NoSQL, MongoDB, Apache Hadoop, and Social Media Analytics

Julius Chan

Table of Contents

Table Of Figures

1.0 Part A: NoSQL

1.1 Introduction

NoSQL: The acronym NoSQL was first used in 1998 by Carlo Strozzi while naming his lightweight, open-source “relational” database that did not use SQL. (Keith Foote, 2018). The term NoSQL refers to “Not Only SQL” or “Not SQL”.

NoSQL is a non-relational database designed to manage and store huge volumes of unstructured and semi-structured data that do not require a fixed schema. This is what makes NoSQL the best option for large data.

MongoDB: MongoDB is a NoSQL database that was developed by 10gen, a company founded by Dwight Merriman and Eliot Horowitz in 2007. Today, MongoDB is used by a wide range of companies and organizations, (Tutor Joe, n.d.)

MongoDB is a document-oriented NoSQL database with high scalability, flexibility, and performance. Unlike traditional relational databases, MongoDB stores data in JSON document format. This makes it simple to work with dynamic and unstructured data.

1.1.1 NoSQL Technical Features and Advantages

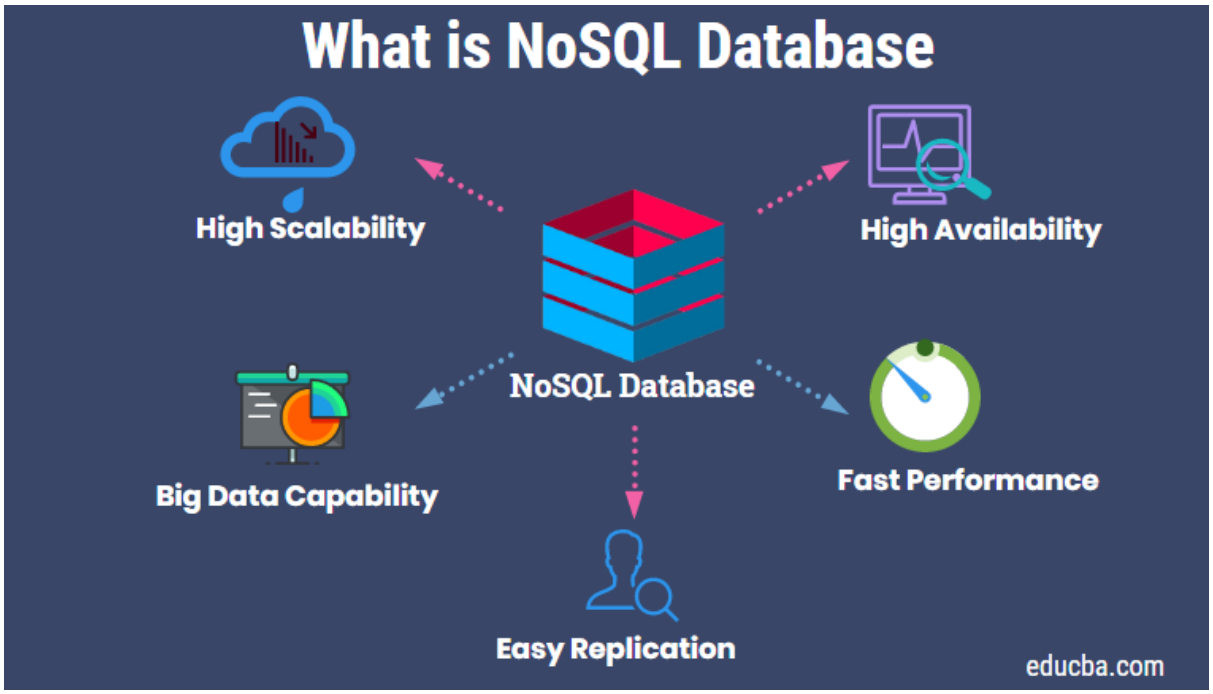

Figure 1: NoSQL Database Features (Ayeshka, (2021) , https://medium.com/introduction-to-application-framework/nosql-bde7d3d2d4ef)

Schema-less Design: Unlike SQL databases, which need a preset schema, NoSQL databases have a schema-free architecture, allowing for flexible and dynamic data structures.

Horizontal Scalability: NoSQL databases scale horizontally by distributing data over numerous servers, allowing them to easily handle enormous amounts of data and heavy traffic, whereas SQL databases generally grow vertically, which can be limiting.

High availability: NoSQL databases include built-in replication and distribution techniques that assure data redundancy and dependability with less manual setup than SQL databases, which frequently require additional settings and maintenance to achieve high availability.

1.1.2 MongoDB Technical Features

Document-Oriented Model: Stores data in BSON format, allowing for embedded documents and arrays.

Indexing: Supports a variety of indexes, including single field, compound, geographic, and text indexes, enabling rapid query processing.

Sharding: Allows for horizontal scaling by distributing data across several servers.

Aggregation Framework: MongoDB provides an Extract, Transform, Load (ETL) architecture that reduces the need for sophisticated data pipelines.

1.1.3 Usage

Internet of Things: Storage and processing of data from a large number of sensors and devices.

Content Management Systems: Managing various forms of material without a strict framework.

Web Applications: Offers a versatile, scalable, and high-performance backend for modern web applications.

1.2 PSB Academy's NoSQL Database

Figure 2: Five collections of PSB Academy Shown in MongoDB (Julius Chan, 2024)

Figure 2 shows the five collections of PSB Academy using the MongoDB database modeling software. The five collections include Departments, Lecturers, Courses, Students, and Contacts.

These five collections were chosen because they reflect the fundamental entities required for simulating the organization and activities of an academic institution such as PSB Academy. Each collection serves an important function in organizing and keeping important data within an educational institution.

1.2.1 Departments

_id: The department's unique identification (created automatically by MongoDB).
department_id : A numerical identification for the department.
department_name: The department's name.

1.2.2 Lecturers

_id: Auto-generated unique lecturer identifier
lecturer_id: Numerical lecturer identifier
name: Lecturer's name
taught_courses: Array of course IDs taught by the lecturer
contacts: Reference to the lecturer's contact information in the Contacts collection.

1.2.3 Courses

_id: Auto-generated unique course identifier
course_id: Numerical course identifier.
course_name: Name of the course.
description: Brief course description.
department_id: Numerical identifier linking the course to a department

1.2.4 Students

_id: Auto-generated unique student identifier
student_id: Numerical student identifier
name: Student's name
enrolled_courses: Array of course IDs the student is enrolled in
contacts: Reference to the student's contact information in the Contacts collection

1.2.5 Contacts

_id: Auto-generated unique contact identifier.
contact_id: Numerical contact identifier.
name: Contact's name.
contact_details: Object with detailed contact information (e.g., phone number, email address).

1.3 MongoDB Aggregation

Figure 3: Input and Output of $lookup aggregation between "Students" and "Contacts" collection. (Julius Chan, 2024)

In Figure 3, the Student's collection is linked to the Contacts collection in a one-to-one relationship through the contacts field. Each student document includes a contacts field that holds the "object_id" of a document from the Contacts collection. This join is carried out via the $lookup aggregate, which incorporates contact information into each student record.

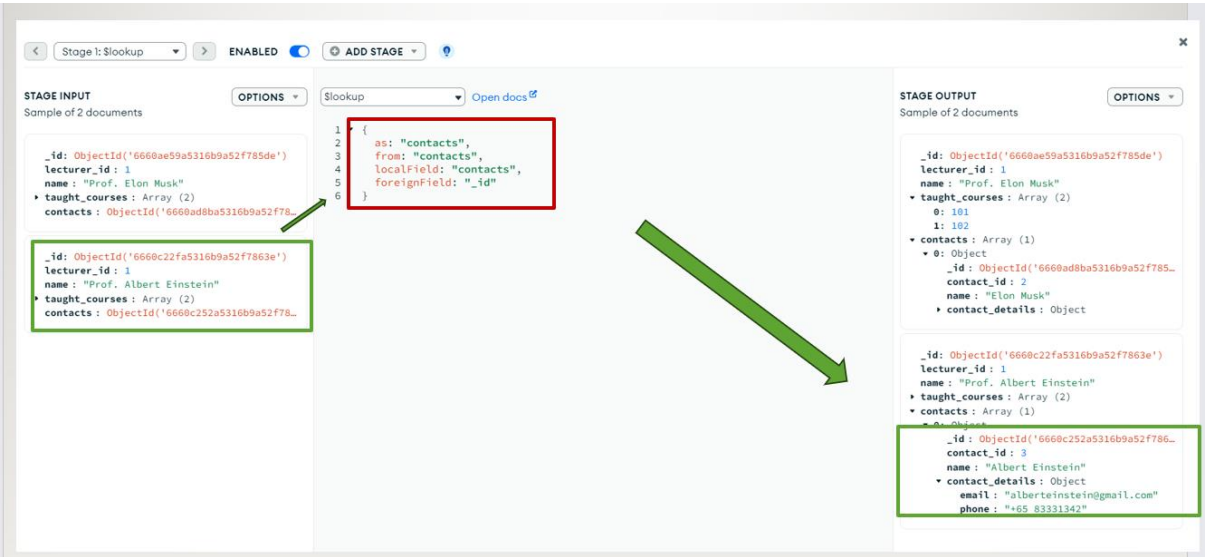

Figure 4: Input and Output of $lookup aggregation between "Lecturers" and "Contacts" collection. (Julius Chan, 2024)

Similarly In Figure 4, the Lecturers collection is linked to the Contacts collection through the contacts' field in a one-to-one relationship. Each lecturer's document

includes a contacts field that holds the “object_id” of a document from the Contacts collection.

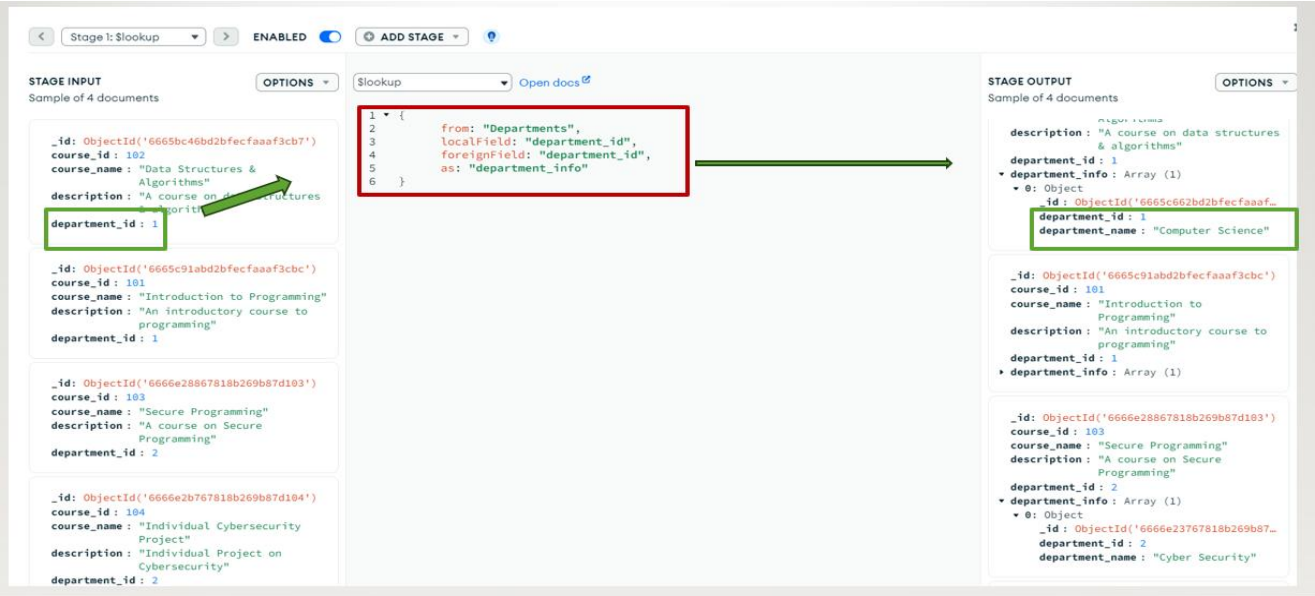

Figure 5: Input and Output of $lookup aggregation between “Courses” and “Departments” collection. (Julius Chan, 2024)

In Figure 5, the “department_id” field establishes a one-to-many relationship between the Courses and Department’s collections. Each course document has a department_id corresponding to the “department_id” in the Department's collection. This join is accomplished using the $lookup aggregate, which contains department information in each course record.

2.0 PART B1: Big Data Analytics using Hadoop

2.1. Introduction

Apache Hadoop started in 2006 as an open-source implementation of Google's file system and MapReduce execution engine. It quickly became a significant part of the Big Data phenomenon. (SJSU, 2019).

Apache Hadoop Ecosystem is a platform that provides a variety of services to help tackle big data challenges. It comprises Apache projects, as well as a variety of tools and solutions. Major Hadoop's components are HDFS, MapReduce, YARN, and Hadoop Common Utilities.

The tools shown in Figure 6 are utilized to assist these major components. All of these tools operate together to provide services such as data acquisition, analysis, storage, and maintenance.

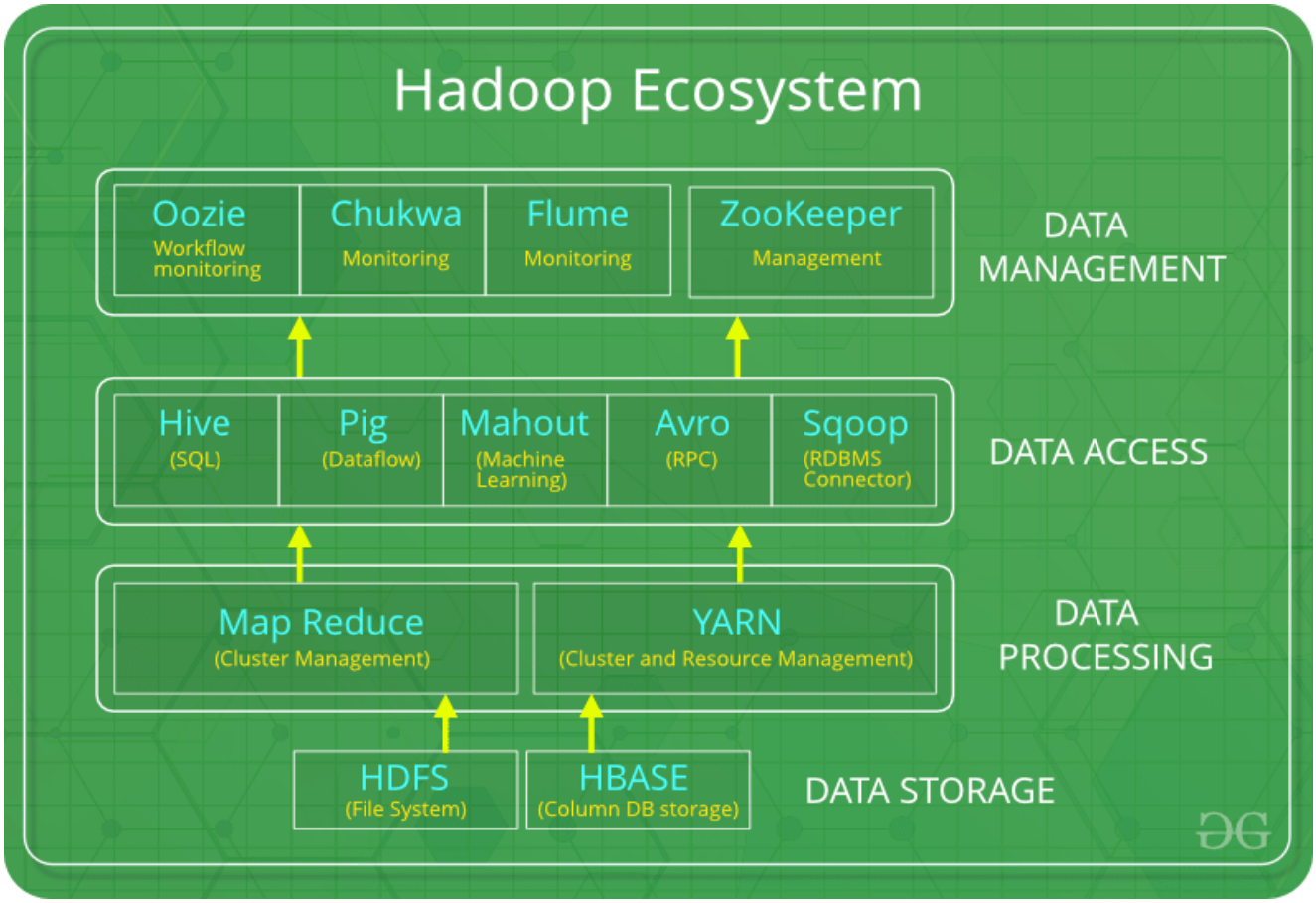

Figure 6: Overview of Hadoop Framework (GFG (2024), https://www.geeksforgeeks.org/hadoop-ecosystem/)

2.2 Apache Sqoop

Developed by Cloudera, Sqoop was later contributed to the Apache Software Foundation (ASF), where it became a top-level project in March 2012. Since then, several major versions have been released, offering improved performance and enhanced stability. (Dremio, n.d.)

In Figure 7, Apache Sqoop is used to export data from the Hadoop file system to relational databases and to import data from relational databases like MySQL and Oracle into the Hadoop ecosystem for additional processing and analysis.

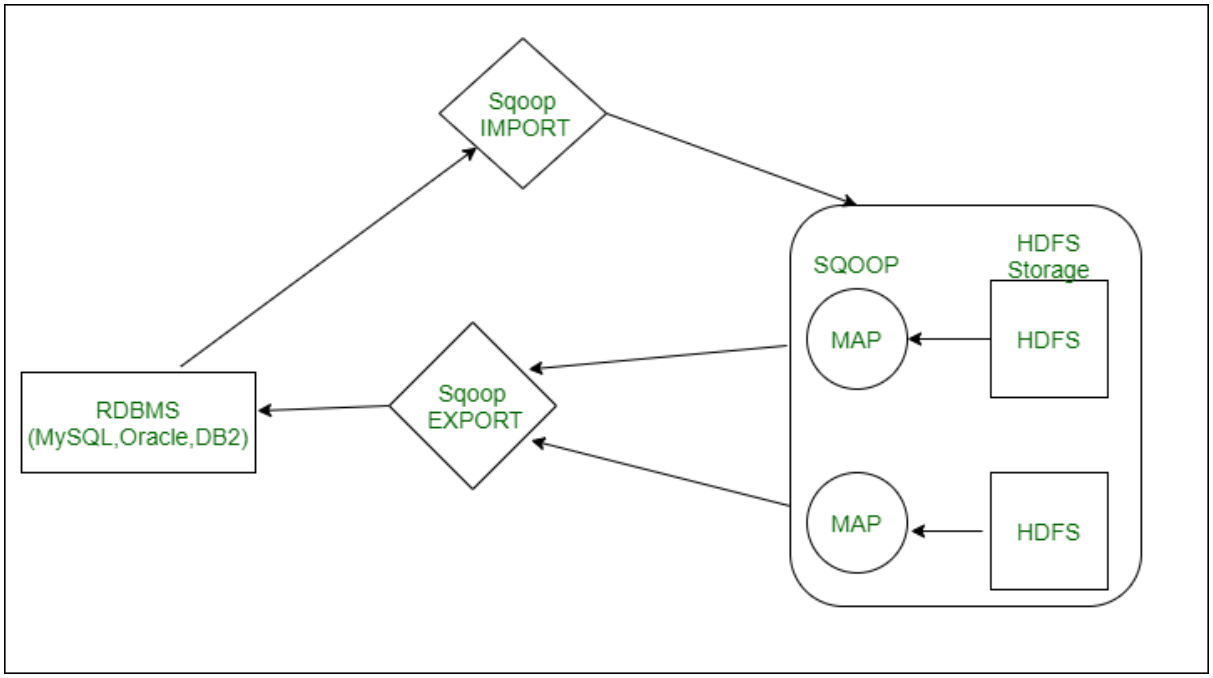

Figure 7: Sqoop in Hadoop Framework (GFG, (2021), https://www.geeksforgeeks.org/overview-of-sqoop-in-hadoop/)

2.2.1 Technical Features and Advantages

Efficient ETL Operations: Sqoop performs ETL operations quickly and cost-effectively.

Parallel Processing: Sqoop's parallel processing increases overall process performance by dividing the burden over numerous threads or nodes.

MapReduce: Sqoop utilizes the MapReduce method for its operations, which also provides recovery and the continuation of the data transmission process without data loss or interruption.

2.2.2 Usage

Data Import into Hadoop: Sqoop allows enterprises to quickly import data from relational databases (e.g., MySQL, Oracle, SQL Server) into the Hadoop environment, namely HDFS (Hadoop Distributed File System).

Integration with Ecosystem: Sqoop works smoothly with other Hadoop ecosystem components like Hive, HBase, and Spark, allowing data engineers to conduct analytics and processing activities on imported data.

2.3 Apache Spark

Apache Spark started as a research project at the UC Berkeley AMPLab in 2009 and was open-sourced in early 2010. Many of the ideas behind the system were presented in various research papers over the years. (Apache Spark, n.d.)

In Figure 8, the Spark Context acts as a hub, coordinating tasks and communicating with the cluster manager, which allocates resources to worker nodes. These nodes run computations on distributed datasets, facilitating efficient large-scale data analytics across the ecosystem.

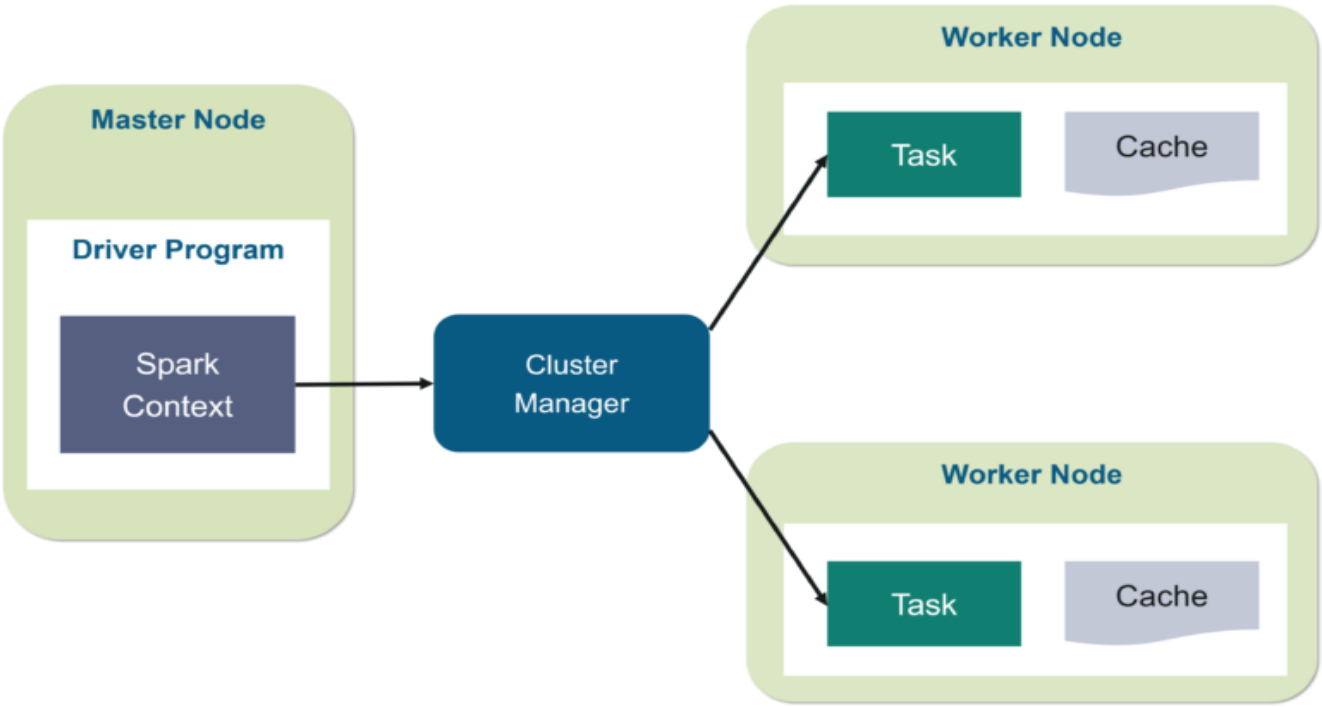

Figure 8: Apache Spark Architecture (Abhijit (2024), https://intellipaat.com/blog/tutorial/spark-tutorial/spark-architecture)

2.3.1 Technical Features and Advantages

Batch Processing: Apache Spark can effectively process massive amounts of data in batch mode, due to its distributed processing capabilities and in-memory computation.

Stream Processing: Allows for real-time stream processing by breaking streaming data into micro-batches and processing them using the same programming model as batch processing.

Graph Processing: Offers a distributed graph processing framework for doing graph analytics and processing large-scale graph data.

2.3.2 Usage

Data processing and transformation: Widely used to process and convert massive amounts of structured and unstructured data. It can effectively perform activities like filtering, aggregating, joining, and sorting.

Machine Learning and Data Mining: A multi-language engine that runs data engineering, data science, and machine learning on single-node computers or clusters.

2.4 Apache Pig

Originally developed by Yahoo Research around 2006, Apache Pig was created to meet the needs of web search platforms that process vast amounts of data. It became part of the Apache Software Foundation in 2007 and made a top-level project in 2008. (Dremio, n.d.)

In Figure 9, Apache Pig streamlines data processing processes by using Pig Latin scripts. Pig translates these scripts into MapReduce tasks, which are subsequently executed on a Hadoop cluster.

This cluster's distributed computing is powered by Hadoop MapReduce. The data is stored and accessed using Hadoop's Distributed File System, allowing for efficient processing of big datasets throughout the system.

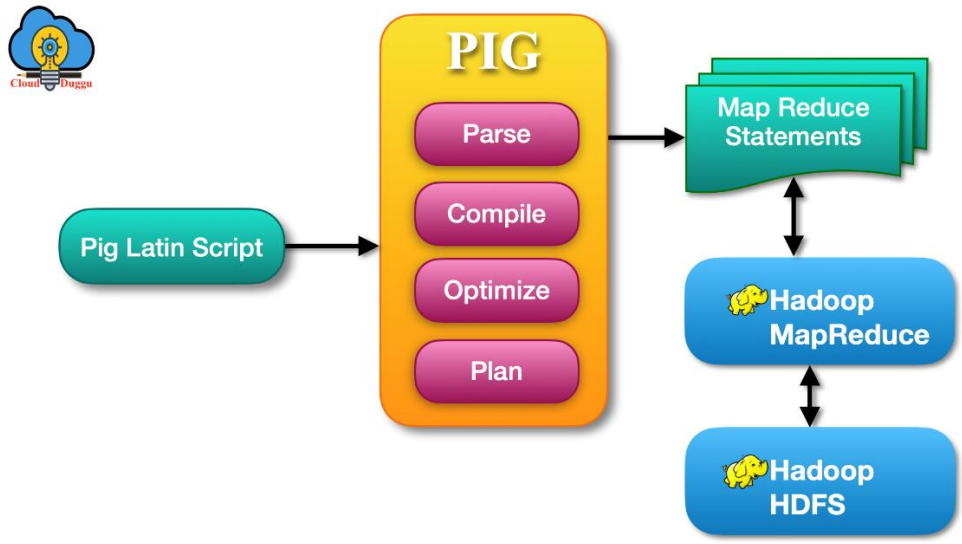

Figure 9: Apache Pig Architecture (Cloudduggu (n.d.), https://www.cloudduggu.com/pig/architecture)

2.4.1 Technical Features and Advantages

Diverse Operators: Apache Pig has a diverse range of operators for conducting various tasks such as filtering, joining, sorting, and aggregation.

Extensions: Apache Pig is extensible, allowing you to create your own processes and user-defined functions (UDFs) in Python, Java, or another computer language.

Simplicity: Apache Pig makes joining simple, and it requires fewer lines of code.

2.4.1 Usage

Data Transformation: It turns raw data into structured representations using a large number of operators, making it easier to examine.

Apache Hadoop Ecosystem: Apache Pig integrates with other Apache Hadoop ecosystem components allowing users to leverage these components' capabilities while transforming data.

2.5 Apache Oozie

Yahoo developed Apache Oozie! and later made accessible to the wider Hadoop community through the Apache Software Foundation. The project became a top-

level Apache project in 2012 and has seen several major versions since its inception. (Dremio, n.d.)

In Figure 10, Apache Oozie manages Hadoop processes using three important components: the Oozie Client, Server, and Database. To connect with the Server, the Oozie Client makes use of command-line tools, Java APIs, and HTTP REST APIs.

The server retrieves and updates task states from an SQL database. This configuration enables Oozie to effectively automate and manage complicated data processing operations inside the Hadoop environment.

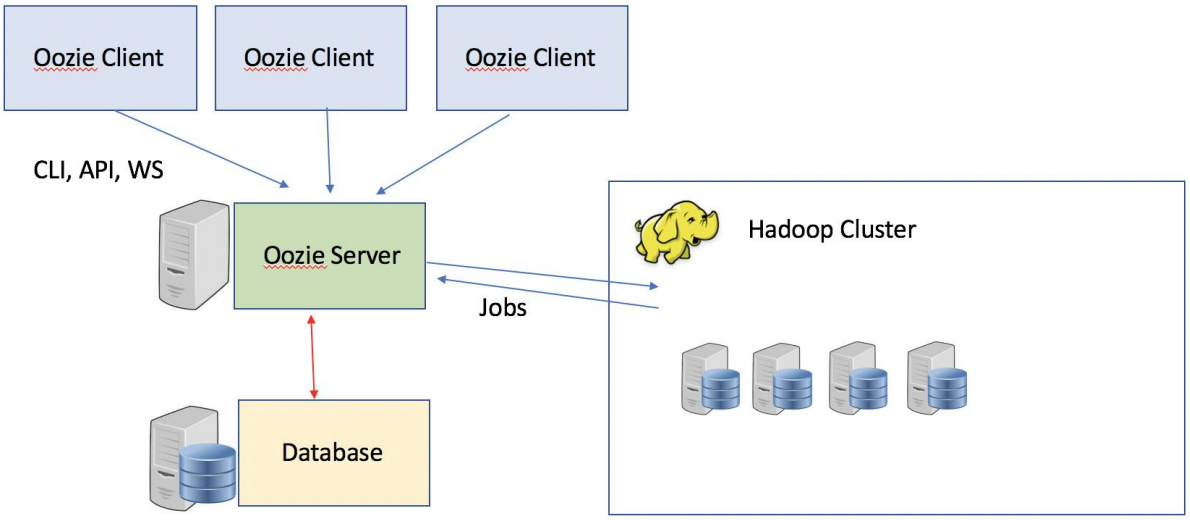

Figure 10: Apache Oozie Architecture (Devstacks (n.d.), https://devstacks.wordpress.com/2017/02/16/oozie-architecture-and-job-scheduling/)

2.5.1 Technical Features and Advantages

Workflow Management: Oozie enables users to create complicated workflows with various Hadoop tasks including MapReduce, Hive, Pig, and Sqoop.

Scheduling and Coordination: Oozie's coordination features enable it to activate workflows in response to specified situations, such as data availability in HDFS, making it perfect for time-sensitive data processing jobs.

Web Console and APIs: Oozie provides a web-based console and REST APIs for task submission, monitoring, and control, making it simple to combine with other systems and interfaces.

2.5.2 Usage

Workflow Orchestration: Oozie is designed to manage complicated processes with several interdependent jobs.

Hadoop Ecosystem: Oozie integrates with Hadoop's many components, such as HDFS, YARN, MapReduce, Hive, Pig, and Sqoop. This strong connectivity makes it easier to manage Hadoop tasks.

2.5 Apache Hive

Built originally by Facebook, Apache Hive evolved as an Apache open-source project in 2010. It was designed to enable easy data summarization, ad-hoc querying, and the analysis of large volumes of data. (Dremio, n.d.)

In Figure 11, Apache Hive shows how it effectively answers user queries, including its interaction with Hadoop and its intentional separation of metadata storage. It highlights the importance of separating metadata storage (Metastore) and data storage (HDFS) for maximum speed and scalability.

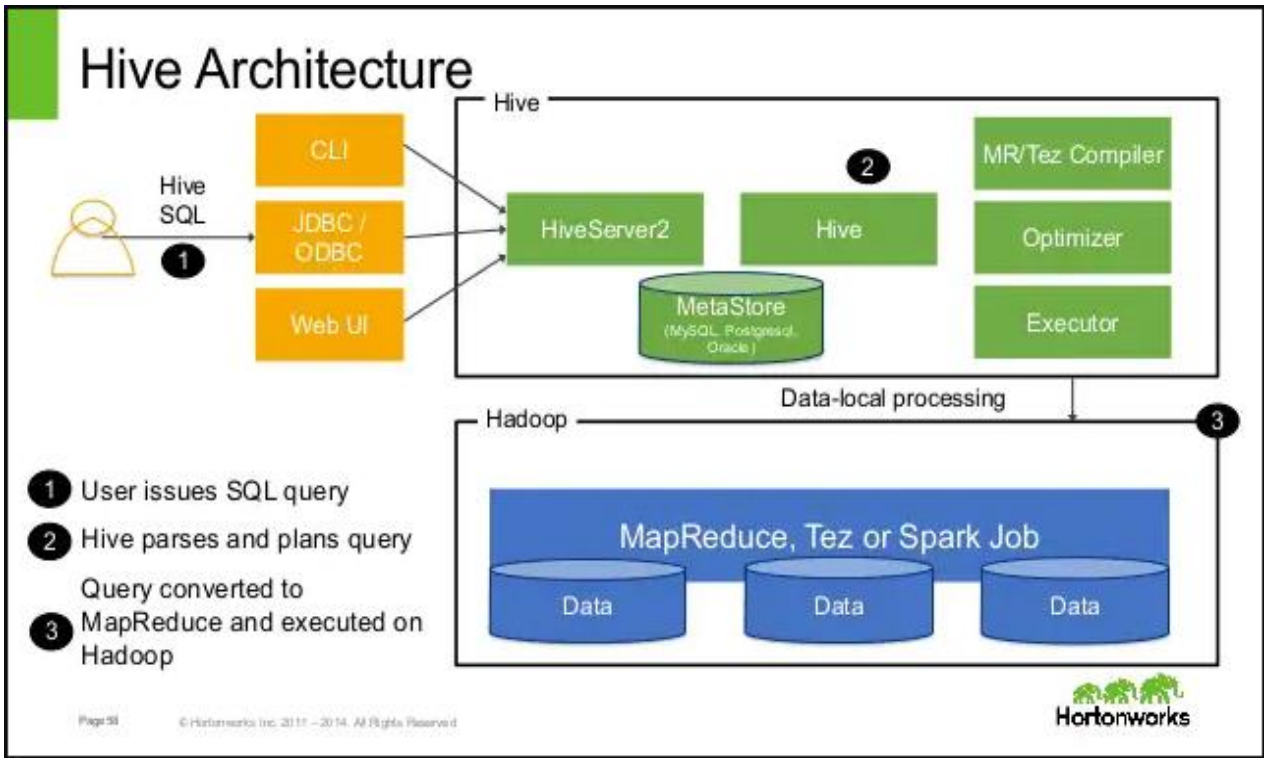

Figure 11: Apache Hive Architecture (Dinesh Gangwar (2021), https://www.linkedin.com/pulse/hive-story-history-architecture-dinesh-gangwar)

2.5.1 Technical Features and Advantages

Efficient Indexing and Metadata Storage: Apache Hive offers efficient indexing and metadata storage, including bitmap indexes introduced in version 0.10 to improve query performance. It saves metadata in an RDBMS, which reduces semantic check time during query execution.

Data Processing and Scalability: Hive's built-in User-Defined Functions (UDFs) provide advanced data processing and scalability, making it ideal for enterprise-scale applications.

User-Friendly Interface: Apache Hive's SQL-like language, HiveQL, provides a familiar interface to SQL users.

2.5.2 Usage

Data Warehousing: Used to create data warehouses and analytical databases, which allows companies to efficiently store, handle, and analyze enormous amounts of structured and semi-structured data.

Hadoop Ecosystem: Hive works well with the Hadoop ecosystem, allowing users to process data with Hadoop tools such as Pig, MapReduce, and Spark.

3.0 Part B2: Social Media Analytics

3.1 Introduction

Social media has become vital to everyday life, providing entertainment and information-sharing opportunities. Businesses increasingly depend on social media for growth and marketing, but organizing and tracking data across numerous platforms may be challenging. Social media analytics platforms solve this issue by including functionality for tracking reach, engagement, and discussions.

They give significant insights into audience interests and behaviors, allowing businesses to improve their social media strategy and drive growth more effectively.

Figure 12: Important KPIs of Social Media Analytics (Emplifi (n.d.), https://emplifi.io/resources/blog/social-media-analytics-the-complete-guide)

3.2 Hootsuite

Hootsuite is a social media management tool founded by Ryan Holmes in 2008. The system's user interface is a dashboard with social network integration for Twitter, Facebook, Instagram, LinkedIn, Pinterest, YouTube, and TikTok.

Hootsuite main function is to enable businesses to better manage their social channels, engage new audiences, and expand their brand, ensuring they reach their target audience efficiently and communicate with them in real-time.

One of Hootsuite's unique advantages is its wide app ecosystem, which permits interaction with over 150 applications, allowing businesses to personalize their dashboard with tools and capabilities customized to their needs.

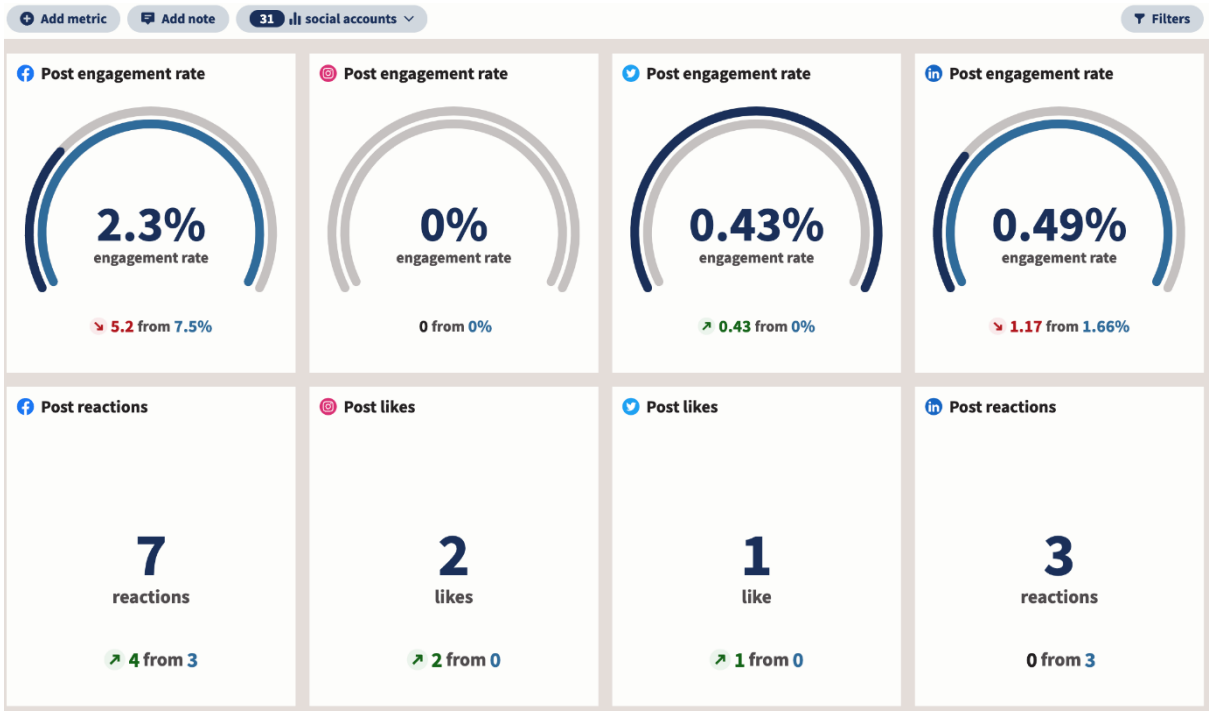

Figure 13: Overview of Analytics of Social Media Channels on Hootsuite (Stanley McLachlan (2023), https://blog.hootsuite.com/social-media-analytics-healthcare)

3.2.1 Technical Features and Advantages

Dashboard Management: Integrates with many social media sites for centralized administration and detailed data. This centralized administration offers users extensive data and analytics, simplifying social media management and increasing productivity.

Analytics Reports: Provides unique reports to monitor social media activity, allowing for strategic insights and optimization which provide strategic insights for enhancing social media strategy and campaigns.

Engagement Monitoring: Monitors multiple social media discussions and trends to enable real-time involvement and connection development. Users may quickly communicate with their audience, encouraging relationship growth.

3.2.2 Industry Application – Healthcare

Hootsuite is one of the most trusted social media management tools for healthcare. It connects with patients and consumers by providing an engaging and compliant social media experience, which can monitor patient topics on social media, health campaign engagement, and uncover return on investment (ROI). This data may then be utilized to personalize marketing campaigns, improve customer service, and boost brand reputation.

3.3 Buffer

Buffer began as a startup project in 2010 and was founded by Joel Gascoigne. It began as a tool for scheduling tweets and eventually grew to handle various social media networks.

Buffer is a social media management and marketing software, it’s main functions include drafting and scheduling posts for distribution on multiple social networks including Facebook, Instagram, Twitter, LinkedIn, and Pinterest.

Its unique ability is to identify the best posting times and give insights into content performance, allowing businesses to optimize the impact of their social media efforts.

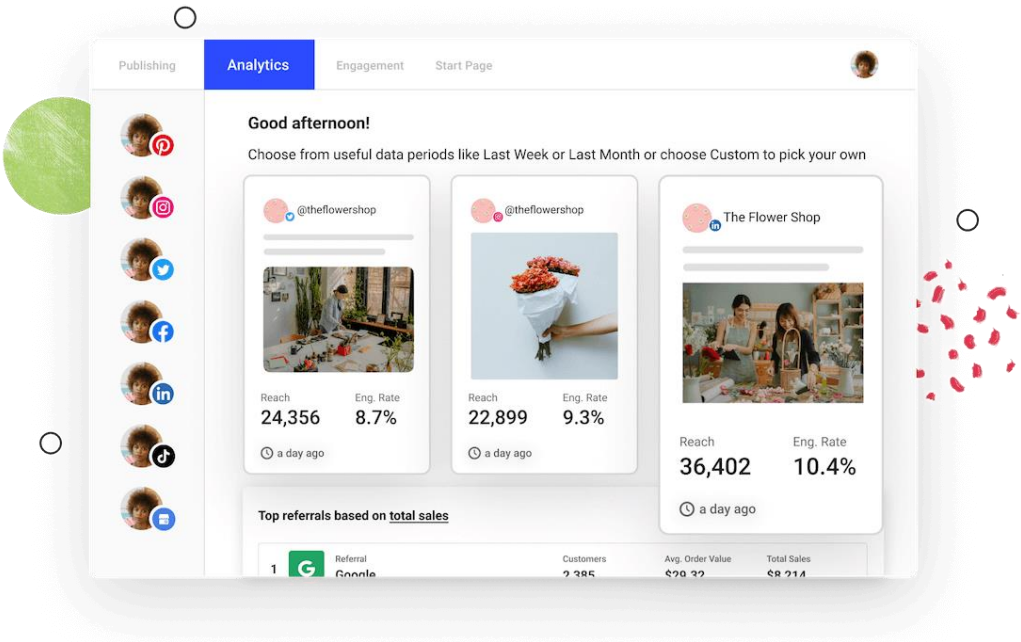

Figure 14: Post Engagement Insights (Buffer (n.d.), https://buffer.com/analyze)

3.3.1 Technical Features and Advantages

Content Scheduling: Allows users to schedule posts across many social media networks, assuring consistent and timely content delivery.

Analytics: Provides performance data and analytics for social media posts, allowing you to optimize your content strategy. A full list of all published posts may be filtered and sorted using various metrics.

Browser Extension Has an extension for browsers that allows users to create new posts from any webpage. Users can continue to use social media without interruption in their work.

3.3.2 Industry Application – Education

Buffer's role in education is important which educational institutions to effortlessly plan and publish information, promoting academic programs, events, and student successes while keeping a constant publishing schedule. This heightened visibility and engagement can translate to higher enrolment rates, improved student satisfaction, and enhanced alumni support all contributing to the institution's commercial success.

3.4 Keyhole

Saif Ajani and Minaz Abdulla co-founded Keyhole in 2013. It is primarily utilized by businesses, marketers, and journalists throughout the world to exchange and analyze data on its platform.

The main function of Keyhole is real-time social media analytics. This entails watching and analyzing social media posts, mentions, hashtags, and accounts in real-time across social-media platforms insights about engagement, reach, and sentiment.

Its unique feature provides real-time tracking and precise performance information for hashtags and keywords, allowing users to make rapid decisions during live events and time-sensitive campaigns.

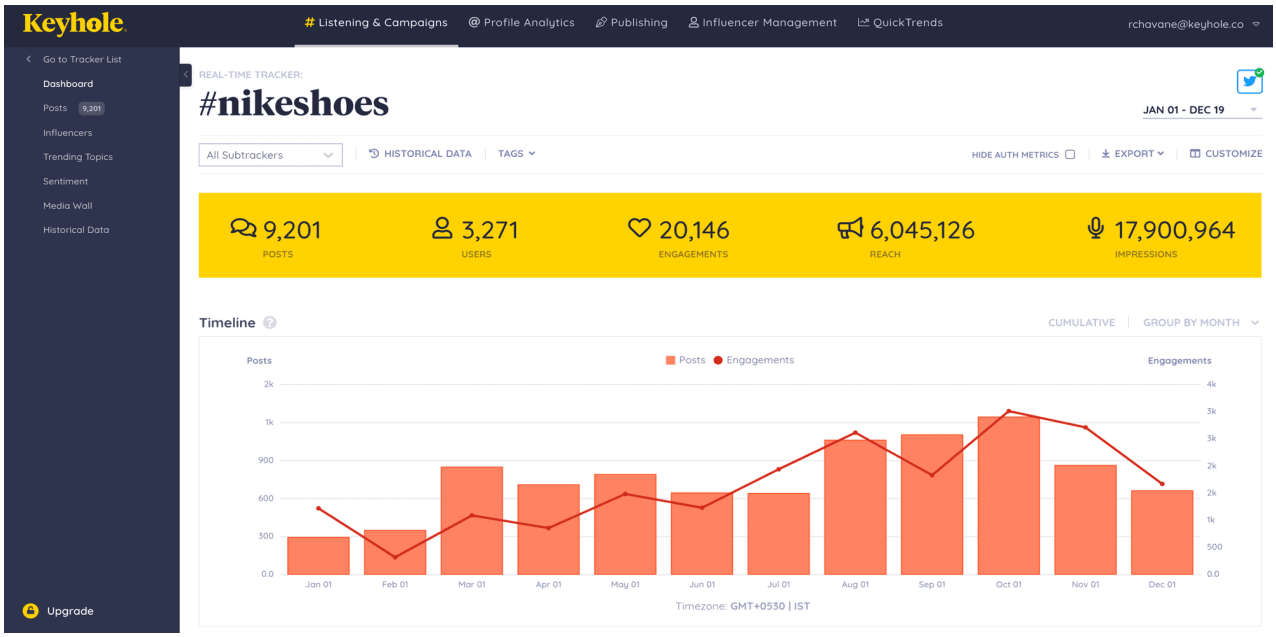

Figure 15: Real-Time Analytics on Hashtags (Keyhole (n.d.), https://keyhole.co/index-2)

3.4.1 Technical Features and Advantages

Hashtag and Keyword Tracking: This feature tracks posts, users, reach, and impressions using hashtags or keywords, allowing for more effective campaign management and optimization.

Influencer Identification: Identifies the most prominent people discussing a specific phrase, allowing for targeted influencer engagement and collaborations.

Advanced search features: The Search Feature identifies and monitors specified keywords, hashtags, @usernames, or URLs to provide precise and relevant insights.

3.4.2 Industry Application – Event Marketing

Keyhole allows organizers and marketers to track the performance of event-related hashtags and keywords in real-time. By monitoring social media discussions and mood, event marketers may gauge engagement and evaluate the effectiveness of promotional activities. These data enable timely adjustments to marketing campaigns, resulting in increased sales and sponsorship possibilities.

3.5 Sprout Social

Sprout Social was founded in 2010 by Justyn Howard, Aaron Rankin, Gil Lara, and Peter Soung. Its platform provides quantitative and contextual data for all social accounts, including Facebook, Twitter, Instagram, LinkedIn, and Google+.

Sprout's main function is to monitor social media conversations, engage with audiences, and analyse success data so that the company can strategically use social media to maximize future commercial impact.

The unique feature of social listening is what sets Sprout Social apart. It goes beyond simple monitoring to evaluate consumer discussions and sentiment, trends, and brand mentions in real-time.

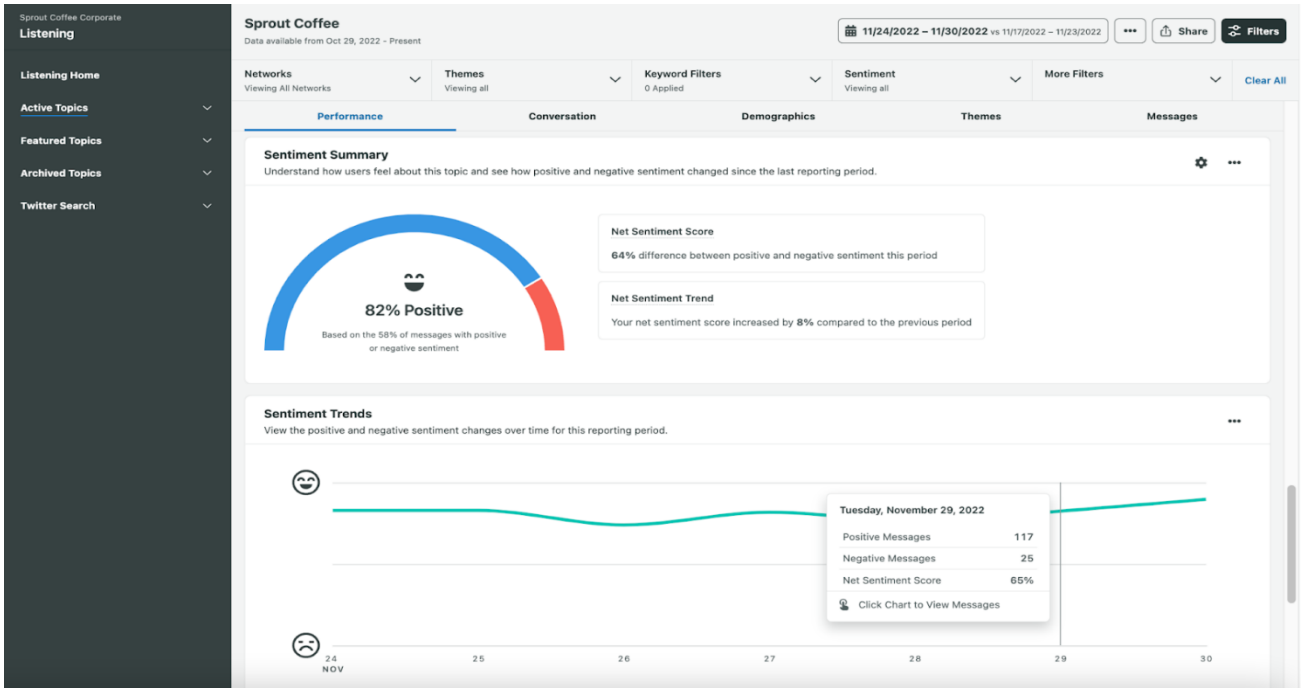

Figure 16: Sprout Social's Sentiment Analysis (Sprout Social (n.d.), https://sproutsocial.com/insights/social-media-listening)

3.5.1 Technical Features and Advantages

Unified Social Inbox: Combines messages, comments, and mentions from many social media sites into a single inbox, increasing productivity by centralizing communication channels and allowing for faster response times.

Audience Engagement: Automated answers facilitate audience involvement, improving customer service by allowing for rapid responses.

Statistics and Analytics: Provides thorough statistics and customizable reports for assessing social media success and tracking important KPIs.

3.5.2 Industry Application – Restaurant Businesses

Understanding customer sentiment and maintaining an online reputation is crucial in the food industry. Sprout Social's social listening features enable businesses to track conversations about their brand, menu items, and dining experiences in real-time. They can spot patterns, assess customer happiness, and respond to negative comments quickly to guarantee customer success.

3.6 Mention

Mention is a social media monitoring and brand management platform created in 2012 by Edouard de La Jonquière and Thibaud Elzière. It enables users to track mentions of their brands across many platforms and industry keywords in real-time.

Mention's main function is to enable real-time monitoring of brand mentions, keywords, and hashtags across several online channels, such as social networking platforms, news sites, blogs, forums, and review websites.

Its unique ability is well-known for its 'alerts' function, which provides real-time messages anytime specific terms are spoken online. This proactive monitoring enables organizations to remain ahead of the curve, respond quickly to new trends, and communicate with their audience effectively.

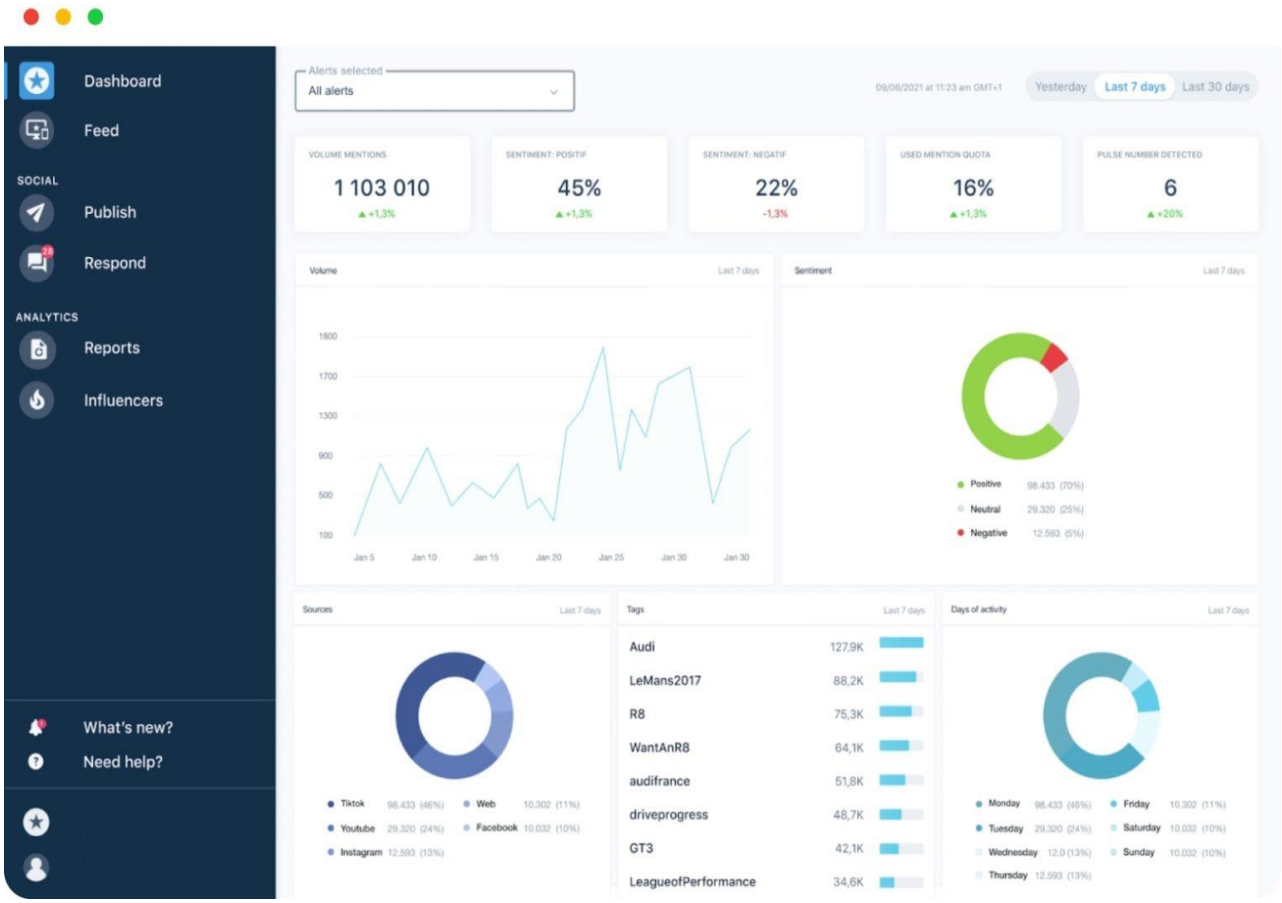

Figure 17: Alert Analytics and Notification on Mention (Mention (n.d.), https://mention.com/en/blog/mention-dashboard-analytics-hub)

3.6.1 Technical Features and Advantages

Real-Time Monitoring: Provides real-time tracking of brand mentions, keywords, and hashtags across many internet platforms. This enables businesses to keep current on important topics and respond quickly to client input.

Competitor Analysis: Allows organizations to examine their competitors' online presence, including brand mentions, social media activity, and audience engagement to measure their performance against industry peers.

Customizable alerts: Provides customized notifications that warn companies when their brand is discussed online. These notifications may be adjusted based on keywords or emotions, allowing organizations to monitor important conversations and reply in real-time.

3.6.2 Industry Application – Customer Service

Mention's real-time monitoring and alerting features make it helpful in businesses requiring consumer feedback and reputation management, such as hospitality, retail, and healthcare. Businesses may improve customer satisfaction and loyalty by responding quickly to queries, complaints, and comments.

4.0 List of References

Dremio. (n.d.). Apache Oozie. Dremio. https://www.dremio.com/wiki/apache-oozie/

Dremio. (n.d.). Apache Pig. Dremio. https://www.dremio.com/wiki/apache-pig/

Dremio. (n.d.). Apache Sqoop. Dremio. https://www.dremio.com/wiki/apache-sqoop/#:~:text=History,additional%20functionalities%2C%20and%20enhanced%20stability.

GeeksforGeeks. (2024). How to Use Buffer to Supercharge Your Social Media Results. GeeksforGeeks. https://www.geeksforgeeks.org/how-to-use-buffer-to-supercharge-your-social-media-results/

GeeksforGeeks. (2024). Introduction to NoSQL. GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-nosql/

GeeksforGeeks. (2020). Overview of Apache Spark. GeeksforGeeks. https://www.geeksforgeeks.org/overview-of-apache-spark/

GeeksforGeeks. (2021) Overview of Sqoop in Hadoop. GeeksforGeeks. https://www.geeksforgeeks.org/overview-of-sqoop-in-hadoop/

Hootsuite Education. (n.d.). Dashboard Overview. Hootsuite. https://education.hootsuite.com/pages/dashboard-overview

Hootsuite Help Center. (2024). What does Hootsuite do? Hootsuite. https://help.hootsuite.com/hc/en-us/articles/6024181624603-What-does-Hootsuite-do

Intellipaat. (2024). Spark Architecture. Intellipaat. https://intellipaat.com/blog/tutorial/spark-tutorial/spark-architecture/

Keyhole. (n.d.). Press. Keyhole. https://keyhole.co/press

Mention. (2022). The Mention Dashboard: Your Analytics Hub. Mention. https://mention.com/en/blog/mention-dashboard-analytics-hub/

MongoDB. (n.d.). NoSQL Explained. MongoDB. https://www.mongodb.com/zh-cn/resources/basics/databases/nosql-explained

San Jose State University Department of Computer Science. (2019, February 27). The Impact of Big Data: A Discussion with Steve Gates. https://www.sjsu.edu/cs/community-and-events/cs-talks/spring-

2019/gates.php#:~:text=Apache%20Hadoop%20started%20in%202006,of%20the%20Big%20Data%20phenomenon.

Simplilearn. (2023). What is MongoDB? Simplilearn. https://www.simplilearn.com/tutorials/mongodb-tutorial/what-is-mongodb

Skelton, J. (2018). A Brief History of Non-Relational Databases. DATAVERSITY. https://www.dataversity.net/a-brief-history-of-non-relational-databases/

Singh, A. (2024). Introduction to NoSQL. GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-nosql/

Shrivastava, R. (n.d.). History of MongoDB. Tutor Joes. https://www.tutorjoes.in/mongodb_tutorial/history_of_mongodb

Sprout Social. (n.d.). About Sprout Social. Sprout Social. https://sproutsocial.com/about/

Sprout Social. (n.d.). Social Media Listening. Sprout Social. https://sproutsocial.com/insights/social-media-listening/