International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.14, No.1 (Jul-23)

http://dx.doi.org/10.12785/ijcds/140124

A State of Art Survey for Big Data Processing and NoSQL Database Architecture

Aqib Ali[1], **Samreen Naeem**[1], **Sania Anam**[2] **and Muhammad Munawar Ahmed**[3]

[1]*College of Automation, Southeast University, Nanjing 210096, China.*
[2]*Department of Computer Science, Govt Associate College for Women Ahmadpur East, Bahawalpur, Pakistan.*
[2]*Department Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan.*

Received 6 Oct. 2022, Revised 30 Apr. 2023, Accepted 8 May. 2023, Published 1 Jul. 2023

Abstract: Internet corporations and other types of enterprises are quickly implementing NoSQL technology rather than merely SQL because of the introduction of Big Data. The simplicity of the design, the capacity to scale horizontally, and tighter control over availability are some benefits. Because of the schema-free data model they use, more businesses are beginning to recognize that NoSQL databases are a superior method for managing bulk amount of structured, semi-structured, and unstructured data currently being collected and handled. As a result, NoSQL database is increasingly viewed as a viable alternative to relational databases. For instance, NoSQL database is regularly utilized to gather and store data about social networking sites. This article targets to introduce the ideas that underpin NoSQL, as well as a review of the pertinent literature, an overview of the many types of NoSQL databases, and reasons both for and against the use of NoSQL. To analyze the known advantages of NoSQL and demonstrate the distinctions between the SQL and NoSQL methods, a simple SWOT analysis was performed. In the final part of the article, we draw some comparative studies of the NoSQL database and conclusions to broaden the scope of our previous work.

Keywords: Big Data, NoSQL, Structured Data, Unstructured Data, SWOT Analysis.

1. INTRODUCTION

Big Data refers to a vast amount and diversity of fast-collected data from several sources. Data is expected to expand 50% per year and 55 times between 2010 and 2022 [1]. Textual data is unstructured with Big Data, Internet organizations and enterprises are using NoSQL more. Simplicity, horizontal scalability, and availability management are benefits [2]. NoSQL databases are viewed as a potential alternative to relational databases because their schema-free data format can handle huge volumes of structured, semi-structured, and unstructured data. NoSQL databases hold social network data, for example. The goal is to introduce NoSQL, highlight its numerous forms, and present reasons for and against its use [3]. Since millions of data have lately been created at intervals of less than a millisecond, one of the most significant technical issues that the world is currently experiencing is the challenge of organizing and storing this data [4]. As a result, handling a vast amount of data is a significant obstacle. As a direct consequence of this, cutting-edge technology for data collecting and administration is necessary to accommodate the expanding human population. The requirement for quick data processing and generation has resulted in more than 2.6 trillion data every day [5]. The applications of Big Data shown in Figure 1.

Considering the way, it is utilized right now, they forecasted that there would be an even more significant exponential growth in the usage of and development of data going ahead in the future [6]. Affirm the obvious, which is that data management and storage in the era of big data, in which conventional software is being pushed to its limit by enormous amounts of data, creating the need for significant change, investing, capturing, and archiving data for future use, which causes several organizations to make a conscious effort to protect and maintain all potential data, is a significant challenge [7]. Even though the data is often unstructured, it can be created from many sources, including postings on social networking platforms, multimedia material with automated files, and so on. Monitoring and application systems for e-mail, search engine inquiries, content management document repositories, sensor data of different sorts, stock market imaging, satellite imagery, and electronic health records, among other things [8]. It is sometimes interpreted as ”Not Just SQL” to express those other technologies used in mass-distributed web applications and relational data technologies. Most notably, NoSQL technologies are essential to meet the high availability requirements of the online service [9]. It is an architecture for globally dispersed storage that incorporates a practical fundamental database management framework.

E-mail address: aqibcsit@gmail.com, samreencsit@gmail.com, sania.anam7@gmail.com, munawar69@iub.edu.pk

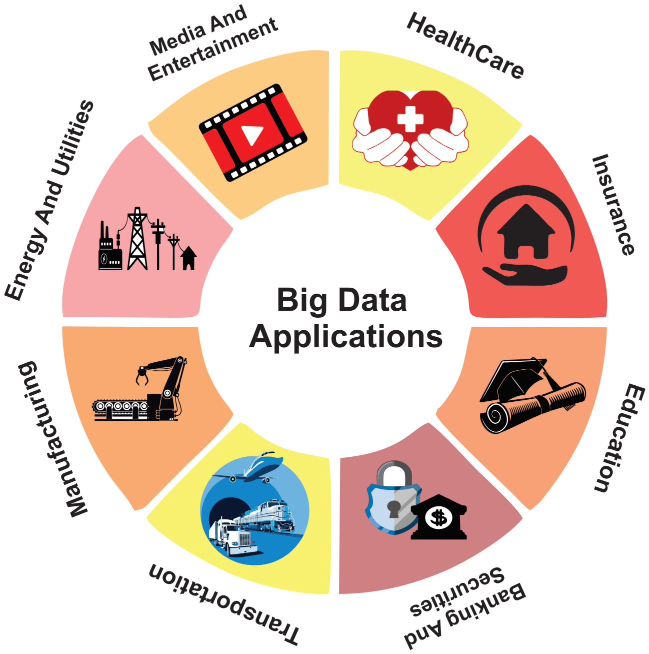


Figure 1. Big data applications

The actual data for the key values are saved in papers, charts, and families of columns and pairs of columns. As a result, numerous different notions of redundancy are offered to reduce the likelihood of mistakes and raise the level of availability of NoSQL database systems [10]. The name ”NoSQL” refers to any method of non-relational data management, and it must fulfill the following characteristics to be considered correct:

- The data are not organized in tables.
- The language known as SQL is not the database itself speaks.

However, the database technology connected to NoSQL technology has to develop to generate continuous worldwide access to the services it provides for applications that handle massive databases or solid online applications [2]. It is helpful to understand the role that the ”Big Data” tool plays in the overall notion of data generating since this knowledge makes it possible to store and manage the limitless quantities of information that are produced every day. Even though relational database management systems (RDMS) for fixed data storage have been around for a while, there is still work on scalability, consistency, system efficiency, data collecting, and data extraction integration [11] [12]. Big data technologies enable us to acquire more meaning from data through machine learning and those that enable us to store more significant amounts of data with better granularity than ever before. These technologies are collectively referred to as ”big data.” Google was a forerunner in developing these Big Data technologies, which eventually made their way into the general IT world in the shape of Hadoop. The term ”big data” does not yet have a definitive meaning, NoSQL works to assist in managing the volume, diversity, and speed needs of big data [13]. On the other hand, most data specialists will concur on the following three tenets: volume for huge volume of data, diversity in the form of numerous formats structured, semi-structured, and unstructured data, and speed for high-speed, real-time data processing. The Reference [14] qualities of big data were categorized as ”5V”, as indicated in the following Figure 2: volume, velocity, variety, veracity, and value.

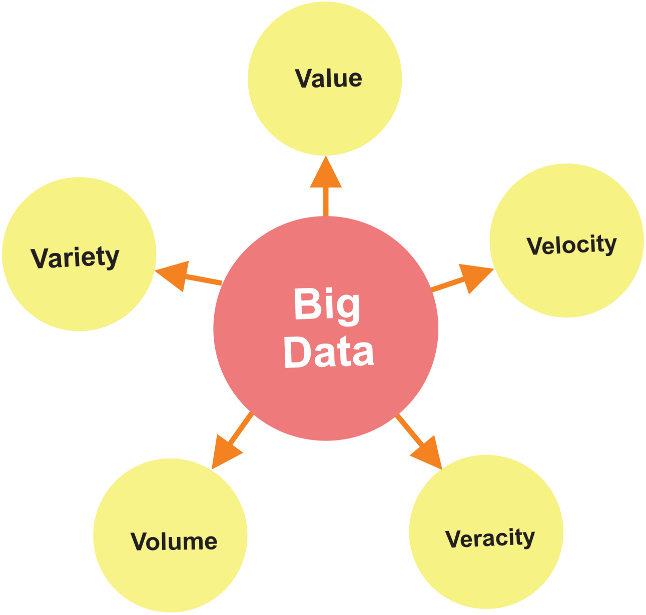


Figure 2. Big data 5 V’s

- **Volume:** Shows massive amounts of data, such as data for mobile devices, used for different functions.
- **Velocity:** specifies the rate or frequency of generation, updating, processing and access to data.
- **Variety:** data accessed through various types of devices, such as videos, photographs, etc.
- **Value:** explains how to draw useful knowledge from massive data sets. The most important aspect of any big data tool is value, as it enables the generation of valuable knowledge.
- **Veracity:** Refers to a huge precision and value of information.

The Reference [15] classifies 1 - 6 v’s of big data, which evolves into data worth, making it 7V’s of big data as indicated in the following Figure 3: Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value.

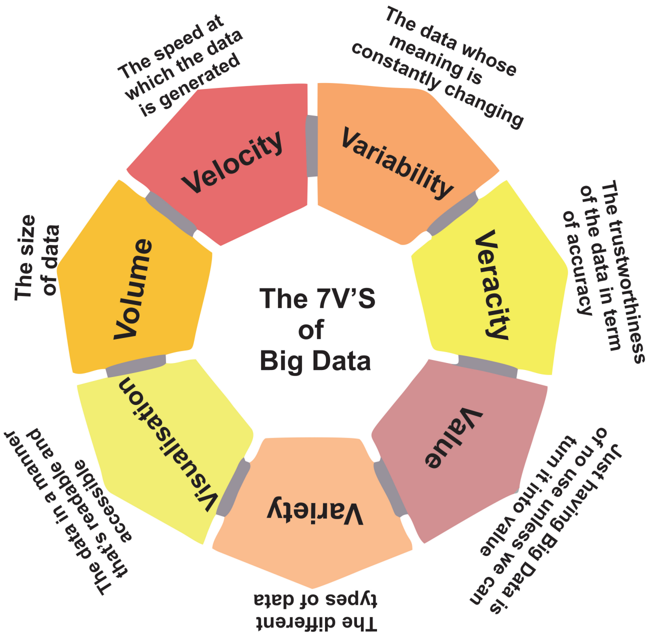

The 7V'S
of
Big Data
Velocity
The speed at which the data is generated
Variability
The data whose meaning is constantly changing
Veracity
The trustworthiness of the data in term of accuracy
Value
Just having Big Data is of no use unless we can turn it into value
Variety
The different types of data
Visualisation
The data in a manner that's readable and accessible
Volume
The size of data

Figure 3. Big data 7 V's

A. Background

The relational model has prevailed since the 1980s, with Oracle Databases, MySQL, and Microsoft SQL Servers as implementations Relational Database Management System (RDBMS) [16]. Increasingly, the usage of relational databases causes challenges with data modeling and parallel scalability on several servers and massive volumes of data as shown in Figure 4. Two tendencies bring the software community's attention to these issues:

- The exponential rise of the user, system, and sensor data, hastened by its concentration in Amazon, Google, and other cloud services [17].
- The Internet, Web 2.0, social networks, and open, standardized access to data sources from many diverse systems are accelerating data dependency and complexity [18].

The programming language known as SQL is utilized for communicating with relational databases. (Relational databases represent the data as records organized in rows and tables, with logical linkages connecting each element.) The term "NoSQL" refers to a category of database management systems (DBMSs) that are non-relational and, in general, do not employ SQL. It is possible to run queries in NoSQL at a significantly slower pace. You have an application that processes many transactions. SQL databases, which are more robust and maintain data integrity, are a better choice for high workloads or complicated transactions. It would help if you guaranteed that ACID (which stands for Atomicity, Consistency, Isolation, and Durability, is a set of attributes that a database transaction should ensure, even in the case of mistakes, power failures, and other similar occurrences) standards are met. In terms of performance, NoSQL is often quicker than SQL, particularly for key-value storage, as demonstrated by our experiment. On the other hand, a NoSQL database could not wholly support ACID transactions, which could lead to data inconsistency [19]. Large organizations with unstructured data are moving to NoSQL databases. NoSQL databases analyze massive volumes and offer better scalability than basic hardware [19]. Big Data Analytics, Business Intelligence, and social networks on petabyte datasets have overwhelmed centralized SQL-like systems. Data storage, Grid, Web 2.0, and cloud applications are difficult to scale with RDBMS. NoSQL databases in cloud computing, notably its scale-out and concurrency approach.

NoSQL databases differ from RDBMS, although they don't ensure ACID. Non-relational databases fueled by Web 2.0 apps [20]. The rigid relational structure might be burdensome for online applications like blogs with many characteristics. Multiple tables must hold text, comments, photos, videos, and source code. Because agile web apps require flexible databases, schema assessment must be straightforward. Adding or removing a blog feature utilizing a relational database makes the system inaccessible. NoSQL systems may store and index arbitrarily huge data sets, enabling many concurrent user queries [21].

2. LITERATURE REVIEW

The Reference [22] identified work that tackles difficulties and provides answers utilizing contradictory Big Data techniques to manage NoSQL databases. Large firms without storage solutions do not need to store and handle enormous data. According to the research, Casandra is a blend of Google's BigTable and Amazon's DynamoDB. The Reference [23] studied NoSQL databases for Big Data processing, covering transactional and structural issues. Big Data processing studies and problems were highlighted. The study provided a fantastic insight into NoSQL Big Data database literature, including structural data, difficulties, and real-time data acquisition. The Reference [24] described Big Data: It is said that data and knowledge expansion relies on user access. Cell phones, laptops, and PCs are readily available. Knowledge quickly dominates as users produce new material to manage, utilize, categorize, and secure data. The article organizes and analyzes Big Data's procedures, possible difficulties, investigations, and efforts and predicts its future expansion. The Reference [25] indicates, most relational databases alter NoSQL for data storage but do not replace SQL. The study compares SQL and NoSQL databases with Big Data Analytics. NoSQL databases, data models, data warehouse modules, features and characteristics, NoSQL vs. RDBMS pros and cons. The Reference [26] described NoSQL databases abandon relational databases' ACID qualities and embrace BASE (basically available, soft state, and eventually consistent). Eric Brewer developed the CAP theorem, which stands for consistency, availability, and partitioning.

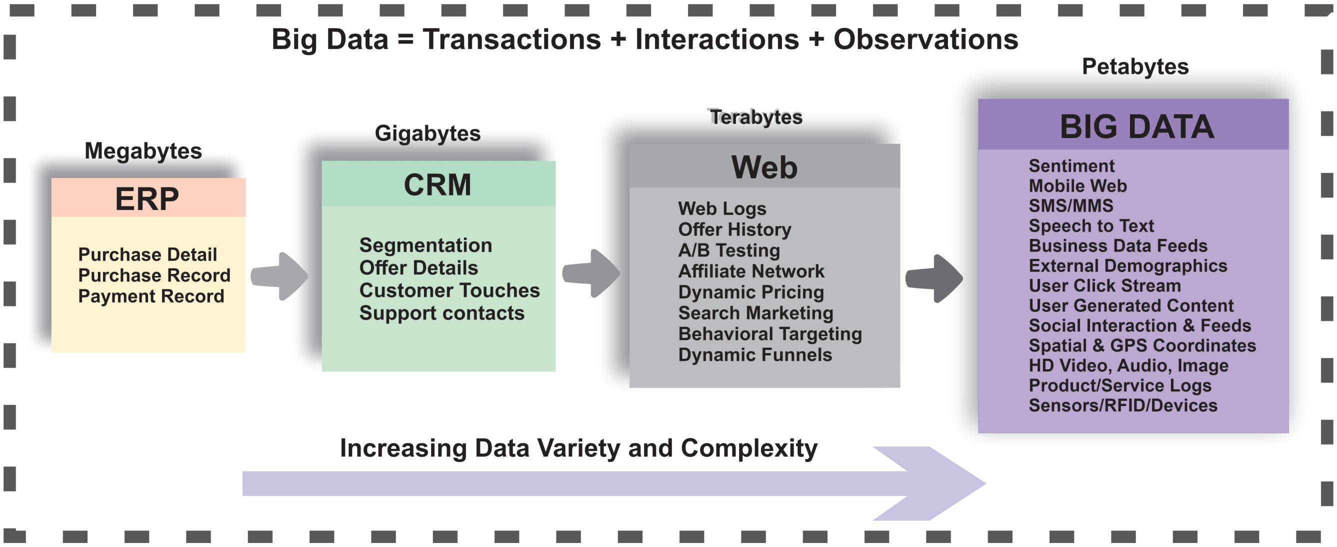


Figure 4. Big data transactions with interactions and observations

The Reference [27] described the BASE provides eventual consistency, enhancing NoSQL performance because data replicas do not need to mirror transactions instantaneously. This will lead us to study NoSQL's fragmentation, replication, and query processing algorithms. The Reference [28] described NoSQL databases provide high scalability using horizontal range partitioning, hash partitioning, or group partitioning. Replication creates data copies that may be accessed anywhere to minimize the volume and speed-related mistakes. Synchronous, asynchronous, primary copy, or update anywhere replication exists. This categorization describes the strategy wherein all replicas will be duplicated following transactions. NoSQL query processing is comparable to distributed database query processing and involves ”query planning” to decrease execution costs. The Reference [29] described planning is essential for aggregations and joins. Choosing NoSQL for an enterprise is difficult. Many studies and polls have been conducted to help corporations identify the best NoSQL database. Key-value, document, broad column storage and graph databases are NoSQL kinds.

3. CHARACTERISTICS OF BIG DATA

Most scientific groups are embracing data science. Facebook, Twitter, Wikipedia, Google, etc., are accountable for the vast fingerprint and data-trace repository. Internet sites have boosted data sources in recent years. Sensors capture temperature data, electricity meter readings, social media, digital video and photographs, traffic data, etc. [30]. These sources generate 5.5 quintillion bytes of data every day. The rapid accumulation of heterogeneous data on a big scale has raised doubts about conventional data models' processing capability. Data size is a significant concern due to rapid data growth. In 2022, a single dataset may grow from a few twenty of petabytes to several zettabytes, and 97% of global data had been created in the previous two years [31]. Existing database management solutions cannot handle their storage and processing needs. Many scientific communities want to use such data. Scientists are bullish about big data but doubt present data access strategies. They support modern data-processing technologies. In this era of data science, ”large, complex, diverse, structured or unstructured data” is dubbed ”BIG DATA”. Big data research is still immature, has more to be demystified, and needs active participation from many scientific groups to make new findings [32].

1) Big Data Challenges

As shown in the preceding discussion, big data is defined by new features, new data models, and new database technologies. Time and cost savings, intelligent decision making, successful product creation and development, and customer relationship support, to mention a few benefits of these new advancements, have helped firms and organizations of all sizes [33]. Despite substantial advancements in big data systems, future research faces several hurdles. Previous research, for example, has identified several problems associated with extensive data. This editorial gives a quick rundown of some of the essential difficulties in big data research. Existing database models cannot manage them owing to storage and processing [34]. This creates ”Big Data” from the database data. Big Data is new and needs further investigation. Big Data addresses data volume, velocity, and variety as shown in Figure 5. All of the three concerns are briefly described here.

- **Data Volume:** All types of data flood today's industry. Terabytes become petabytes quickly. Tweets generate 12 gigabytes of data every day. Power meters produce 350 billion measurements annually [35].
- **Data Velocity:** Companies capture data quickly. A minute's delay can generate parsed output discrep-

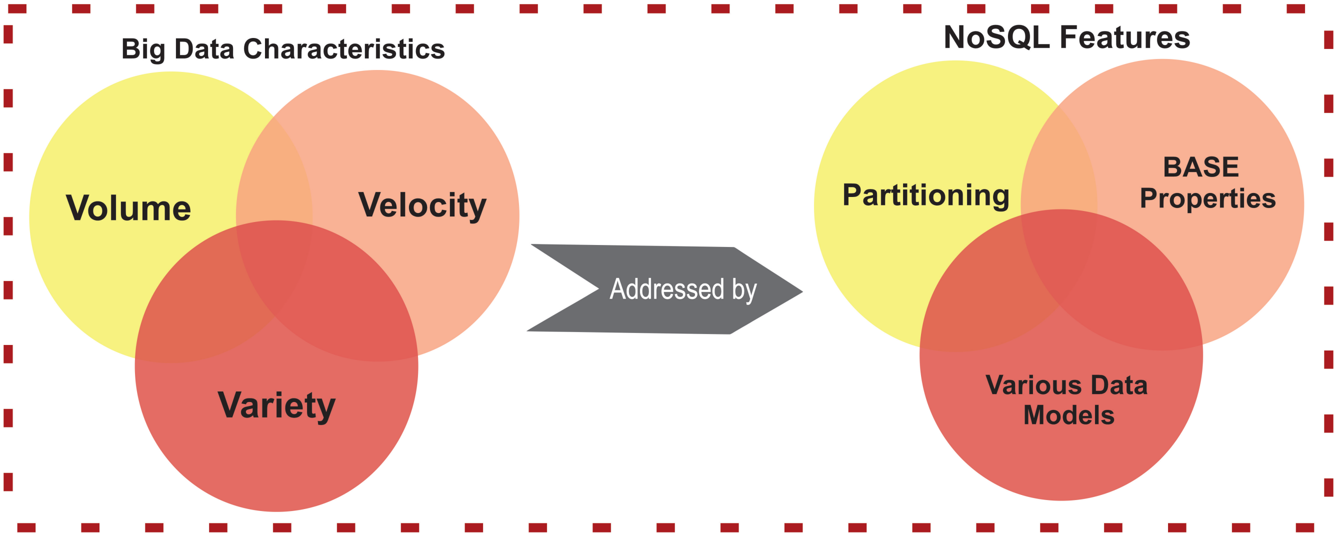


Figure 5. Big data characteristics and NoSQL database features

ancies. For fraud detection, extensive data must be studied as it enters firms for optimal results, such as scanning all 5 million daily company transactions for suspected fraud [36].

- **Data Variety:** Big Data can be organized or unstructured. Text, music, video, sensor data, log files, etc., are data types. Companies analyze extensive data to uncover new insights. Monitoring live video feeds from surveillance cameras to restrict the focus of interest; analyzing films, documents, and photos to increase customer satisfaction [36].

4. CHARACTERISTICS OF NoSQL DATABASES

Transaction-based databases maintain data integrity. This guarantees consistent data handling. These transactional properties are called atomicity, consistency, isolation, and durability (ACID) [37]. Scaling ACID systems proved difficult. Known as the CAP theorem, conflicts emerge between many elements of high obtainability in distributed systems:

- **Consistency:** Even when updating, all clients see the same data. XA transactions and ACID.
- **Available:** Even if one computer in a cluster is down, all consumers can constantly find a replica of the required data.
- **Tolerant Partitions:** Transparent to the client, the Total system saves its characteristics even when spread. According to the CAP theorem, only two of three horizontal scaling characteristics may be fulfilled entirely at once.

For improved partitioning and availability, several NoSQL databases feature lax consistency constraints. This led to BASE systems. These lack traditional transactions and add data model limitations to improve partitioning [38].

A. *Types of NoSQL Databases*

NoSQL databases support unstructured data. It incorporates a SQL database; both may coexist. NoSQL databases contain flexible schemas that may be changed without downtime or service disturbance, unlike relational (SQL) databases. NoSQL was built for large-scale data demands, like Facebook's 500 million users and Twitter's terabytes of data [39]. NoSQL uses horizontal scalability by distributing its load across numerous database servers. NoSQL databases lack the data integrity of relational databases. NoSQL databases manage massive data effectively despite this. NoSQL databases vary from relational databases primarily in their data model. NoSQL types addresses document Database, column, key value and graph database [40] as shown in Figure 6.

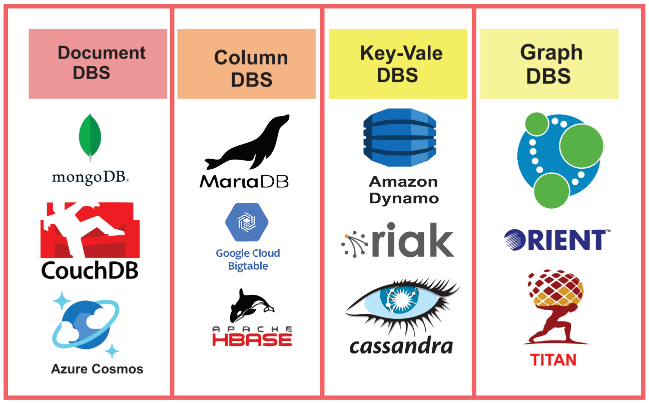


Figure 6. Types of NoSQL database

1) Document Database

Document databases employ XML or JSON files as a dataset. This streamlines data access and lowers the need for complicated joins or operations. Archived documents are scheme-free and comparable, which is beneficial for modeling unstructured data [41].

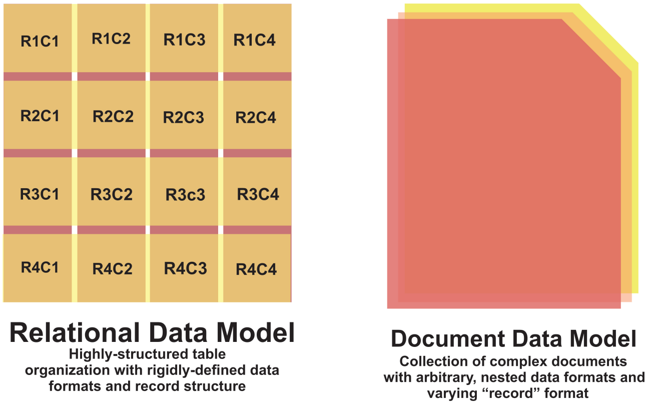

R1C1
R1C2
R1C3
R1C4
R2C1
R2C2
R2C3
R2C4
R3C1
R3C2
R3c3
R3C4
R4C1
R4C2
R4C3
R4C4
Relational Data Model
Highly-structured table organization with rigidly-defined data formats and record structure
Document Data Model
Collection of complex documents with arbitrary, nested data formats and varying "record" format

Figure 7. An example of document database

The Figure 7 illustrates a document database. Like relational databases, document databases may be queried using both keys and values. These databases store and manage large amounts of text documents and e-mails. The following are some examples of such databases:

- **MongoDB:** This open-source NoSQL database is scalable, high-performance, and has document-oriented (JSON-like) storage, complete index support, replication, and quick in-place updates. This product is appropriate for dynamic queries, dynamic data structures built-in C / C ++, and where indexes are preferred over Map / Reduce [42].
- **CouchDB:** Another open-source database, CouchDB focuses on storing data in various JSON documents, each with its schema definitions. ACID semantics provide consistency by preventing database files from being write-locked. This Java-based software is designed for web-based applications that deal with enormous volumes of disorganized data [43].

2) Column Oriented Database

Column-oriented databases contain linked data sets in column families with a row key. Super columns are nested columns that are nested within other columns. This scalable database can handle complicated data sets. Online analytical processing employs this to swiftly calculate data. Google Earth utilizes Big Table to store data. These databases store and analyze large volumes of data in distributed systems, notably versioned data owing to timestamp functions [44]. Figure 8 shows a column-oriented client database.

Here are a couple of such examples:

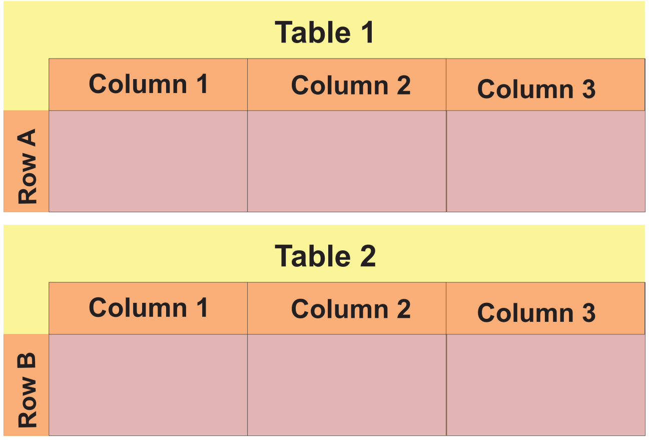


Figure 8. An example of column-oriented database

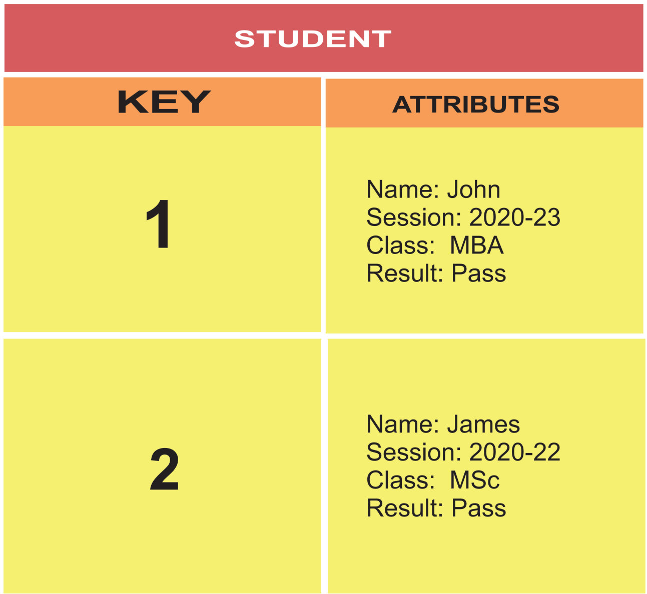
- **HBase:** HBase is a distributed and portable Big Data Store based on the Hadoop database and Google's BigTable technology [45].
- **BigTable:** is a Google product.
- **Cassandra:** is an open-source distributed database management system designed to manage massive amounts of data dispersed over numerous servers without a single point of failure and with a high level of accessibility. This solution, written in Java, is perfect for analyzing non-transactional real-time data with linear scalability and proven fault tolerance and column indexes [45].

3) Key Value Database

Key-value databases utilize a hash table with a unique key and a reference to a collection of values; only the key may be searched. The data isn't organized since key-value pairs lack a pattern. Facebook employs an unorganized, unconnected database. This database form is simple and highly scalable, making it perfect for customer profile maintenance and product name retrieval. Amazon's shopping cart uses DynamoDB. These databases were built for effective distributed data handling. Figure 9 illustrates self-database key-value pairs [46].

Some Example are as follows:

- **DynamoDB:** is a database management system (provided by Amazon) [47].
- **DB Berkeley:** is a database management system (provided by Oracle) [47].
- **REDIS:** Because keys can comprise strings, hashes, lists, sets, and ordered tax, REDIS is also a data structure server. This product is blazingly fast, making it ideal for real-time data collecting. It is written in C / C ++ [47].



STUDENT	
KEY	ATTRIBUTES
1	Name: John Session: 2020-23 Class: MBA Result: Pass
2	Name: James Session: 2020-22 Class: MSc Result: Pass

Figure 9. An example of key value database

- **Riak:** is a robust, distributed, open-source database that expands capacity predictably and facilitates application development through prototyping, fast development, and deployment. This technology, written in Erlang and C, provides transparent fault tolerance and recovery features and a rich and extensible API that is ideal for point-of-sale and factory control systems [48].

- **VoltDB:** NewSQL is a scalable in-memory database that provides complete transactional ACID consistency and ultra-high-speed. This technology is suited for financial markets, digital networks, network services, and online games. It uses segmentation and replication for highly accessible data snapshots and continuous command recording utilizing Java Stored Processes (for crash recovery) [49].

4) Graph Database

Graph databases use nodes, their relationships, and their attributes to describe data as a network of key-value pairs. Graph databases are beneficial when data connections are more important than the data themselves. Location-based facilities utilize this database to discover Facebook friends or the quickest traffic routes (see Figure 10). Graph databases utilize shorter route techniques to make data queries faster than relational databases. These databases are schema-free and horizontally scalable. InfoGrid and Neo4J are graphing databases [50]. NoSQL databases make data management more agile and scalable. NoSQL helps organizations develop new apps, enhance customer experience, and cut expenses. A worldwide insurance firm uses NoSQL to instantly aggregate consumer data from more than 70 systems into a single 360-degree picture. The supplier can handle consumer issues faster. Chicago collects and analyzes geographical data from over 30 agencies utilizing NoSQL to decrease crime and enhance public services. You may analyze 911 calls and reports in a district to predict a crime relapse. NoSQL is impressive at handling unstructured data, but it has some limitations. NoSQL's stability and trustworthy standards support this argument. Since NoSQL data is kept in partitions, availability or consistency cannot be guaranteed, affecting complicated query processing [51]. The following are some examples of such databases:

- **Neo4j:** is the premier graphics database and platform. It is open-source and commercially licensed for enterprise-level security, performance, and reliability through clustering. Cypher, Neo4j's graph query language, is straightforward to understand, and you may utilize the recently published open-source toolkits ”Cypher on Apache Spark (CApS) and Cypher for Gremlin” [52].

- **FlockDB:** Twitter solves fewer problems than other chart databases. It has geared for high-performance, low-latency online settings like webpages [52].

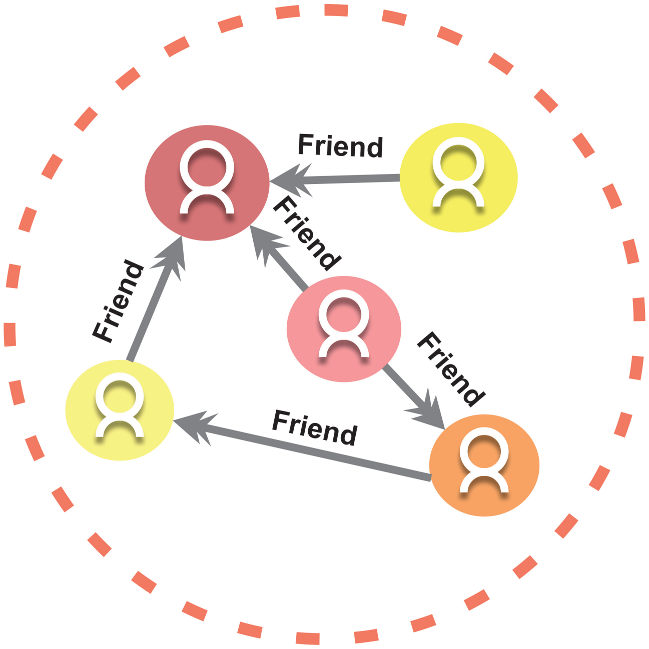


Figure 10. An example of graph database

- **ArangoDB:** needs a database, query language, and three data models. Potential. ArangoDB is a rapidly-growing native multi-model NoSQL database [52].

- **OrientDB:** First multi-model DBMS with True Graph driver. Multiple model NoSQL handles complicated domains efficiently [53].

- **Titan:** is a scalable graph database designed to store and query graphs over a multi-machine cluster. Titan is a transactional database that enables real-time visual walkthroughs by thousands of users [53].

- **DataStax:** helps organizations survive in a fast-changing, aspirational world.
- **Amazon Neptune:** is safe, quick, and includes a fully-managed graph database service to create and execute applications [53].

B. CAP Theorem

NoSQL is an affordable and effective solution to manage Big Data. NoSQL can conduct complicated operations on massive data sets when paired with MapReduce, outperforming traditional data warehouses and BI applications in performance and cost. Figure 11 shows that distributed systems can have at most two consistency, availability and partition tolerance.

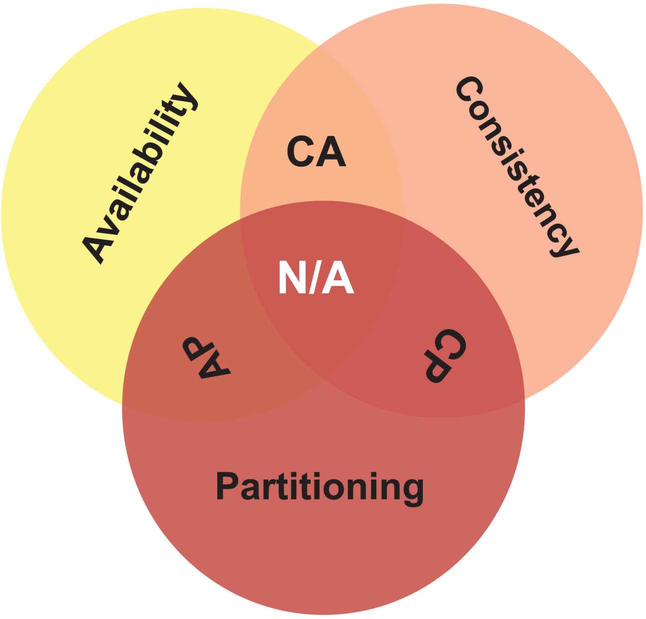


Figure 11. An example of CAP theorem

NoSQL is young. 44% of corporate IT specialists had never heard of NoSQL, and only 1% said it was part of their strategic approach. NoSQL technology enables firms to accomplish their strategic goals and earn new revenue with Big Data [54]. NoSQL databases are becoming a viable alternative to conventional databases as more firms realize schema-free architecture as a superior method for managing semi-structured and unstructured data. NoSQL databases play a significant role alongside traditional databases in many businesses. NoSQL databases don't replace SQL databases; they coexist. NoSQL databases offer interactive applications and are utilized for online, mobile, and cloud business solutions. NoSQL databases power mobile app backends.

Relational databases can't keep up with Facebook's mobile app changes. NoSQL supports many smartphone kinds and includes database updates as the application progresses. Product catalogs hold data for e-commerce and product data management. NoSQL databases may hold objects with various properties in a single database. Its dynamic architecture lets enterprises to add items and new characteristics without interruption; for example, consumers may write product feedback added to the product database [55].

5. BIG DATA PROCESSING USING NOSQL

In Big Data processing, valuable data extraction is an essential issue. The four stages through which this extraction must be processed are shown in Figure 12. Big data becomes a challenge when the quantity of the dataset exceeds the software's capacity to gather, process, retrieve, and manage the data. In Big Data processing, raw data sets from numerous sources are generated [56].

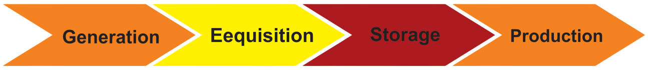


Figure 12. Big data processing life cycle

The second phase is the acquisition, which involves collecting and prepping Big Data for preservation. The archiving step archives Big Data using DFS (DFS). HDFS, GFS, and TFS are available for significant data processing. Big Data generation is crucial for batch-based (MapReduce), BSP-based, or stream-based data processing [57]. Hadoop's distributed structure makes real-time and small-dataset analysis difficult. NoSQL database parses real-time, non-file datasets. To manage the high quantity of datasets, consider the following issues:

- Collecting and processing massive data is challenging.
- There is no way to organize that much-varied data.
- Slower decision-making, enhanced automated data extraction.
- There is no dependable Wifi gateway. Summary of Big Data processing and NoSQL storage management research directions and problems.

Big Data analytics effect. NoSQL databases help address transactional problems and cloud infrastructure gaps. Below, we discuss the fundamental Big Data database properties.

6. COMPARATIVELY ANALYSIS OF NOSQL DATABASE

As seen in Table I, we use a matrix based on a system property to analyze various NoSQL databases and then classify them into one of four groups. Table II displays a matrix that divides NoSQL databases into four distinct groups, each determined by a different design feature. As

TABLE I. A MATRIX COMPARES FOUR TYPES OF NOSQL DATABASES BASED ON A SYSTEM ATTRIBUTES.

	NoSQL Database Model								
Feature	Document Stored		Wide Column Stored				Key Value Store		Graph Oriented
	MongoDB	CouchDB	DynamoDB	HBase	Cassandra	Accumulo	Redis	Risk	Neo4J
Value Size Maximum	16MB	20MB	64KB	2TB	2GB	1EB	-	64MB	-
Operating System	Cross platform	Windows, Mac OS	Cross platform	Cross platform	Cross platform	NIX 32	Windows, Mac OS	Cross platform	Cross platform
Programing Language	C++	C, C++, Python	Java	Java	Java	Java	C, C++	Erlang	Java

TABLE II. A MATRIX COMPARES FOUR TYPES OF NOSQL DATABASES BASED ON A DESIGN ATTRIBUTES.

	NoSQL Database Model								
Feature	Document Stored		Wide Column Stored				Key Value Store		Graph Oriented
	MongoDB	CouchDB	DynamoDB	HBase	Cassandra	Accumulo	Redis	Risk	Neo4J
Data Storage	Volatile	Volatile	SSD	HDFS	-	Hadoop	Volatile	Volatile	Volatile
Query Language	Volatile	JavaScript	API Calls	API Calls	CQL Thrift	-	API Call	Java Script	API Call
Protocol	Custom	HTTP	-	HTTP	CQL3	Thrift	Telnet	HTTP	HTTP
Conditional Entry Updates	Yes	Yes	Yes	Yes	No	Yes	No	No	-
MapReduce	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Unicode	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
TTL for Entries	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	-
Compression	Yes	Yes	-	Yes	Yes	Yes	Yes	Yes	-

TABLE III. A MATRIX COMPARES FOUR TYPES OF NOSQL DATABASES BASED ON AN INTEGRITY ATTRIBUTES.

	NoSQL Database Model								
Feature	Document Stored		Wide Column Stored				Key Value Store		Graph Oriented
	MongoDB	CouchDB	DynamoDB	HBase	Cassandra	Accumulo	Redis	Risk	Neo4J
Integrity Model	BASE	MVCC	ASID	Log Replication	BASE	MVCC	-	BASE	ASID
Atomicity	Conditional	Yes	Yes	Yes	Yes	Conditional	Yes	No	Yes
Consistency	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Isolation	No	Yes	Yes	No	No	-	Yes	Yes	Yes
Durability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-	Yes
Transactions	No	No	No	Yes	No	Yes	Yes	No	Yes
Referential Integrity	No	No	No	No	No	No	Yes	No	Yes
Revision Control	No	Yes	Yes	Yes	No	Yes	No	Yes	Yes

TABLE IV. A MATRIX COMPARES FOUR TYPES OF NOSQL DATABASES BASED ON AN INDEXING ATTRIBUTES.

	NoSQL Database Model								
Feature	Document Stored		Wide Column Stored				Key Value Store		Graph Oriented
	MongoDB	CouchDB	DynamoDB	HBase	Cassandra	Accumulo	Redis	Risk	Neo4J
Secondary Indexes	Yes	Yes	No	Yes	Yes	Yes	-	Yes	No
Composite Keys	Yes	Yes	Yes	Yes	Yes	Yes	-	Yes	-
Full Text Search	No	No	No	No	No	Yes	No	Yes	-
Geospatial Indexes	Yes	No	No	No	No	Yes	-	-	Yes
Graph Support	No	No	No	No	No	Yes	No	Yes	Yes

TABLE V. A MATRIX COMPARES FOUR TYPES OF NOSQL DATABASES BASED ON A DISTRIBUTION ATTRIBUTES.

	NoSQL Database Model								
Feature	Document Stored		Wide Column Stored				Key Value Store		Graph Oriented
	MongoDB	CouchDB	DynamoDB	HBase	Cassandra	Accumulo	Redis	Risk	Neo4J
Horizontal Scalable	Yes	Yes	Yes	Yes	Yes	Yes	-	Yes	Yes
Replication	Yes	Yes	Yes	Yes	Yes	Yes	-	Yes	No
Replication Mode	Master-Slave	-	-	Master-Slave	Master-Slave	-	Master-Slave	Multi Master	-
Sharding	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Shared Nothing	Yes	Yes	Yes	Yes	Yes	-	-	Yes	-

can be seen in Table III, NoSQL databases may be divided into four distinct groups by employing a matrix that is founded on an integrity feature. The NoSQL databases are divided into four groups by employing a matrix and basing it on an indexing property, as shown in Table IV. Table V presents a matrix that divides NoSQL databases into four distinct groups, each of which is determined by a property related to distribution. This comparison will be helpful for anyone who conducts a study in the future.

7. NOSQL SWOT ANALYSIS

An organization may better understand its strengths, weaknesses, opportunities, and threats by doing a SWOT analysis (See Figure 13), which provides a framework for doing so. The fundamental objective of a SWOT analysis is to raise awareness of the elements considered in determining a course of action for a corporation or making a business decision [58].

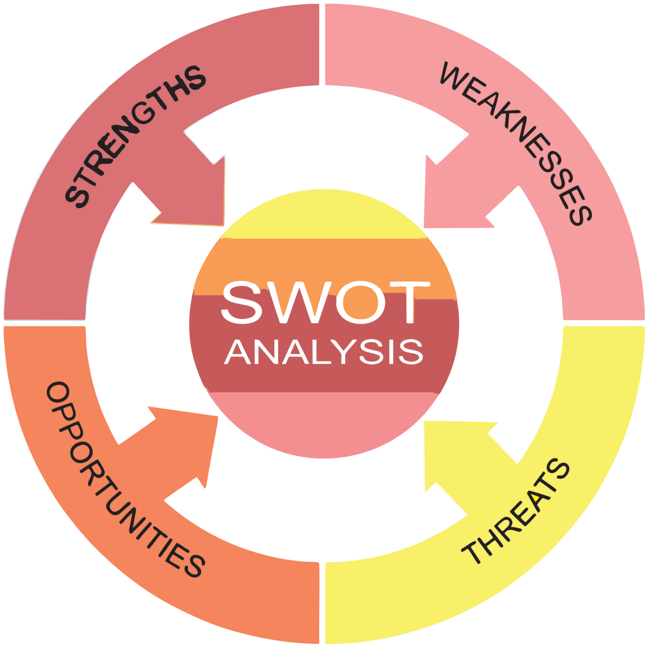


Figure 13. NoSQL SWOT analysis

A. Strengths

- Scalability.
- Safety while maintaining adaptability.
- The power of personal decision.
- Big data computing (maybe unstructured).
- High availability and uninterrupted access are guaranteed.
- Greater freedom; the database format does not conform to the standard SQL conventions.
- A high level of scalability and performance that is compatible with growing amounts of data.
- Open source, which enables the community to collaborate on the NoSQL Progress.
- Auto Shade ensures that data is distributed fairly and equitably among all servers. Even if one of the servers fails, the others will continue to function normally, and there will be no loss of data [59].

B. Weaknesses

- The process of porting applications.
- Normalization facility not available here as well as joins.
- Failed to get updates.
- In the absence of standards and open-source, it is more challenging to obtain assistance and requires more extensive training.
- There is a lack of maturity; it has just been there for the past ten years, yet many still like SQL databases.
- Applications include intricate modifications of pre-existing database programs to work with NoSQL databases [60].

C. Opportunities

- Important financial commitments
- Data-intensive applications

D. Threats

- A paradigm for the commercialization of free software
- Fear, Uncertainty, and Doubt (also known as FUD) among Users

These days, nearly all NoSQL databases have a flaw that might be somewhat problematic. Few people bring it up in conversation. They discuss the efficiency and the simplicity of utilizing schema-free databases concerning the friendly bees. The vast majority are programmers, non-operation administrators, and system administrators. Nobody asks. However, this is where the proverbial ”rubber meets the road” [61].

- **The correction of ad hoc data:** either there is no query language accessible, or there are no skills.
- **Ad hoc report:** no query language available or no internal expertise
- **Data export:** There is not always an API mechanism to retrieve all the data.

8. CONCLUSION

Big Data Analytics, social networks and business Intelligence have compelled centralized SQL databases to their bounds. This steered to distributed, horizontally scalable NoSQL databases. Let us imagine some NoSQL Database uses: Parallel processing in distributed systems, Integrated IR, Expert-level semi-structured data exploration, and data storage capacity. NoSQL is a huge, developing field. Features (characteristics and advantages of NoSQL databases); categorization (four groups on its characteristics); contrast and assessment (using a matrix based on design, integrity, indexing, distribution, system); and NoSQL database adoption. This research provides an unbiased knowledge of the merits and disadvantages of several NoSQL database techniques to enable software that process massive amounts of data and an outline of these non-relational NoSQL databases.

A. *Acknowledgment*

The authors would like to thank the referees for their careful reading and for their comments, which significantly improved the paper. Additionally, thanks to Dr. Salman Qadri, (Associate Professor, Chairman Department of Computer Science, MNS University of Agriculture, Multan, Pakistan) and Dr. Farrukh Jamal, (Assistant Professor, Department of Statistics, The Islamia University of Bahawalpur, Pakistan) for his motivational support.

B. *Conflicts of Interest*

The authors declare no conflicts of interest.

References

[1] N. Deepa, Q.-V. Pham, D. C. Nguyen, S. Bhattacharya, B. Prabadevi, T. R. Gadekallu, P. K. R. Maddikunta, F. Fang, and P. N. Pathirana, "A survey on blockchain for big data: approaches, opportunities, and future directions," *Future Generation Computer Systems*, 2022.

[2] M. A. Kausar and M. Nasar, "Sql versus nosql databases to assess their appropriateness for big data application," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 14, no. 4, pp. 1098–1108, 2021.

[3] J. Ahmed and M. Ahmed, "A study of big data and classification of nosql databases," in *2020 IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)*. IEEE, 2020, pp. 1–8.

[4] A. B. Rashid, M. Ahmed, and A. B. Ullah, "Data lakes: A panacea for big data problems, cyber safety issues, and enterprise security," in *Next-Generation Enterprise Security and Governance*. CRC Press, 2022, pp. 135–162.

[5] A. H. Alsup, "Examining the relationship between query performances when using different data models within relational database systems," Ph.D. dissertation, Colorado Technical University, 2021.

[6] G. Appel, L. Grewal, R. Hadi, and A. T. Stephen, "The future of social media in marketing," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 79–95, 2020.

[7] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *International Journal of Forecasting*, vol. 37, no. 1, pp. 388–427, 2021.

[8] B. Reeves, N. Ram, T. N. Robinson, J. J. Cummings, C. L. Giles, J. Pan, A. Chiatti, M. Cho, K. Roehrick, X. Yang *et al.*, "Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them," *Human–Computer Interaction*, vol. 36, no. 2, pp. 150–201, 2021.

[9] P. Atzeni, F. Bugiotti, L. Cabibbo, and R. Torlone, "Data modeling in the nosql world," *Computer Standards & Interfaces*, vol. 67, p. 103149, 2020.

[10] H. Vera-Olivera, R. Guo, R. C. Huacarpuma, A. P. B. Da Silva, A. M. Mariano, and M. Holanda, "Data modeling and nosql databases-a systematic mapping review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–26, 2021.

[11] Y. Zhan and K. H. Tan, "An analytic infrastructure for harvesting big data to enhance supply chain performance," *European Journal of Operational Research*, vol. 281, no. 3, pp. 559–574, 2020.

[12] M. Bernetti, M. Bertazzo, and M. Masetti, "Data-driven molecular dynamics: a multifaceted challenge," *Pharmaceuticals*, vol. 13, no. 9, p. 253, 2020.

[13] D. Kumar, P. Kumar, A. Ashok *et al.*, "Introduction to multimedia big data computing for iot," in *Multimedia big data computing for IoT applications*. Springer, 2020, pp. 3–36.

[14] J. Liu, T. Li, P. Xie, S. Du, F. Teng, and X. Yang, "Urban big data fusion based on deep learning: An overview," *Information Fusion*, vol. 53, pp. 123–133, 2020.

[15] W. Hattawi, S. Shaban, A. Al Shawabkah, and S. Alzu'bi, "Recent quality models in bigdata applications," in *2021 International Conference on Information Technology (ICIT)*. IEEE, 2021, pp. 811–815.

[16] T. Taipalus, H. Grahn, and H. Ghanbari, "Error messages in relational database management systems: A comparison of effectiveness, usefulness, and user confidence," *Journal of Systems and Software*, vol. 181, p. 111034, 2021.

[17] A. K. Singh, N. Firoz, A. Tripathi, K. Singh, P. Choudhary, and P. C. Vashist, "Internet of things: From hype to reality," in *An industrial IoT approach for pharmaceutical industry growth*. Elsevier, 2020, pp. 191–230.

[18] H. Guo, S. Nativi, D. Liang, M. Craglia, L. Wang, S. Schade, C. Corban, G. He, M. Pesaresi, J. Li *et al.*, "Big earth data science: an information framework for a sustainable planet," *International Journal of Digital Earth*, vol. 13, no. 7, pp. 743–767, 2020.

[19] G. Ekren and A. Erkollar, "The potential and capabilities of nosql databases for erp systems," in *Advanced mis and digital transformation for increased creativity and innovation in business*. IGI Global, 2020, pp. 147–168.

[20] S. Dong, A. Kryczka, Y. Jin, and M. Stumm, "Rocksdb: evolution of development priorities in a key-value store serving large-scale applications," *ACM Transactions on Storage (TOS)*, vol. 17, no. 4, pp. 1–32, 2021.

[21] U. Störl, M. Klettke, and S. Scherzinger, "Nosql schema evolution and data migration: State-of-the-art and opportunities." in *EDBT*, 2020, pp. 655–658.

[22] S. Kalid, A. Syed, A. Mohammad, and M. N. Halgamuge, "Big-data nosql databases: A comparison and analysis of "big-table","dynamodb", and "cassandra"," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. IEEE, 2017, pp. 89–93.

[23] M. R. Ahmed, M. A. Khatun, A. Ali, and K. Sundaraj, "A literature review on nosql database for big data processing," *Int. J. Eng. Technol*, vol. 7, no. 2, pp. 902–906, 2018.

[24] V. Rahmati, "Improved interpolation and approximation through

order manipulation," *International Journal of Emerging Computing Methods in Engineering*, vol. 1, no. 1, 2016.

[25] M. MUS, "Comparison between sql and nosql databases and their relationship with big data analytics," 2019.

[26] A. H. Al Hinai, "A performance comparison of sql and nosql databases for large scale analysis of persistent logs," 2016.

[27] A. K. Zaki, "Nosql databases: new millennium database for big data, big users, cloud computing and its security challenges," *International Journal of Research in Engineering and Technology (IJRET)*, vol. 3, no. 15, pp. 403–409, 2014.

[28] F. Gessert, W. Wingerath, S. Friedrich, and N. Ritter, "Nosql database systems: a survey and decision guidance," *Computer Science-Research and Development*, vol. 32, no. 3, pp. 353–365, 2017.

[29] M. T. González-Aparicio, A. Ogunyadeka, M. Younas, J. Tuya, and R. Casado, "Transaction processing in consistency-aware user's applications deployed on nosql databases," *Human-centric Computing and Information Sciences*, vol. 7, no. 1, pp. 1–18, 2017.

[30] J. Q. Dong and C.-H. Yang, "Business value of big data analytics: A systems-theoretic approach and empirical test," *Information & Management*, vol. 57, no. 1, p. 103124, 2020.

[31] S. Rizvi, I. Williams, and S. Campbell, "Tui model for data privacy assessment in iot networks," *Internet of Things*, vol. 17, p. 100465, 2022.

[32] X. Cid Vidal, L. Dieste Maroñas, and Á. Dosil Suárez, "Modern machine learning: Applications and methods," in *Machine Learning and Artificial Intelligence with Industrial Applications*. Springer, 2022, pp. 19–61.

[33] E. Irannezhad, C. G. Prato, and M. Hickman, "An intelligent decision support system prototype for hinterland port logistics," *Decision Support Systems*, vol. 130, p. 113227, 2020.

[34] P. Mikalef, R. van de Wetering, and J. Krogstie, "Building dynamic capabilities by leveraging big data analytics: The role of organizational inertia," *Information & Management*, vol. 58, no. 6, p. 103412, 2021.

[35] M. Naeem, T. Jamal, J. Diaz-Martinez, S. A. Butt, N. Montesano, M. I. Tariq, E. De-la Hoz-Franco, and E. De-La-Hoz-Valdiris, "Trends and future perspective challenges in big data," in *Advances in intelligent data analysis and applications*. Springer, 2022, pp. 309–325.

[36] A. Baldassin, J. Barreto, D. Castro, and P. Romano, "Persistent memory: A survey of programming support and implementations," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–37, 2021.

[37] D. Khan, L. T. Jung, and M. A. Hashmani, "Systematic literature review of challenges in blockchain scalability," *Applied Sciences*, vol. 11, no. 20, p. 9372, 2021.

[38] S. Bagga and A. Sharma, "A comparative study of nosql databases," in *The International Conference on Recent Innovations in Computing*. Springer, 2021, pp. 51–61.

[39] M. Suriakala, "A comparative study of different types of nosql datamodels and databases," *International Journal of Grid and Distributed Computing*, vol. 13, no. 2, pp. 1249–1257, 2020.

[40] J. D. Guerrero-Sosa, V. H. Menéndez-Domínguez, M.-E. Castellanos-Bolaños, and F. Moo-Mena, "Document database for scientific production," in *Proc. 8th Int. Workshop on ADVANCEs in ICT Infrastructures and Services*, 2020, pp. 129–132.

[41] B. Jose and S. Abraham, "Performance analysis of nosql and relational databases with mongodb and mysql," *Materials today: PROCEEDINGS*, vol. 24, pp. 2036–2043, 2020.

[42] A. C. F. Spengler and P. S. L. de Souza, "The impact of using couchdb on hyperledger fabric performance for heterogeneous medical data storage," in *2021 XLVII Latin American Computing Conference (CLEI)*. IEEE, 2021, pp. 1–10.

[43] A. H. Abed, "Big data with column oriented nosql database to overcome the drawbacks of relational databases," *Int. J. Advanced Networking and Applications*, vol. 11, no. 05, pp. 4423–4428, 2020.

[44] P. Jakkula, "Hbase or cassandra? a comparative study of nosql database performance," *International Journal of Scientific and Research Publications*, vol. 10, no. 3, pp. 808–820, 2020.

[45] G. Xie and Y.-C. Chung, "Bucket-based expiration algorithm: Improving eviction efficiency for in-memory key-value database," in *The International Symposium on Memory Systems*, 2020, pp. 248–259.

[46] E. S. Kumar, S. Kesavan, R. C. A. Naidu *et al.*, "Comprehensive analysis of cloud based databases," in *IOP Conference Series: Materials Science and Engineering*, vol. 1131, no. 1. IOP Publishing, 2021, p. 012021.

[47] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, "A survey of data partitioning and sampling methods to support big data analysis," *Big Data Mining and Analytics*, vol. 3, no. 2, pp. 85–101, 2020.

[48] P. Gupta, A. Mhedhbi, and S. Salihoglu, "Columnar storage and list-based processing for graph database management systems," *Proceedings of the VLDB Endowment*, vol. 14, no. 11, pp. 2491–2504, 2021.

[49] S. Naeem, W. K. Mashwani, A. Ali, M. I. Uddin, M. Mahmoud, F. Jamal, and C. Chesneau, "Machine learning-based usd/pkr exchange rate forecasting using sentiment analysis of twitter data," *CMC-Computers Materials & Continua*, vol. 67, no. 3, pp. 3451–3461, 2021.

[50] R. Wang, Z. Yang, W. Zhang, and X. Lin, "An empirical study on recent graph database systems," in *International Conference on Knowledge Science, Engineering and Management*. Springer, 2020, pp. 328–340.

[51] S. Mondal and N. Mukherjee, "Efficient nosql graph database for storage and access of health data," in *Computer Communication, Networking and IoT*. Springer, 2021, pp. 135–146.

[52] Z. Shen, Z. Zhao, H. Wang, Z. Liu, C. Hu, and C. Zhou, "Pandadb: Intelligent management system for heterogeneous data." *Int. J. Softw. Informatics*, vol. 11, no. 1, pp. 69–90, 2021.

[53] A. K. Pandey and R. Pandey, "Influence of cap theorem on big data analysis," *Int. J. Inform. Technol.(IJIT)*, vol. 6, no. 6, 2020.

[54] A. Ali, W. K. Mashwani, M. H. Tahir, S. B. Belhaouari, H. Alrabaiah, S. Naeem, J. A. Nasir, F. Jamal, and C. Chesneau, "Statistical features analysis and discrimination of maize seeds utilizing

machine vision approach," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 1, pp. 703–714, 2021.

[55] S. Naeem and A. Ali, "Bees algorithm based solution of non-convex dynamic power dispatch issues in thermal units," *Journal of Applied and Emerging Sciences*, vol. 12, no. 1, 2022.

[56] Y. Zhai, J. Tchaye-Kondi, K.-J. Lin, L. Zhu, W. Tao, X. Du, and M. Guizani, "Hadoop perfect file: A fast and memory-efficient metadata access archive file to face small files problem in hdfs," *Journal of Parallel and Distributed Computing*, vol. 156, pp. 119–130, 2021.

[57] M. A. Benzaghta, A. Elwalda, M. M. Mousa, I. Erkan, and M. Rahman, "Swot analysis applications: An integrative literature review," *Journal of Global Business Insights*, vol. 6, no. 1, pp. 55–73, 2021.

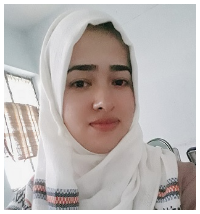
[58] K. A. ElDahshan, A. A. AlHabshy, and G. E. Abutaleb, "Data in the time of covid-19: a general methodology to select and secure a nosql dbms for medical data," *PeerJ Computer Science*, vol. 6, p. e297, 2020.

[59] M. Fruth, K. Dauberschmidt, and S. Scherzinger, "Josch: Managing schemas for nosql document stores," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 2693–2696.

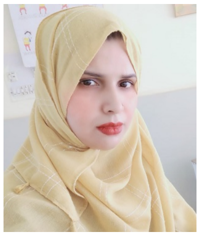
[60] A. Ali and S. Naeem, "The controller parameter optimization for nonlinear systems using particle swarm optimization and genetic algorithm," *Journal of Applied and Emerging Sciences*, vol. 12, no. 1, 2022.

[61] K. Goel and A. H. Ter Hofstede, "Privacy-breaching patterns in nosql databases," *IEEE Access*, vol. 9, pp. 35 229–35 239, 2021.

AQIB ALI obtained his Bachelor's degree in Computer Science (2017), after which he enrolled and completed his M.Phil. Degree in Computer Science from the Islamia University of Bahawalpur, Pakistan (2020). He is also working as a lecturer in computer science and IT at reputed institutes in Pakistan. Now he is doing his Ph.D. degree from Southeast University China to complete his education.



SAMREEN NAEEM got her Bachelor's Degree in Information Technology (IT) from Sargodha University Pakistan. After that, she enrolled and completed her M.Phil Degree in Computer Science from The Islamia University of Bahawalpur, Pakistan. She also works as a computer science IT lecturer in Pakistan's reputed institutes. Now, she is doing her Ph.D. to complete his studies at the Southeast University of China.



Sania Anam got her Bachelor's Degree in Computer (2007),Master degree in Computer (2009) after that he enrolled and completed his M.Phil. Degree in Computer Science (2016) from The Islamia University of Bahawalpur, Pakistan. Since 2016, she is working as Lecturer in Computer Science in Govt Associate College for Women Ahmad pur East ,Bahawalpur, Pakistan.

MUHAMMAD MUNAWAR AHMED completed his Bachelor's Degree in Computer (2005), after that he completed his MSCS Degree in session (2011-13) from The Islamia University of Bahawalpur, Pakistan. He is also working as Lecturer, department of Information Technology at The Islamia University of Bahawalpur, Pakistan. Currently, he is enrolled in Ph.D. program at The Islamia University of Bahawalpur, Pakistan.