Data Curation

Tushar B. Kute,
http://tusharkute.com

Data Curation

- Data curation is the process of creating, organizing and maintaining data sets so they can be accessed and used by people looking for information.
- It involves collecting, structuring, indexing and cataloging data for users in an organization, group or the general public.
- Data can be curated to support business decision-making, academic needs, scientific research and other purposes.

Data Curation

- Data curation is part of the overall data management process and sometimes is incorporated into data preparation work that gets data sets ready for use in business intelligence (BI) and analytics applications.
- In other cases, prepared data may be fed into the curation process for ongoing management and maintenance.
- Some organizations have formal data curator positions -- in ones that don't, data stewards, data engineers, database administrators, data scientists or business users may fill that role.

Data Curation

- The data curation process stems from the centuries-old practice of selecting, organizing and presenting objects as part of collections, such as artwork in a museum or books in a library.
- The term curation dates to ancient times and comes from the Latin word curae, meaning "care for" -- a meaning it still has today, including in relation to data.

Data Curation: Purpose

- In its business sense, data curation is a key component of an enterprise data strategy because it helps ensure the organization can make good use of its data and comply with data-related regulatory and security requirements.

Data Curation: Purpose

- Data curation achieves those objectives because it:
  - makes data findable and accessible;
  - provides the ability to trace information on data lineage; and
  - classifies data by various characteristics, such as whether it's public, proprietary or protected.

Data Curation: Focus

- Data curation focuses partly on understanding and organizing metadata, the set of details that provides information about the data itself.
- As such, data curation practices center on understanding where and how data is generated, as well as where it's stored.
- That includes creating searchable indexes on the data sets being curated; a data catalog also is built in many cases.

Data Curation: Focus

- These features create visibility into the data that's available to use in an organization -- a critical requirement as the volume of data being generated and collected continues to grow.
- In turn, this visibility helps optimize use of the data because BI and data science teams, business executives and other employees can find and access the data they need for analytics applications and operational decision-making.

Data Curation: Focus

- Effective data curation also engenders more trust in the data if users know it's accurate, reliable and up to date.
- That then creates more faith in the accuracy of data-driven decisions and speeds business actions and innovations based on data analytics.

Data Curation: Why?

- In many organizations, data is generated by a growing list of source systems, from conventional business applications to new edge computing devices connected to the internet of things.
- Big data systems often store a combination of structured, unstructured and semistructured data for analysis.
- More data is collected from various external sources for enterprise use.

Data Curation: Why?

- By bringing order to what could otherwise be a chaotic process of data ingestion and use, data curation helps keeps organizations from being overwhelmed by the explosion in data volumes and the proliferation of data sources.
- Without it, an organization could lose track of data sets and users wouldn't be able to get the information they need to do their jobs.
- Ultimately, that could result in wasted resources as users spend more time trying to search for and understand data. It could also lead to inaccurate analytics, flawed business decisions, lost opportunities and other problems that affect business performance.

Data Curation: Steps

- The process of curating data sets includes a variety of tasks, which can be broken down into the following main steps.
  - Identify data that's required for planned analytics applications.
  - Map the data sets and catalog the metadata connected with them.
  - Collect the data sets.
  - Ingest the data into a data warehouse, a data lake or other system.
  - Cleanse the data to fix inconsistencies, anomalies and errors such as invalid entries, missing values, duplicate records and spelling variations.

Data Curation: Steps

- Model, structure and transform the data to format it for particular analytics uses.
- Create searchable indexes of the data sets to make them available to users.
- Maintain and manage the data according to ongoing analytics needs and data privacy and security requirements.

Data Curation: Process

ASPECT	WHAT TO ASK
Operational metadata capturing	When was the data updated?
Data lineage	Where did the data come from?
Execution lineage	Which processes are responsible for the collected data?
Linking to business rules	How was the data converted or collected?
Data profiling	What does the data itself look like?
Data validation	How can I be confident about the data quality?
Data accuracy	Is the data complete, recent, well-defined and coherent? Does it match agreed-upon definitions?

Data Curator

- Some organizations, particularly large ones with mature or expansive analytics programs, have created data curator positions with responsibility for the full range of tasks associated with curating data.
- A data curator typically identifies required data sets and ensures they're collected, cleansed and transformed as needed.
- The curator also is responsible for making the data sets and information about them, such as their metadata and lineage documentation, available to users.

Data Curator

- The data curator's main objective is to ensure users can access the right data for analysis and decision-making.
- Curators also work with other members of the data management team and the IT and security teams to:
  - build required data pipelines;
  - ensure the pipelines are reliable and secure; and
  - set and maintain appropriate data governance, privacy and security standards for each data set.

Data Curator

- An organization may have multiple data curators: some responsible for data sets in specific domain areas, and one or more who are considered lead curators with responsibility for metadata management and overall data curation performance.

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com