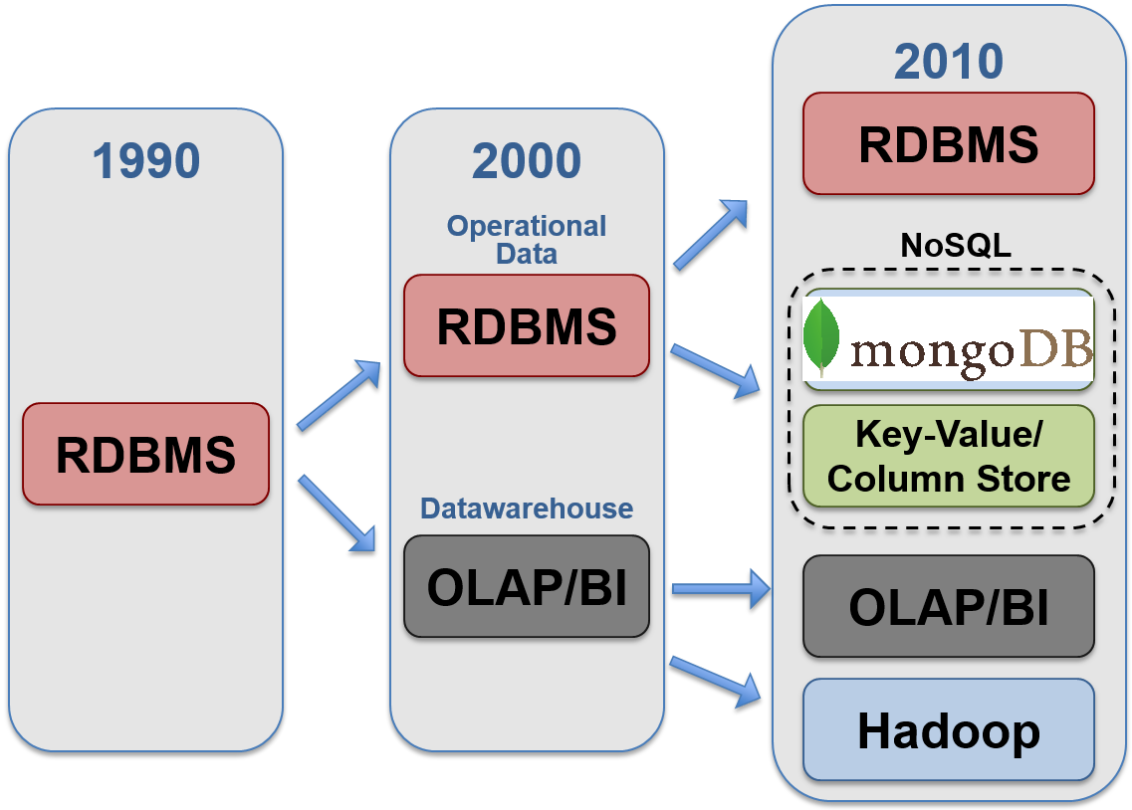
Data Warehousing

- Data Warehouse (DW) was proposed as a new type of database management system which would keep no transactional data but only summarized historical information for decision making purposes.

- W. H Inmon characterized a data warehouse as:
  - **"A subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions."**

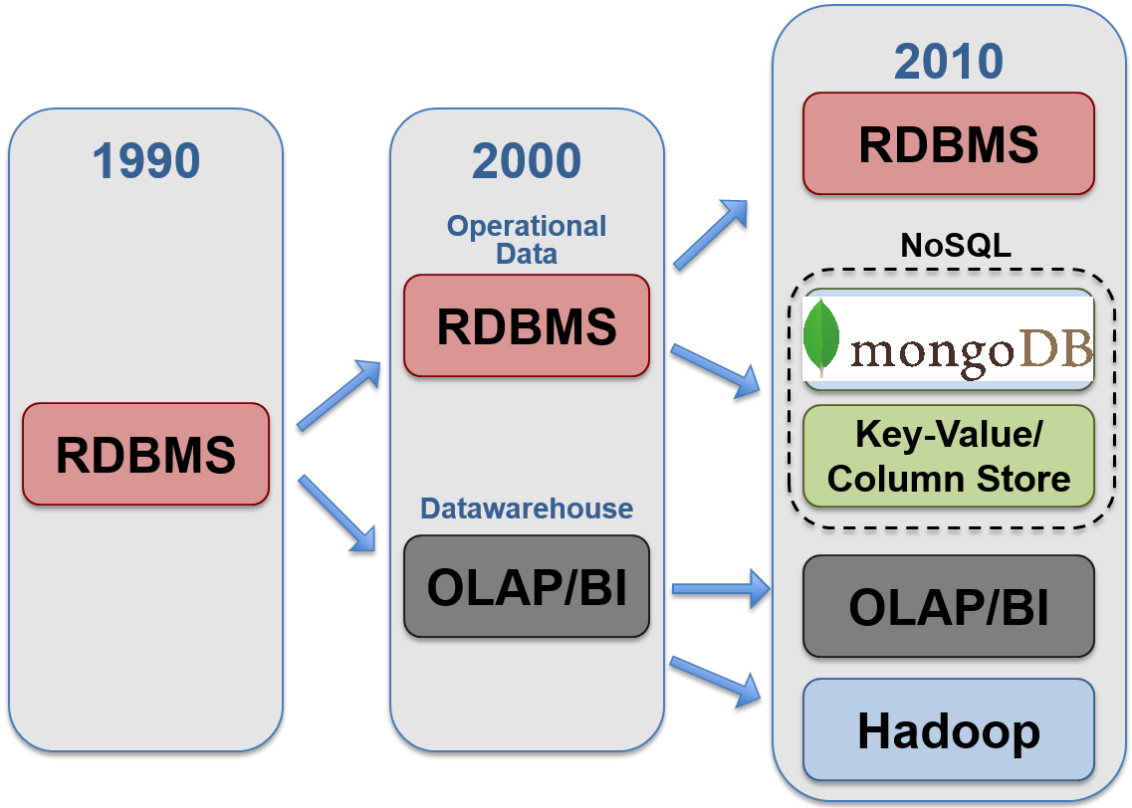
Business Systems Architecture



Decision Making Levels:

- Operational level
  - Day-to-day business decisions
  - Short time frame
  - On-Line Transaction Processing (OLTP)
- Tactical level
  - Middle management decisions
  - Medium-term focus
- Strategic level
  - Senior management
  - Long-term focus

Business Systems Architecture



Business Systems:

- Operational systems
  - Operational level
  - Focus on simple INSERT, UPDATE, DELETE and/or SELECT statements
  - Transaction throughput
- Decision Support Systems
  - Tactical + strategic level
  - Focus on data retrieval by answering complex ad-hoc queries (SELECT statements)
  - Represent data in a multidimensional way
  - Trend analysis

Data Warehouse vs Transactional Systems

	Transactional system	Data Warehouse
Usage	Day to day business operations	Decision support at tactical/strategic level
Data latency	Real-time data	Periodic snapshots, incl. historical data
Design	Application oriented	Subject Oriented
Normalization	Normalized data	(Sometimes also) denormalized data
Data manipulation	Insert/Update/Delete/Select	Insert/Select
Transaction management	Important	Less of a concern
Type of queries	Many, simple queries	Fewer, but complex and ad-hoc queries

Data Warehousing

- Traditional databases are not optimized for data access - they have to balance the requirement of data access with the need to ensure integrity of data.
- DWs provide access for complex analysis of data, knowledge discovery and decision support both through ad-hoc and canned queries.
- Most of the times the data warehouse users need only read access but, need the access to be fast over a large volume of data.
- Most of the data required for data warehouse analysis comes from multiple sources that may include databases from different data models and sometimes files acquired from independent systems and platforms.

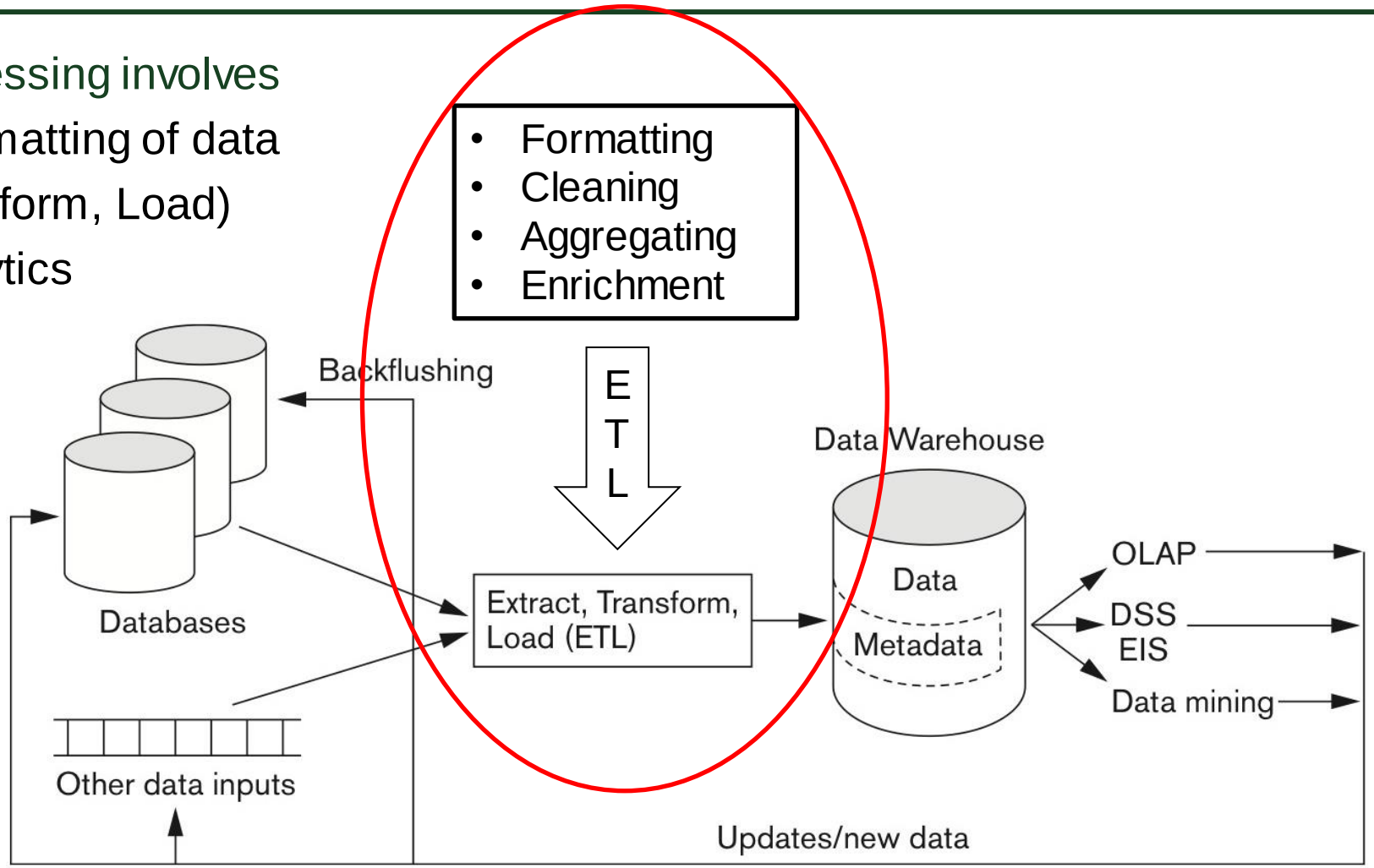
Data Warehouse Terminology

- Data warehouses are databases that store and maintain analytical data separately from transaction-oriented databases for the purpose of decision support
  - Traditional databases support online transaction processing -OLTP .
  - Data Warehouses are for analytical applications- largely OLAP.

- Applications that data warehouse supports are:
  - OLAP (Online Analytical Processing) is a term used to describe the analysis of complex data from the data warehouse.
  - DSS (Decision Support Systems) also known as EIS (Executive Information Systems) supports organization's leading decision makers for making complex and important decisions.
  - Data Mining is used for knowledge discovery, the process of searching data for unanticipated new knowledge.

Data Warehouse Terminology

- Data Warehouse processing involves
  - Cleaning and reformatting of data
  - ETL (Extract, Transform, Load)
  - OLAP – Data Analytics
  - Data Mining



Data Warehouse Functionality

- Functionality that can be expected:
  - Pivot: Cross tabulation (also referred to as rotation) is performed.
  - Roll-up (also Drill-up): Data is summarized with increasing generalization (for example, weekly to quarterly to annually).
  - Drill-down: Increasing levels of detail are revealed (the complement of roll-up).
  - Slice and dice: Projection operations are performed on the dimensions.
  - Sorting: Data is sorted by ordinal value.
  - Selection: Data is filtered by value or range.
  - Derived (computed) attributes: Attributes are computed by operations on stored and derived values.

Data Warehouse Functionality

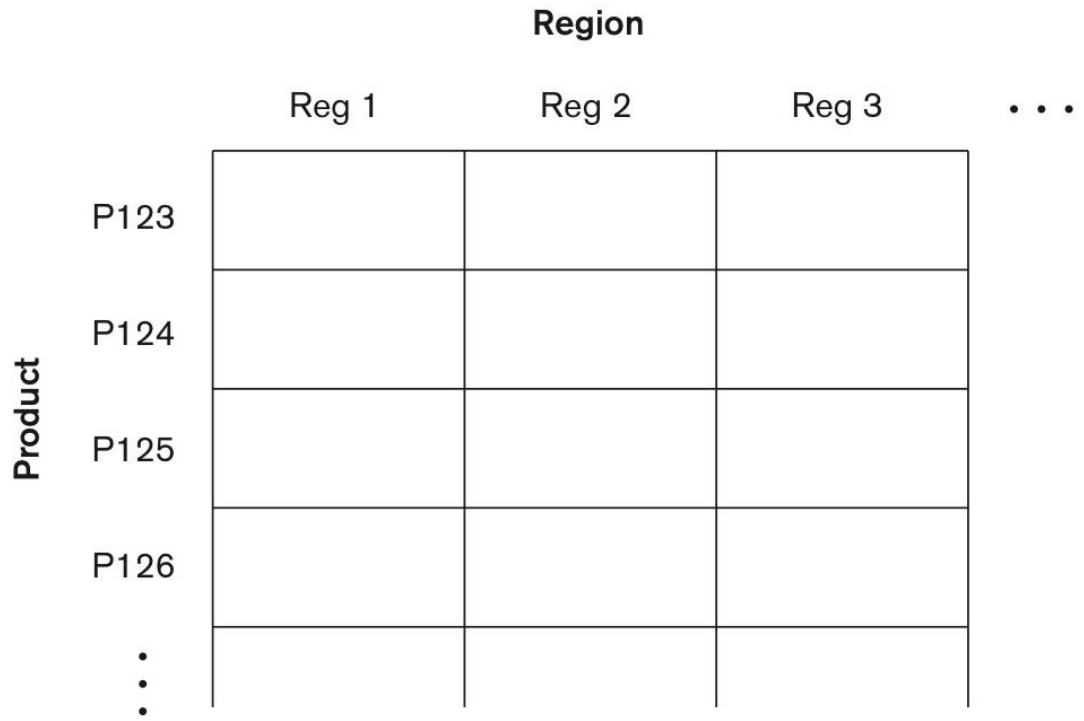


Figure 29.2 A two-dimensional matrix model.

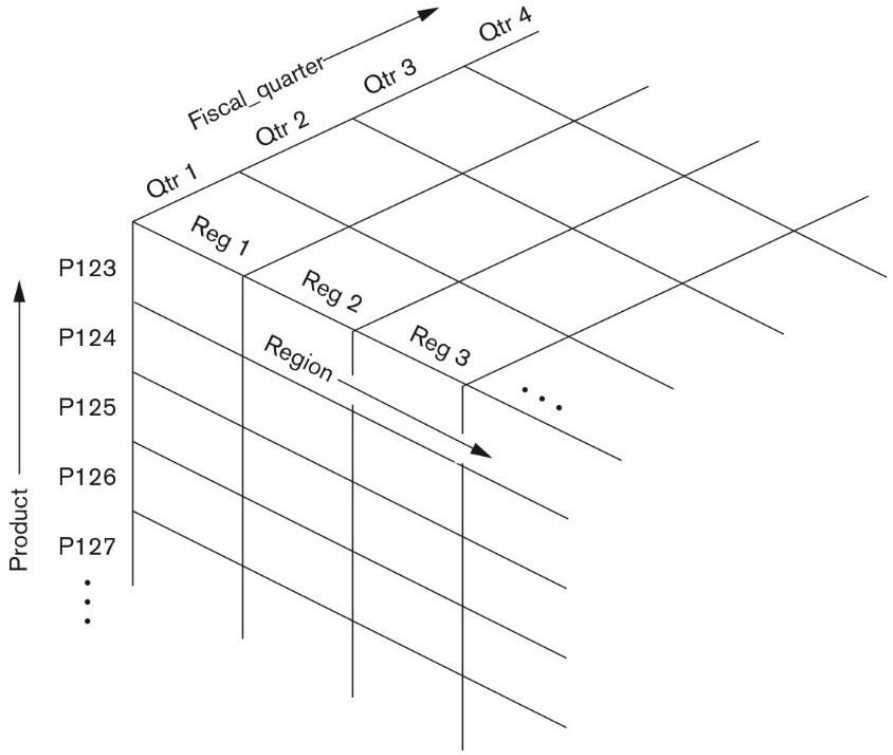


Figure 29.3 A three-dimensional data cube model.

Data Warehouse Design

- Multi-dimensional model (also called ”dimensional model”) includes two types of tables:
  - Dimension table
    - Consists of tuples of attributes of the dimension.
  - Fact table
    - Each tuple is a recorded fact. This fact contains some measured or observed variable (s) and identifies it with pointers to dimension tables. The fact table contains the data, and the dimensions to identify each tuple in the data.
    - A fact table is as an agglomerated view of transaction data whereas each dimension table represents “master data” that those transactions belonged to.

Data Warehouse Design

- The builders of Data warehouse should take a broad view of the anticipated use of the warehouse.
- The design should harmonize dimensions across the whole enterprise and multiple data sources
- The design should support ad-hoc querying
- An appropriate schema should be chosen that reflects the anticipated usage and the business model of the organization.

Data Warehouse Design

Information Subject: Automaker Sales

Dimensions

Hierarchies / Categories

Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		

Facts: Actual Sale Price, MSRP Sale Price, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance

Information Subject: Hotel Occupancy

Dimensions

Hierarchies / Categories

Time	Hotel	Room Type			
Year	Hotel Line	Room Type			
Quarter	Branch Name	Room Size			
Month	Branch Code	Number of Beds			
Date	Region	Type of Bed			
Day of Week	Address	Max. Occupants			
Day of Month	City/State/ Zip	Suite			
Holiday Flag	Construc-tion Year	Refrige-rator			
	Renova-tion Year	Kichen-nette			

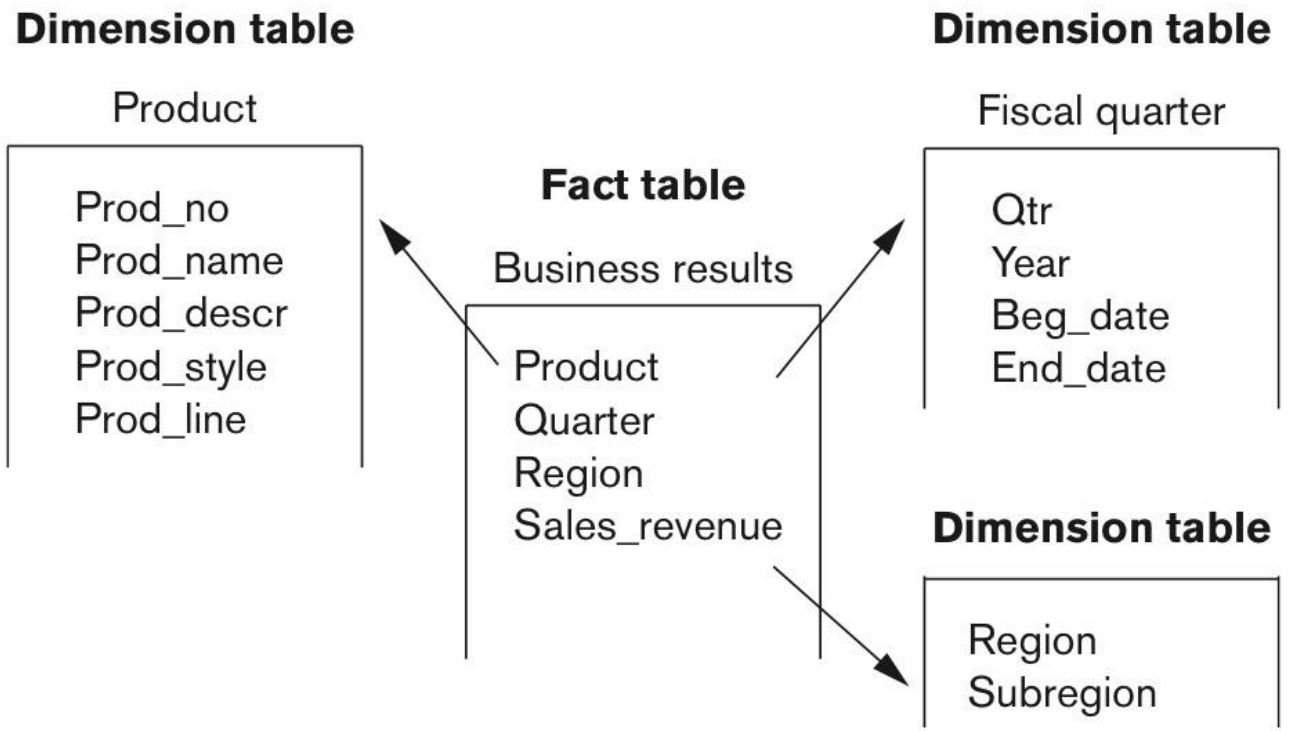
Facts: Occupied Rooms, Vacant Rooms, Unavailable Rooms, Number of Occupants, Revenue

Data Warehouse Schemas

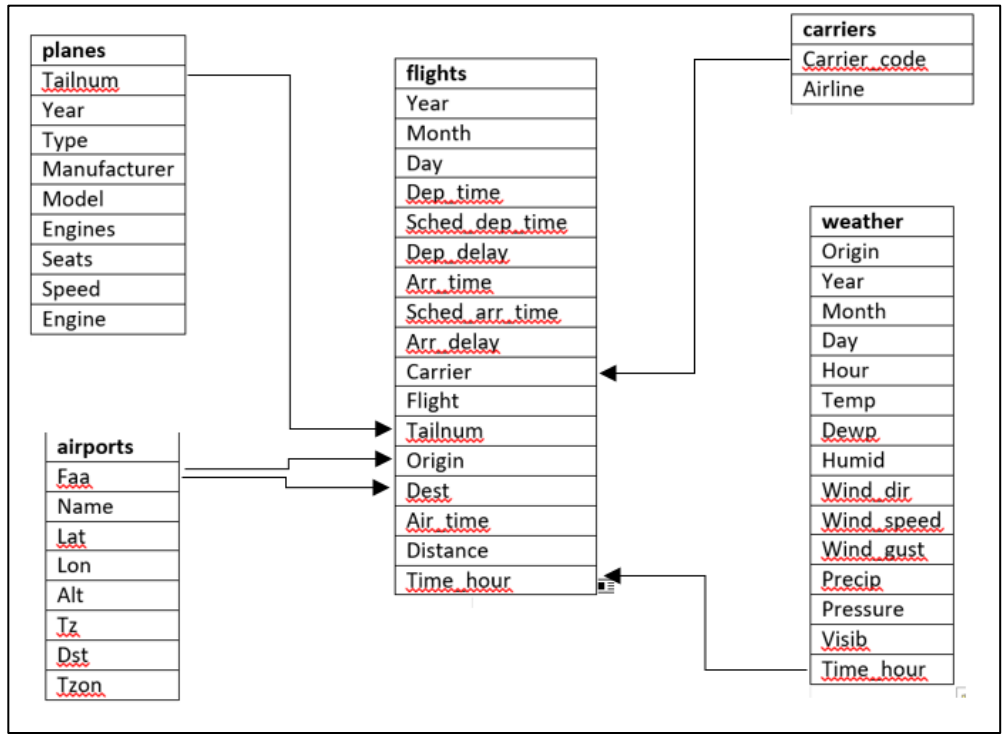
- Two common multi-dimensional schemas are
  - Star schema:
    - Consists of a fact table with a single table for each dimension
  - Snowflake Schema:
    - It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.

Star Schema

Fact table with a single table for each dimension.

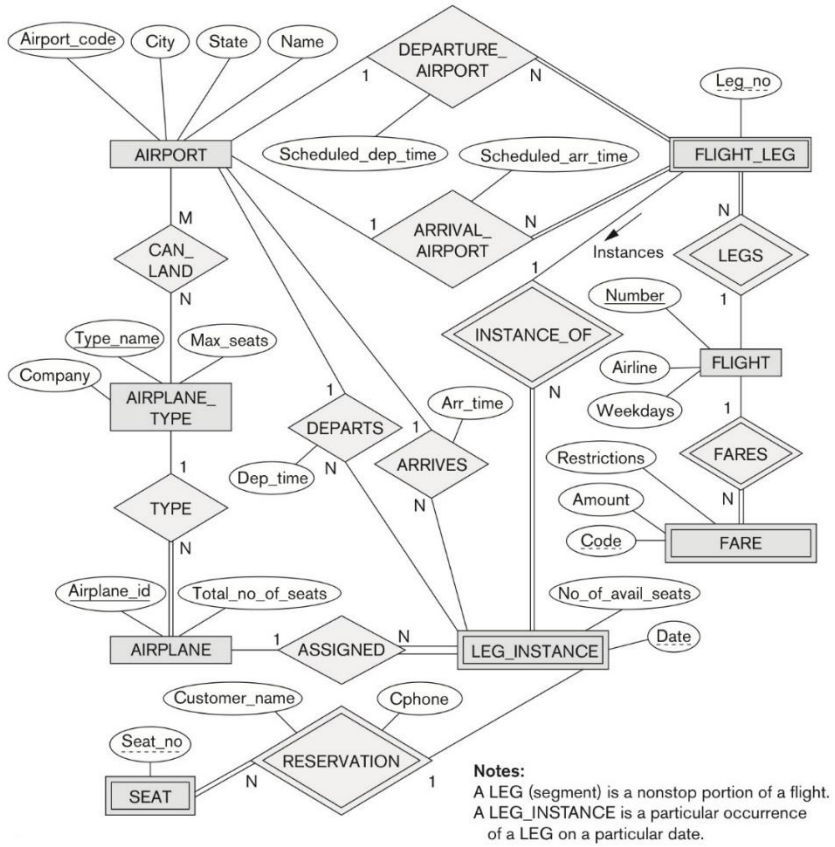


Star Schema



Flights data warehouse

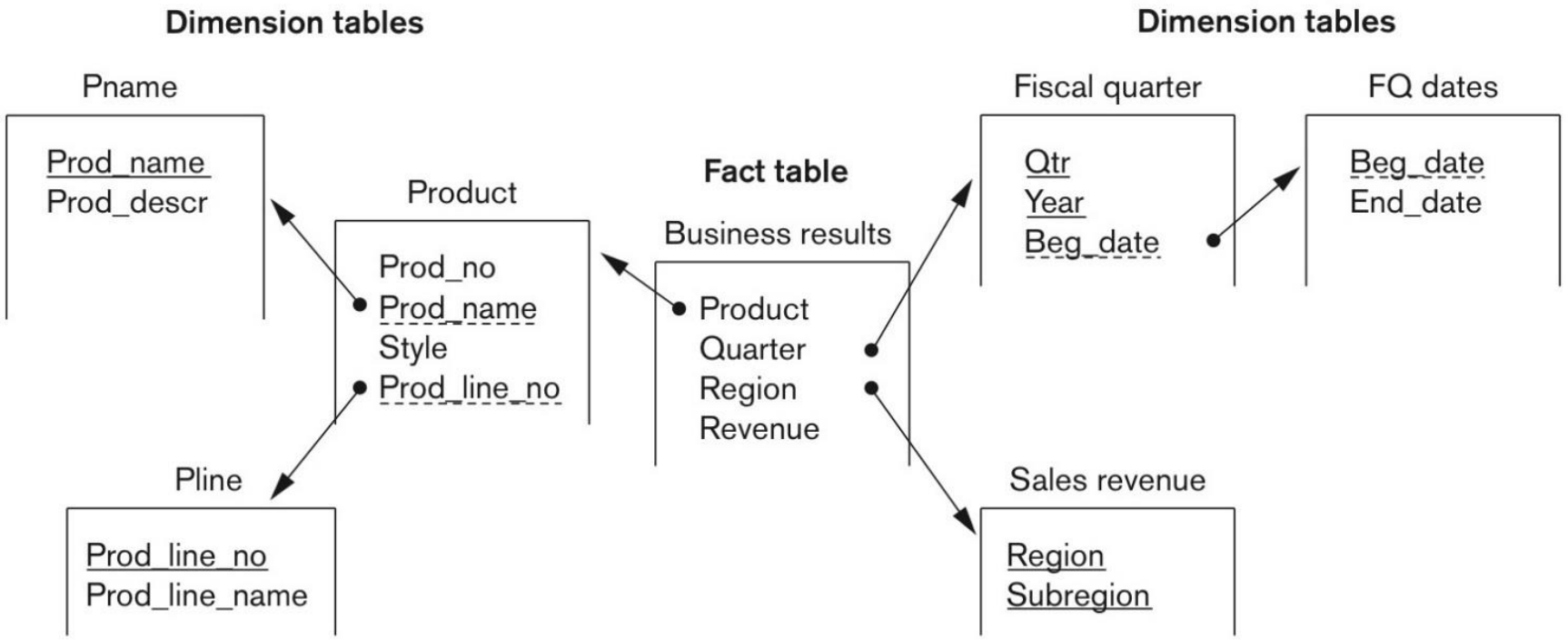
Vs.



Airline reservation (transactional) system

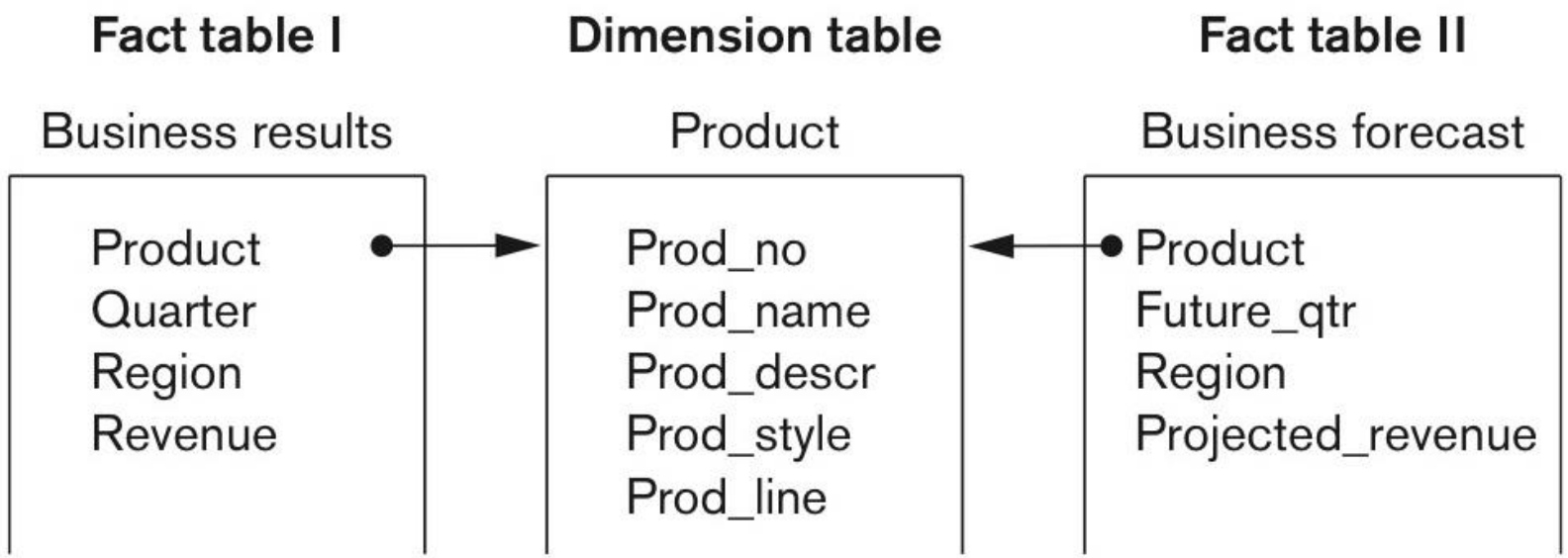
Snowflake Schema

A variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.



Fact Constellation

A set of tables that share some dimension tables. Note: fact constellations limit the possible queries for the warehouse.



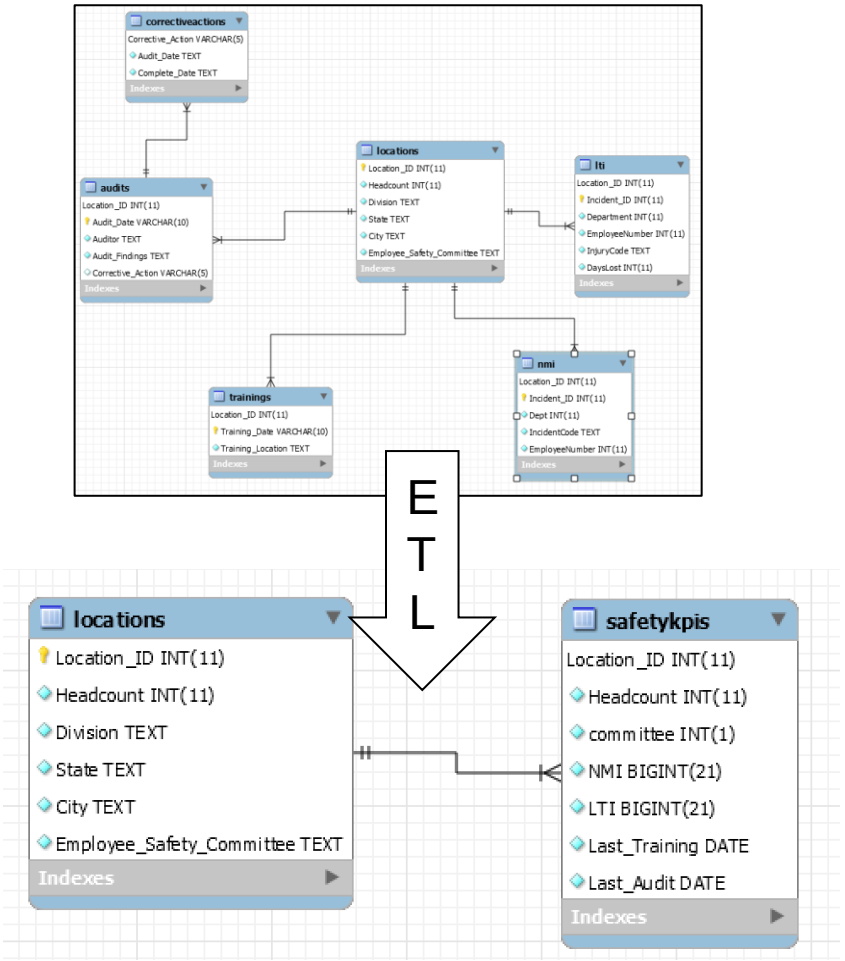
Fact Table Granularity

- Level of detail of one row of the fact able
- Higher (lower) granularity implies more (fewer) rows
- Trade-off between level of detailed analysis and storage requirements
- Examples:
  - One row of the fact table corresponds to one line on a purchase order
  - One row of the fact table corresponds to one purchase order
  - One row of the fact able corresponds to all purchase orders made by a customer

Data Warehouse Creation (ETL)

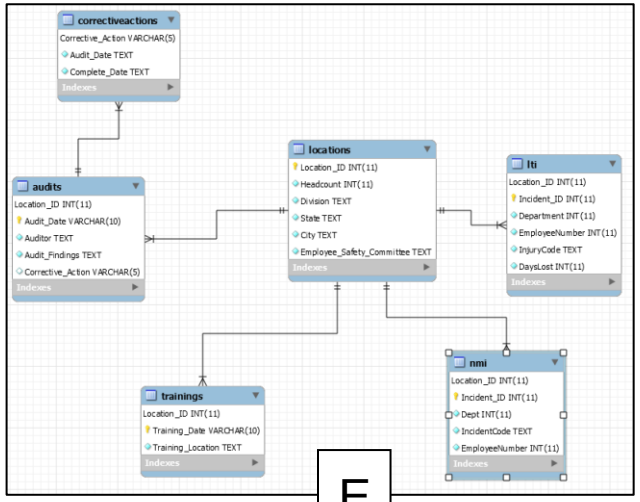
- Acquisition of data for the warehouse
  - The data must be extracted from multiple, heterogeneous sources.
  - Data must be formatted for consistency within the warehouse.
  - The data must be cleaned to ensure validity.
    - Difficult to automate cleaning process.
    - Back flushing: refers to upgrading the source data by returning cleaned data.
  - The data must be fitted into the data model of the warehouse. Data may have to be converted from its source model into a multi-dimensional format.
  - The data must be loaded into the warehouse.
    - Proper design for refresh policy should be considered.

Data Warehouse Example 1 – Safety KPIs

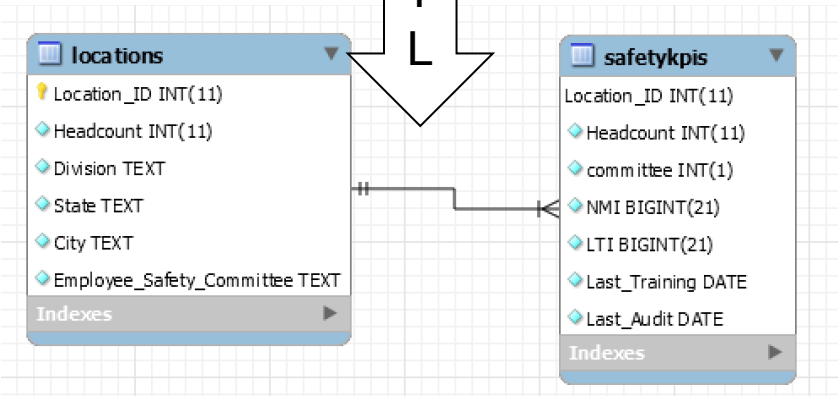



```
224  -- Data Warehouse - kpis view
225  CREATE OR REPLACE VIEW safetykpis AS
226  SELECT locations.Location_ID, Headcount,
227         IF(Employee_Safety_Committee = "yes", 1, 0) AS committee,
228         NMI, LTI, Last_Training, Last_Audit
229  FROM locations
230      JOIN    (   SELECT  Location_ID, COUNT(nmi.Incident_ID) AS NMI
231                  FROM nmi
232                  GROUP BY Location_ID
233              ) AS NMI_query
234          ON locations.Location_ID = NMI_query.Location_ID
235      JOIN    (   SELECT  Location_ID, COUNT(lti.Incident_ID) AS LTI
236                  FROM lti
237                  GROUP BY Location_ID
238              ) AS LTI_query
239          ON locations.Location_ID = LTI_query.Location_ID
240      JOIN    (   SELECT  Location_ID,
241                          MAX(STR_TO_DATE(Training_Date, "%m/%d/%Y")) AS Last_Training
242                  FROM trainings
243                  GROUP BY Location_ID
244              ) AS last_training
245          ON locations.Location_ID = last_training.Location_ID
246      JOIN    (   SELECT  Location_ID,
247                          MAX(STR_TO_DATE(Audit_Date, "%m/%d/%Y")) AS Last_Audit
248                  FROM audits
249                  GROUP BY Location_ID
250              ) AS last_audit
251          ON locations.Location_ID = last_audit.Location_ID;
```

Data Warehouse Example 1 – Safety KPIs



E
T
L



Location_ID	Incident_ID	Department	EmployeeNumber	InjuryCode	DaysLost
2408	1001	15	54095	Laceration	3
2408	1002	16	85971	Sprain	9
2408	1003	12	28325	Sprain	5
2408	1004	21	52011	Laceration	4
2408	1005	20	35602	Head Injury	25
2408	1006	10	94752	Laceration	4
2408	1007	16	73695	Laceration	4
2408	1008	19	20824	Sprain	8
2408	1009	19	38092	Burn	6
2408	1010	16	76448	Laceration	1
2408	1011	10	78585	Sprain	11
2408	1012	19	81880	Laceration	7
2408	1013	20	41819	Laceration	1

Location_ID	Training_Date	Training_Location
2408	11/27/2016	Offsite
2408	12/23/2016	Offsite
2408	12/29/2016	Offsite
2408	12/4/2016	Offsite
2408	3/31/2016	Offsite
2408	7/5/2017	Onsite
2415	12/2/2017	Onsite
2415	3/11/2016	Onsite
2415	8/31/2016	Onsite
2417	12/26/2015	Onsite
2417	3/6/2017	Onsite
2417	4/10/2015	Offsite
2417	5/9/2016	Offsite

Location_ID	Headcount	committee	NMI	LTI	Last_Training	Last_Audit
2408	554	1	81	37	2017-07-05	2017-11-06
2415	181	0	80	47	2017-12-02	2015-04-04
2417	327	1	76	33	2017-03-06	2017-05-09
2440	273	1	79	38	2017-12-27	2015-08-06
2453	228	1	99	45	2016-08-01	2015-07-09
2464	183	1	81	30	2016-09-04	2016-10-11
2468	365	1	105	43	2016-06-17	2016-02-02
2493	430	1	110	48	2016-10-06	2015-11-10
2505	159	1	99	42	2016-07-02	2015-04-09
2528	224	1	63	28	2017-11-04	2016-09-13
2534	422	1	99	37	2016-03-18	2017-07-13
2551	196	1	73	35	2017-08-04	2016-01-16
2554	316	1	118	51	2016-02-18	2015-02-05

See 4_1_safetyDW.sql on Blackboard