BIG DATA ANALYSIS USING IBM CLOUD DATABASE

PHASE 1 SUBMISSION

PROJECT TITLE: Big Data Analysis – Cloud Computing

ABSTRACT:

Platforms like Facebook, Twitter, Instagram, and LinkedIn generate massive amounts of data in the form of posts, comments, likes, shares, and more. Social media data can provide insights into public opinions, trends, and user behaviour. Analysation of those data and provide an improving business model for company and made a user-friendly for user.

PROBLEM STATEMENT:

In the age of digital communication, social media platforms have become rich mediums, reflecting the thoughts, ideas and emotions of millions of users. Analyzing this massive social media data is essential for businesses, governments and organizations to understand public sentiment, identify trends and make informed decisions but the volume, variety and speed of social media data poses significant challenges. Extracting meaningful insights from this data requires advanced big data analytics techniques.

PROBLEM DESCRIPTION:

The aim of this research is to develop a robust big data analytics solution for understanding social media sentiment. The challenges are:

Volume: Social media platforms generate a lot of data every day, including text posts, images, videos and interactions. Effective storage and processing techniques are needed to address this emerging volume of data.

Types: Social media data comes in a variety of formats, including text, multimedia, hashtags, mentions, and emojis. Research must account for this diversity and effectively integrate different types of data.

Speed: Social media data is generated in real time, and requires the ability to process it in real time. Real-time analysis of trends and sentiments is essential for timely decision-making.

Sentiment analysis: There are natural language processing (NLP) techniques for extracting sentiment from text, including sensory polarity analysis, entity recognition, and contextual understanding These techniques should be applied to scale in order to achieve meaningful results.

Data privacy and ethics: Social media data often contains sensitive information. Complying with data protection regulations (such as the GDPR) is paramount to ensure users' privacy when analyzing social media data.

BIG DATA ANALYSIS:

Big data analytics is the use of advanced analytic techniques against very large, diverse big data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

REQUIREMENTS:

1.The platform includes functionality that is designed to help with each of the 5 Vs:

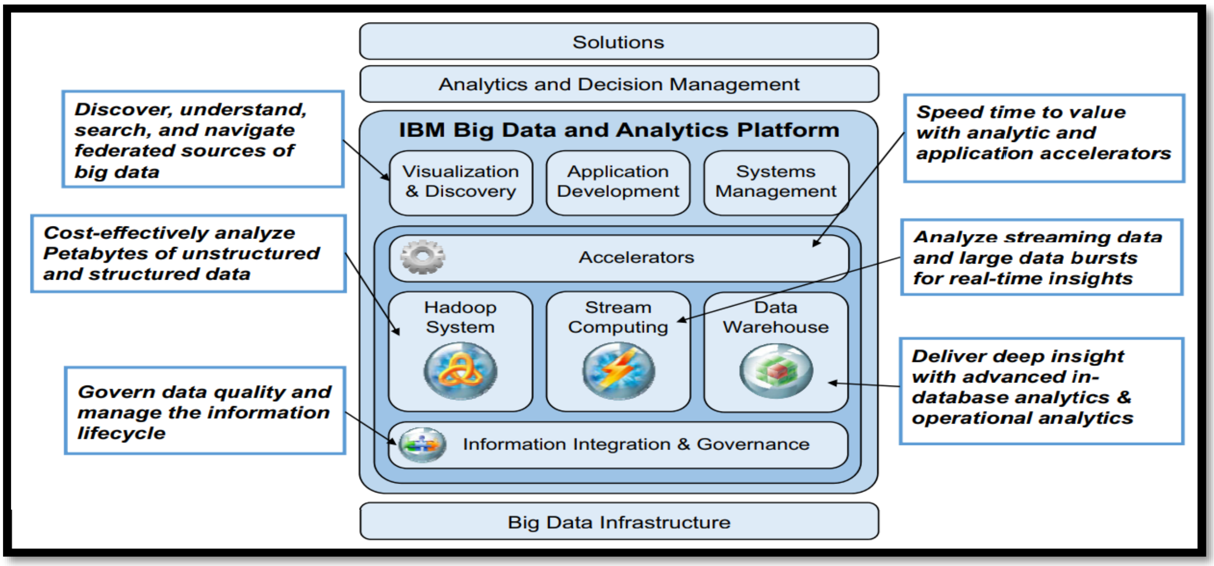
- Variety
- Velocity
- Volume
- Veracity
- Visibility

2.Analytics

3.Ease of use

4.Enterprise-ready

5.Integration

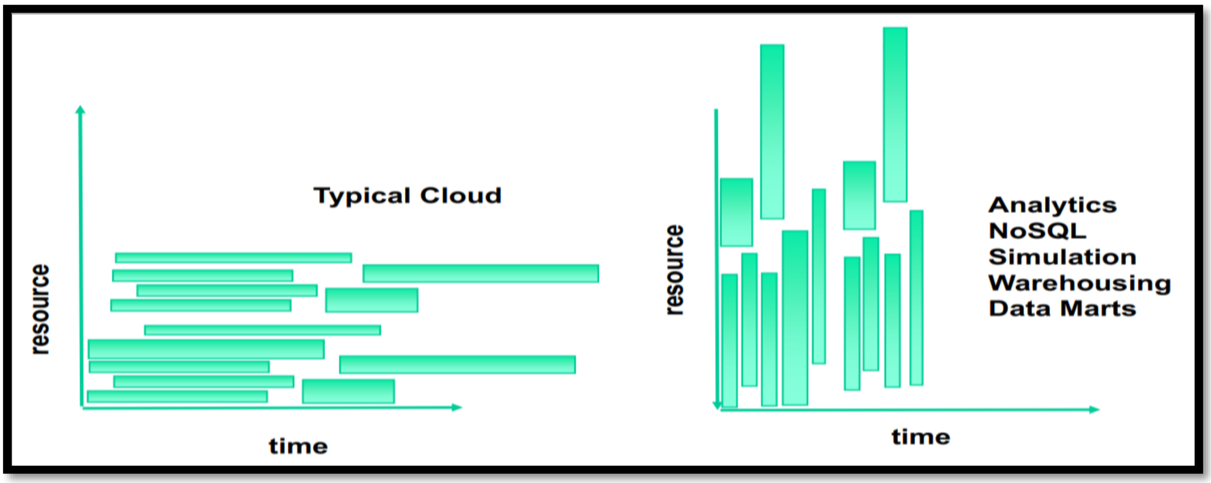

Solutions
Analytics and Decision Management
IBM Big Data and Analytics Platform
Visualization & Discovery
Application Development
Systems Management
Accelerators
Hadoop System
Stream Computing
Data Warehouse
Information Integration & Governance
Big Data Infrastructure
Discover, understand, search, and navigate federated sources of big data
Cost-effectively analyze Petabytes of unstructured and structured data
Govern data quality and manage the information lifecycle
Speed time to value with analytic and application accelerators
Analyze streaming data and large data bursts for real-time insights
Deliver deep insight with advanced in-database analytics & operational analytics

The IBM Big Data and Analytics platform uses the underlying big data infrastructure, which is typically either x86 or Power servers, for running the Hadoop system and Streams components, and data warehousing appliances

The Hadoop system provides a cost-effective way to store large volumes of structured and unstructured data in one place for deep analysis. IBM provides a non-forked, open source Hadoop version and augments it with capabilities, such as enterprise-class storage (by using an IBM General Parallel File System (GPFS™)), security (by reducing the surface area and securing access to administrative interfaces and key Hadoop services), and workload optimization (using the Adaptive Map Reduce algorithm that optimizes execution time of multiple small and large jobs).

TYPICAL CLOUD COMPUTING WORKLOADS VERSUS BIG DATA WORKLOADS:

The graph on the left represents a typical cloud workload. The graph on the right shows a big data workload. The typical cloud uses a few resources over a long period compared to the big data workload that uses many resources over a shorter period.



DESIGN THINKING

1) DATA SELECTION:

Choosing the right data is crucial for social media big data analytics. The selection criteria should be consistent with the research objectives and the specific research questions addressed. Here are the key considerations for data selection in social media big data analytics:

Platform selection: Select social media platforms that align with the research objectives. Platforms cater to different demographics and content. For example, Twitter may be suitable for real-time text analytics, while Instagram may be more suitable for image and video analytics.

Types of data: Social media data comes in various forms such as posts, images, videos, likes, shares, comments, and user profiles. Decide which data types are relevant to the analysis. Textual content is important for sensory analysis, while multimedia content can be important for image analysis tasks.

Timeline: Set a timeline for the analysis. Social media trends are changing rapidly, so timing is important. Real-time analysis should focus on recent events, while historical analysis can span months or years.

Data Volume: Social media platforms generate huge amounts of data on a daily basis. Depending on available computing resources, it decides whether to analyze a data sample, specific subsets, or the entire data set. Sampling methods can be used to control sufficient numbers.

Geography: Consider whether the research should focus on specific regions or countries. Social media platforms can extract data by location, enabling regional or global analysis.

Demographics: Social media users vary in age, gender, interests, and occupation. Understanding population structure

2) DATABASE SETUP:

1. Choose the Database Technology:

a. NoSQL Databases:

Document Stores (e.g., MongoDB): Suitable for storing JSON-like documents. Each social media post, along with metadata, can be stored as a document.

Wide-Column Stores (e.g., Apache Cassandra): Effective for handling large volumes of data with dynamic schema requirements.

Graph Databases (e.g., Neo4j): Ideal for analyzing relationships and networks within social media data.

b. Columnar Databases:

Apache HBase: A distributed, scalable, and big data store that can handle large amounts of sparse data, common in social media datasets.

c. Key-Value Stores:

Redis: Suitable for real-time analytics, caching, and storing key-value pairs like user profiles and their corresponding social media activities.

d. Data Warehouses:

Amazon Redshift, Google BigQuery: Useful for complex queries and aggregations over large datasets. Suitable for business intelligence and deep analytics.

2. Design the Database Schema:

Posts Table: Store individual social media posts with attributes like post ID, content, timestamp, user ID, and platform.

Users Table: Store user profiles with attributes like user ID, name, age, location, and interests.

Interactions Table: Record user interactions (likes, shares, comments) with post IDs, user IDs, interaction type, and timestamps.

Metadata Table: Store additional metadata related to posts, such as hashtags, mentions, and geolocation information.

Sentiment Analysis Table: Store the results of sentiment analysis, including sentiment scores and categories (positive, negative, neutral).

3. Ensure Scalability:

Horizontal Scaling: Distribute data across multiple nodes to handle large volumes. Use sharding and replication for fault tolerance.

Partitioning: Divide large tables into smaller partitions based on certain criteria (e.g., date range) to optimize query performance.

4. Implement Data Processing Pipelines:

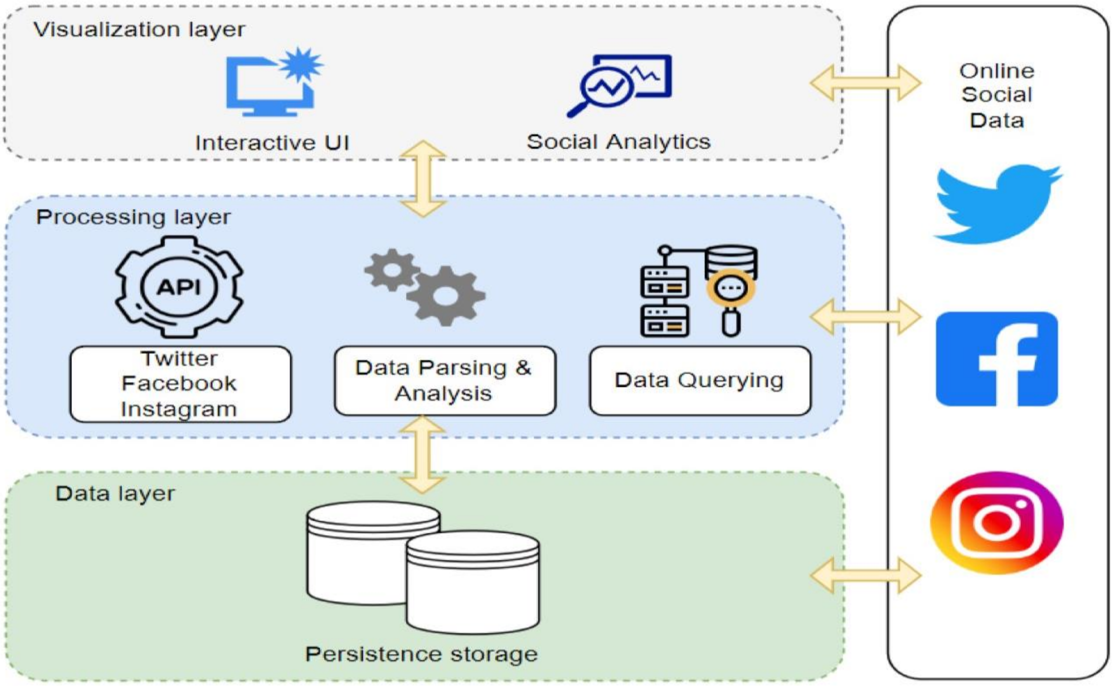
Real-time Data Ingestion: Implement real-time data ingestion pipelines to capture new social media data as it is generated.

Batch Processing: Design batch processing jobs to clean, preprocess, and aggregate historical social media data.

5. Ensure Data Security and Compliance:

Encryption: Implement encryption at rest and in transit to protect sensitive data.

Access Control: Define roles and permissions to restrict access to certain data based on user roles.



6. Backup and Disaster Recovery:

Regular Backups: Schedule regular backups of the database to prevent data loss in case of failures.

Disaster Recovery Plan: Have a disaster recovery plan in place, including failover mechanisms and backup restoration procedures.

3) DATA EXPLORATION:

- Collect social media data.
- Formulate queries to filter relevant information.
- Preprocess and clean the data.
- Extract key features.
- Analyze patterns, sentiment, and networks.
- Use machine learning for advanced analysis.
- Visualize findings for insights and reporting.

4) ANALYSIS TECHNIQUES:

Algorithm used : K – means clustering

- K-Means can be applied to social media data to group similar users or content.
- For instance, it can cluster users based on their behaviour, helping identify different user segments.
- It can also group social media posts or comments with similar content or themes, aiding in content categorization or trend analysis.

5) VISUALIZATION:

Visualization technique used here : Scatter plots

- For visualizing K-means clustering results in social media analysis,we use scatter plots.
- These plots display data points on a 2D or 3D graph, with different colors representing different clusters.
- This visual representation helps you quickly identify cluster separation, outliers, and cluster characteristics, making it easy to interpret your results.

6) BUSINESS INSIGHTS:

Identify Key Insights:

Pinpoint the most significant findings and patterns that emerged from your analysis.

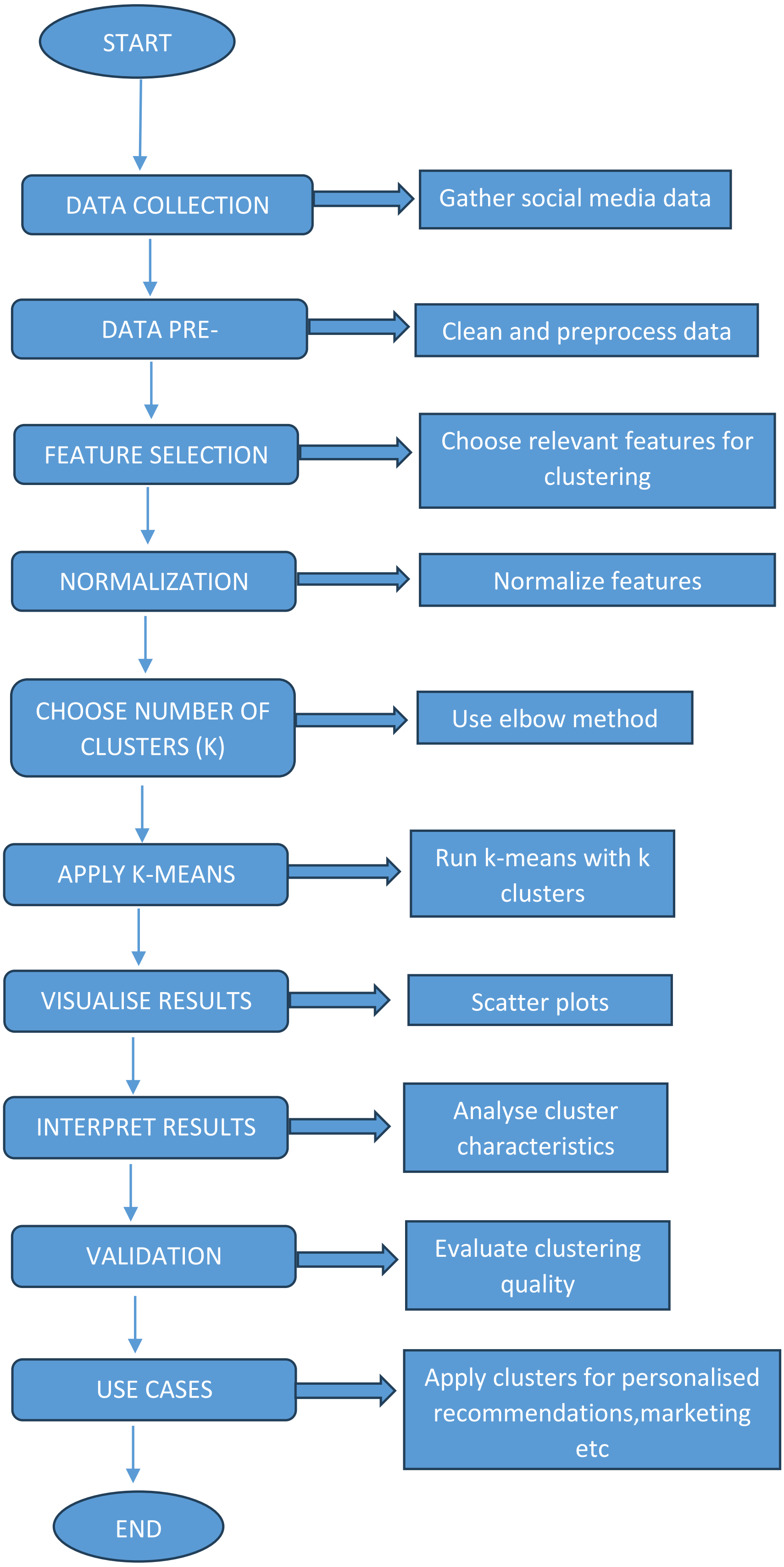
Actionable Recommendations:

Ensure that these recommendations are specific, measurable, and tailored to address the issues or opportunities revealed by the analysis.

Business Impact Assessment:

Consider how these insights may affect your goals, revenue, customer satisfaction, or any other relevant business KPIs.

FLOWCHART:


START
DATA COLLECTION
Gather social media data
DATA PRE-
Clean and preprocess data
FEATURE SELECTION
Choose relevant features for clustering
NORMALIZATION
Normalize features
CHOOSE NUMBER OF CLUSTERS (K)
Use elbow method
APPLY K-MEANS
Run k-means with k clusters
VISUALISE RESULTS
Scatter plots
INTERPRET RESULTS
Analyse cluster characteristics
VALIDATION
Evaluate clustering quality
USE CASES
Apply clusters for personalised recommendations,marketing etc
END

CONCLUSION:

In this project, we embarked on a comprehensive exploration of social media sentiment through the lens of big data analytics. Our goal was to unravel the intricate tapestry of user opinions, emotions, and interactions across various social media platforms. Leveraging advanced technologies, methodologies, and computational tools, we delved deep into the vast expanse of social media data, extracting meaningful insights and patterns that have far-reaching implications for businesses, researchers, and society as a whole.