BDM1213 Data Encoding Principles

Week 02: Data Management

Dr. James Hong

Data management

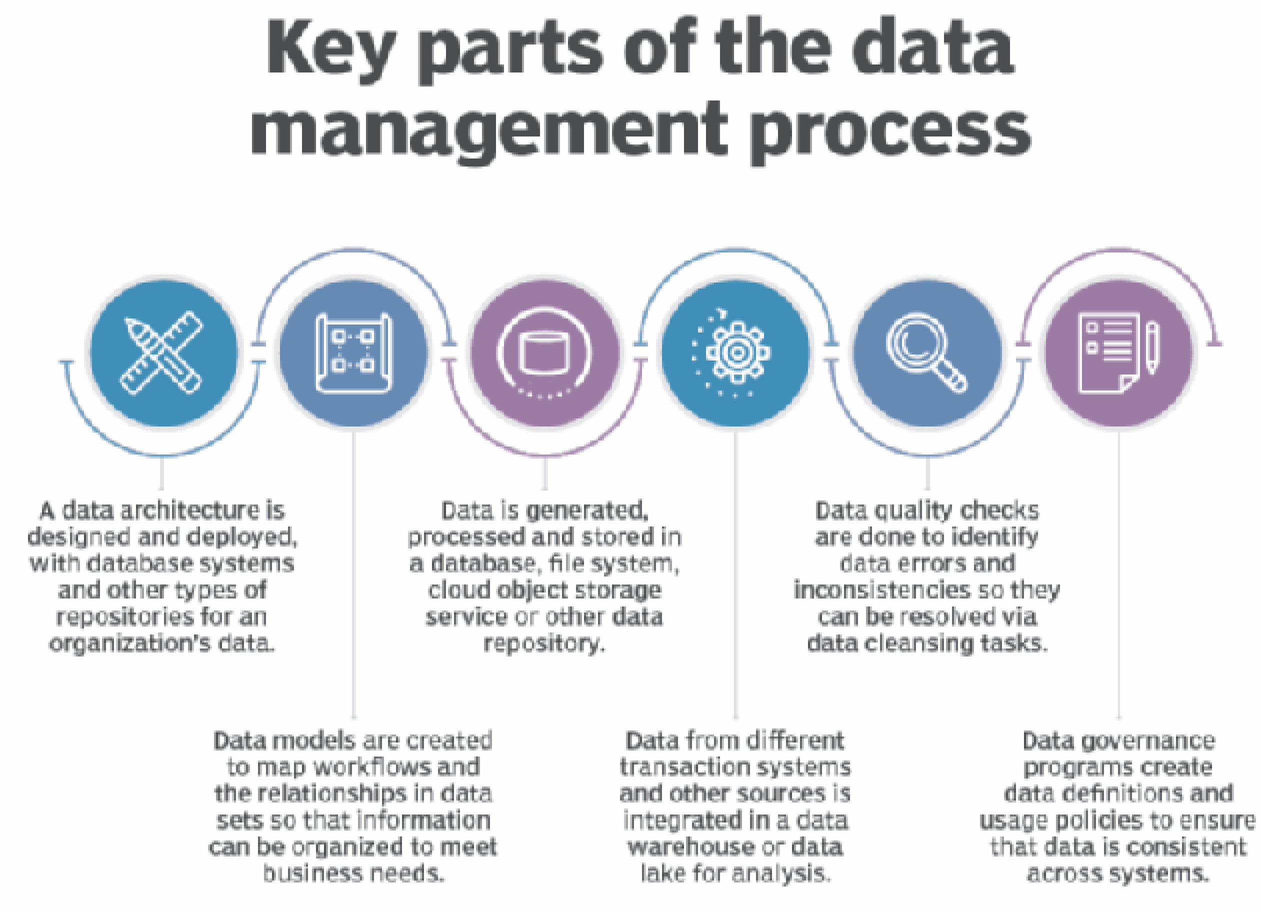
- Process of ingesting, storing, organizing and maintaining the data created and collected by an organization
- Crucial piece of deploying the IT systems that run business applications and provide analytical information to help drive operational decision-making and strategic planning by corporate executives, business managers and other end users

Tasks and roles in data management

- The data management process includes a combination of different functions that collectively aim to make sure that the data in corporate systems is accurate, available and accessible
- Most of the required work is done by IT and data management teams, but business users typically also participate in some parts of the process to ensure that the data meets their needs and to get them on board with policies governing its use

Importance of data management

- Data increasingly is seen as a corporate asset that can be used to make more-informed business decisions, improve marketing campaigns, optimize business operations and reduce costs, all with the goal of increasing revenue and profits.
- A lack of proper data management can saddle organizations with incompatible data silos, inconsistent data sets and data quality problems that limit their ability to run business intelligence (BI) and analytics applications -- or, worse, lead to faulty findings
- Data management has also grown in importance as businesses are subjected to an increasing number of regulatory compliance requirements, including data privacy and protection laws such as GDPR and the California Consumer Privacy Act.
- In addition, companies are capturing ever-larger volumes of data and a wider variety of data types, both hallmarks of the big data systems many have deployed.
- Without good data management, such environments can become unwieldy and hard to navigate.



Key parts of the data management process

A data architecture is designed and deployed, with database systems and other types of repositories for an organization's data.

Data models are created to map workflows and the relationships in data sets so that information can be organized to meet business needs.

Data is generated, processed and stored in a database, file system, cloud object storage service or other data repository.

Data from different transaction systems and other sources is integrated in a data warehouse or data lake for analysis.

Data quality checks are done to identify data errors and inconsistencies so they can be resolved via data cleansing tasks.

Data governance programs create data definitions and usage policies to ensure that data is consistent across systems.

Data management: Data architecture

- The separate disciplines that are part of the overall data management process cover a series of steps, from data processing and storage to governance of how data is formatted and used in operational and analytical systems
- Development of a data architecture is often the first step, particularly in large organizations with lots of data to manage
- An architecture provides a blueprint for the databases and other data platforms that will be deployed, including specific technologies to fit individual applications

Data management: databases

- Databases are the most common platform used to hold corporate data; they contain a collection of data that's organized so it can be accessed, updated and managed
- They're used in both transaction processing systems that create operational data, such as customer records and sales orders, and data warehouses, which store consolidated data sets from business systems for BI and analytics

What is a data warehouse?

- A data warehouse is a repository for data generated and collected by an enterprise's various operational systems.
- Data warehousing is often part of a broader data management strategy and emphasizes the capture of data from different sources for access and analysis by business analysts, data scientists and other end users

Data warehouses

- Typically, a data warehouse is a relational database housed on a mainframe, another type of enterprise server or, increasingly, in the cloud.
- Data from various online transaction processing (OLTP) applications and other sources is selectively extracted and consolidated for business intelligence (BI) activities that include decision support, enterprise reporting and ad hoc querying by users.
- Data warehouses also support online analytical processing (OLAP) technologies, which organize information into data cubes that are categorized by different dimensions to help accelerate the analysis process.

Structure of a data warehouse

- **Basic components of a data warehouse**
- A data warehouse stores data that is extracted from internal data stores and, in many cases, external data sources. The data records within the warehouse must contain details to make it searchable and useful to business users. Taken together, there are three main components of data warehousing:
  1. A data integration layer that extracts data from operational systems, such as Excel, ERP, CRM or financial applications.
  2. A data staging area where data is cleansed and organized.
  3. A presentation area where data is warehoused and made available for use.
- A data warehouse architecture can also be understood as a set of tiers, where the bottom tier is the database server, the middle tier is the analytics engine and the top tier is data warehouse software that presents information for reporting and analysis.

Data management: database administration

- Database administration is a core data management function.
- Once databases have been set up, performance monitoring and tuning must be done to maintain acceptable response times on database queries that users run to get information from the data stored in them.
- Other administrative tasks include database design, configuration, installation and updates; data security; database backup and recovery; and application of software upgrades and security patches.

Database administration: DBMS

- The primary technology used to deploy and administer databases is a database management system (DBMS), which is software that acts as an interface between the databases it controls and the database administrators, end users and applications that access them
- Alternative data platforms to databases include file systems and cloud object storage services; they store data in less structured ways than mainstream databases do, which offers more flexibility on the types of data that can be stored and how it's formatted. As a result, though, they aren't a good fit for transactional applications

Data management: Data modeling, integration and governance

- Other fundamental data management disciplines include data modeling, which diagrams the relationships between data elements and how data flows through systems
- Data integration, which combines data from different data sources for operational and analytical uses
- Data governance, which sets policies and procedures to ensure data is consistent throughout an organization; and data quality management, which aims to fix data errors and inconsistencies. Another is master data management (MDM), which creates a common set of reference data on things like customers and products

Data management tools and strategies

- The most prevalent type of DBMS is the relational database management system (RDMS).
- Relational databases organize data into tables with rows and columns that contain database records; related records in different tables can be connected through the use of primary and foreign keys, avoiding the need to create duplicate data entries.
- Relational databases are built around the SQL programming language and a rigid data model best suited to structured transaction data. That and their support for the ACID transaction properties -- atomicity, consistency, isolation and durability -- have made them the top database choice for transaction processing applications.

NoSQL

- However, other types of DBMS technologies have emerged as viable options for different kinds of data workloads.
- Most are categorized as NoSQL databases, which don't impose rigid requirements on data models and database schemas; as a result, they can store unstructured and semistructured data, such as sensor data, internet clickstream records and network, server and application logs.

Big data management

- NoSQL databases are often used in big data deployments because of their ability to store and manage various data types
- Big data environments are also commonly built around open source technologies such as Hadoop, a distributed processing framework with a file system that runs across clusters of commodity servers; its associated HBase database; the Spark processing engine; and the Kafka, Flink and Storm stream processing platforms
- Increasingly, big data systems are being deployed in the cloud, using object storage such as Amazon Simple Storage Service (S3)

Data warehouses

- Two alternative repositories for managing analytics data are data warehouses and data lakes.
- Data warehousing is the more traditional method -- a data warehouse typically is based on a relational or columnar database, and it stores structured data pulled together from different operational systems and prepared for analysis.
- The primary data warehouse use cases are BI querying and enterprise reporting, which enable business analysts and executives to analyze sales, inventory management and other key performance indicators.

Data marts?

- Data marts are another option -- they're smaller versions of data warehouses that contain subsets of an organization's data for specific departments or groups of users

Data lakes

- Data lakes, on the other hand, store pools of big data for use in predictive modeling, machine learning and other advanced analytics applications
- They're most commonly built on Hadoop clusters, although data lake deployments are also done on NoSQL databases or cloud object storage; in addition, different platforms can be combined in a distributed data lake environment
- The data may be processed for analysis when it's ingested, but a data lake often contains raw data stored as is. In that case, data scientists and other analysts typically do their own data preparation work for specific analytical uses

Data integration

- The most widely used data integration technique is extract, transform and load (ETL), which pulls data from source systems, converts it into a consistent format and then loads the integrated data into a data warehouse or other target system
- However, data integration platforms now also support a variety of other integration methods. That includes extract, load and transform (ELT), a variation on ETL that leaves data in its original form when it's loaded into the target platform
- ELT is a common choice for data integration jobs in data lakes and other big data systems
- ETL and ELT are batch integration processes that run at scheduled intervals. Data management teams can also do real-time data integration, using methods such as change data capture, which applies changes to the data in databases to a data warehouse or other repository, and streaming data integration, which integrates streams of real-time data on a continuous basis.

Data virtualization as a data integration technique

- Data virtualization is another integration option -- it uses an abstraction layer to create a virtual view of data from different systems for end users instead of physically loading the data into a data warehouse

Data governance, data quality and MDM

- Data governance is primarily an organizational process; software products that can help manage data governance programs are available, but they're an optional element. While governance programs may be managed by data management professionals, they usually include a data governance council made up of business executives who collectively make decisions on common data definitions and corporate standards for creating, formatting and using data.
- Another key aspect of governance initiatives is data stewardship, which involves overseeing data sets and ensuring that end users comply with the approved data policies.
- Data steward can be either a full- or part-time position, depending on the size of an organization and the scope of its governance program.
- Data stewards can also come from both business operations and the IT department; either way, a close knowledge of the data they oversee is normally a prerequisite.

Data governance and data quality

- Data governance is closely associated with data quality improvement efforts; metrics that document improvements in the quality of an organization's data are central to demonstrating the business value of governance programs
- Data quality techniques include data profiling, which scans data sets to identify outlier values that might be errors; data cleansing, also known as data scrubbing, which fixes data errors by modifying or deleting bad data; and data validation, which checks data against preset quality rules

MDM

- Master data management (MDM) is also affiliated with data governance and data quality, although MDM hasn't been adopted as widely as the other two data management functions. That's partly due to the complexity of MDM programs, which mostly limits them to large organizations
- MDM creates a central registry of master data for selected data domains -- what's often called a *golden record*. The master data is stored in an MDM hub, which feeds the data to analytical systems for consistent enterprise reporting and analysis; if desired, the hub can also push updated master data back to source systems

Data quality

- Data quality can be a major challenge in any data management and analytics project. Issues can creep in from sources like typos, different naming conventions and data integration problems. But data quality for big data applications that involve a much larger volume, variety and velocity of data takes on even greater importance
- And because quality issues with big data can create various contextual concerns related to different applications, data types, platforms and use cases, Faisal Alam, emerging technology lead at consultancy EY Americas, suggested adding a fourth V for *veracity* in big data management initiatives.

Why data quality for big data is important

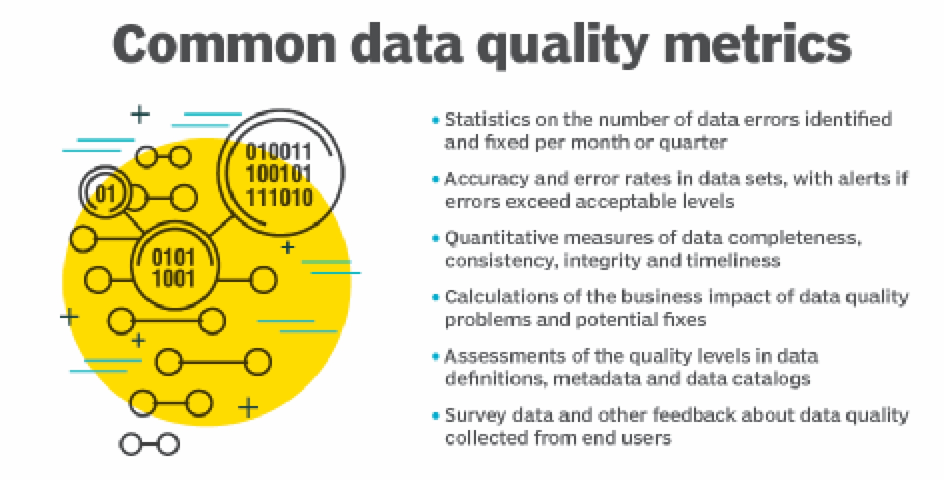
- Big data quality issues can lead not only to inaccurate algorithms, but also serious accidents and injuries as a result of real-world system outcomes.
- At the very least, business users will be less inclined to trust data sets and the applications built on them. In addition, companies may be subject to government regulatory scrutiny if data quality and accuracy play a role in frontline business decisions.
- Data can be a strategic asset only if there are enough processes and support mechanisms in place to govern and manage data quality, said V. "Bala" Balasubramanian, senior vice president of life sciences at digital transformation services provider Orion Innovation.

Bad data quality can accumulate costs

- Data that's of poor quality can increase data management costs as a result of frequent remediation, additional resource needs and compliance issues.
- It can also lead to impaired decision-making and business forecasting.

How data quality differs with big data

- **Scaling issues.** It's no longer practical to use an import-and-inspect design that worked for conventional data files or spreadsheets. Data management teams need to develop big data quality practices that span traditional data warehouses and modern data lakes, as well as streams of real-time data.
- **Complex and dynamic shapes of data.** Big data can consist of multiple dimensions across event types, user segments, application versions and device types. "Mapping out the data quality problem meaningfully requires running checks on individual slices of data, which can easily run into hundreds or thousands," Bansal said. The shape of data can also change when new events and attributes are added and old ones are deprecated.
- **High volume of data.** In big data systems, it's impossible to manually inspect new data. Ensuring data quality for big data requires developing quality metrics that can be automatically tracked against changes in big data applications and use cases.



01
010011
100101
111010
0101
1001

Common data quality metrics

- Statistics on the number of data errors identified and fixed per month or quarter
- Accuracy and error rates in data sets, with alerts if errors exceed acceptable levels
- Quantitative measures of data completeness, consistency, integrity and timeliness
- Calculations of the business impact of data quality problems and potential fixes
- Assessments of the quality levels in data definitions, metadata and data catalogs
- Survey data and other feedback about data quality collected from end users

Big data quality challenges and issues

- **Merging disparate data taxonomies.** Merged companies or individual business units within a company may have created and fine-tuned their own data taxonomies and ontologies that reflect how they each work.
- Private equity investments, for example, can accelerate the pace of mergers and acquisitions, often combining multiple companies into one large organization, noted Chris Comstock, chief product officer at data governance platform provider Claravine.
- Each of the acquired companies typically had its own unique CRM, marketing automation, marketing content management, customer database and lead qualification methodology data.
- Combining these systems into a single data structure to orchestrate unified campaigns can create immense challenges on big data quality.

Consistency in big data

- **Maintaining consistency.** Cleansing, validating and normalizing data can also introduce big data quality challenges. One telephone company, for example, built models that correlated with network fault data, outage reports and customer complaints to determine whether issues could be tied to a geographic location. But there was a lack of consistency among some of the addresses that appeared as "123 First Street" in one system and "123 1ST STREET WEST" in another system.

Variability in data preparation

- **Encountering data preparation variations.** A variety of data preparation techniques is often required to normalize and cleanse data for new use cases. This work is manual, monotonous and tedious. Data quality issues can arise when data prep teams working with data in different silos calculate similar sounding data elements in different ways, said Monte Zweben, co-founder and CEO of AI and data platform provider Splice Machine. One team, for example, may calculate total customer revenue by subtracting returns from sales, while another team calculates it according to sales only. The result is inconsistent metrics in different data pipelines.

Gathering too much data

- **Collecting too much data.** Data management teams sometimes get fixated on collecting more and more data. "But more is not always the right approach," said Wilson Pang, CTO at AI training data service Appen. The more data collected, the greater the risk of errors in that data. Irrelevant or bad data needs to be cleaned out before training the data model, but even cleaning methods can negatively impact results.

No data governance

- **Lacking a data governance strategy.** Poor data governance and communications practices can lead to all sorts of quality issues. A big data quality strategy should be supported by a strong data governance program that establishes, manages and communicates data policies, definitions and standards for effective data usage and to build data literacy. Once data is decoupled from its source environments, the rules and details of the data are known and respected by the data community, said Kim Kaluba, senior product marketing manager at data management and analytics software provider SAS Institute.

Best practices in managing big data quality

- **Best practices on managing big data quality**

1. Gain executive sponsorship to establish data governance processes.
2. Create a cross-functional data governance team that includes business users, business analysts, data stewards, data architects, data analysts and application developers.
3. Set up strong governance structures, including data stewardship, proactive monitoring and periodic reviews of data.
4. Define data validation and business rules embedded in existing processes and systems.
5. Assign data stewards for various business domains and establish processes for the review and approval of data and data elements.
6. Establish strong master data management processes so there's only one inclusive and common way of defining product or customer data across an organization.
7. Define business glossary data standards, nomenclature and controlled vocabularies.
8. Increase adoption of controlled vocabularies established by organizations like the International Organization for Standardization, World Health Organization and Medical Dictionary for Regulatory Activities.
9. Eliminate data duplication by integrating sets of big data through interfaces to other systems wherever possible.