Asking Any Question of All Of Your Data

“Computers are useless. They can only give answers”—Pablo Picasso

Asking questions is a unique human trait. We answer questions to better understand the world, and as computers become more capable and integrated in our lives we ask them more questions. What is a search engine, if not an oracle where we seek answers?

In contrast to the anarchy of data found on the web, organizations have kept a tight rein on the size and structure of the data they store. For example, an e-commerce site will keep a database of inventory and customer transactions. A generation of experience in building data warehouses means it is well understood how to clean, transform and index data to answer specific pre-determined business questions.

This model has served industry well, but has a significant drawback: if the question you want to ask doesn't fit your representation of the data, then you can't ask it. Of course, you could update the structure of your data warehouse, but the overhead of performing that update means you will only do it for the most pressing questions. Analysts get used to asking the questions that they know can be answered.

There's another problem, concerning the data that *isn't* in the warehouse. When data is loaded into the warehouse, it is cleaned and filtered to include only the data of interest. But how do you know what’s interesting? And if it isn't there, you can't query it. Imagine if these drawbacks were lifted, and analysts could formulate any question they liked of a complete dataset (just like a search engine!) and get an answer back in a reasonable time.

A new style of data management is emerging, where *all* the data available is stored without enforcing a pre-defined data structure, and kept online for *any* question to be asked. This approach has been dubbed *schema on read*, since the data is interpreted at query (read) time, rather than when the data is loaded into the data store (*schema on write*). The advantages are that loading data into the database is cheap, since there is no clean/transform/index step and that any question can be asked of the data. But this only works if the underlying data storage and compute engine is powerful enough to operate on a large dataset in a time-efficient manner.

This is where the open source Apache Hadoop project comes in, with its distributed filesystem (HDFS) and compute engine (MapReduce), both based on systems that Google invented. HDFS provides the raw underlying storage on commodity hardware for massive datasets (which may be petabytes in size), and MapReduce provides the programming model to run computations over significant portions of the dataset in parallel. Clusters running a few thousand machines can process ad hoc queries over multi-terabyte datasets in minutes.

There’s an industry growing around the Hadoop platform, building tools and systems that make Hadoop easy to use and deploy in production. The organizations using Hadoop span many sectors. Companies like eBay are installing Hadoop clusters into which all their data is loaded, and that act as data “sandboxes”. (Disclosure: eBay is a Cloudera customer.) This has allowed them to dramatically improve the buyer's search experience and better match them to sellers, thereby improving sales and profits.

Using Hadoop, Yahoo! is able to do research for advertisement systems, optimize the content that users see, and prevent spam. Or consider Facebook, who use Hadoop for reporting and analytics for understanding user behavior. They were able to test the effect of introducing their ‘Like’ button, for example, by testing it on large subsets of users. Their analysis showed it increased participation compared to those who didn’t have it, so they rolled it out to all users. Or see how, with Hadoop, the Tennessee Valley Authority monitors the behavior of the entire east coast electrical grid to reduce dangerous component failures, prevent brown outs and improve overall service quality.

As data becomes critical to achieving business advantage, the successful companies of the future will be the ones that are able to gain insight from all of their data using tools like Hadoop.

Tom White is a committer on the Apache Hadoop project, an engineer at Cloudera, and the author of “Hadoop: The Definitive Guide” (O'Reilly Media).