**What is a data catalog?**

In short, a data catalog is a reference application that enables business users, data scientists, BI analysts, data stewards and other workers to explore data sets, understand their contents, and collaborate and share knowledge about data assets with one another. Ideally, it helps them become more self-sufficient at finding and accessing relevant data to use in operational and analytics applications.

A data catalog collects metadata from databases, data warehouses, data lakes, BI systems and other sources and uses it to create a searchable inventory of data assets. It also provides a single point of reference for enterprise metadata management, which it can handle faster and more effectively than older types of metadata management systems.

**Why are data catalogs important?**

Without a data catalog, useful data may be hidden from end users. As organizations collect more and more data, it commonly sprawls across various data stores. If business and analytics users can't find relevant data, business operations and analytics initiatives will be less effective. That's a big problem when organizations are increasingly looking to make data-driven business decisions.

**Key steps to build a data catalog**

* Build a data dictionary

The data dictionary should contain descriptions and mappings of every data table or file and all their metadata entities. It then becomes the basis for pulling the metadata into the data catalog.

* Discover metadata from databases and other data sources

Data catalogs use metadata to identify data tables and files for users. A catalog searches the company's databases and other data repositories and loads the associated metadata into its inventory of data assets. Before an organization begins building a data catalog, metadata sources must be identified and recorded

* Profile the data to provide statistics for users

These profiles are informative summaries that explain the metadata to the users of a data catalog. For example, the profile of a database often includes the number of tables, files, and row counts.

* Identify relationships among data sources

Discover related data across multiple data stores and build that information into the data catalog so users can understand the relationships. For example, a data analyst may need consolidated customer data for an analytics application.

* Capture information on data lineage

Extract, transfer, and load (ETL) tools are used to extract data from source systems, transform and cleanse it, and load it into a target data repository. In building a data catalog, the metadata collected during the ETL process includes data lineage documentation that tracks where data originated, how it flows through systems and other information.

* Organize the catalog for use by data consumers

Data catalogs and business glossaries should be designed for data consumers -- such as business users and data analysts -- as much as for technologists. Again, their structure should be based on the subject area model that you designed earlier in the process. In addition, these tools should be accessible via PCs, tablets, and smartphones.