## Importing datasets

```
> library(readr)
> Telephone <- read_csv("~/Downloads/ETL_Cyrus
Lentin/datasets-etl/LeadsTelephone.csv")
Parsed with column specification:
cols(
  Sr = col_integer(),
  `Company Name` = col_character(),
  `Contact Name` = col_character(),
  Gender = col_character(),
  Age = col_integer(),
  ProductCategory = col_character()
)
> View(Telephone)
```

column 1: numeric with range 1 - 5

| | | | Contact Name | Gender | Age | ProductCategory |
|---|---|---|---|---|---|---|
| 1 | 1 | ABC Bros | Ravi Raj | Male | 18 | Web Site Hosting |
| 2 | 2 | XYZ Sons | Shyam Patel | Male | 30 | Domain Name |
| 3 | 3 | Chunilal Associates | Sunita Sharma | Female | 43 | Email Service |
| 4 | 4 | Ramlal LLP | Pankaj Gupta | Male | 55 | Web Site Hosting |
| 5 | 5 | Shaan Pvt Ltd | Shaan Reddy | Male | 67 | Domain Name |

```
> library(readr)
> WebChat <- read_csv("~/Downloads/ETL_Cyrus
Lentin/datasets-etl/LeadsWebChat.csv")
Parsed with column specification:
cols(
  Sr = col_integer(),
  `Company Name` = col_character(),
  `Contact Name` = col_character(),
  Gender = col_character(),
  Age = col_integer(),
  ProductCategory = col_character()
)
> View(WebChat)
```

| | Sr | Company Name | Contact Name | Gender | Age | ProductCategory |
|---|---|---|---|---|---|---|
| 1 | 1 | Sony Mony | Ravina Swamy | F | 19 | Web Hosting |
| 2 | 2 | Star Traders | Ria Patel | F | 29 | Domain Registration |
| 3 | 3 | Channel Fashions Ltd | Sunita Sharma | M | 49 | Email Service |
| 4 | 4 | Sunderlal & Sons | Pankaja Munde | F | 59 | Web Hosting |
| 5 | 5 | Khanna & Associates | Rajesh Khanna | F | 68 | Domain Registration |

```
> library(readr)
> WebForm <- read_csv("~/Downloads/ETL_Cyrus
Lentin/datasets-etl/LeadsWebForm.csv")
Parsed with column specification:
cols(
  Sr = col_integer(),
  `Company Name` = col_character(),
  `First Name` = col_character(),
  `Last Name` = col_character(),
  Gender = col_integer(),
  Age = col_integer(),
  ProductCategory = col_character()
)
> View(WebForm)
```

| | Sr | Company Name | First Name | Last Name | Gender | Age | ProductCategory |
|---|---|---|---|---|---|---|---|
| 1 | 1 | Star TV | Narendra | Gandhi | 1 | 69 | Web Hosting |
| 2 | 2 | Tata Sons Trust | Bawi | Tata | 0 | 32 | Domain Name |
| 3 | 3 | Tamil Sarkar LLP | Jaya | Lalita | 0 | 49 | Email Hosting |
| 4 | 4 | Banerjee & Chatterjee | Mamta | Gosh | 0 | 59 | Web Hosting |
| 5 | 5 | Reliance | Aakash | Ambani | 1 | 21 | All Services |

## Merging columns and storing in a new column

```
> WebForm$CustomerName <- paste(WebForm$`First Name`,
WebForm$`Last Name`, sep = " ")
> View(WebForm)
```

| | Sr | Company Name | First Name | Last Name | Gender | Age | ProductCategory | CustomerName |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Star TV | Narendra | Gandhi | 1 | 69 | Web Hosting | Narendra Gandhi |
| 2 | 2 | Tata Sons Trust | Bawi | Tata | 0 | 32 | Domain Name | Bawi Tata |
| 3 | 3 | Tamil Sarkar LLP | Jaya | Lalita | 0 | 49 | Email Hosting | Jaya Lalita |
| 4 | 4 | Banerjee & Chatterjee | Mamta | Gosh | 0 | 59 | Web Hosting | Mamta Gosh |
| 5 | 5 | Reliance | Aakash | Ambani | 1 | 21 | All Services | Aakash Ambani |

```
> WebFormResult <- data.frame(WebForm$Sr,
WebForm$`Company Name`, WebForm$CustomerName,
WebForm$Gender, WebForm$Age, WebForm$ProductCategory)
> View(WebFormResult)
```

| | WebForm.Sr | WebForm..Company.Name. | WebForm.CustomerName | WebForm.Gender | WebForm.Age | WebForm.ProductCategory |
|---|---|---|---|---|---|---|
| 1 | 1 | Star TV | Narendra Gandhi | 1 | 69 | Web Hosting |
| 2 | 2 | Tata Sons Trust | Bawi Tata | 0 | 32 | Domain Name |
| 3 | 3 | Tamil Sarkar LLP | Jaya Lalita | 0 | 49 | Email Hosting |
| 4 | 4 | Banerjee & Chatterjee | Mamta Gosh | 0 | 59 | Web Hosting |
| 5 | 5 | Reliance | Aakash Ambani | 1 | 21 | All Services |

## Metadata:

**Definitions –** There are numerous definitions for metadata. The well known definition for metadata is “Data about data.” The other definitions are as follows:

1. Table of content for the data
2. Catalog for the data
3. Data Warehouse Roadmap
4. Data Warehouse Directory
5. Tongs to handle the data
6. The Nerve Center

Metadata in a data warehouse contains all the answers to the questions about the data in the same data warehouse. Metadata is also information about how the data is used. To understand this definition, consider the following example: "$100.00" is a piece of data. It could be payroll data, personnel data, inventory data, or budget data. Under the expanded definition,

- Metadata is information that provides meaning and context to the piece of data. It tells us that "$100.00" is a monetary amount in U.S. dollars, expressed in terms of dollars and cents.
- Metadata also tells us how to understand the way the data is expressed or represented. Metadata helps us to understand the data.

It keeps the answers of this data in a place called Data Repository. It is useful for using, building and administering your data warehouse. Today, data warehouses are much larger in size, wide in scope. Hence, user critically needs metadata to access and use the data warehouse. Without metadata support, user cannot get information every time they need. The extraction team needs to know the structure and data content in the data warehouse to build it. They must have a proper mapping and data transformation technique assigned. They must prepare a logical structure of data warehouse database. Due to voluminous data and large number of queries, it becomes difficult for the user to access the data they need. Therefore, these issues are to be addressed by metadata. Therefore there is a need to administer the data warehouse appropriately at equal interval of time.

Metadata is essential component of IT. It is essential in each and every process, starting from data extraction till the information delivery. It is essential for addressing various technical issues. Development and deployment of data warehouse is a joint effort between users and IT staff.

**Categories of metadata:**

1. Business Metadata: It has the data ownership information, business definition, and changing policies. It describes the physical nature of the data, how the data was created, and how it is managed. This type of metadata is often machine-readable. Borrowing from the previous example, the fact that $100.00 is a monetary value and how it is expressed is physical data. Other examples of physical metadata might answer questions such as
    - What is the origin of the data? Does it come from an external source, or is it generated internally?
    - Where does the data reside? Is it in a SAS table or some other structure?
    - On which server is the structure stored?
2. Technical Metadata: It includes database system names, table and column names and sizes, data types and allowed values. Technical metadata also includes structural information such as primary and foreign key attributes and indices. It describes business rules and definitions on which the data is based. This type of metadata is often intended for people rather than machines. It is informational metadata that tells us whether the $100.00 value is payroll data, personnel data, inventory data, or budget data. Informational metadata would also answer questions such as
    - Who is responsible for the accuracy of the data? How can I contact him or her?
    - What business process produced this data? How do I execute the business process?
    - Which applications should (and do) have access to this data?
3. Operational Metadata: It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.

**Metadata Management:**

If metadata helps us to understand data, metadata management enables us to use the metadata. Metadata creation is time-consuming and expensive. To be truly useful, once stored, metadata must be centrally available and easy to maintain. The primary goals of metadata management are to promote metadata conformity to enable sharing of metadata by an organization's applications. Metadata that is defined for one application can be copied and easily adapted for use by another application. to provide a common, centralized method of searching and managing distinct collections of metadata

**Challenges for metadata management:**

1. Metadata in a big organization is scattered across the organization. This metadata is spread in spreadsheets, databases, and applications.
2. Metadata could be present in text files or multimedia files. To use this data for information management solutions, it has to be correctly defined.
3. There are no industry-wide accepted standards. Data management solution vendors have narrow focus.
4. There are no easy and accepted methods of passing metadata.

Metadata acts as a nerve center. Metadata is placed in a key position and enables communication among various processes. The fig. below shows the role of metadata in a data warehouse.

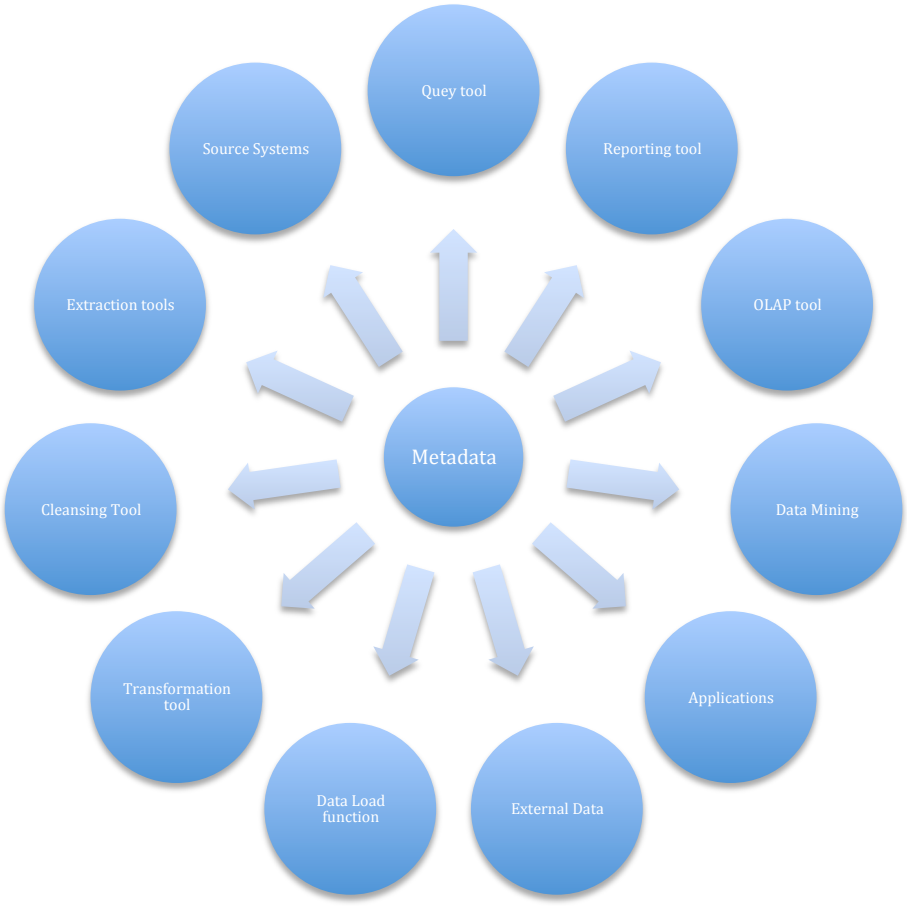


Fig. Metadata as a nerve center