The Evolution of NoSQL – Part 2

RFG POV: Unlike RDMS databases that are architecturally quite similar, NoSQL databases are not and therefore, the classification is a misnomer. Whereas one could count the number of enterprise-class databases (DBs) on one or two hands, the hierachical and relational-database-solves-all-our-data-management-requirements days are being supplemented by the composite NoSQL genre. NoSQL databases in all their varieties are not going away any time soon and IT executives will need to understand the alternatives and select a minimum set that best meets corporate needs.

Part 1 covered the basic definitions and history of NoSQL. This research report addresses the categories, funding and growth. Three more reports will follow that will cover 21 NoSQL innovators worth exploring.

NoSQL Database Categories

As will be seen in the following section, NoSQL DBs simultaneously defy description and define new categories for NoSQL databases. Indeed, many NoSQL vendors possess capabilities and characteristics associated with more than one category, making it even more difficult for users to differentiate between solutions. A good example is the following taxonomy provided by Cloud Service Provider (CSP) Rackspace, which classifies NoSQL DBs by their data model.

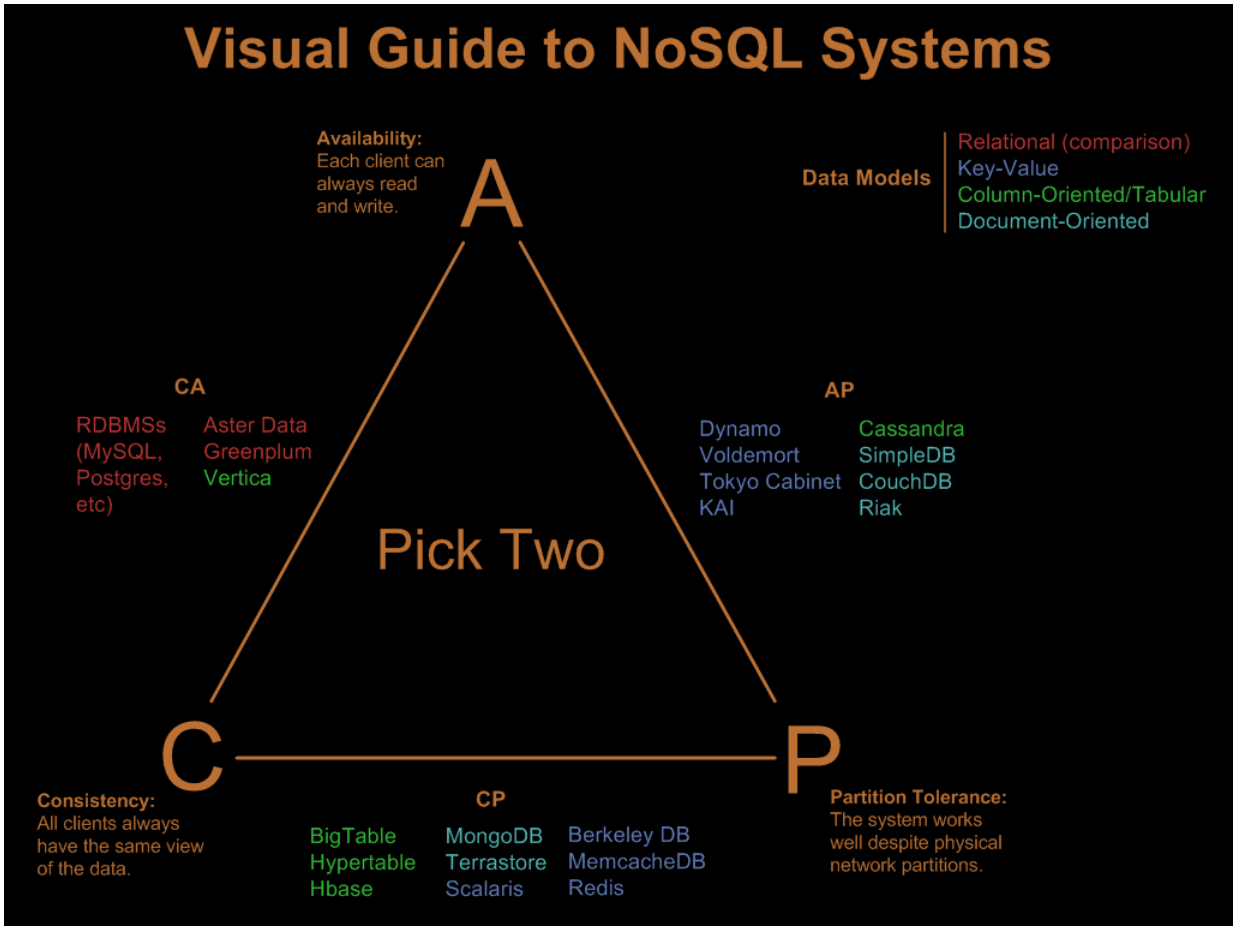
	Data Model	Query API
Cassandra	Columnfamily	Thrift
CouchDB	Document	map/reduce views
HBase	Columnfamily	Thrift, REST
MongoDB	Document	Cursor
Neo4J	Graph	Graph
Redis	Collection	Collection
Riak	Key/value	REST
Scalaris	Key/value	get/put
Tokyo Cabinet	Key/value	get/put
Voldemort	Key/value	get/put

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; *Contact: inquiry@rfgonline.com*

Note: In the original slide, Riak is depicted as a "Document" data model. According to Riak developer Basho, Riak is actually a key-value data model and its query API (application programming interface) is the popular web REST API as well as protocol buffers.

The chart above represents the five major NoSQL data models: Collection, Columnar, Document-oriented, Graph and Key-value. Redis is often referred to as a Column or Key-value DB, and Cassandra is often considered a Collection. According to Technopedia, a Key-Value Pair (KVP) is "an abstract data type that includes a group of key identifiers and a set of associated values. Key-value pairs are frequently used in lookup tables, hash tables and configuration files." Collection implies a way documents can be organized and/or grouped.

Yet another view, courtesy of Beany Blog, describes the database space as follows:



"In addition to CAP configurations, another significant way data management systems vary is by the data model they use: relational, key-value, column-oriented, or document-oriented (there are others, but these are the main ones).

- **Relational** systems are the databases we've been using for a while now. RDBMSs and systems that support ACIDity and joins are considered relational.
- **Key-value** systems basically support get, put, and delete operations based on a primary key.
- **Column-oriented** systems still use tables but have no joins (joins must be handled within the application). Obviously, they store data by column as opposed to traditional row-oriented databases. This makes aggregations much easier.
- **Document-oriented** systems store structured 'documents' such as JSON or XML but have no joins (joins must be handled within the application). It's very easy to map data from object-oriented software to these systems."

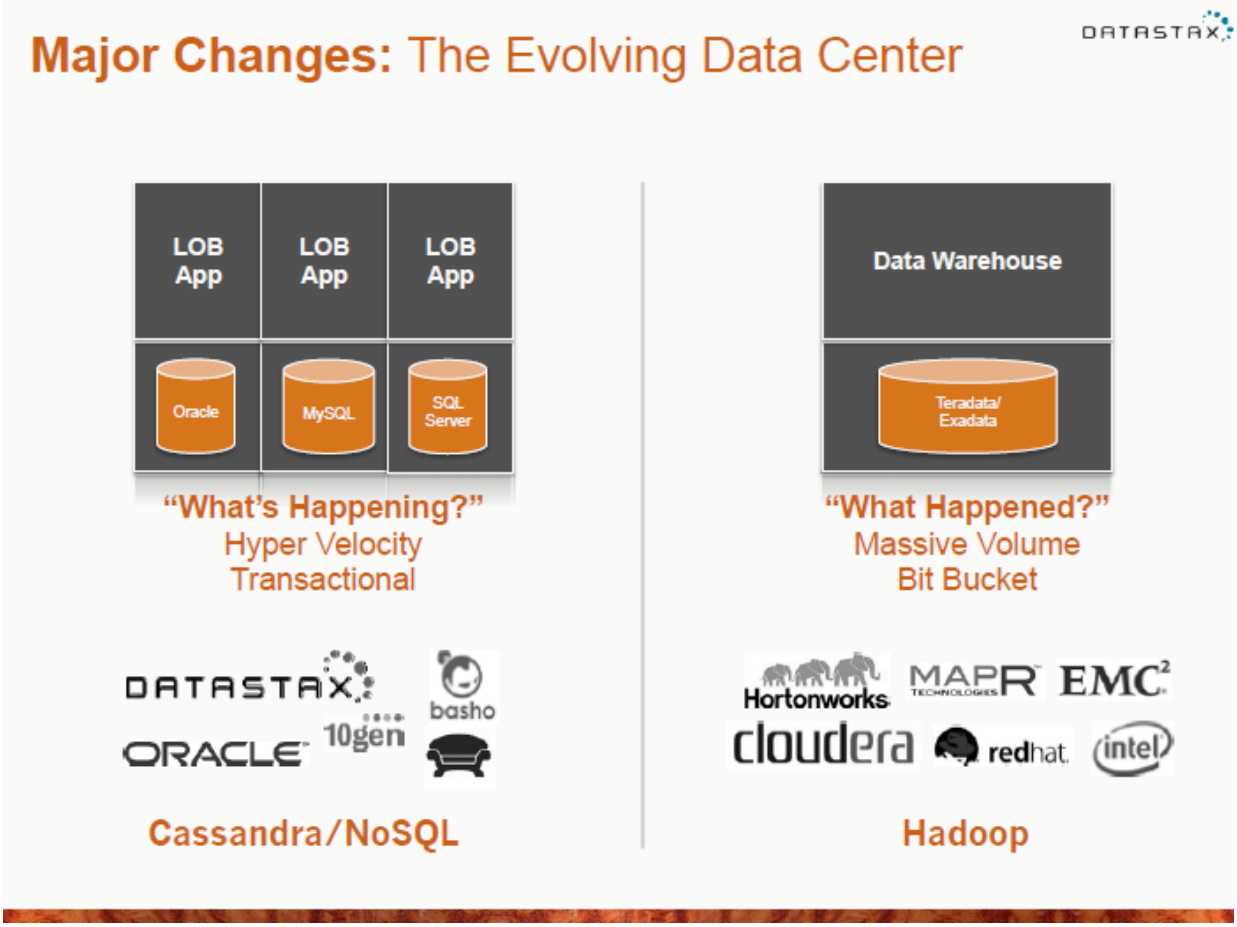
Beany Blog omits the Graph database category, which has a growing number of entrants in the space, including; Franz Inc., Neo4j, Objectivity and YarcData. Graph databases are designed for data whose relations are well represented as a graph – e.g., visual representations of social relationships, road maps or network topologies and representation of "ownership" for documents within an enterprise for legal or ediscovery purposes.

Hadoop and NoSQL

The Hadoop Distributed File System (HDFS) is an Apache open-source platform that enables applications, such as petabyte-scale Big Data analytics projects, to potentially scale across thousands of commodity servers such as Intel standard x86 servers, dividing up the workload.

HDFS includes components derived from Google's MapReduce and Google File System (GFS) papers as well as related open-source projects, including Apache Hive, a data warehouse infrastructure initially developed by Facebook and built on top of Hadoop to provide data summarization, query and analysis support; and Apache HBase and Apache Accumulo, both open-source NoSQL DBs, which, in the parlance of the CAP Theorem, are CP DBs and are modeled after the BigTable DB developed by Google. Facebook purportedly uses HBase to support its data-driven messaging platform while the National Security Agency (NSA) supposedly uses Accumulo for its data cloud and analytics infrastructure.

In addition to the HBase, MarkLogic 7 and Accumulo native integrations of HDFS, several NoSQL DBs can be used in conjunction with HDFS, whether they are open source and community supported or proprietary in nature, including Couchbase, MarkLogic, MongoDB or Oracle's version of NoSQL based on the Berkeley open-source DB. As Hadoop is inherently a batch-oriented paradigm, additional DBs to handle in-memory processing or real-time analysis are needed. Therefore, NoSQL – as well as RDBMS – solution providers have developed connectors for allowing data to be passed between HDFS and their DBs.



The slide above, courtesy of DataStax, illustrates how NoSQL and Hadoop solutions are transforming the way both transactional and analytic data are handled within enterprises with large volumes of data to manage both in real-time, or near real-time, and post-processing or after data is updated or archived.

NoSQL Funding and Growth

A recent note written by Wikibon's Jeff Kelly, *Hadoop-NoSQL Software and Services Market Forecast 2012-2017,* gives a good indication of how well funded and fast growing the market for RDBMS alternatives has become.

"The Hadoop/NoSQL software and services market reached $542 million in 2012 as measured by vendor revenue. This includes revenue from Hadoop and NoSQL pure-play vendors – companies such as Cloudera and MongoDB – as well as Hadoop and NoSQL revenue from larger vendors such as IBM, EMC (now Pivotal) and Amazon Web Services. Wikibon forecasts this market to grow to $3.48 billion in 2017, a 45% CAGR [compound annual growth rate] during this five-year period." Kelly forecasts the NoSQL portion of the market to reach nearly $2 billion by 2017.

Kelly's research also indicates that the top ten companies in the space, measured in amount of funding dollars, received more the $600 million over the last 5 years, with funding increasing dramatically over the last 3 years, including $177 million for 2013 thus far. The top-funded NoSQL DB companies – in order of total funding amount – include DataStax (Cassandra), MongoDB, MarkLogic, MapR, Couchbase, Basho (creator of Riak), Neo Technology (creator of Neo4j) and Aerospike.

Note: On October 4th 2013, MongoDB announced it had secured $150 million in additional funding which would now make it the top-funded company in the space.

Conclusion

Since no one type of NoSQL database neither satisfies all business requirements, innovators and venture capitalists will continue to invest in newer NoSQL iterations and variations. This will just add to the confusion over the next four or five years while all this slowly sorts out. Thus, while the market remains immature and the options are myriad, IT executives cannot wait before selecting the right NoSQL platforms.

RFG POV: The NoSQL wave of database technology is immature and expanding and a myriad of options exist to confound IT executives and slow down decision-making. IT executives and data architects should understand the variety of options and then map them to current and future business and technical requirements for each application type where a NoSQL database might apply. As pointed out in the report, no one solution may meet all the requirements so IT executives should be prepared to act today and adopt and standardize on a minimum set of multiple database solutions.

Additional relevant research is available. Interested readers should contact Client Services to arrange further discussion or interview with Mr. Gary MacFadden, Principal Research Analyst.

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; *Contact: inquiry@rfgonline.com*