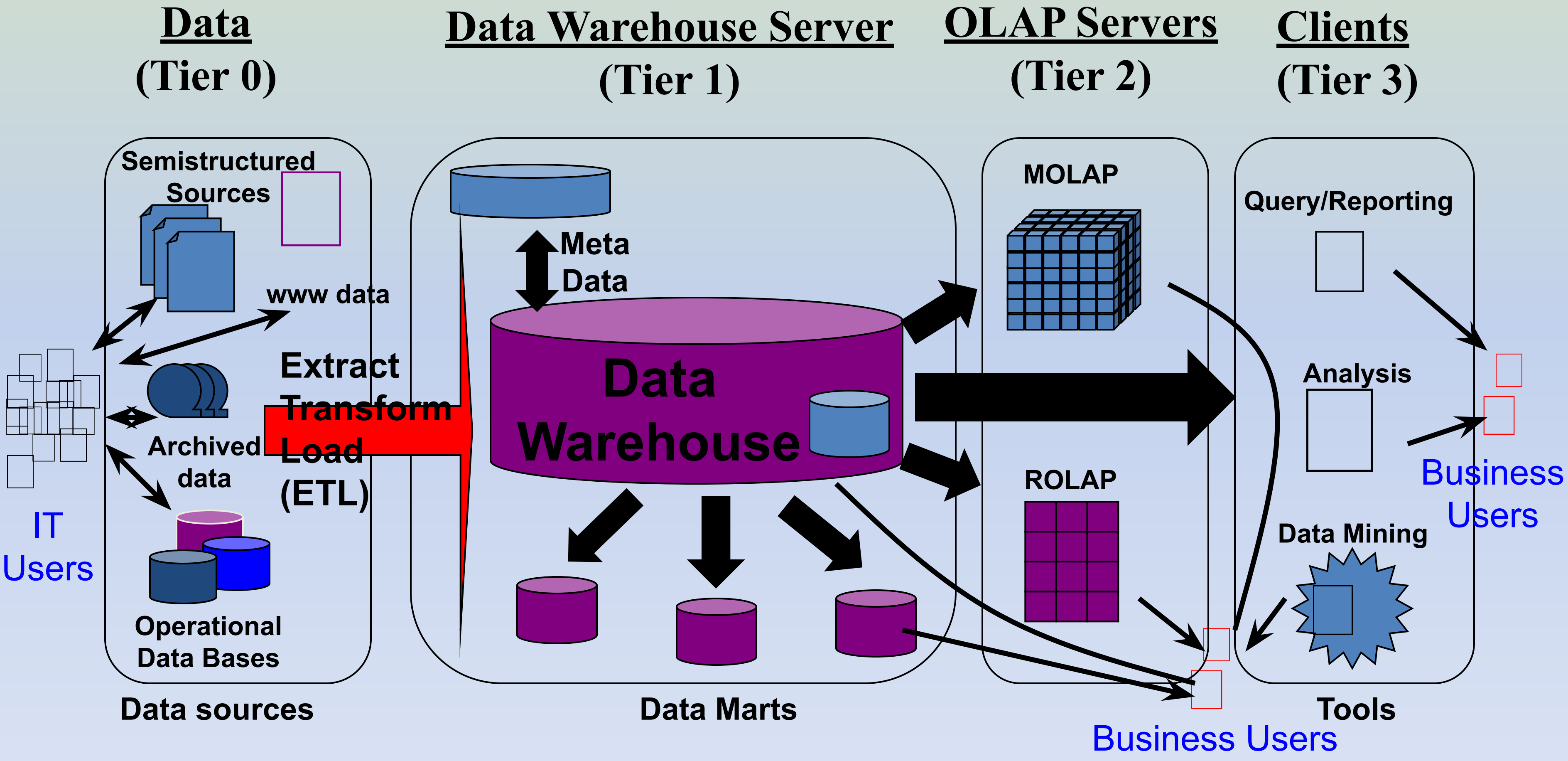
Data Warehousing and Data Mining

Extract Transform Load (ETL)
Part 1 and 2

Putting the pieces together



{Comment: All except ETL washed out look}

The ETL Cycle

EXTRACT

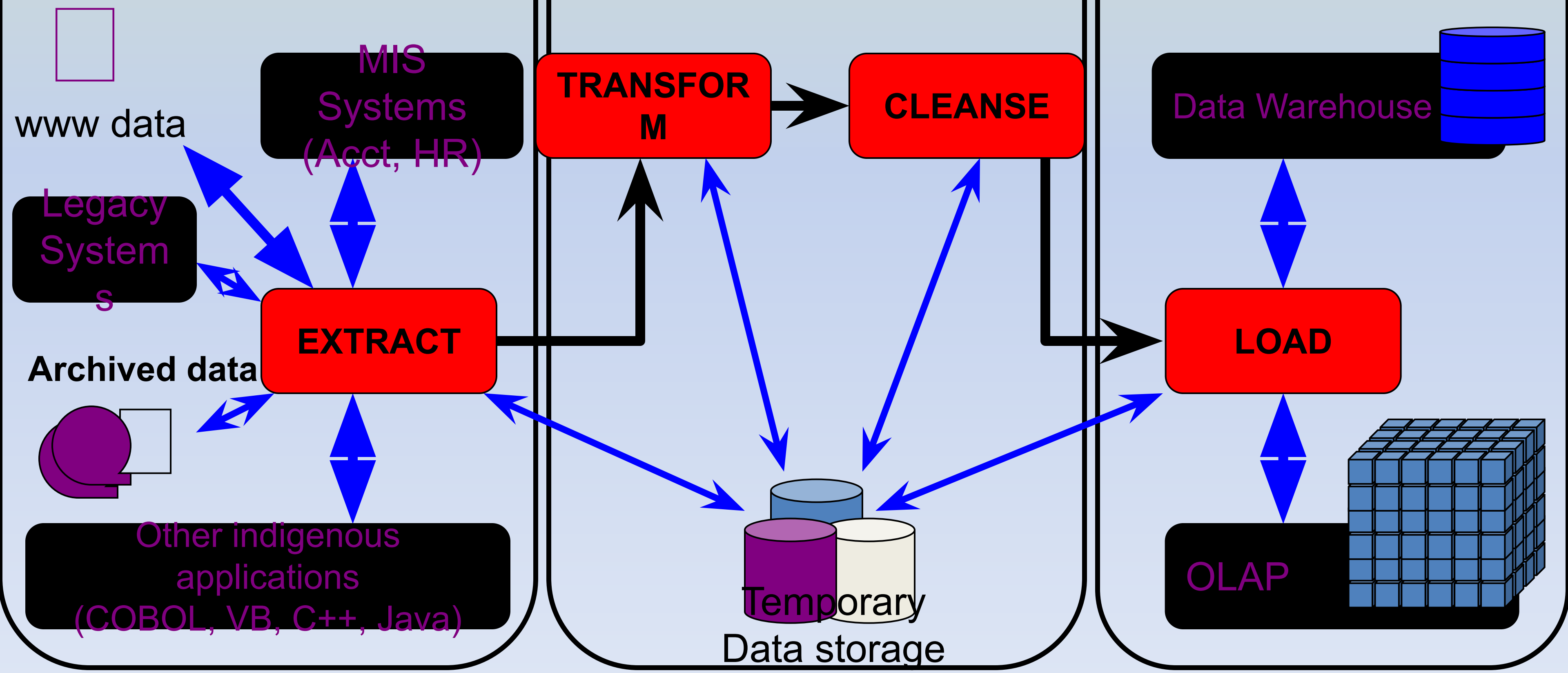
The process of reading data from different sources.

TRANSFORM

The process of transforming the extracted data from its original state into a consistent state so that it can be placed into another database.

LOAD

The process of writing the data into the target source.



www data

MIS Systems (Acct, HR)

Legacy System s

Archived data

EXTRACT

Other indigenous applications (COBOL, VB, C++, Java)

TRANSFOR M

CLEANSE

Temporary Data storage

Data Warehouse

LOAD

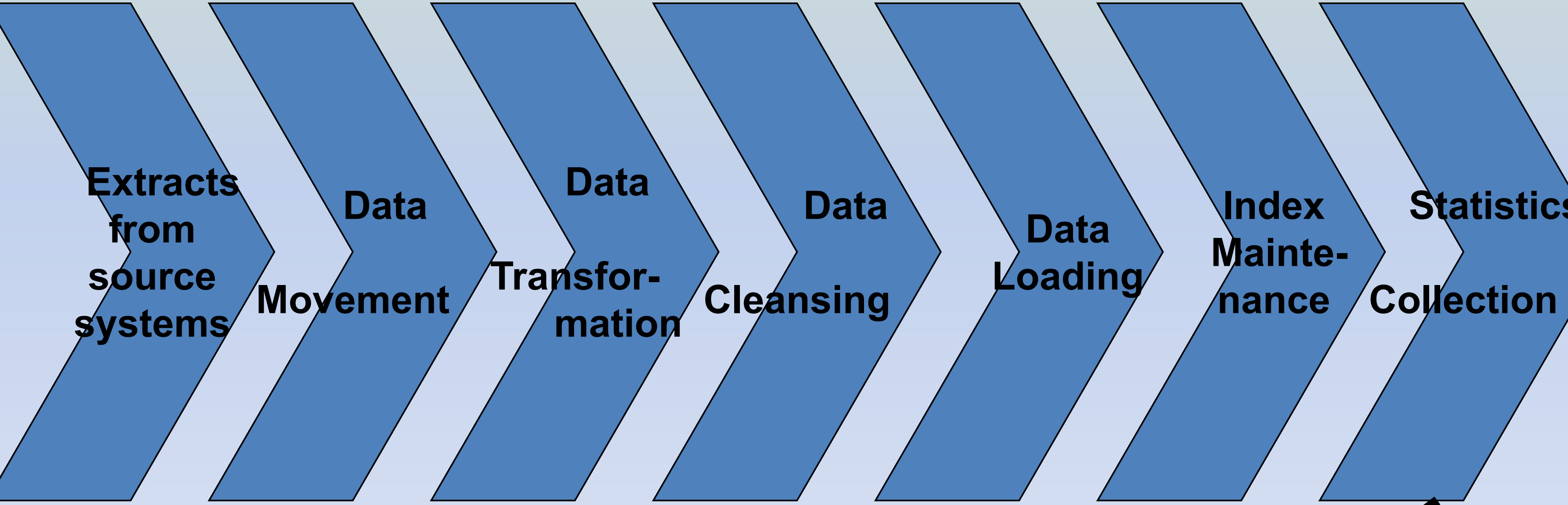
OLAP

ETL Processing

ETL is independent yet interrelated steps.

It is important to look at the big picture.

Data acquisition time may include…



Extracts from source systems

Data Movement

Data Transfor-mation

Data Cleansing

Data Loading

Index Mainte-nance

Statistics Collection

Note: Backup will come as other elements after “Statistical collection”

Backup

Back-up is a major task, its a DWH not a cube

Overview of Data Extraction

First step of ETL, followed by many.

Source system for extraction are typically OLTP systems.

A very complex task due to number of reasons:

- Very complex and poorly documented source system.
- Data has to be extracted not once, but number of times.
-

The process design is dependent on:

- Which extraction method to choose?
- How to make available extracted data for further processing?

Types of Data Extraction

- **Logical Extraction**
  - Full Extraction
  - Incremental Extraction
- **Physical Extraction**
  - Online Extraction
  - Offline Extraction
  - Legacy vs. OLTP

Logical Data Extraction

- Full Extraction
  - The data extracted completely from the source system.
  - No need to keep track of changes.
  - Source data made available as-is with no additional information.
- Incremental Extraction
  - Data extracted after a well defined point/event in time.
  - Mechanism used to reflect/record the temporal changes in data (column or table).
  - Sometimes entire tables off-loaded from source system into the DWH.
  - Can have significant performance impacts on the data warehouse server.

Physical Data Extraction…

- **Online Extraction**
- **Data extracted directly from the source system.**
- **May access source tables through an intermediate system.**
- **Intermediate system usually similar to the source system.**

- **Offline Extraction**
- **Data NOT extracted directly from the source system, instead staged explicitly outside the original source system.**

- **Data is either already structured or was created by an extraction routine.**

- **Some of the prevalent structures are:**
  - **Flat files**
  - **Dump files**
  - **Redo and archive logs**
  - **Transportable table-spaces**

Data Transformation

- **Basic tasks**
  1. Selection
  2. Splitting/Joining
  3. Conversion
  4. Summarization
  5. Enrichment

Data Transformation Basic Tasks

- Selection

Data Transformation Basic Tasks

- Splitting/joining

Data Transformation Basic Tasks

- Conversion

Data Transformation Basic Tasks: Conversion Example-1

- Convert common data elements into a consistent form i.e. name and address.

Field format		**Field data**
First-Family-title	→	Muhammad Ibrahim Contractor
Family-title-comma-first	→	Ibrahim Contractor, Muhammad
Family-comma-first-title	→	Ibrahim, Muhammad Contractor

- Translation of dissimilar codes into a standard code.

Natl. ID	→	NID
National ID	→	NID

F/NO-2 **F-2** **FL.NO.2** **FL.2** **FL/NO.2** **FL-2** **FLAT-2** **FLAT#** **FLAT,2** **FLAT-NO-2** **FL-NO.2**	→	FLAT No. 2

Data Transformation Basic Tasks: Conversion Example-2

- Data representation change
  - EBCIDIC to ASCII
- Operating System Change
  - Mainframe (MVS) to UNIX
  - UNIX to NT or XP
- Data type change
  - Character, numeric and date type.
  - Fixed and variable length.

Data Transformation Basic Tasks

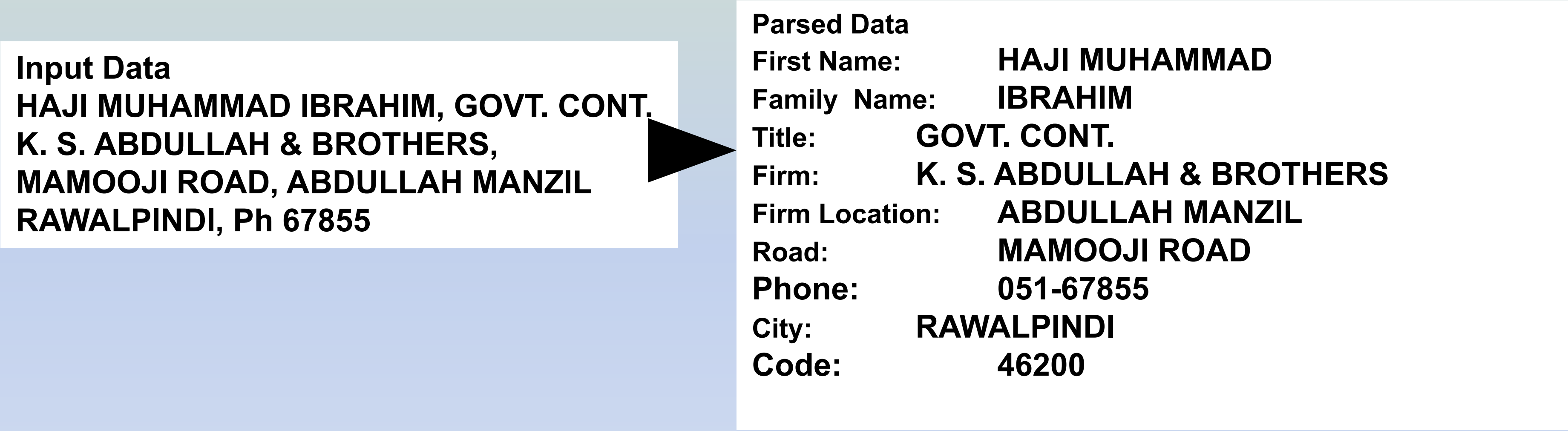
- Summarization

Data Transformation Basic Tasks

- Enrichment

Data Transformation Basic Tasks: Enrichment Example

- Data elements are mapped from source tables and files to destination fact and dimension tables.



Input Data
HAJI MUHAMMAD IBRAHIM, GOVT. CONT.
K. S. ABDULLAH & BROTHERS,
MAMOOJI ROAD, ABDULLAH MANZIL
RAWALPINDI, Ph 67855

Parsed Data

First Name:	HAJI MUHAMMAD
Family Name:	IBRAHIM
Title:	GOVT. CONT.
Firm:	K. S. ABDULLAH & BROTHERS
Firm Location:	ABDULLAH MANZIL
Road:	MAMOOJI ROAD
Phone:	051-67855
City:	RAWALPINDI
Code:	46200

- Default values are used in the absence of source data.

- Fields are added for unique keys and time elements.

Aspects of Data Loading Strategies

Need to look at:

- Data freshness
- System performance

- **Data Freshness**
  - Very fresh-- low update efficiency ✓
  - Historical data-- high update efficiency
  - Always trade-offs in the light of goals
- **System performance**
  - Availability of staging table space
  - Impact on query workload

Three Loading Strategies

- Once we have transformed data, there are three primary loading strategies:

- Full data refresh with BLOCK INSERT or 'block slamming' into empty table.

- Incremental data refresh with BLOCK INSERT or 'block slamming' into existing (populated) tables.

- Trickle/continuous feed with constant data collection and loading using row level insert and update operations.

ETL vs. ELT

There are two fundamental approaches to data acquisition:

ETL: Extract, Transform, Load in which data transformation takes place on a separate transformation server.

ELT: Extract, Load, Transform in which data transformation takes place on the data warehouse server.

Combination of both is also possible