Abstract

This Ebook covers Waterloo Data's must-know terms in the data management ecosystem. If you are a data enthusiast, an aspiring data professional, or dealing with data-related challenges and looking for solutions, your life will be significantly simplified if you know these few terms.

Index

34. Structured Query Language
35. Unstructured Data

Waterloo Data's Must-Know Data Management Terms: A Glossary

1. **Ad Hoc Analysis:** *Ad hoc analysis is a specialized form of data analysis designed to answer unusual or unexpected queries. It can swiftly analyze large amounts of data to help find answers to specific questions and create ad hoc reports.*

   Ad hoc analysis is a process of instantly using business data to find solutions to a question. Usually, these are one-time questions. It's like taking a quick and deep look into the data to answer one specific question by generating a one-time report (ad hoc report) in the form of dynamic dashboards with real-time data. Generally, the answer to these questions helps people make decisions for the short term. For example, a company's chief marketing officer (CMO) can use ad hoc analysis to discover how much the company spends on Google and Facebook ads and compare relative returns on investment.

2. **Big Data:** *"Big data" refers to information assets that are high in volume, velocity, and/or variety. It could be in various formats like structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs), and binary data (images, audio, video).*

   Big data is bigger and more complicated data that demands new, cost-effective ways of processing to drive insights, decision-making, and process automation, especially with net-new data sources. Among others, big data sources include - enterprise-generated data, transaction processing systems, customer databases, papers, emails, medical records, internet clickstream logs, mobile apps, and social networks. These data sets are so big that traditional data processing software can't handle them. But these huge amounts of data can be used to solve business problems that wouldn't have been previously solvable.

3. **Business Intelligence (BI)**: *BI refers to a set of practices supported by technology that analyze data and deliver insights beneficial to business owners, executives, and employees to make better decisions.*

Simply put, business intelligence is putting together all the company's information to give a comprehensive view that drives change, eliminates inefficiencies, and quickly adapts to market fluctuations. It combines business analytics, data mining, visualization, tools and infrastructure, and best practices to help organizations make data-driven decisions. Done properly, a company uses these BI processes or tools to generate reports that facilitate important decisions leading to higher profit and a competitive advantage in the industry. For example, sales managers can monitor revenue targets and sales rep performance through the sales pipeline using dashboards with reports and data visualizations.

4. **Cloud computing** *is a broad term for anything that involves hosting services and delivering them over the internet.*

   Cloud computing delivers computing services like servers, storage, databases, networking, software, analytics, and intelligence over the Internet (the "cloud"). Cloud computing enables faster innovation, more flexible resources, and greater economies of scale than traditional, on-premise, or data center-hosted solutions. In simple terms, cloud computing is when a computer system's resources, like data storage (called "cloud storage") and computing power, are available whenever the user needs them via the Internet, without the user having to manage them directly.

   An example is Dropbox, a system for storing and sharing files. Rackspace offers data, security, and infrastructure services. Salesforce is a cloud-based software company offering Software as a Service (SaaS) focused on sales, customer service, marketing automation, analytics, and application development. Azure, AWS, and Google Cloud are the most well-known hyperscale cloud service providers (CSPs) offering suites of cloud computing services, such as hosting, backup, and disaster recovery services.

5. **Columnar Database** *A columnar database stores data in columns instead of rows. This is good for analytical query processing and, by extension, for data warehouses.*

   A database management system (DBMS) that stores information in columns rather than rows is called a columnar database. A columnar database aims to write and read data to and from hard disk storage as quickly as possible so that the query can be answered faster. While columnar databases are easy to analyze, they are relatively difficult to update because they store all values from each column together, and to insert new data, the entire column has to be rewritten. In contrast, row-oriented databases store all the values in a row together (including the data across all columns that comprise that row). The Casandra, CosmoDB, Bigtable, and HBase databases are examples of columnar databases.

6. **Data governance** *sets internal rules, called "data policies," for how data is collected, stored, processed, and thrown away. It ensures data access to verified users only.*

   Data governance is about the quality and trustworthiness of data. It sets up the rules, policies, and procedures that ensure data is accurate, compliant, and safe. An example of data governance is when an organization starts to define data models, assign roles and responsibilities for using data, keeping old and new data, especially sensitive data, making data standards, putting protections in place, and ensuring the whole enterprise data architecture is secure.

7. **Data Integration** *refers to getting data 'where' and 'when' it is needed, changing it, putting it together with other data, and making it available.*

   Data integration combines different types and formats of data in a single place called a "data warehouse." The end goal of data integration is to create useful information that can be used to solve problems and learn new things. Without data integration, there is no way to use data from one system in another. For example, customer data integration involves getting information about each customer from different business systems like sales, accounts, and marketing. This information is combined into a single view of the customer used for customer service, reporting, and analysis.

8. **Data Lake:** *A data lake is usually a single place where raw copies of source system data, sensor data, social data, and data that has been changed to help with tasks like reporting, visualization, advanced analytics, and machine learning are stored*.

   The data in a data lake is not organized or in a hierarchy. Instead, each piece of data is stored in its own place. It stores data in its most basic form before it has been cleaned or analyzed. A data lake accepts and keeps all data from all sources, works with all types of data, and doesn't apply schemas (the way data is stored in a database) until the data is ready to be used. This implies a data lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs), and binary data (images, audio, video).

9. **Data management** *is ingesting, storing, organizing, and keeping track of the information that an organization makes and collects.*

   In simple terms, data management is collecting, storing, and using information in a safe, efficient way that doesn't cost too much. Data management helps businesses get

the most out of their data to make better decisions. Effective data management enables getting answers and insights from raw data to meet information needs.

10. **Data Management Plan:** *A data management plan, or DMP, is a formal document that shows how data will be handled during and after a research project*.

    A data management plan records how data changes over time. A DMP is a written document describing the data expected to be collected or generated during a research project, how it will be managed, described, analyzed, and stored, and how it will be shared and managed after the project is complete.

11. **Data Mart:** *A data mart is a subset of a data warehouse. It is usually geared toward a specific purpose or primary data subject and can be shared to meet business needs.*

    Data marts make specific data accessible to a certain group of users, like a specific department (HR, sales, finance, etc.) within a large organization. This lets those users quickly get actionable insights without wasting time looking through an entire data warehouse. It simplifies and expedites ad hoc queries, reports, and analytics. Consider a large retail organization, they would have data marts for seasonal products, lawn/garden, and even toys.

12. **Data Modeling** *is the technique of making a picture of a whole information system or just a part of it to show how data points and structures are connected.*

    Data models are graphical representations that show three things: 1) what data an organization collects, 2) where it is collected, and 3) how the data from each section relates to the data from other sections. The process of making these representations is called "data modeling"; it involves three stages - conceptual, logical, and physical. Usually, a data modeler starts by interviewing business stakeholders to gather requirements about business processes. Business analysts may also help design both conceptual and logical models. In the end, the physical data model communicates specific technical requirements to database designers.

    Apache Spark, IBM Infosphere Data Architect, and MySQL Workbench are some data modeling tools. A spreadsheet is also used as a modeling tool. In Excel, data models are used transparently, providing tabular data for PivotTables and PivotCharts. There are seven types of data modeling techniques - hierarchical, network, relational, object-oriented, entity-relationship, dimensional, and graph. A simple example is a data model that might say that the data element representing a car should consist of several other features that, in turn, reflect the car's color, size, and owner.

13. **Data Quality (DQ)** *measures data properties from different perspectives in dimensions such as Accuracy, Completeness, Consistency, Integrity, Reasonability, Timeliness, Uniqueness/ Deduplication, Validity, and Accessibility. In the organization, Data Quality plays a vital role to build trust in the Business Intelligence and Decision support system.*

    Data quality management tools can automate many processes required to ensure data remains fit for purpose across analytics, data science, and machine learning use cases. Organizations must find the best tool to assess existing data pipelines, identify quality bottlenecks and automate various remediation steps.

14. **Data Warehouse** *is an organization's central repository of structured data integrated from disparate sources to enable analytics and other business intelligence.*

    There are three main types of data warehouses - 1) Enterprise Data Warehouse (EDW), which has a broader scope, 2) Operational Data Store (ODS) is preferred for routine activities like storing records of employees, 3) Data Mart is designed for a particular line of business, such as sales or finance. A data warehouse helps organizations derive business insights from the integrated database to make informed decisions.

    Typically the data in data warehouses come from many different places, like marketing, sales, finance, and customer-facing apps, as well as internal apps. Complex queries can be made of the data in a data warehouse to make reports that enhance business efficiency, make smarter choices, and give a business a competitive edge. For example, business intelligence can be applied to a healthcare data warehouse to forecast outcomes, generate treatment reports, and share data with insurance providers, research labs, and other medical units.

15. **Database:** *A database is a set of related information that has been put together in a way that makes it easy to access, retrieve, manage, and update.*

    The database is a collection of data that is easy to find, use, and keep up to date. In simple terms, a database is a place where information is kept. The library is a physical example. There are different kinds of books housed in the library. The library is a collection of data, and the books are the information.

    The early database systems were simple but inflexible. In the 1980s, relational (structured) databases became popular, followed by object-oriented databases in the 1990s. More recently, NoSQL (unstructured) databases were introduced in response to the growth of the internet and the need for faster speed and processing of unstructured data.

16. **Database Administrator (DBA):** *A database administrator, or DBA, keeps databases running and safe and ensures that data is stored and retrieved correctly.*

    The DBA runs, controls, maintains and coordinates the database management system. Database administrators use tailored software (DBMS) to store and organize data. Planning for capacity, installation, arrangement, database design, mobility, performance analysis, security, troubleshooting, backup, and data recovery are all part of the job. Examples of DBMS are MS Access, MySQL, MongoDB, Oracle Database, etc.

17. **Database Management System (DBMS):** *A DBMS is software that stores and retrieves user data while adding a layer of security.*

    A database management system, or DBMS, is a computerized way to keep track of data. It is software used to create and maintain a database that lets users add, read, change, and delete data from the database. Database Administrators (DBAs) use DBMS to coordinate, plan, install, monitor, and efficiently authorize access to the database. MySQL, PostgreSQL, Microsoft SQL Server, Oracle Database, and Microsoft Access are all types of DBMS.

18. **Dimension:** *A "dimension" in data warehousing is a group of reference information about an event that can be measured. Dimensions organize and describe the facts and measures in a data warehouse to find meaningful answers to business questions.*

    A dimension is a structure that sorts facts and measurements into groups so that business questions can be answered. In a data warehouse, dimensions give numeric measures and structured labels. For example, a customer's dimension usually involves their first and last name, gender, date of birth, job, etc. The name and URL of a website are what make up its dimensions.

19. **Enterprise Data Warehouse (EDW):** *An enterprise data warehouse (EDW) is a database, or collection of databases, that centralizes a business's information from multiple sources and applications and makes it available for analytics and BI across the organization. A wider architectural diversity and functionality than a usual data warehouse characterizes the EDW. Because of the complex structure and size, EDWs are often decomposed into smaller databases, so end users are more comfortable querying these smaller databases.*

    Humans make many decisions daily based on what we've done in the past. When we have to decide, our brain uses the hundreds of billions of bits of data saved about what

has happened previously and instantly runs scenarios to predict future outcomes. Historically, business decisions were made similarly. However, with the advent of big data (and even the rapidly increasing volume and variety of data generated in the years preceding it), the human brain no longer suffices as the sole receptacle of data to drive enterprise decision-making. Fortunately, Enterprise Data Warehouses exist. As explained above, EDWs are the largest and most complex kind of data warehouse. They can be housed in an on-premise server or the cloud, though most are now cloud-based for reasons of cost-effectiveness, scale, and ease-of-management.

20. **Extract, Transform, and Load (ETL):** *ETL is a data integration process that takes information from various sources, then processes and loads it into a unified data repository, such as a data warehouse.*

    In short, different types of data are collected and cleaned up during a typical ETL process. To combine data from various sources into one central database: first, EXTRACT information from its original source, then TRANSFORM data by removing duplicates, combining it, and checking its quality, then finally LOAD the data into the target database. Often, after the ETL process, data is sent to a data lake and data warehouses.

21. **Extract, load, transform (ELT)** *is an alternative to extract, transform, load (ETL) used with data lake implementations.*

    In contrast to ETL, in ELT models, the data is not transformed on entry to the data lake but stored in its original raw format. This enables faster loading times. However, ELT requires sufficient processing power within the data processing engine to carry out the transformation on demand to return the results promptly. Since the data is not processed on entry to the data lake, the query and schema do not need to be defined prior; schema can be defined during the data read, also called Schema On Read. In the modern data warehouse, ELT architecture is becoming popular and use Data Lake and Big Data.

22. **Fact:** *A fact is a piece of quantitative information, like a sale or an install, stored in 'fact tables', which have a relationship with several dimension tables through their foreign keys.*

    In a data warehouse, a "fact" is a piece of data that shows a specific event or transaction, like selling a product or getting a shipment from a supplier with a certain number of items.

23. **In-Memory Databases** *are designed to store most of their data in memory. This is different from databases that store information on disks or SSDs.*

    A database that retains the whole set of data in RAM is called an "in-memory database". It means that when such a database is queried or changed, only the main memory is accessed - not physical or virtual media. This is advantageous because the main memory is much faster than any disk. Memcached is a prime illustration of this kind of database.

24. **Master Data Management**: *Master Data Management (MDM) ensures that the organization's shared data is consistent and accurate. This encompasses the people, processes, and systems that keep master data accurate and consistent.*

    Business and IT collaborate to ensure that the enterprise's data assets are uniform, correct, managed, semantically reliable, and accountable. A company can provide correct and unified master data using master data management. Master data, also called the single source of truth, includes information about each person, place, or thing in a business from across internal and external data sources and applications. This information can be de-duplicated, reconciled, and enriched to create a reliable source of business-critical data that helps employees make better-informed business decisions.

    The MDM *discipline* is about the principles of data governance that facilitate the creation of a trusted and consistent master record. While MDM *solutions* automate how business-critical data is governed, managed, and shared across applications used throughout the business. This includes overseeing and assisting in creating, reading, updating, and deleting (CRUD) master data.

25. **Metadata** *means "data about data." It is the information the data gives about one or more of its parts.*

    Metadata is the term for information about other important information. It is a data set, such as how it was collected, when it was collected, what assumptions were made in its method, its geographic scope, if there are multiple files, how they relate to each other, the definitions of the independent variable and, if applicable, what potential solutions there were, the high accuracy of any equipment used during data collection, the version of software used for analysis, etc. Information about the author, date created, and date modified, are basic examples of metadata.

26. **Object-Oriented Database Management System (OODBMS)** *is a database management system that allows data objects to be created and modeled.*

Information is stored in an object-oriented database management system in the same way that information is stored in an object-oriented programming language. Object-oriented programmers can use OODBMS to make products, store them as objects, and then copy or change existing objects to make new ones.

An object-oriented database stores complex and larger data along with the methods to use it (also called data encapsulation) compared to relational databases that store simpler data. OODB is preferred when businesses need high performance on complex data, such as real-time systems, architectural & engineering for 3D modeling, telecommunications, molecular science, and astronomy. The Versant Object Database, Objectivity/DB, ObjectStore, Caché, and ZODB are all examples of OODBMS.

27. **Online Analytical Processing (OLAP):** *OLAP is a way to use computers that allows users to easily and selectively pull out and query data to look at it from different angles.*

    It is software that lets people simultaneously look at data from more than one database system. Since OLAP data has more than one dimension, it can be evaluated by comparing it in many different ways. For example, a company can compare its computer sales in June and July and then compare those outcomes with sales from another location, which might be stored in a different database.

28. **Online Transactional Processing (OLTP):** *Online transactional processing enables the real-time execution of large numbers of database transactions by large numbers of people over the internet.*

    Online Transaction Processing is a data processing or change system that allows multiple transactions to happen concurrently. It is best for web-based transactions due to its defining characteristic of indivisibility or atomicity. This means that the transaction cannot remain in an intermediate or pending stage; it either fails or succeeds as a whole. Examples include online banking, shopping, entering orders, and simply sending text messages. A bank's database that allows Automated Teller Machine (ATM) transactions is another prime example of an OLTP system.

29. **Open Access (OA):** *Open access is a set of guidelines and practices that allow research results to be freely shared online with less restrictive copyright and licensing barriers than traditionally published works.*

    OA means free and unrestricted access to any information and all digital materials widely available via the Internet without licensing or copyright restrictions. Open Access can be used for any kind of digital content, such as articles, journals, books,

conference papers, theses, videos, music, etc. There are six types of OA - Repository-based or Green, Journal-based or Gold, Diamond, Hybrid, Bronze, and Black Open Access.

For example, IEEE Access is a multidisciplinary, online-only, gold, fully open access journal that continuously presents the results of original research or development across all Institute of Electrical and Electronics Engineers (IEEE) fields of interest.

30. **Predictive Modeling** *is a method of predicting what will happen in the future by looking at past and present data.*

    Predictive modeling is a common statistical method for determining how likely a customer will act in a certain way in the future based on how they have behaved in the past. In predictive modeling, data is collected, a statistical model is made, predictions are generated, and the model is validated (or changed) as more data becomes available. For example, risk models can improve underwriting accuracy in insurance by combining member information in sophisticated ways with lifestyle and demographic information from outside sources.

31. **Relational Database Management System (RDBMS):** *It is a database management system that stores and gives access to data points related to each other. Relational databases are built on the relational model, a simple way to show data in tables that make sense.*

    A relational database management system (RDBMS) or SQL database is a set of programs and features that lets IT teams and others add, change, or delete data in relation to other data. It is a system for managing information based on a data model. All the information is stored in tables. SQL Server, Oracle, MySQL, MariaDB, and SQLite are all examples of RDBMS.

    SQL databases are the best fit for heavy-duty transactional-type applications as they enforce business rules at the data layer, adding a level of data integrity not found in a non-relational or NoSQL database. Relationships in the SQL databases have constraints, while NoSQL databases provide scalability and flexibility to meet changing business requirements. Non-relational databases are best for Rapid Application Development. They provide flexible data storage with little to no structure limitations.

32. **Star Schema:** *A star schema is a multidimensional data model used to organize information in a database so that it is simple to understand and analyze.*

A star schema is an architecture model for a data warehouse in which one fact table refers to multiple dimension tables. When viewed as a diagram, the fact table is in the middle, and the dimension tables radiate from it like the points of a star. The star schema is the basic example of a dimensional model used in business intelligence and data warehousing to organize data into dimensions and facts.

33. **Structured Data:** *Structured data adhere to a pre-defined data model that humans and machines can easily access and analyze. It is standardized and often stored in databases. Hence, although it accounts for only 20 percent of data worldwide, it is the foundation of Big Data.*

    Think about data that fits perfectly into fixed rows and columns in spreadsheets and databases that use relational tables. Structured data is mostly kept in databases or other places with clear schemas. It is highly organized and easily understood by machine language. Most of the time, it looks like a table with rows and columns showing what it is and does, like relational databases. For example, an online store might maintain customers' names, addresses, phone numbers, orders, etc.

34. **Structured Query Language (SQL)** *is a standard programming language used to manage relational databases and do different things with their data.*

    Structured Query Language finds, organizes, manages, and changes data in relational databases. Most databases, such as SQL Server, Oracle, PostgreSQL, MySQL, and MariaDB, use this language to handle the data, with some additions and changes.

35. **Unstructured Data** *is data that does not fit into a data model and doesn't have a clear structure. This makes it more difficult to analyze.*

    Unstructured data is not set up according to a data model or schema and cannot be stored in a conventional relational database management system. Unstructured content often comes in the form of text (emails, documents, PDFs) and other types of media (audio and video files, images). 80–90% of the data that organizations make and collect is not structured. There is a tremendous amount of useful information in unstructured data stores that businesses can leverage to make decisions; however, it can be difficult to mine. Artificial intelligence (AI) and machine learning (ML) are making it possible to search through huge amounts of unstructured data to find valuable and useful business intelligence.

    Artificial intelligence simulates human intelligence processes with the help of machines, especially computer systems. AI currently provides three data management and analytics capabilities - prediction, automation, and optimization. Machine

learning is a subset of AI. Machine learning algorithms enable machines to leverage available data to predict outcomes and answer questions.