The new technologies needed for dealing with big data

February 20 2014, by Paul Mccarthy

MongoDB co-founder and chairman Dwight Merriman still writes code. Credit: TechCrunch/Flickr

While much focus and discussion of the so-called "Big Data revolution" has been on the data itself and the exciting new applications it is enabling—from Google's self-driving cars through to CSIRO and University of Tasmania's better information systems for oyster farmers—less focus has been on the underpinning technologies and the talent driving these technologies.

At the heart of the Big Data movement is a range of next generation database technologies that enable data to be amassed and analysed on a

scale and speed hitherto unseen.

Global online services such as Google, Amazon and Facebook that serve billions of people around the world in real time have been made possible due to new technologies that divide tasks and files across banks of thousands of distributed computers.

Storing the data

Traditional database technologies are built around many tables of information like spreadsheets with rows and columns and a way of asking questions of these tables in a structured way.

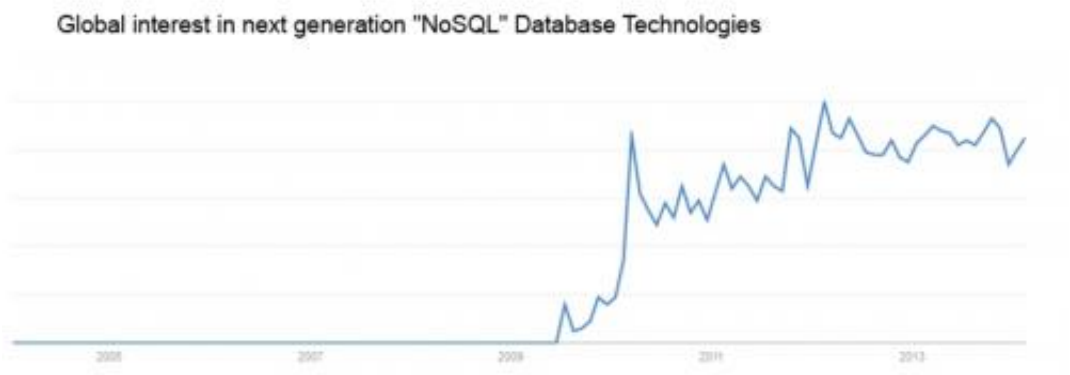
The structured way of asking a question of these data collections was originally named SEQUEL (Structured English Query Language), later shortened to SQL. This is the technology that Oracle pioneered in the 1970s and it has served them well to become the undisputed king of database technology ever since.

If you are familiar with Excel, you'd be familiar with the type of information this kind of technology is suited to representing. Company accounts, marketing and sales figures over time are of course perfect.

But there are other types of data that isn't so easily stored in this way such as storing the relationships in a social network (Facebook), or index of documents stored on the web (Google), or for large collections of digital music and video (Netflix).

Fortunately there are other ways to store information other than in tables such as in trees, graphs, or in lists with an index. And some of these approaches are much better suited for humungous data sets and for data sets that don't naturally fit into a series of tables.

The growing demand to store and analyse very large bodies of information, and information that is not readily suited to storing in tables (unstructured data), has led to a rapid growth in the popularity of these alternative types of database technologies.



Rising Tide. Credit: Google Trends.

Collectively they've become known as NoSQL technologies. Many of the leading technologies in this category are not developed by one company, such as Oracle or Microsoft, but instead are open source - developed by an open network of companies and independent developers and contributors akin to the way Wikipedia or Linux is developed.

Next-generation database technology

There are five key types of next-generation NoSQL data technologies. They are:

1. Document Store—suitable for storing large collections of documents
2. Wide Column Store—for very rapid access to structured or semi

structured data
3. Search Engine—suitable for full text indexing of documents
4. Key-Value Store – suitable for rapid access to unstructured data
5. Graph Database – suitable for storing graph type data such as social networks.

And the leading technologies in each of these categories respectively are:

- MongoDb
- Cassandra
- Solr
- Redis
- Neo4j

Note Apache Hadoop, which is also a leading technology, is not included in this list as it is a framework and file system and not a database technology (but can support many of these).

Where there's talent there's fire

By looking at the companies around the world who have the most employees with skills in each of these these frontier technologies, we can get a unique insight into organisations at the forefront of next generation big data applications.

Where is the the World's Big Data Tech-Talent?
Forty of the World's Leading Enterprises

Technology	Country	Leading Enterprise ranked by specialist talent
MongoDB	Australia & New Zealand	Sensis
MongoDB	Brazil	Globo.com
MongoDB	Canada	The University of British Columbia
MongoDB	France	Amadeus IT Group
MongoDB	India	Infosys
MongoDB	Ukrane	EPAM Systems
MongoDB	United Kingdom	Sky
MongoDB	United States	eBay Inc
Cassandra	Australia & New Zealand	Sirca
Cassandra	Brazil	iG - Publicidade e Conteúdo LTDA
Cassandra	Canada	BlackBerry
Cassandra	France	Orange
Cassandra	India	Cognizant Technology Solutions
Cassandra	Israel	LivePerson
Cassandra	United Kingdom	Sky
Cassandra	United States	Netflix
Solr	Australia & New Zealand	Education Services Australia
Solr	Brazil	Ideais
Solr	Canada	Kobo Inc.
Solr	France	Ekino
Solr	Germany	Adobe
Solr	India	Infosys
Solr	United Kingdom	Ixxus
Solr	United States	Apple
Redis	Australia & New Zealand	iPowow
Redis	Canada	Marvin Canada
Redis	China	Baidu, Inc
Redis	France	Nowapost
Redis	India	MakeMyTrip.com
Redis	Poland	Lunar Logic
Redis	United Kingdom	VisualDNA
Redis	United States	Twitter
Neo4j	Australia & New Zealand	Barnardos Australia
Neo4j	Canada	CGI
Neo4j	France	Capgemini
Neo4j	Germany	Accenture
Neo4j	India	Tata Consultancy Services
Neo4j	Spain	Infojobs
Neo4j	United Kingdom	eSynergy Solutions
Neo4j	United States	Cisco

Data source: Linked In, January 2014; Research by Paul McCarthy.

The table (above) looks at 40 leading global organisations that have the greatest number of specialists in each of the top five next-gen database technologies.

The more detailed country-by-country analysis has revealed some organisations such as Sky in the London, Goldman Sachs in NYC are leaders in the number people they have with skills in these emerging areas.

Source: The Conversation