Data warehouses are dead, long live data warehousing!

Mehul A. Shah, Amazon Web Services

A couple decades ago, an enterprise data warehouse was the information hub that decision making revolved around for a business. Data warehousing was a large, fast growing market with vendors in the eco-system orbiting around platforms like Teradata and Oracle helping to ingest and curate data (ETL) and leverage the engines to help make business decisions (BI). While data warehouses remain an important primitive for analysis, they no longer hold this central position in an enterprise.

There are good reasons for this. First, the kinds of data people want to analyze are too varied and poorly structured to fit into data warehouses, e.g. IoT event streams, ad-tech logs, and social feeds. Second, the kinds of analyses have also broadened, for example, streaming queries for monitoring and alerts, graph analytics for fraud detection, and machine learning for … well … almost everything these days. Finally, with the cloud, technology and cost ratios have changed dramatically. Public clouds offer raw compute and raw storage that is cheap, ubiquitous, and available on-demand.

While these trends curb the usefulness of data warehouses, the need for warehousing data sets company-wide under a single locus of control still exists. People want to collect, share, and integrate disparate data not only for top-level business decisions, but also for communication, collaboration, and innovation at all levels. So, instead, companies are setting up next-generation, communal information hubs – an evolution of enterprise data warehouses – dubbed "data lakes".

"Data lakes" are used differently from data warehouses in three important ways that present new opportunities and challenges for researchers and system developers.

1. **There is no a centralized curation (ETL) process for loading data.** The intent is to allow organizations and users to publish, share, and use data sets independently. While this self-serve nature implies fewer accessibility bottlenecks, it also encourages a data mess. *These platforms, therefore, need ways to automatically characterize and organize their content*, e.g. distribution analysis, schema and type inference, summarization, entity extraction, semantic classification, and more, *as well as ways to automatically discover relevant datasets.*

2. **Users want to run many different kinds of analytics on the same data.** *These platforms, therefore, need mechanisms to efficiently multiplex and materialize data sets across different types of analytics engines.* This goes beyond simply defining open data formats. For example, companies may mine chat interaction logs for fraud and compliance as well as to improve response times and user experience. *In this case, they need efficient tools that maintain several processed representations of raw data*, e.g. one optimized for graph analytics, one for search, and one for predictive analytics. *As a corollary, analytics engines will need to operate efficiently over external data sets stored in shared storage services like Amazon S3 as well as allow for fast ingest and export.*

3. **Users want their data lakes delivered "serverless".** *The cloud and serverless computing are eating the world, and analytics is not immune.* Companies do not want to maintain teams that provision, configure, and tune hardware and software to deliver an analytics solution, whether virtualized or not. Analytics engines are expected to run on-demand, with low-latency, and instantly scale-down to nothing when not needed. We, therefore, need new techniques for auto-tuning dynamically scaling analytics engines based on workload and capacity. We need new methods for automatic physical design in shared storage services.

In this talk, I will briefly share my experiences with why and how users are leveraging AWS Glue to setup and maintain data lakes and describe the key challenges they face.