Journal of Advances in Science and Technology
Vol. 21, Issue No. 1, March-2024, ISSN 2230-9659

A Study the review of Duplicate Data in Cloud Computing

Krishna Kant Tiwari[1]*, Dr. Qaim Mehdi Rizbi[2]

[1] Research Scholar, Shri Krishna University, Chhatarpur, Madhya Pradesh, India

Email: kkit1984@gmail.com

[2] Associate Professor, Department of Computer Science & Application, Shri Krishna University, Chhatarpur, Madhya Pradesh, India

Abstract- In the realm of cloud computing, the management of data redundancy, often referred to as duplicate data, poses a significant challenge due to its implications on storage efficiency, data integrity, and overall system performance. This study provides a comprehensive review of the current strategies employed to detect, manage, and eliminate duplicate data in cloud environments. Duplicate and nearly duplicate web pages cause major issues for web search engines. Duplicate document identification, or the process of finding document pairs that represent the same entity, is a basic component of data cleansing. In order to address the structural heterogeneity issue, data must first undergo a data preparation step that involves parsing, data transformation, and data standardization. The findings aim to aid cloud service providers and users in understanding the importance of effective data deduplication strategies to optimize storage resources and enhance cloud computing performance.

Keywords- Cloud, Duplicate, Data, ETL, Data Cleaning, Documents

----------X----------

INTRODUCTION

An integrated database with subject-oriented, time-variant, & non-volatile data that is structured for multidimensional querying & business decision-making is called a cloud computing. The data models & underlying hardware/software systems of the integrated data sources that make up a cloud computing might vary widely. Star schema-based cloud computing include two main parts: (1) a fact table that stores all of the integrated, topic-oriented, time-variant records, and (2) a series of dimension tables that describe the subject, foreign key properties of the fact table. The "dirty" aspect of the data is a result of the variety in representations present when using data from multiple sources. This means that before it can be stored in the cloud computing, the data must be thoroughly cleansed. The process of identifying and eliminating duplicates in a cloud computing has been a hotspot for investigation in the cloud community for some time. Many studies have been conducted to remedy the data duplication issues brought on by data contamination. In this study, a system is developed to efficiently manage redundant information.

DUPLICATE DOCUMENTS

The Internet hosts numerous duplicates of the same piece of content. Research indicates that 40% or more of all web pages are actually just copies of other pages. The purpose of mirroring some data repositories is to offer redundancy and access dependability, although many of these are true copies as well. It is common practice for search engines to return a prioritized list of documents that are most relevant to a user's query. The user is given the opportunity to review and assess these materials. In order to find the desired results, web surfers must successively examine the titles & snippets from the lengthy list. Users waste time and energy trying to filter search engine results, especially when there are many close duplicates. When dealing with massive amounts of data, efficient near-duplicate detection becomes a critical concern in many applications. They are not identical bit for bit, even though almost identical papers look very similar. Duplicate and nearly duplicate web pages cause major issues for web search engines. These pages make users' lives more difficult by increasing the amount of space needed to store the index, which in turn slows down or increases the cost of serving results. Therefore, modern techniques are purposefully thought to be essential for the identification of these pages.

For many uses of data found on the Internet, duplicate content is an issue. Consider a search engine that crawls every website on the Internet regardless of whether they are duplicates or not. Then it's not out of the question that the first page of search results for a specific query could consist entirely of hits in the same documents. It is hard to see how this search engine may be helpful. With web corpora, it's the same story. When searching in

a corpus that contains duplicate texts, users may encounter numerous duplicate concordance lines. In addition, statistical analysis of corpus data may be skewed by duplicate content if it artificially inflates the frequencies of certain words and expressions. Therefore, in order to use Web data in text corpora, it is crucial to detect & remove duplicates and near-duplicates. One major issue with widespread applications is finding documents that are almost identical or have similar content inside vast collections. The issue has been discussed in several contexts for various data kinds, such as textual documents, spatial points, and relational records. Efficiently identifying almost identical web pages is another modern manifestation of the problem. Due to the high dimensionality of the documents and the massive amounts of data from many areas, this is undeniably a challenging task on a web-scale. A number of applications, including crawling, ranking, clustering, & archival caching, necessitate the identification of duplicate and almost duplicate documents because to the high rate of duplication in Web pages (G. S. Manku 2007; Yi, L., Liu, B 2003; Feetterly 2004; Hung-Chi Chang 2007).

In an effort to save processing & storage overheads, search engines strive to avoid indexing duplicate material. In modern computer-based management systems, databases are super important. Quality data is essential for many organizations. The ever-increasing data volume guaranteed by important data cleansing processes in real-world databases inevitably gives rise to data quality issues. Single data sets, such as files and databases, are prone to data quality issues. The necessity for data cleaning is greatly increased, for instance, due to data entry errors, incomplete or incorrect information that is not purposely removed, or the incorporation of numerous data sources into data warehouses. The goal of data cleaning is to enhance data quality by identifying and eliminating mistakes and inconsistencies. The data mining process relies heavily on data cleaning. Data quality enhancement in a data warehouse is an essential first step before data mining. There is a wide variety of data cleansing procedures used for different objectives.

Duplicate document identification, or the process of finding document pairs that represent the same entity, is a basic component of data cleansing. In order to address the structural heterogeneity issue, data must first undergo a data preparation step that involves parsing, data transformation, and data standardization. This stage ensures that data entries are stored uniformly in the database. The acronym "ETL" stands for "Extraction, Transformation, Loading," which is another word for data preparation. When databases are built from many sources, it's common for different versions of the same material to accumulate. A process called document deduplication is used to identify these many copies. Duplicate documents typically have a higher degree of similarity than random pairs of documents. In duplication detection, all papers with almost identical data in one or more fields are located. Every word—"document linkage," "duplicate detection," etc. denotes the same problem: finding documents that represent distinct entities but have different syntax.

THE ETL (EXTRACT TRANSFORM LOAD) PROCESS

There are four processes that deal with ETL process. They are called as Data staging processes collectively.

a) Extraction: There are various data elements in databases that are useful in decision making process thus facilitating the procedure of data mining. It is important to extract relevant data from huge collection of documents from web that may be relevant or irrelevant. Various tools are available for extracting data from web like ontology, data mining tools, semantic web language etc. The users specify their needs that which files and tables they want to access and extract by using SQL statements.

b) Transformation: After extracting or gathering information about required source, there is need to transform data now into warehouse so that it meets the suitable needs of user queries. It is based on quality criteria and basis of quality criteria, several studies and results have been produced. One of example of transformation is giving name to attributes to remove inconsistencies. Student Name may be STUDENT_NAME in one database, SNAME in the other. So, only one data name is chosen and converted to suitable format. The conversion may produce the following:

- ASCII characters must be converted EBCDIC or vice versa.
- Starting letter of all characters must be converted to uppercase letters.
- No numerical or integer values should be there.
- Programming data must be converted to pseudo code language that is understood by all users.

c) Cleansing: Data cleaning is essential to make it usable by the general public. Concerning data & information quality, it is relevant. Developers are notoriously bad at estimating the worth of data retrieved from the web. It is unable to alter its internal specifications but can alter its exterior characteristics. Data cleansing is the name given to this procedure. As an example, there may be ambiguities or discrepancies in the criteria evaluations, database selection, or data inputs (M.A. Hernandez 1998).

d) Loading: It is defined as transfer of data from machine to database for which data is stored in given data warehouse. It takes place after extraction phase and it took time to load, that why it started immediately after it. Oracle

warehouse builder gives features to perform all four processes in warehouse systems.

NEED FOR DATA CLEANING

Data is growing in quantity but in quality at a slower rate. Data overloading occurs when the quality of the world's data declines due to an excess of data from disparate sources and an absence of quality control procedures. Data mining, often called Knowledge Discovery Databases (KDD) (A. Susaria 2002), has become increasingly popular in the first ten years of this century. Its goal is to glean new information (concepts, patterns, explanations, among others) from existing database data. Technological breakthroughs produce massive amounts of data that are impossible to manually examine and interpret, which is why data mining has become so popular. Databases are frequently utilized without taking into account the potential mistakes and defects present, mostly because of their sheer number. There might be a "garbage-in, garbage-out" scenario if automated data mining and analysis are used to such "dirty data," since the findings could be extremely deceptive. When some of the incorrect results are re-incorporated into the information systems, it further complicates matters and starts a chain reaction of errors. The goal of data cleaning, a relatively new field, is to make data more reliable. This is of the utmost importance in databases that undergo rapid changes, like biological databases or data warehouses used by the financial and telecommunications industries. These databases receive new data every day from all over the world without proper data cleansing procedures or quality assurance checks (M.L. Lee 1999). The "dirty data" grows in quantity & quality when information flows between databases and undergoes transformation in data mining pipelines. Data cleansing is the first and most important phase in data mining, yet it is frequently skipped over because getting high-quality data isn't immediately apparent. "Dirty data" refers to data that is both complex & numerous, and the development of methods to clean this data is still in its early stages.

THE DILEMMA OF DUPLICATE DOCUMENTS

The problem of document is not an isolated issue; there are many ethical and criminal implications for them. Just one example of the problem is in the music industry, where many people have to pay millions of money as royalty to the original author of the work. According to definition of the Royal Academy of the Spanish Language (RAE) (Hussein 2015), the word Document comes from the Latin meaning Plagiarius kidnapping. The problem of document duplication has increased in recent decades because of the ease of access to information.

Some of the negative aspects of duplicate documents are:

- Lowers efficiency of search engines.
- Wastes crawler resources network bandwidth.
- Affects refresh time Increases storage cost.
- Affects the quality of search indexes.
- Increases the load on the remote host that is serving such web pages.
- Affects customer satisfaction.

The solution to this problem is to use a set of techniques or methods to detect the duplicate documents. Those solutions must allow you to discover if a document is indeed original or it was copied from another document.

TYPES OF DUPLICATE DOCUMENTS

A document is a written, drawn, presented or recorded representation of thoughts. When a document is copied, the source is referred to as the original and if the goal is to detect duplicate document and identify those who commit it, it is essential that their types and variations should be distinguished as below:

- **Copy with deliberate intent:** The goal is to do copying work. As indicated by an article "Document Prevention" published in 2005 by Northwestern University, “The deliberate copying is to use the work of others and convert as own work”. This copy can be from several sources: Essays, abstracts, encyclopaedia articles, books, magazines, papers among others.
- **Copy with direct intention:** This document presents some parts, paragraphs or sentences copied verbatim and not cited. The type of copy is intentional but, unlike the copy with deliberate intent, requires a greater effort to discover and analyze the suspect document.
- **Copy unintentionally:** Here the author does not have the knowledge for committing the crime of kidnapping. Copying unintentionally occurs when the author puts references to a date or investigation of any third party implying copying from the references.
- **Copy with structural variations:** The suspected document presents textual changes from the original. Copying with structural variations occurs when copying text has paragraphs, phrases or sentences from the original text modified with the intention to mask the exact copy.
- **Copy with grammatical variations:** Here the suspected document was translated because the original language is different. This type of copy with grammatical variations happens when

the copy document is from an original document, written in another language and translated.

- **Collusion:** Collusion is not exactly document and therefore cannot be classified directly into the group of backup types but refers to an event on a social level where a group of people makes an agreement fraudulently and secretly, for the purpose of committing document among them. According to the definition of Jenny Moon (2012) collusion means taking someone else's work and pass it off as their own, with the difference that the person who is being copied is aware of the copied document. In other words, it is the agreement between the original author and the author for copying the documents.

In Figure 1is shown the agreement between two individuals to interact with each other and copy your documents and this interaction can be modeled mathematically using graph theory G = (V, E) as a directed graph where documents are individual nodes (V) and the degree of copying between documents are the weights C (A, B) of the arcs (E).

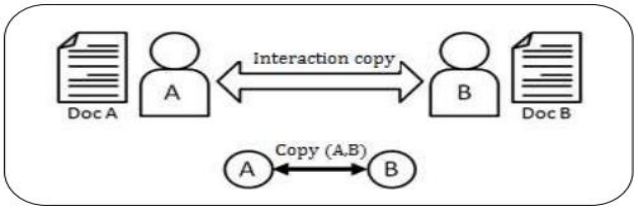

Figure 1: Copy Document

DUPLICATE DETECTION APPROACHES

Ahmed K. Elmagarmid et al. (2007) offer multiple methods for resolving the duplicate detection issue.

Probabilistic Matching Models: In order to divide record pairings into two categories, this model employs a Bayesian strategy. Consider two tables, A and B, that have n fields that are similar to one another. Each tuple pair {α, β >, where α is an element of A and β is an element of B, is allocated to one of the two classes M or U in the instance of a duplicate detection problem. If two entities are comparable ("Match"), then their records should be located in class M. If they are not, then their records should be located in class U. In 2003, Newcombe et al. were the first to identify duplicate detection as an issue of Bayesian inference. In duplication detection literatures, it is now widely employed.

In supervised learning systems, training data is present as record pairings that have already been classified as matching or not. In semi-supervised learning systems, this data is not present. Some supervised learning methods, like the probabilistic methods, handle each record pair (a, b) separately. In 2007, Ahmed K. Elmagarmid popularized the CART algorithm, which is used to create regression and classification trees. "Bootstrapping" is the main strategy for semi-supervised learning, which is also known as weakly supervised learning. Bootstrapping is a method whereby a small set of instances is fed into the system in order to discover related sentences & contextual cues. In an iterative fashion, this technique is used to train the system to uncover new cases that contain clues.

Unsupervised Learning: Using clustering algorithms to gather comparable comparison vectors together is one approach to avoiding manual labeling. The premise upon which the majority of unsupervised learning methods for duplicate detection are based is that classes are represented by comparison vectors that are comparable to one another. Duplicate detection via unsupervised learning is based on the probabilistic model. Learning matching models is accomplished through the use of a clustering-based bootstrapping technique. In training, the underlying principle is to make good use of sparsely labelled data by employing unsupervised learning approaches to appropriately label the data with missing labels. The comparison vector stores the outcomes of field comparisons in its individual entries. Afterwards, the Autoclass clustering program is used to partition the comparison space into clusters. The essential idea is that there are similar comparison vectors in each cluster. So, whether there are matches, no matches, or potential matches, all of the record pairs in the cluster fall into the same category.

Strategies for Engaged Learning: A second option to semi-supervised learning is unsupervised learning. The need for a huge quantity of training examples is an issue with supervised learning methods. Although it's simple to generate many training pairs that are either exact copies or precise non duplicates, it's far more challenging to generate ambiguous examples that would aid in the development of an excellent classifier. In order to identify situations and solicit user feedback, the strategy proposed building several classifiers with slightly varied data or parameters.

Distance Learning: Active learning strategies require training data or human effort to match models. It is not possible to use supervised or active learning methods when there is a lack of training data or human input. A distance measure for records can be defined that does not require adjusting with training data, thus eliminating the necessity for training data. Since distance-based methods compare each record in a single large field, they may miss crucial details that could be useful for duplication identification. A straightforward method involves computing the weighted distance between records after measuring the distance between individual fields with the relevant distance metric. Here, the computation of the weights becomes problematic, and the overall setting resembles the probabilistic scenario. Choosing a good value for the matching threshold is a challenge with distance-based methods. It is feasible to determine the

suitable threshold value because training data is available.

LITERATURE REVIEW

Afsoon Abbasi et al. (2022) In order to collect and analyze massive amounts of individual health records, contemporary healthcare systems depend on sophisticated computing technology, such as cloud-based systems. Protecting sensitive data in health applications while using advanced cloud services technologies like software as a service and application as a service poses a challenge for end-users of cloud systems. Recording and handling the information in a way that prevents any people' identities from being revealed is crucial when publishing data in the cloud. Preserving privacy in the healthcare cloud without sacrificing the quality of publicly available data is, however, a major privacy concern. A common technique for protecting personal information is the K-anonymity technology. Here, we propose a new strategy for finding the best k-anonymity algorithm, one that relies on clustering and the K-means++ method. To further enhance the anonymized data's quality, we also eliminate less common data using the normal distribution function. Extensive testing demonstrate that the suggested approach can significantly decrease execution time by 3.5 times and information loss by 1.5 times when compared to AKA and GCCG algorithms. Compared to others, it is also quite scalable.

Ali Shakarami et al. (2021) Recent years have seen the rise of cloud storage solutions, which promise to store data blocks on multiple servers in the cloud. Data replication is a key component of cloud storage systems, and there are a number of approaches to fixing this problem. By resolving issues with availability, reliability, security, bandwidth, & reaction time of data access, data replication aims to achieve better performance for data-intensive applications. Despite the effects and maturity of cloud data replication, the author is unaware of any systematic, exhaustive, or full survey. This study defines current schemes on the topic & presents outstanding issues using a classical classification. It then gives a detailed review and classification of state-of-the-art data replication strategies among several available cloud computing platforms. Information handling, data auditing, and data deduplication are the three primary categories that make up the offered categorization. A comprehensive analysis of the replication schemes reveals their key characteristics, including the classes used, scheme type, implementation location, evaluation tools, and pros and cons. Lastly, the survey will conclude with a discussion of open topics and future research challenges that have been either neglected or only partially addressed.

Wenjuan Li et al. (2021) With the implementation of virtualization & integration of resources, cloud computing has grown its service area, improved the user experience compared to traditional platforms, and brought about enormous social & economic benefits through its business model. On the other hand, there is a mountain of evidence indicating cloud computing is in the midst of a trust and security crisis, and the creation of a trust-enabled transaction environment is now its most important component. Common problems with the conventional cloud trust model include centralized architecture, excessive administration cost, overloaded networks, and potential failure points. Additionally, not all participants can fully acknowledge the conclusions of the trust rating because there is no traceability or transparency. The distributed ledger technology known as blockchain is an innovative & exciting new framework for decentralization. The operation rules and record traceability are its distinctive features that guarantee the security, integrity, and undeniability of the transaction data. Thus, building a distributed & decentralized trust architecture is a perfect fit for blockchain. This paper provides an extensive overview of cloud computing platforms that utilize blockchain technology for trust. It finds the unanswered questions and points the way for more study in this area by using a unique cloud-edge trust management architecture and a cloud transaction model based on a double-blockchain structure.

Satriyo Wibowo et al. (2020) Along with the IoT and Big Data Analytics, blockchain is one of the latest buzzwords in the digital economy since it provides a secure peer-to-peer link through the use of a decentralized database. Big Data relies on huge data storage in a centralized database, which is not eliminated. Information log, a type of application, is well-suited to blockchain since it requires often updated, dynamic data and the hierarchical hash security features necessary to operate a distributed database system. The CIA Triad, or confidentiality, integrity, & availability, of Big Data is vulnerable to attacks, but the properties of Blockchain have the potential to strengthen this defense. Since information is now essential to running a company, but maintaining large amounts of it presents difficulties in terms of security and reliability, Blockchain technology may be a viable option. Our research shows that Blockchain has the potential to increase the safety of Big Data in several ways: by bolstering the security of data storage; by improving data integrity via digital certificates & chaining blocks via the hash of the previous block; and by increasing data availability via peer-to-peer transmission, distributed nodes, and a consensus method. Better data veracity from token-based validation to improve truth finding & ID decentralization to confirm the identity of the source are two ways in which blockchain can improve the efficiency of Big Data Analytic.

Vaishali Wangikar et al. (2016) Record de-duplication is a process of identification and removal of duplicates from the given dataset in a data warehouse environment. The term record linkage is also used in the same context, the difference between record de-duplication and record linkage is that the former is used when the duplicates are to

be removed from one dataset while the later is used when the duplicates are to be removed from several different datasets that refer to the same entity. Both the processes de-duplication and record linkage are important during data profiling stage of a data warehouse and assure the quality of data without repetition which in turn leads to better decision making. Record de-duplication is focused for the presented research. The Efficiency of Record de-duplication is based on several criteria such as number of comparisons needed, time and cost of comparison, accuracy level of de- duplication, time and space complexity for identification of true duplicates. In this paper we have explored the several indexing techniques which are intended to make less number of comparisons to identify duplicates from the given dataset. Peter Christen has surveyed and experimented six different indexing techniques such as Sorted Neighborhood indexing, Suffix Array indexing, Q Gram based indexing, Canopy Clustering, Threshold based indexing, and String Map based indexing. In this paper, we have studied and implemented Sorted Neighborhood based de-duplication techniques in detail. During this implementation Adaptive and Non-Adaptive Sorted Neighborhood Methods are experimented and validated. Accumulative Adaptive SNM (AASNM), Incrementally Adaptive SNM (IASNM)[16] are adaptive versions of SNM while Duplicate Count Strategy (DCS) is a Non Adaptive SNM. A Group based Accumulative Adaptive Method (GAASNM) is proposed to minimize the record comparisons.

Zhaocong Wen et al. (2014) Deduplication is an important technique to save the storage cost at the cloud storage server. Image is an important data type stored in cloud, but rarely discussed in previous work on deduplication. This paper studies the problem of validating the deduplication of image storage in cloud. In particular, we consider the task of allowing a cloud server to verify the correctness of deduplication. Our scheme consists of several advantages over the previous work, whose framework can be described through the following algorithms. Firstly, before each user uploads an encrypted image, he calculates its hash value as the fingerprint. Secondly, the fingerprint is sent to both cloud servers for checking duplicates. If the storage and verification servers both reply to the user with 'no deduplication', the user transfers his data to the servers. Otherwise, once the fingerprint is consistently found, the user gives up uploading data for deduplication. Specially, when the fingerprint is only found in one server, it implies that the results are inconsistent and at least one of servers is invalid. The security and efficiency analysis are also presented in this paper.

CONCLUSION

Managing duplicate data in cloud computing is crucial for optimizing storage efficiency, reducing costs, and enhancing overall system performance. This study's comprehensive review of deduplication techniques reveals the significant benefits and challenges associated with various approaches, such as client-side and server-side deduplication, as well as inline and post-process methods. Each technique offers unique advantages and trade-offs concerning storage savings, processing overhead, and data security. Future research should focus on developing more advanced algorithms that enhance deduplication accuracy and efficiency while maintaining robust data security. Additionally, exploring the integration of deduplication with emerging technologies, such as machine learning and artificial intelligence, could offer innovative solutions to existing limitations.

REFERENCES

1. Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, Duplicate Record Detection: A Survey, IEEE Transactions on Knowledge and Data Engineering, Volume 19, NO. 1, January 2007.

2. Erhard Rahm, Honghai Do, Data Cleaning: Problems and Current Approaches, IEEE Bulletin of the Technical Committee on Data Engineering, Volume 23, No. 4, Page no. 3–13, 2000.

3. Giri, M. S., Gaur, B., & Tomar, D. (2015). A survey on data integrity techniques in cloud computing. International Journal of Computer Applications, 122(2), 27-32.

4. Huanzhuo, Y., & Di, W. (2010). A Survey of Approximately Duplicate Data Cleaning Method. Data Analysis and Knowledge Discovery, 26(9), 56-66.

5. Ignatov, Dmitry I., Katalin Tünde Jánosi-Rancz, and Sergei O. Kuznetzov. "Towards a framework for near-duplicate detection in document collections based on closed sets of attributes." Acta Univ. Sapientiae 1.2 (2009): 215-233.

6. Kaur, R., Chana, I., & Bhattacharya, J. (2018). Data deduplication techniques for efficient cloud storage management: a systematic review. The Journal of Supercomputing, 74, 2035-2085.

7. Li, Z., Xu, W., Shi, H., Zhang, Y., & Yan, Y. (2021). Security and privacy risk assessment of energy big data in cloud environment. Computational Intelligence and Neuroscience, 2021, 1-11.

8. Teng, Y., Xian, H., Lu, Q., & Guo, F. (2022). A data deduplication scheme based on DBSCAN with tolerable clustering deviation. IEEE Access, 11, 9742-9750.

9. Sonali Agarwal, Neera Singh, Dr. G.N. Pandey, “Implementation of Data Mining and Data Warehouse in E-Governance”, “International Journal of Computer

Applications (IJCA) (0975-8887), Vol.9- No.4, ", November 2010

10. Xiao, Chuan, et al. "Efficient similarity joins for near-duplicate detection." ACM Transactions on Database Systems (TODS) 36.3 (2011)

11. Matthias Friedrich's Blog, Basics of Near Duplicate Detection, http://blog.mafr.de/2011/01/06/near-duplicate-detection/

12. Mohamed, S. M., & Wang, Y. (2021). A survey on novel classification of deduplication storage systems. Distributed and Parallel Databases, 39, 201-230.

13. Chhabra, N., & Bala, M. (2018, December). A comparative study of data deduplication strategies. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 68-72). IEEE.

Corresponding Author

Krishna Kant Tiwari*

Research Scholar, Shri Krishna University, Chhatarpur, Madhya Pradesh, India

Email: kkit1984@gmail.com