Data warehousing is a critical component of any research or data organization; it consists of data storage, a database engine, and a way to organize database assets. Properly implemented, a data warehouse will reduce infrastructure/staffing costs, reduce data bottlenecks, and increase research synthesis. Therefore, academic organizations need to keep abreast of industry trends and best practices to get the most they can out of their data assets.

Here we discuss a few modern data warehousing solutions that we will/have implemented at our organization. Firstly, data is stored in a columnar format such as .parquet which has been optimized for efficient storage and rapid analysis of large amounts of data. Secondly, the intentional utilization of a multilingual and lightweight database engine (e.g. Apache Arrow or DuckDB) that is designed for analytical tasks rather than transactional purposes. Lastly, we orchestrate the data storage and analytics within the context of best-practices frameworks such as Targets (R framework from pharma) or DBT (SQL framework from the health analytics sector). Taken together these tools compose a modern data stack that will not only enable us to effectively utilize big data to rapidly deliver on proposed research goals but also synergize with the staffing and costs parameters within our organization.