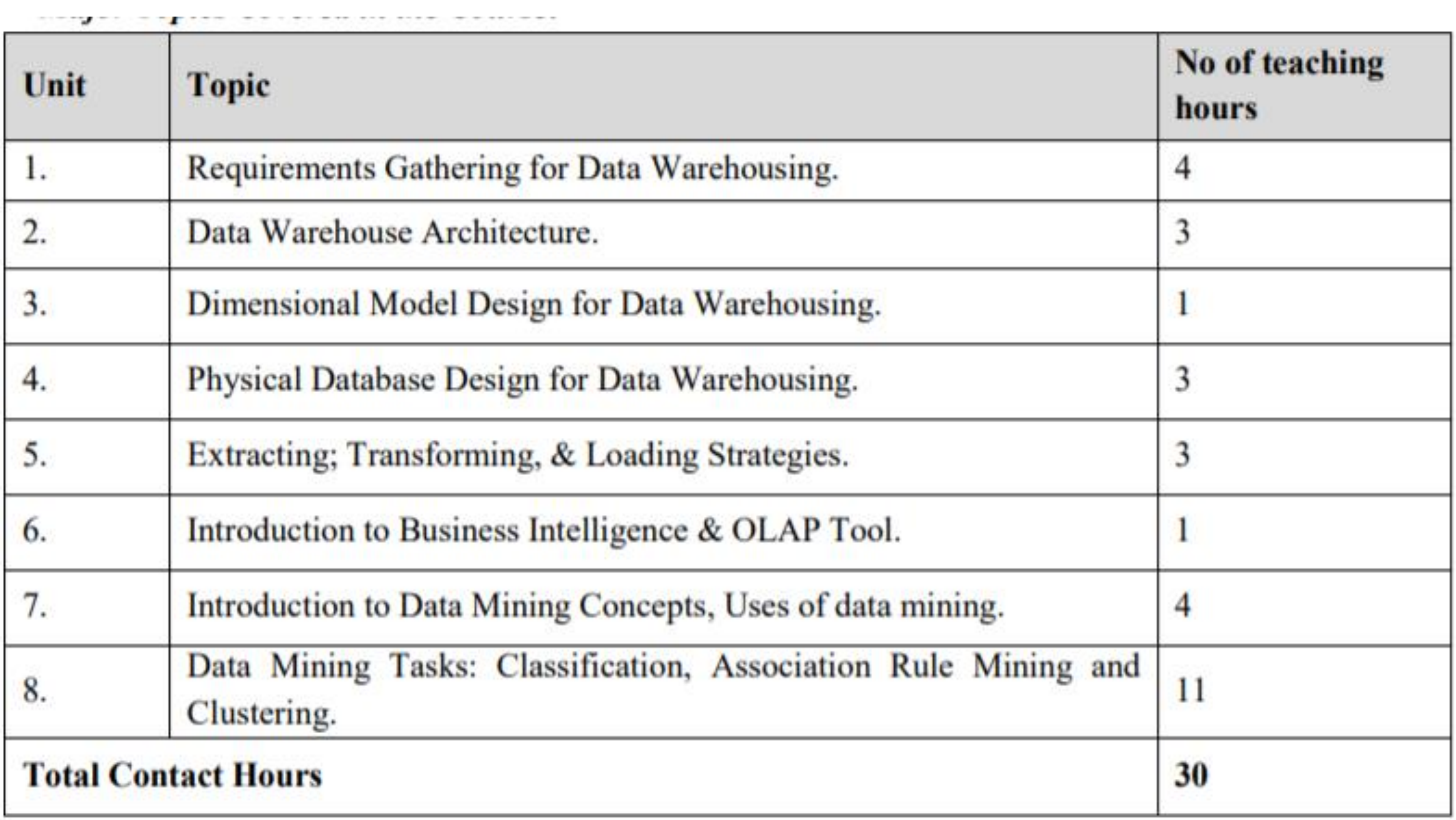
DATA WAREHOUSING AND DATA MINING

Introduction

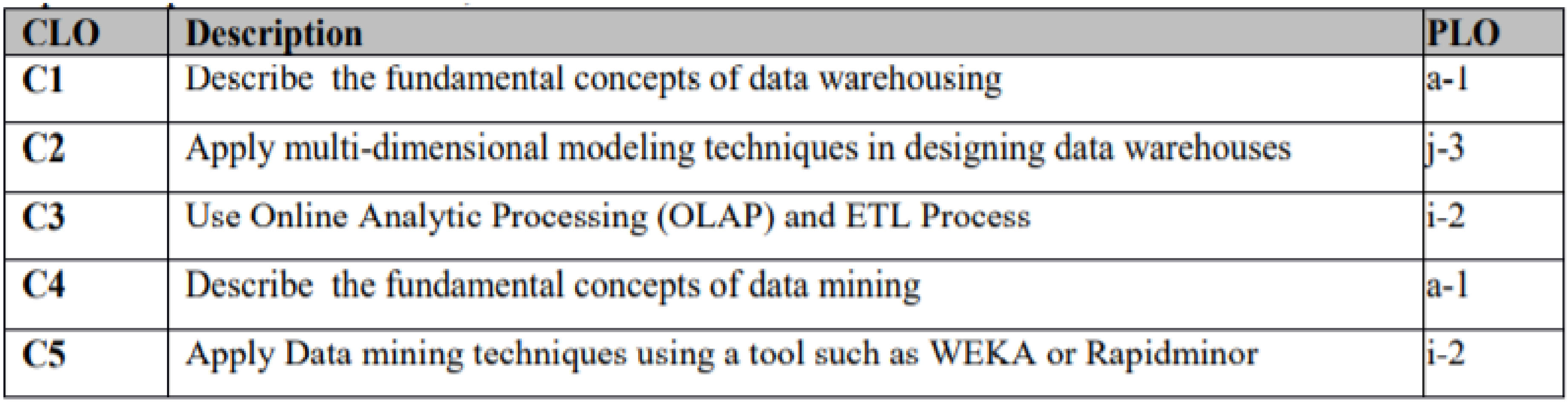
Sajid Majeed

Course outline



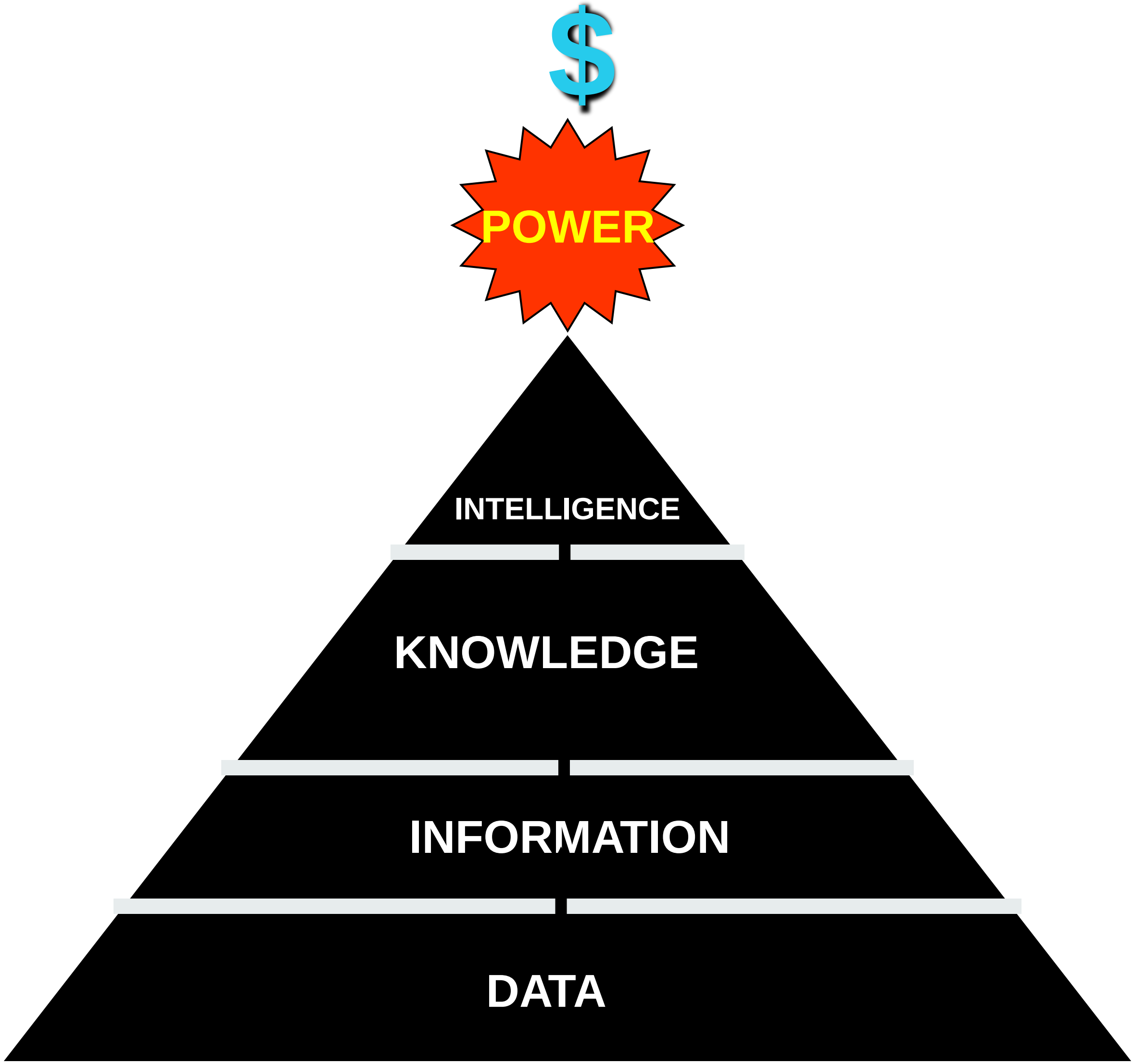
Unit	Topic	No of teaching hours
1.	Requirements Gathering for Data Warehousing.	4
2.	Data Warehouse Architecture.	3
3.	Dimensional Model Design for Data Warehousing.	1
4.	Physical Database Design for Data Warehousing.	3
5.	Extracting; Transforming, & Loading Strategies.	3
6.	Introduction to Business Intelligence & OLAP Tool.	1
7.	Introduction to Data Mining Concepts, Uses of data mining.	4
8.	Data Mining Tasks: Classification, Association Rule Mining and Clustering.	11
Total Contact Hours		30

Course outline



CLO	Description	PLO
C1	Describe the fundamental concepts of data warehousing	a-1
C2	Apply multi-dimensional modeling techniques in designing data warehouses	j-3
C3	Use Online Analytic Processing (OLAP) and ETL Process	i-2
C4	Describe the fundamental concepts of data mining	a-1
C5	Apply Data mining techniques using a tool such as WEKA or Rapidminor	i-2

The need



The need

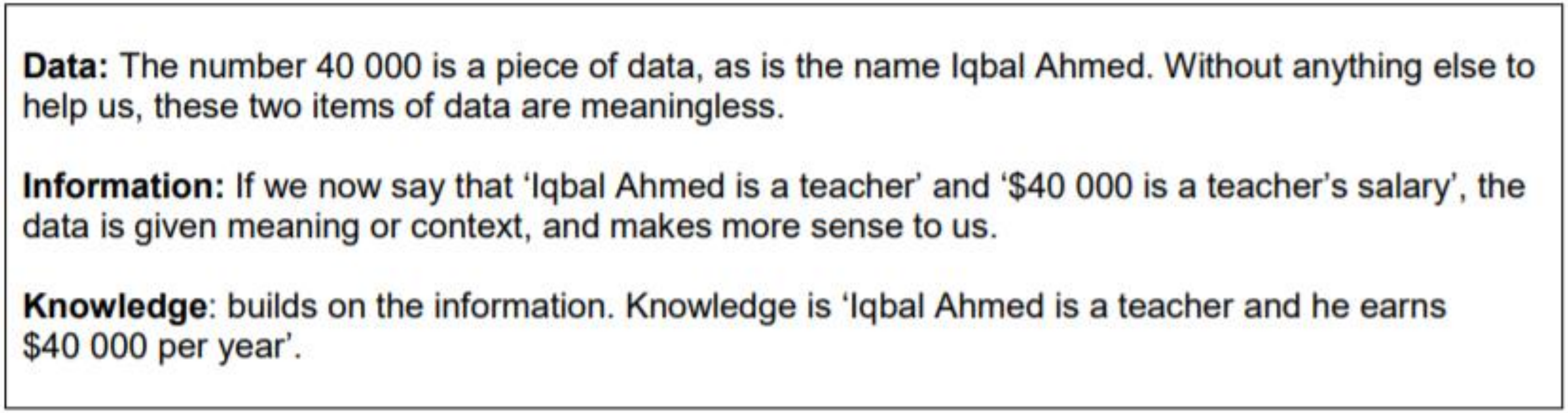
- **Data**: Data is defined as numerical or other facts represented or recorded in a form suitable for processing by computers.

  Information: is processed data that is meaningful. Data that has been processed, e.g. grouped, normally by a computer, to give it meaning and make it interpretable

  Knowledge: Knowledge, is an application of information and data.

-

The need



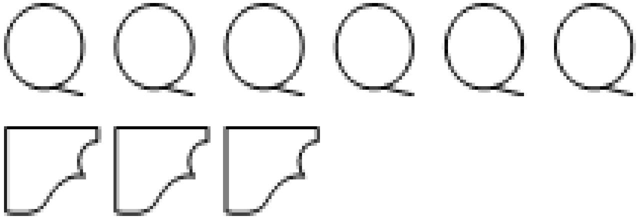
Data: The number 40 000 is a piece of data, as is the name Iqbal Ahmed. Without anything else to help us, these two items of data are meaningless.

Information: If we now say that ‘Iqbal Ahmed is a teacher’ and ‘$40 000 is a teacher’s salary’, the data is given meaning or context, and makes more sense to us.

Knowledge: builds on the information. Knowledge is ‘Iqbal Ahmed is a teacher and he earns $40 000 per year’.

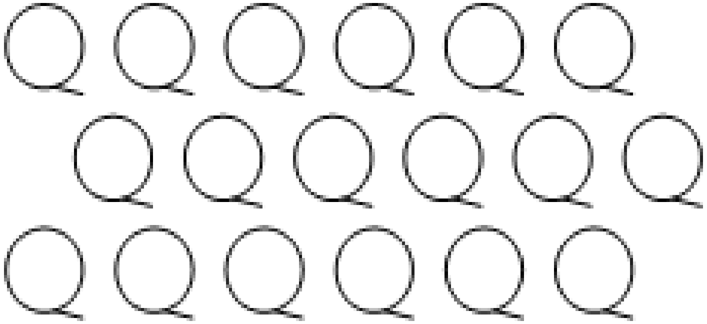
Historical overview

1960



Master files, reports

1965

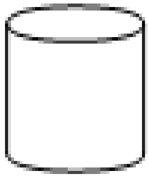


Lots of master files!

- Complexity of–
  - Maintenance
  - Development
- Synchronization of data
- Hardware

Historical overview

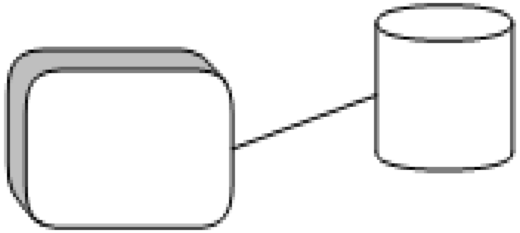
1970



DASD
DBMS

Database–"a single source
of data for all processing"

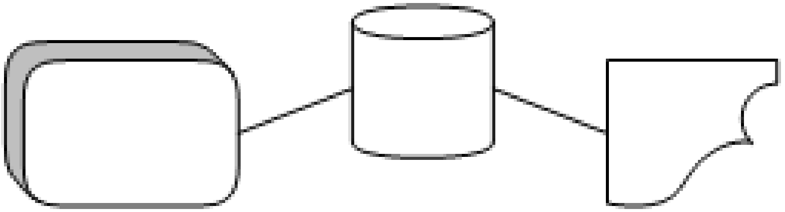
1975



Online, high-performance
transaction processing

Historical overview

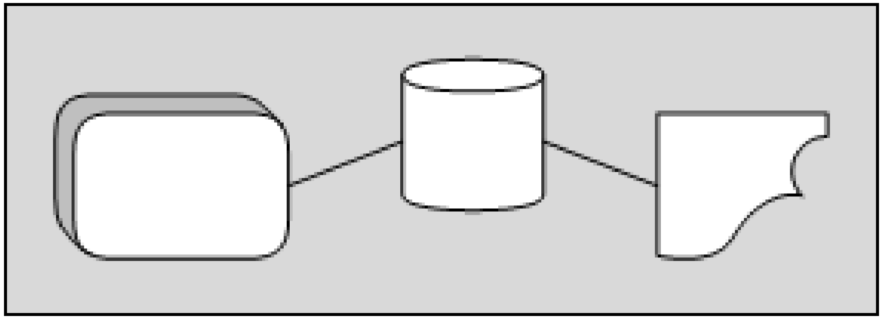
1980



tx processing

MIS/DSS

PCs, 4GL technology

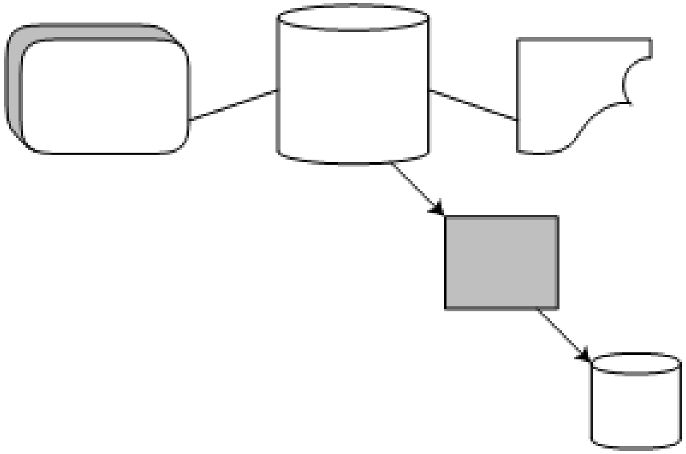


The single-database-serving-all-purposes paradigm

Historical overview

▨ Start with some parameters, search a file based on the satisfaction of the parameters, then pull the data elsewhere.

Extract processing

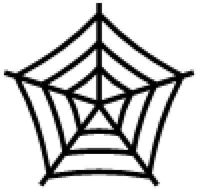


Why extract processing?

- Performance
- Control

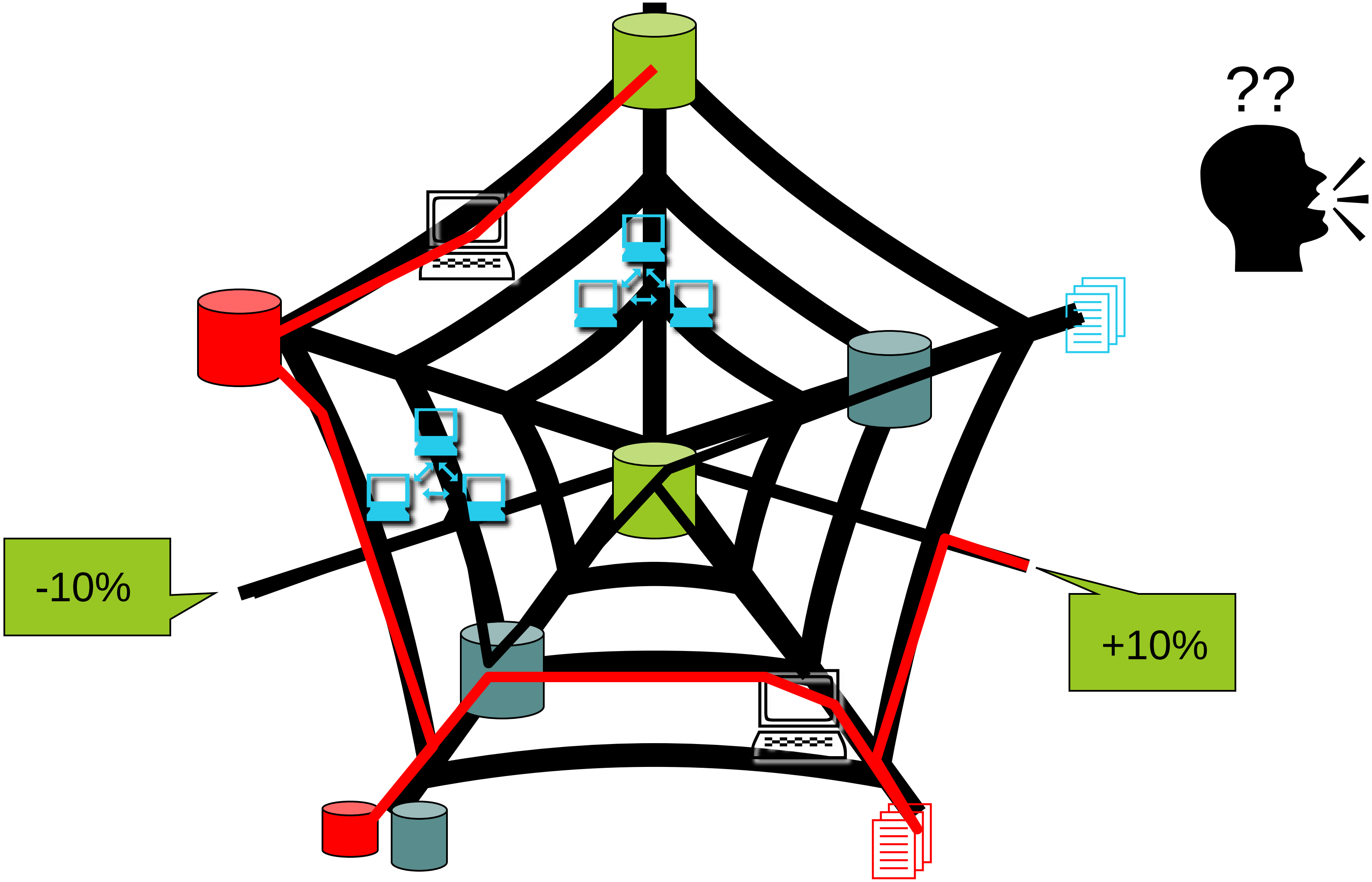
1990

The legacy system's web



Historical overview: Crisis of Credibility

What is the financial health of our company?



Crisis of Credibility

- Department-A which uses a different set of data sources, external reports etc. as compared to Department-B comes with a different answer sales up by 10%, as compared to the Department-B i.e. sales down by 10%.
- This is a typical example of the crisis in credibility because both departments got different view of the business using different sources.