Big Data Analytics Continues to Evolve

Mark A. Anawis

 *Size alone does not define big data — it is best defined as a combination of volume, velocity, variety and value*

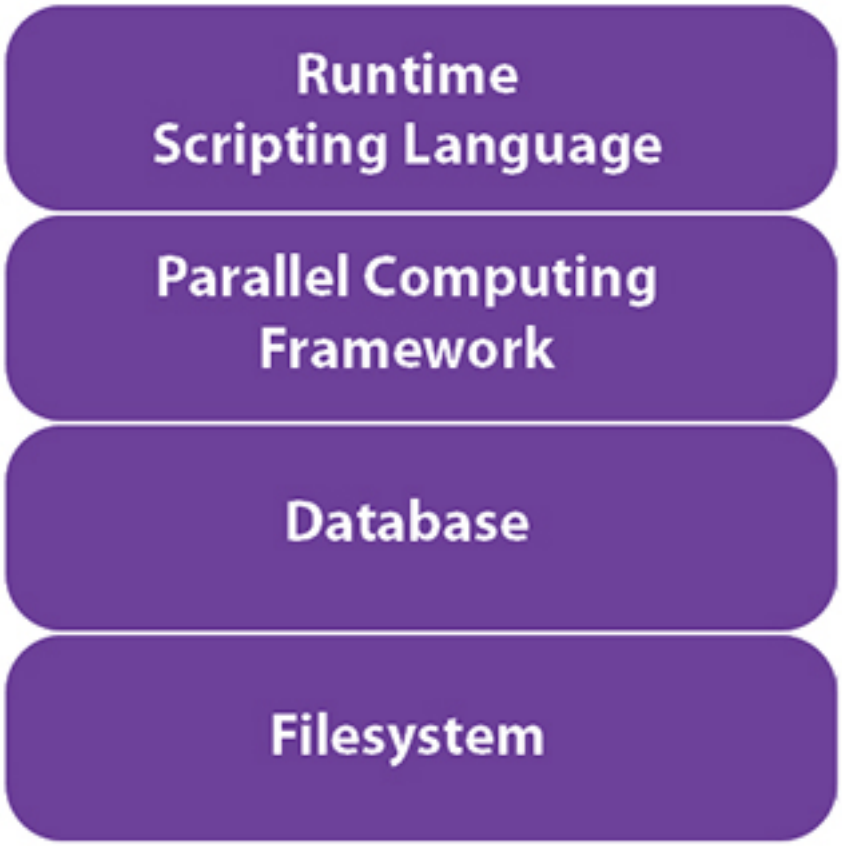
Kevin Geraghty, head of Analytics 360i defined the goal of big data analytics well when he said: “We are trying to listen to what the customer is telling us through their behavior.” The goal of big data analytics is to make the best business decisions possible. Data sources, such as Internet clickstream data, mobile phone data, meteorology

and genomics, accumulate vast amounts of data. The per-capita capacity to store information has been doubling every 40 months for the past 30 years, so that we are at the order of exabytes of data (1018).

Relational	NoSQL
Row/Column structure	HyperTable
Foreign Key	Hypertable Key
Rigid Schema	No Schema
Scale Up	Scale Out
SQL	HiveSQL

However, size alone does not define “big data.” It is best defined as a combination of volume, velocity, variety and value. Sometimes, it is associated with unstructured data, but this is not always the case. More often, the problems facing organizations are the lack of analytical skills and integrating big data with existing data warehouses.

The analytical tools available from data mining can be applied once the underlying architecture is constructed. An important distinction needs to be made between traditional business intelligence and big data analytics. The former relies on descriptive statistics to measure and spot trends while the latter relies on inductive statistics to infer relationships and make predictions.



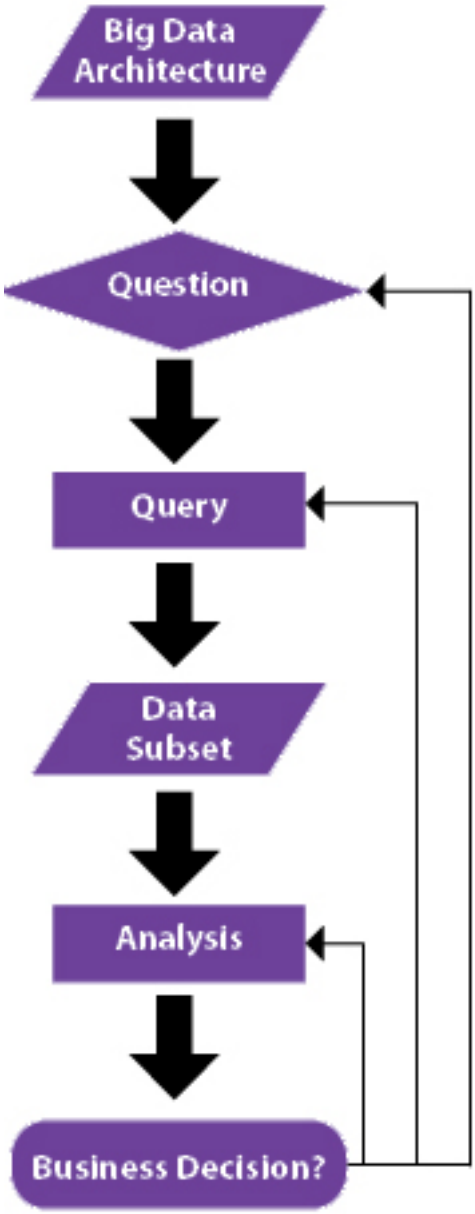
The technologies that are at the heart of the open source software framework are Hadoop/MapReduce, and NoSQL. The data models for the traditional relational databases and NoSQL databases are quite different. A relational database has data in interrelated tables of rows and columns that are referenced through foreign keys stored as columns. NoSQL databases aggregate data into documents using JSON format. Each JSON document is an object. Although its method of aggregation may cause duplication, it allows more flexibility and performance. Another difference is that relational databases have rigid schemas that make changes difficult, whereas NoSQL databases are schema-less, so they allow incorporation of new types of data more easily. Relational databases are expanded by scaling up (adding larger servers), whereas NoSQL databases are expanded by scaling out (servers are added to a cluster) (Figure 1).

The big data architecture (Figure 2) contains a filesystem at the lowest level, which allows creation of files and directories (Hadoop Distributed File System —HDFS — or Google File System). This is very scalable and available due to replication across machines. A hypertable (Google’s BigTable) is the database that creates tables indexed by a primary key. Each row has cells with related information. Each cell contains a row key, a column name (or column family), a column qualifier (a column

instance), and a timestamp.

Within the hypertable system, a distributed file system (DFS) broker handles all filesystem requests. A range server handles reading and writing of data. A master creates and deletes tables, as well as balancing range server loads. The Hyperspace provides a filesystem for metadata.

In conjunction with the hypertable, MapReduce is a parallel computation framework (the algorithm) that processes the data and aggregates it. Hadoop contains a version of this framework. At the top of this architecture is a runtime scripting language (Sawzall, Pig or Hive) that performs statistical analysis. Pig is a procedural language that allows querying of semi-structured data sets using Hadoop. Hive has a simple query language based on SQL that allows summarization, querying and analysis. It is not designed for online transaction processing, but is best used for batch jobs. Complex extract, transform, load (ETL) can be done by either chaining scripts together so that the output of one is the input to another or using a workflow engine like Apache Oozie with actions arranged in directed acyclic graph (DAG) to gate actions. Its definitions are written in hPDL, an XML process definition language. It starts jobs in Hadoop cluster and controls actions through workflows that contain flow and action nodes. Apache Sqoop can be used to transfer data between Hadoop and datastores. It can populate tables in Hive and integrates with Oozie. Apache Flume allows multi-hop, fan-in and fan-out flows and contextual as well as backup routes to provide reliable delivery and manage failures.



Analysis of big data presents its own challenge. Adopt a strategy of breaking up the data into a relevant segment focusing on answering a simple question, and then add data sets where needed, perhaps breaking up the analysis across different teams with complementary analytical skills. Specific analytical tools that can be applied are agent-based modeling, neural networks, factor analysis, cluster analysis and time series analysis.

Agent-based models consist of a system of agents and their relationships. An agent is an autonomous entity that can make its own decisions according to a set of rules. This analysis is applied in complex human systems, such as business and marketing.

Neural networks predict responses from a flexible network of input variables, whereas factor analysis is used to reduce the number of dimensions of the data. Factor analysis is related to principal components where linear combinations of the original variables are created, such that the first component has the most variation, the second component has the next most variation, etcetera.

K-means clustering can be used on large data sets and functions by assigning points to clusters and recalculating cluster points in order to divide data into sets that can be more thoroughly analyzed. Time series analysis also may prove beneficial using auto regressive integrated moving average (ARIMA) or smoothing models with characterization of process disturbances and autocorrelation. The iterative nature of refining the question, query and analysis is represented in Figure 3.

Big Data evolved from the need to accommodate large data sets of varying data types and updating at increasing speed. The key is the development of a suitable architecture and selection of appropriate tools, often from data mining and predictive analytics.

Mark Anawis is a Principal Scientist and ASQ Six Sigma Black Belt at Abbott. He may be reached at editor@SceintificComputing.com.