**Author:** Ken Trinh

**Title:**  Data Relationship and Hidden Value

**Target Audience:** Enterprise Data & Analytics Directors

In the modern world, data is the new gold; Countless of business are creating new opportunities with the help of data science, analytics, machine learning, and Ai. While business operations generate a lot of data over time, the values of these data are still hidden in the historical database. And depending on the business, the data living in the historian is quite dirty or even hard to obtain.

Many data science projects rely heavily on the readiness of data before any value can be extracted. In some industry, the data scientists' time is mostly spent on collecting the necessary data and cleaning it to the appropriate level of readiness. Time spent on data preparatory works can take away time from actual data science workload, where relationships among data are determined, models are built and tested, and complete pipelines are developed. Combining this with the complexity of the project, the expected value obtained would significantly be reduced.

There are two ways to tackle this problem. One, a tool to reduce the time needed for data scientists to gather and clean data. Two, a tool to simplify the complexity of a data science project. Of course, a combination of both is ideal since reducing the time of development and complexity of a data science project means the data science team can approach the same problem through many angles, thereby increasing the value of their deliverables.

A primary goal of this research will focus on any variation of data extracted from traditional relational/non-relational databases, data warehouse, and data lake to determine similarities among the different querying methods and their complexity. A secondary goal is to develop a tool that enhances the data scientist’s ability to gather the necessary data and map the appropriate relationship prior to their work. Additional findings will enhance the seamlessness of the proposed product.