-

Often requiring the appropriate hardware to enable BI and the right personnel often specialist to implement and sometimes operate

> For example, loyalty card BI. In order for the store owner to make use of the BI a particular software needs to be installed on their retail PC. Sometimes the retail PC are out of data and thus can't use it so hardware upgrades are needed and guidance on how to use it

Difficulty of designing and implementation of BI often leading to implementation taking too much time and sometimes ending up with little benefits

> For example, retail company loyalty card needs to persuade users to use it, and this often requires financial incentive. If not, enough people are persuaded then there is no tangible benefit and the financial incentive given would be wasted

Often the introduction of business intelligence to an organisation causes operational disruption

> For example, loyalty card disruption when upgrading retail register PC to enable the loyalty program implementation

Advantages of Business intelligence (impact)

Advantages of Business intelligence include-

Allows for better forecasting making predictions more like to be accurate. This leads to better decision making and a positive organisational outcome

> For example, Loyalty card store the transaction of the customer. This loyalty card thus enables for better predictions when to stock up on an item

Allows decision makers to take on more responsibility confidently since they have more available information

For example, Loyalty card stored the transaction of the customer. This data can then be analysed to help explain why a customer has potentially stopped shopping. This then allows the decision makers such as manager to make decisions that may help remove one of he reasons why the customer left or offer discounts

Reduced (operational/resource) cost since business intelligence can automate a good portion of task that in the past required manual

For example, in the past the transaction would be tracked, and relationship and analysis had to be done manually. Thus, not often ignored, the introduction of BI allowed retail to find out information without straining resources

Components of Business intelligence-

Data is related to the raw numbers, letters and strings that have been captured in some way. Commonly stored in database. The data is collected first.

The data can be structured which is often quantitative allowing for straight forward analysis.

For example- Purchases, enrolments or orders

For example- loyalty program the customer transaction. This is structured data.

The data can be unstructured, but this mean they are difficult to analyses.

For example- Emails, Help desk calls, and social media post

Secondly the data is prepared for analysis. This is where ETL occurs

Analysis is where the prepared data is used and often relationships and information that might be useful for decision makers are fleshed out. The data can be combined with other data, aggregated or summarised

For example- for loyalty program the sales of different types of milk like full cream or fat free milk are worked out. Information such as the current trend of which milk type is most popular is revealed

Presentation is where all the information and relationship discovered during analysis is conveyed in a format that can easily be taken in by the decision makers. This may be in the form of graphs or tables or dashboards. Users will have the information readily available.

For example- for the loyalty program the sales of different types of milk like full cream or fat free milk are worked out. The store manager has access to this information and can easily consume the information since it is presented in an easily visualisable format. Therefore, they are empowered to make decisions like to stock up more on a brand

Critical success factors determine how well a business intelligence is implemented. The factors include-

Data is of integrity and high quality because having bad data in the BI will not only provide decision makers wrong information but will deter adoption going forward

Ongoing support by management

Business case for BI and vision is established clearly

User-oriented change management

ICT394Ans2

Business Intelligence (BI) development life cycle

Business intelligence (project/applications) development life cycle: once deployed the product is improved on and enhanced as a result of the feedback provided by users or product

Business Intelligence (BI) project justification

Justification is where new engineering projects are looked at and the reasons by which the business requires the engineer project is scrutinised. Needs to be business driven and not technology driven since expensive. Must reduce a quantifiable business issue. To justify BI application/project development there are four main business justification components- Business drivers, business analysis issues, cost-benefit analysis, and risk assessments

> ***Business drivers*** is where both the business strategic goals and BI project/application goals are identified. This makes sure there is some sort of alignment between the business strategic goals and BI application goals. Failure to ensure this makes the BI project prone to failure.
>
> > For example- inappropriate business drivers may include the business strategic goal of server company wanting to increase revenue by purchasing older server for customer use. On the other hand BI application goal is to allow for customer to better self-manage server thus increasing customer experience. The issue is that with the older server they may not be able to run the BI application smoothly thus ends up reducing customer experience
> >
> > For example- server company objective could be to reduce the amount of technician required maintain and manage the servers thus saving money. The justification for the BI application is to provide customer information about server health and

uptime so that the customer himself can manage his own server. These two objectives are aligned meaning success is high

Business analysis issues is another reason to justify the implementation of a BI application. Business analysis issues involve information required by stakeholders not being easily available from current systems and data source quality is poor. To justify the implementation of BI application for this context it is important to identify the business issues and possible data sources that provide the information to resolve the identified business issues. Where the justification of BI application may be diminished is when merging and standardising the possible data sources for BI is complex and leads to poor data quality overall. The reason why merging possible data sources may lead to poor quality data is because the data sources come from different types of sources.

Operational data is the internal operational data of subject areas, and it comes from online transaction processing and batch systems. Financial, logistics, sales, order entry, personnel, billing, research and engineering. For example- the business issue such as why aren't customers from New Zealand purchasing our servers? The operational data that could help with this is customer billing information and personal information.

Private data is the internal department data which belongs to business analyst, manager and statisticians. Product analysis spreadsheet, regional product usage spreadsheets, prospective customer databases. For example- the business issue such as why aren't customers from New Zealand purchasing our servers? The private data that could help with this is prospective customer database

External data is the external data which is purchased from vendors who have industry specific information. For example- the business issue such as why aren't customers from New Zealand purchasing our servers? The external data could be competitive data of sales of servers

following:

- Health care statistics
- Customer profile information
- Customer catalog-ordering habits
- Customer credit reports

External data is usually clustered around the following categories:

- *Sales and marketing data:* lists of prospective customers
- *Credit data:* individual credit ratings, business viability assessments
- *Competitive data:* products, services, prices, sales promotions, mergers, takeovers
- *Industry data:* technology trends, marketing trends, management science, trade information
- *Economic data:* currency fluctuations, political indicators, interest rate movements, stock and bond prices
- *Econometric data:* income groups, consumer behavior
- *Demographic data:* age profiles, population density
- *Commodity data:* raw material prices
- *Psychometric data:* consumer profiling
- *Meteorological data:* weather conditions, rainfall, temperature (especially for agricultural and travel industries)

Reasons why merging data may lead to poor quality data

One of the difficulties in merging and standardizing data from different types of data sources is that the data is stored in different file structures on different platforms. What makes the process even more difficult is that the keys for the same objects on different data sources usually do not match, the definitions for the same apparent data are often inconsistent, and the values are often missing or conflicting. In addition, different people in the organization have authority to determine business rules and policies for data from different types of data sources and resolving data conflicts among them or getting clarification is often all but impossible**.**

Cost-benefit analysis is done to help justify the implementation of a BI project. It outlines how the BI project either solves a business problem or facilitates a business opportunity. A BI project should have at least one of the following benefits

Improved customer satisfaction through. In this context the implementation of a loyalty card and increased customer use of a loyalty card provides incentives to the customer to continue shopping this increased customer satisfaction. Furthermore, the BI application can inform manager when to stock up on an item during a specific period

Increased savings. The implementation of BI should have some savings on the business. In this context of a loyalty card retail the BI should increase savings by only purchases stock that is sold. They help decision makers in stock procurement. Furthermore, the implementation of BI application reduces the store manager role of keeping track of sales and present useful information to him automated

Increased market share gain. The implementation of BI could increase the business market share. In this context, the loyalty card aims to provides incentives for customers to continue shopping thus helping retain customers

Increase in profit. The implementation of Bi should increase profit. In this case, the loyalty card BI transform the way the business does promotion. Provides a promotion to the right customers at the right time such as only mailing to customer when BI suggests it.

Whether revenue is increased such as-

1. **Revenue increase**, possibly in the form of:
   - Identification of new markets and niches
   - More effective suggestive selling
   - Faster opportunity recognition
   - Faster time to market
2. **Profit increase**, including possibilities for:
   - Better targeted promotional mailings
   - Early warning of declining markets
   - Identification of under-performing product lines or products
   - Identification of internal inefficiencies
   - More efficient merchandise management
3. **Customer satisfaction improvement** through:
   - Improved understanding of customer preferences
   - Improved customer-to-product matching
   - Up-selling to customers
   - Increased repeat business
   - Faster resolution of customer complaints
4. **Savings increase** through:
   - Reduction in wasted or out-of-date merchandise
   - Reduction in requests for customized reporting
5. **Market share gain** through:
   - Increased numbers of customers who defect from the competition
   - Much higher customer retention rate as compared with previous years and with the competition

Risk assessment looks at six variables that may diminish the justification of BI application

Risk Assessment (refer to textbook)

Risks are factors or conditions that may jeopardize a project. Risks should be assessed for the following six major variables:

1. The technology used for implementing the project
2. The complexity of the capabilities and processes to be implemented
3. The integration of various components and of data
4. The organization and its financial and moral support
5. The project team staff's skills, attitudes, and commitment levels
6. The financial investment in terms of ROI

Planning is the stage where the strategic and tactical plans about how the BI product will be engineered and deployed are fleshed out. This stage involves enterprise infrastructure evaluation and project planning.

The enterprise infrastructure encompassed the technical infrastructure and the non-technical infrastructure-

Technical infrastructure consists of hardware which must be able to handle complex analysis, querying simple and complex data and producing report. Also, scalable. Consist of

middleware which is layered between application and operating system providing interaction between the application and environment e.g. client and server

Non-technical infrastructure ensures that cross organisational questions can be answered. Consist of logical data modelling is where logical data relationships are outlined. Consist of meta data capture is where description of business functions, process and data quality are kept based off guidelines

The components of project planning involve-

Business involvement so who the stakeholders are and if they need to communicate often. Also, the level of involvement this project manager is committing.

Project scope and deliverables are outlining what is produced and how detailed the requirements are. And does the resources and schedule match the scope

Cost benefit analysis has it been completed and what is the return on investment and when to expect a return

Infrastructure is where we check whether our infrastructure has gaps and identify the components required for BI project

Staff and skills look at identifying team members and whether the members have enough skills and training suitable for their assigned roles. And role of project manager whether its full time

Business analysis stage is where the requirement from the business is helped fleshed out as a result of performing a detailed analysis of the business problems and business opportunities

Design stage is where the increased understanding of the business opportunities and business problems is used to help devise the ideal BI product

Construction stage is where the devised BI produced is made within the set time frame and should provide a return on investment

Deployment stage is where the product is rolled out either implemented or sold and measured for its effectiveness and whether the product meets its requirements

Traditional system development project vs BI

Traditional Business intelligence development:

Was built without the idea of systems interacting or working without systems. The system in the past aimed to resolve only one aspect of the problem.

For example- previously at university each department used own reporting software.

Modern times has cross organisational activities thus relying on BI systems working with other systems.

For example- provide non-IT expert ability to dig into data via new precompiled report and answer questions they want answered

The traditional BI development is initiated by outstanding business requirements rather than an opportunity for the business.

For example- previously in University of Konstanz the systems aimed to solve business requirement for a department such as HR and one for student enrolment

The traditional BI development also tends to create BI products that doesn't get continuously iterative development like modern BI development applications

ICT394Ans3

Operational vs Analytical information

Operation information is considered transactional information meaning the information is generated by individual day to day transactions and assist with daily business operations and sometimes supports other organisations. Given the operational information normally consist of daily transactions the information is stored for days and months. Tends to be detailed data. Also, operational information tends to be used for daily business operations by all employees. Retrieved by operational databases

> For example- ATM withdraws, and airline ticket purchases or server purchases, car sales (think no context just sales nothing like what stage of life is the car preferred)

Analytical information utilises operational information or external data such census and generates information that assist with analytical duties. On the other hand, the data is stored yearly because in order to generate analytical information it takes time to collect enough data. Tends to be summarised information. On the other hand, tends to be used for decision making by less employees. Retrieved by data warehouse

> For example- what time of day are servers most used or how does socio economical areas affect the type of server purchased. What car to purchase at a specific time in a person life

Data warehouse (special type of database) enables the storage of detailed and summarised data and the retrieval of such *analytical information* for relevant stakeholders. ~~The purpose is to enable the analysis of specific business subject areas. Contains data that span for a large time horizon. Since its purpose is retrieval of analytical information it is not suitable for direct data entry by users thus data in the data warehouse is not subject to modification, insertion and deletion by users~~. Also, the data warehouse duties may rely on multiple operational data sources, which a single operational database may not be able to facilitate. It is created as a separate entity its own database because the duties of the data warehouse would impact the performance of the operational databases. It is considered a structured repository of integrated, enterprise-wide, historical, subject-oriented and time-variant data.

> **Structured repository** means the warehouse is considered a database that has a structure represented in its metadata and contains analytical information
>
> **Integrated** refers to bringing together the analytical useful data from different operational databases to data warehouse

For example- A server company may have a database for marketing (marketing information) and customer service such as service sales. Potentially the server sales data can be used with the advertisement information data. This allows organisation to analyse the effectiveness of advertisement on sales for example

Subject oriented refers to the distinct purpose of data warehouse which is facilitate the storage and retrieval of analytical information for specific business subject areas

For example- What time of day are servers most used

For example- cost may be the specific subject area. We are able to retrieve a lot of analytical information about cost

Enterprise-wide means the analytical information can be viewed organisation wide

For example- A server company may have an analytical information about cost (business subject area) coming from many operational databases that have data about cost

Time-variant means the data warehouse contains data from different periods of time. Helps with comparing years

For example- server data warehouse they contain customer service sales over 10

Years

Historical means the data in data warehouse has a life span of multiple years

Data warehouse (analytical database) vs Operational database

Operational database enables the storage of operational information which is the information is generated by individual day to day transactions basically transaction level data. The purpose is often to assist with daily business operations. Also, the database contains data that span for a short time horizon. Suitable for direct data entry by users. The data in the database is subject to modification, insertion and deletion by users.

On the other hand, **data warehouse** (special type of database) enables the storage of detailed and summarised data and the retrieval of such *analytical information* for relevant stakeholders. The purpose is to enable the analysis of specific business subject areas. Contains data that span for a large time horizon. Since its purpose is retrieval of analytical information it is not suitable for direct data entry by users thus data in the data warehouse is not subject to modification, insertion and deletion by users. The data warehouse duties may rely on multiple operational data sources

Data warehouse components include-

Source systems are the operational databases and operational data repository that feed operational data that are handy for analytical information to data warehouse. It also includes external data sources such census, market research data and weather data. The data is repurposed for the data warehouse. The operational data is used for the original operational purpose as well

Extraction transformation load (ETL) infrastructure allows for the collection of operational data from the operational database and the loading of the operational data into the data warehouse. The tasks include- collecting only useful operational data that is relevant for analytical information, transforming the data so that it suits the structure of the data warehouse model and is of decent quality, and finally loading the transformed data into the target data warehouse

Data warehouse refers to the destination of operational data useful for analytical information

Front-end applications provide end users access to the data warehouse for their analyses. There can be multiple front-end applications for a single data warehouse based on different requirements from different groups

For example- dash boards or reports created

Data warehouse development process steps-

Data warehouse requirements step is where the desired functions and features of the data warehouse are gathered and written as requirements. The operational data in the internal source systems and potential external data sources must be able to support the requirements that are generally analytical requirements. Furthermore, in order to obtain the requirements different methods such as focus groups, interviews, questionnaire and survey are completed by relevant stakeholders of data warehouse. The collected requirements will need to represented using a conceptual data model such as dimension modelling. Iterative in nature during development because requirements may change throughout process because our understanding of the environment broadens.

Data warehouse modelling step is (next topic 4) the creation of a data model for the data warehouse that will be implemented by DBMS

Creating the data warehouse (topic 5) is completed through the use of a relational DBMS application and it implements the data warehouse based on the created data model

Creating the ETL infrastructure where procedures are created for the ETL tasks. This tends to be the most time and resource intensive step since a lot of details need to be considered such as how to make sure the warehouse has no duplicate operational data. The number of details required to be fleshed out when creating means it is most time and resource intensive part

Creating front end applications is referred to as business intelligent applications and contains a navigation method for different interfaces this step designs and develops it. Needs to wait for data warehouse to be created/deployed in order create front end application by connecting to it

Data warehouse deployment is the step where the business intelligent application and data warehouse is released to end users

Data warehouse use step is once released the end user will be able to indirectly use the data warehouse through the business intelligent application (front end application) or direct use it through the DBMS or OLAP utilities

Data warehouse maintenance and administration step is where actions to support end user and manage technical issues is performed. These include providing security for data warehouse information, ensuring enough storage space is available for data warehouse data and implement ability to backup and recover data

Data mart

ICT394Ans4

Dimensional modelling provides a way to design subject-oriented analytical databases such as data warehouses and data marts. Follows the common relational modelling concepts such as having primary key, foreign keys and abiding by integrity constraints. This type of modelling differentiates dimensions tables and fact tables

Dimension tables vs Fact tables

Dimension tables are tables that provide a description of subject that will be analysed often the subject has something to do with the business organisation or enterprise. For example, sales (subject). Each column in the dimension table is used to describe the subject often either textual for example- produce brand, produce colour or numerical- product weight, cost of product. The dimension table serves information that can be used for further analysis of the subject. The dimension tables tend contain static data. The dimension tables tend to have fewer records than fact tables

Fact tables are tables that provide measurements of the subject of analysis and the foreign key (think table full of calculations and maths). These measurements tend to be numerical and are used for the purpose of quantitative analysis and mathematical computation. For example, sales (subject) then measure would be total sales amount, and profit amount. The fact tables contain data that is continuously added and table grows in size. Facts table contains FK linking to the PK of all the dimension tables. Sometimes contains no measurements columns such as transaction time and identifier but can be used such as grouping time for analysis. Two types of fact tables include-

> **Detailed fact tables** where each record has a single fact. For example- sale fact table
>
> **Aggregated fact tables** are each row sums up multiple facts. Typically, primary key is composite key of all dimensions. So essentially putting facts from other fact tables into one table. For example- sales per customer, day. Store fact table

Star schema vs Snowflake model (schema) vs Multiple facts table (schema)

Multiple facts table/constellation in a dimensional model means multiple subjects to be analysed can all use the same dimension tables. This method allows for easier analysis of different subject thus revealing any links and relationship between them. In addition, since dimension are reused

creating additional multiple facts dimensional models will be quicker. The approach provides fast queries and allows for broad range of information to be retrieved.

Star schema contains facts table at centre which is the subject of analysis and surrounded by dimension tables through primary and foreign key. Often referred to as dimensional schema. Facts table contains FK linking to the PK of all the dimension tables. Given a surrogate key (system generated key often auto incremented) to dimension tables

Snowflake model is where the dimension themselves are normalised for star schema. But rarely used since normalising dimensions make for difficult analysis and not necessary for analytical databases since tend to be read only or append thus no update anomalies which defeats whole purpose of normalising. Normalised means query process is more complex due to more tables/dimensions to be joined

> For example- The process to query is more complex since we are dealing with more tables/dimensions as opposed to same query but with less dimensions (denormalised)

Normalised is where instead of having two tables and combining into one you separate it out

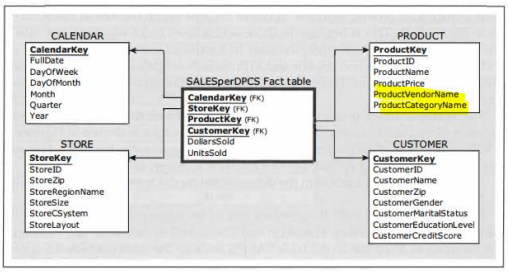
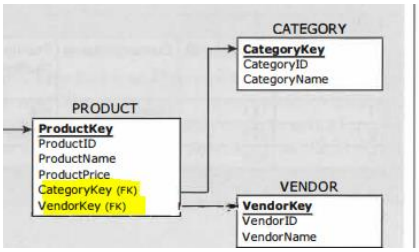
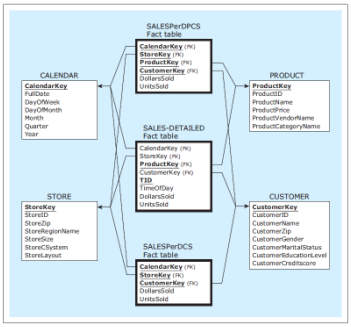


FIGURE 8.24 A dimensional model with an *aggregated* sales fact table: SalesPerDPCS (per day, product, customer, and store).

*Star schema

*Snowflake model





* Multiple facts table

Granularity

Granularity is used to explain what meaning of one row in a fact table may have. So, level of detail of each row. *A detailed fact table* has a fine level of granularity because each record (row) has a single fact. An *aggregated fact table rough/not fine (coarser) granularity* level because each record contains the summarisation of multiple facts rather than a single fact

> For example- SalePerDayPerProductPerCustomer fact table VS Sales fact table where Sales record summarise sales facts only

Advantages of coarser (rough) granularity typically aggregated fact tables include

Provides easier and quicker query of information because the aggregated table tends to store multiple facts rather than a single fact. And the number of rows accessed when responding to the query is less causing improving performance.

For example- So if a customer wanted sales per day, then that aggregated data is designed and available to provide that fact quick

Allows the usage of pre-calculate data from the data warehouse that can be accessed fast

Adds more time to the ETL process

Disadvantage of coarser (rough) granularity typically aggregated fact tables include

Provides limits to what information can be queried from the fact table. So, while aggregated fact table are designed to allow for fast query of specific information the information that was not intended to be queried fast can often not be queried at all. Unlike detailed fact tables, fine granularity table, which was designed to allow for a broad range of queries.

Reduces how flexible fact table itself can be used in different context. Since with aggregated fact tables they are designed for specific purpose, we are unable to reuse it in different context. Can't unaggregated the record

Line-item detailed fact table VS Transaction-level detailed fact table

Transaction-level detailed fact table is where each row represents one transaction. Typically, single primary key. Tends to be the finest granularity

Line-item detailed fact table is where each row represents a line item of a transaction. Typically, composite primary key where one of the keys is TID (transaction key).

Slowly changing dimensions (issue) are dimension in dimension models that have attributes that often change rather than not or rarely change at all. Must be aware of it during our development of dimensional modelling due to it impacting our aggregations. For example- attributes that change often can be salary, tax bracket, and employee age. Attributes that don't change or rarely change include- store size and customer eye color

Ways to deal with slowly changing dimensions (issue)

Approaches to dealing with slowly changing dimensions include-

Type 1 approach is the easiest approach it is where the incorrect attribute record is changed to the correct value. The downside old values are not kept. Commonly used when the incorrect value was arisen because of human error or error

Type 2 approach is where a new record is created with new surrogate key and new value every time dimension record changes. Used when old values need to be kept. Often includes timestamps and row indicators attribute. Timestamps show the time interval when values are correct while row indicator convey whether record is still valid

Type 3 approach is where a current and previous column is created in the table for each column where we expect changes (so column1 becomes previous column1 and current column1). Can be used in situations where only limited history is to be recorded or there is a limit on the changes possible for a column in dimensions. Can be used with timestamps to add additional information

ICT394Ans5

A database management software (DBMS) implements the dimension modelling. The data warehouse may be implemented as a database in the DBMS or spreadsheet

ETL process- (task that is done in order to facilitate the BI three activities)

Three main phases in data preparation

Extract

Extraction is the phase where operational data that could be analytically useful is collected from many operational data sources. For example- customer database, order database, invoice database, product database, or account database. The analytical useful data will be potentially loaded into the BI database. Determining the data to extract is worked out in the requirements and dimension modelling stage (topic 4) which also outlines the data source available. Furthermore, the data model outlines the procedures of the extraction tasks.

Factors influencing extract process

Factors influencing when/how data will be extracted from data source-

Who owns the data source is a factor; this may be that the data source may be owned by another part of the organisation such as human resource section or owned by external source such as different organisation. Thus, limitation placed by the owner on the capacity or part of data to extract from the data source may be placed

> For example- In a server company there may be a human resource unit and feedback by customers may not extracted

The nature of the source system could also be a factor; this could be that the nature of the source system is high throughout transaction processing system such as ATM where many transactions are made thus the capacity at which data can be extracted determines the duration of data extraction. Or if by nature the data source removes recent transaction then the data will be extracted constantly since the data source removes recent transaction.

> For example- In a server company there may be a customer invoice database it could delete invoice after thirty days. If the requirement was to have invoices over longer than 30 days then the extraction will need to occur constantly.

The use of source system could be a factor; if the source system was for online transaction processing system such as ordering item system, then during extraction this may slow down the ordering item system which may negatively impact profit which is not ideal

> For example-

Methods to extract data

Make full copy of the source data. Ideal for small dataset but large dataset require greater data preparation prior to transformation. Also, sometimes with larger data set not everything is used

Considers the following- operational data source will need to be still operating well and extracting from the data source is completed fast

Issues with extract process

Challenges of extraction include-

Source systems may contain a lot of redundant data between them which could make extraction process more difficult

> For example- in a server company there could contain account database and invoice data base. Both of the database contains the name of customer which is redundant data. So, this make extraction more complex in that we need to work out whether customer names located in account database is more accurate than customer names in the invoice database

Limited access to the data source system could make extraction more difficult.

> For example- In a server company there may be a human resource unit and feedback by customers may not extracted even though it is needed

Staging is a challenge of extraction which is working out where extracted data is to be stored prior to loading for transformation. This applies to the use of large data sets and many data sources

Extraction process must produce metadata to keep records of decision. For example-

> Who owns it?
>
> How is it formatted?
>
> What does it *mean*?
>
> It is complete?
>
> How often is it updated?

Transform

Transform is the phase is where the structure of the extracted analytically useful operational data is modified or changed so that it fits the BI database model structure. For example- making data into the same units (CM) or data types or names. Transformation often takes up 80 percent of the ETL process because it involves data cleansing which is the identification and correcting of low-quality data to ensure data is analytical useful and involves making sure the data is consistently high quality. Thirdly, transformation has to deal with the problems in the data source as discussed below. Furthermore, if the operational data is large resolving the problems is even more time consuming.

Issues with transform process

Problems/Issues with transformation include-

One problem could be the data format of operational data could be different from the data format of the BI database. This makes transformation more complex since there is need to work out a method to convert

For example- the date format in the operational data could be 30/3/2022 while in the BI database the date format could be 30-Mar-2022. Thus need to consider

For example- In a server company, the operational data of currency could be in AUD but the BI database is in USD

Another problem could be the data values of operational data could be inconsistent. This makes transformation more complex since there will be a need to reconcile the inconsistent data values and work out which data values is most accurate. So often organisation will duplicate their data set but sometimes when they update a data value in one of their data set it doesn't get updated for all the other duplicated data sets. Thus, creating inconsistent data values which means need to work out which duplicate data set values is most accurate

For example- server company may create a duplicate invoice to customer. However, one invoice gets updated and duplicated doesn't. Customer then pays from the old invoice which means the updated invoice is still created but not paid

Third problem could be operational data name for an element may be a synonym or a homonym for the associated BI database element name. This makes transformation more complex since there is a need to work out a standard name for the data element in the BI.

For example- server company operational data could include data element serverCost while the BI database has data element name price. So standard name could be serverCost

Fourth problem could be operational data primary keys could be different from the BI database primary key. This make transformation more complex since a standard primary key name needs to be decided often a surrogate key for BI database.

For example- server company operational data could include data element serverKey and serverNum both synonyms. The BI database could create a standard primary key serverID which consolidates both

Fifth problem could be the operational data could contain missing data values. The transformation will need to determine how to handle the missing values either ignoring or replacing the record or using a default value

Transformation process must produce metadata to keep records of why a transformation is occurring since transformation can occur many times. Metadata for transformation may include-

Formulae for conversion

Aggregation to a suitable level of granularity

What needs to be done to clean the data for the DW

Replacement of missing values

Load

Load is the phase where the transformed data from the transformation phase is put in the BI database. The insertion is done through a batch process without end-user involvement. This phase must determine the frequency of the data loads

Approaches to loading data

There are three types of ways to load data into BI database-

Initial load is where the current operational data is loaded into an empty BI database. Depending on the time horizon of the operational data this initial load may be process intensive. Must ensure the loaded operational data fits the associated data element in BI database.

Incremental load refers to the loading of current operational data into BI database after the initial load. It refers to any loading that happens after the initial load. Refresh cycle defines when new operational data will be loaded into the data warehouse can be hourly, daily or monthly. For active data warehouse the refresh cycle occurs continuously (real time) delivered in micro batches. The refresh cycle is determined in advanced and is based on users of data warehouse analytical needs and technical feasibility of the system. Extract deltas only is preferred because large volume involved

> For example- university unit enrolment information of each student where data is hourly monthly. The data loaded could be student has changed his unit enrolments by adding new units or deleting units.

Table 9.2. Incremental Load Options

Extract All Records	***Extract Deltas Only***
Extract source data from all operational records, regardless of whether any data values have changed since the last ETL load or not.	Extract source data only from those operational records in which some data values have changed since the last ETL load ("net change").

Historical load is where archived historical operational data is loaded into an empty BI database. The data is older than time horizon of operational system. Must be aware of changes in data formats and types over the time horizon.

> For example- university unit enrolment information of each student spanning over the lifespan of the university. Unit codes format changed with Murdoch ICT206 in 2016 to = BSC275

Load process must produce metadata to keep records of how the data were loaded. Metadata for loading may include-

> What loading schedule?
>
> How much to load?
>
> What happens if repeated data is attempted to be loaded?

Importance of metadata

ICT394Ans6

Online Analytical Processing (OLAP)

Online Analytical Processing (OLAP) is the utilisation of tools that enable multidimensional views of data and allows for their analyses through different windowing methods. OLAP allows for data structures to be multidimensional. “Traditional” query processing is suitable for data that is in two-dimensional data structures

Two-dimensional data structures are where data is kept in columns and rows. Often for basic transactions. Advantage is that it enables simple queries and resources to query is low but if queries are needed for aggregated data, then it becomes resources intensive. For example- relational model (databases) with rows and columns for individual transactions

Multidimensional data structures are where data is kept in not only rows and columns (2D) but more dimensions. Often for aggregated data. Advantage is that it enables querying of aggregated data using only limited resources. Disadvantage is it is more difficult to draw models representing multidimensional data structure

> For example-cube data structure. This data structure has more than just rows and columns. The advantage Is that fastest querying since the extra dimension give certain pre calculated values also usual advanced queries can be simply done on a single table thus easier query

OLAP operations

OLAP operators include- (tools enable simple performance of operators on multidimensional data structures for end-users. Supports analysis required for decision making. Easier than manually writing SQL. Can help design front-end BI applications. Example of OLAP tools is tableau or excel)

- Graphically visualizing the answers
- Creating and examining calculated data
- Determining comparative or relative differences
- Performing exception analysis, trend analysis, forecasting, and regression analysis
- Number of other analytical functions

Slice and dice is where a multidimensional data structure can have dimension attributes added, replaced or eliminated in order to allow the data structure to be examined from different perspective and queried a particular way. The Slice part can essentially pick one dimension and the attribute(s) and restricts all the values based on that attribute(s). The dicing part can essentially pick more than one dimension and attribute(s) and restricts all the values based on that.

> For example- if you have a multidimensional data structure showing five different server companies (dimension 1), name of servers sold (dimension 2), and years sold (dimension 3). And you want to query all five server companies, their name of servers sold, and time sold

then you would need to use slice and dice to replace all the data from years sold with time sold. This is where the dimension years sold is selected and replaced with times sold dimension. Specially a slicing operation

For example- if you have a multidimensional data structure showing five different server companies (dimension 1), name of servers sold (dimension 2), and years sold (dimension 3). And you want to query of only two server companies, and only intel servers sold (name), and years sold. This is where the dimensions different server companies are selected and the attributes for the three companies are eliminated. Also, this is where the name of servers sold dimension is selected and all the attributes except intel is eliminated. Thus, this a dicing operation.

For example- if you have a multidimensional data structure showing five different server companies (dimension 1), name of servers sold (dimension 2), and years sold (dimension 3). And you wanted a query of only two server companies, their name of servers sold, and years sold then you would use slice and dice to remove all the data from the 3 other companies thus only showing two server companies' information. This where the dimension server company is selected and the attributes for the three companies are eliminated. Specially a slicing operation

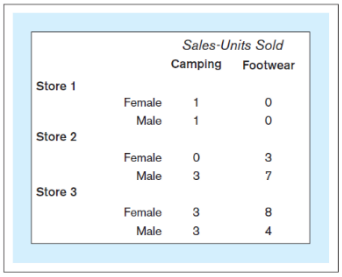
Pivot (rotate) is where a multidimensional data structure has its dimension rotated from one axis (either on x or y) to the other axis. Without removing or change values. Provides a new way to understand and query the data.

For example- if you have a multidimensional data structure showing five different server companies (dimension 1), name of servers sold (dimension 2), and years sold (dimension 3). You can move the name of servers sold dimension axis to years sold dimension axis. This results in a multidimensional data structure showing five different server companies (dimension 1), years sold (dimension 2) and name of servers sold (dimension 3)

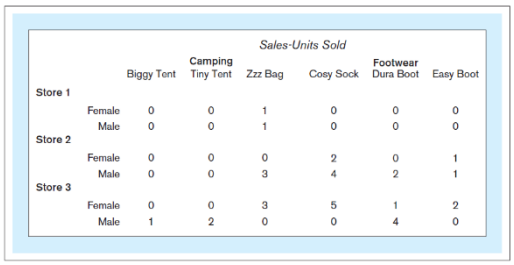
Drill down is where a multidimensional data structure granularity is shaped finer. This is where less detailed record is drilled down to be more detailed often by stepping down a hierarchy. Each record contains a single fact.

For example- if you have a multidimensional data structure showing five different server companies (dimension 1), name of servers sold (dimension 2), and year quarter sold (dimension 3). You can drill down by stepping down year quarter sold dimension to month sold. You can also drill down name of servers sold dimension to server serial number and country name. For both drills down will result in finer granularity. Instead of the year sold quarter sold you get more detailed information about specific month of year. May takes more cognitive load to understand the BI

~~The example below shows this since the first table shows summarisation of multiple facts such as camping and fact about footwear. But in the second table it is shaped finer meaning each record is a single fact about what items were sold to person rather than number of camping and footwear items sold to person. You can also drill down name of servers sold dimension to server serial number a~~



		Sales-Units Sold	
		Camping	Footwear
Store 1			
	Female	1	0
	Male	1	0
Store 2			
	Female	0	3
	Male	3	7
Store 3			
	Female	3	8
	Male	3	4



		Sales-Units Sold					
		Camping			Footwear		
		Biggy Tent	Tiny Tent	Zzz Bag	Cosy Sock	Dura Boot	Easy Boot
Store 1							
	Female	0	0	1	0	0	0
	Male	0	0	1	0	0	0
Store 2							
	Female	0	0	0	2	0	1
	Male	0	0	3	4	2	1
Store 3							
	Female	0	0	3	5	1	2
	Male	1	2	0	0	4	0

Drill up is where a multidimensional data structure granularity is shaped coarser meaning *(coarser)* each record contains the summarisation of multiple facts. This is detailed record is drilled up to be more aggregated often by stepping up a hierarchy.

> For example- you have a multidimensional data structure showing five different server companies (dimension 1), name of servers sold (dimension 2), and year quarter sold (dimension 3). You can drill up by stepping up year quarter sold to year sold. This allows up to view the summarised data of year so granularity is coarser. Takes less cognitive load to understand the BI

ICT394Ans7

Business Analytics

Business Analytics is the use of data to extract insights about why an event has occurred and provides predictions of what may occur in the future ~~(predictive analysis)~~. As opposed to **Business Intelligence** where existing data about current and past events are analysed to inform of what had happened ~~(descriptive analysis)~~. Business intelligence reveal trends and patterns without looking into why or extracting insight on future trends, pattern, or predictions. They provide IT and tools to provide business analytics such as- data warehouse, visualisation, data mining tools and OLAP

> For example- Allrecipes where they keep track clients are doing on and off and what people think so content can be tailored to them. Increased revenue because selling more stuff due to interactions and Audience development due to more effective marketing from data and tailored to customers. Aware of trends in regards to festive events and tailoring content around those events based on location and nationality data of customers

Methods Business Analytics provide insight

Methods business analytics provide insight through

Reporting is where past data is summarised and presented

[Determining] Trending is where time-series data and their patterns is revealed

[Determining] Segmentation is where similarities in the data is revealed

For example- (units are the data and) what units is taken together

[Working out] Predictive modellings is where historical data support prediction of future events

Criteria for BA success-

Requirements for ensuring business analytics is useful (such as producing positive business outcomes)-

Business relevancy

Insight must have practical value meaning whatever insights are generated helps future business operations

For example- organisation can create accurate model but can't be implemented in the organisation system

Capacity to measure value of insights meaning the effectiveness of the business analytics can be constantly monitored to ensure it is still useful or of some use

Business Analytics types

Types of business analytics- (provides insights)

Descriptive analytics looks at historical data that comes from a consolidated data source and determine what has happened or is happening. Reveals trends and patterns but can't be extrapolated in its current state to make predictions about the future. Rather explains the patterns and trends of the past. The descriptive analytics tend to be presented as a visual representation such as- line, bar and pie charts.

For example- Business report where they outline the cash flow, revenue, inventory, expenses and production

Predictive analytics provides insight into what could happen in the future. Uses statistical techniques and other techniques to forecast the probability of what could happen in the future.

For example- Working out which customers have the highest probability of purchasing product thus send marketing campaign to them

Prescriptive analytics provides insight into the best course of action in order to achieve best performance. It weighs up the possible course of action against the different future scenarios. For making sure a system is performing at an optimal level.

For example- Tiktok algorithm for you feed shows videos you want to view

Descriptive analytics report(s)

Descriptive analytics report should contain information relevant to the domain we are interested in. They support management decision making by providing insights. The reports should come from external as well as internal sources. Also, the reports should created in an iterative cycle in order to better reveal the trends based of certain period of time.

The types of descriptive analytic report include-

Metric Management Reports is where performance is compared against desired outcome. They tend to be a snapshot in time. Contain basic visuals about one specific set of data

> For example- (Server level agreement) 100% uptime on server against the performance
>
> For example- (KPI) total sales against KPI

Dashboards show multiple graphs about a topic. All graphs work together to illustrate a story to the audience. The audience is given a quick overview of information and provides a way for audience to drill down further for more information and answer their own questions.

Balanced scorecard is a system that looks at different perspectives in order to determine and measure the success. Recognises that the measure of success of organisation is based on a variety of perspectives and measurements

> For example- customer perspective (customer satisfaction), financial perspective, learning and growth perspective

SMART

SMART ensures we have a full understanding of what we need to solve/achieve for the organisation

> Problem- Users that log in to our site for the first time don't return again
>
> (Figure out the goal which solves the problem)
>
> Problem- The goal of this project is to determine the website changes that will most efficiently increase revenues

Specific refers to making sure the goal we set is more detailed rather than general. (In the goal consider the business metric the measurement value for problem at hand and the measurement for determining when the goal is achieved)

> For example- (How increase revenue or what website changes like redesign website or better connection) Addressing the user interface

Measurable refers to the capacity in which the progress towards our goal is measurement this helps us understand if we are making progress

> Measurable- (increase of revenue) Number of sales or ad revenue, or market share

Attainable refers to the practicality of achieving the goal

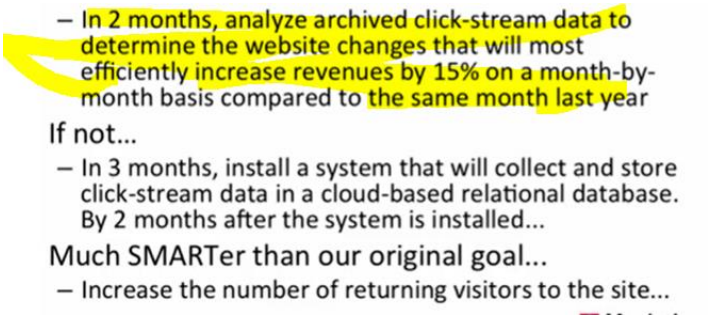
> For example- So whether detecting returning visitor on website is possible

Relevant refers to whether the achieving the goal provides value to the organisation

For example-

Time-bound refers to the time set to achieve the goal and whether it's achievable

For example- (will most efficiently increase revenues) In two months we will see



*NOTE: MORE DESCRIPTION

Determine website changes specifically related to UI that will in three months will improve (sales)

ICT394Ans8

Data mining

Data mining is the process of finding patterns within the large data sets. The patterns identified help explain or identify the behaviour of data given a situation.

For example- (retail) data mining can be used to minimise wastage by predicting the correct amount of inventory level for specific seasons or throughout the year. Data mining can also identify which items have a relationship such as purchased together so they can be position together in the store.

For example- (Customer relationship management) is for management and analyses of customer interactions throughout lifecycle. Data mining uses data collected from interactions such as- sales, inquiries, and demographic to help improve customer retention, maximise marketing campaign effectiveness by using data to see which type of people are susceptible to certain campaigns

For example- (banking) data mining is used to help assist with the automation of processing loan application by making sure defaulters aren't approved. Also identifies fraudulent transactions

For example- (insurance)

The type of patterns in data mining include (patterns are old school way humans identify)-

Association pattern identifies relationships among variables in data set.

For example- in the analysis identifies the units that students take together

Predictions patterns are derived from past information and identifies future event occurrences.

For example- how much revenue will most likely make during Christmas

Sequential relationships patterns identify possible sequence of events occurrence.

For example- a customer will open a cheque account first, then savings account, and finally investment account all within a year

Clusters pattern identifies the items that belong in a group based on features

Data mining tasks (categories)-

Clustering identifies the items that belong in a group based on similar features. The algorithm goes through data set and finds the common features of items and forms a cluster between them. Once clusters are formed, they are analysed by experts in order for the cluster to be considered reasonable. They are used to classify and interpret new data. Cluster analysis ensures each class have common features that best describe them compared to others

Associations identifies relationships among variables in data set. The use of data gathering technology makes it easier to automate and discover relationships

Prediction identifies future events occurrence based off experiences, opinions and other such information. Specific types of prediction include-

Classifications aims to automatically generate a model that point to future occurrences based off analysis of past data stored in the database. It tends to be the most used data mining method and is consider machine learning. The algorithm consists of generalisations over the training data set which helps further define the classes and identify them in unclassified records. For example- target marketing such as a customer is a likely or unlikely customer

The tools to achieve this task include neural networks.

Classification methodology steps-

Model development/training step is where a model is trained based of many data points

Model testing step is where the trained model is used and thus able to be tested on new data points. The output will be compared with the original trained model output.

Classification model assessment criteria-

Predictive accuracy criteria look at the accuracy of a model to be able to label new data points with the right class. The actual correct class labels on test data points are compared with predicted class labels on new data point and an accuracy output is produced based on how accurate it was. This is normally a percentage of cases correctly labelling data point.

Speed criteria measure how much resources are needed to run model

Robustness criteria measures model capacity when faced with noisy data points to make accurate predictions

Scalability criteria measures the ability to create a model efficiently when faced with larger data points

Interpretability criteria measures how insightful the model is

Classification techniques-

Data mining process steps (CRISP-DM)

Data mining process steps (CRISP-DM) stands for cross industry standard process for data mining-

Business/Organisational understanding step is to get a greater understanding of new knowledge needed and to lay out the objectives of the study. This stage involves a budget to support the study and project plan for finding the knowledge and responsible personnel

> For example- What is making customer complain? How to anticipate equipment failure

Data understanding step occurs after we have a strong understanding of the objective of the project. This step identifies the relevant data required for the project in order to answer the specific questions formed in the previous step. This step also considers the location of the data and the way to access and the format the data will be in

> For example- Data in relation to demographic, transaction and sociographic

Data preparation/data pre-processing uses data collected in the previous steps and prepare it for analysis. This takes up 80 percent of the time in data mining projects because the data collected often has issues such as- data being noisy such as containing outliers and incomplete where attributes values are missing, containing aggregated data and not having attributes required. Also, the data may be inconsistent. The four steps to ensure data becomes mineable data includes the following-

> **Data consolidation** step is where the integration of relevant data from sources, the records from multiple data sources and the selected required records occurs
>
> **Data cleaning** step is where data is cleaned this involves removing inconsistencies with the data, fill or ignore missing values and smooth out noisy values such as outliers
>
> **Data transformation** step deals with data being able to be better processed. This may include constructing new attributes or aggregate or normalise data
>
> **Data reduction** is the final step it is reduce the amount of data in the data set. This may include reducing the number of records or columns in file

Modelling/Mode building step is where algorithm identify and output patterns in the data set, they often address the specific business requirement laid out. They are used to both classify and predict future occurrences. The model is built

Testing/Evaluation step is where the model is measured and evaluated in terms of their value and usefulness in achieving businesses objectives. The accuracy and generality often determine this

Deployment step is where the models is deployed and the end user must be able to access and be able to read the mode; May include maintenance because the data collect may no longer be viable for the evolving business activities.

Blunders-

Selecting the wrong problem for data mining

Ignoring what your sponsor thinks data mining is and what it really can/cannot do

Not leaving insufficient time for data acquisition, selection and preparation

Looking only at aggregated results and not at individual records/predictions

Being sloppy about keeping track of the data mining procedure and results

ICT394Ans9

Visualisation is means of showing a story using visuals as opposed to words Allows for patterns to be revealed easily by the visualisation. Also allows for further exploration. The objectives of providing visualisation are the following either for- communication, analysis, monitoring, and planning

Analysis is about using visualisation to identify the conclusions from the data

Communication is about using the visualisation to convey a message

Monitoring is about using the visualisation to compare performance and KPI

Planning is about using visualisation for prediction and forecasting

Simple text visualisation is ideal when message sent is simple such as few numbers rather than using graph. Used for providing emphasis on certain points on the survey can be through font colour and size

Table visualisation is ideal when data needs to be presented and viewed precisely and values of interest in the table are required to be looked up where each member have their own value of interest in a particular row. Table outlining should be faded since emphasis is the data

For example- student results and each student has value of interest in row

Heatmaps visualisation can be thought of as a table but there is a third feature where the values inside the tables are shaded to identify the magnitude or size of values or attract attention to group of values

For example- A table of student with their incomes and shading of the values classifying the

Income as poor, middle class or wealthy

Line graph visualisation is ideal when dealing with continuous data and not ideal categorical data because logically by connecting the values we a suggesting a relationship which could not be true.

Often one axis includes time variable but must be evenly spaced in an interval scale because it would be difficult to draw conclusions if there are time interval missing.

However, if a time interval was missing then the values shouldn't be connected since it is no longer continuous data thus not suitable for line graph

Also the interval must equal in size

For example

Bar charts visualisation is ideal for easy digestion of information since they are well known. Provides us the ability for easy comparison of categorical (discrete) data (x-axis).

Recommended to have a zero baseline for y-axis so that relative height of judging bar chart bars is not deceptive.

Recommended to have y-axis on the left-hand side since information is normally digestion from left to right and placing it on the right means we don't often take it in at first interpretation.

Recommended to have the bars not too wide or else it looks clustered and not to skinny or else difficulty in seeing difference between categories

Recommended to use horizontal bar chart when categorical name is long because with vertical bar chart the names are squashed

Stacked vertical bar charts visualisation is ideal when dealing with comparisons of categories (discrete) data (x-axis) with each category outlining contribution of subcomponents have on the individual category.

Disadvantage is lots of digestion of data needed by the reader and not all

Pie charts visualisation is ideal when dealing with data that can represented as a proportion or percentage.

Not ideal because of the difficulties in comparing the portion since tough to see the magnitude difference between the portion such as how much bigger is portion 1 to portion 2.

Scatterplot visualisation is ideal when dealing with relationship between two continuous variables

> For example- chart of profit and income with line of best fit

Slope graph visualisation is [FILL]

ICT394Ans10

Cognitive load

Cognitive load is the mental effort utilised in order to grasp a new concept. A light cognitive load means it is easy to grasp a new concept because the effort required is less. A heavy cognitive load is when the effort required to grasp a new concept is high. Visualisation in relation to cognitive load is that it provides a way to grasp new concepts by reducing the effort required to understand the concept. It helps create an understandable story explaining the concept

Clutter refers to when a visualisation contains extra components that don't add value to the visualisation. All the extra components do is make visualisation more unnecessarily complex and increase the cognitive load making it harder for new concepts to be grasped.

> Method to reduce clutter could be removal of chart borders, gridlines, axis labels and data markers

Gestalt principles

Gestalt principles of visual perceptions are relevant to us in designing effective visualisations. They can organise content to help make concepts or information easier to grasp. Each of the following gestalt principles can be utilised to make visualisations more effective

Proximity principle is the understanding that objects that are close together are perceived as a group.

> For example- In a chart pie chart if there was three pieces but the labels of each piece were placed outside the piece then we have trouble perceiving the piece with the label making visualisation less effective in allowing audience to understand the information due to the heavy cognitive load. However, if the labels were placed inside pie piece, then the audience ability to pick up the label being associated with the piece is easier.
>
> For example- In a dash board there may be charts with text. By placing he charts closer to their relevant text means it would be easier for the audience to see them as a group. If they were placed further then it would result in the visualisation potentially conveying the wrong message about the graph

Similarity principle is the understanding that objects that are similar size, shape. colour, or orientation are perceived as a group or have a relationship with each other

> For example- In a scatter plot if the x-axis was Age and y-axis was amount of push up and you wanted to show the relationship for Olympians, and non- Olympians. Using the similarity principle such as changing the dot colour for Olympians would make it easier for the audience to see the contrast between Olympians and non-Olympians.
>
> For example- In a dash board, if text of information were same colour blue and the rest were black it was make it easier to see the blue text as a group and thus making it easier to form a relationship.

Enclosure principle is the understanding that objects that are physically enclosed together (such as border) are perceived as a being a group

> For example- in chart or scatter plot draw circle around this is seen as group
>
> For example- In a dashboard there may be charts with text. Placing a border around the two makes it easier for the audience to draw the relationship that those two are a group. This makes visualisation more effective since often viewing the text changes the context of the chart thus makes message less likely to not be grasped by audience.

Connection principle is the understanding that objects that are connected to each other by a line is recognised as a group

Closure principle is the understanding that as long as open structures have the outline of a complete, regular structure it can be perceived as the intended structure

Continuity principle is the understanding that objects that are lined up one after the other are recognised as a group

https://vizzendata.com/2020/07/06/utilizing-gestalt-principles-to-improve-your-data-visualization-design/ ← dash board

https://medium.com/nightingale/how-to-apply-gestalt-psychology-principles-in-data-visualization-6242f4f1a3de <-- charts

Preattentive attributes

Preattentive attributes are object attributes that are seen by the audience without their conscious effort. They enable audience to see information that we want them to see. Visualisations are used to convey specific information and concepts to the audience. Visualisations often have pre-attentive attributes. Manipulating the pre-attentive attributes correctly makes it easier for the audience to grasp the message we are trying to send. The following examples illustrate this.

Form (category)-

Length pre-attentive attribute in bar chart can be altered

> For example- The length attribute in a bar chart is useful in conveying information. An example of a bar chart could be one that had cost as an independent variable and milk brands as dependent variable. The length attribute can help audience identify the cheapest brand of milk clearly. In this case, the smallest bar length clearly represents cheapest brand

Orientation attribute

> For example- The orientation of a bar chart can affect the effectiveness of the visualisation. An example is a bar chart that has lots of texts in the x-axis. Changing the orientation pre-attentive attribute so that the x -axis is vertical will make the longer texts easier to read

Width attribute in visualisation can be altered

> For example- A bar chart with bars that are too skinny are difficult to see. Altering the width of each bar helps with being able to visually separate each bar from one another making visualisations easier to grasp.

Shape attribute in scatter plot can be altered

> ~~For example- A scatter plot chart that had happiness score on y-axis and income on x-axis for USA and China. If the shape of USA and China had the same dot then it would be impossible to see information comparing the two. By altering the shape of one of the dots will make it easier to grasp the information~~.

Size attribute in chart shows degree of difference

Colour (category)-

Hue or colour

> For example- A scatter plot chart that had happiness score on y-axis and income on x-axis for USA and China. If the colon of USA and China had the same hue dot then it would be impossible to see information comparing the two. By altering the hue of one of the dots will make it easier to grasp the information

Intensity combination of lightness and saturation

Spatial Position

> 2-D position attribute in chart shows where data point is up/down and left/right
>
> For example-

https://elias-nordlinder.medium.com/designing-with-your-brain-in-mind-using-pre-attentive-attributes-977a7af54b58#:~:text=Pre%2Dattentive%20attributes%20are%20attributes,are%20even%20aware%20of%20it.

http://daydreamingnumbers.com/blog/preattentive-attributes-example/

ICT394Ans11

For trends over time use line chart. Trend over time conveys how item(s) of interest have changed over define period of time. The time is situated on x-axis (continuous data) and y-axis contains the dependent variable (measure in response to changes with x-axis). The item(s) are distinguished with legends.

Improvements/What's wrong could be having multiple items makes it difficult to see

the trend of overall item/category. So instead sum all the items/categories and have one line

Improvements/What's wrong only have one line and we can't see the contribution of the different item(s)/categories to the trend. So instead use a legend to distinguish each item(s)/categories

Improvement/What's wrong can't see the contribution of different item(s)/categories to the

trend and trend of overall item/category. So instead use a line chart with total

item(s)/category line added OR use a stacked bar chart with each bar showing the contribution of each item OR area chart

Improvements/What's wrong there is an irregular interval meaning a particular time has no data. So, we can't join or connect line. Remove connection

For ranking use bar chart. The x-axis contains the (categorical data) and associated dependent variable (measurement). Ranking is appropriate for bar chart because the audience can clearly see how each bar is different based on the difference of heights with the bar

Improvements/What's wrong has no zero baseline for y-axis (fox news) …

Improvements/what's wrong the bar chart is vertical. x-axis categories (categories) have too much text thus is squished. Change orientation so x-axis is horizontal so implement horizontal bar chart instead. Ensuring the chart is correctly oriented because correct orientation makes the chart easier for audience to digest and cognitive load required is less.

For part to whole use stacked bar chart or pie chart. Deals with the contribution of each item(s) to a whole. A bar will show contribution of each item to the bar.

> Improvement/What's hard to measure or see difference with area or piece of pie in comparison with other pieces. Implement a bar chart since it is easier to see difference with height of bar rather than difference in area
>
> Improvement/What's hard to see the size of each piece. Implement value labels

For correlation use scatterplot. The deals with relationship of two (continuous data/measures) variables or measures.

> Improvement/What's wrong difficult to see trend between two variables. Implement trend line

For distribution use box and whisker or histogram. This deals with how values are disturbed across a define range.

For geographical data use maps paired with another chart. Deals with data associated with geographical data thus a map showing the data at geographical location is effective in highlighting it.

Good/Bad Visualisation Rules

Best visualisation rules- (look here for poor visualisation)

Chart orientation ensuring the chart is correctly oriented because correct orientation makes the chart easier for audience to digest and cognitive load required is less.

> Improvements/what's wrong the bar chart is vertical. x-axis categories (categories) have too much text thus is squished. Change orientation so x-axis is horizontal so implement horizontal bar chart instead

Ensure graphs are not overloaded with information because contains too much irrelevant information that don't help illustrate the story, we are trying to convey

> For example-

Ensure graph does not have too much colours and shape since audience may find it difficult to digest and process. Max is 7-10

Highlight most important data meaning ensuring that x-axis and y-axis illustrates what the information we want to show. Also picking right color, size and shape of important data.

> For example- for the x-asis the size of home is shown instead of lot

Dashboard

Dashboards show multiple graphs about a topic. All graphs work together to illustrate a story to the audience. The audience is given a quick overview of information and provides a way for audience to drill down further for more information and answer their own questions.

Best dashboard rules- (look here for poor visualisation)

Ensure a single dashboard has only three of four graphs to avoid audience missing the story we want to illustrate or overloading their cognitive load

The **most important graphs are located at** top or top left since people read top left to right and end up bottom right

Ensure dashboard graphs have consistent colour schemes because there is a less chance a message is misinterpreted

> For example- if one bar chart has category: dog as brown colour bar then the line graph with dog as a line should also be brown
>
> For example- red green colour scheme is bad for red green colour blindness

Ensure charts are on a single screen the reason why is if a audience is required to click to view another set of charts he tends remember chunks of information at one time so the charts story in the other view may lost when hopping into new chart view. Does not provide consolidated view

> For example- store manager wants to see the sales of each item. If he is required to click to view each items chart(s) it is useless in that it doesn't provide overview of all items

Ensure charts don't require scrolling to see other charts because many view whatever is at the bottom less important and won't bother to scroll

Providing too much information in chart that is irrelevant

> For example- chart has date, time, year and whether leap year

Ensuring chart is appropriate in dashboard

Ensuring dashboard doesn't have useless decorations

> For example**-** fill space with pictures

Ensure legends are next to filters

Colours

> Work together without clashing
>
> Less than 7-10

Fonts

> Consistent use throughout

No more than 3 different ones on a dashboard

Labels

Clear, concise

Placement

Levelling

Tooltips

Useful?

Evaluation questions for judging whether dashboard/visualisation is good-

Are your graphs effective?

Do you have the right chart type for your analysis?

What questions are you trying to answer?

Is your dashboard holistic?

Are there other things you could do to polish your work up?

ICT394Ans9&11

Visualisation is means of showing a story using visuals as opposed to words Allows for patterns to be revealed easily by the visualisation. Also allows for further exploration. The objectives of providing visualisation are the following either for- communication, analysis, monitoring, and planning

Analysis is about using visualisation to identify the conclusions from the data

Communication is about using the visualisation to convey a message

Monitoring is about using the visualisation to compare performance and KPI

Planning is about using visualisation for prediction and forecasting

Good/Bad Visualisation Rules

Best visualisation rules- (look here for poor visualisation)

Chart orientation ensuring the chart is correctly oriented because correct orientation makes the chart easier for audience to digest and cognitive load required is less.

> Improvements/what's wrong the bar chart is vertical. x-axis categories (categories) have too much text thus is squished. Change orientation so x-axis is horizontal so implement horizontal bar chart instead

Ensure graphs are not overloaded with information because contains too much irrelevant information that don't help illustrate the story, we are trying to convey. **Clutter** refers to when a visualisation contains extra components that don't add value to the visualisation. All the extra components do is make visualisation more unnecessarily complex and increase the cognitive load making it harder for new concepts to be grasped.

> Method to reduce clutter could be removal of chart borders, gridlines, axis labels and data markers

Ensure graph does not have too much colours and shape since audience may find it difficult to digest and process. Max is 7-10

Highlight most important data meaning ensuring that x-axis and y-axis illustrates what the information we want to show. Also picking right color, size and shape of important data.

> For example- for the x-asis the size of home is shown instead of lot

Simple text visualisation is ideal when message sent is simple such as few numbers rather than using graph. Used for providing emphasis on certain points on the survey can be through font colour and size

Table visualisation is ideal when data needs to be presented and viewed precisely and values of interest in the table are required to be looked up where each member have their own value of interest in a particular row. Table outlining should be faded since emphasis is the data

> For example- student results and each student has value of interest in row

Heatmaps visualisation can be thought of as a table but there is a third feature where the values inside the tables are shaded to identify the magnitude or size of values or attract attention to group of values

> For example- A table of student with their incomes and shading of the values classifying the
>
> Income as poor, middle class or wealthy

Line graph visualisation is ideal when dealing with continuous data and not ideal categorical data because logically by connecting the values we a suggesting a relationship which could not be true.

For trends over time use line chart. Trend over time conveys how item(s) of interest have changed over define period of time. The time is situated on x-axis (continuous data) and y-axis contains the dependent variable (measure in response to changes with x-axis). The item(s) are distinguished with legends.

> Improvements/What's wrong could be having multiple items makes it difficult to see the trend of overall item/category. So instead sum all the items/categories and have one line
>
> Improvements/What's wrong only have one line and we can't see the contribution of the different item(s)/categories to the trend. So instead use a legend to distinguish each item(s)/categories
>
> Improvement/What's wrong can't see the contribution of different item(s)/categories to the trend and trend of overall item/category. So instead use a line chart with total item(s)/category line added
>
> OR use a stacked bar chart with each bar showing the contribution of each item
>
> OR area chart
>
> Improvements/What's wrong there is an irregular interval meaning a particular time has no data. So, we can't join or connect line. Remove connection. However, if a time interval was missing then the values shouldn't be connected since it is no longer continuous data thus not suitable for line graph
>
> Improvements/What's wrong interval are not equal in size

Bar charts visualisation is ideal for easy digestion of information since they are well known. Provides us the ability for easy comparison of categorical (discrete) data (x-axis).

For ranking use bar chart. The x-axis contains the (categorical data) and associated dependent variable (measurement). Ranking is appropriate for bar chart because the audience can clearly see how each bar is different based on the difference of heights with the bar

> Improvements/what's wrong the bar chart is vertical. x-axis categories (categories) have too much text thus is squished. Change orientation so x-axis is horizontal so implement horizontal bar chart instead. Ensuring the chart is correctly oriented because correct orientation makes the chart easier for audience to digest and cognitive load required is less.
>
> Improvements/What's wrong has no zero baseline for y-axis so relative height of judging bar chart bars is deceptive.
>
> Improvement/What's wrong bars is too wide which makes it look clustered and too skinny which makes it difficult in seeing difference between categories

For single point in time

Stacked vertical bar charts visualisation is ideal when dealing with comparisons of categories (discrete) data (x-axis) with each category outlining contribution of subcomponents have on the individual category.

For part to whole use stacked bar chart or pie chart. Deals with the contribution of each item(s) to a whole. A bar will show contribution of each item to the bar.

> Improvement/What's wrong hard to measure or see difference with area or piece of pie in comparison with other pieces. Implement a bar chart since it is easier to see difference with height of bar rather than difference in area
>
> Improvement/What's wrong is lots of digestion of data needed by the reader

Pie charts visualisation is ideal when dealing with data that can represented as a proportion or percentage.

For part to whole use stacked bar chart or pie chart. Deals with the contribution of each item(s) to a whole. Shows contribution of each piece to the overall pie

> Improvement/What's wrong hard to measure or see difference with area or piece of pie in comparison with other pieces. Implement a bar chart since it is easier to see difference with height of bar rather than difference in area
>
> Improvement/What's wrong hard to see the size of each piece. Implement value labels

Scatterplot visualisation is ideal when dealing with relationship between two continuous variables.

For correlation use scatterplot. The deals with relationship of two (continuous data/measures) variables or measures.

> Improvement/What's wrong difficult to see trend between two variables. Implement trend line
>
> For example- chart of profit and income with line of best fit

Slope graph visualisation is [FILL]

Dashboard

Dashboards show multiple graphs about a topic. All graphs work together to illustrate a story to the audience. The audience is given a quick overview of information and provides a way for audience to drill down further for more information and answer their own questions.

Best dashboard rules- (look here for poor visualisation)-

Ensure a single dashboard has only three of four graphs to avoid audience missing the story we want to illustrate or overloading their cognitive load

The **most important graphs are located at** top or top left since people read top left to right and end up bottom right

Ensure dashboard graphs have consistent colour schemes because there is a less chance a message is misinterpreted

> For example- if one bar chart has category: dog as brown colour bar then the line graph with dog as a line should also be brown
>
> For example- red green colour scheme is bad for red green colour blindness

Ensure charts are on a single screen the reason why is if a audience is required to click to view another set of charts he tends remember chunks of information at one time so the charts story in the other view may lost when hopping into new chart view. Does not provide consolidated view

> For example- store manager wants to see the sales of each item. If he is required to click to view each items chart(s) it is useless in that it doesn't provide overview of all items

Ensure charts don't require scrolling to see other charts because many view whatever is at the bottom less important and won't bother to scroll

Providing too much information in chart that is irrelevant

> For example- chart has date, time, year and whether leap year

Ensuring chart is appropriate in dashboard

Ensuring dashboard doesn't have useless decorations

> For example- fill space with pictures

Ensure legends are next to filters

Colours

> Work together without clashing
>
> Less than 7-10

Fonts

> Consistent use throughout
>
> No more than 3 different ones on a dashboard

Labels

> Clear, concise
>
> Placement
>
> Levelling

Tooltips

Useful?

Evaluation questions for judging whether dashboard/visualisation is good-

Are your graphs effective?

Do you have the right chart type for your analysis?

What questions are you trying to answer?

Is your dashboard holistic?

Are there other things you could do to polish your work up?

Timetable

Timetable-

Sunday (11,10,9)

~~Mon~~

Tuesday ~~(8,7)~~

Wednesday ~~(6,5)~~

Thursday ~~(4,3)~~

Friday ~~(2,1)~~

Saturday (8,7)

Sunday (6,5)

Mon (4,3)

Tuesday ~~(10)~~

Wednesday

Sunday ~~(4, 5)~~

Mon (6, 7)

Tuesday (8)

Wed

BI implementation

Tesco (Topic 1)

University of Konstanz BI (Topic 2)

Royal Liverpool hospital BI (Topic 5)

Allrecipes case study (Topic 7)