Data Catalogue Approaches, Implementation and Adoption: A Study of Purpose of Data Catalogue

Nouby Mahdy Ghazaly
Associate professor, Mechanical Engineering Departments Faculty of Engineering South Valley University Egypt

Abstract: - The amount of data and information present in an organisation is huge and on a daily basis, it increases. All the organisations keep their data and information in an organised manner using various technologies and tools. The amount of data stored in database is more and in some critical scenarios, it takes longer than usual to locate the data needed by the business or organisation to make quick critical decisions. In order to solve this issue, one new approach is being used which is Data catalogue. It is the process of organising the data in such manner that, it is easy to locate the data from the database in which it is being stored. Data catalogue is the process of preparing indexing of the data and information present in the database. Data catalogue is the type of a catalogue which will represent all the type of data and information which is present in the organisation in an organised manner. It makes it easy for the data professionals to collect, gather, arrange data in organised manner. The paper will discuss the process of implementation of data catalogue, its purpose, benefits, challenges, various approaches etc.

Keywords: - Data Catalogue, Types of Data Catalogue, Importance of Data Catalogue, Advantages of Data Catalogue, Challenges Data Catalogue, Implementation of Data Catalogue.

Introduction: - [1]
In any business or organisation, a lot of data and information is stored on a daily basis in the organisation’s database. There are many departments in any business or organisation where each department will have unique set of data and information which are crucial for the organisation. At times, due to large amount of data stored in the database, the professionals find it difficult to obtain particular data on time. They need to search each data set in order to find particular information which is a time-consuming process. If there is an index which will have information about the data sets, then it will be easy to understand and locate the particular information at a much higher speed and it will save time. For example, if we go to a mall to buy electronic appliance like fridge, then instead of checking each fridge one by one, we can refer to the catalogue which will have information like brand, cost, colour description, specifications. This will make it easy for the customer to understand the features and he can select particular variety and then observe it to check the look and feel of the fridge. Similarly, data catalogue is the process used by the data professionals to organise the already existing data of the organisation in an organised way so that any information can be located easily on time. It is a kind of metadata which will have indexing of all the data and information present in all the departments of the business. It can also be considered as library of the data asset of the organisation where proper indexing of data is done, proper listing of data set is done and also it will contain information about which particular data is present at which place in the database. The data catalogue will also give information about the source of the data, origin of data, what type of data is present, quality of data etc. It helps the organisation to perform advance search and the visualisation of the data increases. It is available in a centralised way so that anybody who is authorised to use the data asset of the organisation, can refer to it and is able to locate the data easily. The data professionals, with the help of data catalogue also able to find new possible available data sources which was not possible without it. There are many data catalogue tools available which makes the job easy for the data professionals to search and prepare metadata for already existing data sets of the organisation. Hence, data catalogue is the inventory or a library which is description of the data sets and data assets in an organisation which is useful to find and locate the most accurate data which will be used to perform business analysis or to make critical business decision.

Types of Data Catalogue: - [2]
Data catalogue is a type of metadata which is used to organise the data assets of the organisation in such a way that it makes it easy to find, locate the data on time which can be used for evaluation or taking some business decisions quickly. There are many types of metadata which can be described as following: -
Figure 1. Types of Metadata (Data Catalogue)

1. Process Metadata:
   - Process metadata explains the information as who has the authorised access to use the data present in the database.
   - It is also known as administrative metadata as it gives information about how the data is created and stored.
   - It also gives information about that who has accessed the data at what time, how it has been used, and who was the user who has used it.
   - It also gives information about the history of the data whether it is recently updated, whether the source of data is reliable enough to use it for evaluating business needs etc.
   - Process metadata also gives information about the software used by the users, and hence it can also be used to solve queries of the users.

2. Technical metadata:
   - It is also known as structural metadata as it is used to give information about how the data is visible to the users, and how it is arranged.
   - It is used to represent the data and explains its structure by providing information like tables, rows, columns, indexes, its connections etc.
   - It also helps the professionals to determine that how they can use the available data which means it will explain whether the data can be used as it is or it needs any conversion for making it useful for evaluation etc.

3. Business metadata:
   - It is also known as external metadata, which helps to understand the value of the data as per the business aspects.
   - It describes the value of the data asset/ data set for the business and organisation, it will also help to understand the purpose of the data asset present in the business.
   - It also helps to provide information about compliance, rules and regulations, etc.

Importance of Data Catalogue/Purpose of Data Catalogue: [3]

With the increase in the amount of data, it is becoming crucial to find the correct data at correct time. There are also many rules and regulations which are applicable in order to use data. Data catalogue solves this issue by creating library of the data assets which will have indexing of the data sets present in the organisation. This makes it easy to get information as where the particular data is present, type of source of data, quality of data, whether the data is updated or not. Data catalogue also helps to monitor governance as well as helps to understand who is accessing the data, what is the purpose of data, how it is being accessed, how it is being used. Following are the challenges which are solved by using data catalogue:

- Wastage of time to locate data: -
  This is one of the main challenges which is solved by using data catalogue. It is difficult for the data professionals to locate particular data from large amount of data. Manual way takes a lot of time in order to locate the data. But in data catalogue since, information regarding the location of the data is present, it makes it easy for the data professionals to locate the data in less time.

- Reusing data: -
  In some scenarios, lot of data is duplicated due to lack of awareness of the already existing data. Each time if a user is not able to locate the data he is looking for, he will search it afresh and then store it which makes lot of copies of the data. Data catalogue solves this issue by providing proper details of the type of data and information in a catalogue form which can be referred by any authorised user to check if it is already available in the database or not. If yes, then there will be no need for the user to research and upload new data set. He can easily access the already existing dataset.

- Unnecessary occupied space in database: -
  If the data set is duplicated in the database, then a lot of space will be occupied in the database which can be utilised for other purpose. If data catalogue is used then the duplication of data set can be avoided which in turn helps to free up the space.

- Data Access: -
  Data catalogue not only makes the data set search easy but it also gives detailed information about who is accessing the data, how the data is being accessed, time of data being accessed etc. Only authorised users will be able to use the data catalogue. There are data accessing protocols being used in the data catalogue indexing.
• Data Manipulation: - Significant assessment highlights incorporate capacities to see a dataset, see all related metadata, see client evaluations, read client surveys and caretaker comments, and view information quality data.

• Advanced data set search: - The proper way to search data is by using proper business terms, keywords etc. The users who are not aware of the technical terms uses simple natural language which in some scenarios makes it difficult to locate the data which is already existing. This issue is also resolved by using data catalogue as it has unique common terms and keywords which can be easily understood by all and makes it easy to locate particular data set and data assets.

Implementation process of Data Catalogue: - [4]

1. Deciding Data Strategy: - Following steps are performed in order to decide the data catalogue strategy: a. Identify the goals for data: - The goals and objective to implement data catalogue is decided in this step. It is decided that how the data catalogue will help to improve the growth of the business, how it be used to overcome the challenges of traditional methods of data search, what is the mission and vision of data catalog, etc. b. Determine use cases: - Once the goals and objectives of the data catalogue is determined, then it is time to identify the use cases so that the type of tools used for data catalogue will be determined. c. Decide how data catalogue will be used: - This is the main step once the goals, use cases has been identified. It is very important to decide that how the data catalogue will be used. It is also identified that how it will be in alignment with the business goals and objectives.

b. Data strategy: - In this step offense and defense strategy is identified. It is tried to identify whether the data catalogue will be single point of source or multiple point of source.

2. Preparing for Data Catalogue: -
   a. Documentation of use cases: - The first step for the preparation of data catalogue will be to plan for the particular use cases of the business. These include deciding the data governance plans, analytics part, cloud transformation, risk and compliance etc. Documenting all these factors will help the business understand that what are the key benefits for the users to use data catalogue.

b. Select Data Catalogue: - Once the goals and objectives are identified, use cases are documented, then all this will help to select the correct form of the data catalogue which will help to improve the efficiency of the business.

3. Implementation of Data Catalogue: - This is performed in 4 steps:
   a. Involve users as early as possible: - The managers should decide and find out who are the users who are eager to learn something new. The desired candidates must be selected and trained on the new technology.
   b. Iterative application: - Each person involved in the training must be assigned particular roles and responsibilities so that they can learn it as soon as possible. Iterative application process should be used by the managers to decide the key areas of improvement.
   c. Establish roles: - The roles and responsibilities must be assigned in initial stages so that everybody should put in their complete efforts and they are responsible for documenting and for accuracy of the data.
   d. Outcomes and Benefits: - Execution is a nonstop interaction. The information index needs your social help all through. Share examples of overcoming adversity, illustrations learned, and generally speaking business benefits. The more you report your prosperity during execution, the more information clients will be motivated to utilize the list.
Benefits of Data Catalogue: - [5]
Following are some of the many advantages of data catalogue:

1. Data efficiency: - Since the data catalogue will have proper indexing and information about the location and detailed description of the data, the efficiency to locate and use the data improves.

2. Efficient data Analytics: - The efficiency to perform data analysis increases as the data professionals are able to locate the data easily with the help of data catalogue.

3. Error management: - Since data catalogue will have detailed description about the source of origin of data, information about who has used data, how it is used and when etc, the chances of error and risk is reduced over the period of time by using data catalogue in the business.

4. Wastage of time to find information: - This is one of the fundamental difficulties which is settled by utilizing information list. It is hard for the information experts to find specific information from enormous measure of information. Manual way takes a great deal of time to find the information. In any case, in information inventory since, data in regards to the area of the information is available, it makes it simple for the information experts to find the information quicker than expected.

Conclusion: - How much information and data present in an association is enormous and consistently, it increments. Every one of the associations keep their information and data in a coordinated way utilizing different advancements and devices. How much information put away in data set is more and in a few basic situations, it takes more time than expected to find the information required by the business or association to pursue fast basic choices. To address this issue, one new methodology is being utilized which is Data index. It is the most common way of coordinating the information in such way that, it is not difficult to find the information from the data set in which it is being put away. Information list is the most common way of getting ready ordering of the information and data present in the data set. Information list is the kind of an inventory which will address all the sort of information and data which is available in the association in a coordinated way. It makes it simple for the information experts to gather, accumulate, orchestrate information in coordinated way

References: -
[1]. https://research.aimultiple.com/data-catalog/

[7]. https://research.aimultiple.com/data-catalog/
[8]. https://www.alation.com/blog/what-is-a-data-catalog/