Current tools that facilitate the extract-transform-load (ETL) process focus on ETL workflow, not on generating meaningful semantic relationships to integrate data from multiple, heterogeneous sources. A proposed semantic ETL framework applies semantics to various data fields and so allows richer data integration.

Big data, broadly defined, comprises information available from billions—even trillions—of records generated by (and about) millions of people and stored in myriad sources throughout the cyber universe. The insights revealed by big data are transforming science, industry, and ultimately society itself by fostering more effective business practices and government policies, and by enabling a broad array of applications in engineering, biomedicine, energy, environmental monitoring, genomics, and transportation, to name only a few.

Yet because this data is typically loosely structured and often incomplete, much of it is essentially inaccessible to users. We need technology and tools to find, transform, analyze, and visualize data in order to make it consumable for effective decision making.\(^1\) The research community also agrees that engineering big data meaningfully is crucial.\(^2\) Still, data integration within a complex, schema-less world of heterogeneous databases remains a significant challenge.

**THE BIG DATA ENVIRONMENT**

Much big data research centers on the three V’s:

- **Volume**—finding ways to store the huge amounts of data constantly streaming in from social media,