MSDS7330

Sandesh /Joe Schueder

Daniel Engels

Topic:

Twitter provides a ﬁre hose of data. Automatically ﬁltering, aggregating, analyzing such data can allow a way to harness the full value of the data, extracting valuable information. The idea of this project is investigating stream processing technology to operate on social streams. Storing Data into a nosql database.

Let’s do a specific topic or do the top trends 10 trends by day for 30 days?

Project title. A detailed project title is better than a vague title. A properly formed title will help to focus your energies on the actual problem being addressed in your project.

• Names and email addresses of all investigators for the project. There must be at least two investigators.

• Clear statement of the problem. A one to two sentence statement of the problem followed by a one paragraph clariﬁcation of the problem. The paragraph should identify clearly the research question you are addressing, and proper motivation for the importance of the problem must be provided.

• Clear statement of your research methodology. A multiple paragraph explanation of how you will approach and develop a solution to the research problem under study. Note that you need to identify multiple intermediate milestones. Research, like product development, is not an all or nothing proposition. If you make it all or nothing, you are likely to lose everything. When you strip the problem to its essence, the ﬁrst step becomes clear. And, that ﬁrst step is usually a baby step. Identify that baby step, and then the next one and the next one and so on. It is possible to get an ‘A’ on the research project without completing every step you identify, and it is certainly easier to get a good grade when all the steps you’ve identiﬁed are baby steps.

• A statement of previous work related to the problem. This is a preliminary inquiry into what research has or has not been performed to solve your chosen problem. Your ﬁnal report will contain an extensive discussion of previous and related research. You should have at least ﬁve citations of previous or related work on your topic.

Relational databases have been in place since the 1970s, while NoSQL movement started in 2009. The reason for NoSQL’s emergence was to cover perceived gaps with relational databases. NoSQL started for usage in web development but has also found a tremendous amount of use in data science and analysis since that time. Industry research shows that they both have their uses and that they will both remain in use for the foreseeable future.

The convergence of the creation the new NoSQL technology, data science and machine learning resurgence created a lot of hype surrounding the benefit that could be achieved with their combination. Of course, such hype produced many followers and several skeptics. Both contingents have produced a large amount of research related to comparison of the benefits and drawback of each type of database. Gartner[gartner1] stated that ‘Nonrelational DBMS does not replace relational DBMS, rather augments it and may be a good choice for new applications.’ As shown in figure 1[gartner1] relational DBMS and document and key value DBMS(types on NoSQL) are both viable options.



Figure 1 Gartner DBMS Life Cycle

The key takeaway that Gartner wants to convey is that both are great options depending on the use case. The table shown in Figure 2[gartner2] shows considerations when choosing between the two higher level database tpes

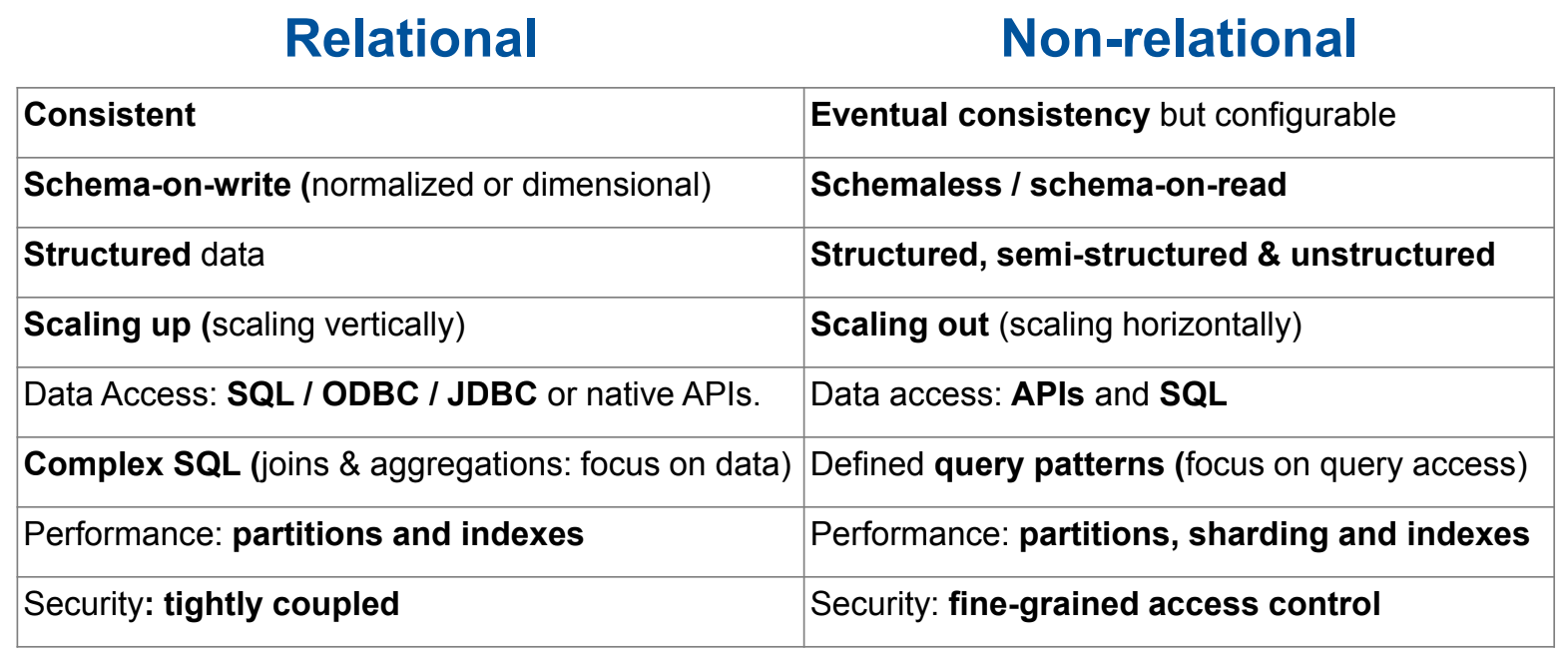


Figure Gartner DBMS Use Cases

The research that has been done has compared performance of read, write, and deletion. The research has compared various relational databases to various NoSQL databases as well as various NoSQL databases to each other.

Most of the database comparisons do comparison of a certain amount of records being written for example one hundred, one thousand, and one million records to a database in one set of transactions. The various comparisons have shown be to be consistent across multiple surveys. In general, it shows that good NoSQL databases are faster than relational databases and ‘bad’ NoSQL databases. An example of a survey presented at The First International Conference on Computational Intelligence and Informatics in 2017, comparing mySQL and Cassandra shows significant differences. “The performance for ‘writes’ is similar in Mysql as in Cassandra, but more time is taken by Mysql showing that Cassandra performs much better ‘writes’ over Mysql”[eval1]

|  |  |  |
| --- | --- | --- |
| Records (no. of inserts) | Cassandra WT (ms) | Mysql WT (ms) |
| 100 | 1 | 5 |
| 200 | 2 | 9 |
| 500 | 4 | 19 |
| 1000 | 8 | 43 |
| 10000 | 60 | 400 |
| 10000 | 456 | 3000 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Database | Number of Operations | | | | | |
| Time in (ms) | 10 | 50 | 100 | 1000 | 10000 | 100000 |
| MongoDB | 8 | 14 | 23 | 138 | 1085 | 10201 |
| RavenDB | 140 | 351 | 539 | 4730 | 47459 | 426505 |
| CouchDB | 23 | 101 | 196 | 1819 | 19508 | 176098 |
| Cassandra | 115 | 230 | 354 | 2385 | 19758 | 228096 |
| Hypertable | 60 | 83 | 103 | 420 | 3427 | 63036 |
| Couchbase | 15 | 22 | 23 | 86 | 811 | 7244 |
| MS SQL | 13 | 23 | 46 | 277 | 1968 | 17214 |

Reads also have had similar results surveyed. This experiment conducted by a group from the University of New Zealand and presented at a conference in 2013 showed: “Sorted by read performance we have the list of databases: Couchbase, MongoDB, SQL Express, Hypertable, CouchDB, Cassandra and RavenDB. Of these Cassandra and Hypertable are column-family databases; and Couchbase, MongoDB, CouchDB, and RavenDB are document-oriented databases. There is no observable correlation between the data model and performance. We also see that the read performance of SQL Express is better than some, but not all, of the NoSQL databases.” [eval2]

In addition, research points out that use cases and design are still important for NoSQL, comparisons show that bad design on the same NoSQL database type can impact performance. Research that was conducted by a group at Universite Grenoble Alpes, LIG France found that schemas may have to be designed for specific scenarios, but generally;” The embedding level of the data has an impact on performances. Accessing data at the first level of a collection is faster and easier than accessing data in deeper levels. … Querying data stored at different embedding levels in a collection may require complex manipulations. … Concerning storage requirements, our experiment revealed that using collections with embedded documents tends to require more storage than using separate ‘flat’ collections and references for the same data.” [schema]

The research gap, that this project will attempt to answer is whether NoSQL or SQL is best for analysis of semi-structured twitter data. This project will also, attempt to expand upon the performance of various database operation for each operation size as the database grows over time.

[schema] - Paola Gómez, Rubby Casallas, Claudia Roncancio. Data schema does matter, even in NoSQL systems! Research Challenges in Information Science (RCIS), 2016 IEEE Tenth International Conference on, Jun 2016, Grenoble, France. ff10.1109/RCIS.2016.7549340ff. ffhal-01482250f

[eval1] Sangeeta Gupta and G. Narsimha. A performance comparison of SQL and NoSQL databases. Springer Science+Business Media Singapore 2017 S.C. Satapathy et al. (eds.), Proceedings of the First International Conference on Computational Intelligence and Informatics, Advances in Intelligent Systems and Computing 507, DOI 10.1007/978-981-10-2471-9\_53

[eval2]Yishan Li and Sathiamoorthy Manoharan Department of Computer Science University of Auckland New Zealand. 978-1-4799-1501-9/13/$31.00 ©2013 IEEE **Published in:**[2013 IEEE Pacific Rim Efficient Query Analysis and Performance Evaluation of the Nosql Data Store for BigDataConference on Communications, Computers and Signal Processing (PACRIM)](https://ieeexplore.ieee.org/xpl/conhome/6602069/proceeding)

[gartner1] Donald Feinberg. The Future of Data Infrastructure in Digital Enterprise. Gartner Data & Analytics Summit. 18 – 21 March 2019 / Orlando, Florida

[gartner2] Sanjeev Mohan. Gartner Zone: Selecting the Optimal Data for Big Datastore Architecture With Gartner’s Technical Assessments. . Gartner Data & Analytics Summit. 18 – 21 March 2019 / Orlando, Florida

[eval3] Gupta, Adity & Tyagi, Swati & Panwar, Nupur & Sachdeva, Shelly & Saxena, Upaang. (2017). NoSQL databases: Critical analysis and comparison. 293-299. 10.1109/IC3TSN.2017.8284494.

• A statement of your research plan and schedule. The timeline and major milestones that you will achieve in completing your project must be explicitly spelled out. You need to timeline your project to convince yourself (and me) that you can complete the project before the end of the semester and that your project isn’t trivial.

• A list of resources needed to accomplish your work, with special emphasis on important pieces you don’t yet have access to. Be as clear and precise as you can in your requirements, and we will work towards getting what you need as quickly as possible. We don’t have a budget, so requests for electron scanning microscopes and new signal generators may not be honored unless they can be borrowed from somewhere else on campus. Even requests for simple things like dedicated computers are unlikely to be met due to the shared nature of the computing environment at SMU. If your request cannot be accommodated for any reason, we will notify you as soon as we ﬁnd out.

• Any other questions or clariﬁcations you need from us.