INTERNATIONAL JOURNAL OF BUSINESS EDUCATION AND MANAGEMENT STUDIES (IJBEMS)
A Peer Reviewed (Refereed) International Journal
Impact Factor 4.308 | http://www.ijbems.com | ISSN:2941-9638

Vol.1. Issue 1. 2022 (SEPTEMBER)

CHALLENGES IN DATABASE THEORY AND PRACTICE

BENJAMIN, UBOKOBONG EFFIONG & OBONGUKO, UBON ASUQUO
Department of Computer Science, Akwa Ibom State Polytechnic,
Ikot Osurua, Ikot Ekpene, Nigeria
benjaminubokobong52@gmail.com | obonguko69@gmail.com

ABSTRACT

This paper focused on challenges in database theory and practice. Database management has undergone more than four decades of evolution producing vast range of research and extensive array of technology solutions. The database research community and software industry has responded to numerous challenges resulting from changes in user requirements. Most recent database challenges arise because there are now hundreds of millions of users and cloud databases need to use novel techniques for managing massive amounts of data, securing data, prevent data duplication while supporting migration to other databases. Challenges highlighted in database theory and practice in this paper are use of primary key and social security numbers, deadlock detection and management, dealing with missing data, data privacy and data auditability challenges. Also presented are ways of overcoming the challenges in database theory and practice.

Keywords: Databases, Database Management System (DBMS), Primary Key, Online Transaction Processing (OLTP), Distributed Database Management System (DDBMS)

INTRODUCTION

Databases, in particular relational databases, are a ubiquitous part of today's computing environment. Database management systems support a wide variety of applications, from business to scientific and more recently various types of internet and electronic commerce applications. Database management systems (DBMS) are a core technology in most organizations today and run mission-critical applications that banks, hospitals, airlines, and most other types of organizations rely on for their day to day operation. Over the last three decades relational DBMS technology has proven to be highly adaptable and has evolved to accommodate new application requirements and the ever-increasing size and complexity of data (Pokorn, Snasel, and Richta, 2010).

But, there are challenges presented by this development. There are indications that some of the recently emerging data-intensive applications (e.g. internet searches) cannot be satisfactorily addressed using existing DBMS technology, and some experts argue that significant innovation is needed (a new database paradigm) to overcome the limitations of the current generation of database technology.

As we move into new frontiers in technologies such as mobile computing, cloud computing and big data analytics, databases are facing more and new issues. The emergence of new and many types of data resources such as social networks, emerging media, semantic web and technology of that nature the variety of data to be processed continues to increase rapidly, this large scale data is called Big data. According to Changqing et al (2012) big data includes data sets with sizes beyond the ability of current technology; methods and theory, to capture manage within a tolerable elapsed time. This presents issues in theory and practice of database management.

Database Concept

Connolly & Begg (2014) defines a database as "a shared collection of logically related data and its description, designed to meet the information need of an organization. Data is a collection of facts such as numbers, text, measurements, observations or a description of things. The data collected are stored in databases in a logical order which are then manipulated by Database Management Systems (DBMS) to create, insert, retrieve, maintain and control the database. Several issues can be identified in database theory and practice during the different phases of the system development life cycle (SDLC) of the database.

The use of databases in the applied sense is subjective, meaning it is strives to capture data specific to organizational need, field or a particular subject or domain. The attributes of the data captured has a factorial significance to the business process or the business organizational structure. Often times, executive management of business organizations want to capture and derive information from data that is being used by the business or upgrade its current legacy systems which are outdated and no longer supported to newer technologies that are supported.

Challenges in Database Theory and Practice

There are various challenges in database theory and practice and they include:

- **Use of Primary Key and Social Security Number (SSN):** One issue that is prevalent with naive databases developers is the assignment of primary keys. According to James (2014), a primary key value should have nothing to do with the data in a row. The primary key should be generated by the system sequentially or randomly based on insertion of a record in a database. Another challenge is the use of SSN. The use of data values such as social security numbers (SSN), employee id or student ids as primary keys in databases possess a serious problem to database integrity during database migration or changing underlying data. The use of SSN in particular as a primary key value in databases jeopardizes data integrity due to the fact that some people do not have SSN. In addition, some people may have more than one SSN in their life time. The practical solution to this problem is the use system generated values as primary keys instead of imposing application data values as primary keys.

- **Deadlock Detection and Management**
  One serious issue that database engineers must strive to avoid in databases is deadlocks. According to Obermacrk(1982), deadlocks occur when one transaction is suspended and

is waiting for a second transaction, and the second transaction is waiting (directly or indirectly) for the first transaction, which result in a circular wait condition. In DDBMS, detection of deadlocks is more complicated because the deadlock cannot be detected by a single site; inter-communication is required to detect deadlocks. (Stokes, 2007) An example of deadlock situation is shown in figure below:

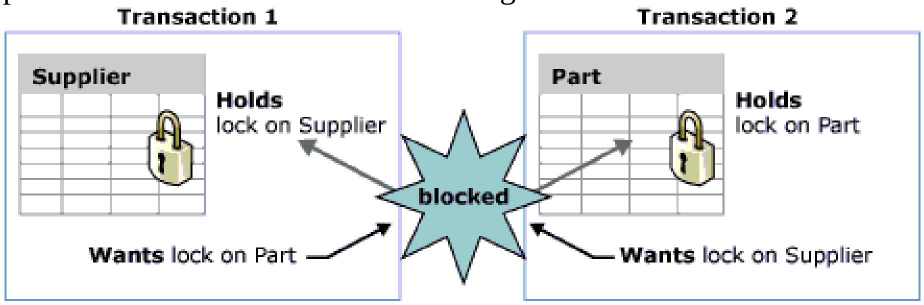

Figure 1. Deadlocking

In figure 1, transaction T1 is waiting on transaction T2 for the Part table lock resource. Similarly, transaction T2 is waiting on transaction T1 for the Supplier table lock resource. Because these dependencies form a cycle, there is a deadlock between transactions T1 and T2. To resolve this type of deadlocks in a distributed database environment, two known approaches can be employed; centralized and distributed. The centralized approach utilizes a global wait-for-graph which is constructed from all local wait-for-graphs within the sites and stored in a centralized detector. The centralized detector allows for bi-directional traffic to the centralized detector. Once a deadlock is detected, a resolution algorithm is initiated by the centralized detector to resolve to the deadlock. In the distributed approach, deadlock detection is equally shared between all sites and is only initiated if the system suspects a deadlock. After a deadlock has occurred, global information is required to determine the solution (Stokes, 2007).

- **Dealing with Missing Data**

Another issue in database theory and practice is how to deal with missing data during the analysis, design and implementation phases of the database SDLC. According to Morrissett (2013), missing data can be categorized into four types:

- Data that is missing because it is applicable but unknown at that time
- The value of the missing data is unreliable or invalid
- The missing data is not applicable to the current entity
- The missing data is unknowable due to understood reason such as law or policy.

The issue of missing data presents a serious challenge as it seriously affects the integrity of the database. As an example, let's evaluate the scenario of a mentally challenged adult male who is homeless. When he was under medication, he usually visits the county clinic to see a psychiatrist and his medical records which includes; name, social security number, address etc. were entered into the county electronic health record system. After becoming homeless, drifted to another city and not taking his medication, he breaks down with severe paranoia and had to be taken to another county clinic by emergency personnel. Barely coherent, he was able to provide his first and last name. Unable to get any other information from the mentally challenged homeless man, the clerk opens up another episode for the client; hence we have duplicate data for the same individual in the database. The example provided here is a real life occurrence that plaques the mental health database systems and it requires a lot of staff time and monetary cost to identify these duplicate records and merge them together in the system.

Bertot et al (2014) point out an ever growing concern with big data in the United States. The authors cite the U.S policy framework and argue that it lags behind technological advancement, raising the question of whether the existing policy addresses the issues raised by big data in an adequate manner. The authors then give recommendations on closing the policy gap such as developing data standards for instance the metadata standards like ISO 19115 (the international standard for geospatial metadata) and the Data documentation initiative (DDI) and by encouraging data sharing policies across sectors.

In addressing the Computation and analysis issues with big data Moniruzzaman & Hossain (2013) show that traditional relational database management systems are being complemented by alternative DBMS such as NoSQL, NewSQL and Search based systems. NoSQL systems according to the authors are distributed, non-relational database designed to process and store large scale data in a large number of servers. NewSQL on the other hand are relational databases that provide Atomicity, Consistency, Isolation, Durability (ACID) compliant real time Online Transaction Processing (OLTP) in big data environments by employing NoSQL features, in-memory processing, and symmetric multiprocessing or massively parallel processing.

- **Data Privacy**

Watson (2014) found that individual privacy issues are on the raise given the collection, storage, analysis and use of big data. The author gives Target as an example of this issue in 2013 when the company mined data in order to identify pregnant women by focusing on women who signed up for the baby registry and building a predictive analysis model. Target then used this data to send out invitations to join its bridal registry.

PR, N. (2013) gave an example of PRISM which was identified as the government spying program. These personal data came from large companies which supplied the National Security Agency (NSA) with information about their customers. Data privacy according to the article will become a legal battleground until new laws are passed, but before that individual organizations must come up with ways to show their customers that their data is protected.

Polonetsky & Tene (2013) proposed an equation that incorporates sizable benefits of big data with the attendant costs focusing on who the beneficiaries of the data are, what the nature of the perceived benefit is and with what level of certainty can those benefits be realized. With This equation the authors offer a way to account for the benefits that business and individuals accrue by the collection of certain data sets. Molnar (2014) cites a paper by John Podesta that contained policy recommendations that ensure student data are only used for educational purposes. Bertot *et al.* (2014) also pointed out existing data policies such as requiring agencies to ensure that confidentiality and privacy guidelines are followed regarding the releasing of data.

JISC Legal also released a code of practice designed to provide individuals with a degree of control over the use of their personal data especially the unforeseen secondary uses of this data. It also provides protection from unwanted or harmful uses of personal data (Charlesworth 2008).

- **Data Auditability**

Wang *et al* (2009) found that cloud computing and big data have brought about many security challenges most notably allowing a third party auditor to verify the integrity of data. Teppler, Warner & Smith (2003) show that data is now susceptible to shifts of reality and give an example of Arthur Anderson and Enron scandal where it was found that the team was amending data to correct records of its reviews, stating that it is likely that some form of digital data alteration is used in almost all sophisticated instances of audit fraud. The authors continue to show the value of data auditing by stating that it provides the means in ascertaining in the physical world that a transaction took place. Auditing also necessitates a human element of supervision and verification in order to validate transactions.

Wang *et al* (2009) proposed a third party Auditor defining it as an entity that has expertise and capabilities that clients do not have and is trusted to asses and expose risk of cloud storage services on behalf of the client upon request.

Cardenas (2013) discusses mutual Auditability that determines which actions were completed by which parties a concept of cloud security adopted from UC Berkeley Tech Report 2010. The author proposed the following items to provide cloud integrity.

- Communications using SSH (Secure shell) and HTTP should be verified
- All data upload should contain a signature to verify its contents

Overcoming the Challenges in Database Theory and Practice

The key to overcoming the challenges in database theory and practice is proper development. This is why it is important that during the requirements phase of the database Software Development Life Cycle (SDLC), it is very important clearly identify the scope and objective of the database in order not to end up creating a database that does not fulfill the purpose it was intended for. It is important during the requirements phase that the following database concerns be met as follows:

- Store all the data the business needs to track
- Must incorporate the business rules for processing that data
- Must protect data security and integrity
- Must be flexible enough to handle exceptions
- Must be flexible enough to allow for growth and change

The analysis, design and implementation phases of the database SDLC is very important as well because it adds the layer of functionality and usage to the requirements phase. Depending on the type of database management system (DBMS) or distributed database management system (DDBMS) that is used to manage the database, several issues have been identified to affect databases managed by DBMS or DDBMS distinctively or collectively. In situations where a database will be managed by a single DBMS at a specific location, it is important for database architects to ensure the integrity database during the design phase.

CONCLUSION

It has been revealed that the rise in the usage of database management systems has given birth to challenges in its theory and practice. Most recent database challenges in database theory and practice include internet-scale databases – databases that manage hundreds of millions of users and cloud databases that use novel techniques for managing massive amounts of data. In addition, challenges such as deadlock management, dealing with missing data, data privacy and data auditability have been identified and these challenges can be dealt with if the database is properly designed from the beginning of its software development cycle.

REFERENCES

Bertot, J. C., Gorham, U., Jaeger, P. T., Sarin, L. C., & Choi, H. (2014). Big data, open government and e-government: Issues, policies and recommendations. Information Polity: *The International Journal of Government & Democracy in the Information Age*, 19(1/2), 5-16. doi:10.3233/IP-140328

Cardenas, C. (2013, June 18). *The Four Keys of Cloud Security: Mutual Auditability*. Retrieved from https://www.joyent.com/blog/the-four-keys-of-cloud-security-mutual-auditability

Changqing, J., Yu, l., Wenming, Q., Yingwei, j., Yujie, x., Uchechukwu, a., & ... Wenyu, q. (2012). Big data processing: big challenges and opportunities. *Journal of interconnection networks, 13(3/4)*, 1-19.doi:10.1142/s0219265912500090

Charlesworth, A. (2008). Code of Practice for the Further and Higher Education sectors on the Data Protection Act 1998.JISC legal.

Connolly, T. M., & Begg, C. E. (2014). *Database Systems: A Practical Approach to Design, Implementation and Management*. New Jersey, NJ: Pearson.

Database Requirements Gathering. (2015) Getting the Scope. PowerPoint Presentation. Retrieved from www.seattlecentral.edu

James, J. (2012). *Five common database development mistakes*. Retrieved from http://www.techrepublic.com/blog/software-engineer/five-common-database-development-mistakes/

Microsoft SQL Server. (2014). *Deadlocking*. Retrieved from http://technet.microsoft.com/en-us/library/ms177433(v=sql.105).aspx

Molnar, M. (2014). Safeguard Use of Student Data, White House Report Urges. *Education Week*, 33(31), 26.

Moniruzzaman, A. M., & Hossain, S. A. (2013). NoSQL Database: New Era of Databases for Big data Analytics-Classification, Characteristics and Comparison. *International Journal of Database Theory & Application*, 6(4), 1-13.

Morrissett, M. R. (2013). *Missing data in the relational model* (Order No. 3561385). Available from ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global. (1367177739). Retrieved from http://search.proquest.com/docview/1367177739?accountid=26967

Obermacrk, R. Distributed Deadlock Detection Algorithm. 1982, *ACM Transactions on Database Systems*, Vol. 7, pp. 187-208.

Polonetsky, J., & Tene, O. (2013). *Privacy and Big Data Making Ends Meet*. Stanford law review. Retrieved from http://www.stanfordlawreview.org/online/privacy-and-big-data/privacy-and-big-data

Pokorn, J., Snasel, V. and Richta, K. (2010): Database Trends and Directions: Current Challenges and Opportunities pp. 163{174, ISBN 978-80-7378-116-3.

PR, N. (2013, August 27). Are Companies Overstepping the Line between Big Data and Big Brother?. PR Newswire US.

Stokes, P. R. (2007). *Design and simulation of an adaptive concurrency control protocol for distributed real-time database systems* (Order No. MR28449). Available from ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global. (304700687). Retrieved from http://search.proquest.com/docview/304700687?accountid=26967

Teppler, S. W., Warner, P. D., & Smith, L. M. (2003).Digital data and the meaning of 'audit.'. *CPA Journal*, 73(2), 70.

Wang, Q., Wang, C., Li, J., Ren, K., & Lou, W. (2009). *Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing*. doi:10.1007/978-3-642-04444-1_22

Watson, H. J. (2014). Addressing the Privacy Issues of Big Data. Business *Intelligence Journal*, 19(2), 4-7.