Infrastructure for Data valuation and management

Eric Solano, Ph.D.

During my wide-ranging career I have experienced the transitioning in the data management infrastructure from on-site relational databases to cloud-based technologies.

As a data scientist at RTI International, I managed large relational databases for government agencies such as the Environmental Protection Agency (EPA). These databases supported research to inform decision making on new regulations or policies. I conducted data collection for science-based research and analyses with very structured data modeling and architectures and detailed metadata and highly enforced relational integrity. Since regulations can be set at the federal, state or regional levels, I collected and stored geo-referenced data for most projects. Geo-referenced data required relational databases to support the storage of GIS derived data. I experienced the emergence of the support for geospatial data in the 2000's and used ORACLE Spatial to create spatial indexes and queries, and to conduct geospatial analyses. I was involved in complex data modeling to assist scientific analyses, to manage data as a resource and to design geospatial databases.

With the emergence of cloud-based technologies, I was involved with both the migration of legacy relational databases as well as with the design of cloud based relational databases, non-SQL databases and data lakes. One of my most important responsibilities related to design and maintenance of infrastructure for data management is the Telemetry Backbone (TBB) at Continental. The TBB's main purpose is to support Data Scientists with a solid and easily accessible layer to all available and enriched telemetry data. The TBB uses Apache Cassandra as the persistence layer without a single point of failure. The TBB scales linearly and consists of a cluster of nodes interconnected with a ring architecture. Data is currently being ingested via stream (mainly from the ContiConnect Kafka Broker) or via Batch process from different batch engines. I have installed Apache Spark collocated with the Apache Cassandra cluster to provide a very strong analytical layer for distributed data. The Spark SQL and Spark MLlib (machine learning) modules can be used to perform very efficient analysis of telemetry big data.

I also conduct projects with Amazon S3 storage-based architectures. Continental manufacturing data has very high dimensionality and Amazon S3 buckets provide a very efficient way to store and retrieve data in different compressed formats such as parquet. The datasets contain billions of records with many columns. Big data technology was required to analyze these datasets. The Apache Spark platform, the Hadoop Distributed File System (HDFS) and the Apache Mahout framework. An Amazon EMR cluster with 5 nodes was used. Each EMR cluster node is of type 'm5a.xlarge' (4 CPUs, 16 GB RAM).

I was also in charge of developing the Modern Data Analytics Platform (MoDAP) at Continental. MoDAP is a composition of cloud services from Amazon Web Services to enable developers to build and schedule data pipelines, deploy machine learning models and store data from different sources. It abstracts the complexity of using AWS services. MODAP users can author their own ETL jobs using Apache Airflow for job orchestration and scheduling.

Additionally, I have tested and used other data management platforms in multiple projects. These platforms include Denodo, KNIME and Matillion.