DS-306 Data Warehousing and Business Intelligence

Topic 2: Overview of DW and BI

Dr. Khurram Shahzad

Operational Sources (OLTP's)

- Operational computer systems did provide information to run day-to-day operations, and answer's daily questions, but…
- Also, called online transactional processing system (OLTP)
- Data is read or manipulated with each transaction
- Transactions/queries are simple, and easy to write
- Usually for middle management

Typical decision queries

- Decision-making require complex questions from integrated data
- Enterprise-wide data is desired
- Decision makers want to know:
  - Where to build new oil warehouse?
  - Which market they should strengthen?
  - Which customer groups are most profitable?
  - How much is the total sale by month/ year/ quarter for each offices?
  - Is there any relation between promotion campaigns and sales growth?
- **Can OLTP answer all such questions, efficiently?**

Failure of old OLTPs

- Inability to provide strategic information
- IT receive too many ad hoc requests
- Requests are not only numerous, they change overtime, because for more understanding more reports
- Users are in spiral of reports
- Users have to depend on IT for information
- Can't provide enough performance, slow
- Strategic information have to be flexible and conductive, analysis driven (not report driven analysis)

Expectations of new soln.

- DB designed for analytical tasks
- Data from multiple applications
- Easy to use
- Ability of what-if analysis, (analysis driven)
- Read-intensive data usage
- Direct interaction with system, without IT assistance
- Periodical updating contents & stable
- Current & historical data
- Ability for users to initiate reports

DW meets expectations

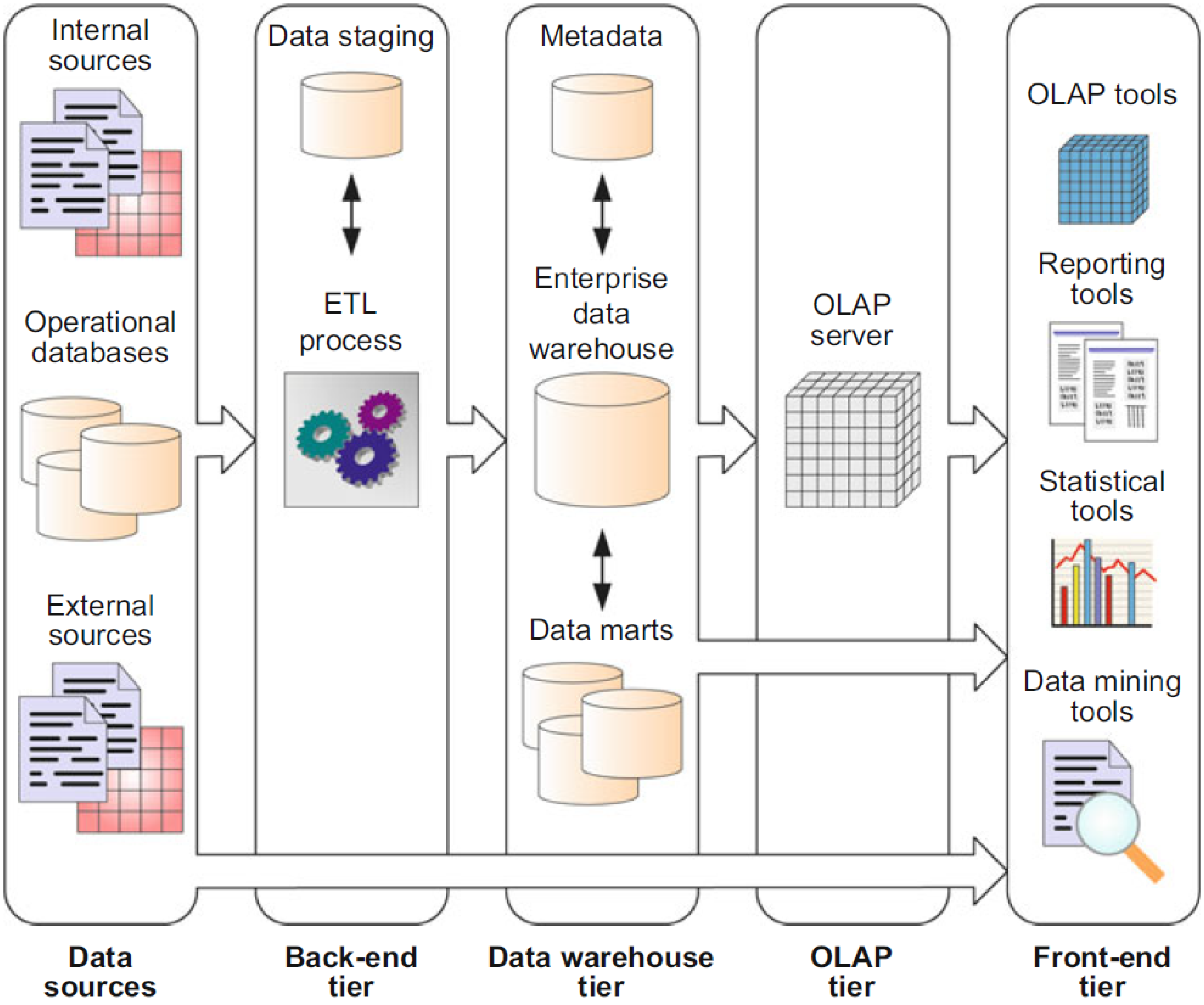
- Provides enterprise view
- Current & historical data available
- Decision-transaction possible without affecting (locking) operational source
- Reliable source of information
- Ability for users to initiate reports
- Acts as a data source for all analytical applications

OLTP vs. BI

Trait	OLTP	BI
User	Middle management	Executives, decision-makers
Function	For day-to-day operations	For analysis & decision support
DB (modeling)	E-R based, after normalization	Star oriented schemas
Data	Current, Isolated	Archived, derived, summarized
Unit of work	Transactions	Complex query
Access, type	DML, read	Read
Access frequency	Very high	Medium to Low
Records accessed	Tens to Hundreds	Thousands to Millions
Quantity of users	Thousands	Very small amount
Usage	Predictable, repetitive	Ad hoc, random, heuristic
DB size	100 MB-GB	100GB-TB
Response time	Sub-seconds	Up-to min.

Architecture of BI solution

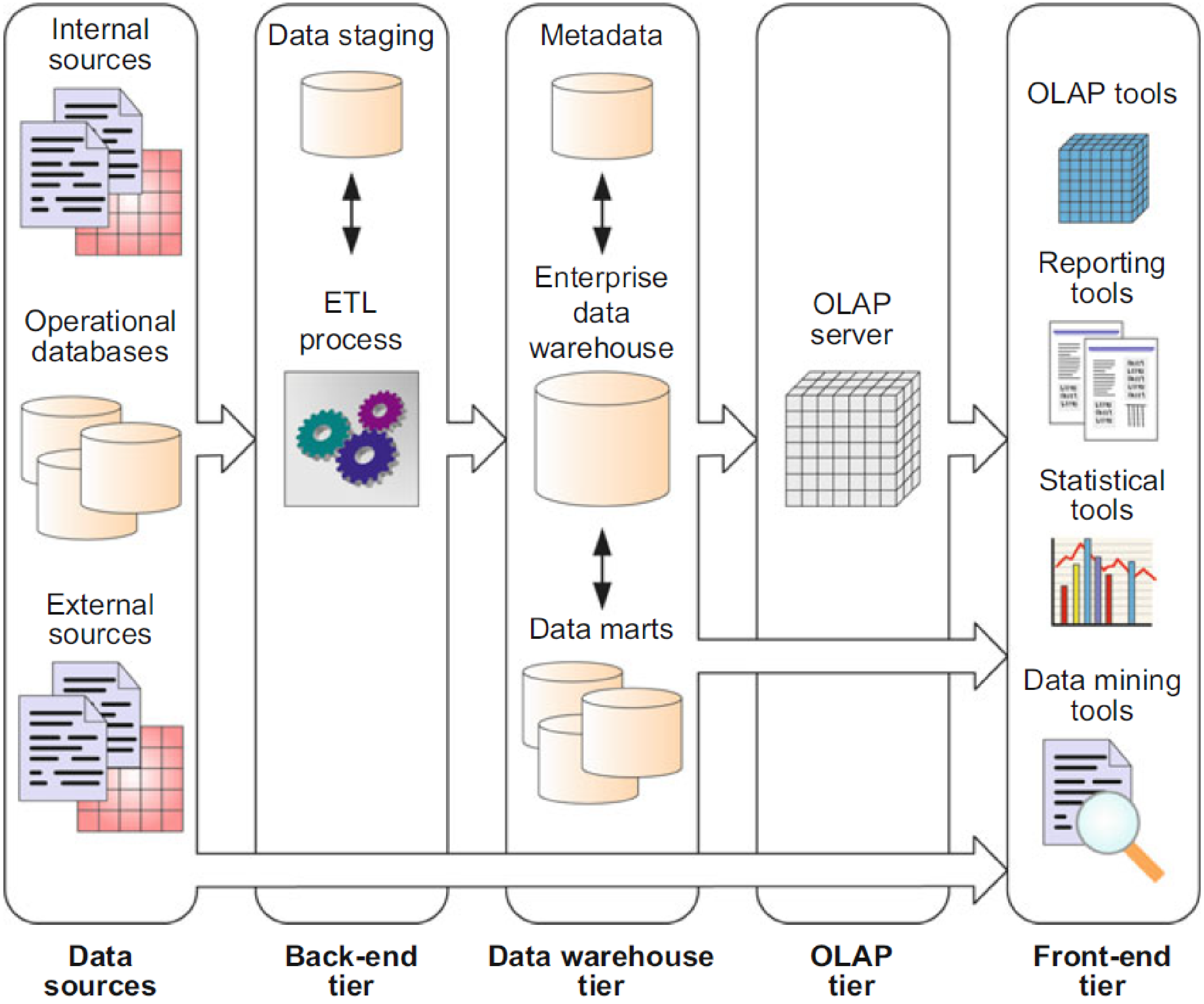
Typical DW architecture



1. Data Sources

- Source data can be grouped into 4 components
  - Production data
    - Comes from operational systems of enterprise
    - Some segments are selected from it
    - Narrow scope, e.g. order details
  - Internal data
    - Private datasheet, documents, customer profiles etc.
    - E.g. Customer profiles for specific offering
    - Special strategies to transform ‘IT’ to DW (text document)
  - Archived data
    - Old data is archived
    - DW have snapshots of historical data
  - External data
    - Executives depend upon external sources
    - E.g. market data of competitors, car rental require new manufacturing. Define conversion

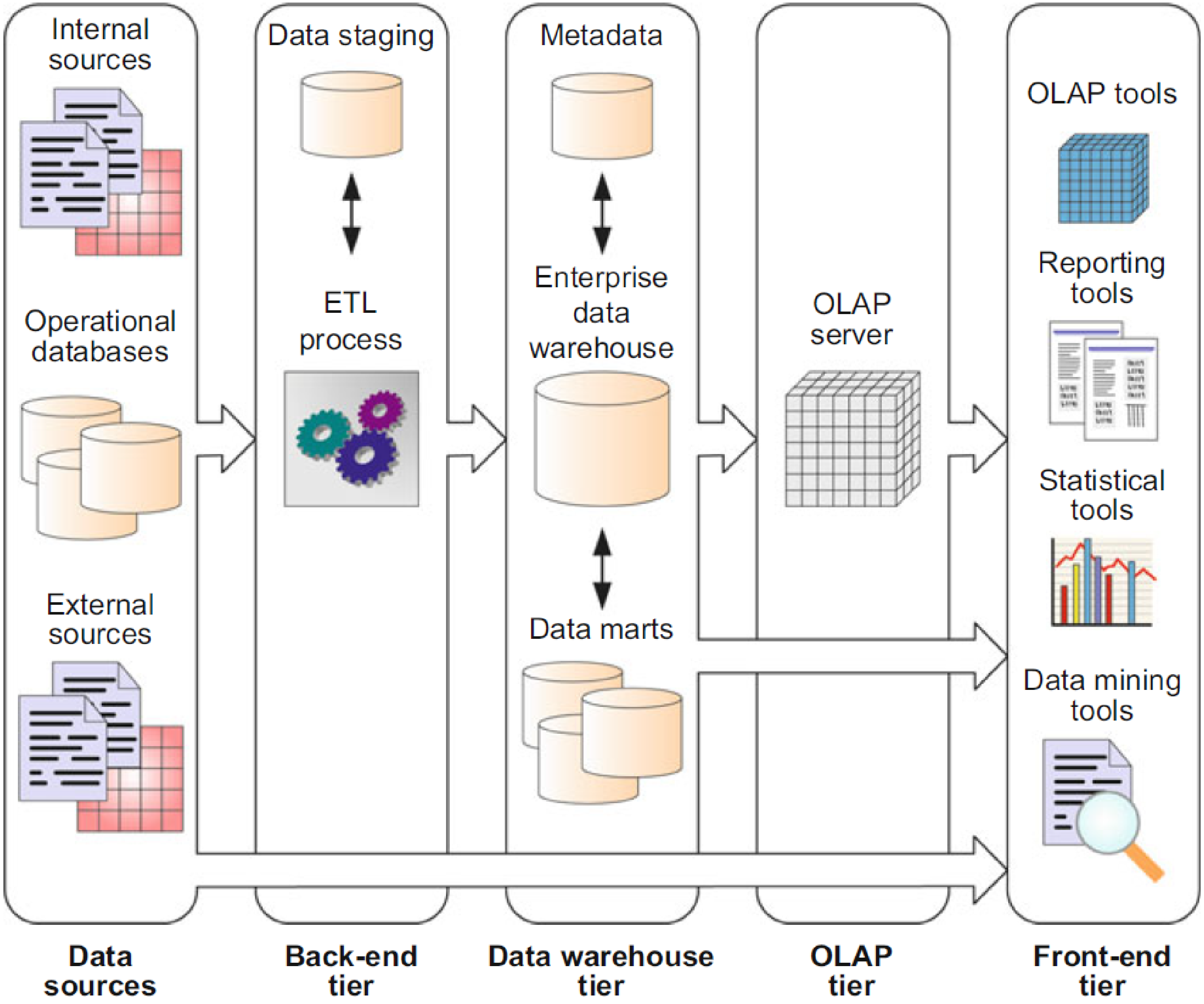
Typical DW architecture



2. Back-end Tier

- After data is extracted, data is to be prepared
- Data extracted from sources needs to be changed, converted and made ready in suitable format
- Three major functions to make data ready
  - Extract
  - Transform
  - Load
- Staging area provides a place and area with a set of functions to
  - Clean
  - Change
  - Combine
  - Convert

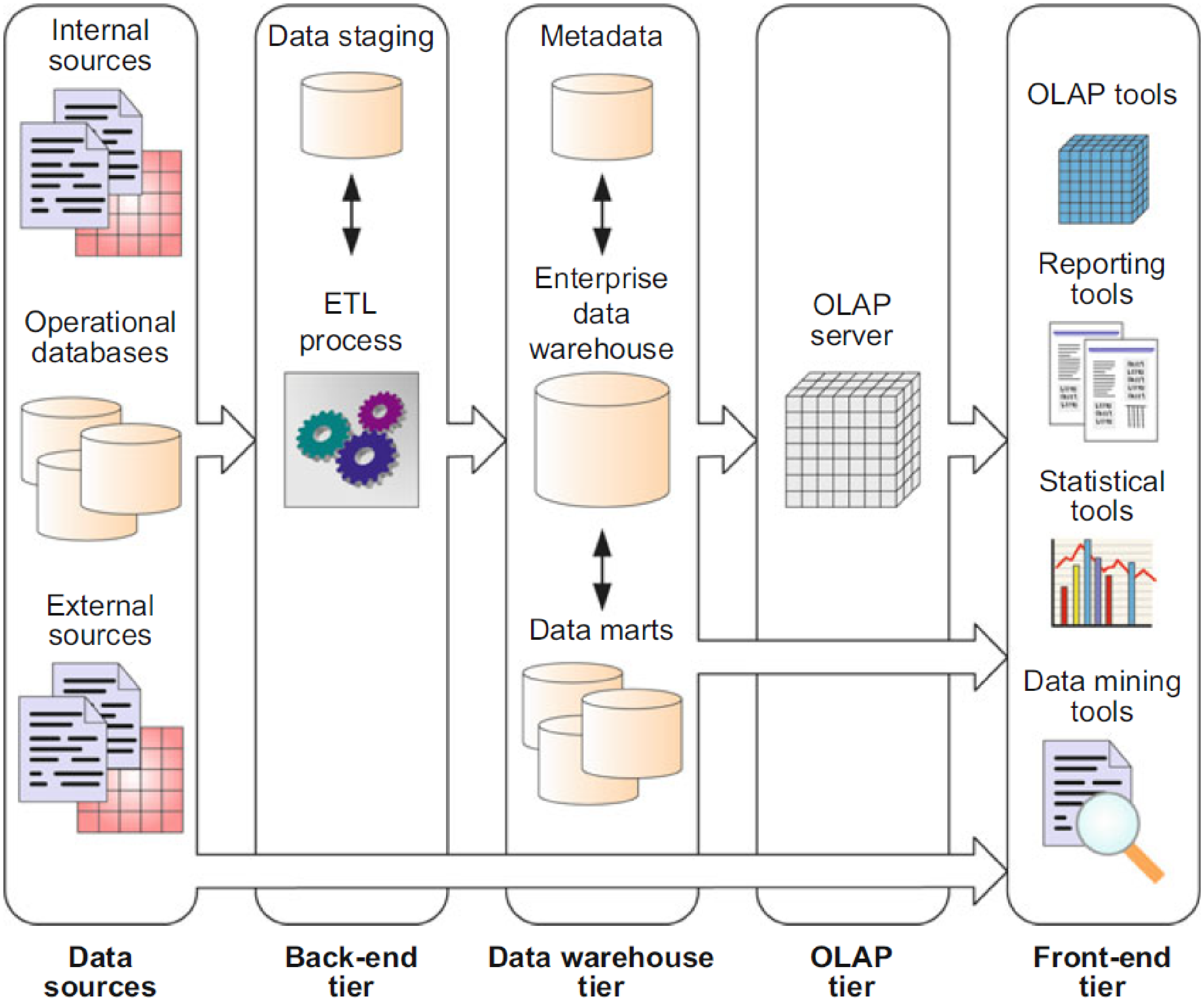
Typical DW architecture



3. Data Warehouse Tier

- Separate repository
- Data structured for efficient processing
- Redundancy is increased
- Updated after specific periods
- Only read-only

Typical DW architecture



4. Information Delivery Component

- Authentication issues
- Active monitoring services
  - Performance
  - User performance
  - Aggregate awareness
  - E.g. mining, OLAP etc

Definition of DW

Data Warehouse

- A data warehouse is a particular database targeted toward decision support.

- It takes data from **various operational databases** and other data sources and **transforms** it into **new structures** that **fit** better for the task of performing business analysis.

- DWs are based on a **multidimensional model**, where data are represented as **dimensions** corresponding to the various business perspectives and cube cells containing the **measures** to be analyzed.

Definition of DW

Inmon defined

“A DW is a subject-oriented, integrated, non-volatile, time-variant collection of data in favor of decision-making”.

Kelly said

“Separate available, integrated, time-stamped, subject-oriented, non-volatile, accessible”

Four properties of DW

Subject-oriented

- In operational sources data is organized by applications, or business processes.
- In DW, subject is the organization method
- Subjects vary with enterprise
- These are critical factors, that affect performance
- Example of Manufacturing Company
  - Sales
  - Shipment
  - Inventory, etc.

Integrated Data

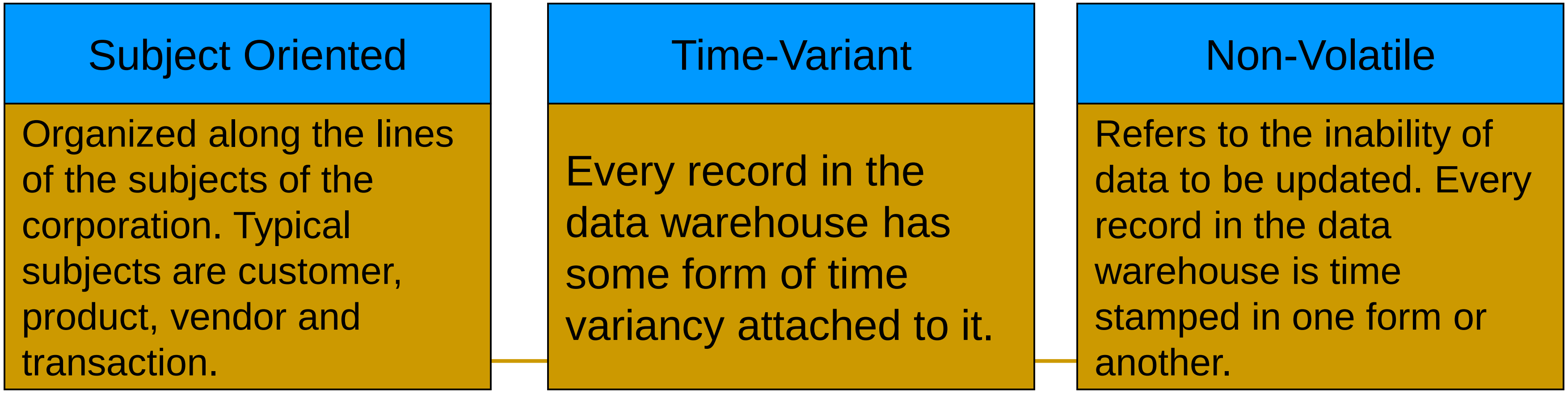
- Data comes from several applications
- Problems of integration comes into play
  - File layout, encoding, field names, systems, schema, data heterogeneity are the issues
  - Bank example, variance: naming convention, attributes for data item, account no, account type, size, currency
- In addition to internal, external data sources
  - External companies data sharing
  - Websites
  - Others
- Removal of inconsistency
- So process of extraction, transformation & loading

Time variant

- Operational data has current values
- Comparative analysis is one of the best techniques for business performance evaluation
- Time is critical factor for comparative analysis
- Every data structure in DW contains time element
- In order to promote certain products, analyst has to know about current and historical values
- The advantages are
  - Allows for analysis of the past
  - Relates information to the present
  - Enables forecasts for the future

Non-volatile

- Data from operational systems are moved into DW after specific intervals
- Data is persistent/ not removed i.e. non volatile
- Every business transaction don’t update in DW
- Data from DW is not deleted
- Data is neither changed by individual transactions
- Properties summary



Subject Oriented

Organized along the lines of the subjects of the corporation. Typical subjects are customer, product, vendor and transaction.

Time-Variant

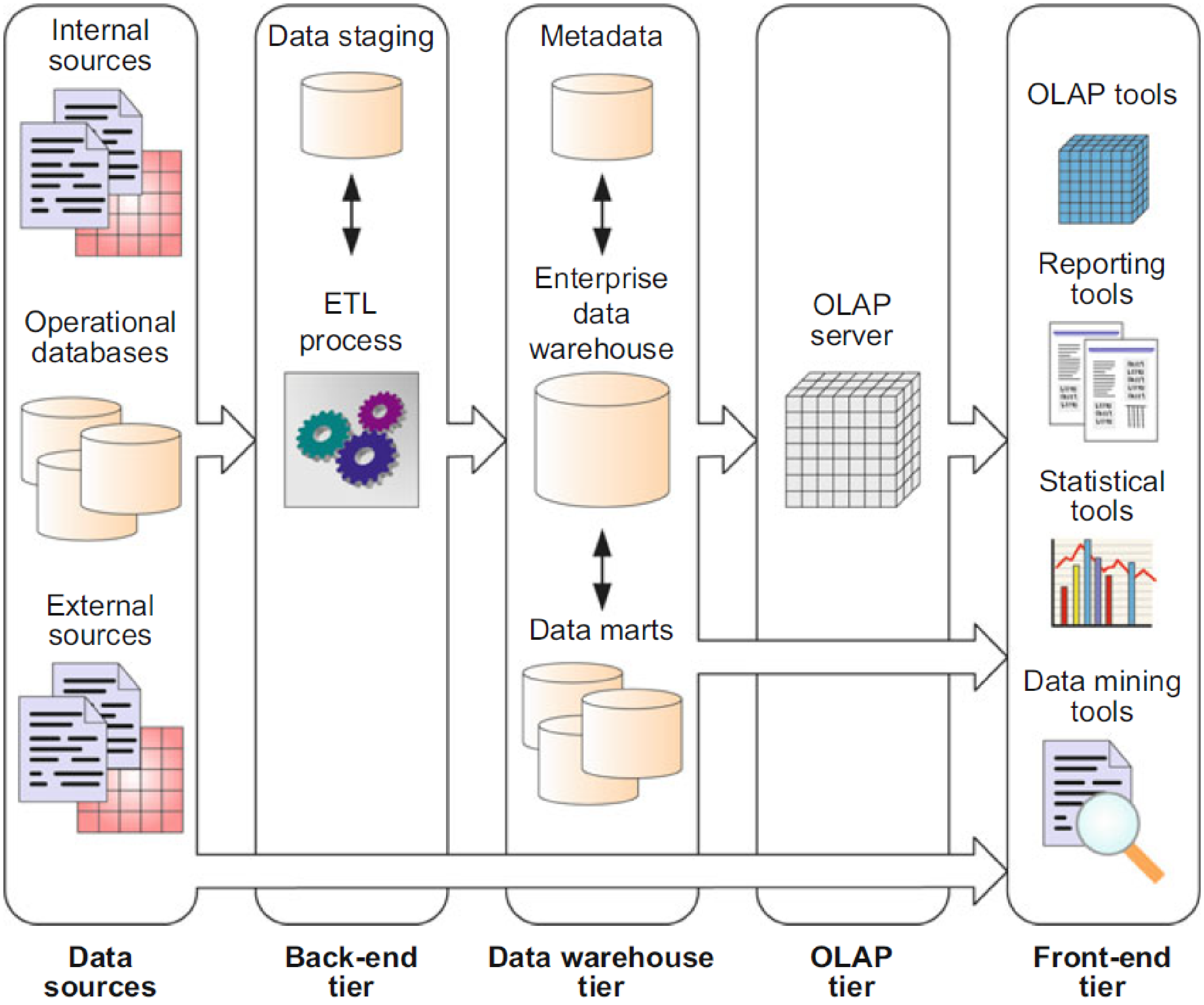
Every record in the data warehouse has some form of time variancy attached to it.

Non-Volatile

Refers to the inability of data to be updated. Every record in the data warehouse is time stamped in one form or another.

Multidimensional Model

Typical DW architecture



Dimensional Model

- Data warehouses and OLAP systems are based on **multidimensional model**
- Dimensional modeling focuses subject-orientation, critical factors of business
- Critical factors are stored in **facts**

Dimensional Model

- Logical design technique for high performance
- Each model represent a subject in DW
- Is the modeling technique for storage

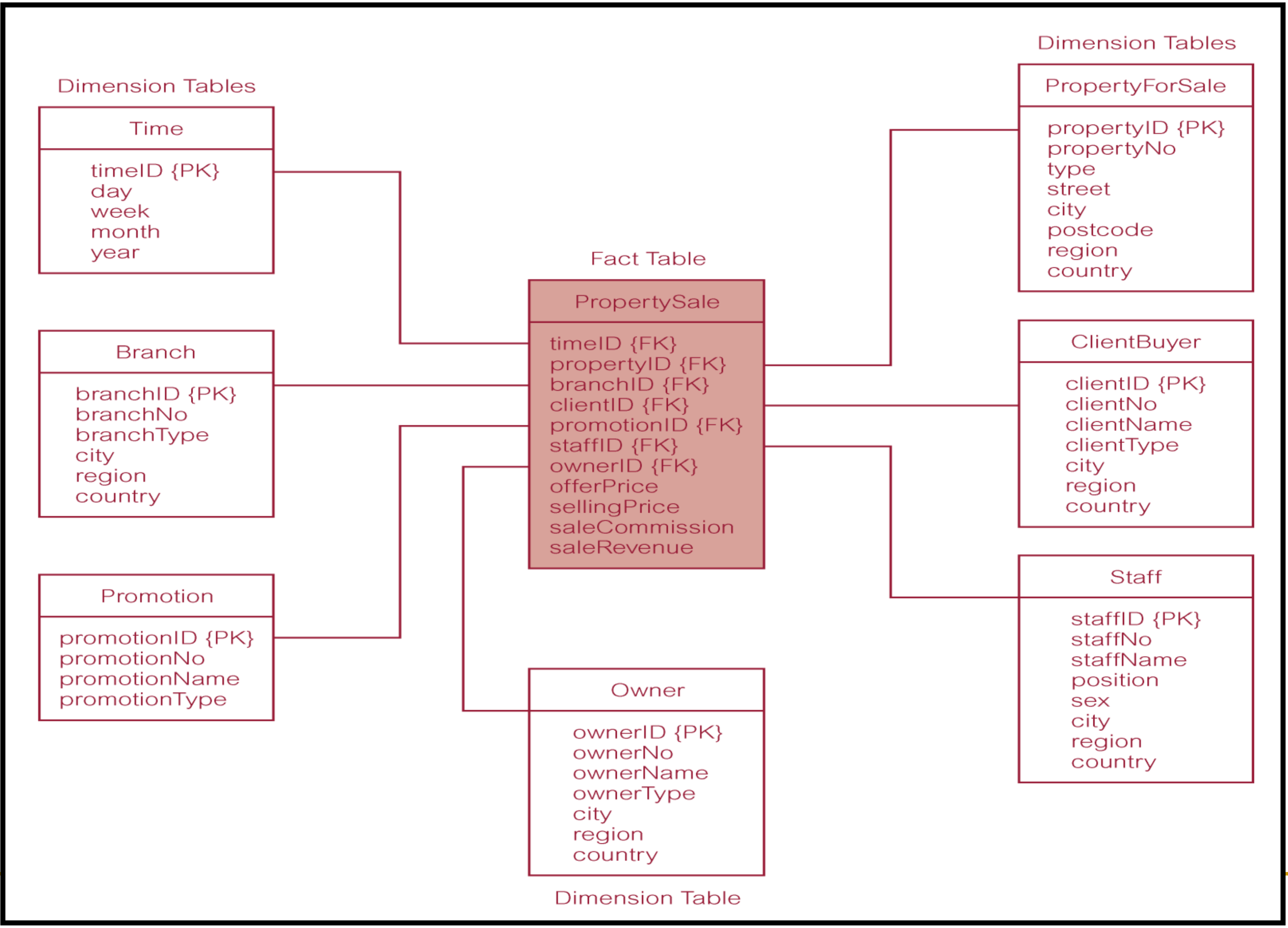
Dimensional Model

- Two important concepts
  - Fact
    - Numeric measurements, represent business activity/event
    - Are pre-computed, redundant
    - Example: Profit, quantity sold
  - Dimension
    - Qualifying characteristics, perspective to a fact
    - Example: date (Date, month, quarter, year), product(type, category)

Dimensional Model

- Every dimensional model (DM) is composed of one (or more) fact tables, and a set of smaller dimension tables.
- Look on Fact table through one (or more) dimensions.
  - What is the sale amount in Consumer Product category, for elderly customers in the second quarter of 2004?
- Forms ‘star-like’ structure, which is called a star schema or star join.

A Typical Dimensional Model



Dimension Tables

Time

timeID {PK}
day
week
month
year

Branch

branchID {PK}
branchNo
branchType
city
region
country

Promotion

promotionID {PK}
promotionNo
promotionName
promotionType

Fact Table

PropertySale

timeID {FK}
propertyID {FK}
branchID {FK}
clientID {FK}
promotionID {FK}
staffID {FK}
ownerID {FK}
offerPrice
sellingPrice
saleCommission
saleRevenue

Owner

ownerID {PK}
ownerNo
ownerName
ownerType
city
region
country

Dimension Table

Dimension Tables

PropertyForSale

propertyID {PK}
propertyNo
type
street
city
postcode
region
country

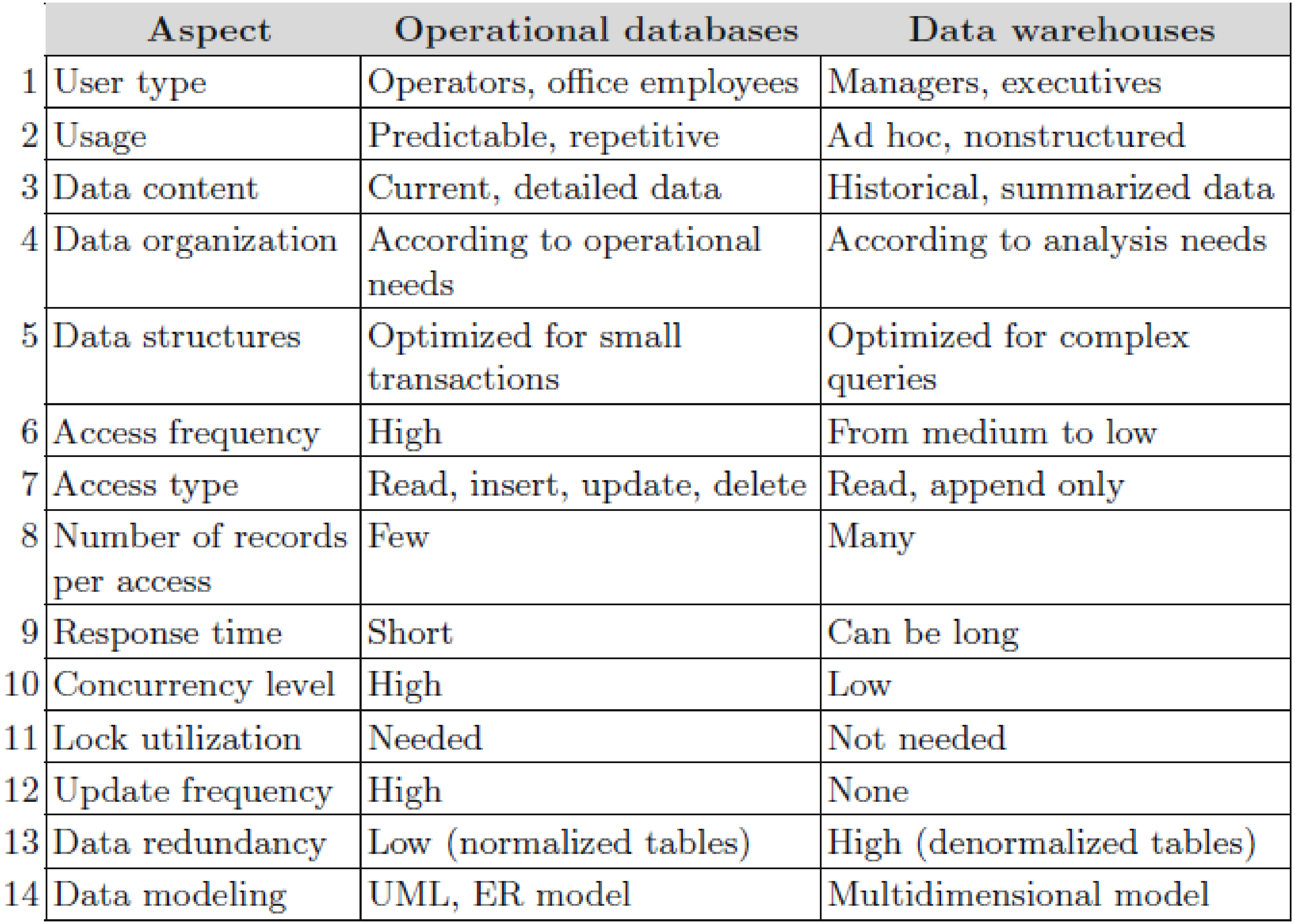
ClientBuyer

clientID {PK}
clientNo
clientName
clientType
city
region
country

Staff

staffID {PK}
staffNo
staffName
position
sex
city
region
country

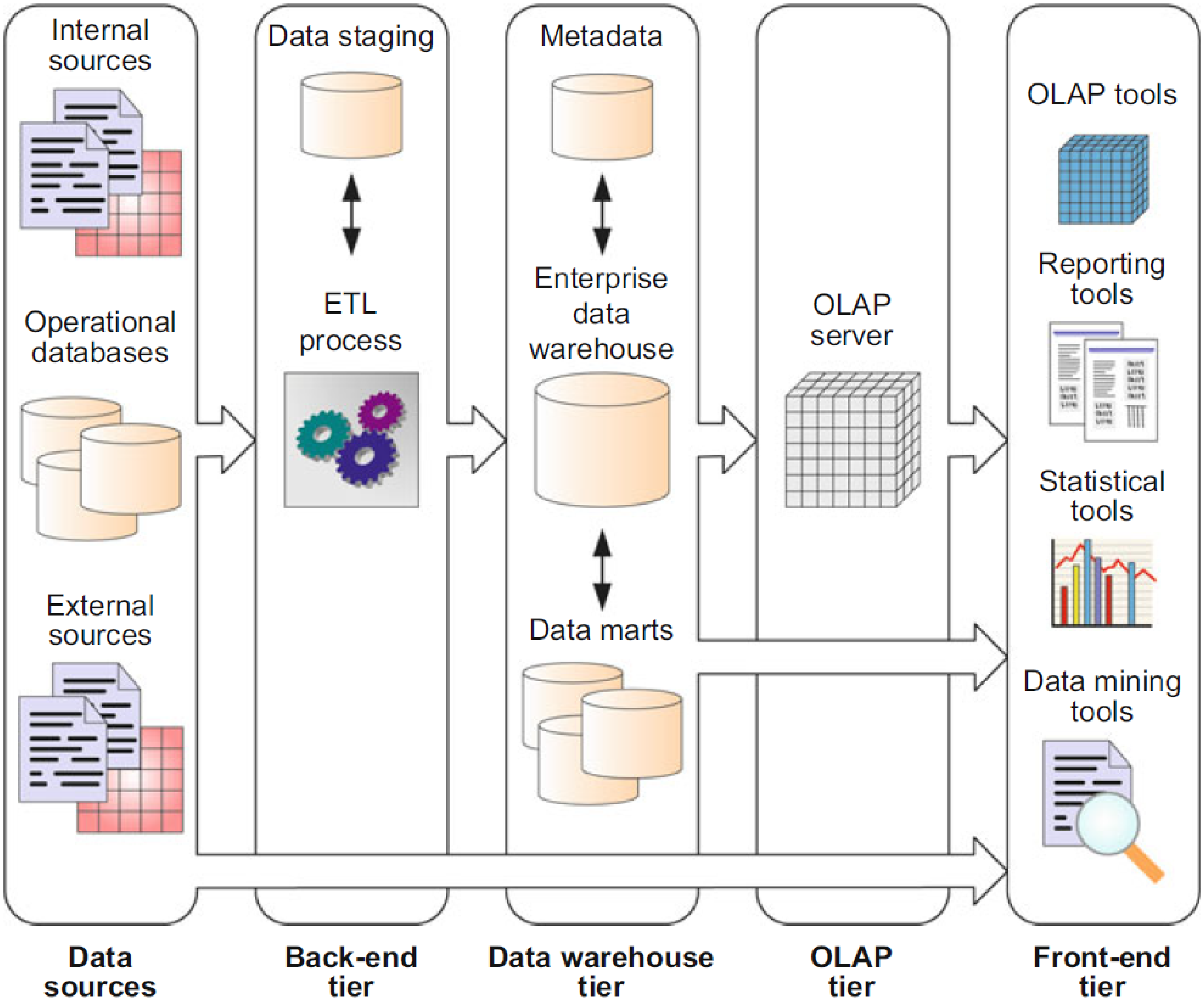
Operational DB vs DW



	Aspect	Operational databases	Data warehouses
1	User type	Operators, office employees	Managers, executives
2	Usage	Predictable, repetitive	Ad hoc, nonstructured
3	Data content	Current, detailed data	Historical, summarized data
4	Data organization	According to operational needs	According to analysis needs
5	Data structures	Optimized for small transactions	Optimized for complex queries
6	Access frequency	High	From medium to low
7	Access type	Read, insert, update, delete	Read, append only
8	Number of records per access	Few	Many
9	Response time	Short	Can be long
10	Concurrency level	High	Low
11	Lock utilization	Needed	Not needed
12	Update frequency	High	None
13	Data redundancy	Low (normalized tables)	High (denormalized tables)
14	Data modeling	UML, ER model	Multidimensional model

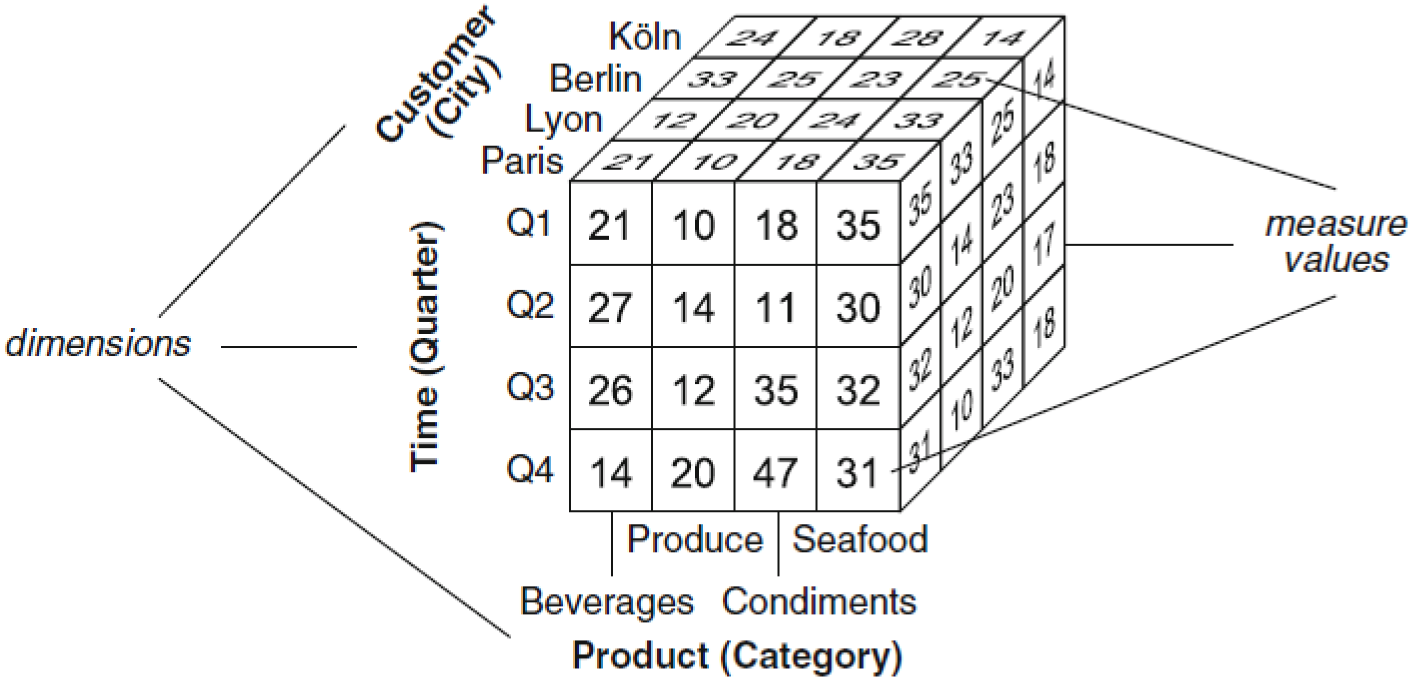
OLAP

Typical DW architecture



Data Cube

- A data cube is defined by **dimensions** and **facts**



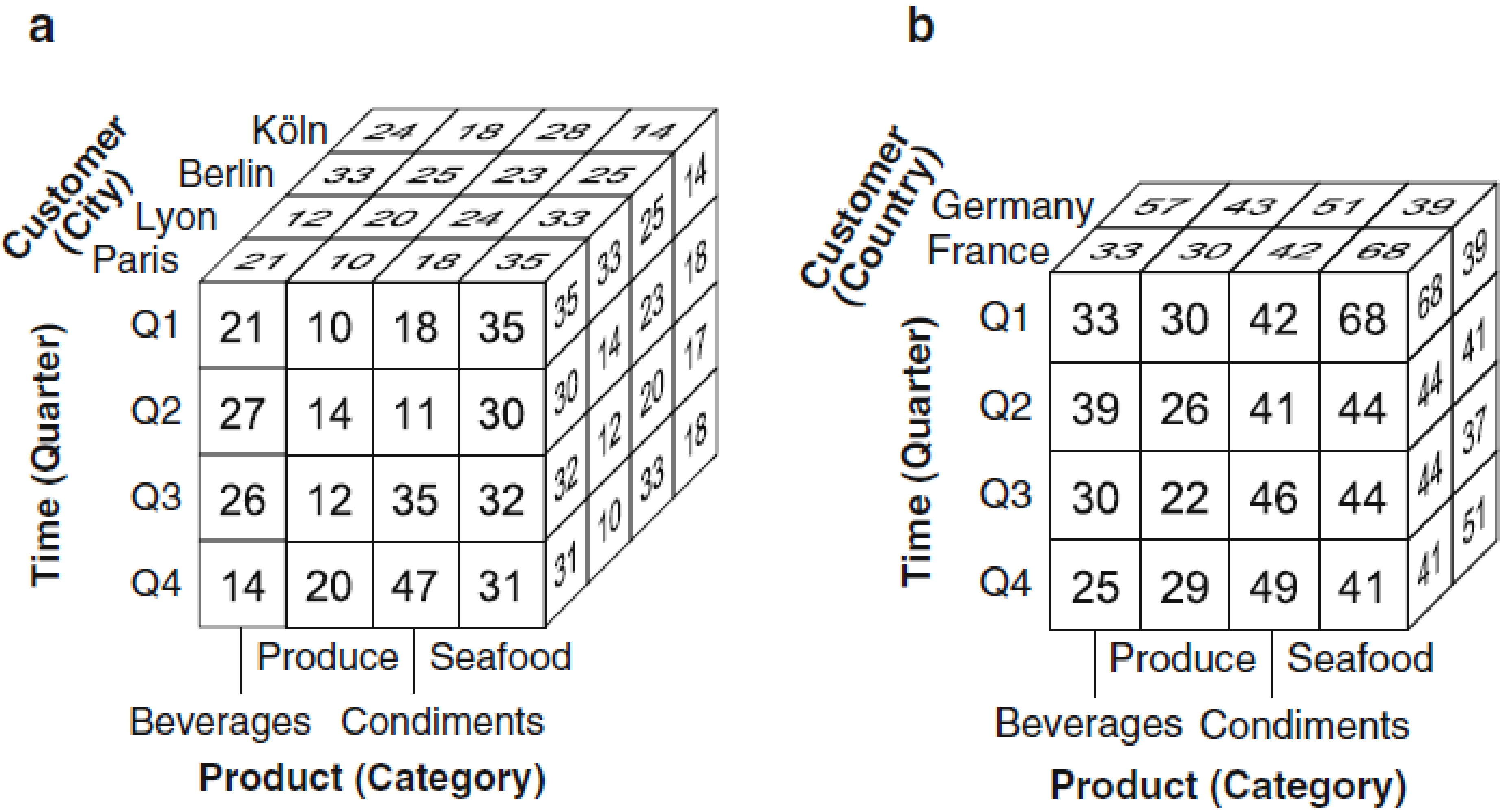
	Beverages	Produce	Condiments	Seafood
Q1	21	10	18	35
Q2	27	14	11	30
Q3	26	12	35	32
Q4	14	20	47	31

Cube

- The cube has three dimensions
  - Product
  - Time
  - Customer city

- The cells of a data cube, or **facts**, have associated **numeric values**

OLAP Operations



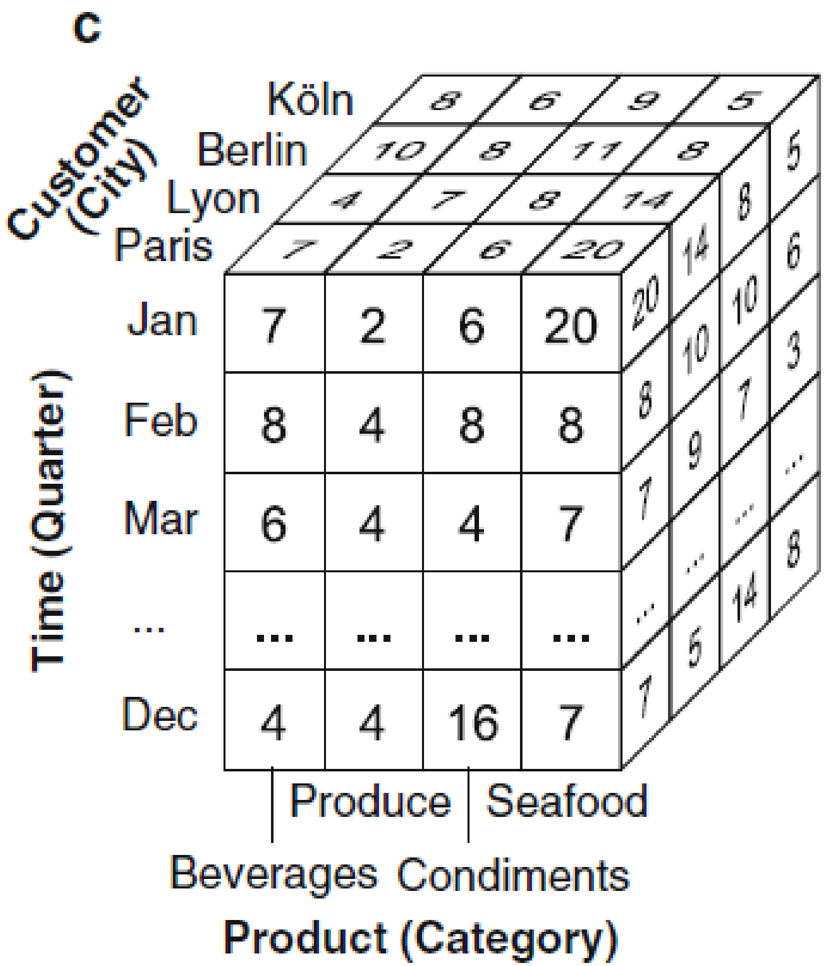
a

Time (Quarter)	Beverages	Produce	Condiments	Seafood
Q1	21	10	18	35
Q2	27	14	11	30
Q3	26	12	35	32
Q4	14	20	47	31

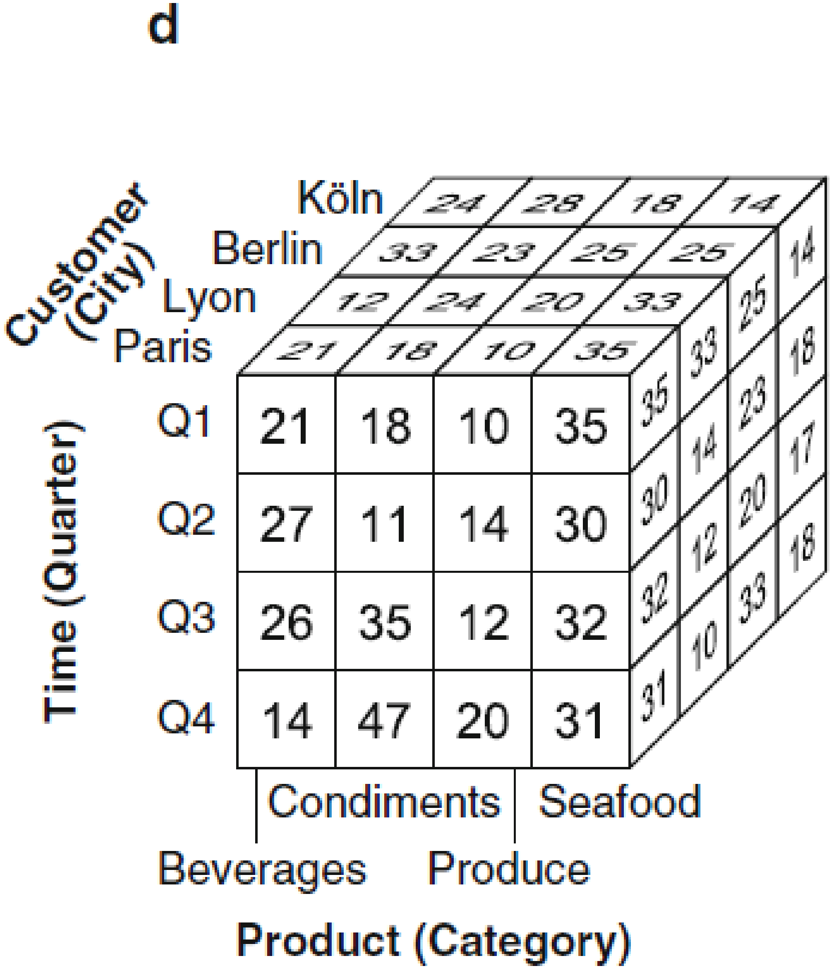
b

Time (Quarter)	Beverages	Produce	Condiments	Seafood
Q1	33	30	42	68
Q2	39	26	41	44
Q3	30	22	46	44
Q4	25	29	49	41

OLAP Operations



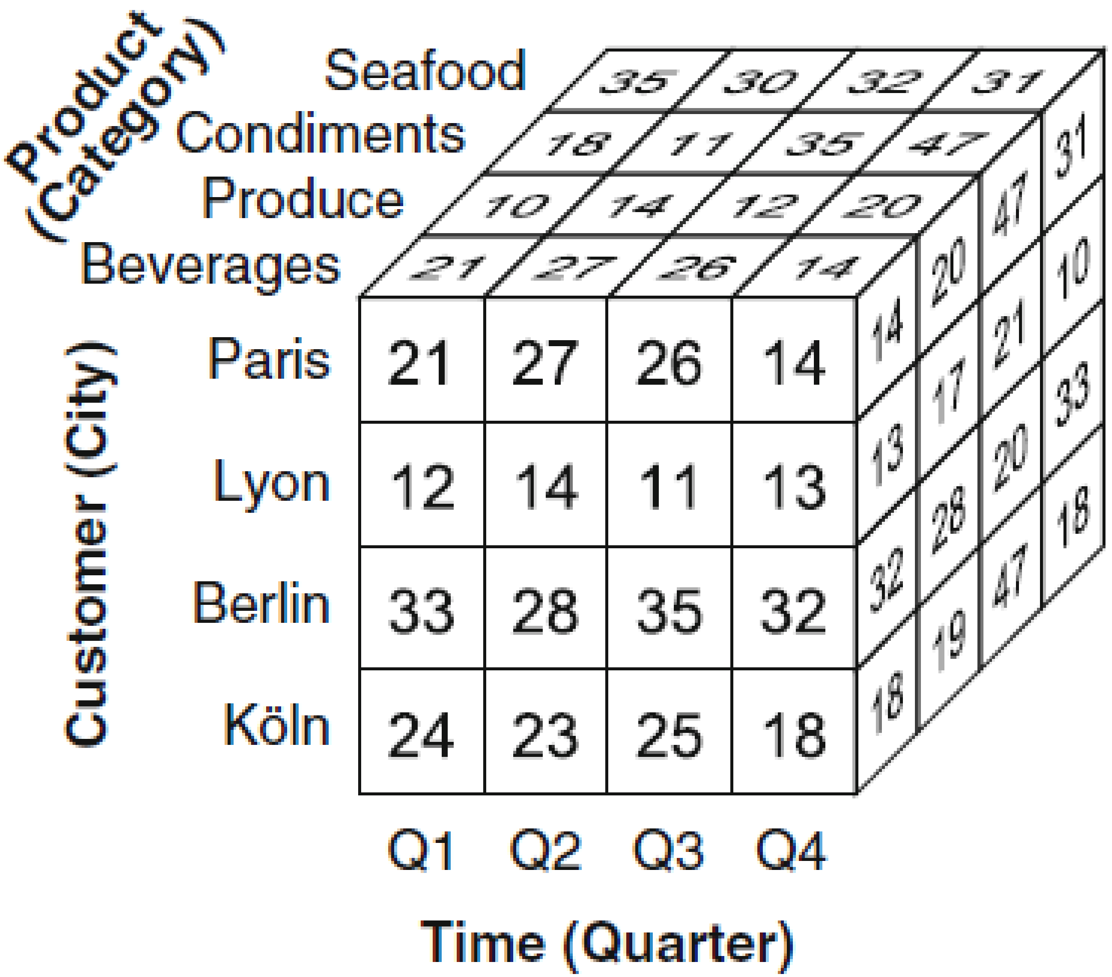
	Beverages	Produce	Condiments	Seafood
Jan	7	2	6	20
Feb	8	4	8	8
Mar	6	4	4	7
...	...	...	...	...
Dec	4	4	16	7



	Beverages	Condiments	Produce	Seafood
Q1	21	18	10	35
Q2	27	11	14	30
Q3	26	35	12	32
Q4	14	47	20	31

OLAP Operations

e



	Q1	Q2	Q3	Q4
Paris	21	27	26	14
Lyon	12	14	11	13
Berlin	33	28	35	32
Köln	24	23	25	18

Business Intelligence Tools

BI tools

- Microsoft
- Oracle
- IBM
- Teradata
- SAP
- Microstrategy
- Targit

Microsoft SQL Server tools

- Database Engine
- Integration services (SSIS)
- Analysis Services & Reporting Services