The Power and Limits of Relational Technology In the Age of Information Ecosystems

Michael L. Brodie and Jason T.Liu

For over three decades relational technology has been the most efficient data management solution on the planet and the data management bedrock of business information processing. Due to increasingly sophisticated and unavoidable data modelling and integration requirements of today's Information Ecosystems that exceed the modelling power of the relational data model, relational technology is less than optimal. Data modelers have known this for decades yet continue to design and integrate databases under assumptions that underlie the relational data model. Making the relational assumptions explicit assists in resolving inherent data integration challenges and can lead to more efficient design, development, and execution of relational data modelling and integration solutions.

Relational databases are an optimal fit for semantically homogenous worlds that typically contain a small number of databases and in which all views of the same entity are consistent. Homogeneity is key to the optimal data modelling and integration solutions that depend on the relational data model and its assumptions that have guided the data modelling and integration world for decades. The relational world assumes that data naturally fits in tables; tables are easily and uniquely identified by a relational key, and that all views of the same tuple are consistent which leads the underlying belief in a single version of truth and the concept of a global schema. The phrase "single version of truth" seems intuitively correct and may provide assurance in a confusing world but is almost entirely false in the real world. Data management vendors promote the "single version of truth" assumption as a highly desirable objective and something that they products can provide. Not only is the claim false, but also the objectives are unrealistic. The basic assumption of the relational world is not just semantic homogeneity but also ontological homogeneity while in reality semantic heterogeneity dominates. As a result, most data modelers and data integrators work in a relational world and adopt the relational assumptions that are not always true.

The dramatic success of relational technology has propelled data modelling and management requirements beyond the modelling and processing capabilities of the relational technology. Our Digital Universe is no longer a semantically homogeneous set of a few databases but Information Ecosystems of 100s or 1,000s of semantically heterogeneous databases to be managed and integrated collectively. As a result relational data integration solutions required for apparently inconsistent data descriptions of the same entity are developed manually with much effort and little guidance. Over a decade of experience with these issues arising in extremely large-scale applications reveals that *the chief problems arise from the conflict between the inherent semantic heterogeneity of data to be integrated and the inherent semantic homogeneity of relational modelling tools*. As semantic heterogeneity increases so do the consequent data integration challenges that decrease the efficiency and increase the cost of designing, developing, and executing relational integration solutions.

This talk focuses on inherent data integration challenges that arise from inconsistencies between relational data descriptions of the same entity that arise from differences in their corresponding ontologies. We investigate the power and limitations of the relational data model to model and execute the resulting relational data integration solutions and the logical fallacies of the underlying relational assumptions. We close by providing a simple conceptual basis for identifying and resolving inherent and complex data integration challenges. The conceptual basis, called the Shadow Approach, arises from the truths underlying Plato's Cave.