Empirical Investigation of Trends in NoSQL-based Big-data Solutions in the Last Decade

Harshit Gujral, Abhinav Sharma, Parmeet Kaur
Department of Computer Science Engineering & IT
Jaypee Institute of Information Technology
Noida, India
harshitgujral12@gmail.com, sharmaabhinav1997@gmail.com, parmeet.kaur@jiit.ac.in

***Abstract*— The usage and popularity of NoSQL databases have sharply risen over the past decade due to their ability to handle a huge amount of data by employing scalable architecture, high availability and better performance than traditional relational database systems (RDBMS). In addition to reporting dynamics in NoSQL-database world, this paper focuses on presenting results from the perspective of developers. Stack Overflow provides a comprehensive technical niche with about 15 million technical questions, 8.1 million users and 25 million answers. In this paper, we aim to study variation in yearly trends of 20 NoSQL databases. To reveal the interest of the programmers we have investigated questions-asked and presented an unbiased Normal Interest Score by employing three parameters, first, the number of questions asked, second, mean views on a question and third, the mean score on a question. MongoDB, Cassandra, Redis, and Neo4j emerged as most popular databases in their respective families while NIS of all four of them is decreasing 2015 onwards. Additionally, we have also discussed how real-world events like publications, open-sourcing, mention in critical bills, version-release, acquiring ventures etc affect the interest corresponding to NoSQL databases over Stack Overflow. Results of this work will help database developers, database administrators for database selection, upgradation and maintenance.**

Keywords*— *NoSQL; Big Data; MongoDB; Cassandra; Database; Stack Overflow

I. INTRODUCTION

The evolution of cloud-computing changed multiple trends in the IT industry. The most prominent change is visible in data-storage solutions. Processing of large distributed data generated by internet every day is not feasible by traditional DBMSs [1]. Data is growing at a rate of 40 percent annually for an average investing firm with a current storage of 3.8 PBytes of data and less than 1000 employees [2]. This has driven the development and adoption of non-relational NoSQL databases, over the last decade. NoSQL databases implement eventual consistency instead of strong consistency. Study of the in-memory performance of various databases proved that even when a RDBMS stores its data in memory, its overall performance is worse than NoSQL databases [3]. Hence, Dynamic scalability is the key of cloud-computing. Authors in [4] employed NoSQL-cloud to implement dynamic, scalable and exhaustive 3D-hash data structure to facilitate designed content search application. Massive scalability, low latency, high availability are predominant requirements of any growing IT firm that traditional RDBMS fails to facilitate in a cost-effective way.

II. RELATED WORKS

Authors in [5, 6] studied a wide structured list of NoSQL databases with the open and closed source. DB-Engine ranking (https://db-engines.com/en/) provides monthly popularity of DBMS with the aid of mentions on search engines; Google trends; frequency of technical discussion over Stack Overflow and DBA Stack Exchange; job offers etc. Study in [7] evaluates the performance of five most popular NoSQL databases Cassandra, HBase, MongoDB, OrientDB, and Redis. Study in [8] compares NoSQL databases by their data models, query possibilities, concurrency controls, partitioning and replication opportunities by presenting a use case oriented survey on NoSQL.

III. METHODOLOGY

Our research objective involves finding trends of NoSQL related questions on Stack-Overflow. Users provide tags to questions to the best of their knowledge and these tags form the basis of classification of questions used in our study. We have considered a question related to our database *X* if at least one of its tags has *X* in it. Our Empirical analysis involves investigation of yearly trends of number of questions asked or α, Mean views per question or β, and Mean question score or γ on Stack Overflow.

A **Normal Interest Score** (NIS) or Δ is observed by calculating the expectation value (or mean) of normalized parameters α, β and γ (unbiased) by

$$\Delta = E[\partial(\alpha, yr) + \partial(\beta, yr) + \partial(\gamma, yr)]$$

where $\Delta \varepsilon [0,1]$, where 1 is most popular and 0 is least popular in terms of interest gathered over Stack Overflow and $\partial(a, yr)$ is min-max normalization operator for year yr for any variable a that is calculated by

$$\partial(a, yr) = (a - a_{min}) / (a_{max} - a_{min})$$
$$\partial(a, yr) \varepsilon [0, 1]$$

where amax and amin are the maximum and minimum value of a in year yr. Code and further details of this analysis is publically available at www.github.com/newtein/nosql.

IV. RESULTS

NoSQL databases can be categorized into four classes [9].We have investigated trends of four-parameters i.e. α, β, γ and Δ on various NoSQL databases divided into four families. We have divided 10-year phase into 3 smaller-phases such that

Phase-I: 2008-2010, **Phase-II**: 2011-2013 and **Phase-III**: 2014-2017 for fine analysis.

A. Column

Phase-I: Few questions were asked in this phase but overall-popularity of Hadoop based Column store, HBase was highest in 2008. The most prominent feature of HBase is that it runs on the top of HDFS [10]. The other column family store, Cassandra was initially released in July 2008. Cassandra's popularity kept on increasing since 2009. Cassandra provides a pragmatic solution to the famous consistency problem i.e. a tunable consistency and this degree of consistency is governed by client-application and can be modified by the client for any read-write operation. Additionally, Cassandra was developed to power inbox search at Facebook, this add-on to the popularity of Casandra [11].

Phase-II: This phase is marked by a significant increase in popularity of Vertica and Accumulo, although a total number of questions asked is significantly lesser than Cassandra and HBase while mean-views per question and mean-score per question of Vertica and Accumulo became credible in this phase. 2011 is marked by a significant increase in Vertica's NIS. Vertica's acquisition by HP in March 2011 [12] and publication by Lamb at al. [13], could be one of the reasons for the interest of programmers in Vertica. The most prominent feature of Accumulo is that it adds a new element called column visibility to the key. This feature finds its application in Cell-level security. In June 2012, Apache Accumulo is referred in a critical draft released by Senate Armed Services Committee (SASC). This draft is a DoD Authorization Bill that demanded DoD to assess if Accumulo could attain commercial viability [14]. Specific criteria for evaluation is not mentioned but this could be attributed to success factor.

Phase-III: Druid project started in 2011 and open-sourced in 2012. Druid is frequently used in production by web-based technology giants namely Alibaba, Netflix etc (http://druid.io/). Although the maximum number of questions in Column NoSQL family concerns Cassandra, it is evident from the NIS that popularity of Cassandra is slighting decreasing from 2015 to 2017 while the number of questions concerning HBase is approximately constant between 2015 and 2017 (see Fig 1).

B. Key-Value

Phase-I: BerkeleyDB and DBM are one of the oldest databases of NoSQL family. The popularity of both of these databases is evident in this phase while it gradually decreases and then become stagnant in later phases. Redis and Riak are released in May 2009 and Aug 2009 respectively. Soon after release, Riak acquired highest NIS in 2009 on the other hand Redis emerged as a late bloomer and acquired highest popularity score in 2010. Redis is a profound example of in-memory databases.

Phase-II: Amazon released Dynamo in 2012, it is evident from Fig. 1 that soon after release, it gathered interests of programmers on Stack Exchange. Dynamo is implemented in order to provide availability and partition tolerance on the cost of consistency [15].

Phase-III: In the year 2014, Aerospike was open-sourced, the following two years witnessed growth in Aerospike related interest on Stack Overflow. An unexpected increase in interest is observed corresponding to MemcacheDB in the year 2016 and 2017 while last stable version of MemcacheDB is released in 2008 and its website was last updated in the year 2009 (http://memcachedb.org). Although, [3] studies that in terms of elapsed time, MemcacheDB provides the best write performance, while Redis is designed in order to use memory more efficiently than NoSQL variants. From 2014 to 2017, the number of questions regarding Redis and Dynamo is increased by 1.43 and 4.31 folds respectively.

C. Document

Phase-I: 2008 is dominated by CouchDB, CouchDB uses B-tree based data-structure. Although, MongoDB is initially released in 2009 but we can find mongo tag in the year 2008 also. The underlying reason is that in order to increase the reach of the question, users answering a question tends to add additional tags to the original question over the time. This question concerns Django Sessions and is available at (https://stackoverflow.com/q/50568/). One of the comment focuses on using MongoDB instead of RDBMS.

Phase-II: In year February 2011, a major BaseX 6.51 (beta) version was released with multiple feature enhancement and bug fixed (http://basex.org/2011/02/03/basex-6.51-beta-released/). This ostensibly seems to be the reason of a boom in interest corresponding to BaseX. It should be noted that while MongoDB has most questions asked in overall-NoSQL database category, its score is equivalent to other NoSQL databases on parameters of mean view count and mean question score. Increase in size of data corresponds to a significant decrease in performance of MongoDB due to the locking mechanism of MongoDB [3]. RethinkDB is open-sourced in 2012 [16], soon a boom of interest is evident from NIS.

Phase-III: Presence of Cosmos DB is evident from the Fig.1, It was initially released in 2010 while in 2014, it builds and extends the previous Azure DocumentDB [17] Hence, interest around Cosmos DB is observed in and after 2014. RethinkDB announced its shut down in the year 2016, rights of its source-code is then purchased by CNCF and contributed to the Linux Foundation [16].

D. Graph

Phase-I: Neo4J was initially released in the year 2007, although its version –I was released in February 2010. Soon, Neo4J started gathering discussion on Stack Overflow. Trivial presence of AllegroGraph is also evident. Although, the Virtuoso project is born in 1998 but its presence is evident after 2010.

Phase-II: Significant increase in a number of questions is observed for Neo4J. A number of questions concerning Neo4j increased over 12 folds, from 178 questions in the year 2011 to 2,190 in the year 2013. Increase in a number of questions is rarely accompanied by an increase in mean views and score. NIS is decreased from 2010 to 2011, gradually increased then become stagnant.

Phase-III: Total number of questions concerning Neo4j researched their highest in the year 2015 and then started decreasing since. A dip in NIS is also observed from 2015 to 2016 while no significant change in trends of Virtuoso and AllegroGraph is observed (see Fig.1).

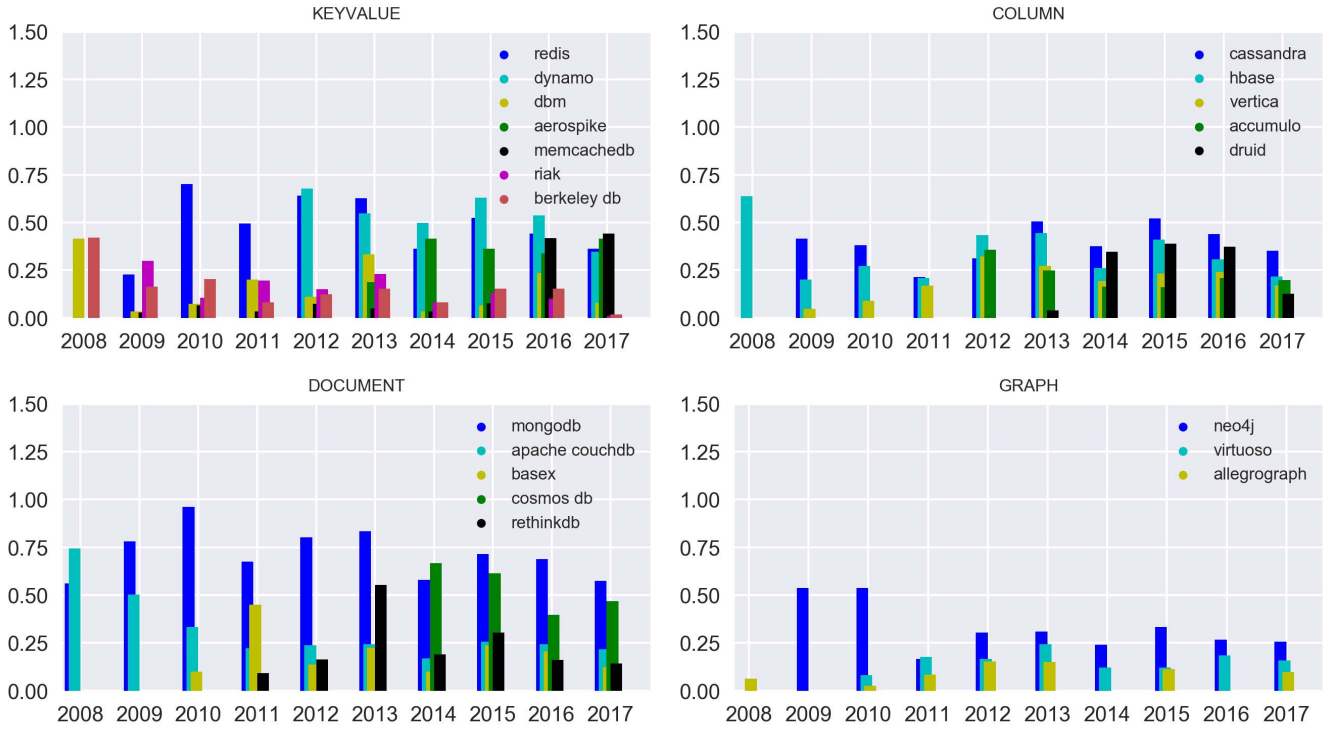

Fig. 1. Normal Interest Score of NoSQL-families between 2008-2017

V. CONCLUSION

NIS depicts that decade started with a boom in interest around NoSQL-based solutions but 2015 onwards interest seems to gradually decrease. We have witnessed that events like acquisitions or transfers (by HP Vertica-2011, by CNCF/Linux Foundation RethinkDB-2016-17), publication of paper (Vertica-2012), mention in US DoD Authorization Bill (Accumulo-2012), open-sourcing (Aerospike-2012, RethinkDB-2012) and version-release (BaseX-2011) ostensibly contributes to gather interests of programmers corresponding to databases on Stack Overflow. Interest corresponding to old NoSQL solutions like DBM and BerkeleyDB is stagnant over the decade while MemcacheDB stopped development of stable-releases since 2008 but it still accumulates interests of programmers over Stack Overflow (best write performance in terms of elapsed time). Cassandra, MongoDB, Redis, and Neo4j are most popular databases of their respective families. Additionally, an interesting point to outline is that NIS of these 4 databases is decreasing from the year 2015. There can be multiple underlying reasons for the cause, eg. the evolution of NewSQL, Spark or HDFS-based solutions.

REFERENCES

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58.

[2] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

[3] Kabakus, A. T., & Kara, R. (2017). A performance evaluation of in-memory databases. Journal of King Saud University - Computer and Information Sciences, 29(4), 520–525.

[4] Gujral, H., Sharma, A., & Mittal, S. (2017, August). No-escape search: Design and implementation ofcloud based directory content search. In Contemporary Computing (IC3), 2017 Tenth International Conference on (pp. 1-6). IEEE.

[5] Cattell, R. (2011). Scalable SQL and NoSQL data stores. Acm Sigmod Record, 39(4), 12-27.

[6] Intersimone, D. (2010). The end of SQL and relational database?(Part 2 of 3). Computerworld, 10..

[7] Abramova, V., Bernardino, J., & Furtado, P. (2014). Which NoSQL database? a performance overview. Open Journal of Databases (OJDB).

[8] Hecht, R., Jablonski, S. (2011). NoSQL evaluation: a use case oriented survey. In: Proceedings – 2011 International Conference on Cloud and Service Computing, CSC 2011, pp. 336–341.

[9] Indrawan-Santiago, M. (2012). Database research: are we at a crossroad? Reflection on NoSQL. In: Proceedings of the 2012 15th International Conference on Network-Based Information Systems, NBIS 2012, Melbourne, Australia, pp. 45–51.

[10] Pokorny, J. (2013). NoSQL databases: a step to database scalability in web environment. International Journal of Web Information Systems.

[11] Hamilton, James (July 12, 2008). "Facebook Releases Cassandra as Open Source". http://perspectives.mvdirona.com/2008/07/12/FacebookReleasesCassandraAsOpenSource.aspx

[12] HP News Release (2011). "HP Completes Acquisition of Vertica Systems, Inc." http://www8.hp.com/us/en/hp-news/article_detail.html?compURI=tcm:245-907883&pageTitle=HP%20Completes%20Acquisition%20of%20Vertica%20Systems,%20Inc.

[13] Lamb, A., Fuller, M., Varadarajan, R., Tran, N., Vandiver, B., Doshi, L., & Bear, C. (2012). The vertica analytic database: C-store 7 years later. Proceedings of the VLDB Endowment, 5(12), 1790-1801.

[14] Metz, Cade. (2012) NSA Mimics Google|. https://www.wired.com/wiredenterprise/2012/07/nsa-accumulo-google-bigtable/

[15] DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., ... & Vogels, W. (2007, October). Dynamo: amazon's highly available key-value store.

[16] RethinkDB Team (2012). RethinkDB is out: an open-source distributed database. https://www.rethinkdb.com/blog/

[17] CrawCour, Ryan (21 August 2014). "Introducing Azure DocumentDB – Microsoft's fully managed NoSQL document database service". https://blogs.msdn.microsoft.com/documentdb/2014/08/21/introducing-azure-documentdb-microsofts-fully-managed-nosql-document-database-service