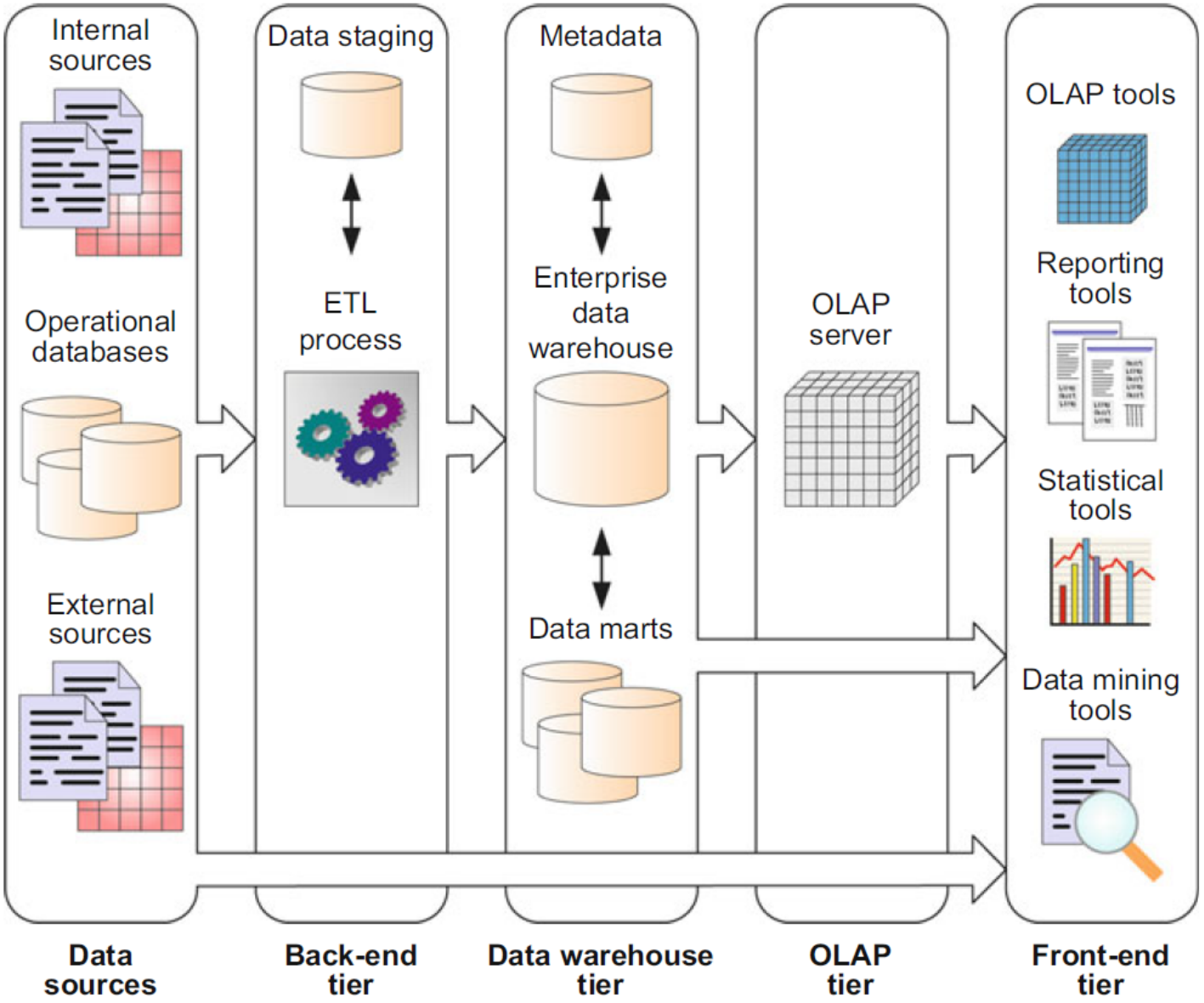
DS-306 Data Warehousing and Business Intelligence

Topic 7: ETL Fundamentals

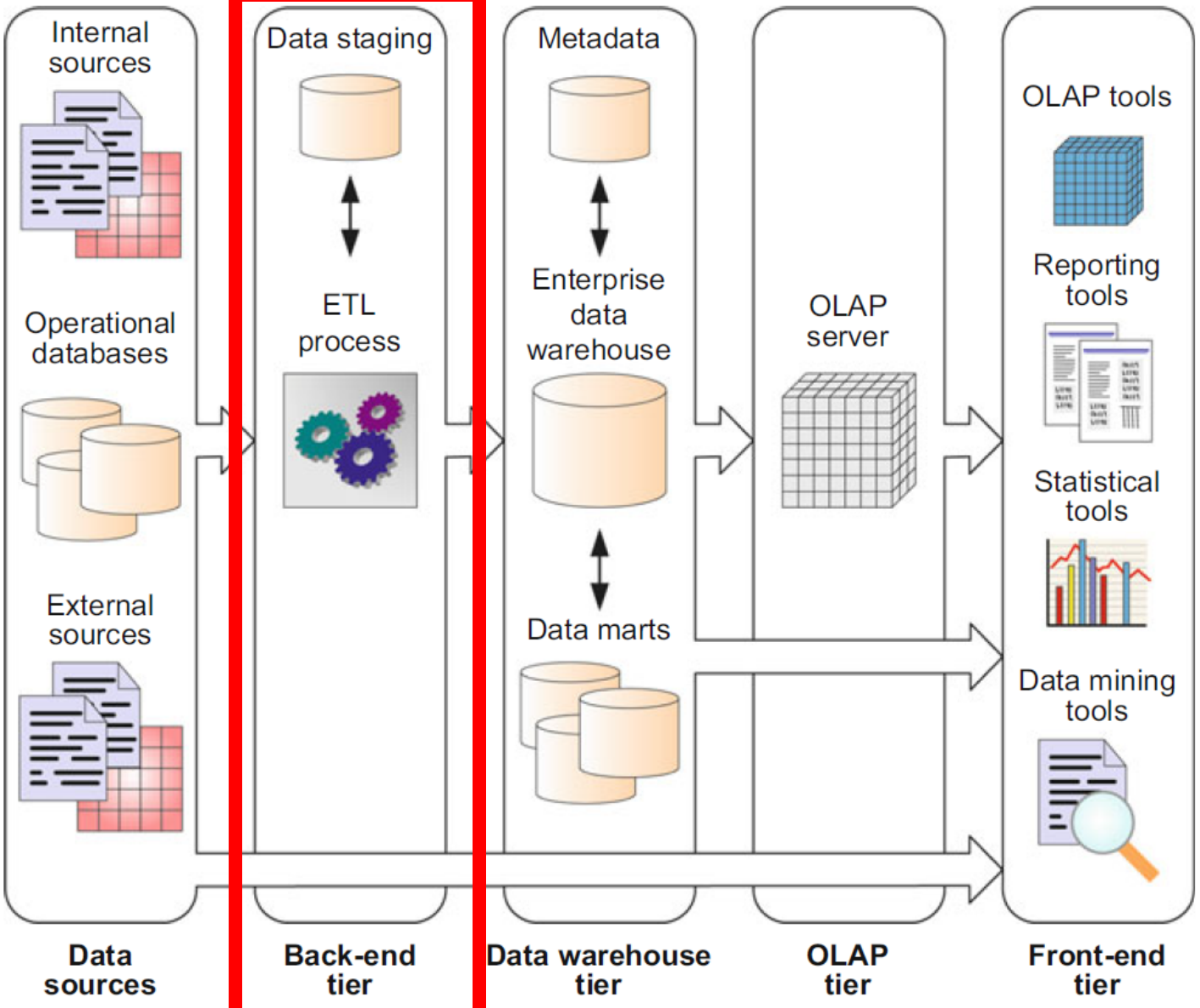
Dr. Khurram Shahzad

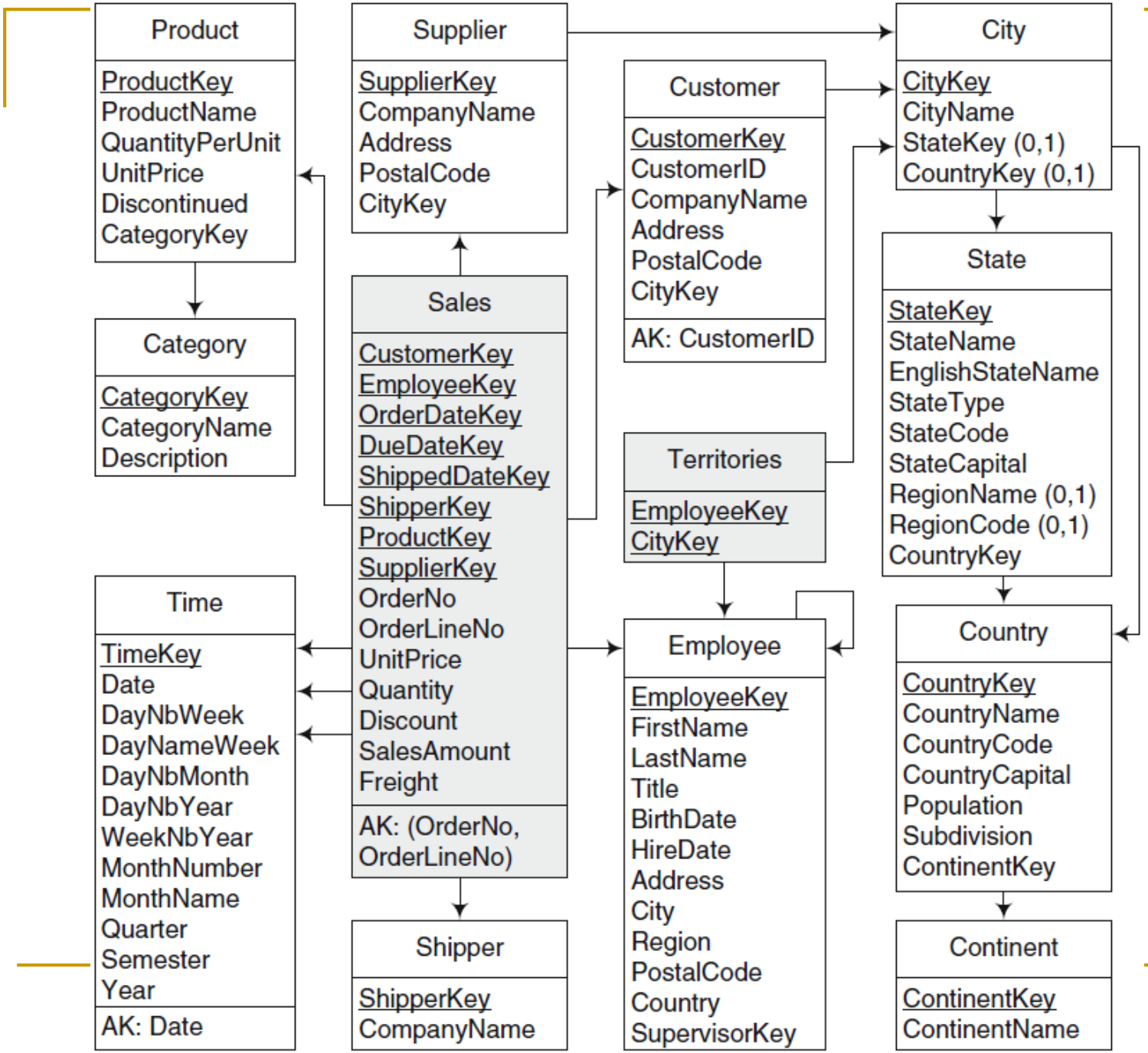
https://www.youtube.com/watch?v=0ikNnenDyNw

Typical DW architecture



Typical DW architecture




Product
ProductKey
ProductName
QuantityPerUnit
UnitPrice
Discontinued
CategoryKey
Supplier
SupplierKey
CompanyName
Address
PostalCode
CityKey
City
CityKey
CityName
StateKey (0,1)
CountryKey (0,1)
Customer
CustomerKey
CustomerID
CompanyName
Address
PostalCode
CityKey
AK: CustomerID
State
StateKey
StateName
EnglishStateName
StateType
StateCode
StateCapital
RegionName (0,1)
RegionCode (0,1)
CountryKey
Category
CategoryKey
CategoryName
Description
Sales
CustomerKey
EmployeeKey
OrderDateKey
DueDateKey
ShippedDateKey
ShipperKey
ProductKey
SupplierKey
OrderNo
OrderLineNo
UnitPrice
Quantity
Discount
SalesAmount
Freight
AK: (OrderNo, OrderLineNo)
Territories
EmployeeKey
CityKey
Time
TimeKey
Date
DayNbWeek
DayNameWeek
DayNbMonth
DayNbYear
WeekNbYear
MonthNumber
MonthName
Quarter
Semester
Year
AK: Date
Employee
EmployeeKey
FirstName
LastName
Title
BirthDate
HireDate
Address
City
Region
PostalCode
Country
SupervisorKey
Country
CountryKey
CountryName
CountryCode
CountryCapital
Population
Subdivision
ContinentKey
Shipper
ShipperKey
CompanyName
Continent
ContinentKey
ContinentName

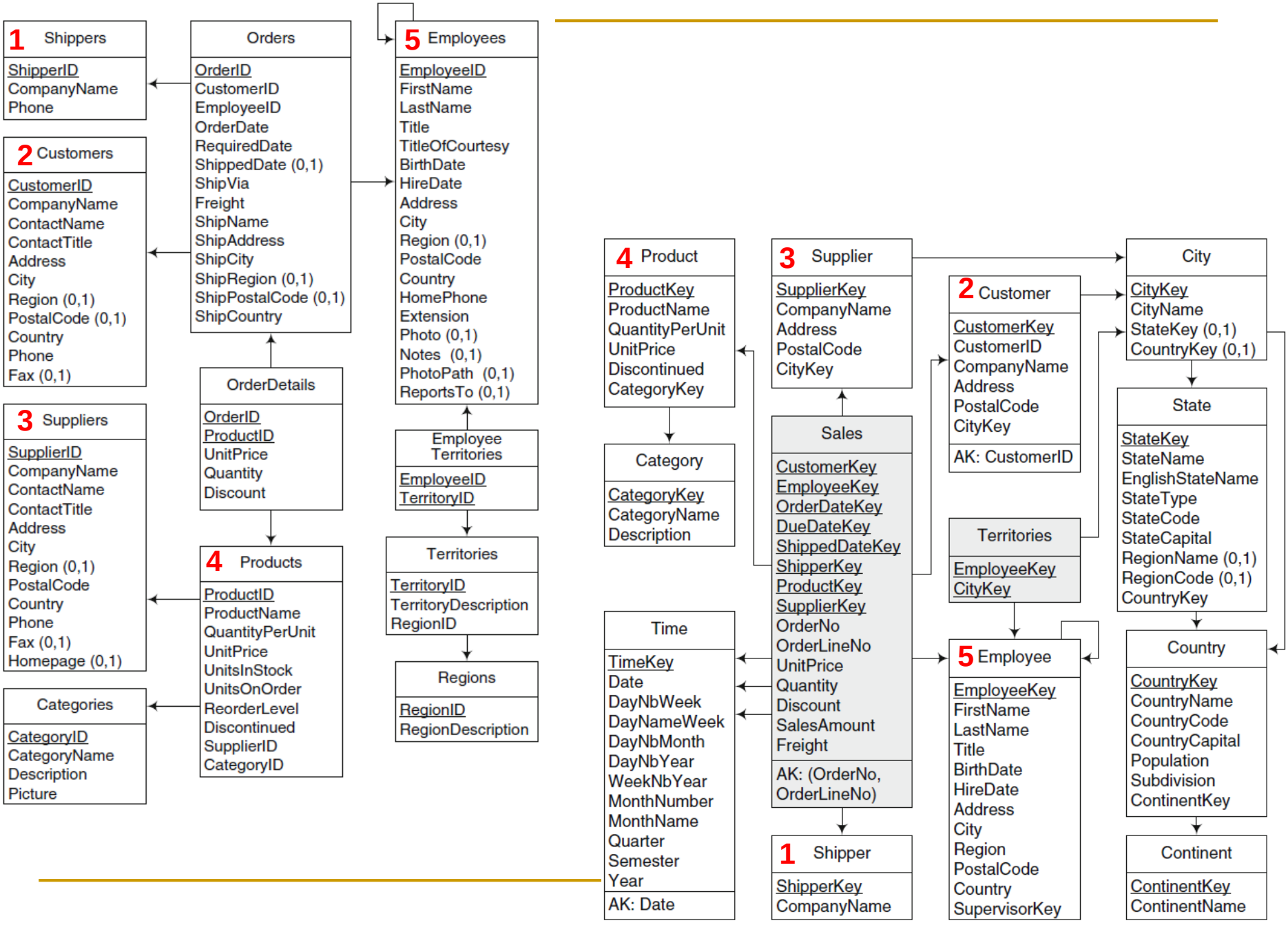
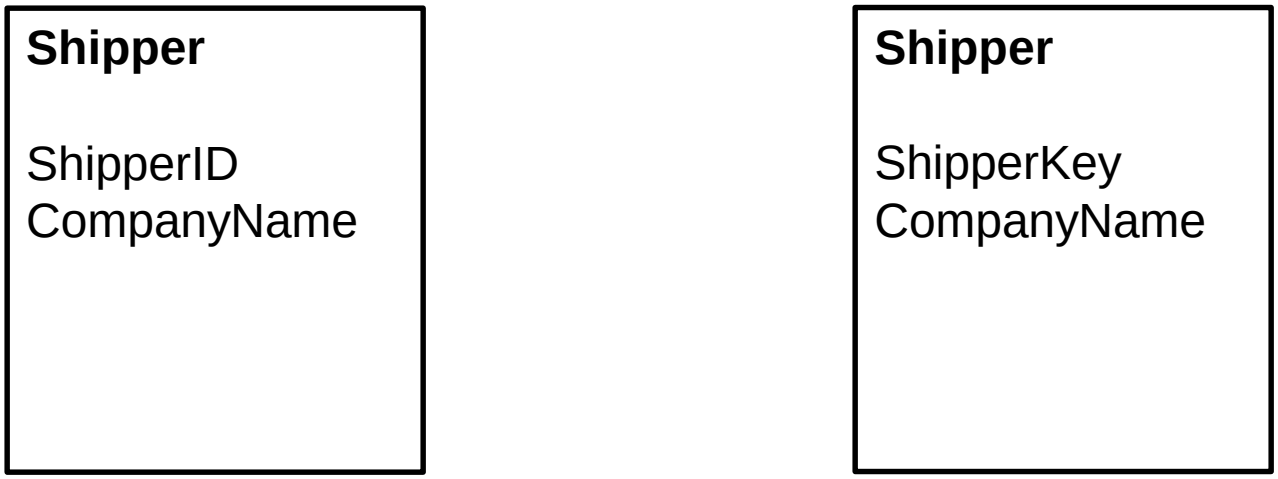
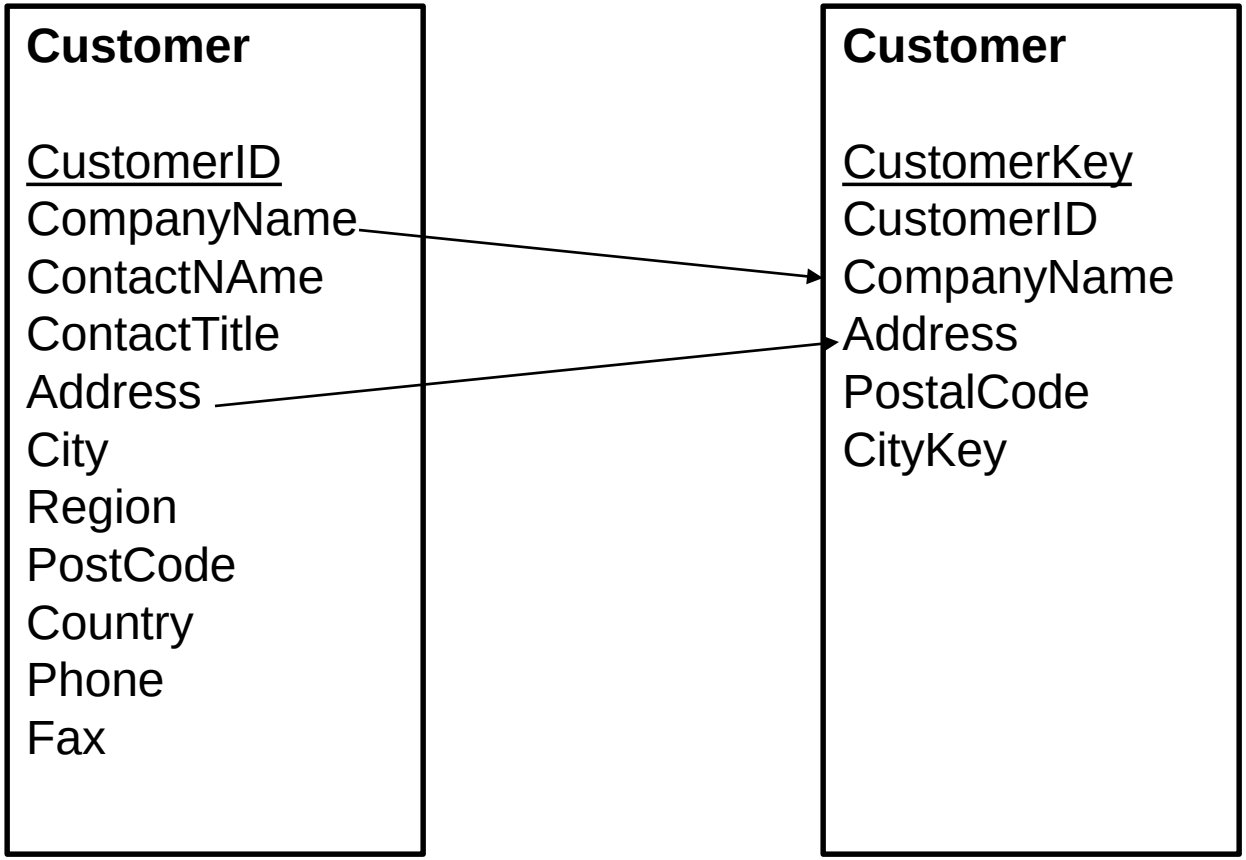

1 Shippers
ShipperID
CompanyName
Phone
2 Customers
CustomerID
CompanyName
ContactName
ContactTitle
Address
City
Region (0,1)
PostalCode (0,1)
Country
Phone
Fax (0,1)
3 Suppliers
SupplierID
CompanyName
ContactName
ContactTitle
Address
City
Region (0,1)
PostalCode
Country
Phone
Fax (0,1)
Homepage (0,1)
Categories
CategoryID
CategoryName
Description
Picture
Orders
OrderID
CustomerID
EmployeeID
OrderDate
RequiredDate
ShippedDate (0,1)
ShipVia
Freight
ShipName
ShipAddress
ShipCity
ShipRegion (0,1)
ShipPostalCode (0,1)
ShipCountry
OrderDetails
OrderID
ProductID
UnitPrice
Quantity
Discount
4 Products
ProductID
ProductName
QuantityPerUnit
UnitPrice
UnitsInStock
UnitsOnOrder
ReorderLevel
Discontinued
SupplierID
CategoryID
5 Employees
EmployeeID
FirstName
LastName
Title
TitleOfCourtesy
BirthDate
HireDate
Address
City
Region (0,1)
PostalCode
Country
HomePhone
Extension
Photo (0,1)
Notes (0,1)
PhotoPath (0,1)
ReportsTo (0,1)
Employee Territories
EmployeeID
TerritoryID
Territories
TerritoryID
TerritoryDescription
RegionID
Regions
RegionID
RegionDescription
4 Product
ProductKey
ProductName
QuantityPerUnit
UnitPrice
Discontinued
CategoryKey
Category
CategoryKey
CategoryName
Description
Time
TimeKey
Date
DayNbWeek
DayNameWeek
DayNbMonth
DayNbYear
WeekNbYear
MonthNumber
MonthName
Quarter
Semester
Year
AK: Date
3 Supplier
SupplierKey
CompanyName
Address
PostalCode
CityKey
Sales
CustomerKey
EmployeeKey
OrderDateKey
DueDateKey
ShippedDateKey
ShipperKey
ProductKey
SupplierKey
OrderNo
OrderLineNo
UnitPrice
Quantity
Discount
SalesAmount
Freight
AK: (OrderNo, OrderLineNo)
1 Shipper
ShipperKey
CompanyName
2 Customer
CustomerKey
CustomerID
CompanyName
Address
PostalCode
CityKey
AK: CustomerID
Territories
EmployeeKey
CityKey
5 Employee
EmployeeKey
FirstName
LastName
Title
BirthDate
HireDate
Address
City
Region
PostalCode
Country
SupervisorKey
City
CityKey
CityName
StateKey (0,1)
CountryKey (0,1)
State
StateKey
StateName
EnglishStateName
StateType
StateCode
StateCapital
RegionName (0,1)
RegionCode (0,1)
CountryKey
Country
CountryKey
CountryName
CountryCode
CountryCapital
Population
Subdivision
ContinentKey
Continent
ContinentKey
ContinentName

Table level mapping

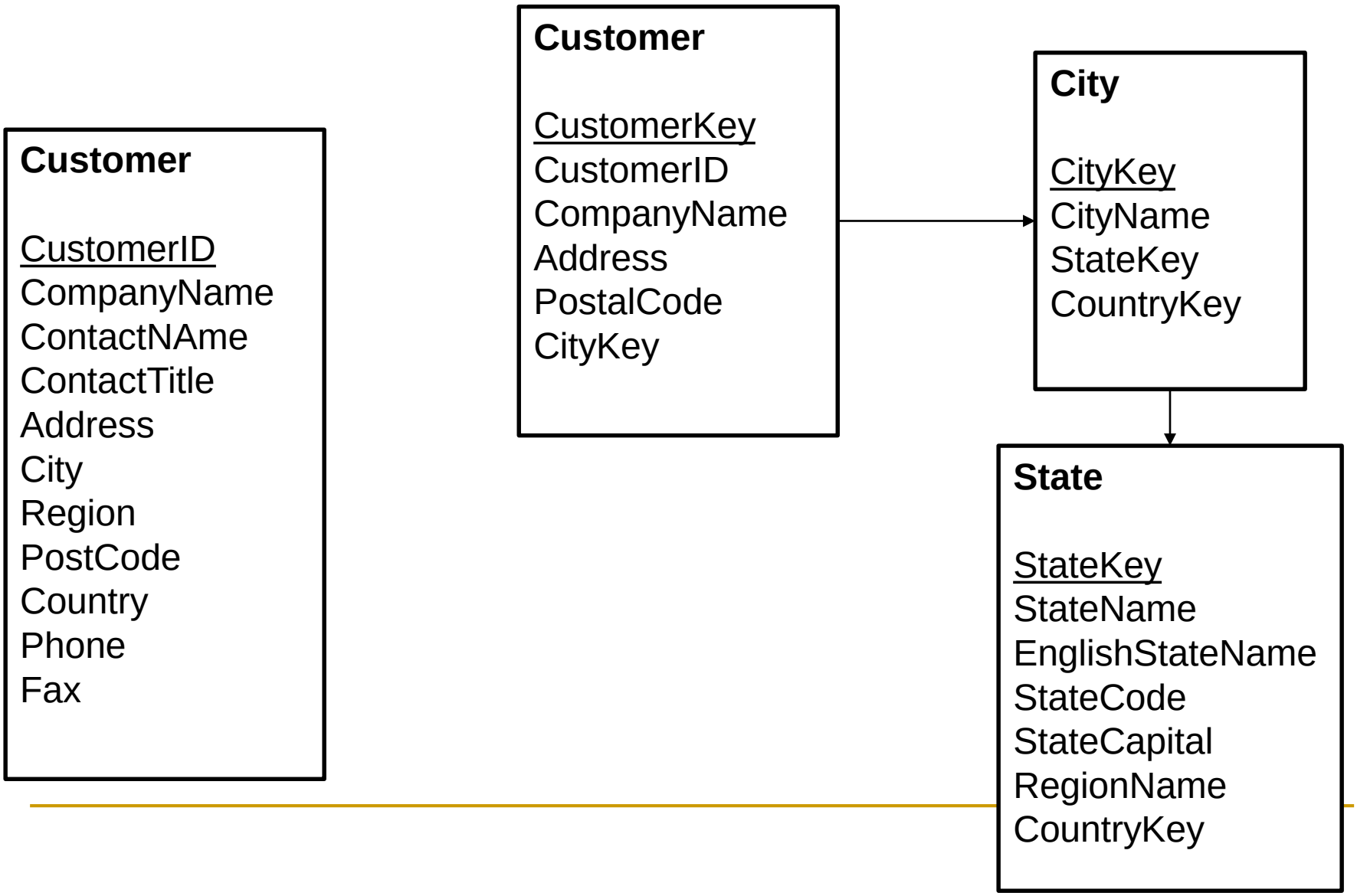
Shipper - Shipper



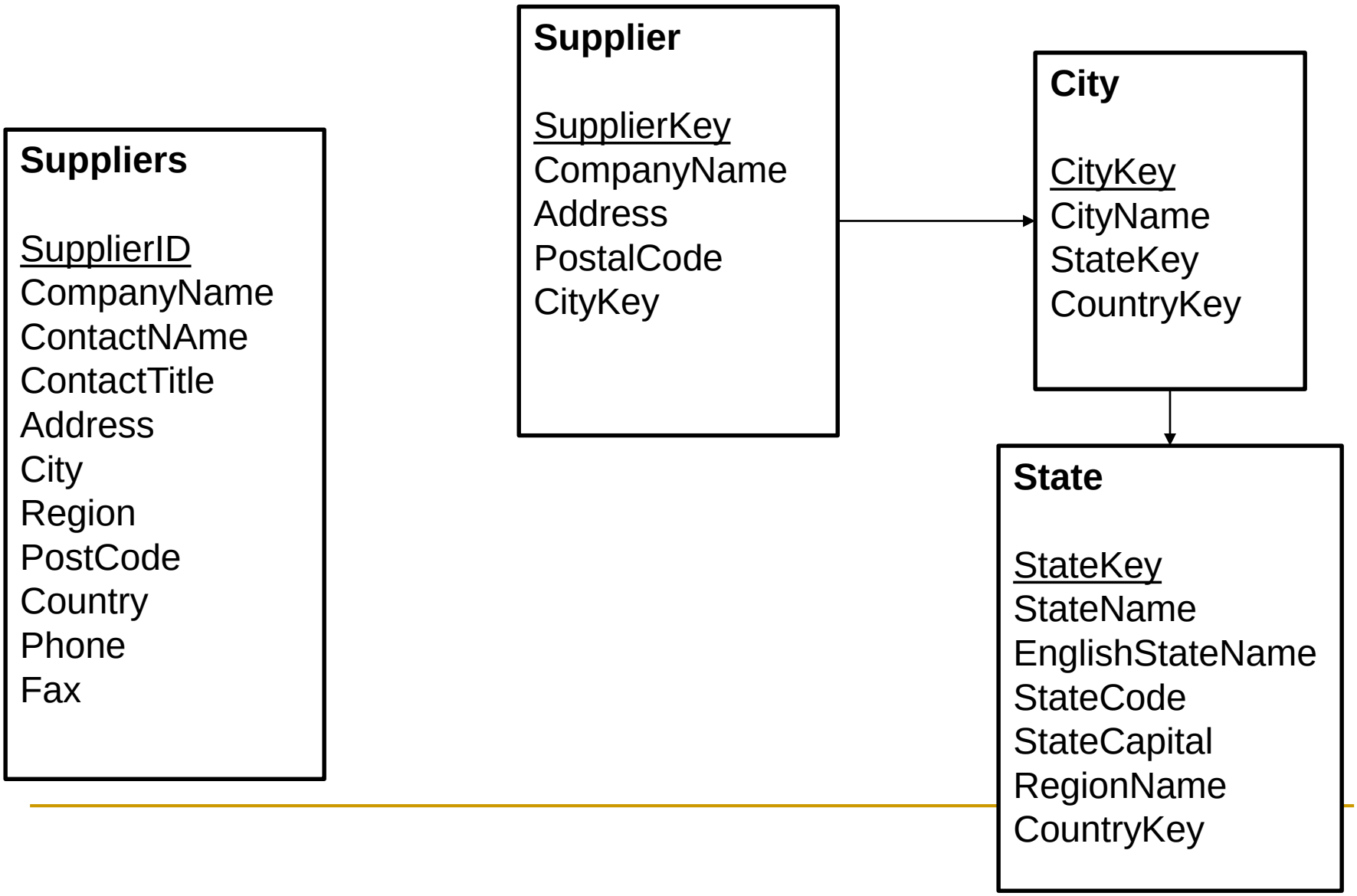
Customer - Customer



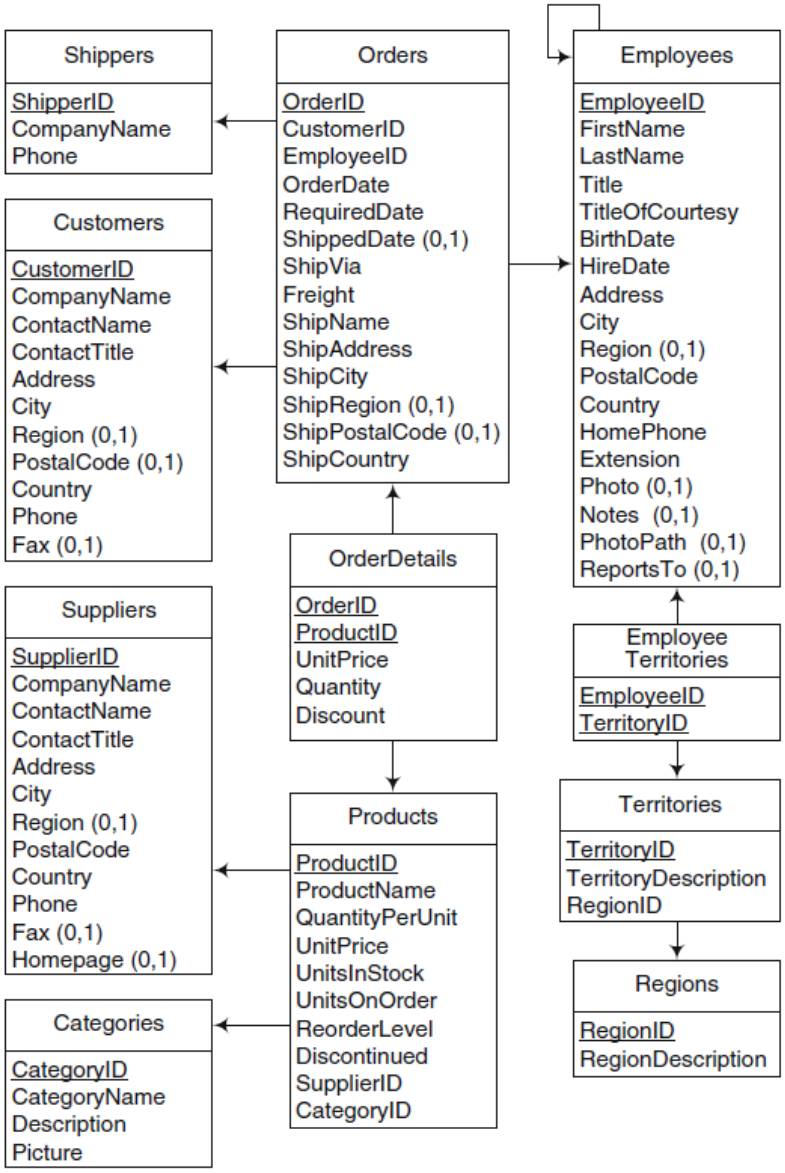
Customer - Customer



Supplier - Supplier



Fact Table



Sales

CustomerKey
EmployeeKey
OrderDateKey
Duedatekey
Shippeddatekey
Shipperkey
OrderNo
OrderLine No
UnitPrice
Quantity
Discount
SalesAmount
Freight

ETL function

- Re-shape relevant data from source systems into useful information for DW.
- Is a combination of
  - Extraction
  - Transformation
  - Loading
- Most important and challenging
  - Because of the diverse and heterogeneous nature of source systems
  - Up to 80% of the total effort of BI implementation

Why ETL is challenging?

- Reasons
  - Source systems are diverse and disparate
  - Source systems are on multiple platforms, different operating systems
  - Many sources are older legacy applications on obsolete database technologies
  - Typically, historical data changes are not preserved in sources
  - Dubious data quality
  - Sources keep changing overtime

ETL function

- Reasons (cont.)
  - Gross lack of consistency among sources systems in common
    - Same data is presented differently in various systems
  - Lack of means to resolving mismatches escalates the issue
  - Source systems may not represent data in types or formats that are meaningful (unclear attributes)

CMS Case Study

Analytical Questions

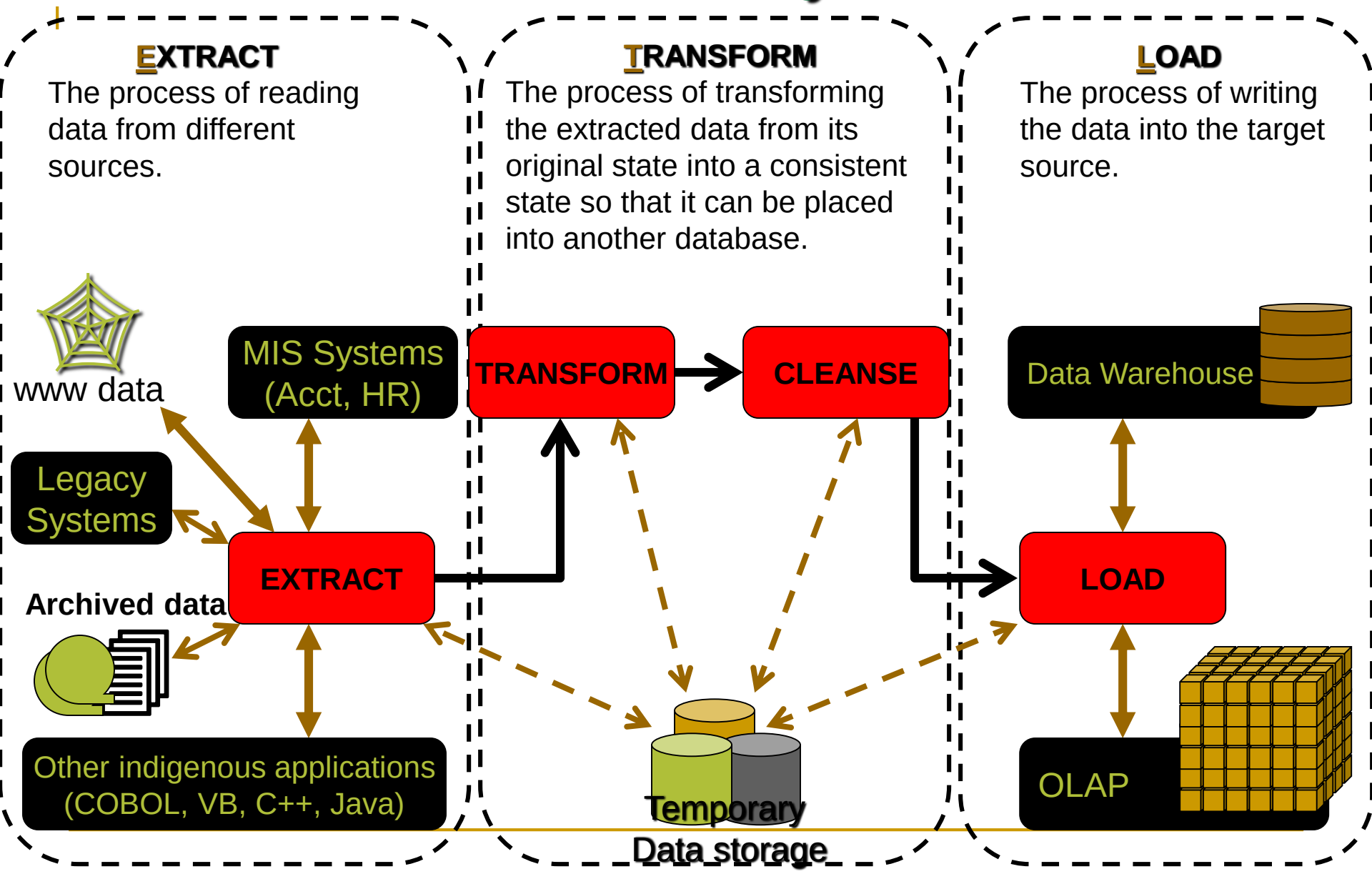
- What is the average CGPA of Class?
- What is the average marks given by a Regular Faculty and Visiting Faculty?
- What is the grades comparison of faculty with respect to course load?
- What is the average grades if semester Calendar is of 14 weeks vs 16 weeks?

The ETL Cycle

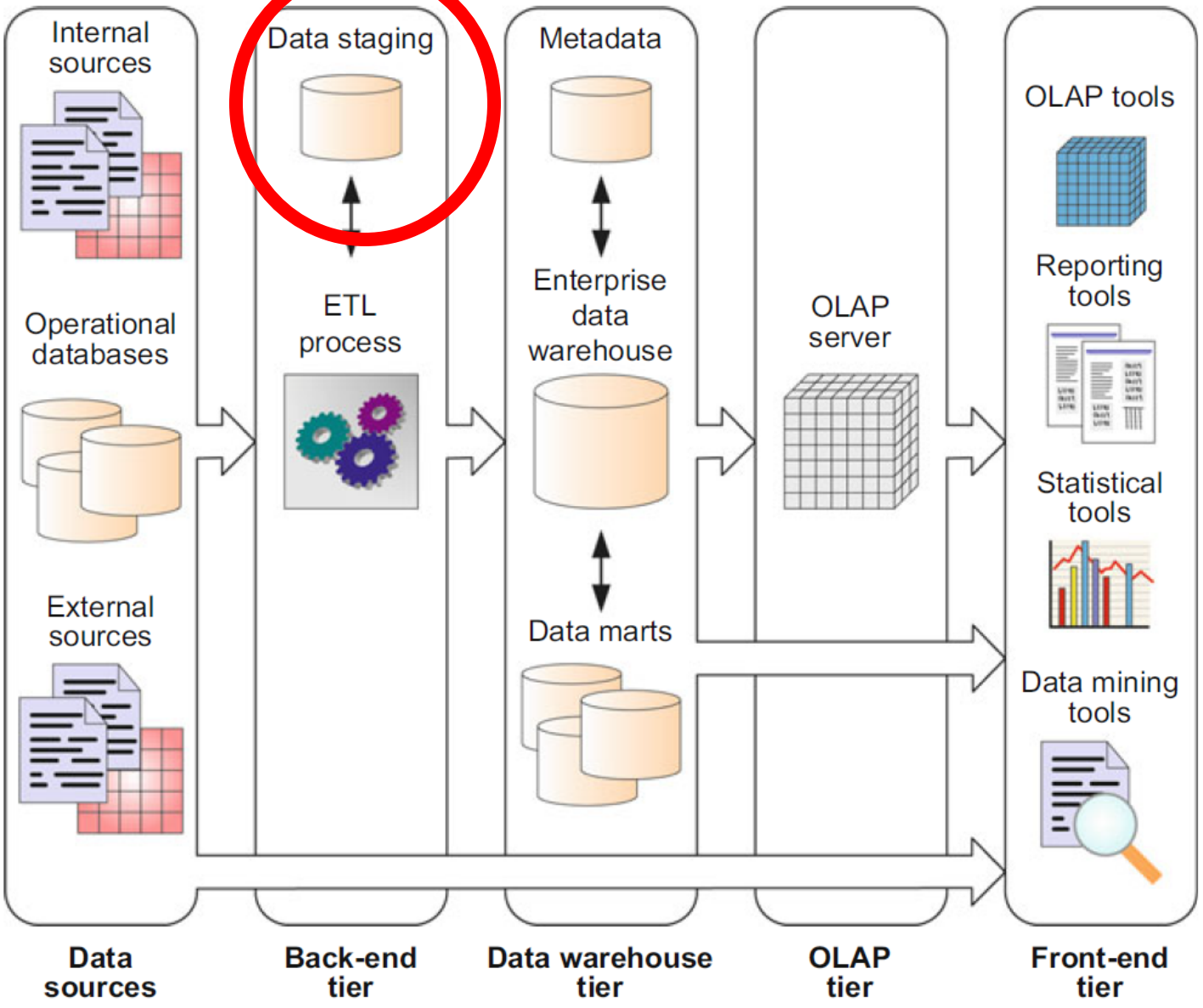
EXTRACT
The process of reading data from different sources.

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.



Typical DW architecture

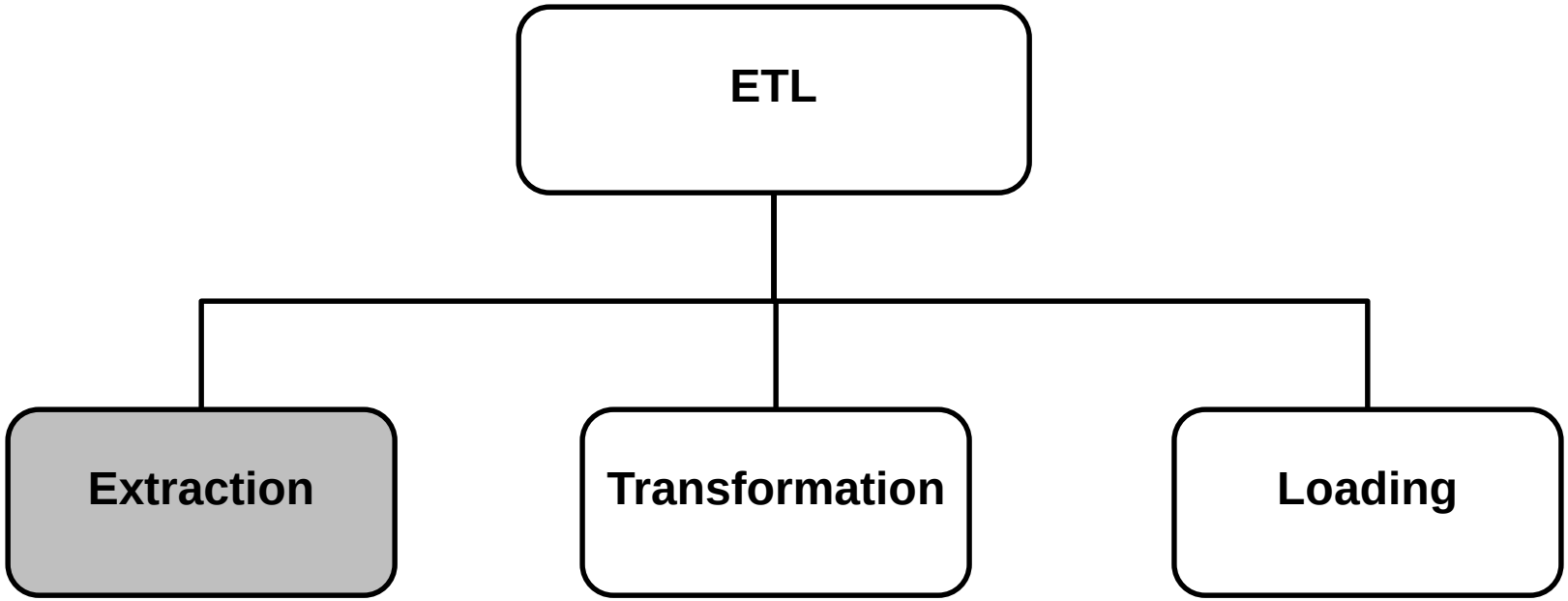


Data Stagging, Staging area, landing zone

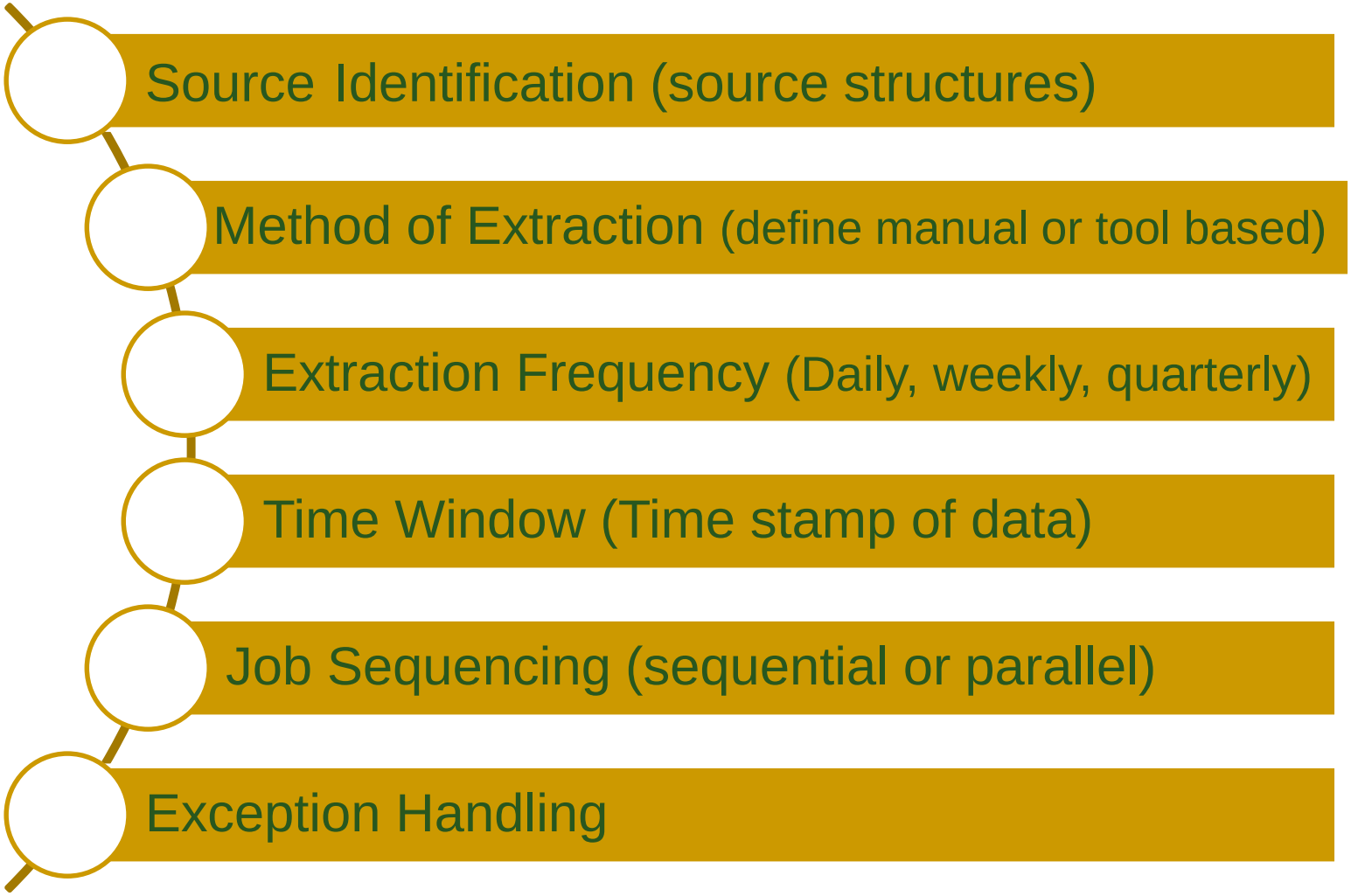
- An intermediate storage area used for **raw data**
- Between data source (OLTP) and data target (DW)
- Temporary storage
- Implementation
  - Various formats, relational databases, JSON, XML format, spread sheets, etc.
  - Not in Star format
- Can be in more than one format

Reasons for Data Stagging

- Reasons
  - Data consolidation
  - Enhanced data quality
  - Schema alignment
  - Data Integration
  - Performance efficiency
  - Data Security
    - Compliance with privacy and security protocols
  - Audit and tracking
    - Comparing input and the output files


ETL
Extraction
Transformation
Loading

Data Extraction **High-level Issues**



Data Acquisition (specific issues)

- Diversity in source systems and platforms

Platform	OS	DBMS	MIS/ERP
Main Frame	VMS	Oracle	SAP
Mini Computer	Unix	Informix	PeopleSoft
Desktop	Win NT	Access	JD Edwards
	DOS	Text file	

  - Dozens of source systems across organizations
  - Numerous systems within an organization
  - Need specialist for each

Data Acquisition (specific issues)

Same data, different representation

Date value representations

Examples:

970314
03/14/1997
March 14 1997
1997-03-14
14-MAR-1997
2450521.5 (Julian date format)

Gender value representations

Examples:

- Male/Female
- 0/1
- M/F
- PM/PF

Data Acquisition (specific issues)

- Car number plats

Data Acquisition (specific issues)

- Multiple sources for the **same data elements**
  - Need to establish **precedence** between source systems on a per data element basis.
  - Take data element from source system with highest precedence where element exists
  - Must sometimes establish “group precedence” rules to maintain data integrity
  - Guessing gender from name
  - First, middle and family name from two systems of different rank. People using middle name as first name

Data Acquisition (specific issues)

- Rigidity and unavailability of legacy systems
  - Need efficient and easy ways to deal with incompatible formats
  - Need specialized / outdated skills
  - Absence of metadata or documentation of systems

Data Acquisition (specific issues)

- Volume of legacy data
  - Talking about not weekly data, but data spread over years
  - Historical data on tapes that are serial and very slow to mount etc.
  - Need lots of processing and I/O to effectively handle large data volumes.
  - Need efficient interconnect bandwidth to transfer large amounts of data from legacy sources to DWH.

Data Acquisition (specific issues)

- Web scrapping
  - Lot of data in a web page but mixed with junk
  - Problems
    - Limited query interfaces
    - Free text fields
    - Rapid changes without notice

Types of Data Extraction

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

Logical data extraction

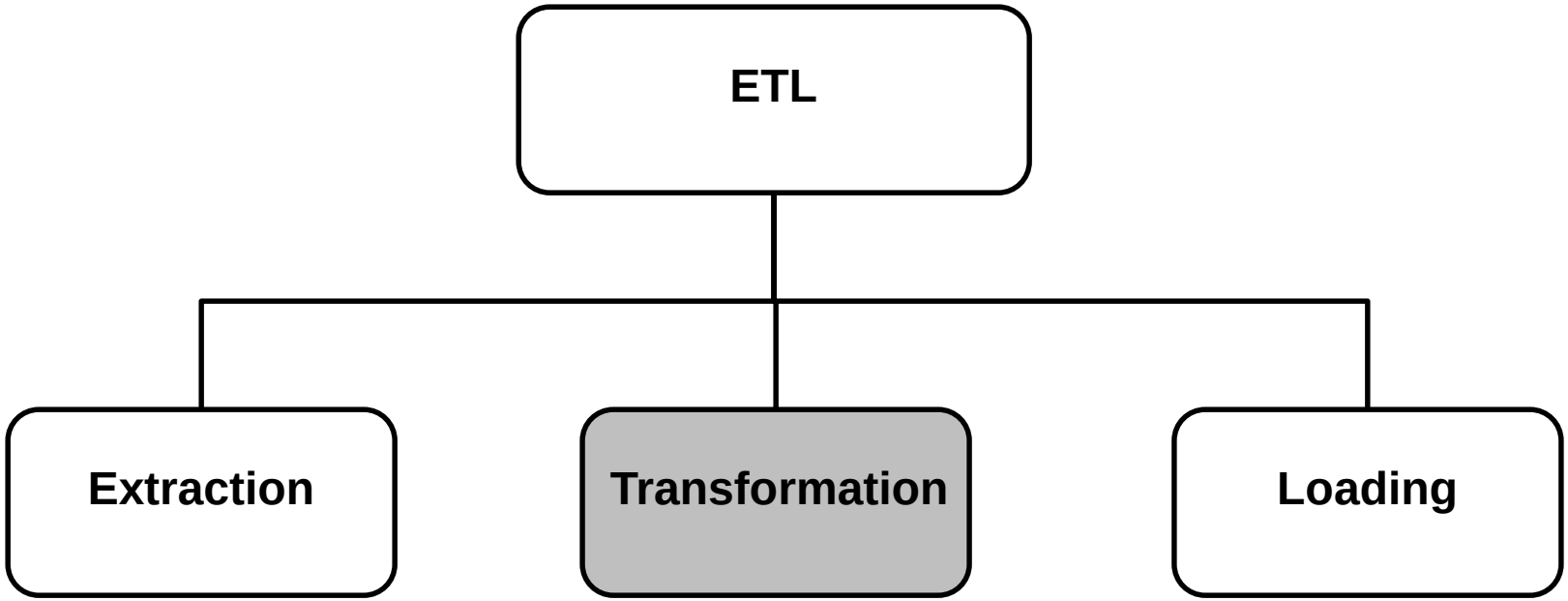
- Full Extraction
  - The data extracted completely from the source
  - No need to keep track of changes.
  - Source data made available as-is with any additional information
- Incremental Extraction
  - Data extracted after a well-defined point/event
  - Mechanism used to reflect/record the temporal changes in data

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

Physical data extraction

- 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


ETL
Extraction
Transformation
Loading

Where Transformation is required?

- Where is Data Transformation required?
  - **Migrate** legacy systems to modern applications
  - Translate from one **data model** to another
  - Integrate **heterogeneous systems**
  - Achieve enterprise-wide **integration**

1. Data Transformation (Basic Tasks)

- Basic transformation: basic tasks
  - Selection (part of Source)
  - Decoding of fields
  - Splitting/joining of fields
  - Conversion (like date format, units)
  - Summarization (aggregates)
  - Enrichment (adding new data or modifying existing data)

2. Major Transformation Types

- Major transformation types
  - Format revisions (data type, field length)
  - Decoding of fields
  - Calculated & derived values
  - Splitting of single field
  - Merging of information
    - Not really means combining columns to create one column
    - Info for product coming from different sources merging it into single entity

2. Major Transformation Types

- Conversion of units of measurements
  - For companies with global branches Km vs. mile or lb vs Kg
- Data and time conversion
  - June 4, 2005 as 6/4/2005 in US and 4/6/2005 in the British format.
  - This date may be standardized to be written as 14 NOV 2005.
- Key restricting
- Duplication

2. Major Transformation Types

- Key restricting
  - 92424979234 changed to 12345678
  - +92-334-4633353, 00923344633353, PK(+92)-(0)334-

92	42	4979	234
Country_Code	**City_Code**	**Post_Code**	**Product_Code**

- ****92**334*4633353
- +92-334-4633353,
- 00923344633353,
- PK(+92)-(0)334-4633353

3. Data Transformation (integration)

- Data integration and consolidation (major challenges)
  - Entity identification problem
  - Multiple sources problem
  - Transformation for dimension attributes
  - How to implement transformation
    - Using transformation tools
    - Using manual techniques
- BSDSF21A011 Fatima Khurram
- MSDSF26M001 Fatima Hassan

Data Transformation Tasks: 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 → FLAT No. 2
FL/NO.2
FL-2
FLAT-2
FLAT#
FLAT,2
FLAT-NO-2
FL-NO.2

Data Transformation Tasks: Example-2

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

Data Transformation : 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.

Input Data

Name: HAJI MUHAMMAD IBRAHIM,
Locaiton: GOVT. CONT.
Company:K. S. ABDULLAH & BROTHERS,
MAMOOJI ROAD, ABDULLAH MANZIL
City:RAWALPINDI,
Ph 67855

Data content defects

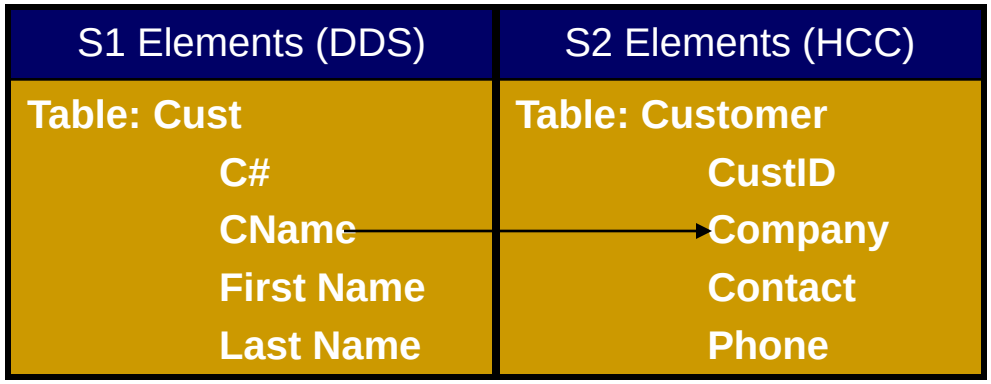
- Domain value redundancy
  - Unit of Measure (Dozen, Doz., Dz., 12)
- Non-standard data formats
  - Phone Numbers (1234567 or 123.456.7)
- Non-atomic data fields
  - Names and addresses
    - Muhammad Khurram Shahzad, PhD
    - Dr. Muhammad Khurram Shahzad
- Embedded meaning
  - RC, AP, RJ
    - Received, approved, rejected

Schema Mapping for ETL

Schema Integration and Mapping

- Schema Integration & Matching
- Fundamental schema matching operator is *'MATCH'*
  - *Input:* Multiple sources
  - *Output:* Mappings
- A mapping is a set of *mapping elements*, each of which indicates that certain elements of a schema, say S1, are mapped to certain elements in the other schema, say S2.
- Each mapping element can have a *mapping expression* which specifies how the S1 and S2 elements are related.

Mapping example



Customer
CKEY
CompanyName
Phone

- Mapping Example
  - Mapping element relating Cust.C# to Customer.CustID
  - Mapping expression Cust.C# = Customer.CustID
- Match operation is a function that takes two schemas S1 and S2 as input and returns a mapping between those two schemas, called the *match result*.

Schema level approach

1. **Problems**
2. Granularity of match
3. Match cardinality
4. **Solutions**
5. Linguistic approaches
6. Constraint-based approaches

1. Granularity of match

- Element-level matching
  - Determines the matching elements in the second input schema.
  - In the simplest case, only elements at the finest level of granularity are considered.
  - e.g. Address.ZIP = CustomerAddress.PostalCode

- Structure-level Matching
  - Matching combinations of elements that appear together in a structure.
  - In the ideal case, all components of the structures in the two schemas fully match.
  - Alternatively, only some of the components may be required to match (i.e., a partial structural match).

1. Granularity of match

S1 elements	S2 elements	
Address Street City State Zip	CustomerAddress Street City USState PostalCode	Full structure match of Address and CustomerAddress
AccountOwner Name Address Birthdate TaxExempt	Customer Cname CAddress Cphone	Partial structural match of AccountOwner and Customer

2. Match cardinality

Local match cardinalities	S1 element(s)	S2 element(s)	Matching expression
1:1	Price	Amount	Amount = Price
n:1	Price, Tax	Cost	Cost = Price * (1 + Tax/100)
1:n,	Name	FirstName, LastName	FirstName, LastName = Extract(Name, …)
n:1, structure-level (n:m element-level)	B.Title, B.PuNo, P.PuNo, P.Name	A.Book, A.Publisher	A.Book, A.Publisher = select B.Title, P.Name from B, P where B.PuNo = P.PuNo

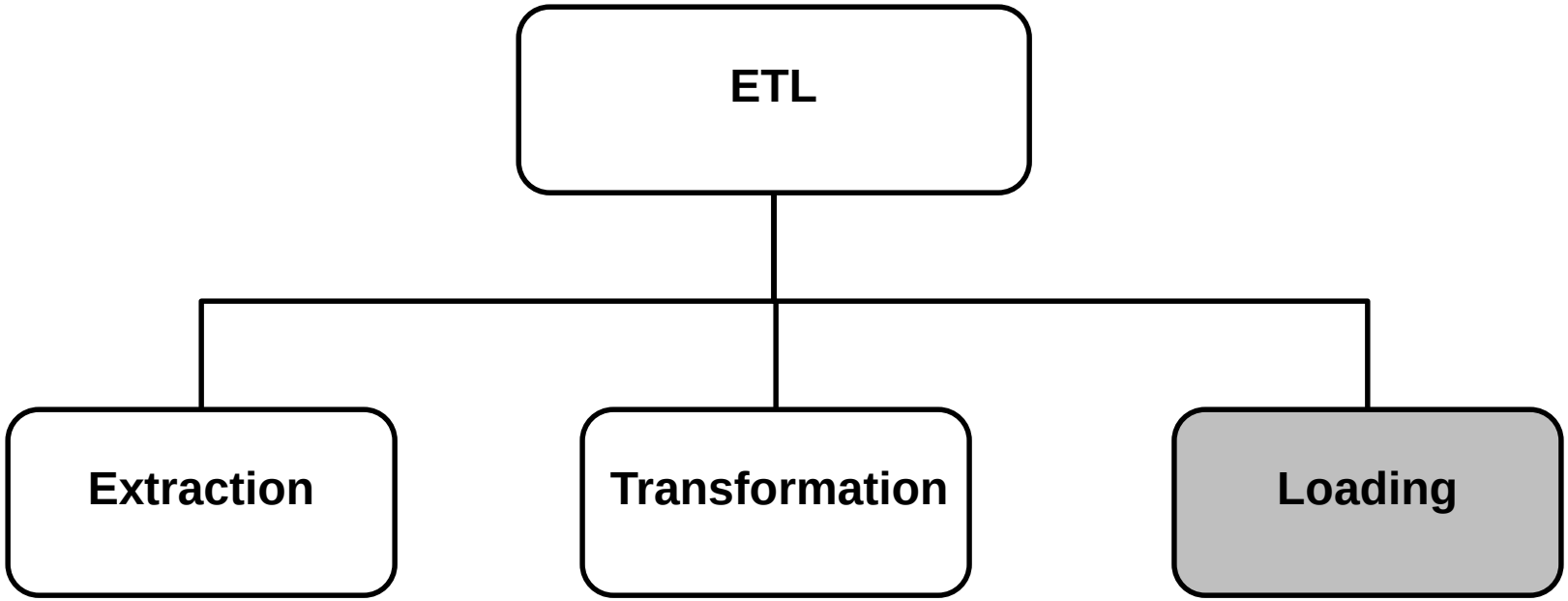
3. Linguistic approach

- Use **names and text** (i.e., words or sentences) to **find semantically similar schema elements**

- **Name Matching**
  - Equality of names (be aware of Homonyms)
  - Equality of canonical name representations (e.g., CName → customer name, and
    EmpNO → employee number)
  - Equality of synonyms
  - Equality of hypernyms (E.g., book *is-a* publication and article *is-a* publication imply book=publication, article=publication, and book = article)

3. Linguistic approach

- **Description Matching**
  - Schemas contain comments in natural language to express the intended semantics of schema elements.
  - Ex. S1: empn //**employee name**
  - Ex. S2: name //**name of employee**
  - Analysis could be as simple as extracting keywords from the description
  - Or it could be as sophisticated as using natural language understanding technology to look for semantically equivalent expressions.


ETL
Extraction
Transformation
Loading

Data Loading

- Two types, **initial** loading and **subsequent** refresh
- There are three update loading strategies, for subsequent loading
  - **Trickle/continuous** feed with constant data collection using row level insert and update operations
  - **Incremental load** – applying ongoing changes as necessary in periodic manner
    - Block slamming into existing populated tables
    - Trickle feed with continuous data acquisition using row level insert and update operations
  - **Full refresh** – completely erasing content of one or more tables and reloading with full refresh

Full refresh strategy

- Completely re-load data on each refresh
- Strategy for small tables when large rows are changed on each refresh
- Remove referential integrity constraints
- Consider using shadow table to allow refresh to take place without impacting query workloads

Loading strategies

- Choice in loading strategy depends on tradeoffs in
  - Data freshness, performance
  - And, data volatility characteristics.
  - Define goal
    - Increased data freshness
    - Increased data loading performance

Data Cleansing

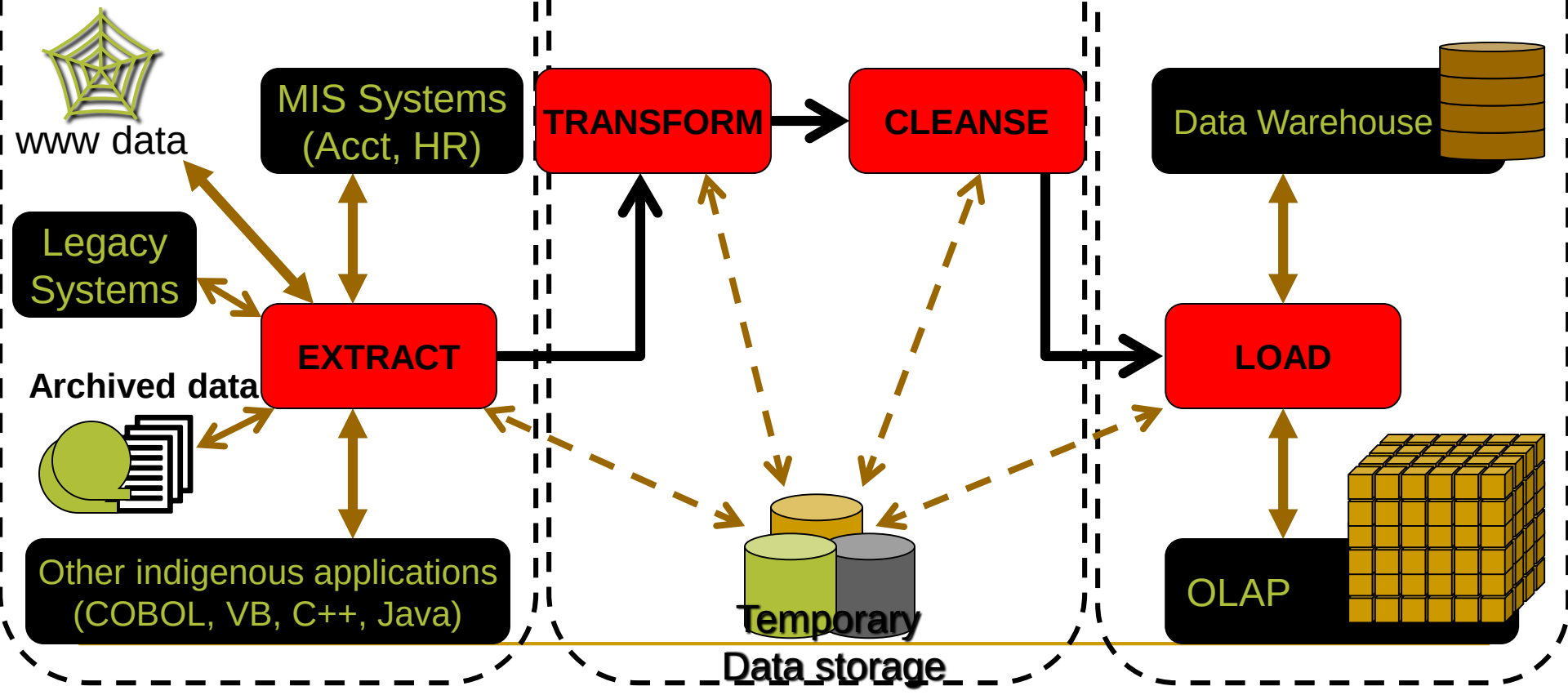
An important part of ETL

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.



Reasons for Dirty data

- Dummy Values
- Absence of Data
- Multipurpose Fields
- Contradicting Data
- Inappropriate Use of Address Lines
- Violation of Business Rules
- Reused Primary Keys
  - Reusing PK after deleting data from previous records
- Non-Unique Identifiers
- Data Integration Problems

Three types of anomalies: Requires cleaning

- Syntactically Dirty Data
  - Lexical Errors
  - Irregularities
- Semantically Dirty Data
  - Integrity Constraint Violation
  - Business rule contradiction
  - Duplication
- Coverage Anomalies
  - Missing Attributes
  - Missing Records

1. Syntactically dirty data

- Lexical Errors
  - Discrepancies between the structure of the data items and the specified format of stored values
    - e.g. number of columns used are unexpected for a tuple (mixed up number of attributes)
- Irregularities
  - Non-uniform use of units and values, such as only giving annual salary but without info i.e. in US$ or PK Rs?

2. Semantically dirty data

- Integrity Constraint Violation
- Business rule contradiction
- Data Duplication
  - Data duplication can result in costly errors, incorrect facts

3. Coverage anomalies

- Missing Attribute
  - Result of omissions while collecting the data.
- Missing record
  - A constraint violation if we have null values for attributes where NOT NULL constraint exists.
  - Equipment malfunction (bar code reader, keyboard)
  - Inconsistent with other recorded data
  - Data not entered due to misunderstanding
  - Data not considered important at entry (2002 to 02)

Handling missing records

- Dropping records
- Manually filling missing values
- Using attribute mean, or median as filler
- Filling global constant as filer
- Using most probable values as filler

Key based classification of problems

- Primary key problems
- Non-primary key problems

Primary key problems

- Same primary key but different data
- Same entity with different key
- Primary Key in one system but not in the other
- Same Primary Key but in different formats

Data Duplication: Non-Unique PK

- Duplicate Identification Numbers

  - Multiple Customer Numbers

Name	Phone Number	Cust. No.
M. Ismail Siddiqi	*021.666.1244*	*780701*
M. Ismail Siddiqi	*021.666.1244*	*780203*
M. Ismail Siddiqi	*021.666.1244*	*780009*

  - Multiple Employee Numbers

Bonus Date	Name	Department	Emp. No.
Jan. 2000	*Khan Muhammad*	*213 (MKT)*	*5353536*
Dec. 2001	*Khan Muhammad*	*567 (SLS)*	*4577833*
Mar. 2002	*Khan Muhammad*	*349 (HR)*	*3457642*

Non-primary key problems

- Required fields left blank
- Data erroneous or incomplete
- Data contains null values

Data Duplication: House Holding

- Group together all records that belong to the same household.

.....	S. Ahad	440, Munir Road, Lahore
.....	.....	.....
.....	Shiekh Ahad	No. 440, Munir Rd, Lhr
.....	.....	.....
.....	Shiekh Ahed	House # 440, Munir Road, Lahore

Why bother ?

Data Duplication: Individualization

- Identify multiple records in each household which represent the same individual

.....	M. Ahad	440, Munir Road, Lahore
.....	.....	.....
.....	Maj Ahad	440, Munir Road, Lahore

Address field is standardized.

By coincidence ??

Steps Data Cleansing

- Parsing
- Correcting
- Standardizing
- Matching
- Consolidating

Steps Data Cleansing

- Parsing
  - Locating and identifying individual data elements from sources
- Correcting
  - Correcting individual data using algorithms and secondary sources
- Standardizing
  - Applying conversion routines to transform data into custom business rules

Steps Data Cleansing

- Matching
  - Searching and matching records within and across parsed, corrected and standardized data
  - Means, eliminating duplicates
- Consolidating
  - Analyzing and identifying relationships between matches records and merging into one representation

Automatic Data Cleansing

- Statistical approach
- Pattern based approach
- Clustering approach
- Associate rules-based approach

Automatic Data Cleansing…

- Statistical Methods
  - Identifying outlier fields and records using the values of mean, standard deviation, range, etc.
- Pattern-based
  - Identify outlier fields and records that do not conform to existing patterns in the data.
  - A pattern is defined by a group of records that have similar characteristics (“behavior”) for p% of the fields in the data set, where p is a user-defined value (usually above 90).
  - Techniques such as partitioning, classification, and clustering can be used to identify patterns that apply to most records.

Automatic Data Cleansing

- Clustering
  - Identify outlier records using clustering based on Euclidian (or other) distance.
  - Clustering the entire record space can reveal outliers that are not identified at the field level inspection
  - Main drawback of this method is computational time.
- Association rules
  - Association rules with high confidence and support define a different kind of pattern.
  - Records that do not follow these rules are considered outliers.

Formal definition & Nomenclature

- Problem statement:
  - “Given two databases, identify the potentially matched records Efficiently and Effectively”
- Many names, such as:
  - Record linkage
  - Merge/purge
  - Entity reconciliation
  - List washing and data cleansing.

ETL vs ELT

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

ELT

- First, load ‘raw’ data into empty tables using RDBMS
- Next, use SQL/python to transform the ‘raw’ data into a form appropriate to target table
- Ideally, the SQL is d\generated using a meta data driven tool rather than hand coding
- Finally, use insert-select into the target table for incremental loads or view switching if a full refresh strategy is used.

Operational sources 1,2

Customer

CustID	Fname	Lname	Title
C11	Anna	Jennings	Ms
C12	Robert	Walters	Mr
C13	Lina	Skoglund	Ms
C14	Tobias	Almquist	Mr

Orders

OrderID	CustID	ServID	Date	Time	Amount
1221	C11	S1	24/04/2005	10:33	9
1222	C12	S2	24/04/2005	11:48	13
1223	C13	S1	24/04/2005	12:24	24
1224	C11	S2	25/04/2005	10:13	15
1225	C12	S1	25/04/2005	11:23	32
1226	C13	S2	25/04/2005	11:23	17
1227	C14	S1	25/04/2005	12:23	21

Service

ServID	Stype
S1	Local call
S2	Mobile call

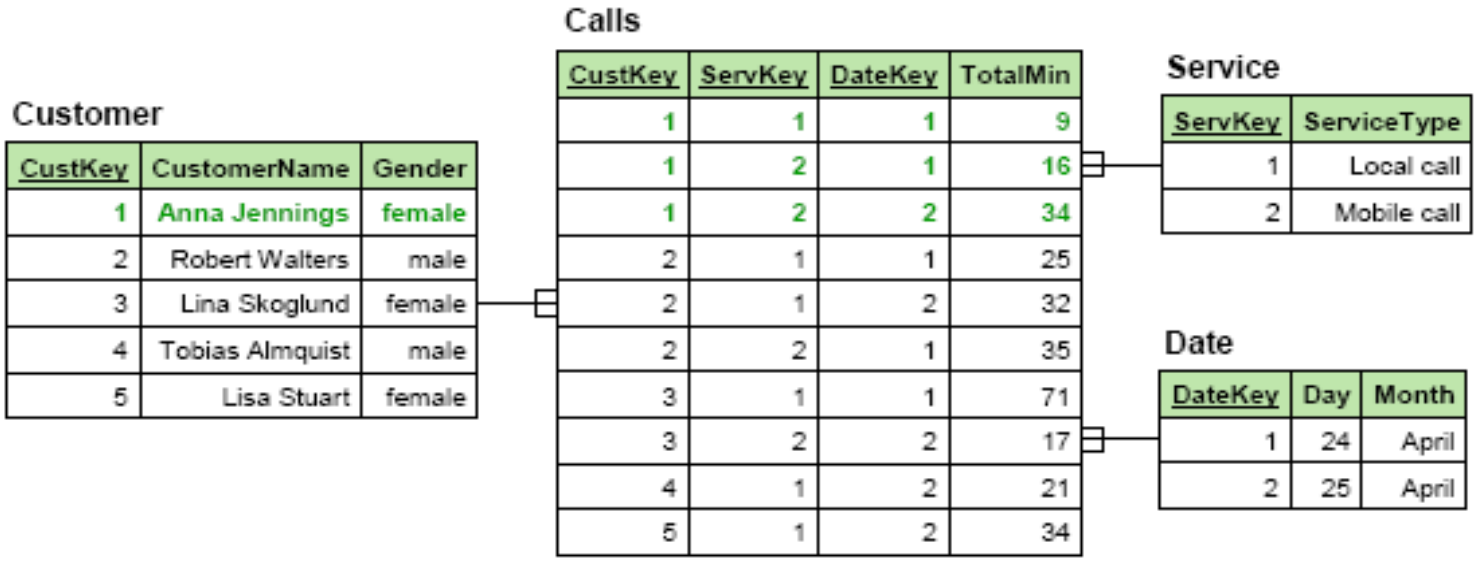
Person

PNR	PNAME	F/M
P134	SKOGLUND, LINA	F
P135	WALTERS, ROBERT	M
P136	JENNINGS, ANNA	F
P137	STUART, LISA	F

Calls

CNR	PNR	CTYPE	CDATE	CTIME	CMIN
1567	P134	LC1	050424	0917	18
1568	P135	LC2	050424	1022	25
1569	P136	MC2	050424	1134	16
1570	P134	LC2	050424	1149	29
1571	P135	MC1	050424	1234	22
1572	P136	MC2	050425	1123	19
1573	P137	LC1	050425	1246	34

Dimensional model ex.



Customer

CustKey	CustomerName	Gender
1	Anna Jennings	female
2	Robert Walters	male
3	Lina Skoglund	female
4	Tobias Almquist	male
5	Lisa Stuart	female

Calls

CustKey	ServKey	DateKey	TotalMin
1	1	1	9
1	2	1	16
1	2	2	34
2	1	1	25
2	1	2	32
2	2	1	35
3	1	1	71
3	2	2	17
4	1	2	21
5	1	2	34

Service

ServKey	ServiceType
1	Local call
2	Mobile call

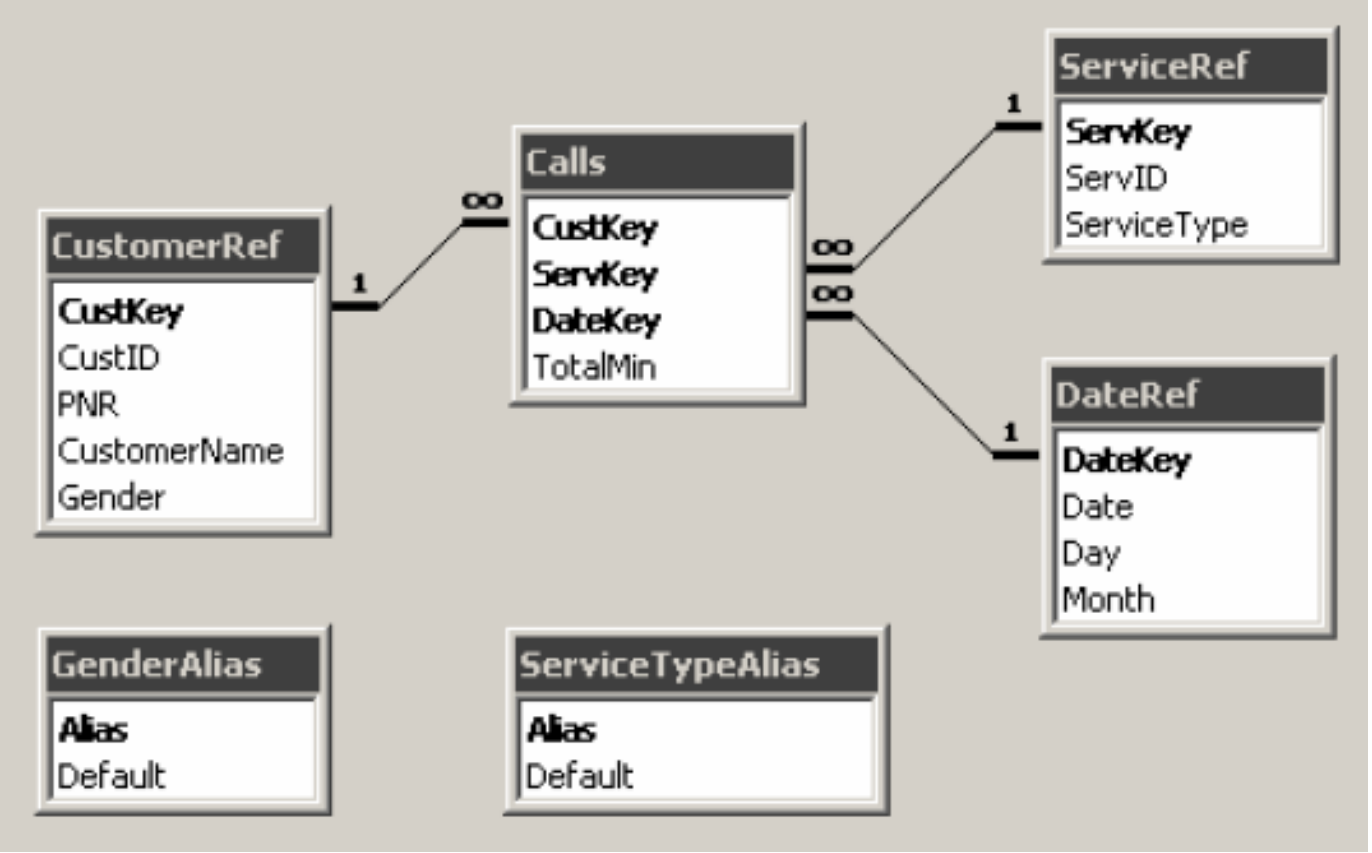
Date

DateKey	Day	Month
1	24	April
2	25	April

Requirements:

1) the identical objects in the two operational databases should be matched and assigned the same surrogate key in the DW database

2) all attributes required for record matching should be contained in the data staging area

3) no changes can be done in the operational databases to facilitate the ETL process

Staging area