Brij Kishore Pandey

THE COMPLETE MODERN DATA DICTIONARY

A CLEAR GUIDE TO MODERN DATA CONCEPTS AND TERMINOLOGY

AUTHOR

BRIJ KISHORE PANDEY

swipe ▶

Table of Contents

swipe ▶

Table of Contents

Basic Data Concepts

Data

Raw numbers, text, or files that haven't been processed yet. The basic building blocks of all information.

Information

Data that has been processed to be meaningful and useful. Helps answer questions or make decisions.

Metadata

Basic details about other data, like when it was created and by whom. Helps track and manage data effectively.

Dataset

A collection of related data gathered for a specific purpose. Contains multiple records or observations.

Data Point

A single piece of information in a dataset. Like one customer record or one sales transaction.

Data Element

The smallest single piece of meaningful data. Examples include a name, date, or amount.

Types of Data

Structured Data

Information organized in fixed fields within records or files. Easy to search and analyze, like spreadsheets and databases.

Unstructured Data

Information that doesn't fit a predefined model. Includes text documents, emails, videos, and social media posts.

Semi-structured Data

Information that has some organizational properties. Has tags or markers to separate elements but isn't strictly structured.

Quantitative Data

Numerical information that can be measured and analyzed. Can be used for calculations and statistical analysis.

Qualitative Data

Descriptive information based on characteristics or qualities. Focuses on descriptions rather than numbers.

Time-Series Data

Information collected at regular time intervals. Shows how things change over time.

Data Properties

Data Format

The specific way data is arranged and stored in a file or system. Common formats include CSV, JSON, or XML.

Data Schema

A blueprint that defines how data is structured in a database. Shows what fields exist and how they relate to each other.

Data Model

A framework showing how data elements connect and interact. Helps understand relationships b/w different pieces of info.

Data Hierarchy

The organization of data in parent-child relationships. Shows which elements are subordinate to others.

Data Relationship

The connections between different pieces of data. Shows how different data elements influence or relate to each other.

Data Attribute

A specific characteristic or property describing an item. Examples include price, color, or size.

Database Types

Relational Database

Stores data in tables with rows and columns that connect to each other. Best for structured data with clear relationships.

NoSQL Database

Stores and retrieves data without using traditional tables. Better for handling varied types of data that change often.

Graph Database

Focuses on relationships between data elements. Ideal for complex networks of connected information.

Document Database

Stores data in flexible, self-contained document formats. Good for content management and user profiles.

Time-Series Database

Optimized for handling time-stamped data. Perfect for monitoring systems and tracking changes over time.

In-Memory Database

Keeps data in computer memory for faster access. Used when speed is critical.

Storage Concepts

Data Lake

Central storage for raw data kept in its original format. Allows for flexible analysis of large amounts of data.

Data Warehouse

Central storage for structured, filtered data ready for analysis. Optimized for business reporting and analysis.

Data Mart

A subset of a data warehouse focused on a specific business area. Provides relevant data for specific departments or uses.

Data Swamp

A data lake that's become difficult to use due to poor organization. Results from lack of proper data management.

Data Platform

Complete infrastructure for storing, managing, and using data. Includes tools for collection, storage, processing, and analysis.

Data Repository

A central place where data is stored and maintained. Managed for specific purposes like reporting or archiving.

Architecture Types

Data Fabric

A unified architecture that connects data across an org. Provides consistent data management regardless of location.

Data Mesh

Approach where different teams manage their own data domains. Treats data as a product owned by business teams.

Lambda Architecture

Handles both real-time and batch data processing together. Has the speed of real-time with accuracy of batch processing.

Kappa Architecture

Processes all data as real-time streams. Simplifies data processing by using a single path for all data.

Medallion Architecture

Processes data through bronze, silver, and gold quality levels. Improves data quality through progressive refinement.

Hub-and-Spoke

Central data hub connected to multiple endpoint systems. Distributes data from one central point to many locations.

Infrastructure Elements

Data Center

Physical facility housing computing and storage systems. Contains servers, networking, and security equipment.

Cloud Infrastructure

Computing resources accessed over the internet. Provides flexible, scalable computing and storage.

Edge Computing

Processes data near where it's created instead of centrally. Reduces delays by bringing computing closer to data sources.

Hybrid Infrastructure

Combines local systems with cloud services. Balances control of local systems with cloud flexibility.

Data Grid

Network of servers working together as one system. Shares processing and storage across multiple machines.

Data Node

Individual server or storage unit in a larger system. Handles a portion of the total workload.

Pipeline Components

Data Pipeline

System moving data from source to destination with processing steps. Automates data collection, transformation, and loading.

ETL (Extract, Transform, Load)

Takes data from various sources, cleans it up, and puts it where it needs to go. Standard process for preparing data for analysis.

ELT (Extract, Load, Transform)

Loads data first, then transforms it where it lands. Modern approach taking advantage of powerful storage systems.

Data Flow

Path data takes as it moves through systems. Shows how data moves and changes in your system.

Data Stream

Continuous flow of data records as they're created. Handles data that arrives constantly in small amounts.

Data Batch

Group of data records processed together periodically. Processes data in chunks rather than continuously.

Processing Types

Batch Processing

Handles large groups of data at scheduled times. Best for tasks that don't need immediate results.

Stream Processing

Processes each piece of data as soon as it arrives. Ideal for real-time insights and immediate actions.

Parallel Processing

Multiple processors working on different parts of the same task. Speeds up processing by dividing work across systems.

Distributed Processing

Spreads processing tasks across multiple computers. Handles large workloads by sharing work between machines.

Event Processing

Analyzes and responds to data events as they happen. Triggers actions based on specific data conditions.

Query Processing

Retrieves and processes data based on specific requests. Turns user questions into database operations.

Engineering Concepts

Data Integration

Combines data from different sources into a unified view.
Makes different data systems work together smoothly.

Data Migration

Moves data from one system or storage type to another.
Includes planning, moving, and validating transferred data.

Data Replication

Creates and maintains copies of data in different locations.
Ensures data availability and backup protection.

Data Synchronization

Keeps multiple copies of data consistent and up-to-date.
Makes sure all systems have the same current information.

Data Orchestration

Coordinates multiple data processes and workflows. Manages timing and dependencies between data tasks.

Data Transformation

Changes data from one format or structure to another.
Prepares data for different uses and systems.

Analysis Types

Descriptive Analytics

Shows what has happened using historical data. Provides insights about past performance and trends.

Diagnostic Analytics

Examines why something happened. Uses techniques like drilling down to find root causes.

Predictive Analytics

Forecasts what might happen in the future. Uses historical patterns to make predictions.

Prescriptive Analytics

Suggests actions to achieve desired outcomes. Recommends best steps based on analysis.

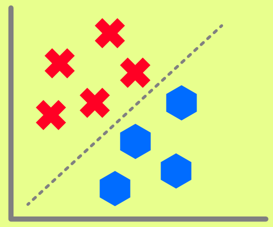
Exploratory Analysis

Initial investigation to find patterns in data. Helps understand basic trends and relationships.

Statistical Analysis

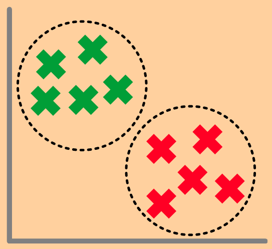
Uses mathematical methods to interpret data. Tests hypotheses and validates findings.

Machine Learning



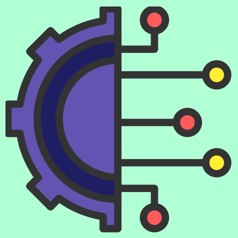
Supervised Learning

Trains on labeled data to make predictions on new data. Works with data where correct answers are known.



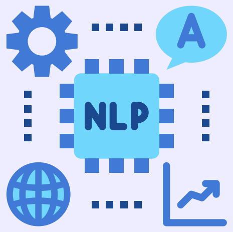
Unsupervised Learning

Finds hidden patterns in data without predefined labels. Discovers natural groupings and relationships in data.



Deep Learning

Uses multiple processing layers to learn high-level patterns. Handles tasks like image recognition and language processing.



Natural Language Processing

Enables computers to understand human language. Processes text for translation, sentiment analysis, and chatbots.

Computer Vision

Helps machines understand and process visual information. Enables image recognition and visual data analysis.

Feature Engineering

Creates better variables for machine learning models. Transforms raw data into more useful format for analysis.

Data Science Tools

Data Mining

Examines large datasets to discover patterns and relationships. Extracts meaningful insights from raw data.

Data Modeling

Creates statistical models to represent data relationships. Helps predict outcomes and understand data connections.

Data Visualization

Creates visual representations of data and insights. Presents data in charts, graphs, and interactive displays.

Data Clustering

Groups similar items based on their characteristics. Identifies natural categories within data.

Data Classification

Assigns items to predefined categories. Organizes data points into known groups.

Data Regression

Predicts numeric values based on other variables. Estimates relationships between different factors.

Quality Metrics

Data Accuracy

Measures how correctly data represents reality. Shows if values are true and error-free.

Data Completeness

Checks if all required data is present. Ensures no important information is missing.

Data Consistency

Verifies data is uniform across all systems. Confirms the same information appears everywhere it should.

Data Timeliness

Measures how current and relevant data is. Ensures data is available when needed.

Data Validity

Confirms data meets defined rules and formats. Checks if data makes logical sense.

Data Integrity

Ensures data remains accurate throughout its lifecycle. Maintains data quality over time.

Governance Elements

Data Stewardship

Manages and oversees data assets. Ensures proper data use and maintenance.

Data Lineage

Tracks data's origin and journey through systems. Documents how data changes over time.

Data Catalog

Lists and describes available data assets. Helps users find and understand available data.

Data Dictionary

Defines and explains data elements and terms. Provides common understanding of data meanings.

Data Policy

Sets rules and guidelines for data handling. Defines how data should be used and protected.

Data Standard

Establishes consistent formats and rules. Creates uniformity in data handling.

Security Concepts

Data Encryption

Converts data into coded form for protection. Prevents unauthorized access to sensitive information.

Data Masking

Hides sensitive data while maintaining format. Protects private information during testing and sharing.

Data Authentication

Verifies user identity for data access. Ensures only authorized users can access data.

Data Authorization

Controls what data users can access. Manages permissions for different users and roles.

Data Auditing

Records who accesses data and what changes are made. Tracks data usage and modifications.

Data Backup

Creates copies of data for disaster recovery. Protects against data loss.

Privacy Standards

GDPR

European Union's comprehensive data protection law. Sets strict rules for personal data handling.

CCPA

California's consumer privacy protection law. Gives California residents control over their personal data.

HIPAA

U.S. healthcare data privacy regulation. Protects medical information privacy.

Data Protection

Safeguards data from unauthorized access and misuse. Implements security measures to protect sensitive information.

Data Privacy

Ensures appropriate use of personal information. Controls how personal data is collected and used.

Data Compliance

Follows regulations and standards for data handling. Ensures legal and regulatory requirements are met.

Modern Tools

Data Build Tool (dbt)

Transforms data in a warehouse using SQL. Creates reliable data transformations and models.

Apache Airflow

Schedules and manages data pipelines. Automates complex data workflows.

Snowflake

Cloud platform for storing and analyzing data. Provides scalable data warehouse solutions.

Databricks

Platform for processing and analyzing big data. Combines data warehouse and machine learning capabilities.

Tableau

Creates interactive data visualizations. Turns complex data into understandable visuals.

Looker

Delivers business intelligence and analytics. Helps explore and share data insights.

Modern Concepts

DataOps

Improves speed and reliability of data analytics. Combines automated testing with monitoring and quality control.

MLOps

Standardizes machine learning system deployment. Manages the lifecycle of machine learning models.

Data Versioning

Tracks changes in datasets over time. Maintains history of data modifications.

Data Observability

Monitors data systems for problems. Ensures data reliability and quality.

Data Discovery

Helps users find and understand relevant data. Makes data assets findable and usable.

Data Democratization

Makes data accessible to all users. Removes barriers to data access and understanding.