# Performance Comparison between Amazon DynamoDB and Amazon Relational Database Service

|  |  |  |
| --- | --- | --- |
| 1st Kevin Thompson  *Southern Methodist University*  Dallas, Texas, USA kthompson395@hotmail.com | 2nd Sandesh Ojha  *Southern Methodist University*  Dallas, Texas, USA sandesho@smu.edu | 3rd Joe Schueder  *Southern Methodist University*  Houston, Texas, USA jjschued@gmail.com |

## PROBLEM STATEMENT

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. This project will attempt to determine which is more efficient for Twitter Analysis.

## II. METHODOLOGY

Using python to connect to twitter, we will gather a large volume of tweets with the hashtag #metoo and store them in DynamoDB and in Amazon Relational Database Service (RDS) Databases. We will then compare the performance of each database by timing writes and queries to the databases. We anticipate DynamoDB to have better performance.

We will use the tweepy python library to gather a large volume of tweets and store them into a DynamoDB database and an Amazon RDS database. We plan on having these databases hosted on a free-tier Amazon Web Services (AWS) instance. We hope to obtain 9 gigabytes of tweets for each database. We will use python to time the execution of various reads and writes and then load the data into pandas dataframe for analysis.

## III. RELATED WORK

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 stated that nonrelational DBMS does not replace relational DBMS, rather it may be a good choice for new applications. Mohan (2019) argues that both relational DBMS and document and key value DBMS(types on NoSQL) are both viable options.

Gupta and Narsimha (2017) compare mySQL and Cassandra, showing significant differences in performance. They found that the Cassandra outperformed mySQL when writing data. Gupta, et. al (2017) found that mySQL underperformed against many NoSQL databases like Cassandra, while outperforming NoSQL databases like MongoDB. Though researchers still point out that use cases and design are still important for NoSQL. Gomez, Cassallas, and Roncancio (2016) find that schemas while schemas have to be designed for specific scenarios, in general, the embedding level of the data will have an impact on performances. The research question that this project will attempt to answer is whether Amazon’s relational database service or their DynamoDB service is better for the analysis of semi-structured twitter data.

## IV. SCHEDULE

Given that there is a daily cap on how many tweets we can download every day, we believe that it will take days or weeks for us to reach the 18 gigabyte target. Once we have the necessary tweets, we will immediately begin our comparison and finalize project paper.

## V. RESOURCES

We will only need a free-tier AWS account and three twitter developer accounts for this project.

|  |
| --- |
| REFERENCES   1. P. Gomez, R. Casallas, C. 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 101.1109/RCIS.2016.7549340. hal-01482250 2. S. Gupta and G. Narsimha, “Efficient Query Analysis and Performance Evaluation of the Nosql Data Store for BigData,” in *Proceedings of the First International Conference on Computational Intelligence and Informatics*, S. Satapathy, V. Prasad, B. Rani, S. Udgata, K. Raju, Eds. Singapore: Springer, 2017. 3. L. Yishan and M. Sathiamoorthy, “A performance comparison of SQL and NoSQL databases,” in *2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, Victoria, BC, Canada, 2013. Victoria: IEEE, 2013 pp.15-19. 4. A. Gupta, S. Tyagi, S. Sachdeva, U. Saxena, “NoSQL databases:Critical analysis and comparison,”in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, Gurgaon, India, October 12-24, 2017. Gurgaon: IEEE, 2017 |