University of Wales Trinity Saint David

Web Development

ACCA7004

Elliot Short - 1702797

Write a report that considers the main principles of Big Data and NoSQL. Choose a NoSQL database, evaluate its use and provide scenarios which would benefit from its inclusion.

Large amounts of data are created each day by individuals, let alone groups. This data ranges from social media content, profiles built by companies to gauge a users personalities, Google searches etc. This large amount of data being generated every minute requires different practices than traditional computing to handle it, the field around this is called big data[1]. When handling this amount of data traditional data handling methods aren't often used by companies like Amazon. Relational databases built using SQL are often slow when handling large amounts of data, leading to slow response times. While this wasn't as big a problem over a decade ago when companies could rely on upgraded hardware to handle the increased workload, the amount of data being uploaded has increased almost exponentially. An example of this is YouTube, in 2013 100 hours were uploaded a minute to YouTube, last year this figure was 500[2][3]. Relational databases adhere to a system called ACID (Atomicity, Consistency, Isolation and Durability). These are the main principles that govern how relational databases function and because of this it makes scaling out difficult. One reason for this is from the atomicity and isolation principles. Amazon will have many servers for data handling, so if a request is made to these servers and the data requested is stored across multiple different physical machines, because of atomicity requiring a transactions and isolation requiring the database doesn't change more than the parameters of a transaction, having to spend extra time gathering data from multiple physical machines doesn't only slow down that one transaction but subsequent ones after it[4]. The way that NoSQL solves a lot of these issues is by ditching the traditional ACID and using a system called BASE (Basically Available, Soft state and Eventually consistent). By giving up ACID's model BASE is able to scale better to larger systems.

An example of an organisation that felt that the vast amount of data it collected required a switch to NoSQL is the DVLA. The DVLA switched to using MongoDB in 2015 because of the vast amount of data they handle. A big reason for the switch stated in a blog by MongoDB is that the DVLA wanted to make driver licence data available online for people to check on their own data[5]. In England alone from 2015 to 2019 an average of 74% of people had a full driving licence, this is over 35 million people in England alone. When handling this amount of data and allowing the public to query the database that holds this data it's important that the database

works quickly and efficiently. Using ACID in this situation would only cause the service to slow to a halt if there was ever an influx in usage.

In scenarios like the DVLA it's easy to see how using NoSQL would largely benefit the organisation and the people using its services. British government agencies have done a good job ensuring their services stay modern, MongoDB also works with the National Archives, HM Revenues & Customs, UK Met Office and the Gov.uk website in general[6]. It's clear that government agencies that work with vast amounts of data benefit from the inclusion of NoSQL. So it begs the question of why in a country nearly 5 times larger than the UK do they not take advantage of this technology. MongoGB works with one US government agency, the Department of Veterans Affairs, a much smaller agency than a number of others in government. After reviewing the principles of relational and non-relational databases it's clear that in large systems that are open to the public to use NoSQL is the clear winner in terms of database technology. UK government departments have realised this, and similar US departments would benefit greatly by switching to this technology.

References

[1]J. Mashey, Big Data and the Next Wave of InfraStress Problems, Solutions, Opportunities. Monterey: {USENIX} Association, 1999.

[2]"Here's to eight great years", blog.youtube, 2013. [Online]. Available: https://blog.youtube/culture-and-trends/heres-to-eight-great-years. [Accessed: 03- May- 2021].

[3]S. Wojcicki, "YouTube at 15: My personal journey and the road ahead", blog.youtube, 2020. [Online]. Available: https://blog.youtube/news-and-events/youtube-at-15-my-personal-journey. [Accessed: 04- May- 2021].

[4]D. Abadi, "The problems with ACID, and how to fix them without going NoSQL", Dbmsmusings.blogspot.com, 2010. [Online]. Available: http://dbmsmusings.blogspot.com/2010/08/problems-with-acid-and-how-to-fix-them.html. [Accessed: 04- May- 2021].

[5]"DVLA Digitizes Vital Great Britain Driving License Data and Makes it Available Online With MongoDB | MongoDB Blog", MongoDB, 2015. [Online]. Available: https://www.mongodb.com/blog/post/dvla-digitizes-vital-great-britain-driving-license-data-and-makes-it-available-online. [Accessed: 04- May- 2021].

[6]"Government Customers", MongoDB. [Online]. Available: https://www.mongodb.com/industries/government. [Accessed: 04- May- 2021].