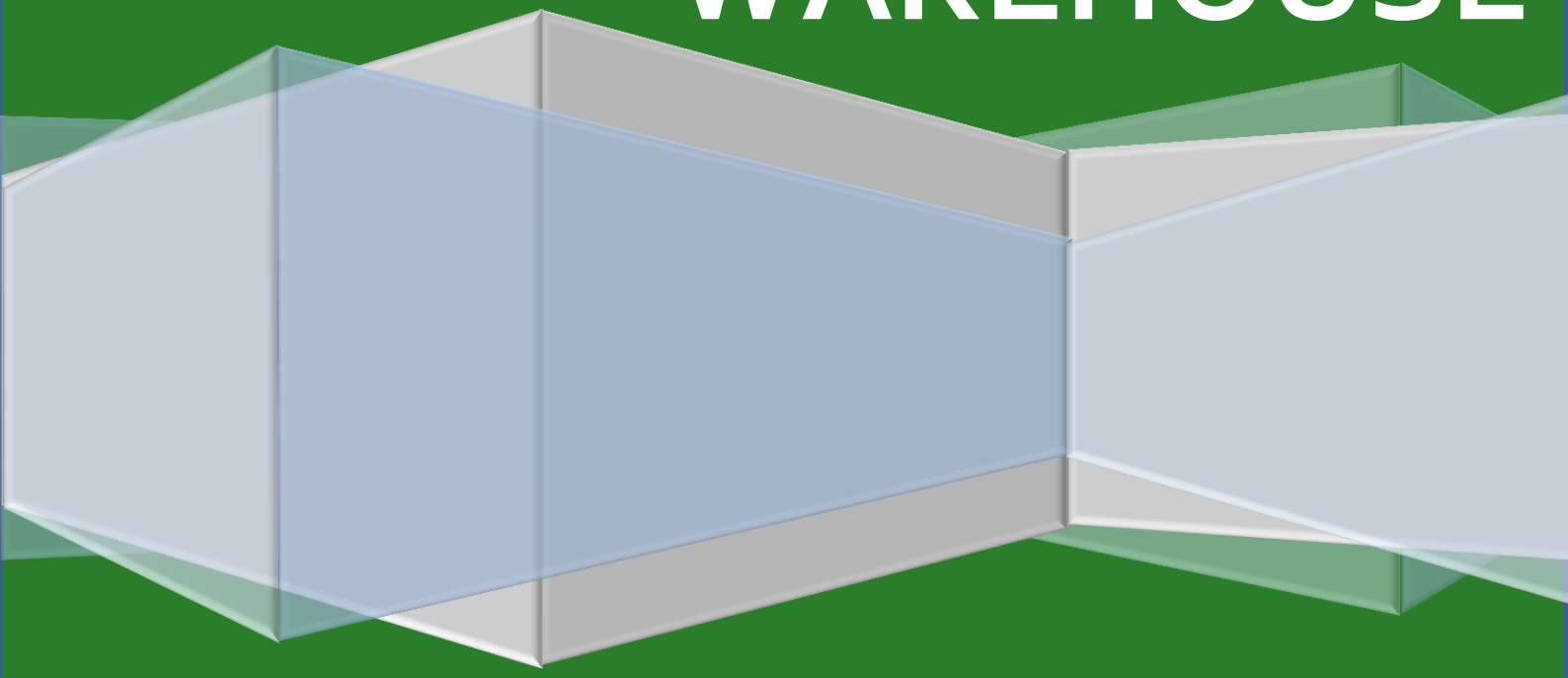
OIC ACCREDITATION CERTIFICATION PROGRAMME FOR OFFICIAL STATISTICS

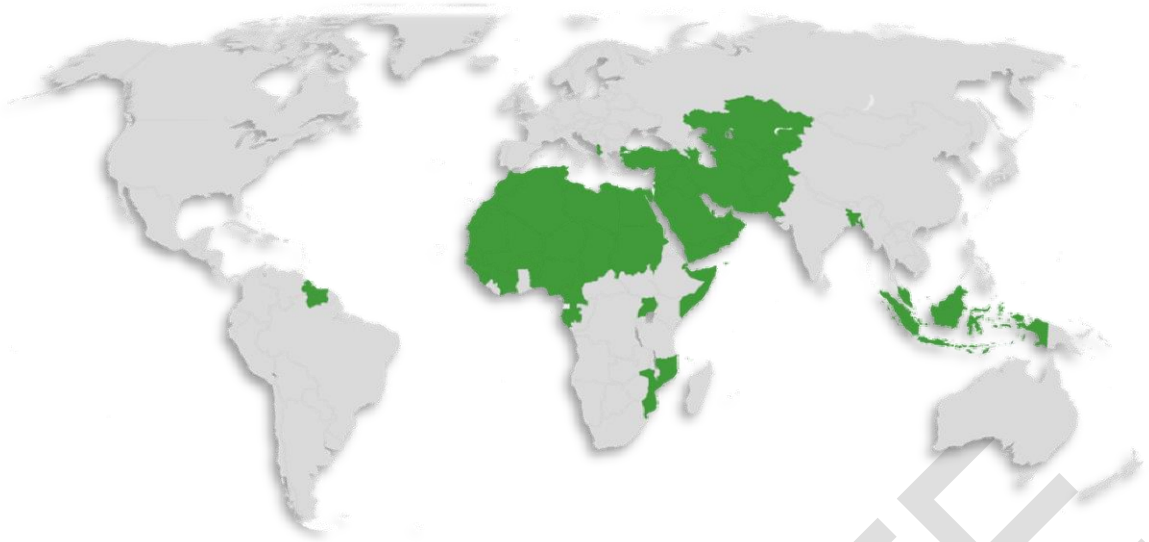
INTRODUCTION TO DISSEMINATION AND DATA WAREHOUSE



ORGANISATION OF ISLAMIC COOPERATION

STATISTICAL ECONOMIC AND SOCIAL RESEARCH
AND TRAINING CENTRE FOR ISLAMIC COUNTRIES

OIC ACCREDITATION CERTIFICATION PROGRAMME FOR OFFICIAL STATISTICS



INTRODUCTION TO DISSEMINATION AND DATA WAREHOUSE

TEXTBOOK

{{AHMED UDDIN, KABIR}}

ORGANISATION OF ISLAMIC COOPERATION

STATISTICAL ECONOMIC AND SOCIAL RESEARCH AND TRAINING CENTRE FOR ISLAMIC COUNTRIES

Kudüs Cad. No: 9, Diplomatik Site, 06450 Oran, Ankara – Turkey
Telephone +90 – 312 – 468 6172
Internet www.sesric.org
E-mail statistics@sesric.org

ISBN: xxx-xxx-xxxx-xx-x

Cover design by Publication Department, SESRIC.

For additional information, contact Statistics Department, SESRIC.

CONTENTS

UNIT 1: Data dissemination

1.1 Introduction

Simply speaking, data dissemination is the release to users of information obtained through a statistical activity. It consists of distributing or transmitting statistical data to users. Various release media are possible; for example: electronic format including the internet, CD-ROM, paper publications, files available to authorized users or for public use; fax response to a special request, public speeches, press releases.

Wikipedia defines data dissemination as the distribution or transmitting of statistical, or other, data to end users. There are many ways organizations can release data to the public, i.e. electronic format, CD-ROM and paper publications such as PDF files based on aggregated data.

Under the Special Data Dissemination Standard (SDDS), the dissemination formats are divided into two categories: (i) Hardcopy publications and (ii) Electronic publications

Some examples of Hardcopy publications:

- Yearbook
- Quarterly report
- Monthly review
- Trends
- Pocketbook
- Periodical

Some examples of electronic copy publications:

- CD Rom
- Webpage
- PDF
- Downloadable Databases for private use in 3rd party software applications

1.2 Data, microdata and metadata

Data:

Data are facts and figures which relay something specific, but which are not organized in any way and which provide no further information regarding patterns, context, etc. In general, data is unprocessed facts and figures without any added interpretation or analysis. "The price of crude oil is $50 per barrel."

Microdata:

Microdata consist of the data directly observed or collected from a specific unit of observation. That is, a microdata file contains organized raw data wherein the lines represent a specific unit of measure (usually an individual, household or family) and the information about the lines are the values of variables.

Usually survey collects information from each unit of observation (e.g., individual, household, etc.). It processes these answers by coding them using a specific number to identify the respondent's answer. For example, we often use a "1" to represent males and a "2" to represent females. The microdata file is created by coding and electronically recording each survey respondent's responses to all relevant questions.

A microdata file consists of rows of numbers and letters– each row represents the respondent's responses to the questionnaire. It also consists of one logical record per respondent, where the logical record includes all responses made by a single respondent to the questionnaire.

Microdata allow researchers to use any variable in the file for analysis. With microdata files, researchers can analyses any variable in the file, and can construct the tables they need, rather than choosing from the pre-tabulated information presented in an aggregated file.

To make the microdata anonymous variables may be collapsed (e.g., age groups instead of individual years of age); collapse variables into one variable (e.g., multiple language questions collapsed into one language variable for analysis); suppressing variables; and removing outliers (removing cases that are extremes - often used with income). By using these techniques to anonymise the files, combining variables will not result in the user identifying a respondent.

Metadata:

The term metadata defines all information used to describe other data. A very short definition of metadata then is "*data about data*". Metadata descriptions go beyond the pure form and contents of data. They are used to describe administrative facts about data (who creates them, and when), how data were collected and processed before they were

disseminated or stored in a database. In addition, metadata facilitate efficient searching and locating of data. It is needed by people or systems to make proper and correct use of the real statistical data, in terms of capturing, reading, processing, interpreting, analyzing and presenting the information (or any other use). The information usually includes in metadata are the definition of variables and description of their classification schemes, the description of the methodology used in collecting, processing and analyzing the data, and information on the accuracy of the data.

Metadata can consist of many different documents including survey questionnaires, instructions to interviewers, codebook, user's guide, record layout, data dictionary, frequency file, cv tables, etc. It may be noted here that codebooks, record layouts, user guides and data dictionaries have overlapping properties.

"The Definitions, data sources and methods" of Statistics Canada includes the following items viz. status, frequency, questionnaire and reporting guide, description, data sources, methodology, data accuracy, target population, instrument design, sampling, error, imputation, estimation, quality evaluation, and disclosure control.

- **Questionnaire**: This tool is helpful to assess the questions posed to the respondent and how the questions were formulated.
- **Interviewer instructions**: Interviewer instructions give an indication of how the data was collected and also provides an indication of skip patterns in the questionnaire (which helps explain why the population for certain variables may be lower than the total population).
- **User's guide**: The user's guide contains information to help the user interpret the survey data. It has overlapping properties with the data dictionary, record layout and codebook as it often contains all the documentation pertaining to a survey (such as the sampling methodology, population sampled, variable descriptions, position, labels, etc.).
- **Codebook**: A codebook is a generic term often used to describe the user's guide, record layout and data dictionary or combinations of these documents. In its earliest usage, the codebook contained the rules for assigning numeric codes to the responses for questionnaire items. However, it typically provides variable-specific metadata - question text, response values, missing value declarations, variable universe, etc.
- **Record layout**: The record layout provides variable names, column positions in the data file, and number of decimals. It is often distributed in .xls format - and hence, can be exported to ASCII and used to create SPSS/SAS/Stata command files.
- **Data dictionary**: The data dictionary is an excellent source to find general information about the variables in a survey, the codes for variables, missing value assignments, and frequency counts. This document has overlapping properties with the codebook, user's guide and record layout.

1.3 Quality dimension of microdata and metadata

Best practice and standards in microdata and metadata quality dimension are — documentation, cataloging, anonymization and dissemination, and preservation.

- **Documentation**. Compliance with international metadata standards is crucial for ensuring exchangeability of metadata and for promoting collaboration or coordination in the development of microdata curation tools. The standard most commonly used for the documentation of microdata is the Data Documentation Initiative (DDI) by the DDI Alliance. The standard is used by most of the large social science data archives around the world, by many national statistical agencies, and by international organizations (including FAO, ILO, UNICEF, WFP, WHO, and the World Bank). A number of free tools exist for creating DDI metadata, such as a free DDI editor developed by NESSTAR Ltd and supported by the IHSN. The DDI standard complements the SDMX standard. The (long) process of gaining ISO certification for the standard has recently been initiated. In the meantime, adopting the DDI as a UN recommended standard for microdata documentation would bring the UN institutions in line with current best practice. Ideally, the DDI standard should be complemented by the adoption of a common, multilingual taxonomy of topics.
- **Cataloguing**. Publishing detailed metadata in on-line searchable catalogs is important to make data discoverable. Compliance with the DDI standard makes it considerably easier. Open source DDI compliant cataloging applications already exist such as the open source National Data Archive (NADA) developed by the IHSN or DataVerse developed by Harvard University.
- **Anonymization**. This is an area where academically well-grounded methods exist. Tools exist for the measurement and reduction of disclosure risk, such as the open source, R-based sdcMicro application (Templ et al, 2012) or µArgus (a freeware, soon to be open source). But no international standards exist for their implementation. Institutional and national practices are typically not documented or shared. The reasons for this are that the methods used for anonymizing microdata are very contextual to the type of data being anonymized, and that disclosing detailed information on the methods may provide useful information to those trying to defeat the protections. Eurostat publishes broad information on their anonymization methods on their website. Such transparency is good practice as it makes researchers aware of the content and limitations of the microdata. Useful but somewhat outdated information can also be found in the Report on Statistical Disclosure Limitation Methodology by the US Federal Committee on Statistical Methodology (2005). An international

review of disclosure risk management practices by statistical agencies, research centers and other data repositories would be tremendously useful.

- **Dissemination**. Cognizance needs to be given to the fact that not all data are the same. Dissemination policies need to be developed that are clear but also flexible enough to cover the full range of issues such as data ownership, legal and ethical responsibilities and sensitivity of data. Some data are more sensitive than others, and as such dissemination policies must provide for multiple access policies to accommodate various types of datasets. Typically, five levels of accessibility are considered: open access (no restriction), direct access or Public Use Files (some restrictions on use, but no screening of users), Research Use Files (or Scientific Use Files, or Licensed Files), availability only in an enclave, and no access authorized.
- **Preservation**. Organizations which disseminate microdata and the related metadata are also often responsible for their long term preservation. Preserving digital content is not a trivial exercise. Procedures and infrastructures must be put in place to protect data against hardware and software obsolescence (regular migration of datasets to new media and formats), system failures, human errors and other hazards. The IHSN Working Paper on Principles and Good Practice for Preserving Data by the Interuniversity Consortium for Political and Social Research (2009), provides guidelines and multiple references to recommended practices and standards such as the Open Archival Information System (OAIS).

1.4 Key Issues of microdata dissemination policy

Many national and other agencies refer to the United Nations Fundamental Principles of Official Statistics when setting up their data dissemination policies. Best practices should be compatible with these principles. The sixth principle governing International Statistical Activities states that "Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes." The strict confidentiality is often invoked as a reason not to share any microdata. In 2006, a task force set up by the Conference of European Statisticians assessed the implications of the principle and produced Guidelines and Core Principles of Confidentiality and Microdata Access (UNECE, 2007) in which they suggested that:

- "It is appropriate for microdata collected for official statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected. (...) Making available microdata for research is not in

contradiction with the sixth UN Fundamental Principle as long as it is not possible to identify data referring to an individual."

- Microdata should only be made available for statistical purposes: The aim must be to derive statistics that refer to a group (of persons or legal entities), not to specific individuals. • Provision of microdata should be consistent with legal and other necessary arrangements that ensure that confidentiality of the released microdata is protected.
- The procedures for researcher access to microdata should be transparent, and publicly available. This is important "to increase public confidence that microdata are being used appropriately and to show that decisions about microdata release are taken on an objective basis."

1.5 Principles related to microdata dissemination

The Organization for Economic Co-operation and Development (OECD) defined a set of core principles related to microdata dissemination in their Principles and Guidelines for Access to Research Data from Public Funding (OECD, 2007). The list of principles below is adapted from these guidelines:

- **Openness**: Openness should not be understood as "unrestricted access" or "open data"; it means that access must be provided on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination.
- **Transparency**: Detailed metadata must be provided, and the specifications of conditions attached to the use of the data should be internationally available in a transparent way, ideally through the Internet. UNICEF, WHO and the World Bank disseminate microdata as Public Use Files or Licensed Files, accessible to registered users. In all cases, the terms of use associated with microdata include obligations and restrictions (e.g., prohibiting attempts to re-identify respondents or selling or transferring the data to others).
- **Legal conformity and protection of privacy**: National laws and international agreements, as they pertain to the protection of privacy, directly affect data access and sharing practices. These must be taken into account in the formulation of data access arrangements. Microdata are typically obtained from households, individuals, firms or facilities against a commitment to keep the data confidential. Unless formal consent has been provided by the respondent (which is rarely the case) data can only be disseminated after being properly treated to ensure the risk of disclosure is minimal, i.e. the data are anonymized.

- **Protection of intellectual property**: Data access arrangements must consider the applicability of copyright or of other intellectual property laws that may be relevant. Note: Microdata obtained (and in some cases published) by international organizations are often produced through data collection activities implemented and funded by multiple national and international partners. The ownership of the resulting data is not always clearly identified. The rights or obligations to disseminate microdata should be explicitly defined in funding agreements and contracts. Good practice and models are available.
- **Interoperability**: Technological and semantic interoperability is a key consideration in enabling and promoting international and interdisciplinary access to and use of research data. Access arrangements, should pay due attention to the relevant international data documentation standards. Note: The DDI metadata standard is the standard adopted by most specialized microdata libraries and many statistical agencies. It was also adopted by several international organizations.
- **Quality**: The value and utility of data depends, to a large extent, on the quality of the data itself. Data managers, and data collection organizations, should pay particular attention to ensuring compliance with explicit quality standards.
- **Security**: Specific attention should be devoted to supporting the use of techniques and instruments to guarantee the integrity and security of data.
- **Accountability**: The performance of data access arrangements should be subject to periodic evaluation by user groups, responsible institutions and funding agencies.

1.6 Electronic dissemination (Internet) and statistical portals

Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens entitlement to public information. Statistics make it easier for people, politicians and the business community to stay informed and to make informed decisions. Official

statistics must be presented in such a way that the main results can be understood without expert knowledge of statistics, coherent and easily accessible.

NSOs are interested in information who is visiting their websites, what visitors are looking for, what problems they have with obtaining information and what suggestions they have for improving the websites. Most NSOs use feedback to determine new statistics, type of products and the mode of dissemination. NSOs communicate with different types of users. Users differ in their attitudes (statistical literacy), efforts, time and money to give away to obtain statistics. The general public appeared as an important user group that influenced a range of new products and services (storytelling, writing for web, metadata databases, and data visualization tools).

Typical products on the web:
Typical products on the web are electronic versions of paper documents, electronic-only documents, tables, databases of aggregated data and micro data, metadata and knowledge databases, spreadsheets, static and animated graphs and maps and podcasts. NSOs present data and metadata, data release dates, business information for reporting units and users, school curriculum materials for teachers and students, research and information papers and podcasts promoting statistical literacy and access to micro data for researchers through the web.

1.7 Communicating with the media and press

Organizations and individuals recognize the importance of using statistical findings to make evidence-based decisions. Therefore, it is critical that the statistical organization communicates effectively with the media to achieve three important dissemination objectives:

- To inform the general public about the latest releases of official statistics and reports on the social, economic and general conditions of the country.
- To demonstrate the relevance of statistical information to both the general public and to public and private-sector organizations and businesses to inform decision-making throughout society more effectively.
- To increase public awareness of and support for statistical programmes and services.

Dissemination of statistical information to the media is based on the same core principles that underlie the general dissemination activities of the NSO.

- **Relevance**: The information should be relevant to the social, economic and general conditions of the country and meet the needs of both public and private decision makers. For the media, relevance translates into newsworthiness. However, the statistical organization must be careful to present information in a way that does not trivialize the data or findings. The goal is to inform citizens about the availability of the data or information. Media coverage is desirable because it enlarges the audience for the message and will increase knowledge and stimulate debate among the broader public.
- **Confidentiality**: The NSO must protect the confidentiality of individual respondents, whether persons or businesses, for all data collected. The organization should not release any information that identifies an individual or group without prior consent. Nor must the organization reveal information that undermines the confidentiality of its respondents. This restriction applies to the media the same way it does to any other customer of the organization.
- **Independence and objectivity:** Information should be presented in an objective and impartial manner, and be independent of political control or influence. The Fundamental Principles of Official Statistics4 set criteria by which independence and objectivity can be judged.
- **Timeliness**: Information should be current and released as soon as possible after the reference period. The timeliness of information will influence its relevance.
- **Accessibility and clarity**: In principle, all users should have equal access to data as well as to metadata. Information should be publicly available in appropriate formats through appropriate delivery channels, and be written in plain and understandable language adapted to the level of understanding of the main user groups. The statistical organization should ensure that the media, like other clients, are able to access and correctly interpret information on statistical methods, concepts, variables and classifications used in producing statistical results.
- **Coherence**: The use of standard concepts, classifications and target populations promotes coherence and credibility of statistical information, as does the use of common methodology across surveys.

Adherence to these core dissemination principles will enhance the credibility of the NSO and build public trust in the reliability of its information.

UNIT 2. Data warehouse

2.1 Introduction

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but can include data from other sources. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources. Businesses have collected operational data for years, and continue to accumulate ever-larger amounts of data at ever-increasing rates as transaction databases become more powerful, communication networks grow, and the flow of commerce expands. Data warehouses collect, consolidate, organize, and summarize this data so it can be used for business decisions.

Different people have different definitions for a data warehouse. The most popular definition came from Bill Inmon, who provided the following: A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

- **Subject-Oriented**: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.
- **Integrated**: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.
- **Time-Variant**: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.
- **Non-volatile**: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

Data warehouses reside on servers dedicated to this function running a database management system (DBMS) such as SQL Server and using Extract, Transform, and Load (ETL) software such as SQL Server Integration Services (SSIS) to pull data from the source systems and into the data warehouse.

2.2 Key characteristics of a data warehouse

The key characteristics of a data warehouse are as follows:

- Data is structured for simplicity of access and high-speed query performance.
- End users are time-sensitive and desire speed-of-thought response times.
- Large amounts of historical data are used.
- Queries often retrieve large amounts of data, perhaps many thousands of rows.
- Both predefined and ad hoc queries are common.
- The data load involves multiple sources and transformations

2.3 Importance of data warehouse

There are many reasons why you would want to use a data warehouse:

- ❖ You need to integrate many different sources of data in near real-time. This will allow for better business decisions because users will have access to more data. Plus this will save users lots of time because they won't waste precious time retrieving data from multiple sources
- ❖ You have tons of historical data that you need to gather in one easily accessible place in which it will have common formats, common keys, common data model, and common access methods.
- ❖ You need to keep historical records, even if the source transaction systems does not
- ❖ You can restructure the data and rename tables and fields (i.e. numbers without decimal points) so if makes more sense to the users.
- ❖ You need to use master data management to consolidate many tables, such as customers, into one table
- ❖ Users are running reports directly against operational systems, causing performance problems. Instead, create a data warehouse so users can run reports off of that. Plus, the data warehouse is optimized for read access, resulting in faster report generation
- ❖ The data warehouse can be housed on a server built specifically for a data warehouse, resulting is much quicker access than hardware designed for handling transactions
- ❖ There is a risk that BI users might misuse or corrupt the transaction data
- ❖ Having an easy to use data warehouse allows users to create their own reports without having to get IT involved.
- ❖ A data warehouse is a convenient place to create and store metadata
- ❖ Improve data quality by cleaning up data as it is imported into the data warehouse (providing more accurate data) as well as providing consistent codes and descriptions

2.4 Differences between data warehouses and operational databases

Operational databases and data warehouses are mostly based on the same technological support: both are data collections, both function based on keys, indexes and views, both are based to a data model. Nevertheless, the two systems are different, as the criteria described below shows.

1) From a functional point of view: operational databases process transactions, providing answers to operational requirements, while data warehouses are used based on adhoc queries, mainly for management purposes.
2) Functional requirements are different: operational databases mainly focus on data security and coherence, which makes queries slow, special ad-hoc, mainly in the case of unpredicted criteria, while in data warehouses is usually. These queries, specific to economic analysis, may significantly compromise the performance of the operational system, due to the lack of predictable indexes, as is the case of data warehouses.
3) Although most operational systems and data warehouses are built on relational technologies, their design is substantially different, as their purpose is also different. Operational databases are designed for online transaction processing and their main objective refers to the efficient storing of a large amount of transactional data. They include current information on day-to-day activities and process-oriented information which is subject to updating. As a result, data is dynamic and thus, very volatile. The tasks of such systems are structured and repetitive and are made up of current, short and isolated transactions, which include detailed data. These transactions read or update few recordings – tens at most, mainly accessed based on their primary keys. Operational databases reach sizes from hundreds of megabytes to gigabytes. Their consistency is essential and refers to rapid transaction processing. As opposed to transactional databases, data warehouses are designed to be the support of decision-making systems. They are designed to facilitate data analysis, not efficient storing, and the only operations performed refer to the initial data loading, data access and its refreshment. As data is static and non-volatile, the size of data warehouses may reach in time hundreds of gigabytes, terabytes or even petabytes. Many ad-hoc queries may be made and millions of records may be accessed, as many joins and aggregations may be performed. Information is subject-oriented, as data warehouses provide a multidimensional view on data, based on an intuitive model, designed to meet the requirements of data analysis and decision makers.

4) Another difference refers to the status shown by data. Data warehouses show the status of data at different moments in time, thus providing historical outlook. This is different from operational databases, where data shows the current status at the time of access.
5) Optimisation is an issue for both systems, but in a different way. While operational databases are designed to provide data processing optimisation and security, data warehouses optimise analyses and the economic significance of data. Operational databases can be said to be optimised for writing, while data warehouses are optimised for reading. To this purpose, multidimensional modelling is used for the design of data warehouses to make queries for the analysis and summary of large amounts of data more efficient. The structure of a data warehouse is simple, intuitive and easy to understand by non-expert users, as opposed to the structure of an operational database, which is designed based on the entity-relationship model, by specific techniques which are complex and difficult to understand. On the contrary, multidimensional modelling involves the denormalisation of tables. This enters controlled data redundancy, allows analyses from different points of view and different levels of detail.
6) Categories of users are different. Operational databases are meant for a large number of users, from different categories. Automated processes are repetitive, as processing requirements are known before the initial development. The system should immediately provide answers to any query or to any new transaction. As opposed to these systems, data warehouses are used by a small number of users, namely by managers and business analysts. Processes are heuristic, as requirements are not completely known before the initial development. Response requirements are more lax compared to operational systems. Depending on the complexity of processing requirements, response times ranging from several seconds to days are allowed.
7) Data integrity is seen differently. Integrity constraints are established for the verification of input data in operational databases. These constraints are not necessary in the case of data warehouses, as data has been verified and filtered before loading, and historical data will not be updated following its loading in the data warehouse. In operational databases, a transaction must lead the data collection from one consistent status to another consistent one. This involves complex mechanisms for data integrity maintenance systems: data logs, data restore, detection of blockings, backup and recovery. These mechanisms are useless in the case of data warehouses. Treating of updating anomalies are not as important as in the case of transactional systems, as data warehouses are specialised and optimized for the fast retrieval of large volumes of data, and updating refers only to the regular add of new data.

8) Another difference between the two types of systems refers to the mechanisms required for users' concurrent access. As data warehouses are not updated, transaction management, concurrent access management and other such mechanisms integrated in the database management system are used only in the initial loading stage and for subsequent add, due to the fact that they are expensive from the point of view of response time. These mechanisms may be disabled during the current use of data warehouses. The freedom thus generated may be employed for the optimization of data access by: denormalisation, summarization, data access statistics, index dynamic reorganization etc.
9) Backup and recovery strategies are different for the two types of systems. Most data in data warehouses is historical data, which is non-variant and does not require repetitive saving. New data can be saved at the time of loading. It is advisable for data to be saved from intermediary databases in certain cases in order to minimize impact on the performance of data warehouses. Recovery policies may also be different in the case of data warehouses as opposed to operational databases, depending on how critical permanent, seamless access to data warehouses is for the organization. In actual database backup and recovery task is for DBMS. In actual data warehouse this task is for database administrator.
10) Not only data organization is different in data warehouse from operational databases, but the interfaces used as well. Data warehouses support analytical processing by OLAP – On-Line Analytical Processing, which differs from a functional point of view from transactional processing applications. The attempt to perform analytical processing and comprehensive queries in operational databases will only reduce performance.

The differences shown above are part of the reasons why data warehouses are built separately from operational databases. The separation of the two systems ensures the scalability of business intelligence solutions as well as their ability to answer rapidly and efficiently to queries on the company. Data warehouses allow comprehensive analyses, as the structures of data collections: are more simple – only necessary information is retain, are standardized – structures are well documented, and are denormalised – there are fewer joins between data collections.

2.5 Data extraction, transformation, and loading

Data extraction:

Extraction is the operation of extracting data from a source system for further use in a data warehouse environment. This is the first step of the ETL (Extract-Transform-Load) process. After the extraction, this data can be transformed and loaded into the data warehouse. Designing and creating the extraction process is often one of the most time-consuming tasks in the ETL process and, indeed, in the entire data warehousing process. The source systems might be very complex and poorly documented, and thus determining which data needs to be extracted can be difficult. The data has to be extracted normally not only once, but several times in a periodic manner to supply all changed data to the data warehouse and keep it up-to-date. Moreover, the source system typically cannot be modified, nor can its performance or availability be adjusted, to accommodate the needs of the data warehouse extraction process.
Designing this process means making decisions about the following two main aspects:

- Which extraction method do I choose?
  This influences the source system, the transportation process, and the time needed for refreshing the warehouse.
- How do I provide the extracted data for further processing?
  This influences the transportation method, and the need for cleaning and transforming the data.

The extraction method you should choose is highly dependent on the source system and also from the business needs in the target data warehouse environment. Very often, there is no possibility to add additional logic to the source systems to enhance an incremental extraction of data due to the performance or the increased workload of these systems.
This section contains the following topics:

- Logical Extraction Methods
- Physical Extraction Methods
- Change Tracking Methods

Transportation in Data Warehouses:

Transportation is the operation of moving data from one system to another system. In a data warehouse environment, the most common requirements for transportation are in moving data from:

- A source system to a staging database or a data warehouse database
- A staging database to a data warehouse
- A data warehouse to a data mart

Transportation is often one of the simpler portions of the ETL process, and can be integrated with other portions of the process. You have three basic choices for transporting data in warehouses:

- Transportation Using Flat Files
- Transportation Through Distributed Operations
- Transportation Using Transportable Tablespaces

Loading and Transformation in Data Warehouses:

Data transformations are often the most complex and, in terms of processing time, the most costly part of the extraction, transformation, and loading (ETL) process. They can range from simple data conversions to extremely complex data scrubbing techniques. Many, if not all, data transformations can occur within an Oracle database, although transformations are often implemented outside of the database (for example, on flat files) as well.

You can use the following mechanisms for loading a data warehouse:

- Loading a Data Warehouse with SQL*Loader
- Loading a Data Warehouse with External Tables
- Loading a Data Warehouse with OCI and Direct-Path APIs
- Loading a Data Warehouse with Export/Import

2.6 Issues/challenges of data warehousing

Consider the following data warehouse challenges before building a data warehouse:

1. **Data Quality** – In a data warehouse, data is coming from many disparate sources from all facets of an organization. When a data warehouse tries to combine inconsistent data from disparate sources, it encounters errors. Inconsistent data, duplicates, logic conflicts, and missing data all result in data quality challenges. Poor data quality results in faulty reporting and analytics necessary for optimal decision making.
2. **Understanding Analytics** – The powerful analytics tools and reports available through integrated data will provide credit union leaders with the ability to make precise decisions that impact the future success of their organizations. When building a data warehouse, analytics and reporting will have to be taken into design considerations. In order to do this, the business user will need to know exactly what analysis will be performed. Envisioning these reports will be difficult for someone that hasn't yet utilized a BI strategy and is unaware of its capabilities and limitations.
3. **Quality Assurance –** The end user of a data warehouse is using Big Data reporting and analytics to make the best decisions possible. Consequently, the data must be 100 percent accurate or a credit union leader could make ill-advised decisions that are detrimental to the future success of their business. This high reliance on data quality makes testing a high priority issue that will require a lot of resources to ensure the information provided is accurate. The credit union will have to develop all of the steps required to complete a successful Software Testing Life Cycle (STLC), which will be a costly and time intensive process.
4. **Performance** – Building a data warehouse is similar to building a car. A car must be carefully designed from the beginning to meet the purposes for which it is intended. Yet, there are options each buyer must consider to make the vehicle truly meet individual performance needs. A data warehouse must also be carefully designed to meet overall performance requirements. While the final product can be customized to fit the

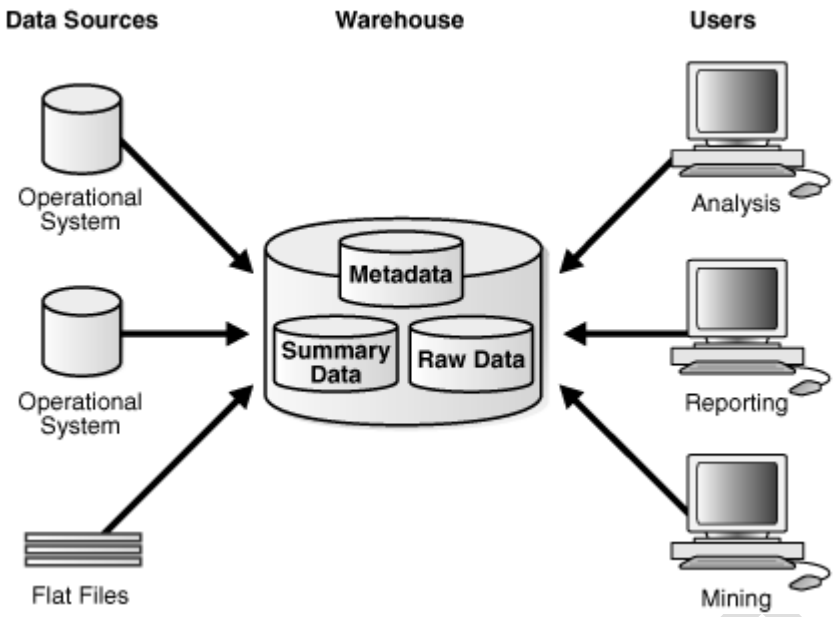
performance needs of the organization, the initial overall design must be carefully thought out to provide a stable foundation from which to start.

5. **Designing the Data Warehouse –** People generally don't want to "waste" their time defining the requirements necessary for proper data warehouse design. Usually, there is a high level perception of what they want out of a data warehouse. However, they don't fully understand all the implications of these perceptions and, therefore, have a difficult time adequately defining them. This results in miscommunication between the business users and the technicians building the data warehouse. The typical end result is a data warehouse which does not deliver the results expected by the user. Since the data warehouse is inadequate for the end user, there is a need for fixes and improvements immediately after initial delivery. The unfortunate outcome is greatly increased development fees.
6. **User Acceptance** – People are not keen to changing their daily routine especially if the new process is not intuitive. There are many challenges to overcome to make a data warehouse that is quickly adopted by an organization. Having a comprehensive user training program can ease this hesitation but will require planning and additional resources.
7. **Cost** – A frequent misconception among credit unions is that they can build data warehouse in-house to save money. As the foregoing points emphasize, there are a multitude of hidden problems in building data warehouses. Even if a credit union adds a data warehouse "expert" to their staff, the depth and breadth of skills needed to deliver an effective result is simply not feasible with one or a few experienced professionals leading a team of non-BI trained technicians. The harsh reality is an effective do-it-yourself effort is very costly.

2.7 Understanding data warehouse architecture

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but can include data from other sources. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources.

Figure 1 shows a simple architecture for a data warehouse. End users directly access data derived from several source systems through the data warehouse.



This figure illustrates three things:

- Data Sources (operational systems and flat files)
- Warehouse (metadata, summary data, and raw data)
- Users (analysis, reporting, and mining)

2.8 Use of data warehouse

There are hundreds of reasons why a data warehouse is useful, the following list be a good starting point:

- Real-time issues – your current systems aren't enabled to ***integrate*** disparate sources of data ***and*** keep historical records of those integrations, ***in near real-time.***
- Scalability issues – you have tons of historical data you need to gather in to an easily accessible place, common formats, common keys, and common access methods. AND you need to ensure that the system is scalable over the next 3 to 5 years.
- Avoidance of Siloed Solution Sets – if you have many different or disparate solutions already in existence, yet your corporation is unable to answer common questions requiring consistency across your enterprise.
- Enterprise Class System of Record – across historical and integrated data sets, if you have a need to do this, you probably need an enterprise data warehouse
- Disparate Source Systems along with Internal and External Data Sets – if you need to ingrate all of these for a single enterprise vision WITH HISTORY, then you need a data warehouse.
- Self-Service BI – if you have a need to eventually reach this goal, where users can "visualize" and construct their own reports, then you probably need an enterprise data warehouse, along with it's highly integrated historical facts from all the different sources in your organization.
- Kick start for a Master Data Management initiative. If you want Master Data, then it is important to understand the nature of your history – where the problems exist, how the data does and does not align with business perception, and basically where to "get" the golden copies of records you want to begin populating your MDM solution (remember: MDM is NOT just a tool, it's people, process, governance, and so on).. Yes, you can build MDM solutions without a Data Warehouse, but how good is your confidence that the data you selected is truly "gold copy" if you don't have historical evidence to back it up?
- If you do ANY sort of data mining, you need a data warehouse. Data Mining is becoming (or already is) the heart-and-soul of better decision making in BI. And of course, the mining engine is only as smart as the domain of information that you provide to it, along with the model that is designed. Statistics say: you can project for 1/2 as much time as you have history for. So: with 2 years, you can project (with some accuracy) only 1 year out. The same goes for Data Mining initiatives, AND the better interconnected the data set is (by Business Keys across the enterprise) the better your Data Mining confidence ratings will be.

Data Warehouse Users

The success of a data warehouse is measured solely by its acceptance by users. Without users, historical data might as well be archived to magnetic tape and stored in the basement. Successful data warehouse design starts with understanding the users and their needs. Data warehouse users can be divided into four categories: Statisticians, Knowledge Workers, Information Consumers, and Executives.

- **Statisticians:** There are typically only a handful of sophisticated analysts—Statisticians and operations research types—in any organization. Though few in number, they are some of the best users of the data warehouse; those whose work can contribute to closed loop systems that deeply influence the operations and profitability of the company. It is vital that these users come to love the data warehouse. Usually that is not difficult; these people are often very self-sufficient and need only to be pointed to the database and given some simple instructions about how to get to the data and what times of the day are best for performing large queries to retrieve data to analyze using their own sophisticated tools. They can take it from there.

- **Knowledge Workers:** A relatively small number of analysts perform the bulk of new queries and analyses against the data warehouse. These are the users who get the "Designer" or "Analyst" versions of user access tools. They will figure out how to quantify a subject area. After a few iterations, their queries and reports typically get published for the benefit of the Information Consumers. Knowledge Workers are often deeply engaged with the data warehouse design and place the greatest demands on the ongoing data warehouse operations team for training and support.

- **Information Consumers:** Most users of the data warehouse are Information Consumers; they will probably never compose a true ad hoc query. They use static or simple interactive reports that others have developed. It is easy to forget about these users, because they usually interact with the data warehouse only through the work product of others. Do not neglect these users! This group includes a large number of people, and published reports are highly visible. Set up a great communication infrastructure for distributing information widely, and gather feedback from these users to improve the information sites over time.

- **Executives:** Executives are a special case of the Information Consumers group. Few executives actually issue their own queries, but an executive's slightest musing can generate a flurry of activity among the other types of users. A wise data warehouse designer/implementer/owner will develop a very cool digital dashboard for executives, assuming it is easy and economical to do so. Usually this should follow other data warehouse work, but it never hurts to impress the bosses.