Term Paper

“Making Big Data Discoverable At Data-Driven Companies”

By,

Mugdha Kamat

BIA 678 A

1. Abstract:

The digitally generated data is projected to grow from 33 ZB in 2018 to 175 ZB by 2025. This unprecedented growth leads to increased complexity in the data-driven organization's ecosystems and engineering architecture. Having Big Data spread across hundreds of different data stores is nearly not enough. Navigating through this data and finding specific data points in this chaos that will generate key insights for business decisions has become an important aspect as well. The challenges associated with handling Big Data are productivity and compliance. The solution for this problem indeed lies with the metadata and not data. This paper aims to discuss the concept of metadata, important standards associated with it and the ways big technology companies utilize this metadata for actual data exploration.

2. Introduction:

When we shop for grocery on Amazon, we use metadata, when we listen to our playlist on YouTube Music, we use metadata, when we pick some book from the library, again we use metadata. What initially appeals us to select that particular item/song or book is not the entity itself but perhaps the attractive image, renowned artist, or the interesting title. These bits and pieces of information are nothing but metadata.

3. Metadata:

The catchy literal definition of metadata is data about data. But this definition seems to be obscure and needs more exploration. The Data, Information, Knowledge and Wisdom pyramid consisting of data at the bottom and wisdom at the apex explains the logical stages of processing carried out at each layer eventually leading towards better illuminated decisions. Here, data is considered something as raw and unprocessed, for example, it could be merely observations collected from machines/equipments etc. On the other hand, information may evolve from the data analysis and the ‘What’, ‘When’, ‘Who’ questions could be answered here. Incidentally, information conveys a communicative aspect. Therefore, we can say that data is potential information antecedent to anyone being informed about it. This leads us to a modified definition of Metadata which is a Statement about a Potentially Informative Object [1].

Metadata can be categorized into the various types according to its purpose. Complementing to the 3Vs of Big Data – Velocity, Volume, and Variety, metadata can be classified according to three key sources of data context such as Applications, Behavior and Change, in short ABC of metadata [3]. Each of them represents a critical shift from the elementary metadata of traditional enterprise management.

i. Applications Context Metadata: This is the core information that is required by humans to interpret data. The applications context consists of basic descriptions (schemas, tags), semantics associated with the data.

ii. Behavior Metadata: This contains information of data creation and consumption over time. It consists of details of ownership, frequent users and processes, lineage, common patterns of use, basically logs of usage such as ‘digital exhaust’ cast off from computations on the data.

iii. Change Metadata: This information consists of history of versions, changes in the structure and content of the data over time. This is particularly useful for enabling debugging and auditing as the version history of the entire data pipeline can be tracked.

Now, in most of the organizations, the metadata described by the above ABC context consists of, but not limited to the below items:

Data Stores	Dashboards, Reports	Streams	Processing
Structured Stores : Hive, Presto, MySQL, Snowflake Unstructured stores: S3, Google Cloud Storage	Saved queries and reports from tools such as: Tableau, Looker, Apache Superset	Apache Kafka, AWS Kinesis	ETL, Streaming jobs, Machine learning workflows

Table 1: Sources of metadata

When the data users are provided with such relevant metadata ready at hand, it makes them even more productive to further take suitable actions right away based on the insights.

4. Metadata Standards:

When we talk about metadata, it is important to stick to standard vocabularies, standard schemas etc. It becomes even more challenging when organizations need to migrate to a newer platform and the data does not match between systems. It is indeed a difficult process to achieve seamless and concise retrieval of the metadata. Carefully structured and mapped metadata plays a vital role in order to reach this objective. Need of standardization helps search tools work better, ensures reusability, interoperability, and access to data. We will discuss some of the important metadata standards in the next part.

4.1. Dublin Core Metadata Standard:

The Dublin Core Metadata is one such lucid and widely used metadata schema serving various physical (books, art works) and digital resources (video, images, web pages, etc.) [6]. Dublin Core metadata may be utilized for variety of reasons, such as plain resource description, integrating metadata vocabularies of disparate metadata standards, and offering interoperability for metadata vocabularies in semantic applications.

The two main principles of Dublin Core are Dumb Down and One-to-One. The Dumb Down principle states that if the element is irrelevant for describing, then it is safe to ignore it of the metadata record. The One-to-One principle asserts that one-to-one relationship must be maintained between object and metadata record to refrain from ambiguity.

The three common terms often used with metadata are Elements, Values and Records. A metadata schema is a way of controlling the kind of statements to be made and according to Dublin Core, 15 kinds of statements are allowed. An element is a category of the statement allowed to make using a metadata schema and the value is the data provided in the statement. Broadly the characteristics of the Dublin Core records are that the metadata elements are optional, repeatable and can be displayed in any order.

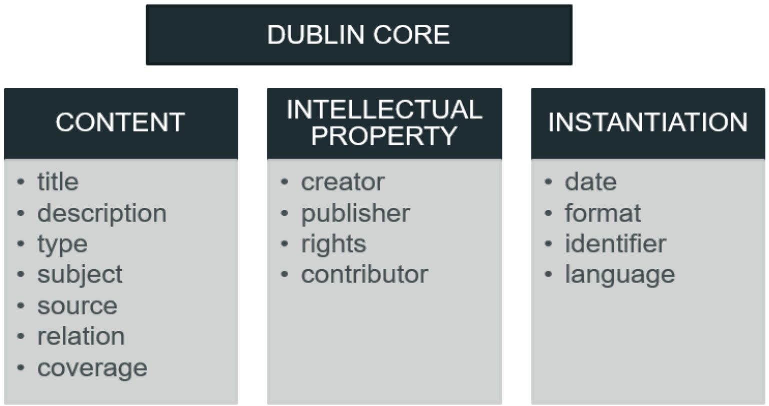


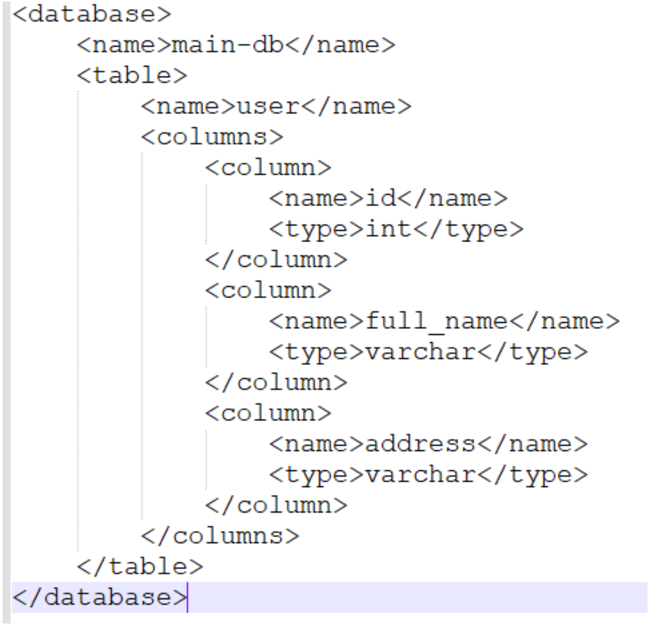
Fig 1. Dublin Core metadata elements

The above figure lists 15 metadata elements of Dublin Core categorized as Content, Intellectual Property, and Instantiation. They offer information regarding the catalog for easier search.

4.2. Common Warehouse Metamodel Standard:

The Common Warehouse Metamodel (CWM) is another standard from the Object Management Group (OMG) for modelling metadata for data warehouses supporting relational, non-relational and multidimensional objects in the environment [7]. CWM permits organizations and tool vendors to illustrate their metadata, models, and processes in a standard format for enabling simple exchange of metadata between different tools and platforms in distributed heterogeneous environments. It supports other formats such as

Unified Modeling Language (UML), Meta Object Facility (MOF) and XML Metadata Interchange (XMI).



```xml
<database>
    <name>main-db</name>
    <table>
        <name>user</name>
        <columns>
            <column>
                <name>id</name>
                <type>int</type>
            </column>
            <column>
                <name>full_name</name>
                <type>varchar</type>
            </column>
            <column>
                <name>address</name>
                <type>varchar</type>
            </column>
        </columns>
    </table>
</database>
```

Fig 2. Common Warehouse Metamodel Standard

The above figure demonstrates a CWM description of a table from a relational database with metadata description of the columns.

Following up on the metadata standards, we will now discuss two case studies of Netflix and Lyft who built their own tools to make metadata across various data stores explorable to their data teams. These tools are a solution to the basic questions like : Where can one find the data? Whom can one ask for accessing it? Is this available data trustworthy? And so on.

5. Netflix Metacat:

Netflix being a data driven company has approximately 213 million paid subscribers worldwide, with around 100 million monthly active users. At peak hours, Netflix records 8 million events per second. Netflix data warehouse consists of hundreds of petabytes of data stored in various data sources such as Amazon S3 (via Hive), MySQL, Druid, Elasticsearch, Redshift, and Snowflake. Considering the scalability issues, Netflix designed its core architecture of the big data platform such that it could operate together as a 'single data warehouse' across diverse datasets. Thus, Metacat [5] was built – a mediator that can track data across infrastructure making data discovery and management easier.

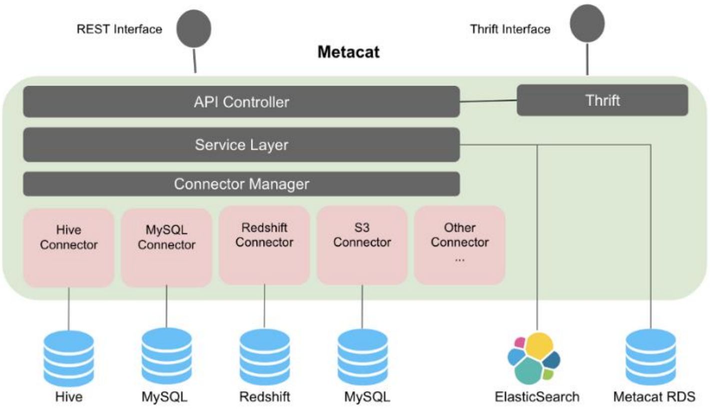


Fig 2. Metacat in Netflix data ecosystem

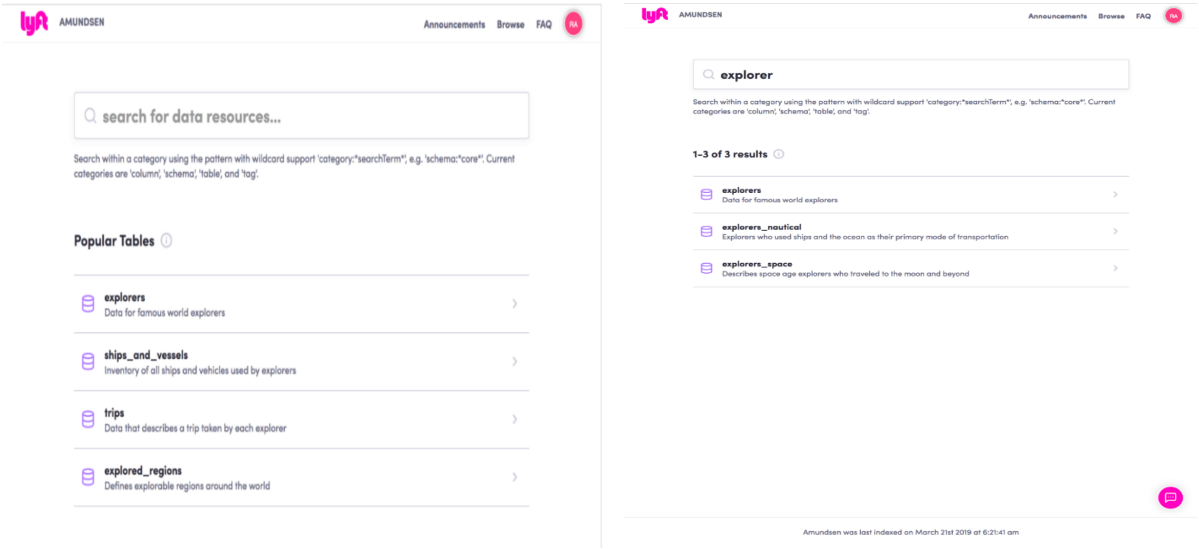
The above figure explains the role of Metacat filling a key gap in Netflix ecosystem. It exposes federated API for indexing resources through the search engine. When end users (for example data analysts, data scientists) search any term with a search query on UI, Metacat responds back appropriate results with the help of REST APIs and ElasticSearch. ElasticSearch utilizes full-text search, auto-suggest and auto-complete functionalities by syncing schema metadata along with user defined metadata with Metacat. Having ElasticSearch search, which essentially is built on top of Apache Lucene engine as a distributed data store, along with advanced search functionalities improves the speed and aids in faster metadata discovery. Metacat also has a push notification system that notifies the users of the schema changes, or when a table is dropped, for the data clean up in the infrastructure.

6. Lyft Amundsen:

Lyft also has their own data discovery tool known as Amundsen [4], code of which has been open-sourced from February 2019, thereby empowering the end-users(data analysts, data scientists) with the relevant metadata. The motivation behind building this tool was to have a holistic perceptive of the entire data ecosystem for making informed decisions.

Amundsen contains a DataBuilder framework inspired by Apache Gobblin, which extracts metadata across the dashboards, databases and HR systems. The search functionality is based on the Elasticsearch and neo4j is used as the backend database solution.

Amundsen also has diverse set of integrations with various data stores, dashboards, Airflow and so on, mostly of which have been contributed by the community.



search for data resources...

Search within a category using the pattern with wildcard support 'category:*searchTerm*', e.g. 'schema:*core*'. Current categories are 'column', 'schema', 'table', and 'tag'.

Popular Tables

explorers
Data for famous world explorers

ships_and_vessels
Inventory of all ships and vehicles used by explorers

trips
Data that describes a trip taken by each explorer

explored_regions
Defines explorable regions around the world

explorer

Search within a category using the pattern with wildcard support 'category:*searchTerm*', e.g. 'schema:*core*'. Current categories are 'column', 'schema', 'table', and 'tag'.

1-3 of 3 results

explorers
Data for famous world explorers

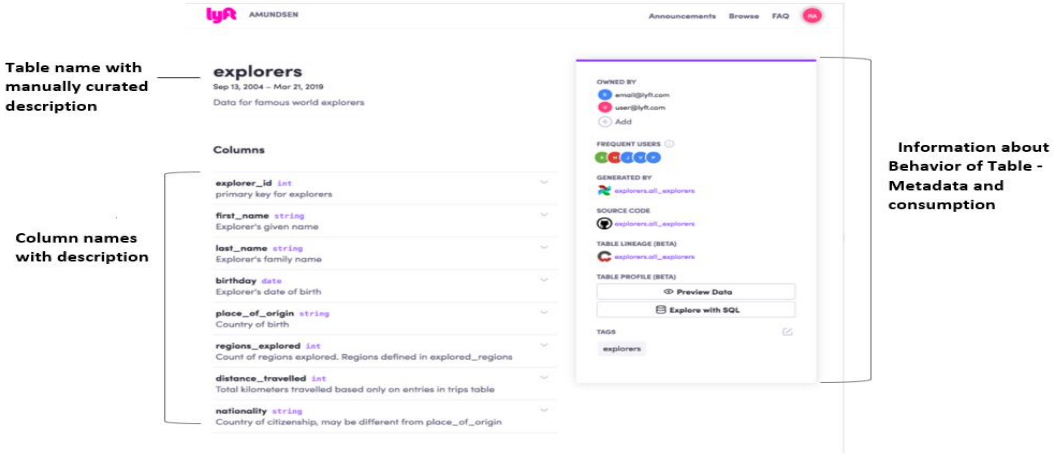
explorers_nautical
Explorers who used ships and the ocean as their primary mode of transportation

explorers_space
Describes space age explorers who traveled to the moon and beyond

Amundsen was last indexed on March 21st 2019 at 6:21:41 am

Fig 3. Amundsen Landing page

The above figure is a snapshot of landing page of Amundsen where the users can search for the desired data resources along with recommendations. The search ranking is based upon an algorithm similar to Google's Page Rank where popular queries show up above in the list. The right side of the above figure shows the results of the search 'explorer' with some in-line metadata.


AMUNDSEN
Announcements
Browse
FAQ
RA
Table name with manually curated description
Column names with description
Information about Behavior of Table - Metadata and consumption
Preview Data
Explore with SQL

explorers
Sep 13, 2004 – Mar 21, 2019
Data for famous world explorers

Columns

explorer_id int
primary key for explorers

first_name string
Explorer's given name

last_name string
Explorer's family name

birthday date
Explorer's date of birth

place_of_origin string
Country of birth

regions_explored int
Count of regions explored. Regions defined in explored_regions

distance_travelled int
Total kilometers travelled based only on entries in trips table

nationality string
Country of citizenship, may be different from place_of_origin

OWNED BY
email@lyft.com
user@lyft.com
Add

FREQUENT USERS

GENERATED BY
explorers.all_explorers

SOURCE CODE
explorers.all_explorers

TABLE LINEAGE (BETA)
explorers.all_explorers

TABLE PROFILE (BETA)

TAGS
explorers

Fig 4. Amundsen Details page

The above figure shows the Detail page after selecting the desired result from the searches which captures both programmatically and manually curated metadata such as column lists with description, behavior metadata such as owners and consumers of the data and the associated tags of the data.

Amundsen has been successfully adopted by various technical as well as business and marketing teams in Lyft bringing down the time for asset discovery to 5% of the baseline thereby increasing their productivity by 20%.

7. Next Challenges:

Though, with metadata services, organizations are more actionable now than before, there is still a scope to expand by building new robust functionalities. For enhancing the datawarehouse experience, Netflix is now working to incorporate schema and metadata versioning features to be able to track the trends in metadata changes over time along with building a pluggable metadata architecture. Lyft’s roadmap further includes integration with HR systems like Workday and adding streams like Apache Kafka, AWS Kinesis to Amundsen. Along with this all organizations are working to tackle the issue of maintaining integrity of the metadata.

8. Conclusion:

We learned that for overcoming the challenges of big data, companies need to evolve and devise innovative data management techniques. It is interesting to see that utilizing big data seeks the need of metadata exploration. Through this paper we also covered Netflix’s

Metacat and Lyft's Amundsen metadata exploration platforms which empower all the data users from new hires to experienced Data Scientists, Data Analysts and increase their productivity by faster data discovery. Most of the times these solutions are customized according to organization's existing data infrastructure and have tremendous scope to scale up in future. Integrating ElasticSearch for full-text search/auto-complete is a great solution in terms of speed but it comes with some synchronization challenges. It would be interesting to know how Netflix or Lyft are handling the synchronization issue that reflects changes in the metadata instantaneously. Metacat's code is available on GitHub as an open-source project maintained by Netflix but currently there's no documentation or record of any other organization using it. Lyft's Amundsen is more popular with documentation available for users to start and test locally via Docker. In spite of being fairly new, companies such as Asana, Instacart and Square have already adopted it. To conclude, data discovery is indeed a critical step in the data science overflow and needs attention.

9. References:

[1] Marcia Lei Zeng, & Jian Qin. (2016). Metadata: Vol. 2nd edition. ALA Neal-Schuman.

[2] Creating a data-driven enterprise with DataOps : insights from Facebook, Uber, LinkedIn, Twitter, and eBay

[3] Joseph M. Hellerstein, Vikram Sreekanti. Ground: A Data Context Service

[4] https://eng.lyft.com/amundsen-lyfts-data-discovery-metadata-engine-62d27254fbb9

[5] https://netflixtechblog.com/metacat-making-big-data-discoverable-and-meaningful-at-netflix-56fb36a53520

[6] https://www.dublincore.org/specifications/dublin-core/usageguide

[7] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_900