

KCA UNIVERSITY RUARAKA MAIN CAMPUS

SCHOOL OF TECHNOLOGY

UNIT: ADVANCED DATABASES

CODE: BSD3108

ASSIGNMENT 3: DISCUSS:

WEB SERVICES, DATABASE & LOGIC, DATA STREAMS, BIG DATA, SOCIAL MINGING

TRIMESTER: MAY – AUGUST 2022

LECTURER: MR. DAVID K MARANGU

GROUP 5 MEMBERS:

|  |  |  |  |
| --- | --- | --- | --- |
|  | STUDENT NAME | REGISTRATION NO. | COURSE |
| 1. | STELLA KABURA | 17/04393 | BSD |
| 2. | AGEY HOPE | 19/02249 | BSD |
| 3. | MYRA SIKA | 20/04165 | BSD |
| 4. | RONOH EMMANUEL KIPKOECH | 19/03434 | BSD |
| 5. | TOM MWAMBIRE | 19/03962 | BSD |
| 6. | HARUN DEROW | 19/02760 | BSD |
| 7. | PRUDENCE MUTHONI | 19/02733 | BSD |

Table of Contents

[WEB SERVICES 2](#_Toc108195277)

[DATABASES AND LOGIC 5](#_Toc108195278)

[DATA STREAMS 8](#_Toc108195279)

[BIG DATA 12](#_Toc108195280)

[SOCIAL MINING 18](#_Toc108195281)

[REFERENCES 21](#_Toc108195282)

## WEB SERVICES

What are web services?

Independent, customized, distributed web, flexible applications that can be defined, published, available, or requested by the network to create products, processes, and supply chains. These applications can be local, distributed, or web-based. Web services built on open standards such as TCP / IP, HTTP, Java, HTML, and XML.

However, the web service is any service:

1. Available online or on private (intranet) networks.
2. Uses standard XML messaging system.
3. It is not tied to any single operating system or programming language.
4. Do you describe yourself in the standard XML syntax?
5. Available for easy access

Components of Web Services

The basic web services platform is XML + HTTP. All standard web services run using the following components:

1. Simple Object Access Protocol (SOAP)
2. Global Definition, Discovery and Integration (UDDI)
3. Web Services Description Language (WSDL)

How does the web service work?

Below is a description of how it works:

1. Web-Service acts as an application response pattern that enables communication between different applications through open ports, for example, HTML, XML, WSDL, and SOAP.
2. The feature will ask for some partner support to become a specialized service provider. If necessary, the specialist co-op will respond with a response message. Therefore, there are two messages combining one application message (XML) and one reply message (XML).
3. Managers use XML to record information, INSIPHO to deliver a message at the end, WSDL to indicate access to management systems.

Web Services Features

1. *XML-Bases*: The web service uses XML to represent data and data transfer layers.
2. *Loosely Coupled*: Web service is not tied to the concept of a server directly. The web service interface can change over time without compromising the client's ability to interact with the service.
3. *Ability to be synchronous or asynchronous*: Synchronization refers to binding a client to perform a service. For compatible applications, the client blocks and waits for the service to complete its operation before proceeding. Concurrent functionality allows the client to request a service and perform other functions. Asynchronous clients get their result over time, while compatible clients get their result when the service is completed.
4. *Remote Processing Support (RPCs)*: web services that allow clients to request procedures, functions and methods for remote objects using the XML-based protocol. Remote control processes specify input and output parameters that a web service must support.
5. *Support Document Exchange*: web services represent many complex documents in a standard way. Supports transparent document exchanges to facilitate business integration.

Architecture of Web Services

There are two ways to look at web services architecture:

1. Examine the role of each service actor.

2. Check the emerging web service protocol stack.

Web Service Roles

The three main roles within the web service are:

1. *Service Provider*: this is a web service provider and uses the service and makes it available online.

2. *Service Requester*: This is a web service client and uses an existing web service by opening a network connection and sending an XML request.

3. *Service Registry*: This is a logical and comprehensive list of services and provides a central location where developers can publish new services or acquire existing ones.

Web Service Protocol Stack

The four main layers are:

1. *Transport Service*: this is the layer responsible for transporting messages between applications. Includes Hyper Text Transport Protocol (HTTP), Simple Mail Transfer Protocol (SMTP), and File Transfer Protocol (FTP).
2. *XML Messaging*: This layer is responsible for encoding messages in a standard XML format so that messages can be understood anywhere. Includes XML-RPC and SOAP.
3. *Service Description*: This layer is responsible for defining a public interface in a particular web service. Hosted in the Web Service Definition (WSDL) language.
4. *Service Discovery*: This layer is responsible for integrating resources into the same registry and providing the publishing / acquisition function. Hosted with Universal Description, Discovery, and Integration (UDDI)

Web service security

There are three specific security issues with web services:

1. Confidentiality
2. Authentication/Verification
3. Network Security

*Confidentiality*

A single web service may have a series of applications. For example, one major service may combine services from three other applications. In this case, SSL is not enough; messages need to be encrypted in each area of ​​the service path, and each node represents a link that could be weak on the chain. Currently, there is no agreed solution to this problem, but one promising solution is the W3C XML Encryption Standard. This standard provides encryption and decryption for all XML documents or just parts of the XML document.

*Authentication/Verification*

The following options may be considered but there is no clear agreement on a robust verification system.

1. HTTP incorporates built-in basic authentication and Digest, and services can be protected in the same way as HTML documents are currently protected.
2. SOAP Digital Signature (SOAP-DSIG) uses public key cryptography to sign SOAP messages. Allows the client or server to verify the identity of the other party.
3. The Organization for the Advancement of Standardized Information (OASIS) operates in the Language of Security Certification (SAML).

*Network Security*

At the moment there is no easy answer to this problem, and it has been a great debate. In the meantime, if you really want to filter SOAP or XML-RPC messages, then you may have to filter all HTTP POST requests that have set their content type to text/xml.

Another alternative is to filter the SOAP Action HTTP header attribute. Firewall vendors are currently developing tools explicitly designed to filter web service traffic.

**Reference:**

<https://www.tutorialspoint.com/webservices/what_are_web_services.htm>

## DATABASES AND LOGIC

In this section, we are going to take an in-depth look at databases and how logic plays a huge role in how they function. Databases are structures set of data that is stored and accessed electronically. Logic on the other hand is a system or set of principles underlying the arrangements of elements in a computer or electronic device so as to perform a specified task. The two are often intertwined and this creates the strong basis for the database systems that we use today. This is because logic provides a unifying framework in which data management tasks can be studied.

The main uses of logic in databases are;

* Logic is used as a database query language. This enables easier expression of questions asked against the database.
* Logic is used to express integrity constraints in a database

Logic as a database query language.

In this section, we are going to take a look at how logic is used as a database query language and to do this we are going to use the relational database model.

What are the relational databases?

This is a database that is built upon the mathematical concept of relational model of data. In this model, a relation, a table with row and columns, is used to describe and represent data. This relation holds related data fields which a data instance belonging to the relation has.

The relational data model has 2 different query languages and these are;

1. **Relational algebra.**

This is a procedural language in which queries are expressed by applying a sequence of operations to relations.

Relational algebra consists of all expressions obtained by combining five basic operations in syntactically correct ways. These 5 basic operations are;

* Union- Combines the rows of two relations and outputs a new relation that has both input relations’ rows in it.
* Difference- Splits the row of a relation into two different relations.
* Cartesian Product- Taking the cross product of two relations essentially taking every combination of the two tables’ tuples.
* Projection- taking a vertical subset from the columns of a single table that retains the unique rows.
* Selection- When a selection operation is applied to a relation it returns rows in the relation that satisfy the condition.

1. **Relational Calculus.**

This is a declarative language in which queries are expressed as formulas of first order logic.

It comes in 2 different forms which both provide only the description of the query but it does not provide the methods to solve it and these are;

* Tuple relational calculus
* Domain relational calculus

Although they are 2 different query languages, relational algebra and relational calculus have the same equivalence and this means they have very similar expressive power. Expressive power is the ability to represent ideas or problems.

In this case we can see that logic provides a framework in which the database can be queried and information can be drawn from it so that it meets user defined conditions and this is achieved through relational calculus and relational algebra that provides a basis as to which can be used by the database. This means that in the context of an SQL relational database, the SQL used to query the database has to be converted into one of these low-level languages for results to be achieved.

Logic is used to express integrity constraints in a database.

Logic plays an important role in a database so as to ensure that the database is consistent and that the data adheres to set logic rules so that they maintain their consistency and their accuracy. The following are the types of integrity constraints that logic forces upon a database;

1. Entity Integrity

Logic ensures that every relation must have a unique identifier field. This field should be unique to all the records in the relation or those that are to be added on in future in the relation. This is because if there is ever a null value it means that some rows cannot be identified and this will cause a huge issue in the relational data mode. To enforce this, logic dictates that a primary key must be added to a field that is not going to be null and that every record entered into it should be unique.

1. Referential Integrity.

Referential integrity necessitates that a foreign key must have a matching primary key or it must be null. This is dictated by logic so as to ensure that if data in one relation is related or has some relationship to data in another relation then data must be valid. This can also be traced back to the relational model whereby if one relation is to get data from another relation. The initial relations data needs to be verified that it can exist in another table without ruining the accuracy and consistency of the database.

1. Cardinality and Connectivity Constraints

Cardinality defines the relationship between two data tables by expressing the minimum and maximum number of entity occurrences associated with one occurrence of a related entity. Connectivity describes if and how two data tables are going to share entities across each other and what is to be shared. Logic helps introduce and enforce the rules so as to control the number of entity occurrences to only those that follow the set procedures and rules so as to ensure data remains consistent throughout the database and thus ensuring it remains accurate.

**References**

1. Logic and Databases <https://simons.berkeley.edu/sites/default/files/docs/5225/simons162.pdf>
2. IBM Informix Guide to SQL Reference – Manualzz <https://manualzz.com/doc/6983261/ibm-informix-guide-to-sql-reference>
3. Chapter 9 Integrity Rules and Constraints – Database Design <https://opentextbc.ca/dbdesign01/chapter/chapter-9-integrity-rules-and-constraints>

## DATA STREAMS

What is a data stream?

A data stream is a series of objects that is already present, continuous, and ordered (explicitly or implicitly by timestamp or entrance time). Both controlling the arrival order of units and locally capturing the entire stream are impractical. There are vast amounts of data, and things come in quickly.

Data stream types:

*Data stream*

A succession of tuples (potentially unchained) is referred to as a data stream. Comparable to a row in a database table, each tuple was made up of a set of attributes.

*Transactional data stream*

A log interconnection between entities is what is known as a transactional data stream.

Purchases made by consumers from producers using credit cards

Telephone calls placed by callers to the parties on a called number

Clients accessing information from servers via the web.

Measurement data streams

Road traffic IP Network traffic and measurement data streams from Sensor Networks and Sensor Networks are examples of physical natural phenomena.

Temperature and humidity readings from weather stations reflect the environment on Earth.

Examples of Stream Sources

1. *Sensor data* is utilized by navigation systems. Imagine a temperature sensor floating in the water and delivering a readout of the surface temperature to a base station every hour. This sensor produces a stream of real numbers as its output. We need to consider what can be kept going and what can just be archived as we receive 3.5 terabytes of data daily.
2. *Image Data:* Daily down-to-earth streams comprising many terabytes of photos are often sent by satellites. Satellites produce photographs with higher resolution than surveillance cameras, but there might be many more of them, each delivering a stream of images every 1 second.

Data streams' characteristics include:

Streaming data differs from traditional, historical data in a number of ways and comes through sensors, web browsers, and other monitoring systems. These are a few of the main traits of stream data.

*Sensitive to time*

A data stream has a time stamp for each component. The data streams have a limited shelf life and become irrelevant after a while. For instance, to remain relevant, data from a home security system that shows a suspicious movement should be examined and treated quickly.

*Continuous*

Streaming data has no beginning or conclusion. Although data streams are ongoing and occur in real-time, they aren't usually acted upon right away depending on the needs of the system.

*Heterogeneous*

The stream data frequently comes from tens of thousands of distinct sources, some of which may be spread across vast distances. The stream data may be a mixture of several formats due to the diversity of the sources.

*Imperfect*

A data stream may contain data pieces that are missing or corrupted because of the multiplicity of their sources and various data transport technologies. A stream's data components could also arrive out of sequence.

*Unpredictable and Variable*

Repeated transmission of a stream is challenging because real-time data streaming occurs. Retransmission is allowed, however the new data could not be the same as the previous one. The data streams are consequently very unstable.

Data stream applications include

The term "Internet of Things" refers to a vast array of gadgets that use sensors to gather data and send it in real time to a data processor. Stream data is produced by IoT data. Data is generated and streamed in real-time via wearable health monitors like watches, home security systems, traffic monitoring systems, biometric scanners, connected home appliances, cybersecurity, and privacy systems.

Monitors of the stock market in real time: Streams of real-time financial data are frequently transmitted. Organizations can quickly reach important decisions thanks to the processing and analysis of financial data (such as stock prices and market movements).

Activity and transaction logs: A significant source of real-time stream data is the internet. Web browsers create activity records when users access websites or click links. Credit card purchases and other online financial transactions can produce time-critical data that can be streamed and processed for real-time actions.

How stream processing works

You require a procedure that is very different from conventional batch processing in order to process streaming or live data. A stream processor gathers, examines, and presents a steady stream of data. Of course, you must start with data flowing in order to process. The first stage of stream processing is data streaming. Often in real-time, stream processing is utilized to take in the data streams and extract insights from them. Because flowing data is distinct, a stream processor must adhere to the following standards:

*Low Latency*

A stream processor should process continuous data streams fast. Processing speed is a major issue for two causes. One reason is that data enters in a continuous stream, making it impossible for a sluggish processor to catch up on lost information. Second, streaming data quickly becomes irrelevant. Any processing lag decreases the value of the data.

*Scalability*

The volume of streaming data varies from time to time. For instance, sensors may frequently provide small volumes of data, yet sporadically the data may rise. The processor should be able to scale up to handle enormous volumes of data if necessary because the volume of data is unpredictable.

*Availability*

Long downtimes are unaffordable for stream processors. Real-time data is continuously arriving in the stream. A processor must be fault-tolerant, which means it must be able to keep running even if some of its parts malfunction. Additionally, a stream processor must be able to gather, analyze, and send the insights directly to an upper layer for presentation.

Data streams' advantages include:

*Benefits of Data Streaming and Processing*

Data in general can be extremely valuable to organizations. Organizations can gain an advantage by using real-time stream processing techniques to analyze time-sensitive data and respond swiftly to emerging problems. It enables financial institutions to stay current on market movements. Organizations can speed up their response to critical events by using robust visualization systems and a real-time stream processing infrastructure.

*Reduced infrastructure cost*

Data warehouses are frequently used in traditional data processing to store massive amounts of data. The expense of these hardware and storage systems is frequently a strain on the enterprises. Since data isn't kept in massive quantities with stream processing, processing systems require less hardware.

*Cut back on avoidable losses*

Organizations can continuously monitor their business ecosystem thanks to real-time data feeds. They keep businesses informed of potential security breaches, production problems, client complaints, financial collapses, or impending damage to their reputations. Organizations can avert such situations through constant data flow and processing.

*Boost customer satisfaction and competition*

Organizations may proactively address potential problems before they arise with real-time data processing. They gain time and an advantage over rivals as a result. Customer satisfaction is also increased through data processing and streaming since problems with customers can be resolved immediately. The delay brought on by data waiting to be processed in the warehouses is eliminated with continuous, real-time data processing.

Data streams have drawbacks

Systems for processing and streaming data deal with continuous, real-time data that is very variable. Stream data is frequently diverse and lacking. The very nature of stream data presents a number of difficulties for processing and broadcasting it. Including the following drawbacks:

*Data Diversity and Volume*

Massive amounts of real-time, continuous data are dealt with in data streaming. Data streaming frequently faces problems with data loss and corrupted data packets. Since it comes from several applications and geographical places, stream data is frequently heterogeneous. Due to the nature of the data, handling it presents a problem for programmes that stream and process data.

*Timeliness*

Stream data loses its usefulness over time. A system for streaming and processing data must be quick enough to examine data while it is still pertinent. Due to the stream data's time-sensitive nature, a high-performing, fault-tolerant system is required.

*Elasticity*

The amount of stream data is growing daily. The stream processing systems must dynamically adjust to the load in order to maintain a specific degree of service quality. It's possible that stream data providers don't always send large amounts of data. The processing systems must only consume the bare minimum of resources in such situations. The system should dynamically allocate more resources as demand rises. Another difficulty presented by stream processing systems is the requirement for elasticity.

*Mistake Tolerance*

Real-time stream processing occurs continually. It is impossible to replicate or perfectly retransmit the streamed data. As a result, downtime is unaffordable for stream processing systems. The period between data gathering and processing is very short. Systems must operate continuously and be accessible at all times. any component of the system falters, it should not affect the rest of the processing system.

**References**

<https://www.tibco.com/reference-center/what-is-data-streaming>

<https://towardsdatascience.com/tagged/data-streaming>

## BIG DATA

What is big data?

Big data can be defined as a group of structured, semi structured, and unstructured data that is collected by organisaitons which can be mined for information and used in machine learning projects, predictive modelling among other advanced analytical applications.

Big data transfers usually include terabytes, petabytes and hematite data created and collected over time, even if the big data does not match any particular volume of data. Big data can be compared to small data, a term sometimes used to describe data sets that can be easily used for self-help and analysis. The axiom often cited is, "Big data for machines; small data for humans."

Originally identified in 2001 by Doug Laney, Big Data is usually seen with three Vs:

a) *Variety* - a wide variety of data types that are usually stored in large data systems.

b) *Volume* - large amount of data in multiple locations. A large data center does not need to have a large amount of data, but most do because of the nature of the data collected and stored on it. Clickstreams, system logs and streaming systems are among the sources that typically produce large amounts of data continuously.

c) *Velocity* ​​- the speed at which the data is produced, collected and processed. In most cases, big data sets are updated in real time or near real-time, instead of daily, weekly or monthly updates performed at most traditional data storage facilities. Managing data speed is also important as big data analysis progresses to machine learning and artificial intelligence (AI), where analytics processes automatically detect patterns in data and use them to generate data.

Other “V” features include:

a) *Veracity* - the level of accuracy in data sets and how reliable they are. Raw data collected from various sources can cause data quality issues that may be difficult to identify. If left unmanaged, they can be left astray and lose the right path.

b) *Variability* - this applies to large data sets, which may have multiple meanings or formatted differently to different data sources - features that continue to be difficult for big data management and analysis.

c) *Value* - the level of usability of data sets and their value.

What are some examples of big data?

Big data comes from many sources - some examples are:

a) Processing systems, customer databases, documents, emails, medical records, online click logs, mobile applications and social media platforms.

b) Mechanically generated data, such as network log files and server as well as data from sensors in the production machine, industrial equipment and internet of things.

c) In addition to data from internal systems, large data centers typically include external consumer data, financial markets, weather and traffic conditions, location information, scientific research and more.

d) Images, videos and audio files are big data types, too, and many big data applications involve streaming data and continuously collected.

Big data importance:

Big data is important as more companies use it in their plans to provide services, better customer service, improve personal marketing campaigns and jobs that can increase revenue and profit. It gives them a competitive advantage over those who do not use it as they are able to make better and faster business decisions.

Big data can be used to provide valuable information to customers that companies can use to refine their marketing, advertising, and promotions to improve customer engagement and negotiation levels. Real-time data can also be analyzed to assess the emerging interests of customers or buyers of companies, thus enabling the business to respond to customer needs.

Big data is also used by medical researchers to diagnose symptoms and risk factors and physicians to help diagnose diseases and medical conditions in patients. In addition, a combination of data from electronic health records, social networking sites, the web and other sources provides health care organizations and government agencies with up-to-date information on infectious disease threats or outbreaks.

In the energy industry, big data helps oil and gas companies identify potential mines and monitor pipeline performance; similarly, resources use it to track electrical grids.

Financial services firms use big data systems to manage risk and real-time market data analysis.

Manufacturers and transport companies rely on big data to manage their supply chain and improve delivery routes.

Other government functions include emergency response, crime prevention and smart city programs.

Big data covers many types of data, including the following:

a) Organized data, such as transactions and financial records;

b) Informal data, such as text, documents and mixed media files; and

c) Slightly structured data, such as web server logs and streaming data from sensors.

Different types of data may need to be stored and managed together in large data systems. In addition, large data applications typically include multiple data sets that may not be previously integrated. For example, a major data analysis project may attempt to predict product sales by linking previous sales data, returns, online reviews and customer service calls.

How is big data stored and processed?

Big data is usually stored in the data pool. Although data warehouses are usually built into related databases and contain only structured data, data pools can support a variety of data and are usually based on Hadoop collections, cloud storage services, NoSQL databases or other large data platforms.

Many big data areas include multiple systems in a distributed structure; for example, a central data pool may be integrated with other domains, including a related website or data repository. Data on large data systems may be left in its immature state and then sorted and sorted as needed for the application of certain analyzes. In some cases, it's pre-processed through the use of data mining tools and data preparation software so it's ready for applications that are run regularly.

Big data processing places difficult demands on basic computer infrastructure. The required computer power is usually provided by integrated systems that distribute workloads to hundreds or thousands of asset servers, using technologies such as Hadoop and the Spark processing engine.

Finding that type of processing capacity is a challenge. As a result, the cloud is a popular site for large data systems. Organizations can use their cloud-based systems or take advantage of large-scale data delivery from cloud providers. Cloud users can upgrade the required number of servers long enough to complete large data analysis projects. The business only pays for storage and the time it takes, and cloud conditions can be closed until they are needed again.

How big data statistics work

To get valid and relevant results in large data systems, data scientists and other data analysts must have a detailed understanding of available data and a sense of what they are looking for. That makes data editing, which includes profiling, cleaning, validation and modification of data sets, an important first step in the analysis process.

Once the data has been collected and prepared for analysis, various data science and advanced analytics guides can be used to operate different applications, using tools that offer great features for data analysis and power. Those courses include machine learning and its beginnings for in-depth learning, predictable modelling, data mining, mathematical analysis, broadcast analysis, text mining and more.

Using customer data as an example, the various branches of statistics can be made up of big data sets that include the following:

1. *Comparative analysis*. This examines the metrics for customer behavior and customer engagement in real time to compare company products, services and brands with competitors.

2. *Listening to social media*. This analyzes what people are saying on social media about a business or product, which can help identify potential problems with the target audience of marketing campaigns.

3. *Marketing statistics*. This provides information that can be used to improve marketing campaigns and promotional offerings for products, services and business plans.

4. *Emotional analysis*. All customer data can be analyzed to reflect how they feel about the company or product, customer satisfaction levels, potential problems and how customer service can be improved.

Big data management technology

1. Hadoop, a distributed open source analysis framework released in 2006, was originally among the largest data structures. The development of Spark and other search engines has pushed MapReduce, a Hadoop-based engine, to the extreme. The result is an ecosystem of big data technology that can be used for different applications but is often used together.

2. Large data centers and managed services provided by IT vendors integrate most of those technologies into a single package, primarily for cloud use. At present, that includes these donations, listed below:

3. Amazon EMR (formerly Elastic MapReduce)

4. Cloudera Data Platform

5. Google Cloud Dataproc

6. HPE Ezmeral Data Fabric (formerly MapR Data Platform)

7. Microsoft Azure HDInsight

8. For organizations seeking to supply large data systems themselves, either in buildings or in the cloud, the technology available over Hadoop and Spark covers the following categories of tools:

9. Storage storage, such as Hadoop Distributed File System (HDFS) and cloud storage services including Amazon Simple Storage Service (S3), Google Cloud Storage and Azure Blob Storage;

10. Collection management structures, such as Kubernetes, Mesos and YARN, Hadoop's built-in resource manager and operations planner, meaning Yet Another Resource Negotiator but generally known as a summary;

11. Distributed search engines, such as Flink, Hudi, Kafka, Samza, Storm and Spark Streaming and Spark Stable Modules built into Spark;

12. NoSQL database including Cassandra, Couchbase, CouchDB, HBase, MarkLogic Data Hub, MongoDB, Neo4j, Redis and various other technologies;

13. data platform forums, among which are Amazon Redshift, Delta Lake, Google BigQuery, Kylin and Snowflake; and

14. SQL query engines, such as Drill, Hive, Impala, Presto and Trino.

Big data challenges

1. *Design*: In terms of processing power issues, designing a large data architecture is a common challenge for users. Large data systems should be tailored to the specific needs of the organization, the DIY function that requires IT and data management teams to integrate a customized set of technologies and tools.

2. *Deployment*: Developing and managing large data systems also requires new skills compared to those that webmasters and developers focusing on related software typically have. Both of these issues can be identified through cloud management, but IT administrators need to monitor cloud usage to ensure that costs do not go out of hand.

3. *Complexity*: Also, moving local data sets and processing workload to the cloud is often a complicated process. The process of assembling large data sets is often difficult, especially if data variability and speed are factors.

4. *Accessibility*: managing large data systems includes making data accessible to data scientists and analysts, especially in distributed areas that combine a combination of different platforms and data stores. To help analysts find relevant data, data management and statistics teams are increasingly compiling data catalogs that integrate metadata management and data line functions.

The keys to a great effective data strategy

In an organization, building a large data strategy requires understanding the business principles and data currently available for use, as well as an assessment of the need for additional data to help meet the objectives.

The following steps need to be taken to include the following:

a) prioritizing planned use cases and applications;

b) identify new programs and tools required;

c) develop a distribution guide; and

d) assess internal skills to determine whether retraining or recruitment is required.

To ensure that big data sets are clean, consistent and efficient, the data management system and associated data quality management systems should also be a priority. Some of the best practices for managing and analyzing big data include focusing on the business needs of information technology available and using data recognition to assist with data acquisition and analysis.

Procedures and rules for large data collection

As the collection and use of big data increases, the chances of data misuse increase. A public complaint about data breaches and other privacy violations has led the European Union to ratify the General Data Protection Regulation (GDPR), a data privacy law that came into effect in May 2018.

The GDPR restricts the types of data organizations can collect and requires selection. - with the permission of the public or to comply with any other reasons stated for the collection of personal data. It also includes the provision of the right to forget, allowing EU citizens to request companies to delete their data.

To ensure compliance with such laws, businesses need to manage the data collection process carefully. Controls should be set to identify controlled data and prevent unauthorized employees from accessing them.

**Reference:**

Sheldon et al. (2022) , TechTarget: What is Big Data? Retrieved from <https://www.techtarget.com/searchdatamanagement/definition/big-data>

## SOCIAL MINING

Also known as social media data mining, it is used to reveal hidden patterns and trends on social media platforms such as Twitter, LinkedIn, Facebook, Instagram etc. This is usually done by studying machine, mathematical, and mathematical techniques.

How does a data mining mine work?

Public data first needs to be collected and processed. This is publicly available and accessible data, which may include age, gender, race, location, occupation, networks, and more.

Generally, data represents people's attitudes, interactions, behaviors, and feelings about a particular topic, product, or service. Depending on the social network used, this data may include the number of followers, comments, likes, or sharing, if targeted contact data is from Facebook, Twitter retweets or number of comments, or Instagram engagement and hashtag usage. These types of data should be collected in order to develop marketing and engagement strategies.

Once the data has been collected and processed, the following is the use of various data mining techniques that allow for the identification of common patterns and the association of various data points on large databases. Some of these strategies include classification, integration, tracking patterns, forecasting analysis, keyword releases, sentiment analysis, and market or trend analysis. Mining data mining forums uses software solutions to improve the mining process. Advanced data mining software solutions include the following: Microsoft SharePoint, Sisense, IBM Cognos and RapidMiner.

The final step in the mining process is to make a tangible representation of the information obtained throughout the targeted information delivery process. This is usually done using social media statistics or various data viewing tools, such as Infogram, ChartBlocks, Tableau, and Datawrapper, etc.

How is social media used and who uses it?

The mining of social media data has gained widespread use and is increasingly recognized as a valuable asset in many sectors. Although primarily used for business purposes, this procedure is now accepted by researchers and government agencies as well. Businesses purchase data sets from data mining companies to help them customize customer information, develop marketing strategies and service satisfaction, and improve their businesses, in general.

The methods used are:

*Advertising*

Targeted advertising on social media is growing as companies find better ways to identify and target specific audiences. Advertising managers can also use data mining techniques to determine which messages work best between certain groups of people or the best time of day to launch an ad in a particular digital area.

*Influencer Marketing*

Social media data mining tools in identifying promoters with strong fan bases and engagement levels on specific social media platforms. Businesses will use strong marketing to help bring attention to products and services. The promoter may be a high-profile company official, a celebrity, a blogger, or an external product reviewer who can play hot songs and click through a previously unused marketing channel. Careful analysis of public data can help companies find the right promoter to promote their offerings. This helps the business to grow its sales more effectively.

*Market Research*

Companies use social media data mines to find out more about customer preferences, preferences, and bias. For example, an organization may want to explore the statistical features of emerging client groups or determine public sentiment about a particular product, logo or even a politician and religious groups. Social media data mining can be used in a variety of research fields, including social science research, health research, and technical research.

*Public spam detection*

Social media data mining allows easy and quick detection of spam and bots on social media platforms such as Instagram and Twitter.

*Guessing Statistics or Trend Analysis*

Complex algorithms and machine learning strategies can help with predictive modeling techniques, allowing companies to anticipate future trends for customers. Businesses use social media data mining to get important details of current trending keywords, topics, and topics on social media to know what to offer in a timely manner. The information is also useful to the medical community in tracking and predicting the course of an outbreak.

*Government institutions*

The extraction of social media data is increasingly being used by government agencies for the purpose of socially focused interventions. The way this is done is to track the movements of citizens as they record their activities in marked areas throughout the day. Clearly, social media mining can be a powerful tool that can help improve the lives of residents and the safety of communities.

Is social media data mining ethical?

The mining of social media data has raised concerns and concerns about its use for a variety of purposes. One of the major problems with its use lies in the question of whether the social data mining practice is ethical.

As mentioned earlier, social media data mining uses a large amount of user-generated data available publicly, meaning users have decided to publicly share their personal data and are aware of the fact that anyone can view their posts, comments, or shared them on social media. This becomes a problem in distinguishing what is ‘public’ or ‘private’. The difficulty of verifying anonymity, or obtaining permission by selecting information, all poses challenges for data miners.

In addition, the lack of a framework for clear working principles and communication platform data adds to the complexity of the whole process of collecting, analyzing, and visualizing user-generated data.

Recently, several privacy breaches occurred that involved the use of social media data mining due to unclear principles, rules, and regulations on the use of this powerful tool. Large social media companies, including Facebook and Google, have been implicated in disputes over their improper use of social data, and news about Facebook’s Cambridge Analytica scandal in 2018 has been circulated worldwide.

In conclusion, clear ethical principles, and laws regarding the use of social data need to be established so that businesses, researchers, and government agencies can enjoy the benefits offered by social media data mining.

**References**

1. Ausrine (2021, June 22). Social Media Data Mining: What It Is, How It Works, and How to Use It. https://whatagraph.com/blog/articles/social-media-data-mining.
2. Anonymous (2020, April 3). Social Media Data Mining: Understanding What It Is and How Businesses Can Use It. https://www.sandiego.edu/blogs/business/detail.php?\_focus=76022.

## REFERENCES

WEB SERVICES:

1. Tutorials Point (2022). What are Web Services? Retrieved from <https://www.tutorialspoint.com/webservices/what_are_web_services.htm>

DATABASES AND LOGIC

1. Berkeley S. (2022) Logic and Databases. Retrieved from <https://simons.berkeley.edu/sites/default/files/docs/5225/simons162.pdf>
2. IBM Informix Guide to SQL Reference – Manuals. Retrieved from <https://manualzz.com/doc/6983261/ibm-informix-guide-to-sql-reference>
3. Chapter 9 Integrity Rules and Constraints – Database Design. Retrieved from <https://opentextbc.ca/dbdesign01/chapter/chapter-9-integrity-rules-and-constraints>

DATA STREAMS

1. What is data streaming? Retrieved from <https://www.tibco.com/reference-center/what-is-data-streaming>
2. Data Streaming. Retrieved from <https://towardsdatascience.com/tagged/data-streaming>

BIG DATA

1. Sheldon et al. (2022) , TechTarget: What is Big Data? Retrieved from <https://www.techtarget.com/searchdatamanagement/definition/big-data>

SOCIAL MINING

1. Ausrine (2021, June 22). Social Media Data Mining: What It Is, How It Works, and How to Use It. <https://whatagraph.com/blog/articles/social-media-data-mining>
2. Anonymous (2020, April 3). Social Media Data Mining: Understanding What It Is and How Businesses Can Use It. <https://www.sandiego.edu/blogs/business/detail.php?_focus=76022>