Chapter 3:

Data warehouse and Analytical Processing

Data Mining

Instructor:

Debesh Jha, Ph.D.,

Visiting Assistant Professor,

University of South Dakota,

Vermillion, SD

September 19, 2024

Outline

- Role of Data Analytics
- Data warehouse
- Data warehousing
- Online Analytical Processing
- Data warehouse modeling: schema and measures
- OLAP operations
- Data cube computation
- Data cube computation methods

Data center

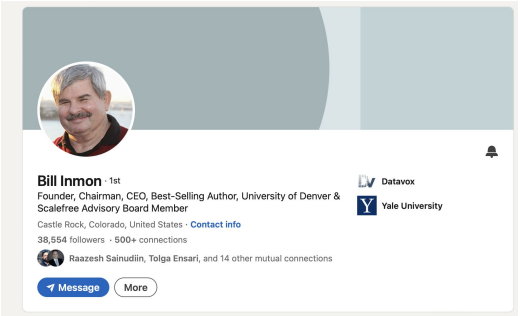
Data Analytics in Business Intelligence

- **Data analytics:** It is the science of analyzing raw data to make conclusions about that information.
- **Role in strategic decision-making:** It helps organizations harness their data and use it to identify new opportunities, leading to smarter business moves, more efficient operations, and higher profits.
- **Integration with data warehousing:** Data warehouses provide a clean, consolidated data source for analytics tools, enhancing the quality and speed of data analysis.

Example: An e-commerce platform analyzes **customer purchase history** and **browsing behavior** from their data warehouse to craft personalized marketing campaigns, enhancing customer engagement and boosting sales.

Data warehouse

- It is a specialized data management system that is built to support business intelligence (BI), particularly analytics.
- It provide companies the ability to analyze data and change over time, create insights and arrive at conclusion that help shape business decision.
- “ It is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—Bill Inmon (W.H Inmon)

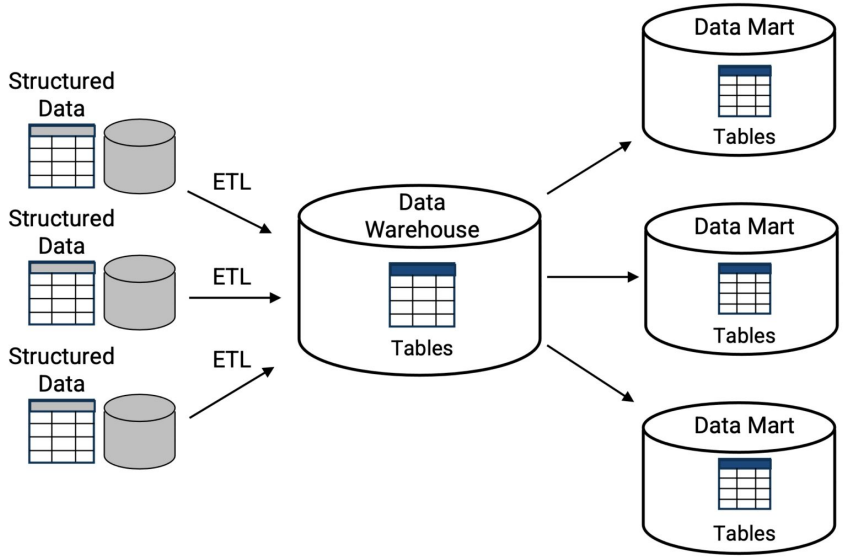


Overall it is a large collection of organized and clean business data, ready to help organization make decision.

Classic Data warehouse Architecture

Sources of data

- Transactional databases
- Customer sales
- Supply chain data
- External source data
- Public dataset
- Machine and IoT data
- Data from APIs
- Real-time or streaming data



Examples of classic data warehouse include

Snowflake, Google BigQuery, Oracle, Teradata, Amazon Redshift, Azure Synapse Analytics, IBM Db2 Warehouse, Firebolt

Data warehouse: Subject-Oriented

- Organized around major subjects, such as customer, product and sales
  - Examples: In retail business, a data warehouse might be focused on customer demographics, product categories, and sales performance.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
  - Analytical queries performed to understand market campaign looking at sales before and after.
- Exclusion of non-supportive dataset: minute transactional details or operational logs, are typically not included in the data warehouse.
  - Daily transaction logs excluded from the data warehouse where the main focus is quarterly sales data and trend analysis.
  - Provides a simple and concise view around particular subject

Data Warehouse - Integrated

- Integration of diverse data sources
  - Data warehouses combine data from multiple heterogeneous sources, such as relational databases, flat files, and online transaction records.
- Data cleaning and integration technique:
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources,
  - Ex. Hotel price: differences on currency, tax, breakfast covered, and parking
- Consistency across data sources:
  - Data brought into the warehouse is consistent in terms of format, scale, and context
  - Example: DD-MM-YYYY vs. MM-DD-YYYY
- Conversion and transformation:
  - As data is moved to the warehouse, it often needs to be converted or transformed to fit the warehouse schema.

Data warehouse: Time Variant

- Longer time horizon
  - Stores historical data significantly longer than operational systems.
  - Provide insights from past, necessary for trend analysis and forecasting.
- Time-stamped data
  - Every record includes a time element, explicit (dates) or implicit (versions).
  - Facilitates retrieval and analysis of data from specific time periods.
- Comparison: operational vs. analytical use:
  - Operational databases focus on current, real-time data.
  - Data warehouses maintain extensive historical records for deep analysis.

Example business application:

- Financial institutions analyze decades of loan data to **refine credit risk models** and **adjust offerings based on historical default patterns**.

Data Warehouse—Nonvolatile

- Non Volatility
  - Data in the warehouse does not change once it is loaded, except for periodic updates.
  - Ensures historical accuracy and consistency for reliable analysis.
- Separation from operational databases (e.g, MYSQL, Microsoft SQL, Oracle database)
  - Physically and logically separated from operational databases.
  - This separation ensures that the analytical processes do not impact the performance of the transactional systems and vice versa.
- Limited data operations
  - Support two types of operations: the initial loading of data and the accessing of data for analysis.
  - Unlike transactional systems, data warehouses **do not require capabilities** such as real-time data updating, concurrency control, or recovery mechanisms.

Data Warehousing

The process of constructing (collecting, storing and managing large volumes of data) and using data warehouse for decision making process.

Benefits:

- Improved decision making
- Enhanced data quality and consistency
- Faster query performance: optimized for read-heavy operations
- Historical data storage
- Scalability (for growing organization)
- Cross-department collaboration (sales, marketing, finance being integrated in data warehouse)

Data warehousing real world use case

- **Retail and e-commerce**
  - Amazon and walmart use data warehouse to track sales data, customer behavior, and inventory levels
- **Finance**
  - Banks use data warehouses to analyze historical transaction data, identify trends, and manage risks.
- **Healthcare**
  - Hospitals use data warehouses to integrate patient records, treatment histories for improved patient outcomes and operational efficiency.

Data warehousing real world use case and Limitation

- **Telecommunication**
  - Telecom companies use data warehouses to analyze *call records*, *customer service data*, and *network usage* to optimize customer experience and network performance.

Limitation of data warehouse:

- High initial costs
- Time-consuming ETL
- Not-suitable for real-time data
- Data overload.

Database and datawarehouse

Feature	Database	Data warehouse
Primary use	Real-time transactional processing (OLTP)	Analytical processing and reporting (OLAP)
Data type	Current, real-time data; structured and small volumes	Historical data; large volumes from multiple sources
Purpose	Day-to-day operations (e.g., orders, payments)	Historical data; large volumes from multiple sources
Data sources	Single source or operational systems	Multiple sources consolidated (CRM, ERP)
Query complexity	Simple, fast queries for single records	Complex queries across multiple dimensions
Best for	Operational systems like banking, e-commerce, POS	Business intelligence, forecasting, trend analysis

Scenarios for using database and datawarehouse

Retail Company:

- Record real-time sales, manage inventory, update customer profiles, and process online orders.
- Monthly reports on sales trends, product performance, and customer behavior over the past year for future trend analysis.

Hospital:

- Manage real-time patient records, appointments, billing, and prescriptions with instant updates.
- Analyze patient outcomes, treatment effectiveness, and resource usage over five years across departments (lab, radiology, surgery) for policy decisions.

When to use database?

- To handle real-time operations like customer orders, payments, and inventory management.
- The focus is on short-term data and transactional consistency.
- Perform CRUD operations (**Create, Read, Update, Delete**) frequently.
- The data is constantly changing or being updated.
- You require real-time access to current data for operational tasks.
- Your queries involve retrieving a small subset of records (e.g., one customer's order history).
- Ecommerce transactions, banking systems, CRM systems etc are application.

When to use Data Warehouse?

- To store large volumes of historical data for analysis over time.
- The focus is on business intelligence (BI), reporting, and data analysis.
- Run complex queries across multiple data sources and dimensions (e.g., analyzing sales trends over the last five years).
- Require data consolidation from various operational databases (e.g., CRM, ERP, sales) to generate insights.
- The data is mostly read-only and not frequently updated.
- You need to support decision-making by analyzing trends, performing aggregations, and generating reports (e.g., quarterly financial reports)
- Example application include business reporting, historical data analysis, data consolidation etc.

Customer Relation Management and Enterprise Resource Planning

Feature	CRM	ERP
Purpose	Manages customer interactions and relationships, improving sales.	Manages internal business processes, improving operational efficiency etc.
Focus area	Sales, marketing, customer support.	Finance, HR, supply chain, manufacturing.
Primary function	Lead tracking, customer data management, sales pipelines.	Financial management, inventory, HR, procurement.
Users	Sales, marketing, customer service teams.	Finance, HR, operations, procurement teams.
Integration	Integrates with marketing tools, social media, and email platforms.	Integrates with internal systems like accounting, HR, and production.
Examples	Salesforce, HubSpot CRM, Zoho CRM, Shopify, Magento, Square.	SAP, Oracle ERP, Microsoft Dynamics 365.

Traditional Data warehouse and Cloud data warehouse

Feature	Traditional Data Warehouse	Cloud Data Warehouse
Deployment	On-premises, physical hardware	Cloud-based, accessible via the internet
Cost	High upfront hardware and maintenance	Pay-as-you-go, lower upfront costs
Scalability	Limited, requires new hardware	Highly scalable, adjusts on demand
Setup Time	Time-consuming, physical setup	Quick setup, ready in minutes
Maintenance	Requires dedicated IT staff	Managed by cloud provider
Performance	Limited by hardware	High performance, scales with needs
Security	On-site control, perceived as secure	Cloud provider-managed, remote access risk
Storage	Limited by physical capacity	Near-infinite capacity
Elasticity	Fixed capacity	Automatic resource adjustment

Leading Cloud-Based CRM and ERP Solutions Providers

- CRM Solution provider
  - Salesforce
  - HubSpot
  - Zendesk

- ERP Solutions Providers
  - SAP SE
  - Oracle
  - Workday

Cloud Based Data Warehousing companies

1. Snowflake
2. Amazon web service (Amazon Redshift)
3. Google Cloud Platform - Google BigQuery
4. Amazon
5. Microsoft - Azure Synapse Analytics
6. Oracle Autonomous Data Warehouse
7. IBM - IBM Db2
8. Cloudera Data Warehouse
9. Teradata

Leading Tech Leaders Across Cloud and Data Solutions

1. Microsoft Corp. (Cloud, Data Warehousing (Azure), Enterprise Software
2. Amazon (Cloud Infrastructure (AWS), E-commerce)
3. Salesforce Inc. (CRM, Cloud-based Software)
4. Nvidia Corp (Data Warehousing (GPUs for AI and Analytics), Semiconductors)
5. Snowflake Inc. (Cloud Data Warehousing)
6. Oracle Corp (Cloud ERP, Database Software)
7. Palantir Technologies (Data Analytics/Big Data)
8. Datadog Inc. (Cloud Monitoring/Data Analytics)
9. IBM

Scenario: Data Warehouse Integration in Healthcare

- Organization: A large hospital
  - Hospital Goals: Enhance Patient Care, Boost Efficiency, and Advance Treatment Research

- Data sources
  - Electronic Health Records (EHRs): Contains patient medical histories, lab results.
  - Patient Administration Systems (PAS): scheduling, admissions, discharges, and transfers.
  - Laboratory Information Systems (LIS): Stores laboratory test results.
  - Radiology Information Systems (RIS): Contains imaging reports and schedules.
  - Pharmacy Information Systems (PIS): Tracks medication orders and dispenses history.
  - Finance Systems: Manages billing, insurance claims, and procurement details.

Data Integration Process and Data warehouse framework

- **Data Extraction**: Collecting data from various hospital systems.
- **Data Transformation**: Cleansing and standardizing data for consistency.
- **Data Loading**: Nightly updates to the data warehouse.

- Structure and Storage:
  a. **Data Models:** Utilizes a star schema for efficient querying and data retrieval, centered around patients, treatments, physicians, and costs.
  b. **Storage:** Hosted on a secure, scalable cloud platform to manage large data volumes and ensure reliability.

Analytics and importance

- **Analytics and Reporting:**
  - BI Tools: Integrates tools like Microsoft Power BI for OLAP operations, supporting complex, multidimensional analyses.
  - Dashboards: Creates customized dashboards and reports highlighting efficiency, care metrics, financials etc.
- **Benefits and Impact**
  - Improved Patient Outcomes
  - Increased Operational Efficiency
  - Enhanced Research Capabilities

Online analytical processing (OLAP)

Software technology designed for multi-dimensional analysis of business data, providing the ability to perform complex calculations and trend analysis.

- Combines and groups this data into categories for actionable insights.
- Convert raw data into actionable insights for strategic planning and operational improvement.
- OLAP systems integrate data from a variety of sources such as websites, applications, smart meters, and internal systems.
- These diverse datasource consolidates to provide a comprehensive view of business operation and customer interaction.

Purpose of OLAP

- What are the **sales trends** over the **last 5 years**?

- Which products are performing best in **different regions**?

- OLAP tools enable users to view data from different perspectives, such **as by time** (monthly, yearly) or **by product category** (e.g., electronics, furniture). It is possible to "slice and dice" the data to see detailed insights and perform aggregations (e.g., totals, averages).

- Example: A company wants to know which product sold the most in the last quarter across different countries. OLAP tools help them generate this report by **querying and summarizing large datasets**.

Organization of data in OLAP

- **Senario:** A retailer stores product data including color, size, cost, and location, and separately records customer purchases, detailing item names and total sales values.
- **Findings:** OLAP **combines the datasets** and derive insights to answer for best selling colors and optimal shelf placement for increased sales.
- **Impact:** It lead to targeted inventory stocking and store layout adjustments driving higher sales and customer satisfaction.
  - Analyzing **detailed product** and **customer purchase data** allowed the retailer to **optimize its inventory** and **store layout strategically**.
  - This targeted approach resulted in **increased sales** and **improved customer satisfaction**.

Example of OLAP systems are Snowflake, Microsoft SQL server analysis service, Google Bigquery, Amazon redshift

The main goal of OLAP is **data analysis** and not data processing.

The main goal of OLTP is **data processing** not data analysis.

Benefits of OLAP

- Faster decision making
  - Utilized by businesses to expedite and refine decision-making processes.
- Non-technical user support
  - Simplifies complex data analysis for users without technical expertise.
  - Business professionals can create complex analytical calculations and generate report without the need to master database operations.
- Integrated data view
  - Provides a unified platform that combines marketing, finance, production etc.
  - Managers and decision-makers can understand the broad effects of their strategies.
- OLAP services enforce security measures to protect data.
- They provide a multidimensional view of data, enabling diverse operations.
- OLAP services ensure data consistency and support efficient calculations.

Drawback of OLAP/OLAP Service

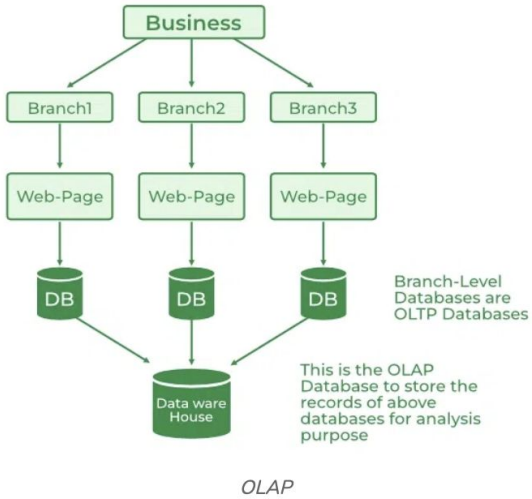
- Complexity and professional expertise
  - Requires skilled professionals due to the complex modeling and data handling procedures.
- High cost of implementation
  - Expensive to implement and maintain, especially with large datasets and infrastructure requirements.
- Rigid data structure
  - Data cubes need to be predefined, making it less flexible for adapting to new data types or analyses.
- Delayed data analysis
  - Analysis can only be performed after data extraction and transformation, causing delays.
- Periodic updates
  - OLAP systems are **updated periodically, making them less efficient** for real-time decision-making.
- Limited integration with unstructured data
  - Primarily designed for structured data, **difficult to analyze unstructured formats** (e.g, text, images)

OLAP examples

- **Spotify (OLAP Example)**
  - Spotify analyzed songs by users to come up with a personalized homepage of their songs and playlists.
- **Netflix movie recommendation system**
  - Content performance analysis: Netflix uses OLAP to analyze viewer habits, such as which shows are most watched, viewer retention rates, and the effectiveness of recommendations.
  - Trend analysis: By aggregating data from millions of viewers, Netflix can predict which content shows will be popular in certain regions.
- **Healthcare data analytics**
  - Hospitals and healthcare institutes use OLAP (Oracle OLAP) for patient outcome analysis, resource allocation and cost management.
- **Sales and Marketing**
  - Companies use OLAP (Microsoft SQL Server Analysis Services (SSAS)) for sales trend analysis and customer segmentation.

OLAP Structure: Data Warehousing from Multiple OLTP Sources

- The image shows a **business** with several branches (Branch1, Branch2, Branch3).
- Each branch has a **web-page interface** connected to a **database (DB) to handle daily transactions (operational task)**.
- **Real-time data updates** with a focus on fast, efficient query processing for operational needs.
- A centralized **Data Warehouse** consolidates data from the **multiple OLTP** systems.
- Here, **OLAP system** allows business to analyze combined data, perform complex analysis, reporting, and strategic decision-making using multidimensional data.



OLAP

Purpose:

- OLTP databases handle **transactional data** (e.g., customer orders),
- OLAP database allows **complex queries, reporting, and analysis** using that collected data.

Image source: https://www.geeksforgeeks.org/difference-between-olap-and-oltp-in-dbms/

Outline

- OLAP and OLTP
- Multi-tier architecture of data warehouse
- Data lake
- Data warehouse modeling: schema and measures
  - Data cube: a multidimensional data model
  - Schemas for multidimensional data models: stars, snowflakes, and fact constellations
  - Concept hierarchies
  - Measures: categorization and computation
- OLAP operations
- Data cube computation
- Data cube computation methods

Online Transaction processing (OLTP)

It is a system that manages **real-time transactional data**.

- It supports high transaction volume with fast query processing.
- Commonly used for day-to-day operations like banking, retail, and reservations.

Key Features of OLTP

- Real-time processing
- High concurrency (supports many users)
- Short, simple queries: Handles small, quick queries (e.g., CRUD operations)
- Data Integrity: Ensures accuracy and consistency with ACID properties (Atomicity, Consistency, Isolation, Durability).

OLTP Advantages

- **Fast data operations**:
  - OLTP enable quick read, write, and delete operations, ensuring efficient data handling.
- **Real-time data access**:
  - By supporting a high number of users and transactions, OLTP systems allow real-time access to current and accurate data.
- **Enhanced security**:
  - Multiple security layers help protect sensitive data during transactions.
- **Improved decision-making**:
  - Real-time, accurate data allows for better and faster decision-making.
- **Data integrity and consistency**:
  - Ensures the accuracy and reliability of transactions by adhering to **ACID properties** (Atomicity, Consistency, Isolation, Durability).

OLTP Drawback

- **Limited analytical capability**:
  - OLTP systems are optimized for transaction processing, not for complex analysis or reporting, which requires separate systems like OLAP.
- **High Maintenance Costs**:
  - Due to the need for frequent backups, updates, and recovery, maintaining OLTP systems can be costly.
- **Vulnerability to Hardware Failures**:
  - In the event of hardware failure, online transactions can be disrupted, leading to system downtime or data loss.
- **Data Duplication and Inconsistencies**:
  - OLTP systems can sometimes experience issues with duplicate or inconsistent data, especially during high transaction volumes or system errors.

OLTP examples: ATM withdrawal

ATM Transactions:

User authenticates by entering PIN and requesting cash.

- **Authentication First**: OLTP system verifies account balance and ensures funds are available.
- **Real-time transaction:** Deduct balance and update balance instantly
- **Instant Feedback: Machine dispense money,** and we see updated account balance. Thanks to OLTP System

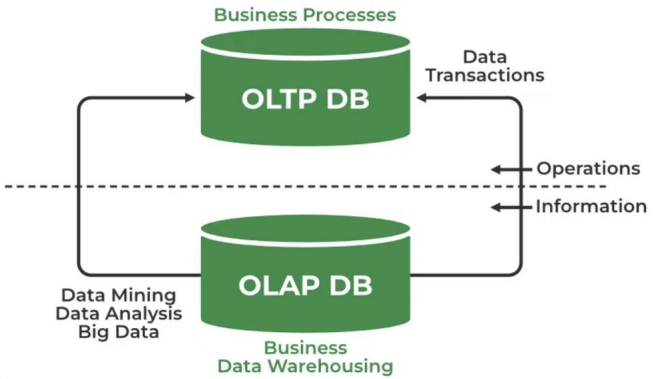
Key Features:

- **ACID Properties**: Ensures data integrity and consistency.
- **Instant Feedback**: Transactions are processed and reflected immediately.

OLTP & OLAP relationship within Organizational framework

OLTP Database:

- **OLTP DB** handles **business processes** and **real-time transactions**.
- It is optimized for day-to-day operations, such as customer transactions or order processing.
- The flow of **data transactions** shows that the **OLTP system is responsible** for **updating operational data.**



OLAP Database:

- Centralized data warehouse for analysis, reporting, and strategic planning.
- The arrows connecting the **OLTP and OLAP databases show** how **operations and information** flow between them.
- Data is periodically transferred from the OLTP system to the OLAP system for analysis and reporting purposes.

Purpose: OLTP is focused on **transactional operations**, while OLAP is focused on **analyzing large volumes of data** stored in the data warehouse.

Other OLTP examples

- **Online Banking**: Transferring money, paying bills, or checking your balance—each transaction happens instantly.

- **Airline Ticket Booking**: Booking a flight seat is processed immediately, ensuring no double bookings.

- **Text Messaging**: Sending a message involves real-time data handling, where the OLTP system manages each text's delivery status.

- **E-commerce**: When you add a book to your shopping cart or make a purchase, OLTP systems handle the transaction and update stock levels in real time.
- **Oracle database, MySQL, Microsoft SQL server, IBM Db2 are OLTP systems**

Similarities between OLAP and OLTP

- Both are database management systems for storing and processing data in large volumes.
- Both require efficient and reliable IT infrastructure to ensure smooth operation.
- Both support data-driven decision-making in an organization.
- Most companies use OLTP and OLAP systems together to meet their business intelligence requirements.
- However, the approach to and purpose of data management differ significantly between OLAP and OLTP.

OLTP and OLAP differences

- OLTP: Online transactional processing
  - DBMS operations
  - Query and transactional processing

- OLAP: Online analytical processing
  - Data warehouse operations
  - Drilling, slicing, dicing, etc.

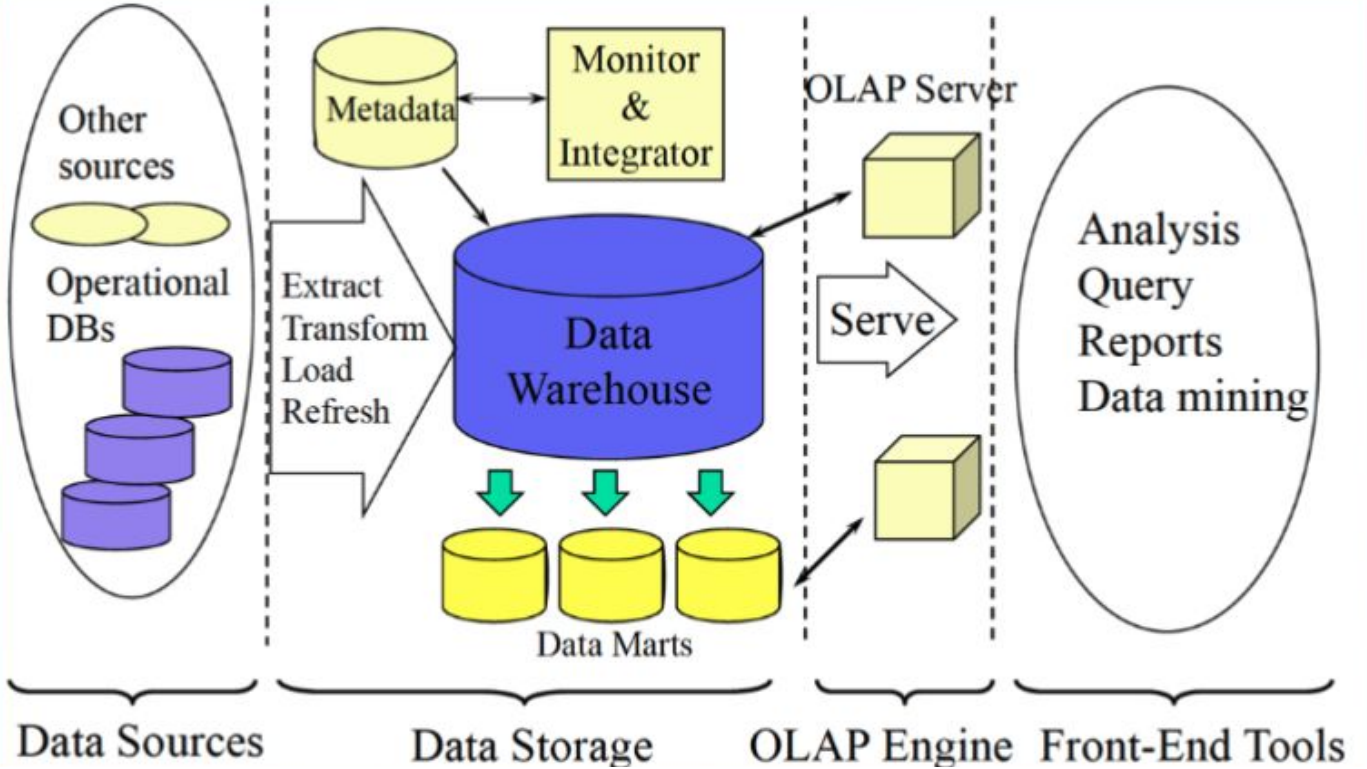
	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

OLTP and OLAP differences

Criteria	OLAP (Analytical)	OLTP (Transactional)
Purpose	Analyze large datasets for decision-making.	Manage and process real-time transactions.
Data Source	Aggregated, historical data from multiple sources.	Real-time data from a single source.
Data Structure	Multidimensional (cubes), relational databases.	Relational databases
Data Model	Star, snowflake schemas (analytical models).	Normalized/denormalized (transaction models)
Data Volume	Large (TB to PB)	Smaller (GB)
Response Time	Slower (seconds to minutes).	Fast (milliseconds).
Example Use	Trend analysis, predicting customer behaviour, profitability.	Payment processing, customer data management, order processing.

Multi-tier architecture of Data Warehouse

- Top Tier: Front-End Tools
- Middle Tier: OLAP Server
- Bottom Tier: Data Warehouse Server
- Data



- ERP Databases
- CRM systems
- POS Systems
- Human Resource Management Systems

Three Data Warehouse Models

- Enterprise warehouse
  - Collects all of the information about subjects spanning the entire organization

- Data Mart
  - A subset of corporate-wide data that is of value to a specific groups of users.
  - Its scope is confined to specific, selected groups, for e.g marketing data mart.
    - Independent vs. dependent (directly from warehouse) data mart
      - E.g A company's sales department maintains its own data mart for customer and sales data.
- Virtual warehouse (a logical layer)
  - It refers to the compute clusters that drive modern data warehouses, providing on-demand processing power for real-time data access.
  - Instead of creating physical data marts, they create virtual data marts for various departments, where each department can access only the relevant data through specific queries or views.

Extraction, Transformation, and Loading (ETL)

- **Data extraction**
  - Get data from multiple, heterogeneous, and external sources.
- **Data cleaning**
  - Detect errors in the data and rectify them when possible.
- **Data transformation**
  - Convert data from legacy or host format to warehouse format.
- **Load**
  - Sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions.
- **Refresh**
  - propagate the updates from the data sources to the warehouse

Metadata Repository

- **Metadata (data about the data)** is the data defining warehouse objects.
- Description of the structure of the data warehouse
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- Extraction and transformation metadata and end user metadata (for e.g, data labels and description)
- The algorithms used for summarization (for e.g averaging sales data by quarter or year)
- The mapping from operational environment to the data warehouse
- Data related to system performance
  - How quickly queries are processed.
  - How efficiently data is loaded into the warehouse.
- Business data
  - business terms and definitions, ownership of data, charging policies (cost-based access model)

OLAP Engine and Front-End Tools

Monitor & Integrator

- Monitors the **performance and integration** of the data warehouse.
- Ensures **data consistency**, **accuracy**, and **integrity** across the entire system.

OLAP Engine

- **OLAP Server** provides data to end-users in a **multidimensional format** for complex analysis.
- Enables **fast querying**, **reporting**, and **data analysis** by **pre-aggregating** data.
- Communicates with the data warehouse to **fetch data and deliver** it to the **front-end tools**.

Front-End Tools

- Tools for **Analysis**, **Query**, **Reports**, and **Data Mining**.
- Used by **analysts**, **decision-makers**, and **data scientists** for interacting with data.
- Examples: **Tableau**, **Power BI**, and custom data mining algorithms.

Data lake

- It is a centralized repository that allows organizations to store, manage, and analyze large volumes of structured, semi-structured, and unstructured data in its raw format.

Key Characteristics

- Raw data storage
- Scalability (ex: Amazon S3, Google Cloud Storage, or A
  Data Lake Storage)
- Schema on road (data structure applied only when necessary)
- Cost-effective
- Multiple data source (variety of data)
- Data analysis (using python, R , SQL etc)

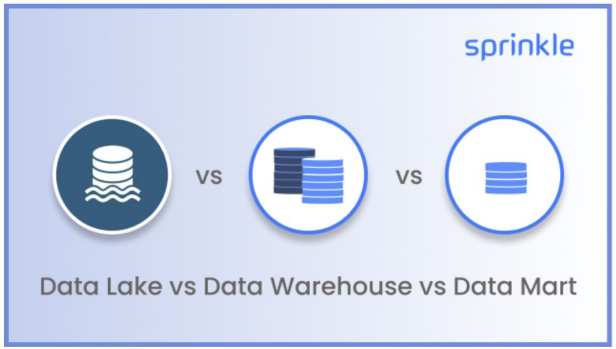
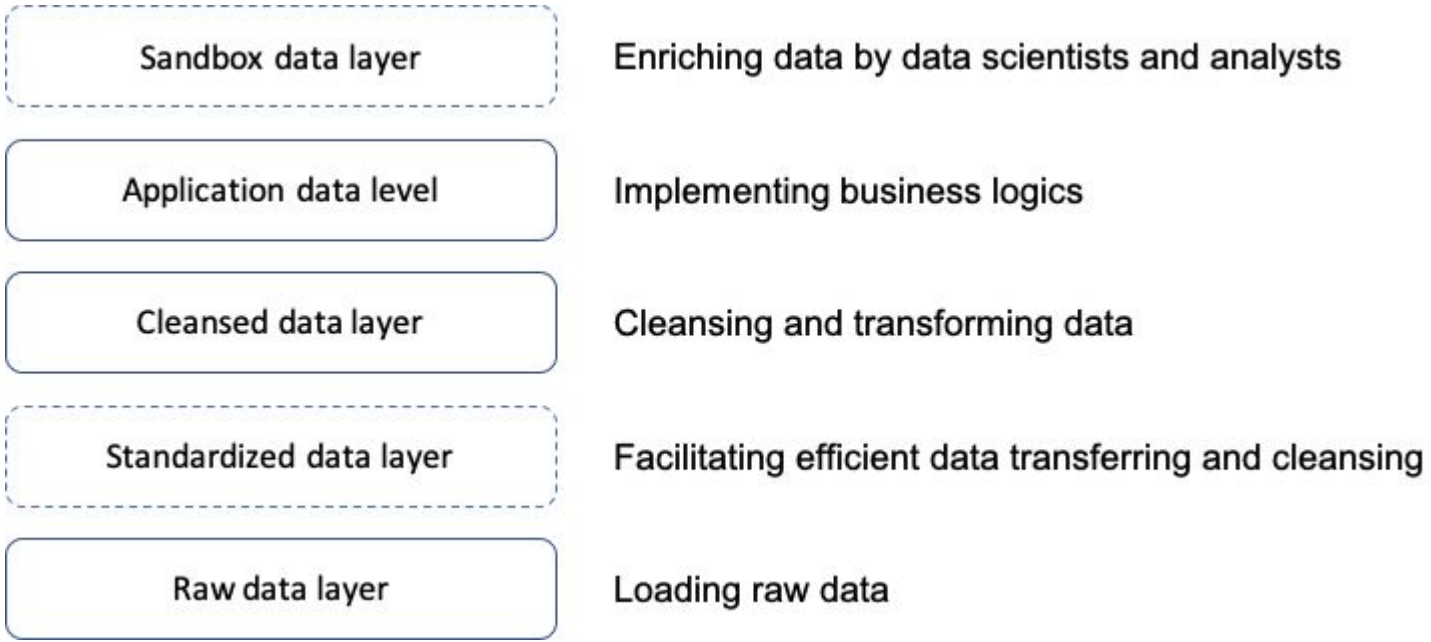


Image source: https://www.sprinkledata.com/blogs/data-lake-vs-data-warehouse-vs-data-mart

Layers of Storage



Data warehouse and data lake

Aspect	Dare Warehouse	Data lake
Usage	Business intelligence and analytics	Large-scale data storage
Data type	Clean and organized data	Structured, semi-structured, and unstructured data
Schema	Schema-on-write: structure decided before loading	Schema-on-read: structure specified during query
Data quality	High data quality with on duplication or cleansing issue	Inconsistent data quality due to lack of processing
Cost	More expensive due to complexity	Low-cost with low processing and maintenance
Latency	Low latency is pre-processed	High latency
Use cases	Core reporting, BI	ML, predictive analytics and real-time analysis

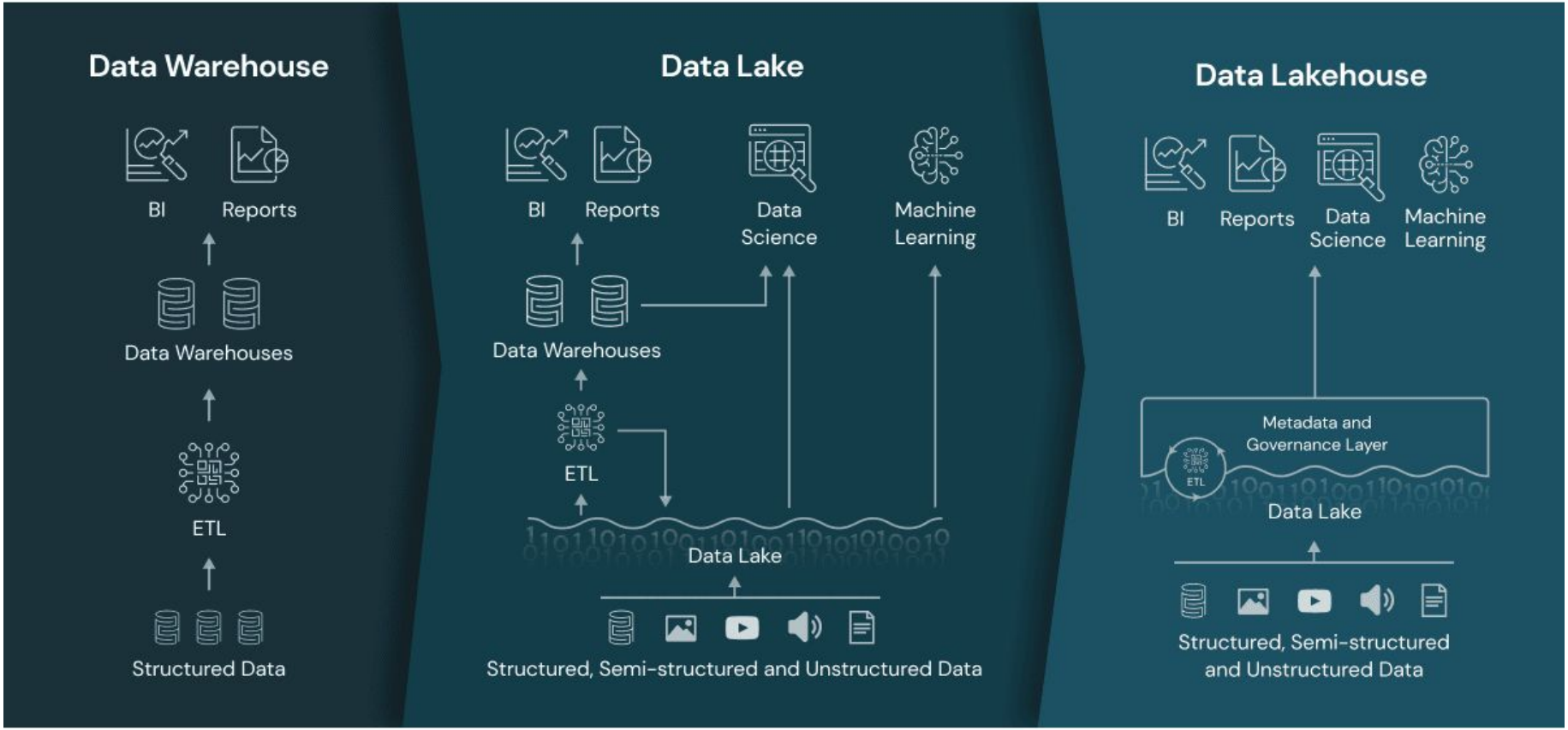
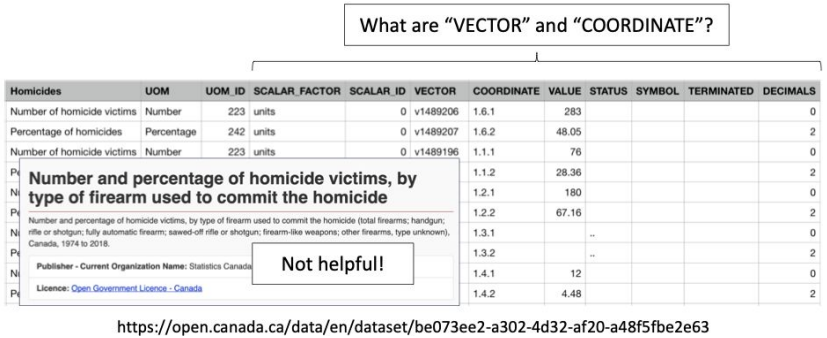


Image Source: https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

Data Lake Challenges

- Many data sets
  - hundreds of thousands and more with different structures.
  - Lot of coding, command line tools.
- Complicated queries
  - keyword queries, finding joinable data sets, finding relationship between different features,
- Core task in building data lake
  - metadata management
- Multiple architectures used by teams.



https://open.canada.ca/data/en/dataset/be073ee2-a302-4d32-af20-a48f5fbe2e63

Data Cleaning in Data Lakes

- Data cleaning is the No. 1 most cited task in data lake
- Over 85% considered it either major or critical to the business
- Example: Incorrect fare amounts in sales data (as shown) can distort revenue reports and lead to poor business decisions.

tpep_dropoff_dat...	passenger_count	trip_...	Rate...	store...	PULo...	DOLo...	pay...	fare_amount
2017 Jan 09 11:25:45 AM	1	3.3	1	N	263	161	1	12.5
2017 Jan 09 11:36:01 AM	1	0.9	1	N	186	234	1	5
2017 Jan 09 11:42:05 AM	1	1.1	1	N	164	161	1	5.5
2017 Jan 09 11:57:36 AM	1	1.			36	75	1	6
2017 Jun 20 10:39:16 PM	1	0.			41	141	2	630,461.82
2017 Jan 19 09:29:44 AM	3				39	264	2	625,900.8
2017 Jan 01 02:57:09 AM	1	0	1	N	232	243	3	538,579.2
2017 Apr 01 07:45:43 P...	1	0	1	N	90	264	2	538,481.03
2017 Oct 10 04:33:07 P...	1	1,178.6	1	N	170	170	2	404,093.97

Are these values correct?

https://data.cityofnewyork.us/Transportation/**2017-Yellow-Taxi-Trip-Data**/biws-g3hs

Data Lakehouse

- A data lake house combines the benefits of data lakes and data warehouses, supporting diverse data types and workloads while ensuring data quality.
- Key features include ACID transactions, schema enforcement, BI tool compatibility, decoupled storage and compute, open formats, and streaming capabilities.
- Lake Houses simplify data infrastructure, support real-time analytics, and handle modern applications like machine learning and AI.

Evolving Data in Data Lakes

- Many duplicates as data sets are often being copied for new projects.
- Data sets are constantly being updated, having their schema altered, being derived into new ones, and disappearing/reappearing
- Data set versioning is to maintain all versions of datasets for storage cost-saving, collaboration, auditing, and experimental reproducibility

Diversity in Data Lakes

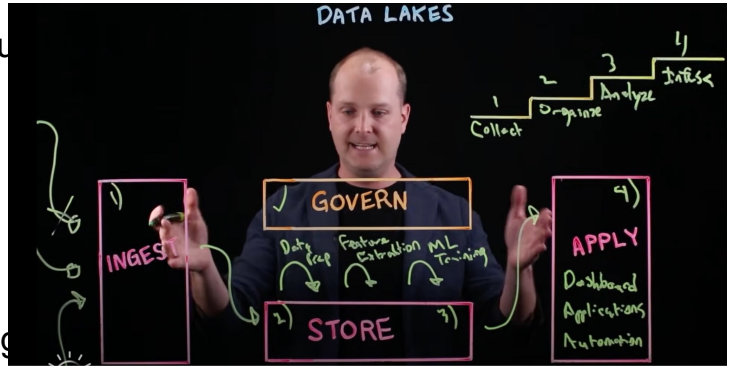
- Dataset formats in the open world can be highly heterogeneous.
- Ingestion and extraction is the task of bringing structured datasets into data lake
  - Ingest already-structured data sets
  - Extract structured data from unstructured and semi-structured data sources (e.g, table from pdf)
- Data Integration is the task of finding joinable or union-able tables or of on-demand population a schema with all data from the lake that conforms to the schema
  - Example: enrich Electronic Health Records (EPR) using data from va
  - rious non-standard personal health record data sets for **better predicting health risks**.
- Joining or “union-ing” tables from different data sets and sources

Example

- Step 1: Ingestion of Structured Data
  - Ingest structured EHR data (patient demographics, medical history, lab results) into a central data lake.
- Step 2: Extraction of Non-Standard Data
  - Extract structured data from wearables (heart rate, steps) and health apps (text entries) using NLP and API methods.
- Step 3: Data Integration
  - Join EHR data with wearable and app data based on patient ID.
  - Union similar fields (e.g., heart rate) for a richer dataset.
- Outcome:
  - Detect early signs of **heart disease** from unusual spikes in heart rate patterns.
  - Predict a patient's **diabetes risk** by correlating sedentary behavior with rising blood sugar levels from the EHR.

Common Tasks in Building Data Lakes

- Ingestion
  - Loading CSVs, JSON files, and images into the system
- Extraction (Type Inference)
  - Converting text files into structured tables
- Metadata Management
  - Cataloging datasets with descriptions, timestamps, and sou
- Cleaning
  - Removing duplicate rows
- Integration
  - Joining customer data from CRM systems with purchase history
- Discovery
  - Querying for healthcare datasets to analyze patient trends
- Versioning
  - Versioning a dataset after each batch update to track chang over time



Source: https://www.youtube.com/watch?v=LxcH6z8TFpI&t=12s

Data Lakehouses

- A data lakehouse is a new, open data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data.” (Databricks)

Lake Houses uses

- Transaction support
- Schema enforcement and governance
- BI support
- Decoupling between storage and compute
- Open storage formats
- Support for diverse data types, ranging from unstructured to structured
- Support for diverse workloads: data science, machine learning, SQL and analytics
- End-to-end streaming: real-time reports

Data Mesh

- A data mesh creates multiple domain-specific systems, each specialized according to its functions and uses, thus bringing data closer to consumers.

- Data Mesh is a specific architecture pattern **focused on data management**.

Data Virtualization

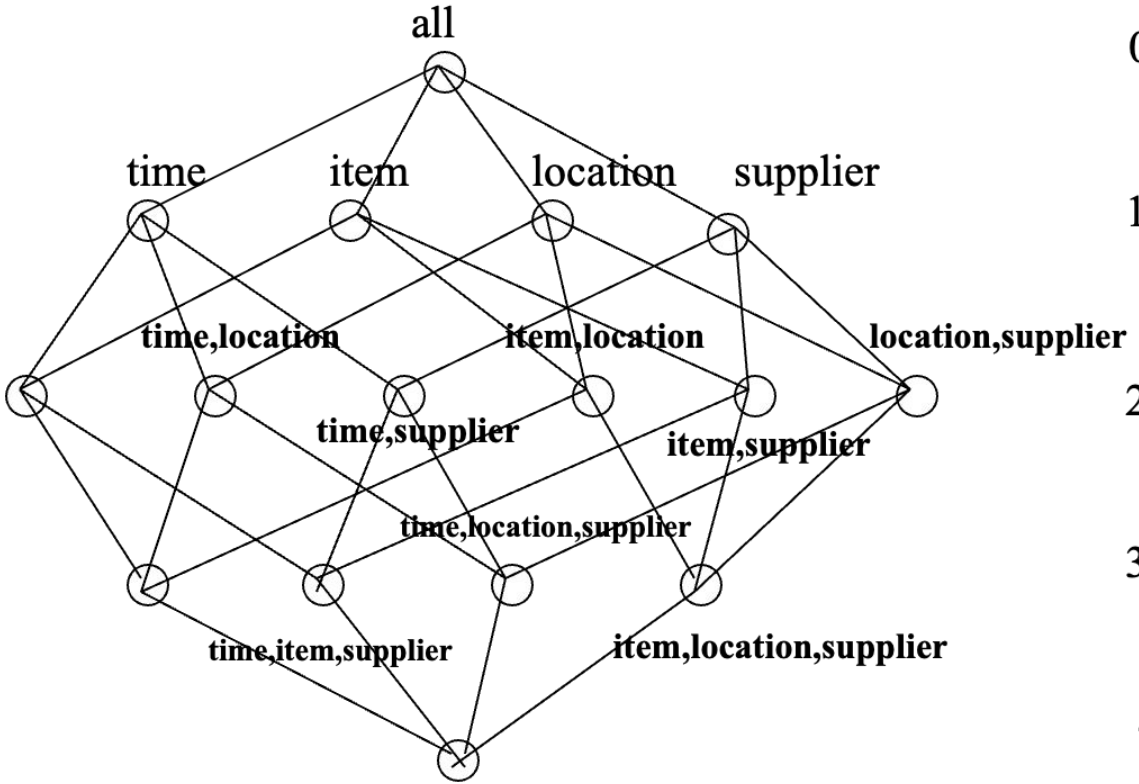
- Data virtualization is one of the technologies that enables a data fabric approach. Rather than physically moving the data from various on-premises and cloud sources using the standard ETL (extract, transform, load) processes, a data virtualization tool connects to the different sources, integrating only the metadata required and creating a virtual data layer.

- Data Virtualization can be a component or a layer in a data fabric architecture.

From Tables and Spreadsheets to Data Cubes

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube.
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - **Fact table** contains measures (such as dollars_sold) and keys to each of the related dimension tables
- Data cube: A lattice of cuboids
  - In data warehousing literature, an n-D base cube is called a base cuboid.
  - The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid
  - The lattice of cuboids forms a data cube.

Data Cube: A Lattice of Cuboids



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

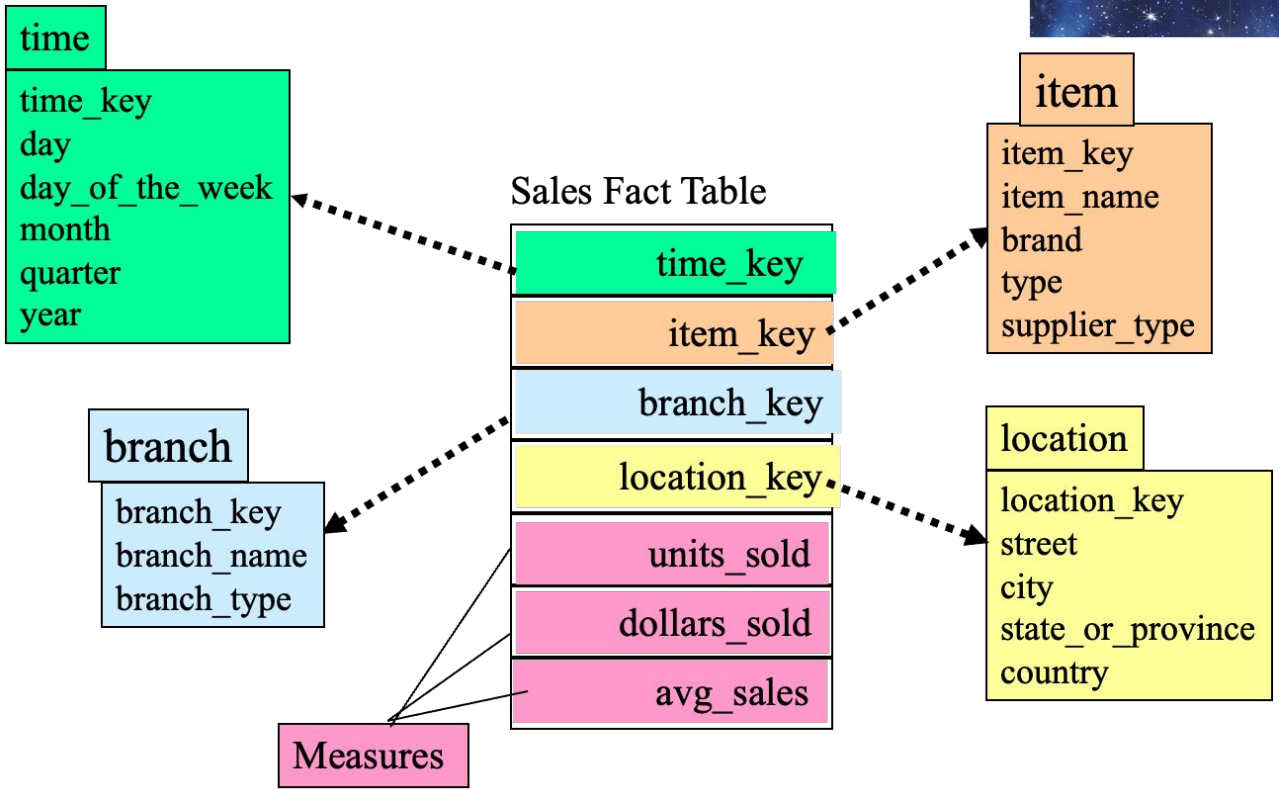
3-D cuboids

4-D (*base*) cuboid

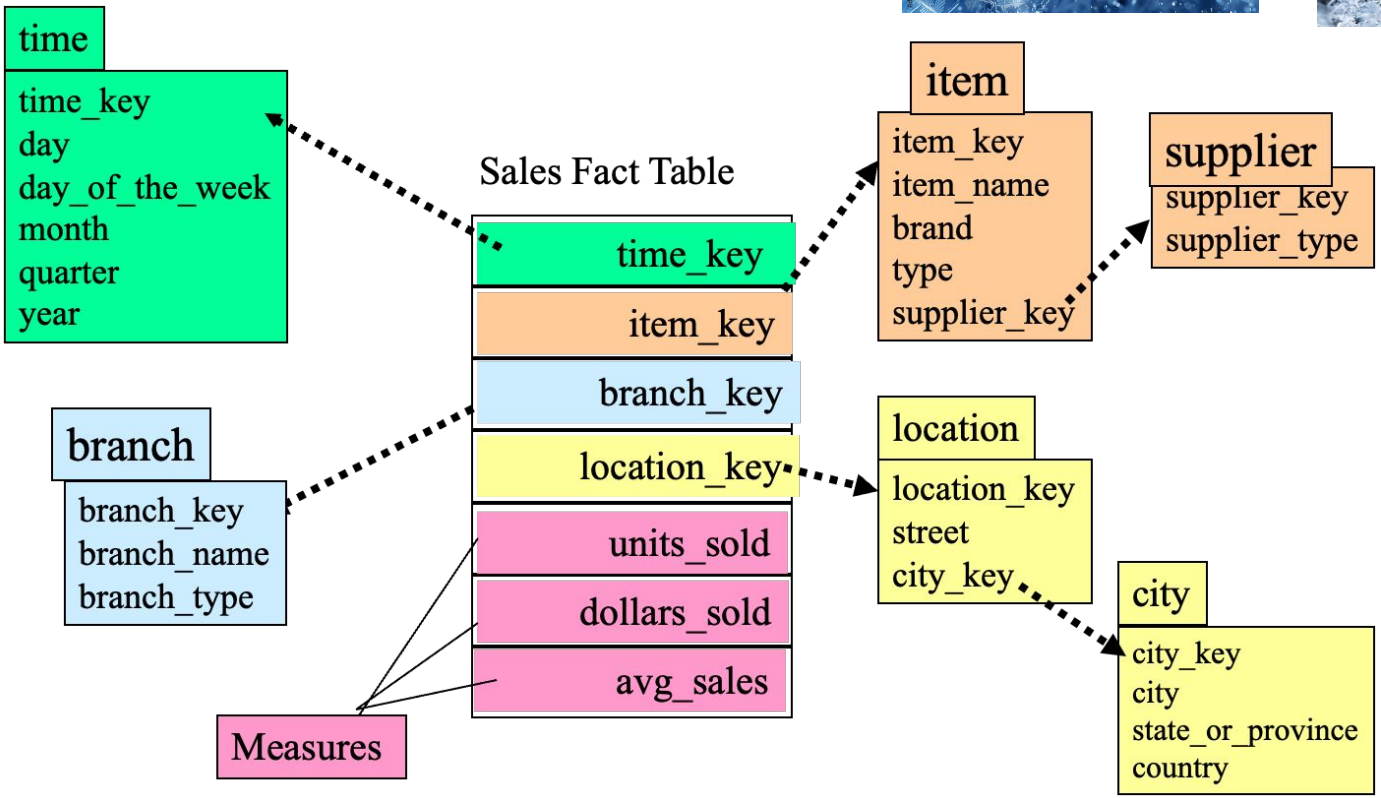
Conceptual Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
- Star schema: A fact table in the middle connected to a set of dimension tables.
- Snowflake schema: A refinement of star schema where some dimensional hierarchy is **normalized** into a set of smaller dimension tables, forming a shape similar to snowflake.
- Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called **galaxy schema** or **fact constellation**.

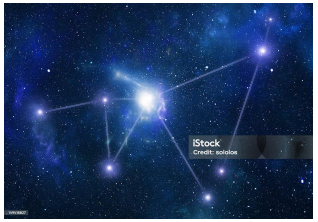
Star Schema: An Example

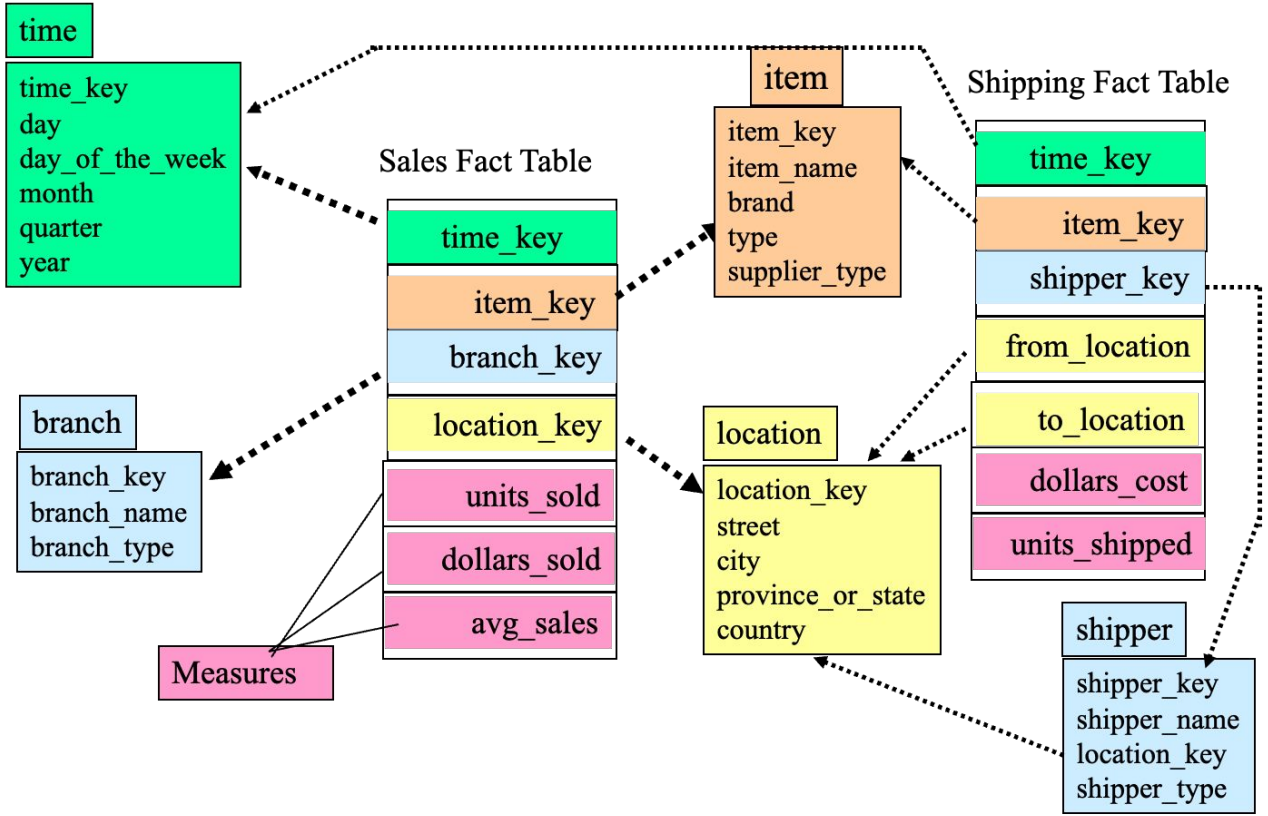


Snowflake Schema: An Example

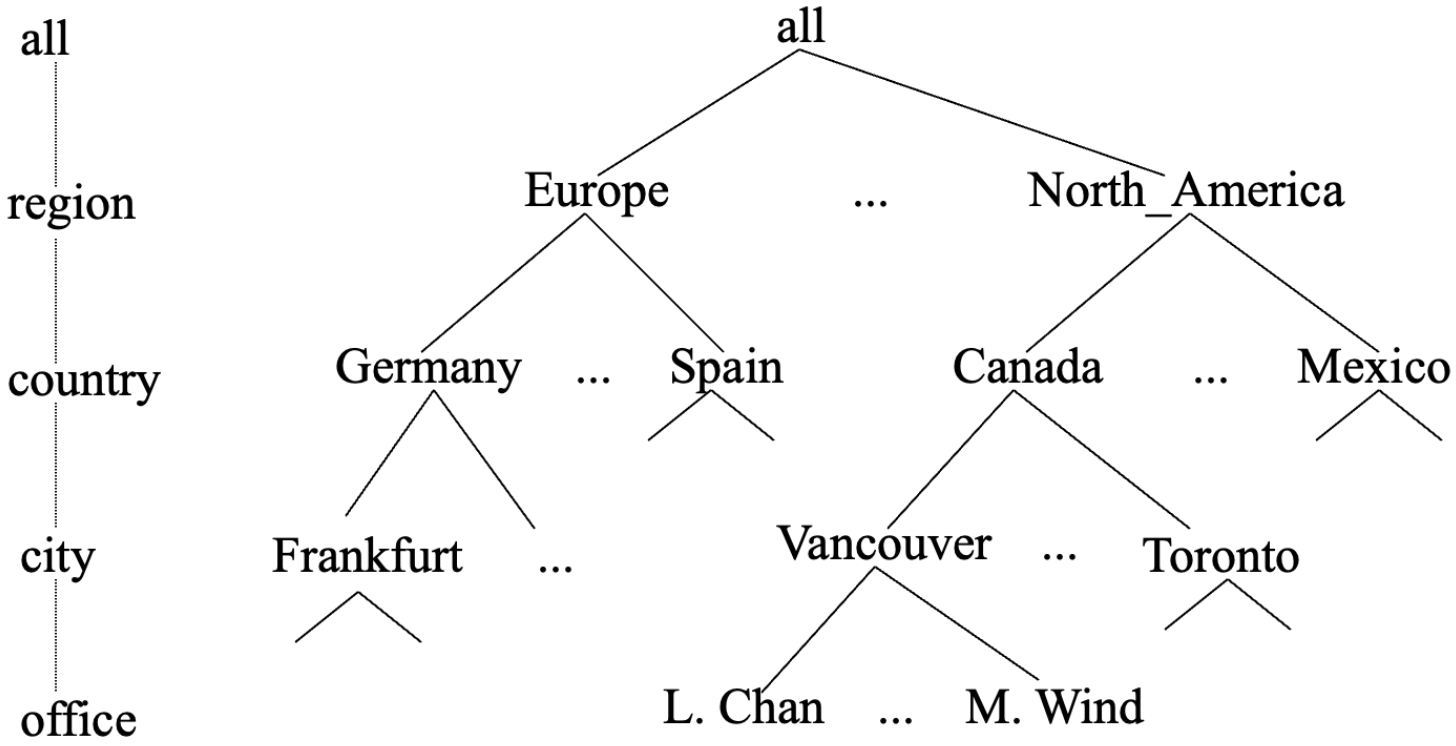


Fact Constellation: An example





A Concept Hierarchy for a Dimension (location)

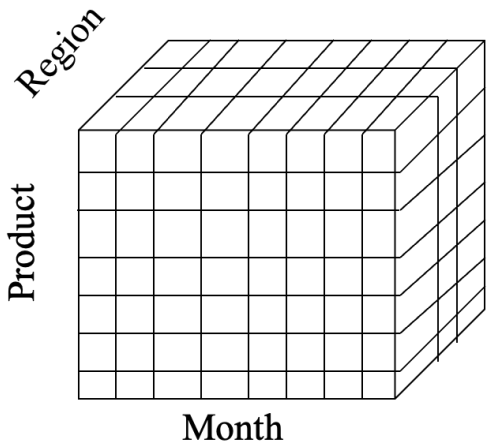


Data Cube Measures: Three Categories

- Distributive: if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning
  - E.g., count(), sum(), min(), max()
- Algebraic: if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a distributive aggregate function
  - avg(x) = sum(x) / count(x)
- Is min_N() an algebraic measure? How about standard_deviation()?
- Holistic: if there is no constant bound on the storage size needed to describe a subaggregate.
  - E.g., median(), mode(), rank()

Multidimensional Data

•Sales volume as a function of product, month, and region



Dimensions: *Product, Location, Time*
Hierarchical summarization paths

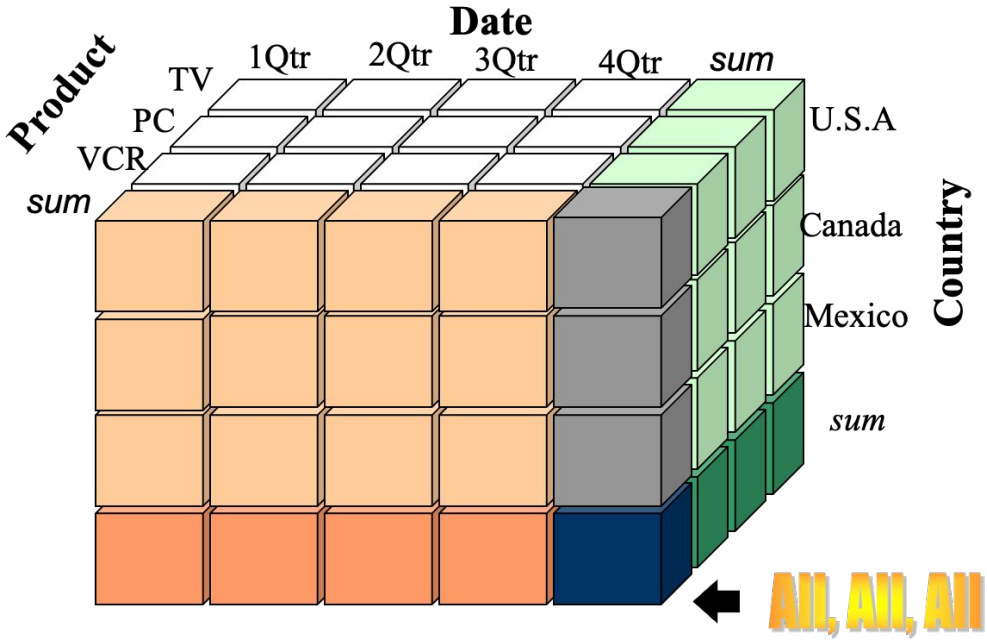
Industry Region Year

Category Country Quarter

Product City Month Week

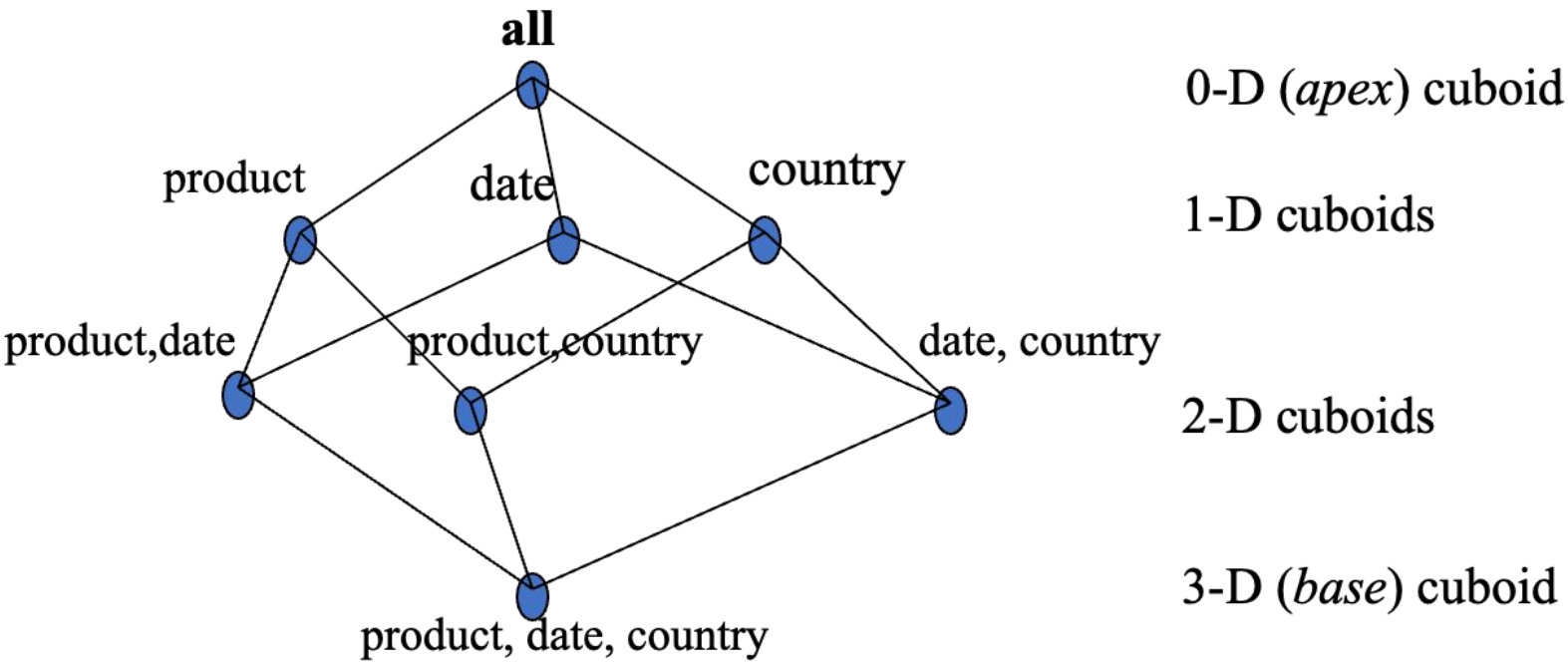
Office Day

A Sample Data Cube



Total annual sales of TVs in U.S.A.

Cuboids Corresponding to the Cube



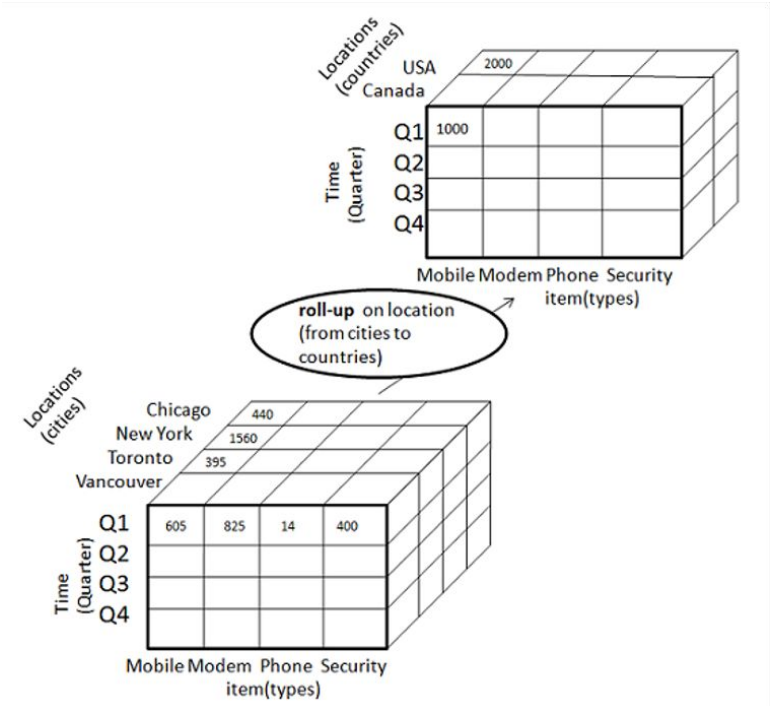
Online Analytic Processing (OLAP)

- Conceptually, we may explore all possible subspaces for interesting patterns
- Some fundamental problems in analytics and data mining
  - What patterns are interesting?
  - How can we explore all possible subspaces systematically and efficiently?
- Aggregates and group-bys are frequently used in data analysis and summarization
  - SELECT time, altitude, AVG(temp)
  - FROM weather GOUP BY time, altitude;
  - In TPC, 6 standard benchmarks have 83 queries, aggregates are used 59 times, group-bys are used 20 times
- Online analytical processing (OLAP): the techniques that answer multi-dimensional analytical (MDA) queries efficiently

OLAP Operations

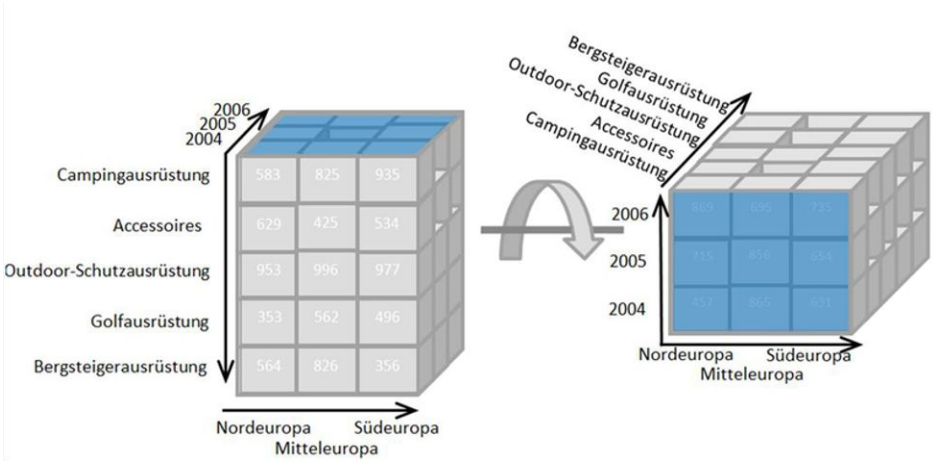
- Roll up (drill-up): summarize data by climbing up hierarchy or by dimension reduction
  - (Day, Store, Product type, SUM(sales) à (Month, City, *, SUM(sales))
  - Highly detailed to less detailed
- Drill down (roll down): reverse of roll-up, from higher level summary to lower level summary or detailed data, or introducing new dimensions
  - Drill down operation leads to hilgy detailed data.

Country level data from city

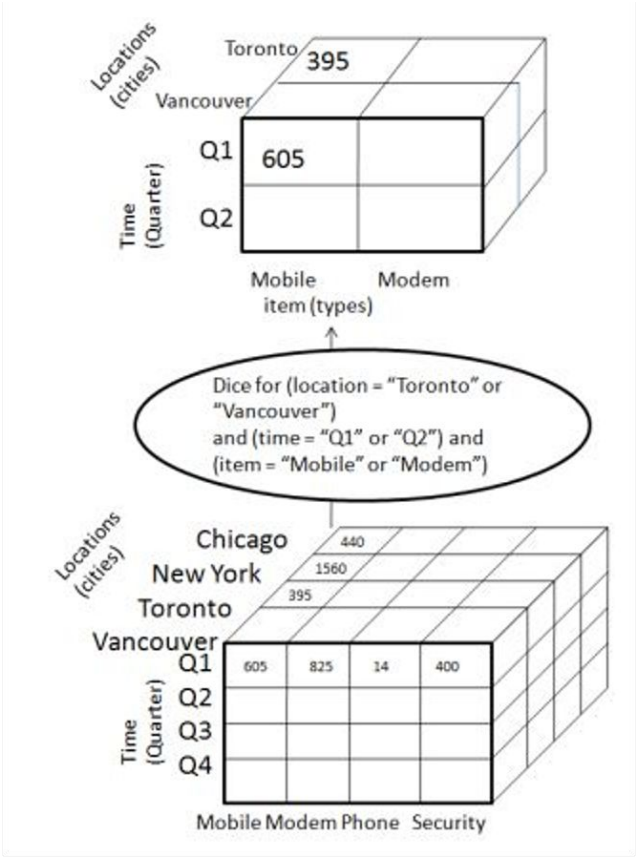


Other Operations

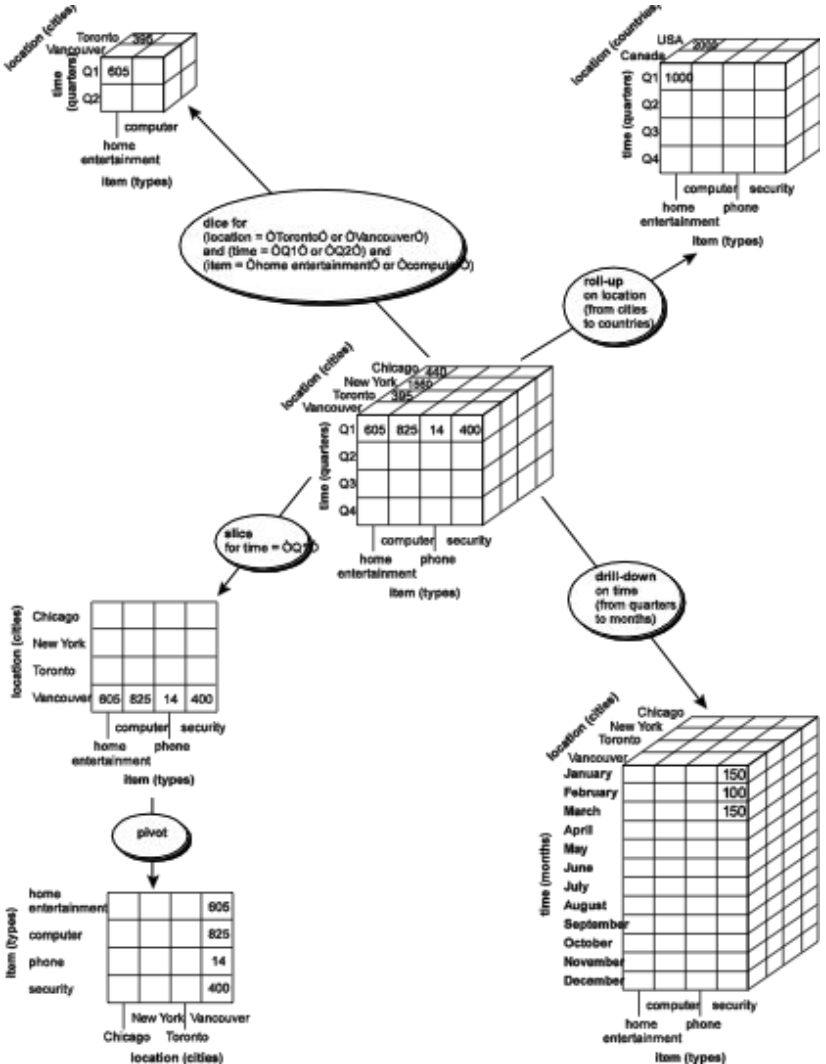
- Dice: pick specific values (piece of information) or ranges on some dimensions.
- Pivot: “rotate” a cube – changing the order of dimensions in visual analysis
- Slice



Rotate to get new view



Typical OLAP Operations



OLAP Server Architectures

- Relational OLAP (ROLAP)
  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware
  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
  - Greater scalability
- Multidimensional OLAP (MOLAP)
  - Sparse array-based multidimensional storage engine
  - Fast indexing to pre-computed summarized data
- Hybrid OLAP (HOLAP) (e.g., Microsoft SQLServer)
  - Flexibility, e.g., low level: relational, high-level: array
- Specialized SQL servers (e.g., Redbricks)
  - Specialized support for SQL queries over star/snowflake schemas

References

https://hanj.cs.illinois.edu/bk4/bk4_slidesindex.htm

https://aws.amazon.com/compare/the-difference-between-a-data-warehouse-data-lake-and-data-mart/

https://aws.amazon.com/what-is/olap/#:~:text=help%20with%20OLAP%3F-,What%20is%20online%20analytical%20processing%3F,smart%20meters%2C%20and%20internal%20systems.

https://www.snaplogic.com/blog/data-warehouses-data-lakes-data-lakehouses-everything-you-need-to-know

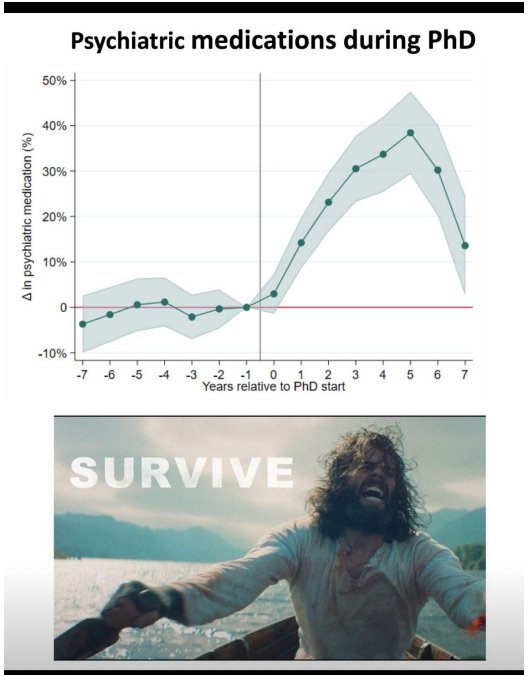
https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-data-lake#:~:text=risks%20more%20efficiently.-,What's%20the%20difference%20between%20a%20data%20lake%20and%20a%20data,as%20specific%20BI%20use%20cases.

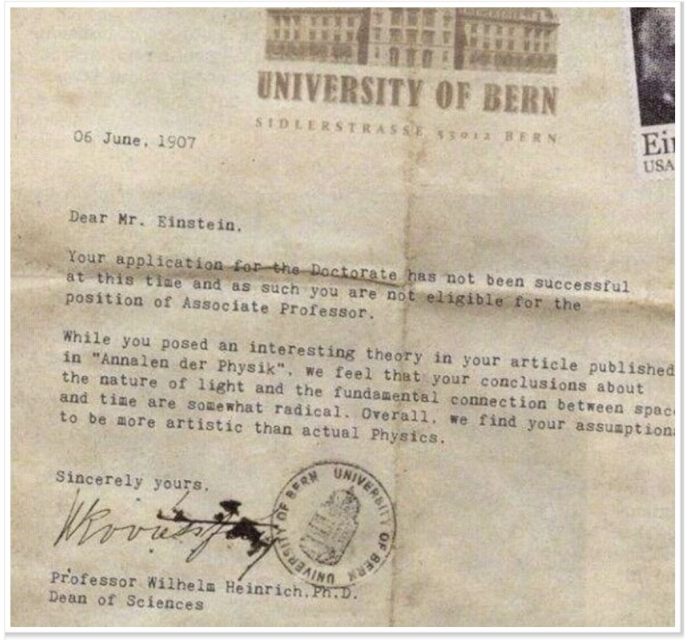
End of the Presentation

Thank you!

Note: Some content are enhanced by Large Language Models (LLMs) for educational purposes only . This is to make content more interactive for the students. All original materials remain the intellectual property of the presenter.

Bonus Slide




UNIVERSITY OF BERN

SIDLERSTRASSE 5 3012 BERN

06 June, 1907

Dear Mr. Einstein,

Your application for the Doctorate has not been successful at this time and as such you are not eligible for the position of Associate Professor.

While you posed an interesting theory in your article published in "Annalen der Physik", we feel that your conclusions about the nature of light and the fundamental connection between space and time are somewhat radical. Overall, we find your assumptions to be more artistic than actual Physics.

Sincerely yours,

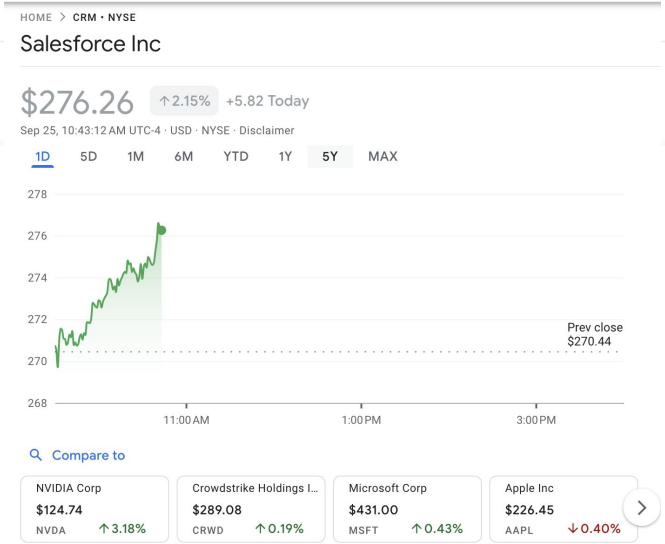
Professor Wilhelm Heinrich, Ph.D.
Dean of Sciences

His Theory of Relativity was rejected as being much less Physics, more Art !

Market trends

Most active | Gainers | Losers | Trending | More

NVDA	NVIDIA Corp	Dow Jones Futures Waver As Nvidia Hits Buy Trigger; KB Home Dives On ... Investor's Business Daily • 1 hour ago	$124.16	↑2.72%
SNOW	Snowflake Inc	Snowflake stock drops on $2 billion debt plan MarketWatch • 1 day ago	$114.13	↑1.45%
NIO	Nio Inc - ADR	Is NIO Inc. (NIO) Among the Best EV Stocks to Buy for the Long Term? Yahoo Finance • 4 hours ago	$5.61	↓5.64%
RITM	Rithm Capital Corp	Stocks making the biggest moves after hours: KB Home, Rithm Capital an... CNBC • 16 hours ago	$11.32	↓4.19%
F	Ford Motor Co		$10.38	↓4.51%
NVD	GraniteShares 2x Short NVD...		$1.65	↓5.17%



The Best 7 Cloud ERP Software of 2024

- **Microsoft Dynamics 365 Business Central:** Best overall
- **SAP Business One Professional:** Best for customization
- **SYSPRO:** Best for manufacturing businesses
- **QT9:** Best for real-time reporting
- **Epicor Prophet 21 ERP:** Best for distributors
- **Oracle NetSuite OneWorld:** Best for global companies
- **Acumatica:** Best for easy pricing

BEST OVERALL

Microsoft Dynamics 365 Business Central

5.0 Forbes ADVISOR

Starting price	Free trial	Built-in CRM
$70 per user (monthly)	30 days	No

Learn More

Read Forbes' Review

Editor's Take

Pros & Cons

BEST FOR CUSTOMIZATION

SAP Business One Professional

4.5 Forbes ADVISOR

Starting price	Free trial	Built-in CRM
$56 per user per month	None	Yes

Learn More

Read Forbes' Review

Editor's Take

Pros & Cons

Example Data Mining Project

- **Prediction of Student Academic Performance using Educational Data Mining**
  (Exploring factors influencing student success using demographic and academic records.)
- **Anomaly Detection in Cybersecurity**
  (Developing data mining techniques for detecting intrusions and unusual patterns in network traffic.)
- **Predicting Customer Churn for Subscription-based Services**
  (Using transactional and behavioral data to forecast customer retention.)
- **Fraud Detection in Financial Transactions**
  (Creating machine learning models to identify fraudulent activities in financial systems.)
- **Health Data Mining for Predictive Healthcare**
  (Analyzing EHR to predict disease outbreaks or patient readmission.)

Example Data Mining Project

Social Media Sentiment Analysis on Public Health Issues

- Use Twitter/Facebook data to analyze **public sentiment** during health crises (e.g., COVID-19).

Climate Data Mining for Extreme Weather Prediction

- Analyze climate data to identify trends and predict events like hurricanes or floods.

Recommender System for Personalized Online Learning

- Build recommendation systems to tailor learning paths for students based on behavior data.

Crime Prediction and Prevention

- Mine crime data to predict hotspots and suggest strategies for prevention.

Genomic Data Mining for Disease Prediction

- Analyze genetic data to predict susceptibility to diseases such as cancer or Alzheimer's.

Healthcare Data Mining Projects

Predicting Diabetes Risk Using EHR

- Analyze patient data to forecast diabetes risk using demographic, genetic, and lifestyle factors.

Predicting Heart Disease

- Use clinical and lifestyle data to predict the onset of heart disease and assist in early intervention.

Hospital Readmission Prediction

- Mine EHR data to identify patients likely to be readmitted within 30 days of discharge.

Early Diagnosis of Alzheimer's Disease

- Leverage longitudinal healthcare data to detect early signs of Alzheimer's.

Wearable Data for Diabetic Patient Monitoring

- Analyze wearable device data to monitor blood sugar, activity, and sleep in diabetic patients.

Want to apply for Data Scientist/ Machine learning Engineer?

Daniel Lee • 1st
Data/AI Jobs with Datainte...
Visit my website
18m • Edited •

Machine learning has 8 major problem areas:

- Regression
- Classification
- Clustering
- Optimization
- Computer Vision
- Forecasting
- Recommender System
- NLP / LLM

Differences Between Data Mining, Machine Learning, and Computer Vision

Aspect	Data Mining	Machine Learning	Computer Vision
Focus	Discovering patterns and insights in large datasets	Building predictive models from data	Understanding and interpreting visual data (images, videos)
Techniques	Clustering, Association Rules, Decision Trees	Supervised/Unsupervised Learning, Neural Networks	CNNs, Object Detection, Image Segmentation
Applications	Fraud detection, Market basket analysis, Healthcare insights	Spam detection, Recommender systems, Predictive analytics	Facial recognition, Medical image analysis, Object tracking
Data Type	Structured (tabular) and unstructured (text, social media)	Structured, semi-structured, and unstructured	Visual data (images, videos, 3D scans)
Goal	Extract knowledge from data	Learn patterns for prediction and decision-making	Analyze and extract meaning from visual content

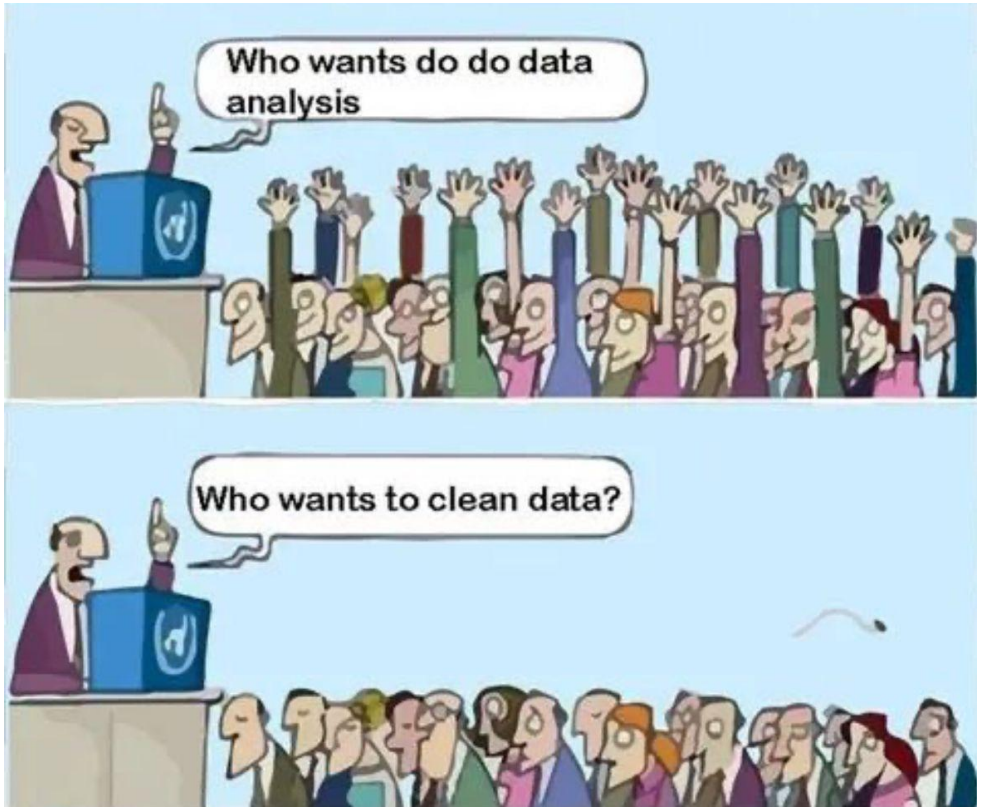
Table generated with assistance from GPT-4.o.

Similarities Between Data Mining, Machine Learning, and Computer Vision

Aspect	Data Mining	Machine Learning	Computer Vision	Similarity
Data-driven	Analyzes large datasets to extract insights	Learns from data to make predictions	Uses data (images/videos) to interpret visual content	All rely on analyzing and learning from data
Algorithms	Uses algorithms (e.g., clustering, decision trees)	Employs algorithms (e.g., neural networks, SVM)	Applies algorithms (e.g., CNNs, object detection)	All use computational algorithms for pattern recognition
Automation	Automates pattern discovery in data	Automates predictive model building	Automates visual content analysis	All aim to automate tasks traditionally done manually

Table generated with assistance from GPT-4.o.


The whole hospital
Radiology


Who wants do do data analysis
Who wants to clean data?


ET AL.
FIRST AUTHOR


ET AL.
FIRST AUTHOR
imgflip.com
@PHD_GENIE


ME LOOKING CLEVER
ON AN EMAIL
CHATGPT