Data lakes

Data lakes vs Data warehouse

Dimension	Data Warehouse	Data Lake
The nature of data	Structured, processed	Any data in raw/native format
Processing	Schema-on-write (SQL)	Schema-on-read (NoSQL)
Retrieval speed	Very fast	Slow
Cost	Expensive for large data volumes	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile, flexible configuration
Novelty/newness	Not new/matured	Very new/maturing
Security	Well-secured	Not yet well-secured
Users	Business professionals	Data scientists

Data Warehouse (OLAP)	Operational Database(OLTP)
It involves historical processing of information.	It involves day-to-day processing.
OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
It is used to analyze the business.	It is used to run the business.
It focuses on Information out.	It focuses on Data in.
It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.

Fact Types

Feature	Transaction	Periodic	Accumulating
Grain	1 row/transaction	1 row/time-period	1 row/entire event stages
Date Dimension	Lowest granularity	End-of-period granularity	Multiple date
Facts	Transaction activities	Periodic activities	Defined lifetime activities
Size	Largest	Medium	Smallest
Update	No	No	Yes, after stage finished

Fact Types

Property	Transaction	Periodic	Accumulating
Grain	One row per transaction	One row per time period	One row per lifetime of an event
Date dimension	Lowest level of granularity	End-of-period granularity	Multiple per row
Number of dimensions	Least	Average	Most
Facts	Transaction related	Period related	Numerous events over lifetime
Measurement	Additive	Not additive, average	Need to derive
Conforming dimensions	Yes	Yes	Yes
Conforming facts	Yes	Yes	Yes
Database size	Largest	Smaller	Smallest

Data warehouse vs data mart

Data Warehouse	Data Mart
Scope • Application independent • Centralized, possibly enterprise-wide • Planned	**Scope** • Specific DSS application • Decentralized by user area • Organic, possibly not planned
Data • Historical, detailed, and summarized • Lightly denormalized	**Data** • Some history, detailed, and summarized • Highly denormalized
Subjects • Multiple subjects	**Subjects** • One central subject of concern to users
Sources • Many internal and external sources	**Sources** • Few internal and external sources
Other Characteristics • Flexible • Data oriented • Long life • Large • Single complex structure	**Other Characteristics** • Restrictive • Project oriented • Short life • Start small, becomes large • Multiple semi-complex structures, together complex

-Permeates redundant data in every data mart .

Data Warehouse Development Approaches

	Inmon	Kimball
Overall approach	Top-down	Bottom-up
Architecture structure	Enterprise-wide (atomic) data warehouse "feeds" departmental databases	DMs model a single business process, and enterprise consistency is achieved through a data bus and conformed dimensions
Complexity of the method	Quite complex	Fairly simple
Comparison with established development methodologies	Derived from the spiral methodology	Four-step process; a departure from RDBMS methods
Data orientation	Subject or data driven	Process oriented

OLAP

OLAP models

Characteristics	MOLAP	ROLAP	HOLAP
Query performance	Fastest	Fast	Faster
Well-defined model, definitions and rules	Yes	Yes	Yes
Data volumes	Limited by physical cube	Limited by relational database	Limited by relational database
Needs to be rebuilt when data change	Yes	No	Only aggregations
Primary disadvantages vs relational	Cube "explosion," i.e., must build many cubes because of size limitations	Need to manage both relational DBMS and metadata layer	Need to manage both relational DBMS and metadata layer

MODELING

Operational Systems	BI and Analytics
Normalized models are standard for OLTP	Dimensional models are standard for BI and OLAP
Highly volatile	Generally not updated
Transaction throughput (updating and maintaining numerous records) is critical	Query performance (gathering and aggregating large sets of records) is critical
Characteristics supporting use of normalized models:	Characteristics supporting use of dimensional models:
Minimal redundancy (normalization)	Increased redundancy (denormalization)
Limited index use	Increased index use
Efficient use of storage space	Increased storage space
Eliminate inconsistent data	Consolidate inconsistent data
Few maintenance concerns	Increased maintenance issues

Schema Types

Star	Snowflake
Dimension represented by one-table	Dimension tables are expanded into multi-tables
Fact table surrounded by dimension tables	Fact table surrounded by Hierarchy of dimension tables
Less join	Requires many joins
Simple Design	Very Complex Design
De-normalized Data structure	Normalized Data Structure
High level of Data redundancy	Very low-level data redundancy
Maintenance is difficult	Maintenance is easier
Good for datamarts with simple relationships (1:1 or 1:many)	Good for core to simplify (many: many)

Processing

Criteria	OLTP	OLAP
Purpose	To carry out day-to-day business functions	To support decision making and provide answers to business and management queries
Data source	Transaction database (a normalized data repository primarily focused on efficiency and consistency)	Data warehouse or DM (a nonnormalized data repository primarily focused on accuracy and completeness)
Reporting	Routine, periodic, narrowly focused reports	Ad hoc, multidimensional, broadly focused reports and queries
Resource requirements	Ordinary relational databases	Multiprocessor, large-capacity, specialized databases
Execution speed	Fast (recording of business transactions and routine reports)	Slow (resource intensive, complex, large-scale queries)