Chapter #4

User-Requirement Analysis

The process of gathering information about the needs and requirements of the users of the data warehouse

⇒ After investigating the available information resources in the data source analysis and reconciliation phase, approach end users again to get more accurate information.

→ The user requirement analysis phase attempts to collect end users needs for data mart applications and usage.

→ **Importance:** This phase holds a strategic significance for designing data marts, because it influences every decision made during the project.

→ it plays necessary role in:

- Defining Conceptual schemata of data
- Data staging design
- Requirements for data analysis applications
- Data warehouse system architecture
- Instructions for data warehouse system maintenance
- Upgrades

Data come from this Phase:

Challenges:

→ The user-requirement analysis phase delivers ambigious, incomplete and shod-l... requirements in data warehousing projects.

- Long-term nature: These are long-term projects and it is difficult to get and ana... every requirement from the beginning.
- complexity of decision making: The information requirements of data warehouse applications are difficult to explain because decision-making process has very flexible structure.
- Difficulty in explaining information: Decision making process requirements frequently make reference to a set of information that is not available in suitable format and needs to be deriv... from data sources.

Approaches to user-requirement analysis phase:

⇒ There are two different approaches:

~~User requirement analysis~~

1. Informal approach
2. Formal approach

1: Informal approach:

→ Requires glossaries to support designers in the conceptual design phase.

→ Perfectly fits into a data-driven design framework that emphasizes deriving data mart schemata from operational source schemata.

2. Formal approach:

- Based on Tropos formalism for user-requirement analysis.
- Tropos formalism helps in designing system based on the needs and goals of users.
- This approach emphasizes that system design should be directly derived from the user's requirements.
- It integrates user requirements with practical limitations, like existing rules or processes to create a balanced design.

Ways of Requirement

1: Interviews

1. Business users:

The primary sources of requirement for data warehouse projects are the end users, also known as business users.

2. Challenges in interviews:

Interviews can be difficult due to language differences between designers and end users, and end-users limited understanding of data warehousing

3. Accurate requirements:

Accurate requirements are essential for ensuring end user satisfaction & the successful implementation of the data warehouse.

4. Technical know-how:

Information system administrators and data processing centers staffs can provide valuable insights into technical constraints.

5. Constraints:

Constraints on data warehouse design, such as performance, scalability and security must be considered during the requirement analysis phase.

Interview methods:

→ Two basic methods for conducting user requirement analysis are:

- Interviews
- Requirement analysis

1: Interviews:

Can be conducted with individual users or small groups and offers the advantage of active participation and detailed specifications.

2: Facilitated sessions:

it involve large, heterogeneous groups lead by facilitors, who are responsible for setting up common language for all interviewees.

Can be helpful for encouraging creative brainstorming and setting general priorities.

Preliminary activities:

The main preliminary activities involved in interviews are:

1. Pre-interview research:

- Conduct research on the business issues and the specific application domain to gain awareness among interviews.
- Also analyze previous data-warehousing projects to understand why they were unsuccessful.

2: Interview selection:

- Select a representative sample of end-users to obtain a wide range of information on the analysis needs in the data-mart specific application field.

3: Interview question development:

- Develop guidelines for interviews, but they should vary by interview job functions and levels.

4: interview sheduling:

- Shedule the first meeting with the project sponsor to get their support and involvement.

5: Interviewee Preparation:

- Clarify interview goals and plans with all participants.

Types of questions during interview:

Combination of open-ended, closed-questions and evidential questions to gather information.

1. Open-ended questions:

- Allow interviewees to express their opinions and ideas freely.

2. Closed-ended questions:

- Focus the interview on specific topics and obtain direct answers.

3. Evidential questions:

- Ask interviewees to provide examples or evidence to support their answers.

Advantages of open-ended questions:

- Provide rich and detail information
- Allow interviewees express ideas freely

Disadvantages of OF questions:

- Time-consuming
- Lead to digressions

Advantages of closed questions:

- Shorten the interview time
- Focus interviewees on relevant topics
- Make answers comparable

Disadvantages of closed questions:

- Can sound boring
- Limit the interviewee's freedom
- Assume that the interviewer knows the relevant topics.

Advantages of evidential questions:

- Allow the interviewer to learn about the interviewee's experience and knowledge.

Disadvantages of evidential questions:

- Can be time-consuming
- May not be relevant to the interview if the interview lacks experience.
- Can make interviewees nervous because the questions are probing.

Types of interviews

→ There are two types or approaches for doing interviews:

- Pyramid-Shaped
- Funnel-Shaped

1. Pyramid-Shaped interviews:

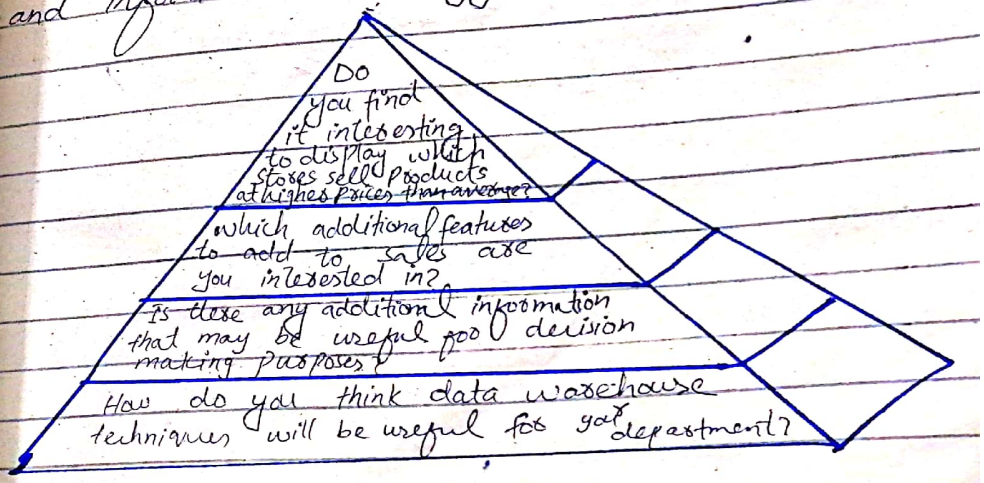
- it is an inductive approach. Start with detailed questions and gradually become more general.
- Usually with closed questions.

Advantages:

overcomes interview hesitation and allows the interviewer to avoid getting deeply involved in the interview.

Example:

Start with specific questions about data marts and then ask more general questions about decision-making processes and information technology.



2. Funnel-Shaped interviews:

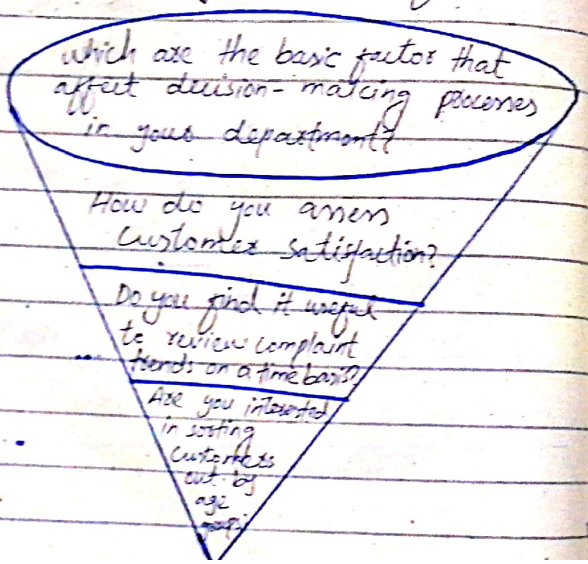
- Deductive approach.
- Start with general questions and gradually become more specific.

Advantages:

Useful if interviewees are unsure or too differential, as general questions don't have wrong answers and can make interviewees feel more comfortable.

Example:

- Start with general questions about the company goals
- Then ask more specific questions about data analysis and information needs.



2. Glossary-based Requirement Analysis:

based on → fact
→ Preliminary workflow

→ A method of gathering information from users about what they need from the data warehouse. This method involves creating a glossary of terms that describe the data and how it should be used.

→ It is informal approach suitable when using data driven approach.

→ In 2002, Lechtenbörger proposed user-requirement analysis approach based on glossaries.

⇒ **Glossaries:**

- These are like dictionaries that define the specific terms and concepts related to the data warehouse.

Derivation table:

- This table shows how the data in the data warehouse will be derived from the original data sources.

Usage table:

- This table explains how the data will be used for analysis and reporting.

→ it links textual description to attribut and specifies attribute roles such as analysis dimension and measure.

⇒ Both tables should list the attributes for data modeling.

Structure table:

- This table defines how the data will be organized and structured within the data warehouse.
- Specifies whether the attributes shou be modeled as dimensions or attributes li to dimensions or measure.

(i) Facts

→ Facts are fundamental units of data in data warehouse.

→ They represent measureable events or accu that are the primary focus of analysis.

→ Facts are the basic units of inform that users will analyze.

Dynamic feature: Facts should be able to cha ones time (e.g sales figures etc.)

Relevance: A fact may be relevant f one data warehouse but not for another.

Users interest: The choice of facts depends on what users want to analyze.

Data sources: The fact should be based on the data available in the original data sources.

Example of fact in various application fields, along with various data marts:

1. In Healthcare application field, under the Division data mart, Some typical facts include admissions, discharges, transfers, surgical operations, diagnosis and prescriptions. These facts can be used to analyze patient flow, treatment effectiveness and resource utilization within a healthcare facility.
2. In Business manufacturing, under Supplies data mart, Some typical facts are purchases, stock inventory and distribution.
- Under demand management data mart. Some facts are Sales, invoices, orders. Shipments and complaints.

Fact granularity:

Refers to level of detail at which data is to be represented in a data mart.

for example: Patient data can be stored at a granular level of individual medical records or aggregated to higher level, like monthly patient visits.

Trade-off:

- Choosing a granularity level involves balancing flexibility and performance.
- Higher granularity allows for more detailed analysis but can be computationally expensive, while lower granularity might be faster to process but less flexible.

Important points:

- One of the most important features of data warehouse is representation of historical data. That help to track how data changes over time
- historical interval: Every fact needs a historical interval to be defined. means each fact should have a specific time range associated with it.

→ The value of this interval typically range from three to five years.

Example of user-Requirement

Glossary:

→ Requirement glossary table created by data mart designers focused on sales and inventory data.

		measure	
Fact	Dimensions		High
STOCK INVENTORY	Product, date, warehouse	stocked quantity	1 Ye
SALES	Product, date, store	Sold quantity, receipts, discount	5 Yea
ORDER LINES	Product, date, supplier	ordered quantity, received quantity	3 Yea

(ii) Preliminary Workload

→ This involves identifying the facts, dimensions, and measures that will be included in the data warehouse. Basically, Specification on analysis queries data must include

→ **Fact identification:**

- Primary measurements within data warehouse.

Dimension & measure identification:

- it involves identifying attributes that describe facts and measures that are quantifiable values associated with facts.

Workload:

- Workload can be expressed in natural language that will help the designer to overcome the doubt about fact granularity and and relevant measures.

Workload refinement:

- Do after, aim of workload refinement is to validate conceptual schema created by designers.
- Detail information is collected after the conceptual design phase.

Example of preliminary work

Fact	Query
STOCK INVENTORY	what is [Average quantity ? availability of product ?] in data
SALES	what are [Total sales ? Daily sales ? Sales by product type or the
ORDER LINES	what is [order details ? Total orders ? order amount ?

⇒ The **STOCK INVENTORY** fact is concerned:

- The data should be organized by product type and time (day, week, month).
- The quantity of products in stock should be calculated by adding up the quantities in different warehouses.
- Average and maximum quantities are calculated along the time.

⇒ The **SALES** fact is concerned:

- The data should be organized by product type, product category, store and time (day, week, Year).
- Also calculate average price per unit sold.
- The total quality sold should be calculated by adding up the quantities sold in different stores and over time.

⇒ For **ORDER LINES** same as STOCK INVENTORY

Importance of:

(i) Hierarchies:

- way to organize data in levels like tree structure.
- Ensures attributes are consistent in different facts
- Hierarchies build on dimensions.

(ii) Aggregations:

- Calculations that summarize data.
- Choice of aggregation operator can significantly affect the logical and concep design of data warehouse.

⇒ In short, while designing a data wa Considering both hierarchies and aggregat to ensure that the data can be effectivel analysed and used to answer busine questions.

3: Goal oriented Requirement Analysis

⇒ This methodology focuses on understanding the goals of decision-makers in an organization rather than just collecting a list of requirements.

Abstraction sheets:

⇒ These sheets are used to analyze goals separately and considers how they relate to enterprise-specific activities.

⇒ Goal oriented approach based on:

Tropos formalism:

⇒ Framework used to represent user requirements and includes both decision-making modeling and organizational modeling.

Organizational analysis:

⇒ Focuses on the stakeholders and actors involved in managing the organization.

Fact, dimension & measure analysis:

⇒ Involves identifying:

- facts dimensions measures that are relevant to the goals and organizational structure.

3-1: Introduction to TROOPS

→ Developed by Bresciani at 2004
→ Based on modeling framework Yu is
→ Software designing methodology.
→ Focuses on understanding the goal of people and systems and how the interact.

Two main features:

(i) Agent-oriented methodology:

Troops uses agents, which are like actors or objects, to represent people and systems.

(ii) Goal-based approach:

Troops emphasizes understanding the goals of agents and how they work together to achieve those goals.

→ Troops mangges four Software develo stages:

1. Early requirement analysis:

- This phase focuses on understand the goals of agents and how they work together to achieve those goals.

Design

2. Early requirement analysis:

- This phase focuses on defining the system functions and features.

3. Architecture design:

- This phase involves designing the overall structure of the system.

4. Working plan:

- This phase involves creating a detailed plan for each Software module.

Notations of Troops

Some notations of Troops methodology that applies to requirement analysis for data masts:

1) Actors:

- Represent people, Software agents involved in a system.
- Business stakeholders.
- Circles are used to graphically represent the actor.

2) Strategic dependencies:

- These are agreements between actors about their goals, tasks or resources.
- Represented by ovals.

3. Actor diagram:

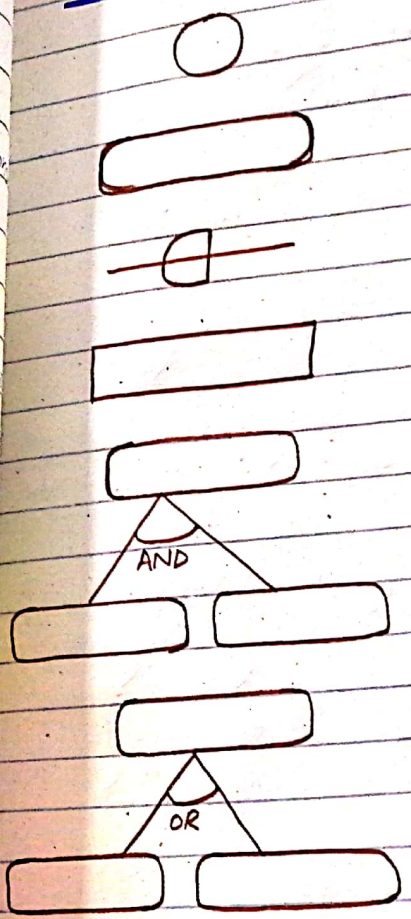
- This diagram shows how actors interact with each other based on their strategic dependencies.
- Used to describe how actors depend on each other.

4. Rationale diagram:

- This diagram shows the relation between actors and their goals.
- Logical relationships shows.
- it looks like a bubble.
- Goals are the objectives that actors want to achieve.
- Goals broken down into small subgoals using AND Semantics (every subgoal must be reached) or the OR semantics (any one of those subgoals must be reached).
- Positive or negative contributions to goals are represented by arrows.
- That indicated one goal must be achieved before another.

- Contribution show how one goal can help or hinder another goal.
- The +, ++, - and -- marks are used to rate the contributions.

Graphical Notations



Actor

Goal

Dependency

Fact

AND - decomposition

OR - decomposition

⇒ To apply Tropos approach to data warehouse following concepts should be considered.

Facts:

→ From stand point of organizational modeling:
- a fact models a group of events taking place when an actor reaches a goal.

→ From standpoint of decision-making modeling:
- Facts are potential analysis items.

→ Rectangles connected to goals graphically mark facts in diagrams.

Attributes:

→ Field that is a given value recorded a fact after reaching a goal.

→ Small diamonds connected to goals graphically mark facts in diagrams.

Dimensions:

→ Property of fact.

→ Help decision makers gain different perspectives on facts to reach analysis goals.

→ Small circles connected to goals graphically mark dimensions in diagrams.

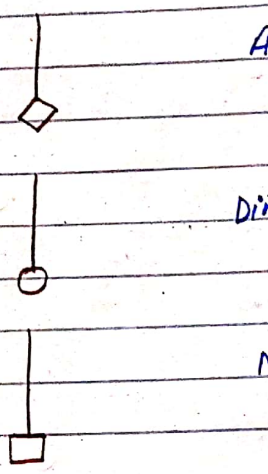
Measure:

→ Numerical property of fact.

→ Specifies quantitative feature of a fact.

→ Small squares connected to goal graphically mark measures in diagrams.

Graphical Notations



3.2: Organizational Modeling

→ Process used to understand and analyze an organization's structure, goals and relationships between different stakeholders.

→ To model organizations, a three phase process is follows.

→ Each phase consists of an iterative process whose inputs are the diagrams resulting from the previous step.

→ Three phases are as follows:

1. Goal analysis:

- Actor and rationale diagrams are created. These diagrams help understand goals and relationships of different people or groups involved in the organization.

2. Fact Analysis:

- In second phase, more information added to the rationale diagrams.
- This information is about facts that are happening in the organization.

3. Attribute Analysis:

- In third phase, rationale diagrams are extended further by adding more information.
- In this phase, speific characteristics qualities of facts are specify.

3-2-1: Goal Analysis:

→ Once goals are analyzed, then actor diagram is created and rationale diagram.

(1) Actor diagram:

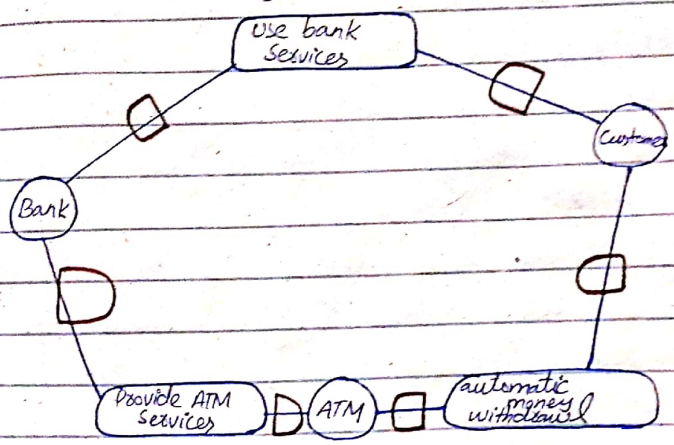
→ This diagram shows the different stakeholders involved in the organization both internal and external.

→ it also shows how these stakeholders interact with each other and the organization.

→ Actors in actor diagram should represent agents, roles and positions in organization.

Example

Actor diagram of Bank.



Stakeholders (Actors):

- Customer
- ATM
- Bank

Relationships and interactions:

- Customer actor depends on Bank actor to be able to use bank services.
- Customer actor also depends on the ATM actor, so that able to withdraw money.
- The Bank actor depends on the ATM actor, so that bank will be able to provide ATM services to its customers.

Goals:

- **Customer:**
  To access bank services conveniently and efficiently.
- **ATM:**
  Provide a reliable and secure way to access bank services. Bridge b/w customers & bank.
- **Bank:**
  Satisfy customers needs by providing ATM services.

Rationale Diagrams:

→ The second step is constructing rationale diagram.

→ Goal of each stakeholder (actor) is determined and rationale diagram is created for each actor/stakeholder in the organization.

→ The goal analysis phase is completed when all the relevant goals for each actor are specified and all the dependencies b/w them are identified.

Example

- The rationale diagram for a bank.
- Goals of bank:
  - Gaining new customers
  - Managing loans
  - Providing services to customers

(I)

- The "gain new customers" goal for the bank is AND-decomposed into two smaller goals:
  1. Search new customers
  2. Enter new customers
- The "Enter new customers" goal is further AND-decomposed into three smaller goals:

1: Specify contract details
2: Enter customer info
3: Enable customer services

→ The "enter customer info" goal is the AND-decomposed into three even smaller goals:

1. Collect customer info
2. enter personal data
3. Enter financial situation

→ New dependency is discovered, the Bank actor depends on its customers in order to collect their data.

(II)

→ The "manage loans" goal is AND-decomposed into four smaller goals:

1. enter info
2. check financial situation
3. Agree on loan specifications
4. Grant a loan.

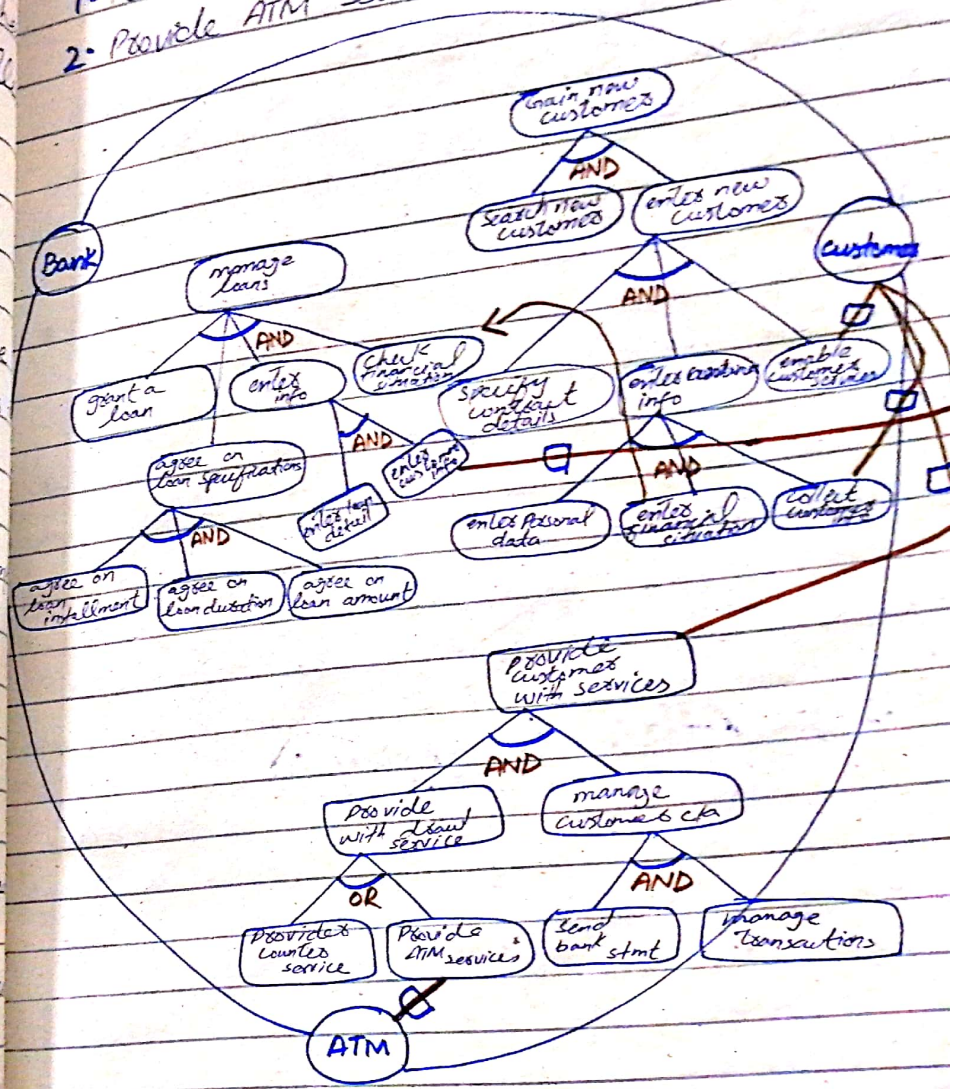
→ "Check financial situation" goal benefits from the enter "enter financial situation" goal.

(III) The "customer services" is AND-decomposed into:

1) Manage Customer cla 2) Provide withdrawl se

→ The customer service also OR-decomposed into:

1. Provide customer service
2. Provide ATM service.



3.2.2: Fact analysis :-

→ In this phase identify all those facts relevant to the organization.

→ Then these facts "links" to the goals dependencies identified in the previous Pha

→ Fact analysis should be done care and systematically.

→ Explore all the rationale diagram and goals in order to identify all the relevant facts.

→ In previous page example of ration diagram for bank includes facts such transactions, deposits, withdrawls and paym

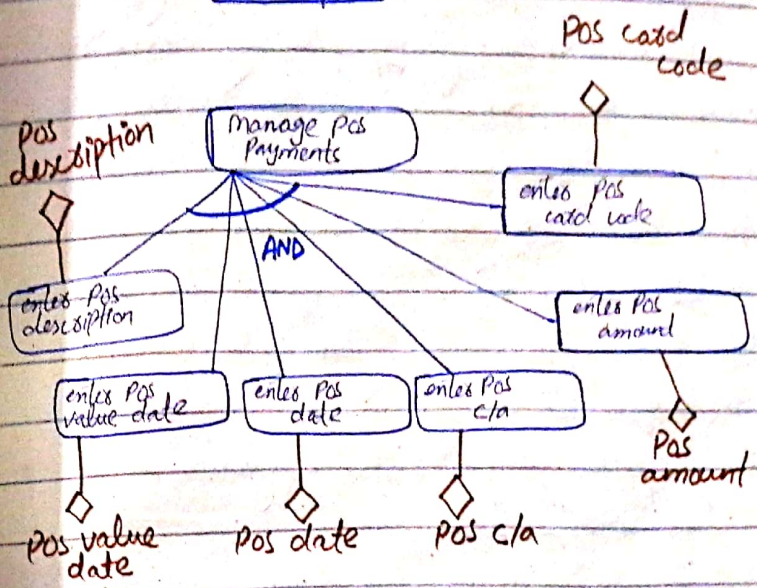
→ These facts are linked to the goals of bank, such as managing transactions an providing services to customers.

3.2.3: Attribute analysis:

In this phase, identify all the attribute of the facts that identified in the previ Phase. Attributes are the specific details that describe the facts.

Review all the rationale diagram created in the previous phase (for every actor) and should als explores the subgraphs of the goals identify the attributes that are relevant each goal.

Example



→ The attributes added to the "manage Pos payments" goal. are :

- Pos value date
- Pos date
- Pos description
- Pos c/a
- Pos amount
- Pos card code

3.3: Decision-making Model

→ Process used to understand and anal
how decisions are made within the organiz
→ It involves identifying the key decision
makers, their goals, and the factors tha
influence their choices.
→ it follow a four step process:

Goal analysis:

- Understanding the goals of decision makers.
- Leads to rationale diagrams.

Fact analysis:

- Specifies facts or additional information in order to creat extended rationale diagram.

Dimensional analysis:

- Adds dimensions to extended ration diagrams.

Measure analysis:

- Adding measurements to extende rationale diagrams.

3-3-1: Goal analysis:

→ First step in decision making modeli process.

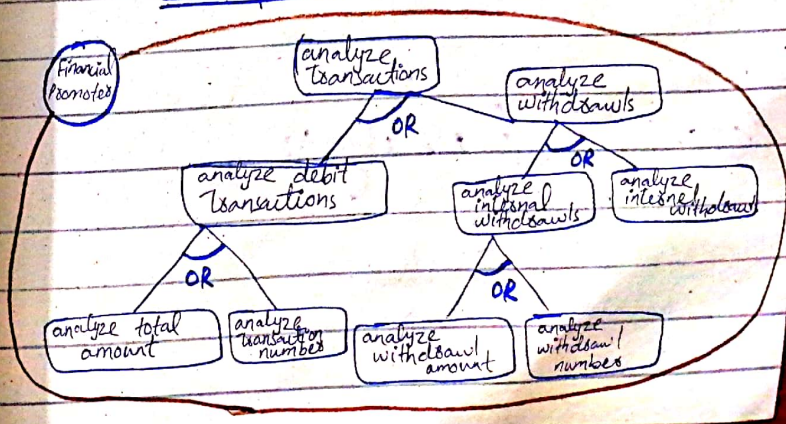
It involves:

- Examining actor diagrams - that show the roles of different people involved in decision-making.
- Identifying decision makers
- Studying decision makers dependencies

Goal decomposition:

→ Goals of each decision maker break down into smaller, more specific goals and studied them closely in order to draw a rationale diagram.
→ Goals can be completely different from those detected in the organizational modeling Phase.
→ They are integral part of the decision-making process.

Example:



→ This diagram is a rationale diagram for the financial promotes decision maker.

→ Various maing goals and have subgoals of each goal.

Main Goal:

- Analyze transactions

Subgoals:

1. Analyze debit transactions.
   - (i) Analyze total amount
   - (ii) Analyze transaction number
2. Analyze withdrawls
   - (i) Analyze internal WD
     - Analyze withdrawl amount
     - Analyze withdrawl number
   - (ii) Analyze External WD

→ All goals are OR-decomposed.

3-3-2: Fact Analysis:

→ In fact analysis :-

Analysts should extend the rationale diagrams by identify facts and match those facts to decision-makers goals

→ Facts are usually imported from extended rationale diagram of the organizational modeling phase.

Example:

→ Extended rationale diagram of financial promotes on next page.

→ Show the Transactions fact, which was detected in the organizational modeling Phase, to analyze transactions goal.

→ Sometimes analysts can also add new facts as they directly analyze rationale diagrams of decision makers.

3-3-3: Dimensional Analysis:

→ In this phase, each fact is connected to those dimensions that decision maker considers important for every decision-making to be reached.

Example:

→ Considers extended rationale diagram of financial promotes on Next page.

- the c/a and month dimensions are specified to reach the "analyze total amount" goal.

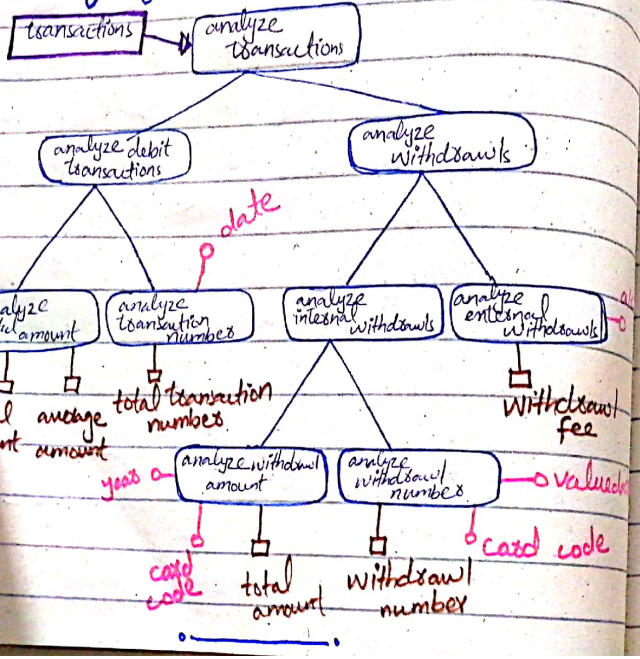
3.3.4: Measure analysis:

→ In this phase analysts should:
Connect a set of measures to every fact previously identified.

Example:

→ The average amount and total amount measures are identified for the "analyze amount." fact.

Extended Rational diagram of financial promotes:



4: Additional Requirements:

→ Beyond the basic specifications of facts and preliminary workload, some additional specifications are essential for various phases of datawarehouse house projects.
→ They should be collected from end users and information system administrators.
→ Those specifications concern the following:

1. Logical and physical design:

- These requirements relate to how the data warehouse will be structured and implemented.
- Both logical and physical design are influenced by the constraints posed by data warehouse designers and corporate IT staff.

→ Logical design refers to the conceptual structure of the data warehouse, including the entities, attributes and relationships between them. it is like creating blueprint for the warehouse.
→ Physical design refers to the actual implementation of the data warehouse,

including the choice of hardware, software and database technologies. it is like building the warehouse based on the blueprint.

2. Data-staging design:

→ This refers to how data is initially stored and processed before being loaded into the data warehouse.

→ The frequency of data staging depends on the needs of end users and the capabilities of operational sources.

→ Frequency of data staging refers to how often data is extracted from operational systems and loaded into the staging area of a data warehouse.

→ Typically, frequency values of data staging range from a day to a week.

→ Special real-time applications can also require hourly updates.

3. Data quality:

This refers to the accuracy and trustworthiness of the data.

it is important to set criteria and thresholds for discarding faulty data and to establish during ETL a reliable workflow for ensuring data quality.

4. Data warehouse architecture:

→ This refers to the overall structure and design of the data warehouse, including the:

- number of layers
- Use of data marts
- Reconciled layer materialization techniques

5. Data analysis applications:

→ This refers to the types of queries and reports that end users will need to perform on the data warehouse.

→ it is important to consider the features of the analysis tools that will be used.

6. Startup schedule:

→ This refers to the timeline for implementing the data warehouse.

→ Schedule for startup a data warehouse decided with support from DBA who provided their expertise and ensure their complete collaboration with end-users.

7. Training schedule:

→ it is important to provide training to users so that they can effectively the data warehouse.

→ Designers should create a training and offer follow-up sessions as need

Chapter #5
Conceptual Modeling

Challenges of designing data mart

→ Entity-relationship model is widely used for database design, but it is not well suited for multidimensional data, which is common in data marts (data mart is based on multidimensional view of data).

→ The ERM expressive enough to represent the majority of concepts necessary for data modeling.

→ However, the basic ERM is not able to accurately highlight the distinctive features of the multi-dimensional modeling.

→ That's why its use for data mart is not suitable.

Challenges of designing data marts using star schema:

Designers focus on the structure and relationships of the data within the data mart without considering the underlying