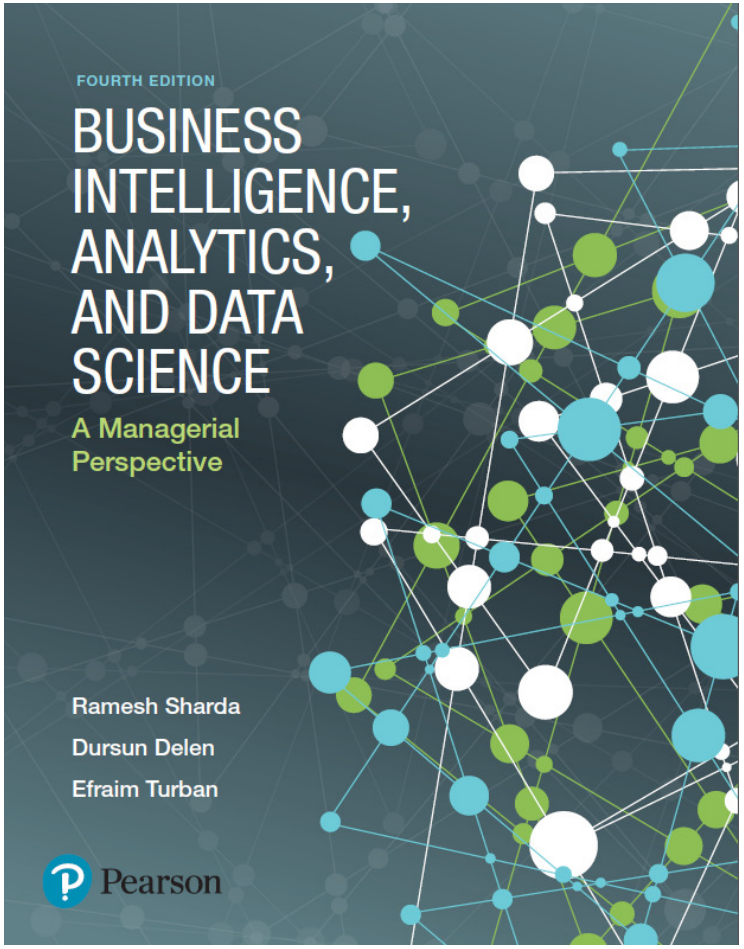
Business Intelligence, Analytics, and Data Science: A Managerial Perspective

Fourth Edition



Chapter 3 – Part B

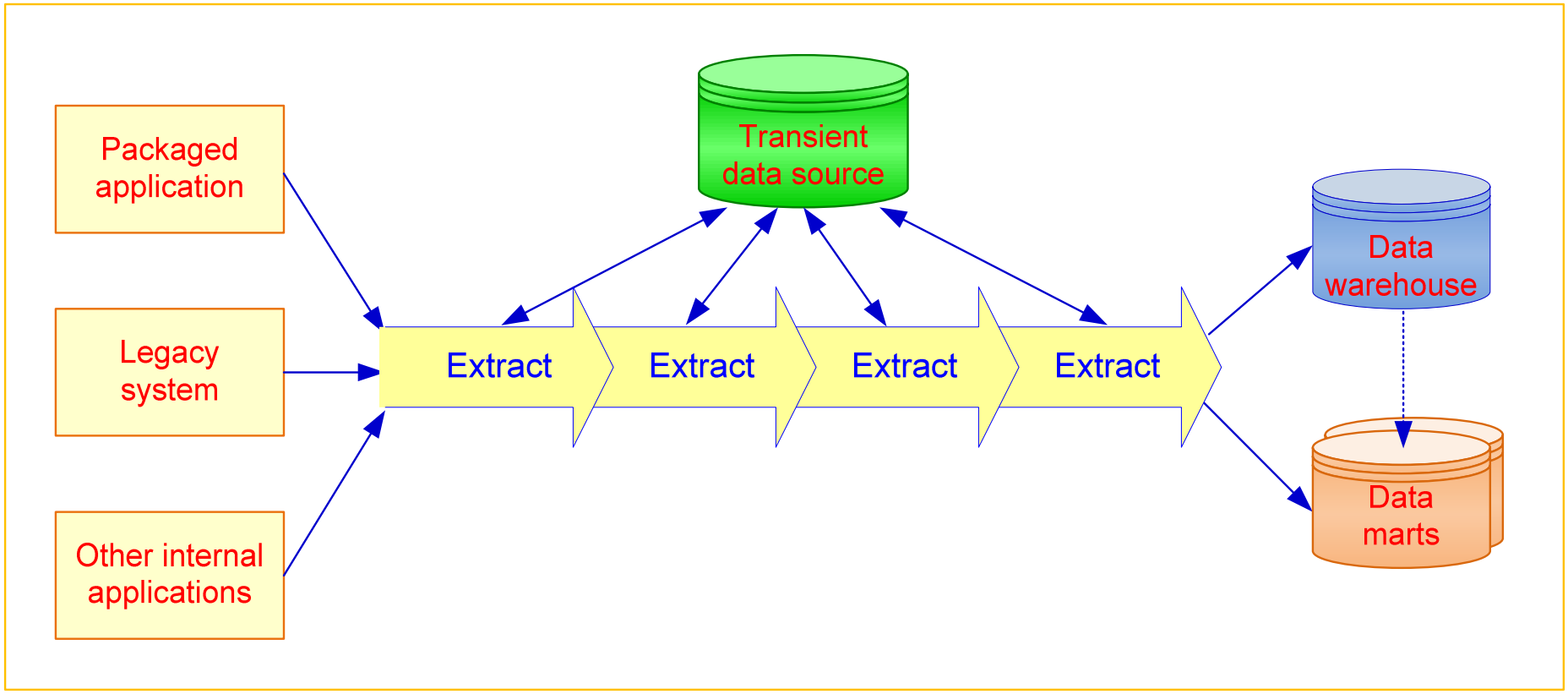
Descriptive Analytics II: Business Intelligence and Data Warehousing

Pearson

Data Integration and the Extraction, Transformation, and Load Process

- ETL = Extract Transform Load
- Data integration
  - Integration that comprises three major processes: data access, data federation, and change capture.
- Enterprise application integration (EAI)
  - A technology that provides a vehicle for pushing data from source systems into a data warehouse
- Enterprise information integration (EII)
  - An evolving tool space that promises real-time data integration from a variety of sources, such as relational or multidimensional databases, Web services, etc.

Data Integration and the Extraction, Transformation, and Load Process



ETL (Extract, Transform, Load)

- Issues affecting the purchase of an ETL tool
  - Data transformation tools are expensive
  - Data transformation tools may have a long learning curve
- Important criteria in selecting an ETL tool
  - Ability to read from and write to an unlimited number of data sources/architectures
  - Automatic capturing and delivery of metadata
  - A history of conforming to open standards
  - An easy-to-use interface for the developer and the functional user

Application Case 3.2
BP Lubricants Achieves BIGS Success

Questions for Discussion

1. What is BIGS?
2. What were the challenges, the proposed solution, and the obtained results with BIGS?

Data Warehouse Development

- Data warehouse development approaches
  - Inmon Model: EDW approach (top-down)
  - Kimball Model: Data mart approach (bottom-up)
  - Which model is best?
- Table 3.3 provides a comparative analysis between EDW and Data Mart approach
- Another alternative is the hosted data warehouses

Comparing EDW and Data Mart

TABLE 3.3 Contrasts between the DM and EDW Development Approaches

Effort	DM Approach	EDW Approach
Scope	One subject area	Several subject areas
Development time	Months	Years
Development cost	$10,000 to $100,000+	$1,000,000+
Development difficulty	Low to medium	High
Data prerequisite for sharing	Common (within business area)	Common (across enterprise)
Sources	Only some operational and external systems	Many operational and external systems
Size	Megabytes to several gigabytes	Gigabytes to petabytes
Time horizon	Near-current and historical data	Historical data
Data transformations	Low to medium	High
Update frequency	Hourly, daily, weekly	Weekly, monthly
Technology		
Hardware	Workstations and departmental servers	Enterprise servers and mainframe computers
Operating system	Windows and Linux	Unix, Z/OS, OS/390
Databases	Workgroup or standard database servers	Enterprise database servers
Usage		
Number of simultaneous users	10s	100s to 1,000s
User types	Business area analysts and managers	Enterprise analysts and senior executives
Business spotlight	Optimizing activities within the business area	Cross-functional optimization and decision making

Application Case 3.3

Use of Teradata Analytics for SAP Solutions Accelerates Big Data Delivery

Questions for Discussion

1. What were the challenges faced by the large Dutch retailer?
2. What was the proposed multivendor solution? What were the implementation challenges?
3. What were the lessons learned?

Additional DW Considerations
Hosted Data Warehouses

- Benefits:
  - Requires minimal investment in infrastructure
  - Frees up capacity on in-house systems
  - Frees up cash flow
  - Makes powerful solutions affordable
  - Enables solutions that provide for growth
  - Offers better quality equipment and software
  - Provides faster connections
  - … more in the book

Representation of Data in DW

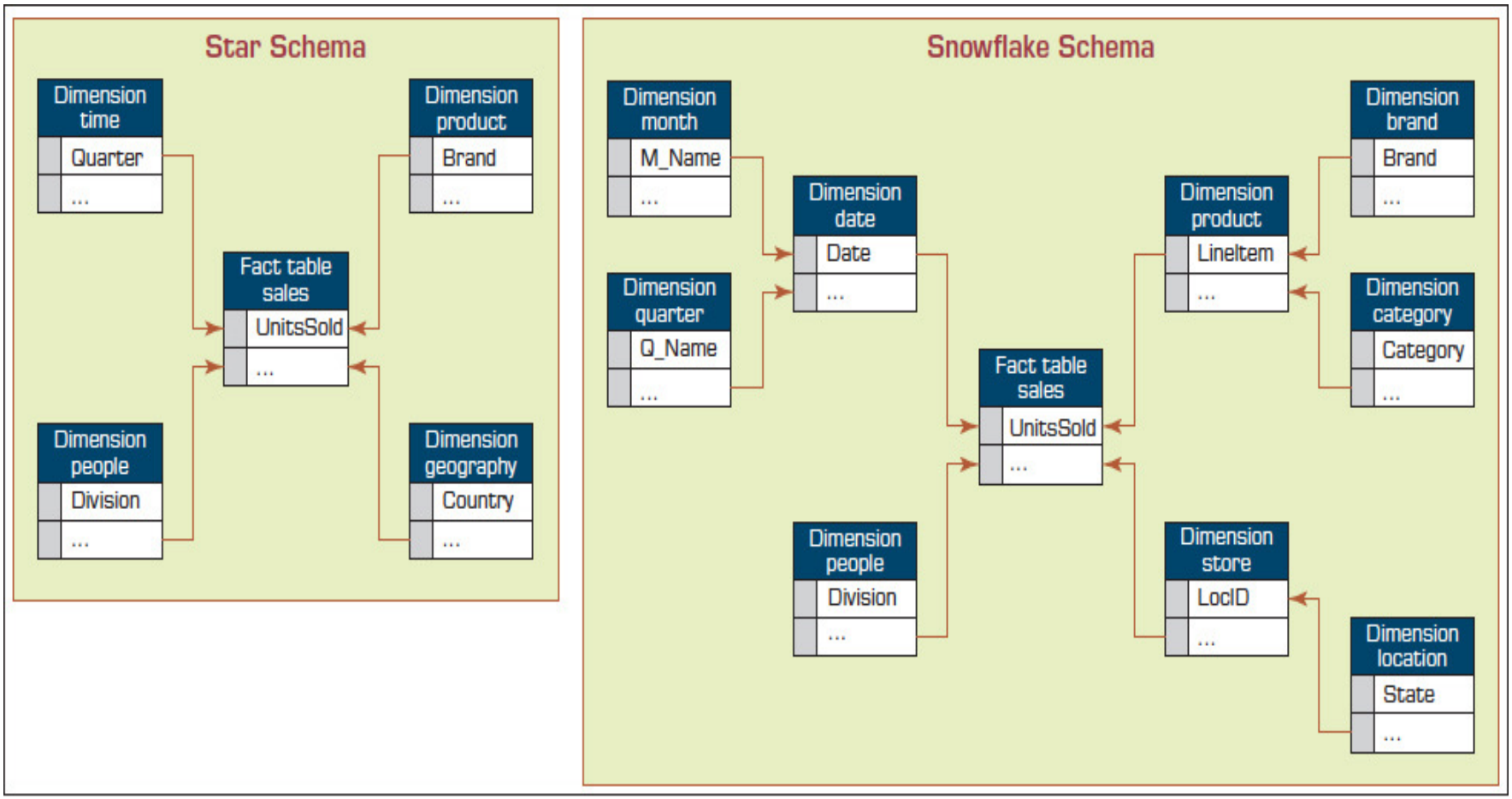
- Dimensional Modeling
  - A retrieval-based system that supports high-volume query access
- Star schema
  - The most commonly used and the simplest style of dimensional modeling
  - Contain a fact table surrounded by and connected to several dimension tables
- Snowflakes schema
  - An extension of star schema where the diagram resembles a snowflake in shape

Multidimensionality

The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions)

- Multidimensional presentation
  - Dimensions: products, salespeople, market segments, business units, geographical locations, distribution channels, country, or industry
  - Measures: money, sales volume, head count, inventory profit, actual versus forecast
  - Time: daily, weekly, monthly, quarterly, or yearly

Star Schema versus Snowflake Schema



Analysis of Data in DW

- OLTP vs. OLAP…
- OLTP (Online Transaction Processing)
  - Capturing and storing data from ERP, CRM, POS, …
  - The main focus is on efficiency of routine tasks
- OLAP (Online Analytical Processing)
  - Converting data into information for decision support
  - Data cubes, drill-down / rollup, slice & dice, …
  - Requesting ad hoc reports
  - Conducting statistical and other analyses
  - Developing multimedia-based applications
  - …more in the book

OLAP vs. OLTP

TABLE 3.5 A Comparison between OLTP and OLAP

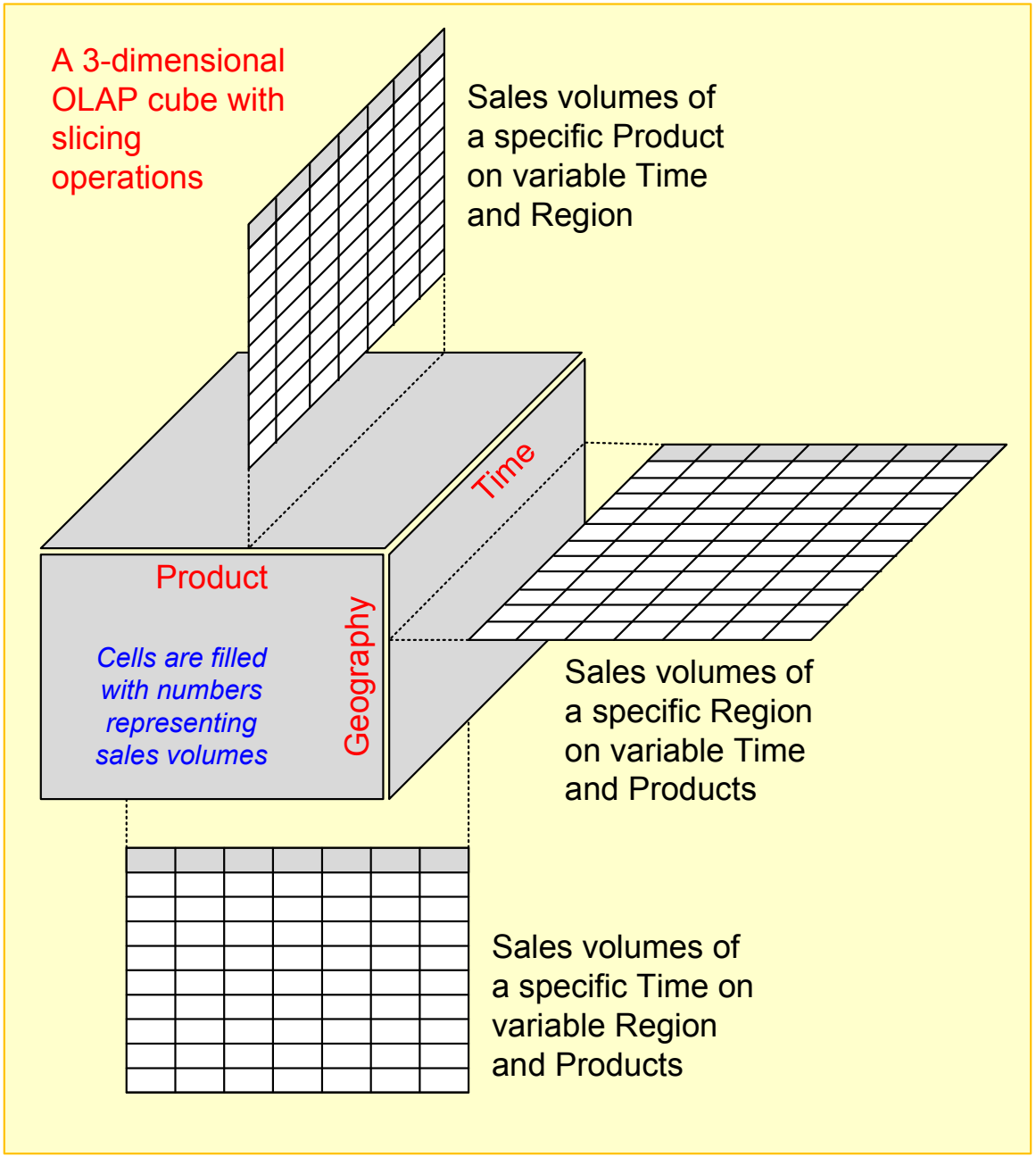
Criteria	OLTP	OLAP
Purpose	To carry out day-to-day business functions	To support decision making and provide answers to business and management queries
Data source	Transaction database (a normalized data repository primarily focused on efficiency and consistency)	Data warehouse or DM (a nonnormalized 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)

OLAP Operations

- Slice - a subset of a multidimensional array
- Dice - a slice on more than two dimensions
- Drill Down/Up - navigating among levels of data ranging from the most summarized (up) to the most detailed (down)
- Roll Up - computing all of the data relationships for one or more dimensions
- Pivot - used to change the dimensional orientation of a report or an ad hoc query-page display

OLAP

Slicing Operations on a Simple Tree-Dimensional Data Cube



Successful DW Implementation Things to Avoid

- Starting with the wrong sponsorship chain
- Setting expectations that you cannot meet
- Engaging in politically naive behavior
- Loading the data warehouse with information just because it is available
- Believing that data warehousing database design is the same as transactional database design
- … more in the book

Massive DW and Scalability

- Scalability
  - The main issues pertaining to scalability:
    - The amount of data in the warehouse
    - How quickly the warehouse is expected to grow
    - The number of concurrent users
    - The complexity of user queries
  - Good scalability means that queries and other data-access functions will grow linearly with the size of the warehouse

Application Case 3.4
EDW Helps Connect State Agencies in Michigan

Questions for Discussion

1. Why would a state invest in a large and expensive IT infrastructure (such as an EDW)?
2. What is the size and complexity of the EDW used by state agencies in Michigan?
3. What were the challenges, the proposed solution, and the obtained results of the EDW?

DW Administration and Security

- Data warehouse administrator (DWA)
  - DWA should…
    - have the knowledge of high-performance software, hardware, and networking technologies
    - possess solid business knowledge and insight
    - be familiar with the decision-making processes so as to suitably design/maintain the data warehouse structure
    - possess excellent communications skills
- Security and privacy is a pressing issue in DW
  - Safeguarding the most valuable assets
  - Government regulations (HIPAA, etc.)
  - Must be explicitly planned and executed

The Future of DW

- Sourcing…
  - *Web, social media, and Big Data*
  - Open source software
  - SaaS (software as a service)
  - Cloud computing
  - Data lakes

- Infrastructure…
  - Columnar
  - Real-time DW
  - Data warehouse appliances
  - Data management practices/technologies
  - In-database & In-memory processing New DBMS
  - New DBMS, Advanced analytics, …

Data Lakes

- Unstructured data storage technology for Big Data
- Data Lake versus Data Warehouse

TABLE 3.6 A Simple Comparison between a Data Warehouse and a Data Lake

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