CURRENT TOPICS IN COMPUTER SCIENCE

Business Intelligence Systems and Analytics
ETL AND DATA INTEGRAITON CHALLENGES

Trong Nhan Phan, PhD

OUTLINE

- ETL overview
- Data integration
- Data quality
- SSIS Demo
- Summary
- References

DATA WAREHOUSE CONCEPT AND ARCHITECTURE

DWH ARCHITECTURE

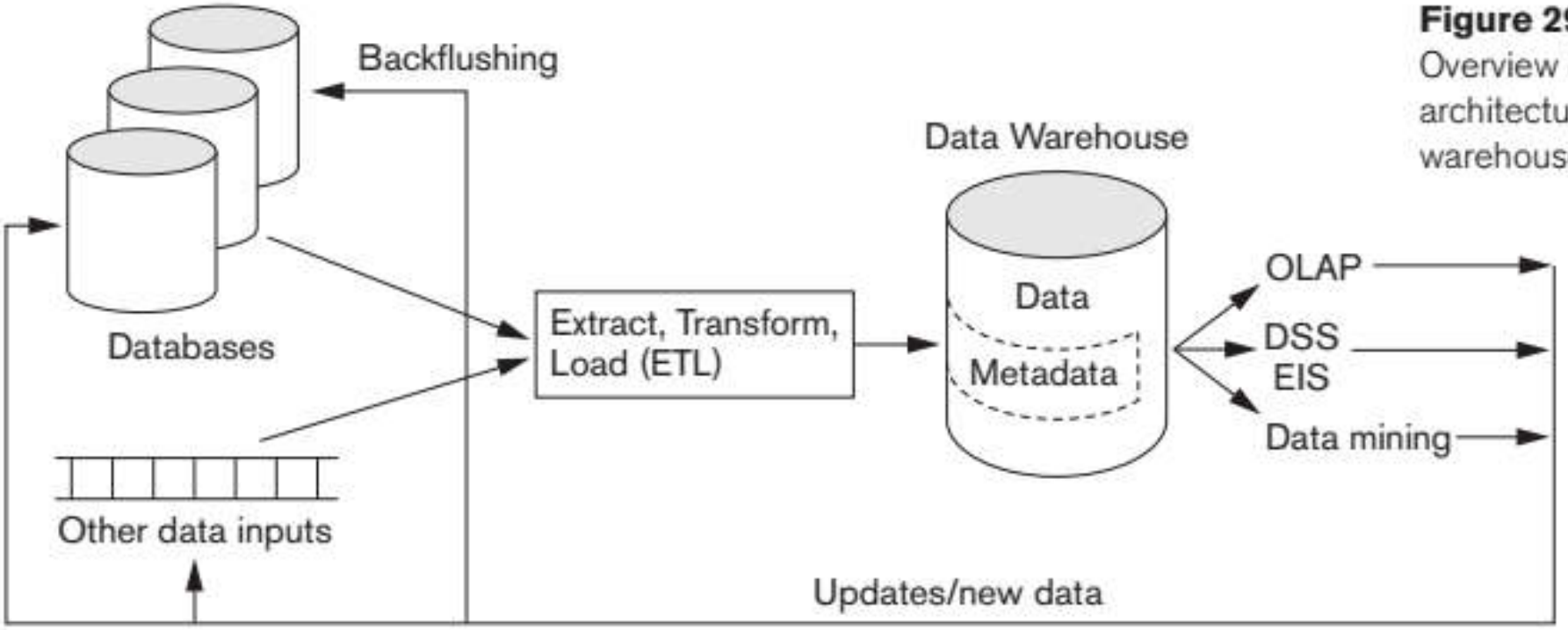
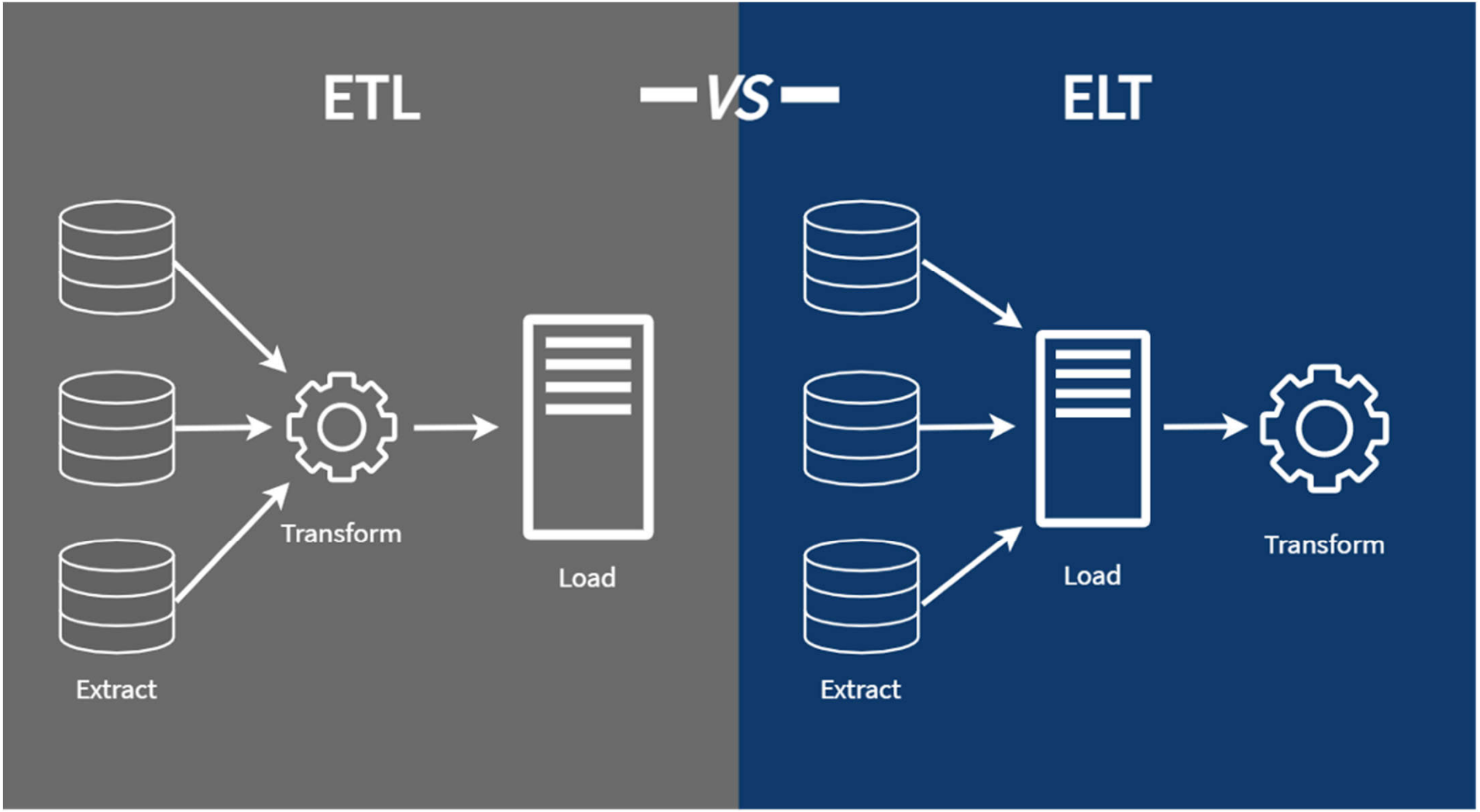


Figure 29.1
Overview of the general architecture of a data warehouse.

[5]

ETL vs. ELT



https://miro.medium.com/max/1200/1*-6tNymvTTqGIWJlzQHwBaw.png

DISCUSSION

ETL VS. ELT
WHICH ONE WOULD YOU PREFER?

HOW TO REFRESH

- By snapshot
- By application code
- By log or audit file
- By timestamped
- By user-define
- Etc.

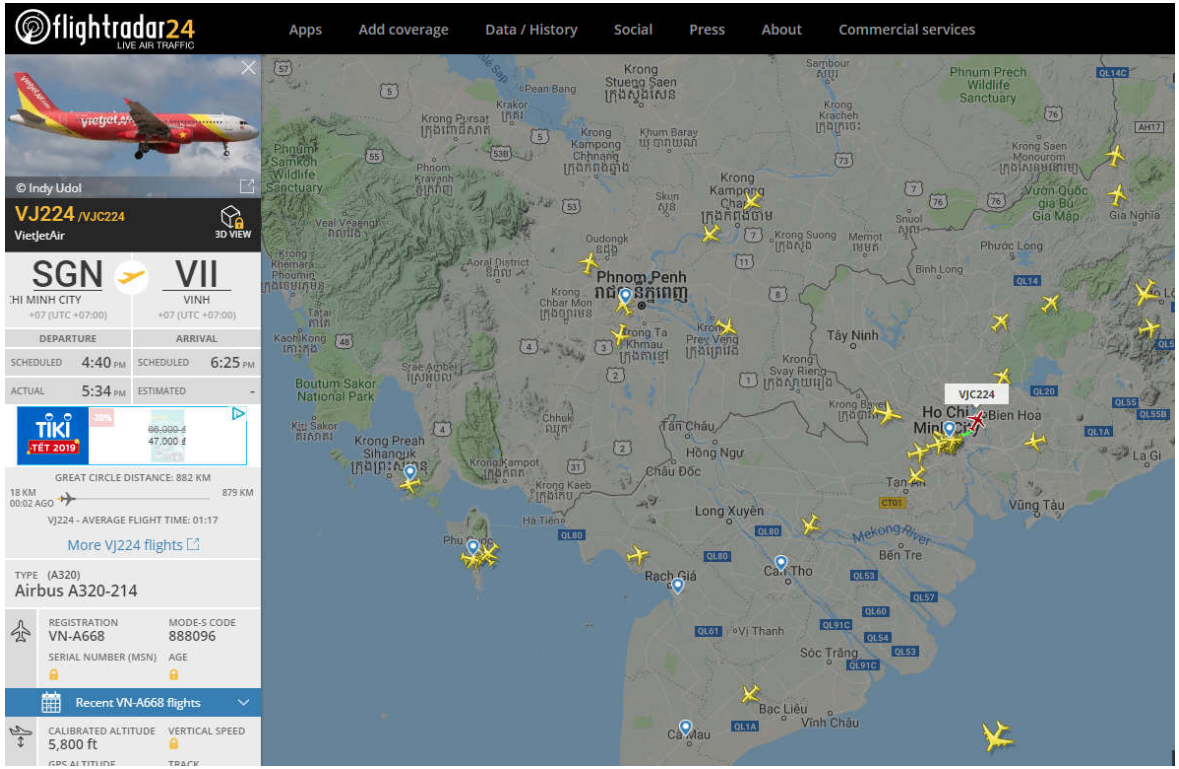
DATA INTEGRATION TOOLS

- SSIS
- OpenTalend
- Pentaho
- Holistics
- Python + Airflow
- Kafka + Greylog + Elasticsearch
- Skyvia
- CloverETL
- Alooma
- Information Builders
- Syncsort
- Etc.

DATA INTEGRATION

PRETTY MUCH EVERYWHERE

- Business
- Science
- Web
- Government
- Etc.



Data integration should be machine work

https://www.flightradar24.com

INTEGRATION ISSUES (1/4)

Table 1 Classification of integration problems based on integration objects

Integration Layer		Integration Category	Possible Integration Problems
Strategy		Business segment	1. Lack of overlap and/or complementarity (Sperling 2007)
			2. Different business operating models (Bakker and Helmink 2000)
		Goal system	3. Conflicting goals (Sperling 2007)
			4. Different performance indicators
Processes		Roles	5. Diverging number of roles
			6. Different function scope
		Activities/tasks	7. Different granularity
			8. Different sequence of activities (control flow) and process interaction (Weigand and van den Heuvel 2002; McAfee 2005)
		Business objects (including business documents)	9. Semantic heterogeneity between business objects (Legner and Vogel 2008): existence of synonyms / homonyms and differing structures
Systems	Applications	Application function	10. External access to data and functions (Boles et al. 2004; Linthicum 2000)
			11. Function granularity (Papazoglou and van den Heuvel 2006)
Data	Data models / schemas		16. Data model heterogeneity (Batini et al. 1986)
			17. Syntactic heterogeneity (Leser and Naumann 2007): different data types (numeric, alpha-numeric, etc.), field lengths and value ranges
			18. Structural heterogeneity (Spaccapietra et al. 1992): different understanding of entities as attribute or as (data) object, different cardinalities between entities
			19. Semantic heterogeneity (March et al. 2000): different understanding of data objects, existence of synonyms and homonyms
			20. Attributes are defined differently (e.g. as mandatory or optional) (Riehm 1997)
	Data elements / data objects		21. Inadequate data quality and/or data value conflicts, e.g. accuracy, missing attributes (Kim and Seo 1991)

A. Schmidt, et al. Integrating information systems: case studies on current challenges. Electron Markets (2010) 20: 161-174

INTEGRATION ISSUES (2/4)

- Heterogeneous data sources
  - E.g., different DBMSs and files
- Data mapping
  - E.g., different schema Employee vs Emp
- Data conflicts
  - E.g., data types, values, format, unit, precision
- Data redundancy
  - E.g., duplicates

INTEGRATION ISSUES (3/4)

- Entity resolution
  - E.g., the same entity.

HCMC University of Technology, 268 Ly Thuong Kiet Street, Ward 14, District 10, Ho Chi Minh City, Vietnam

University of Technology, VNU-HCM, 268 Ly Thuong Kiet Street, Ward 14, District 10, Ho Chi Minh City, Vietnam

University of Technology, VNU-HCM, 268 Ly Thuong Kiet Str., Dist. 10, HCMC, VN

INTEGRATION ISSUES (4/4)

- Constraints violation
  - E.g., primary key constraints, semantic constraints
- Data quality
  - E.g., accuracy, completeness, uniqueness, timeliness, consistency
- Communication heterogeneity
  - E.g., different interfaces such as web, API

Mohamed Eltabakh, "Data integration", slides, 2012.

AN EXAMPLE

S1	S2
Cust	**Customer**
CNo	CustID
CompName	Company
FirstName	Contact
LastName	Phone

AIRLINE SOURCES (1/5)

- Airline1.Schedule(Flight Id, Flight Number, Start Date, End Date, Departure Time, Departure Airport, Arrival Time, Arrival Airport)

	FI	FN	SD	ED	DT	DA	AT	AA
r_{11}	123	49	2013-10-01	2014-03-31	18:05	EWR	21:10	SFO
r_{12}	234	49	2014-04-01	2014-09-30	18:20	EWR	21:25	SFO
r_{13}	345	55	2013-10-01	2014-09-30	18:30	ORD	21:30	BOS
r_{14}	346	55	2013-10-01	2014-09-30	22:30	BOS	23:30	EWR

- Airline1.Flight(Flight Id, Departure Date, Departure Time, Departure Gate, Arrival Date, Arrival Time, Arrival Gate, Plane Id)

	FI	DD	DT	DG	AD	AT	AG	PI
r_{21}	123	2013-12-21	18:45	C98	2013-12-21	21:30	81	4013
r_{22}	123	2013-12-28	21:30	C101	2013-12-29	00:30	81	3008
r_{23}	345	2013-12-29	18:30	B6	2013-12-29	21:45	C18	4013
r_{24}	346	2013-12-29	22:35	C18	2013-12-29	23:35	C101	4013

Dong X.L., Srivastava D. (2015). Big Data Integration. Morgan & Claypool Publishers, p. 198.

AIRLINE SOURCES (2/5)

- Airline2.Flight(Flight Number, Departure Airport, Scheduled Departure Date, Scheduled Departure Time, Actual Departure Time, Arrival Airport, Scheduled Arrival Date, Scheduled Arrival Time, Actual Arrival Time)

	FN	DA	SDD	SDT	ADT	AA	SAD	SAT	AAT
r_{31}	53	SFO	2013-12-21	15:30	16:00	EWR	2013-12-21	23:35	00:15 (+1d)
r_{32}	53	SFO	2013-12-22	15:30	16:15	EWR	2013-12-22	23:35	00:30
r_{33}	53	SFO	2014-06-28	16:00	16:05	EWR	2014-06-29	00:05	23:57 (-1d)
r_{34}	53	SFO	2014-07-06	16:00	16:00	EWR	2014-07-07	00:05	00:09
r_{35}	49	SFO	2013-12-21	12:00	12:35	EWR	2013-12-21	20:05	20:45
r_{36}	77	LAX	2013-12-22	09:15	09:15	SFO	2013-12-22	11:00	10:59

Dong X.L., Srivastava D. (2015). Big Data Integration. Morgan & Claypool Publishers, p. 198.

AIRLINE SOURCES (3/5)

- Airport3.Departures(Air Line, Flight Number, Scheduled, Actual, Gate Time, Takeoff Time, Terminal, Gate, Runway)

	AL	FN	S	A	GT	TT	T	G	R
r_{41}	A1	49	2013-12-21	2013-12-21	18:45	18:53	C	98	2
r_{42}	A1	49	2013-12-28	2013-12-28	21:29	21:38	C	101	2

- Airport3.Arrivals(Air Line, Flight Number, Scheduled, Actual, Gate Time, Landing Time, Terminal, Gate, Runway)

	AL	FN	S	A	GT	LT	T	G	R
r_{51}	A2	53	2013-12-21	2013-12-22	00:21	00:15	B	53	2
r_{52}	A2	53	2013-12-22	2013-12-23	00:40	00:30	B	53	2
r_{53}	A1	55	2013-12-29	2013-12-29	23:35	23:31	C	101	1
r_{54}	A2	49	2013-12-21	2013-12-21	20:50	20:45	B	55	2

Dong X.L., Srivastava D. (2015). Big Data Integration. Morgan & Claypool Publishers, p. 198.

AIRLINE SOURCES (4/5)

- Airfare4.Flight(Flight Id, Flight Number, Departure Airport, Departure Date, Departure Time, Arrival Airport, Arrival Time)

	FI	FN	DA	DD	DT	AA	AT
r_{61}	456	A1-49	Newark Liberty	2013-12-21	18:05	San Francisco	21:10
r_{62}	457	A1-49	Newark Liberty	2014-04-05	18:05	San Francisco	21:10
r_{63}	458	A1-49	Newark Liberty	2014-04-12	18:05	San Francisco	21:10
r_{64}	460	A2-53	San Francisco	2013-12-22	15:30	Newark Liberty	23:35
r_{65}	461	A2-53	San Francisco	2014-06-28	15:30	Newark Liberty	23:35
r_{66}	462	A2-53	San Francisco	2014-07-06	16:00	Newark Liberty	00:05 (+1d)

- Airfare4.Fares(Flight Id, Fare Class, Fare)

	FI	FC	F
r_{71}	456	A	$5799.00
r_{72}	456	K	$999.00
r_{73}	456	Y	$599.00

Dong X.L., Srivastava D. (2015). Big Data Integration. Morgan & Claypool Publishers, p. 198.

AIRLINE SOURCES (5/5)

- Airinfo5.AirportCodes(Airport Code, Airport Name)

	AC	AN
r_{81}	EWR	Newark Liberty, NJ, US
r_{82}	SFO	San Francisco, CA, US

- Airinfo5.AirlineCodes(Air Line Code, Air Line Name)

	ALC	ALN
r_{91}	A1	Airline1
r_{92}	A2	Airline2

Dong X.L., Srivastava D. (2015). Big Data Integration. Morgan & Claypool Publishers, p. 198.

INTEGRATION NEEDS

- Linking
  - Airline (e.g., Airline 1, Airline 2) + Airport (e.g., Airport 3)
  - Airline + Airport + Airfare
- Benefits
  - Reasons for flight delays
  - Flight patterns
  - Flight booking
  - Single point access

"For each airline flight number, compute the average delays between scheduled and actual departure times, and between actual gate departure and takeoff times, over the past one month."

Dong X.L., Srivastava D. (2015). Big Data Integration. Morgan & Claypool Publishers, p. 198.

DATA INTEGRATION CHALLENGES

SEMANTIC AMBIGUITY

- Conceptual information
  - The same but modeled differently
    - Time information: departure time and arrival time
    - Airline1: Gate departure and arrival times
    - Airline2: Takeoff and landing times
  - Different but modeled similarly
    - Departure date by Airline1
    - Departure date by Airfare4

INSTANCE REPRESENTATION AMBIGUITY

- The same data instance
  - Flight numbers in Airline1, Airline2, and Airfare4
  - Departure and arrival airports in Airline1, Airline2, and Airfare4
  - Searching problem: string matching
    - “Newark Liberty” in Airfare4.Flight with “Newark Liberty, NJ, US” in Airinfo5.AirportCodes

FOR EXAMPLE

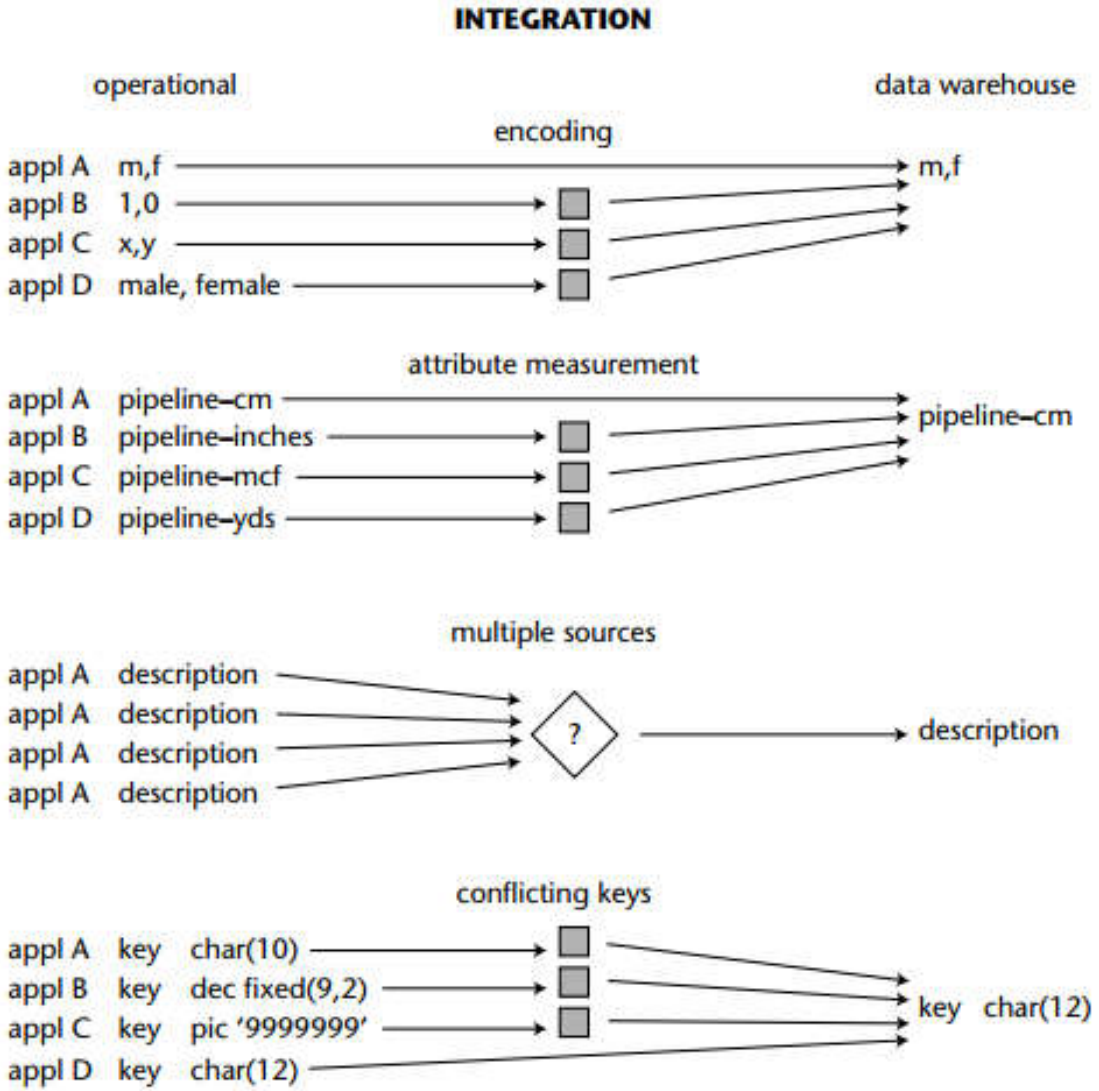


Figure 2-2 The issue of integration.

[1]

DATA INCONSISTENCY (1/3)

- E.g., records r32 in Airline2.Flight and r52 in Airport3.Arrivals

	FN	DA	SDD	SDT	ADT	AA	SAD	SAT	AAT
r_{31}	53	SFO	2013-12-21	15:30	16:00	EWR	2013-12-21	23:35	00:15 (+1d)
r_{32}	53	SFO	2013-12-22	15:30	16:15	EWR	2013-12-22	23:35	00:30
r_{33}	53	SFO	2014-06-28	16:00	16:05	EWR	2014-06-29	00:05	23:57 (-1d)
r_{34}	53	SFO	2014-07-06	16:00	16:00	EWR	2014-07-07	00:05	00:09
r_{35}	49	SFO	2013-12-21	12:00	12:35	EWR	2013-12-21	20:05	20:45
r_{36}	77	LAX	2013-12-22	09:15	09:15	SFO	2013-12-22	11:00	10:59

	AL	FN	S	A	GT	LT	T	G	R
r_{51}	A2	53	2013-12-21	2013-12-22	00:21	00:15	B	53	2
r_{52}	A2	53	2013-12-22	2013-12-23	00:40	00:30	B	53	2
r_{53}	A1	55	2013-12-29	2013-12-29	23:35	23:31	C	101	1
r_{54}	A2	49	2013-12-21	2013-12-21	20:50	20:45	B	55	2

Dong X.L., Srivastava D. (2015). Big Data Integration. Morgan & Claypool Publishers, p. 198.

DATA INCONSISTENCY (2/3)

- E.g., record r62 in Airfare4.Flight and record r12 in Airline1.Schedule

	FI	FN	DA	DD	DT	AA	AT
r_{61}	456	A1-49	Newark Liberty	2013-12-21	18:05	San Francisco	21:10
r_{62}	457	A1-49	Newark Liberty	2014-04-05	18:05	San Francisco	21:10
r_{63}	458	A1-49	Newark Liberty	2014-04-12	18:05	San Francisco	21:10
r_{64}	460	A2-53	San Francisco	2013-12-22	15:30	Newark Liberty	23:35
r_{65}	461	A2-53	San Francisco	2014-06-28	15:30	Newark Liberty	23:35
r_{66}	462	A2-53	San Francisco	2014-07-06	16:00	Newark Liberty	00:05 (+1d)

	FI	FN	SD	ED	DT	DA	AT	AA
r_{11}	123	49	2013-10-01	2014-03-31	18:05	EWR	21:10	SFO
r_{12}	234	49	2014-04-01	2014-09-30	18:20	EWR	21:25	SFO
r_{13}	345	55	2013-10-01	2014-09-30	18:30	ORD	21:30	BOS
r_{14}	346	55	2013-10-01	2014-09-30	22:30	BOS	23:30	EWR

Dong X.L., Srivastava D. (2015). Big Data Integration. Morgan & Claypool Publishers, p. 198.

DATA INCONSISTENCY (3/3)

- E.g., record *r*65 in Airfare4.Flight and record *r*33 in Airline2.Flight

	FI	FN	DA	DD	DT	AA	AT
r_{61}	456	A1-49	Newark Liberty	2013-12-21	18:05	San Francisco	21:10
r_{62}	457	A1-49	Newark Liberty	2014-04-05	18:05	San Francisco	21:10
r_{63}	458	A1-49	Newark Liberty	2014-04-12	18:05	San Francisco	21:10
r_{64}	460	A2-53	San Francisco	2013-12-22	15:30	Newark Liberty	23:35
r_{65}	461	A2-53	San Francisco	2014-06-28	15:30	Newark Liberty	23:35
r_{66}	462	A2-53	San Francisco	2014-07-06	16:00	Newark Liberty	00:05 (+1d)

	FN	DA	SDD	SDT	ADT	AA	SAD	SAT	AAT
r_{31}	53	SFO	2013-12-21	15:30	16:00	EWR	2013-12-21	23:35	00:15 (+1d)
r_{32}	53	SFO	2013-12-22	15:30	16:15	EWR	2013-12-22	23:35	00:30
r_{33}	53	SFO	2014-06-28	16:00	16:05	EWR	2014-06-29	00:05	23:57 (-1d)
r_{34}	53	SFO	2014-07-06	16:00	16:00	EWR	2014-07-07	00:05	00:09
r_{35}	49	SFO	2013-12-21	12:00	12:35	EWR	2013-12-21	20:05	20:45
r_{36}	77	LAX	2013-12-22	09:15	09:15	SFO	2013-12-22	11:00	10:59

Dong X.L., Srivastava D. (2015). Big Data Integration. Morgan & Claypool Publishers, p. 198.

DATA INTEGRATION ARCHITECTURE

SCHEMA ALIGNMENT

- Outcomes
  - A mediated schema
  - Attribute matching
  - Schema mapping
- Challenges
  - The same domain with different schemas (e.g., Arrival Date in Airline1.Flight, Actual Arrival Date in Airline2.Flight, and Actual in Airport3.Arrivals)
  - The same attribute but different meaning (e.g., Actual in Airport3.Departures and Actual in Airport3.Arrivals)

RECORD LINKAGE

- Outcome
  - Identifying the records that refer to a distinct entity
- Challenges
  - The same entity described in different ways (e.g., records *r*11 in Airline1.Schedule and *r*21 in Airline1.Flight should be linked to record *r*41 in Airport3.Departures)
  - The same information described in different ways (e.g., the alternate ways of representing airports)
  - The comparison for every pair of records may be infeasible

DATA FUSION

- Outcome
  - Avoiding conflicting values (e.g., mis-typing, incorrect calculation, out-of-date information)
- Challenges
  - Data quality

MORE CHALLENGES WITH BIG DATA

- Volume
- Variety
- Velocity
- Veracity

IMPLEMENTATION CHALLENGES

- Time, effort, and cost
- Technology variance
- Cross-divisional collaboration
- Business requirement change

DATA SOURCE EVOLUTION

- Delivery system
  - Order, store, product, customer
  - Shipping (internal vs. external drivers)
  - E-voucher
  - Store open/close and boundary
  - Etc.

MAINTENANCE AND MANAGEMENT

- Like a DBA
- Management team

DATA WAREHOUSE MONITORING (1/2)

- Identifying what growth is occurring, where the growth is occurring, and at what rate the growth is occurring
- Identifying what data is being used
- Calculating what response time the end user is getting
- Determining who is actually using the data warehouse
- Specifying how much of the data warehouse end users are using
- Pinpointing when the data warehouse is being used
- Recognizing how much of the data warehouse is being used
- Examining the level of usage of the data warehouse

[1]

DATA WAREHOUSE MONITORING (2/2)

- What data is being accessed?
  - When?
  - By whom?
  - How frequently?
  - At what level of detail?
- What is the response time for the request?
- At what point in the day is the request submitted?
- How big was the request?
- Was the request terminated, or did it end naturally?
- Etc.

[1]

DATA PROFILE

- A catalog of all tables in the warehouse
- A profile of the contents of those tables
- A profile of the growth of the tables in the data warehouse
- A catalog of the indexes available for entry to the tables
- A catalog of the summary tables and the sources for the summary

[1]

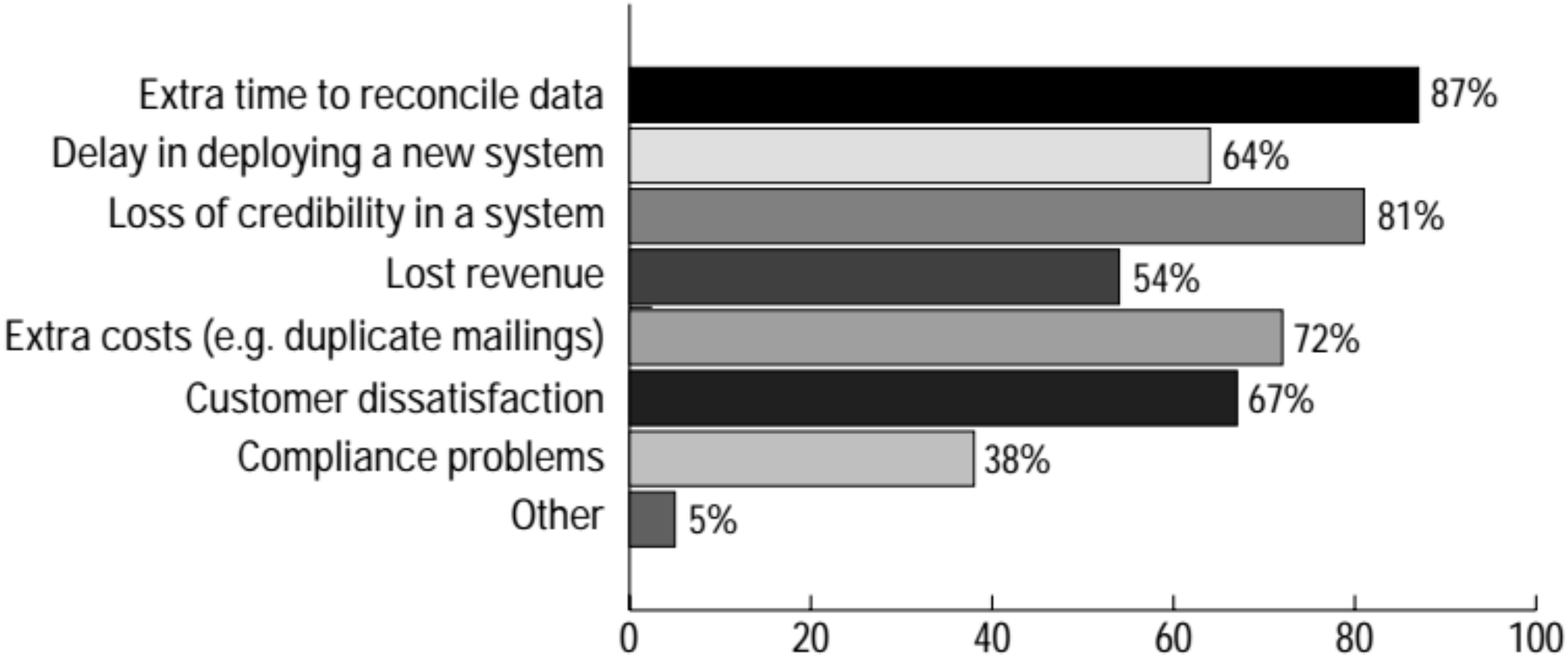
DATA QUALITY

WHAT DO REAL-WORLD NUMBERS SAY?

- An insurance company
  - **2 millions** claim per month
  - Each claim has **377** data elements
  - If the error rate is **1/1000**, there are **754.000** errors per month → **9.04 million** errors per year
  - If **10%** of those data elements are critical, about **1 million** errors should be fixed
  - Assume that **$10** per error (staff time, errorneous payouts, the loss of customer trust and loyalty), the firm costs **$10 millions** per year

Wayne W. Eckerson, Achieving Business Success through a Commitment to High Quality Data, in DATA QUALITY AND THE BOTTOM LINE, TDWI REPORT SERIES, The Data Warehousing Institue, 2002

PROBLEMS DUE TO POOR DATA QUALITY



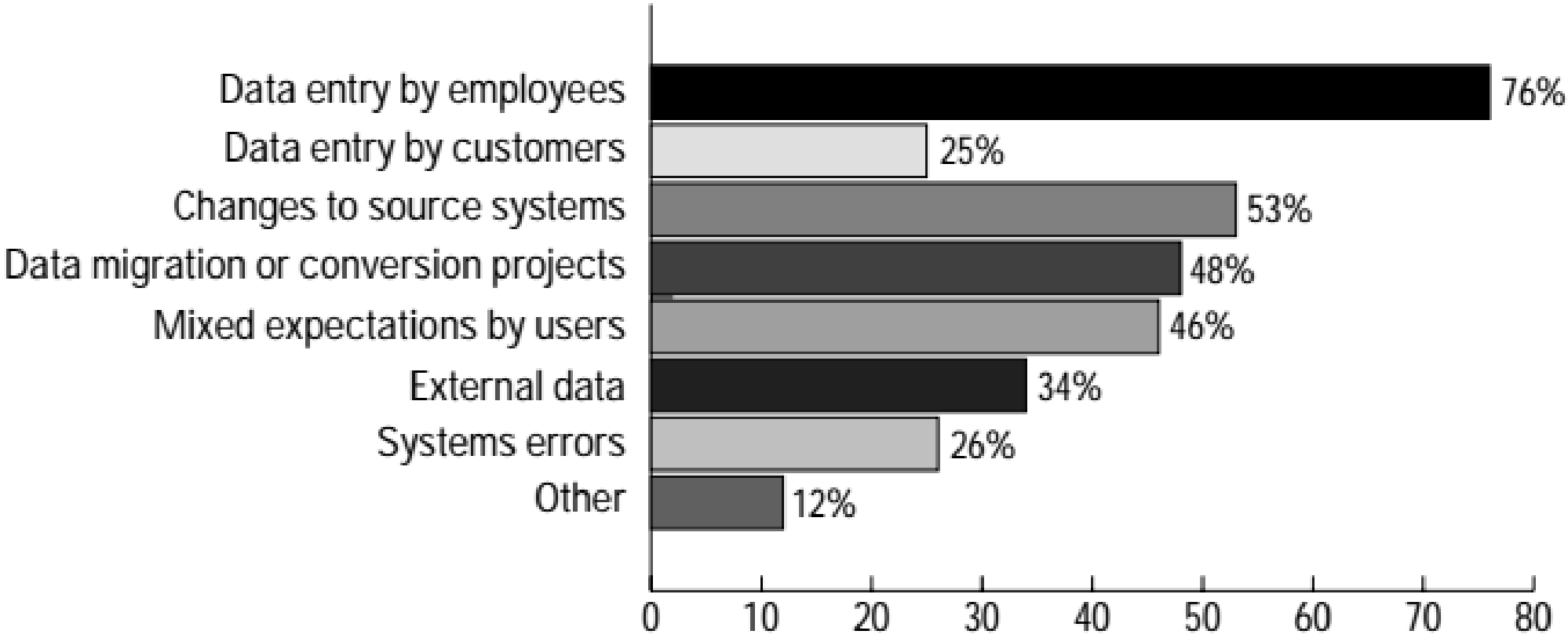
Wayne W. Eckerson, Achieving Business Success through a Commitment to High Quality Data, in DATA QUALITY AND THE BOTTOM LINE, TDWI REPORT SERIES, The Data Warehousing Institue, 2002

REASONS FOR BAD QUALITY DATA

- Historical changes
  - E.g., the date of birth of customers of an insurance company
- Data usage
  - E.g., customer profession of those who buy some shares from a bank
- Company mergers
  - E.g., a bank and a insurance company are merged together into larger holdings
- Dormant data
  - E.g., 10% of the population moves to a new address every year
- Data enrichment
  - E.g., enriching internal data with external sources from a large hardware supplier with its B2B application

ir. Henri Theuwissen, Nancy Croonen, The Horror of Bad Data Quality, paper 162-28, SUGI 28, Data Warehousing and Enterprise Solutions, 2003

SOURCES OF DATA QUALITY PROBLEMS



Wayne W. Eckerson, Achieving Business Success through a Commitment to High Quality Data, in DATA QUALITY AND THE BOTTOM LINE, TDWI REPORT SERIES, The Data Warehousing Institue, 2002

HOW TO MANAGE DATA/INFORMATION QUALITY?

- Data quality methodology

Wayne W. Eckerson, Achieving Business Success through a Commitment to High Quality Data, in DATA QUALITY AND THE BOTTOM LINE, TDWI REPORT SERIES, The Data Warehousing Institue, 2002

DATA QUALITY DIMENSIONS (1/3)

- Data quality dimensions
  - **Completeness:** *"The proportion of stored data against the potential of '100% complete'."*
  - **Uniqueness:** *"No thing will be recorded more than once based upon how that thing is identified."*
  - **Timeliness:** *"The degree to which data represent reality from the required point in time."*
  - **Validity:** *"Data are valid if it conforms to the syntax (format, type, range) of its definition."*
  - **Accuracy:** *"The degree to which data correctly describes the 'real world' object or event being described."*
  - **Consistency:** "*The absence of difference, when comparing two or more representations of a thing against a definition."*

The Six Primary Dimensions for Data Quality Assessment-Defining Data Quality Dimensions, DAMA UK Working Group, 2013

DATA QUALITY DIMENSIONS (2/3)

- Data quality dimensions
  - **Accuracy**: *"inaccuracy implies that the information system represents a real world state different from the one that should have been represented"*
  - **Timeliness**: *"the delay between a change of the real-world state and the resulting modification of the information system state"*
  - **Completeness**: *"the ability of an information system to represent every meaningful state of the represented real world system"*
  - **Consistency**: *"inconsistency would mean that the representation mapping is one-to-many"*
  - **Interpretability**: *"concerns the documentation and metadata that are available to interpret correctly the meaning and properties of data sources"*
  - **Accessibility**: *"measures the ability of the user to access the data as from his/her own culture, physical status/functions and technologies available"*
  - **Usability**: *"measures the effectiveness, efficiency, satisfaction with which specified users perceive and make use of data"*
  - **Trustworthiness**: *"measures how reliable is the organization in providing data sources"*

C. Batini, B. Pernici. Data Quality Management and Evolution of Information Systems. In IFIP International Federation for Information Processing, volume 214, The Past and Future of Information Systems: 1976-2006 and Beyond, eds. D. Avison, S. Elliot, J. Krogstie, J. Pries-Heje, Boston: Springer, 2006, pp. 51-62

DATA QUALITY DIMENSIONS (3/3)

Content	Time	Form
• Accuracy • Relevance • Completeness • Conciseness • Scope • Performance	• Timeliness • Currency • Frequency • Time Period	• Clarity • Detail • Order • Presentation • Media

(O'Brien & Marakas, 2010)

INFORMATION MANAGEMENT

- Information management should be taken with different aspects
  - Information collection
  - Information organization
  - Information storage
  - Information manipulation
  - Information processing
  - Information protection

DISCUSSION

SSIS DEMO

SSIS DEMO (1/6)

- Requirements

1. SQL Server Standard/Developer/Enterprise/Evaluation edition.
   - Click here to download SQL Server 2022
2. MS Visual Studio.
   - Click here to download MS Visual Studio 2022
3. SQL Server data tools.
   - Click here to download SQL Server Data Tools for visual studio.

SSIS DEMO (2/6)

- Migrate data from one database to another database
  - E.g., RetailSales

SSIS DEMO (3/6)

- Input data
  - SalesDetails_NorthAmerica.csv
  - SalesDetails_Others.csv
- Calculated data
  - Total quantity
  - Total sales = Sales + Tax Amount
  - Tax Amount = 8% * unit price
- Lookup data
  - CompanyX database
    - Lookup product name
    - Lookup territory name

SSIS DEMO (4/6)

- SalesDetails_NorthAmerica.csv

SalesOrderI	ProductID	OrderQty	UnitPrice	TerritoryID
43659	776	1	2024.994	5
43659	777	3	2024.994	5
43659	778	1	2024.994	5
43659	771	1	2039.994	5
43659	772	1	2039.994	5
43659	773	2	2039.994	5
43659	774	1	2039.994	5
43659	714	3	28.8404	5

- SalesDetails_Others.csv

SalesOrderI	ProductID	OrderQty	UnitPrice	TerritoryID
43698	773	1	3399.99	7
43701	773	1	3399.99	9
43703	749	1	3578.27	9
43704	778	1	3374.99	9
43705	771	1	3399.99	9
43708	764	1	699.0982	10
43709	752	1	3578.27	9

SSIS DEMO (5/6)

- Sales_NorthAmerica

SQLQuery6.sql - LA...ilSalesDW (sa (88)) | SQLQuery5.sql - LA...ilSalesDW (sa (87))* | SQLQuery3.sql - L...CompanyX (sa (95))

```
      ,[Sales]
      ,[Territory ID]
      ,[Tax Amt]
      ,[Total Sales]
      ,[Territory ID]
      ,[Product ID]
      ,[ProductName]
      ,[TerritoryName]
  FROM [RetailSalesDW].[dbo].[Sales_NorthAmerica]
```

121 %

Results | Messages

	SalesOrderID	Product ID	OrderQuantity	Sales	Territory ID	Tax Amt	Total Sales	Territory ID	Product ID	ProductName	TerritoryName
1	49085	808	2	24.2945	6	3.5227	27.8172	6	808	LL Mountain Handlebars	Canada
2	71830	884	17	29.6945	6	4.3057	34.0002	6	884	Short-Sleeve Classic Jersey, XL	Canada
3	65112	869	1	69.99	4	10.1486	80.1386	4	869	Women's Mountain Shorts, L	Southwest
4	48709	797	1	1000.4375	4	145.0634	1145.5009	4	797	Road-550-W Yellow, 38	Southwest
5	44113	777	6	2024.994	4	293.6241	2318.6181	4	777	Mountain-100 Black, 44	Southwest
6	52802	707	1	34.99	4	5.0736	40.0636	4	707	Sport-100 Helmet, Red	Southwest
7	50298	729	1	202.332	4	29.3381	231.6701	4	729	LL Road Frame - Red, 60	Southwest
8	51839	799	3	672.294	6	97.4826	769.7766	6	799	Road-550-W Yellow, 42	Canada
9	55282	970	4	728.91	4	105.692	834.602	4	970	Touring-2000 Blue, 46	Southwest
10	49540	806	2	61.374	4	8.8992	70.2732	4	806	ML Headset	Southwest
11	65257	864	4	38.10	2	5.5245	43.6245	2	864	Classic Vest, S	Northeast
12	46645	784	5	1229.4589	6	178.2715	1407.7304	6	784	Mountain-200 Black, 46	Canada
13	72882	708	1	34.99	1	5.0736	40.0636	1	708	Sport-100 Helmet, Black	Northwest
14	57120	708	4	20.994	6	3.0441	24.0381	6	708	Sport-100 Helmet, Black	Canada

Query executed successfully. | LAPTOP-8G6IDACM (16.0 RTM) | sa (87) | RetailSalesDW | 00:00:(

SSIS DEMO (6/6)

- Sales_Others

SQLQuery6.sql - LA...ilSalesDW (sa (88))* | SQLQuery5.sql - LA...ilSalesDW (sa (87))* | SQLQuery3.sql - L...CompanyX (sa (95))

```
        ,[OrderQuantity]
        ,[Sales]
        ,[Territory ID]
        ,[Tax Amt]
        ,[Total Sales]
        ,[Tax Amt]
        ,[ProductName]
        ,[TerritoryName]
    FROM [RetailSalesDW].[dbo].[Sales_Others]
```

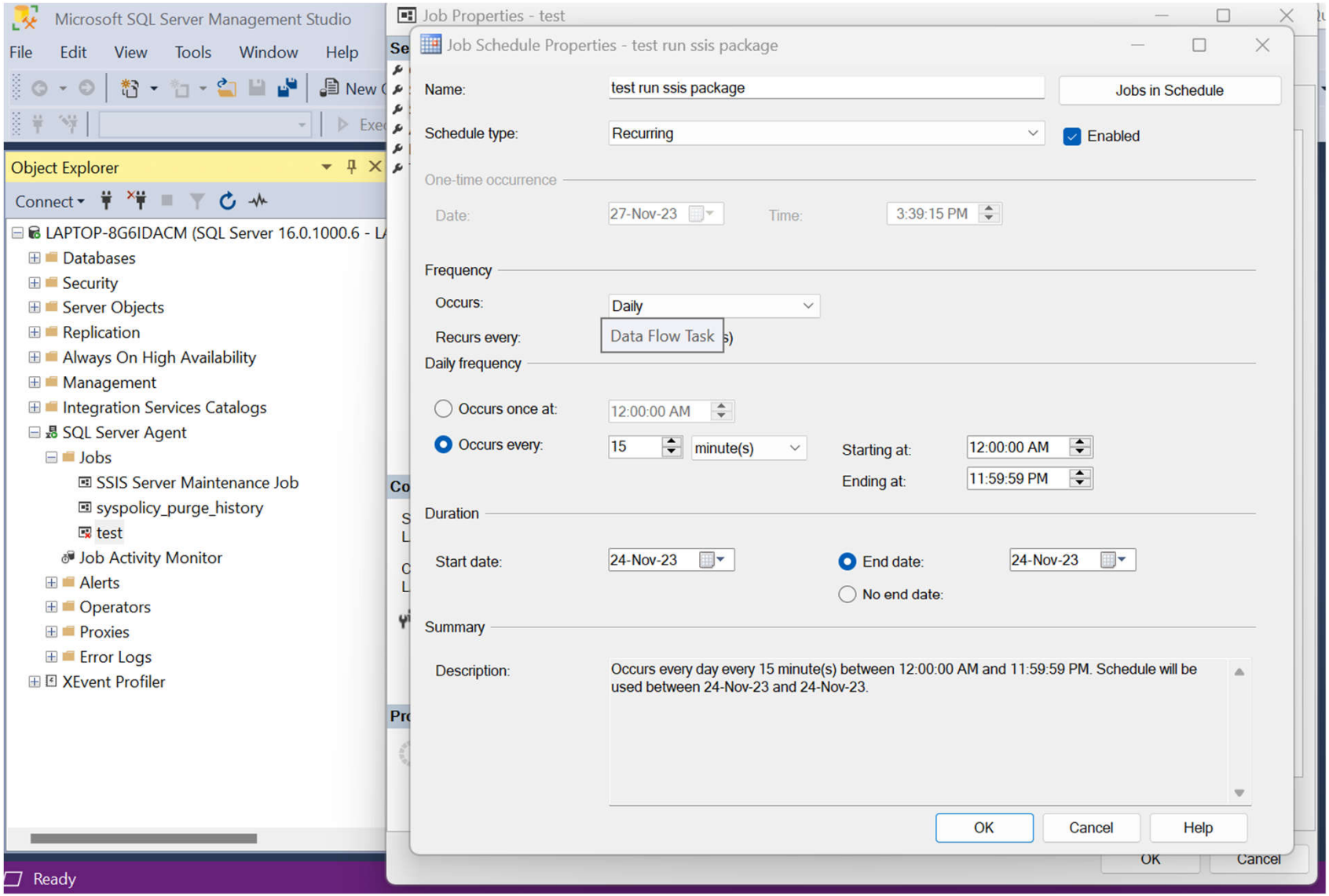
121 %

Results | Messages

	SalesOrderID	Product ID	OrderQuantity	Sales	Territory ID	Tax Amt	Total Sales	Tax Amt	ProductName	TerritoryName
1	50508	793	1	2181.5625	9	316.3266	2497.8891	316.3266	Road-250 Black, 44	Australia
2	71838	864	3	38.10	7	5.5245	43.6245	5.5245	Classic Vest, S	France
3	63465	880	1	54.99	7	7.9736	62.9636	7.9736	Hydration Pack - 70 oz.	France
4	60322	712	1	8.99	10	1.3036	10.2936	1.3036	AWC Logo Cap	United Kingdom
5	52670	708	1	34.99	9	5.0736	40.0636	5.0736	Sport-100 Helmet, Black	Australia
6	49175	791	1	2443.35	9	354.2858	2797.6358	354.2858	Road-250 Red, 52	Australia
7	58931	870	7	2.994	9	0.4341	3.4281	0.4341	Water Bottle - 30 oz.	Australia
8	61599	921	1	4.99	9	0.7236	5.7136	0.7236	Mountain Tire Tube	Australia
9	45243	766	1	699.0982	9	101.3692	800.4674	101.3692	Road-650 Black, 60	Australia
10	63523	867	1	69.99	10	10.1486	80.1386	10.1486	Women's Mountain Shorts, S	United Kingdom
11	64645	713	1	49.99	7	7.2486	57.2386	7.2486	Long-Sleeve Logo Jersey, S	France
12	56886	880	1	54.99	7	7.9736	62.9636	7.9736	Hydration Pack - 70 oz.	France
13	65253	974	3	1020.594	7	147.9861	1168.5801	147.9861	Road-350-W Yellow, 42	France
14	44918	777	1	3374.99	8	489.3736	3864.3636	489.3736	Mountain-100 Black, 44	Germany

Query executed successfully. | LAPTOP-8G6IDACM (16.0 RTM) | sa (88) | RetailSales

ETL SCHEDULER



ETL UPSERT

- In SQL Server
  - Using MERGE

```
MERGE dbo.FactBuyingHabits AS Target
USING (SELECT CustomerID, ProductID, PurchaseDate FROM dbo.Purchases) AS Source
ON (Target.ProductID = Source.ProductID AND Target.CustomerID = Source.CustomerID)
WHEN MATCHED THEN
    UPDATE SET Target.LastPurchaseDate = Source.PurchaseDate
WHEN NOT MATCHED BY TARGET THEN
    INSERT (CustomerID, ProductID, LastPurchaseDate)
    VALUES (Source.CustomerID, Source.ProductID, Source.PurchaseDate)
OUTPUT $action, Inserted.*, Deleted.*;
```

https://learn.microsoft.com/en-us/previous-versions/sql/sql-server-2008-r2/bb522522(v=sql.105)?redirectedfrom=MSDN

SUMMARY

- ETL overview
- Data integration
- Data quality
- SSIS Demo

QUESTIONS AND ANSWERS

Picture from: http://philadelphiasculpturegym.blogspot.com/2013/09/save-date-free-talk-and-q-on-affordable.html

REFERENCES

1. Tobias Zwingmann, “AI-Powered Business Intelligence,” Kindle Edition, O'reilly Press, 2022
2. Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques,” Third Edition, Morgan Kaufmann Publishers, 2012.
3. David L. Olson, Dursun Delen, “Advanced Data Mining Techniques,” Springer-Verlag, 2008.
4. Jeen Su Lim, John Heinrichs, “Digital Business Intelligence Management with Big Data Analytics," Kindle Edition, O'reilly Press, 2021.
5. William H. Inmon, “Building the Data Warehouse,” Fourth Edition, Wiley Publishing, Inc., 2005.
6. R. Kimball, M. Ross, "The Data Warehouse ToolKit," 3rd Edition, Wiley Publishing, Inc., 2013.
7. Turban, E., Aronson,J.E., “Decision Support Systems and Intelligent Systems” - 7th Edition, Prentice-Hall, 2005.
8. Ramesh Sharda, Dursun Delen, Efraim Turban, “Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support,” 7th Edition, Pearson Education, Inc., 2020.