KIMBALL GROUP KIMBALL UNIVERSITY

www.kimballgroup.com Number 141, December 6, 2011

Design Tip #141 Expanding Boundaries of the Data Warehouse

By Ralph Kimball

There is never a dull moment in the data warehouse world. In the past decade, we have seen operational data come thundering in, then an enormous growth of interest in customer behavior tracking, and in the last two years Big Data. At the same time, there has been a steady stream of software and hardware changes impacting what we have to think about. We have big shifts in RDBMS architectures that include massively parallel processing, columnar store databases, in-memory databases, and database appliances. Data virtualization threatens to change where the data warehouse actually resides physically and where the processing steps occur. Big Data, in particular, has ushered in a whole competing paradigm to traditional RDBMSs named MapReduce and Hadoop, as well as data formats outside of the traditional comfort zone of relational tables. And let's not forget our increased governance responsibilities including compliance, security, privacy, and records retention. Whew! No wonder we get paid so much. Just kidding...

It is fair to ask at this juncture what part of all this IT activity is "data warehouse?"

Whenever I try to answer this question I go back to the data warehouse mission statement, which can be said in four words: Publish The Right Data. "Publish" means to present the data assets of the organization in the most effective possible way. Such a presentation must be understandable, compelling, attractively presented, and immediately accessible. Think of a high quality conventional publication. "Right Data" means those data assets that most effectively inform decision makers for all types of decisions ranging from real-time tactical to long term strategic.

Taking the mission statement seriously means that the data warehouse must encompass all the components necessary to publish the right data. Yes, this is an expansive view! At the same time, the data warehouse actually has well defined boundaries. The data warehouse is NOT responsible for original data generation, or defining security or compliance policies, or building storage infrastructure, or building enterprise service oriented architecture (SOA) infrastructure, or implementing the enterprise message bus architecture, or figuring out software-as-a-service (SAAS) applications, or committing all of IT to the cloud, or building the enterprise master data management (MDM) system. Does that make you feel better?

All of the above mentioned exclusions (shall we say headaches?) are necessary parts of the IT ecosystem that the data warehouse absolutely needs and uses. We data warehousers need to focus on the key pieces that enable us to publish the right data. We must own and control the extraction interfaces to all of the data needed to fulfill our mission. That means considerable influence over the source systems, both internal and external, that provide us with our data. We must own and control the data virtualization specs, even if they sit right on top of operational systems. We must own and control everything that makes up the "platform" for BI, including all final presentation schemas, user views, and OLAP cubes. And finally it must be clear to management that the new pockets of Big Data analytic modelers sprouting up in end user departments need to participate in the data warehouse mission. The new Big Data tools including MapReduce, Hadoop, Pig, Hive, HBase, and Cassandra are absolutely part of the data warehouse sphere of influence.

From time to time, vendors try to invalidate tried and true approaches so that they can position themselves as doing something new and different. Don't let them get away with that! Data warehousing has an enormous and durable legacy. Keep reminding IT management and senior business management of the natural and expected expanding boundaries of the data warehouse.