The Future of Data Management

David B. Lomet, Microsoft Research

Fifty years from today, we will live in a data-immersive world, doing things we have never done before.

Data management is driven by the ever-shrinking size and cost of a bit of storage, resulting in an increasing volume of data from which we can create value. We’ve grown accustomed to terabytes of data (with occasional petabytes), and exabytes are looming. Lower costs of data acquisition, storage, and management translate to increased data-driven opportunities. In turn, cloud vendors are growing their datacenter infrastructure at an astounding rate. Data needs to always be accessible and available, and must be easy to request. Online retailing with diverse shoppers, products, vendors, and means of delivery is an iconic example of this.

So, what does this mean for data management, and can we project what the world of data will look like in 50 years? As the famed saying goes, making predictions is difficult, especially about the future. Making a 50-year prediction is a fool’s errand, but we can identify some long-term trends and possibilities. Let’s first look back 50 years, both for inspiration and as a sanity check.

In the late 1960s, data management mainly meant business data processing. Online data management arrived around then (with some rounding), enabled by disk drives, and online transaction processing (OLTP) was born. Then, two great abstractions appeared that have served the data-management community well and should be relevant for the next 50 years: (1) transactions and (2) relational data models and algebra.

In a transaction, a user has the illusion that he is acting alone on the data. The result of the actions of all concurrent transactions is as if each transaction were executed serially without concurrency. If each transaction is correct in isolation, the overall result will be correct. Relational algebra transformed a collection of informal data-engineering operations into a mathematically sound closed system working on aggregated data (relations). We could then speak precisely of correct transformations and answers, equivalent but faster executions, and so on. Particularly important were ways to bring data together (joins) that permitted separating physical data formats from logical views and enabled diverse physical formats (column, row, hierarchical, tree, semistructured, and so on) to be treated within this algebra. Looking ahead, these two fundamental abstractions should continue to serve the data-management community well.

However, large portions of data management have needs that go beyond these abstractions. For example, dirty data and representation diversity are huge barriers to coherent access and manipulation of data from multiple sources. Web search solves this problem when human eyes are the target and “good enough” is good enough, but not when the target is another machine and when
logical precision is a requirement. A new (well-paid) field called data science has evolved to cope with this.

Sensors produce astoundingly large dataflows that might need close to real-time processing and data reduction. Machine learning is being applied to ever-larger datasets to exploit the data for profit and for health, science, and the planet. Disasters happen, resulting in a need for data-disaster survival and for real-time, location-based information relevant to emergency responders. Social networks need data management at a global scale that is responsive to users everywhere, consistent, and available, with low latency and easy-to-use interfaces. New levels of well-being are possible when large parts of our environment can exploit data to serve us better—this is part of the promise of the Internet of Things (IoT).

We need to successfully grow the size and number of datacenters, which will transform the data-management landscape. Cloud vendors automate system operations, provide mechanisms for data security, assured data availability, and will shorten latency via edge servers, which will perhaps eventually be built into the Internet-switching infrastructure. Integrating geographically distributed data, applications, and disaster protection might well be enabled by this same infrastructure. However, the speed of light makes some latency unavoidable, and the CAP theorem (consistency, availability, and the resilience to network partitions) suggests that “having it all” will remain a challenge. But never underestimate clever engineering.

A revolution is beginning on how we query data. Computer cognoscenti might be able to cope with formal languages, but most users can’t. For an end user, some queries can frequently be derived from context (for example, restaurants near your current location). But more sophisticated queries will increasingly be expressed at least partially verbally using increasingly conversational natural language. We are starting to see this with Cortana, Siri, and Alexa. The capability of this technology should steadily increase, and its use will eventually become second nature. The challenge is to turn conversational query into formal language query—but this will come.

Fifty years ago, we entered a data-assisted world. Today, the world is data-dependent—we can’t check out at a store if their data systems are down. Fifty years from today, we will live in a data-immersive world, doing things we have never done before via data’s ubiquitous integration into every facet of our lives. This has already begun. Enjoy the ride.

**DAVID B. LOMET** is a principal researcher at Microsoft Research and currently serves as the first vice president and treasurer of the IEEE Computer Society. Contact him at lomet@microsoft.com.

Want to know more about the Internet? This magazine covers all aspects of Internet computing, from programming and standards to security and networking. www.computer.org/internet