PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY MOHAMED BOUDIAF - M'SILA

FACULTY: Mathematics and Computer Science

DEPARTMENT: Computer Science

N°: ……

DOMAIN: Mathematics and Computer Science

BRANCH: Computer Science

OPTION: SI-GL

Dissertation submitted to obtain Master degree

By: Mr. FODIL Youssouf Islam & Mr. MOKRAN Abdelrrahim

SUBJECT

Real-time data Analytics Apache Druid

Supported before the jury composed of:

……	University of M'sila	President
Mr. Hichem Debbi	University of M'sila	Supervisor
……	University of M'sila	Examiner
……	University of M'sila	Examiner

Academic year: 2019/2020

ACKNOWLEDGMENT

First and foremost, heartfelt gratitude and praises go to the Almighty Allah who guided me through and through.
This work could not have reached fruition without the unflagging assistance and participation of so many people whom I would never thank enough for the huge contribution that made this work what it is now.
I would like to thank profoundly Mr Hichem Debbi for his scientific guidance and corrections, suggestions and advice, pertinent criticism and pragmatism, and particularly for his hard work and patience. I am very grateful to him, for without his help an important part of this work would have been missed.
I would also like to thank all the Jury Members, who have agreed to review this work.
I thank all the teachers who guided us throughout our journey.
Thanks to all who helped me
Thanks

DEDICATION

This dissertation is dedicated to My parents and for their encouragement, prayers, motivations and being there.
, all of my Family members near or far.
All of my friends and colleagues.

CONTENT

Contents

TABLE OF FIGURES

TABLE OF FIGURES

GENERAL INTRODUCTION

With over 1.5 billion users worldwide, social media is a gold mine of information in the form of real-time, interactive communications offered through tweets, blogs, status, images, and videos. Not surprisingly, organizations are relying more and more on social media to understand and work more effectively and responsively with customers, employees, and vendors, and even analyze competitors. However, mining and analyzing the immense volumes of structured and unstructured data generated by social media in real-time is no easy task.

This work is structured like the following:

In the first chapter: we start by an introduction to the data world and how the Big data term was created, we define its main characteristics, problems, its application in many domains like the industry and marketing, and finally how we can use it and bring forth some tools to do it [1] .

In the second chapter: we define the real-time analytics by giving it a brief introduction, describing its characteristics, challenges, components, advantages, and even some of its disadvantages, also we clarify the term real-time Big data analytics (the relation between the real-time and the Big data analytics) and also explain the difference between the real-time and historical analytics and at last we publicize some big companies real-time analytics architectures.

In the third chapter: we introduce the social media then define the social media analytics and light its benefits, we explain the data analytics methods because of its big value and the variety of data types coming from the social media platforms, we explain the social media analytic process and introduce the key areas for analyzing the social media and finally present some native tools for the social media analytics and their most important insights.

In the fourth chapter: we introduce the Technology stack KDM (Kafka, Druid, Metatron) and the cause of picking its components, its functional and non-functional advantages like the performance and the effectiveness to analyze the social media data, and at last, we point the challenges we faced and some perspectives for further development.

I. CHAPTER 1: BIG DATA

1. Introduction

Every day, a wealth of digital information is being generated, information has great potential value for many purposes if captured and aggregated effectively. Previously, data warehouses were largely an enterprise phenomenon, with large companies being unique in recording their day-to-day operations in databases, and warehousing and analyzing historical data to improve their businesses. Nowadays, researchers in a deferent sectors and areas are seeing an important potential value and insight to be gained by warehousing the emerging wealth of digital information, popularly referred to as “big data”.

2. What is big data?

2.1 A bit of big data history

It is fair to say that the IT world has been facing big data challenges for over four decades it’s just that the definition of “big” has been changing. In the 1970s, big meant megabytes; over time, big grew to gigabytes and then to terabytes. Today, the IT notion of big has reached the petabyte range for conventional, high-end data warehouses, and exabytes are presumably waiting in the wings. In the world of relational database systems, the need to scale databases to data volumes beyond the storage and/or processing capabilities of a single large computer system gave birth to a class of parallel database management systems known as “shared-nothing” parallel database systems [2]. As the name suggests, these systems run on networked clusters of computers, each with their own processors, memories, and disks. Data is spread over the cluster based on a partitioning strategy usually hash partitioning, but sometimes range partitioning or random partitioning and queries are processed by employing parallel, hash-based divide-and-conquer techniques. A first generation of systems appeared during the 1980s, with pioneering prototypes and the first commercial offering to come to the web and the resulting need to index and query its burgeoning content created big data challenges. For research companies. The processing needs in the research world were quite different, however, and SQL was not the answer, although uncharged clusters again emerged as the hardware platform of choice. Google responded to these challenges by developing the Google File System (GFS), providing a familiar file view based on a byte stream of data partitioned randomly across hundreds, if not thousands, of nodes in a cluster [3]. GFS was then coupled with a programming model, MapReduce, to allow programmers to process big data by writing two user-defined functions, map and collapse [4]. The MapReduce framework applied these functions in parallel to individual data instances (Map) in GFS files and to sorted groups of from Teradata Corporation. The past decade has seen the emergence of a second major wave of these systems, with a number of startups delivering new parallel database systems that were then swallowed up through acquisitions by the industry’s major hardware and software vendors. Because high-level, declarative language (SQL) front relational databases, users of parallel database systems have been shielded from the complexities of

parallel programming. As a result, until quite recently, these systems have arguably been the most successful utilization of parallel computing. During the latter 1990s, while the database world was admiring its "finished" work on parallel databases and major database software vendors were busy commercializing the results, the world of distributed systems starts facing its own set of big data challenges. The quick growth of the deep instances that share a common key (Reduce) similar to the partitioned parallelism used in shared-nothing parallel database systems. Yahoo and other big Web companies such as Facebook and Tweeter created an Apache open source version of Google's big data stack because of the now highly popular Hadoop platform with its associated HDFS storage layer.

Just like the big data back-end storage and analysis dichotomy, the historical record for big data also has a front- story worth noting. As enterprises in the 1980s and 1990s started automating more and more of their day-to-day operations using databases, the database world had to scale up its online transaction processing (OLTP) systems as well as its data warehouses.

Companies like Tandem Computers reacted with fault-tolerant, cluster-based SQL systems. The same situation, but later in the distributed systems world, large Web companies were driven by an expansive user bases to find solutions to achieve very fast simple lookups and updates to large, keyed data sets such as collections of user profiles. Monolithic SQL databases built for OLTP were rejected as being too expensive, too complex, and/or not fast enough, and today's "NoSQL movement" was born [5]. Again, companies such as Google and Amazon developed their own answers (BigTable and Dynamo, respectively) to meet this set of needs, and again, the Apache open-source community created corresponding clones like HBase and Cassandra.

2.2 A definition of big data

A couple of definitions were given to this concept one of the themes is an enabler of its scientific development. As Ronda-Pupo and Guerras-Martin suggest, the level of consensus shown by a scientific community on a definition of a concept can be used as a measure of the progress of a discipline. Big Data has instead evolved so fast and disorderly that such universally accepted formal statement denoting its meaning does not exist. There have been many definitions of Big Data. However, none of these proposals has prevented authors of Big Data-related works to extend, renew, or ignore previous definitions and propose new ones. Although the concept of Big Data is still a relatively young one, it certainly deserves an accepted vocabulary of reference that enables the proper development of the discipline among cognoscenti and practitioners.

Consensual Definition

By looking at both the existing definitions of Big Data and at the main research topics associated with it, we can affirm that the nucleus of the concept of Big Data can be expressed by:

- 'Volume', 'Velocity' and 'Variety', to describe the characteristics of Information involved.

- Specific ‘Technology’ and ‘Analytical Methods’, to clarify the unique requirements strictly needed to make use of such Information.
- Transformation into insights and consequent creation of economic ‘Value’, as the principal way Big, Data is impacting companies and society.

We believe that the “object” to which Big Data should refer to in its definition is ‘Information assets’, as this entity is clearly identifiable and is not dependent on the field of application.

Therefore, we believe in the following formal definition:

“Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.”

Such a definition of Big Data is compatible with the existence of terms like “Big Data Technology” and “BigData Methods” that should be used when referring directly to the specific technology and methods mentioned in the main definition.

2.3 Information, the Fuel of Big Data

One of the essential reasons for the existence of the Big Data phenomenon is the current extent to which information can be generated and made available.

Take the case of mass digitization which represents the attempt to convert libraries of entire printed books into digital collections using optical character recognition (OCR) software to minimize human intervention (Coyle, 2006.). One of the most well-known attempts of this is the ***GPLP*** (Google Print Library Project), the beginning was in 2004, that aimed at digitizing more than 15 million volumes held in multiple university libraries, (Harvard, Stanford, Oxford, etc.). More recently it has been proposed a subtle differentiation between digitization and its next step, datafication[1] (Mayer-Schönberger, V. & Cukier, K., 2013.). The fundamental difference is that digitization enables analog information to be transferred and stored in a more convenient digital format while the aim of datafication is to organize digitized version of analog signals in order to generate insights that would have not been inferred while signals were in their original form. In the previous mentioned case of the Google mass digitization effort, the value of datafication came when researchers showed they were able to provide insights on lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology by using Google Books data (Michel, J.-B. et al, 2011.).

Digitization and datafication have become pervasive phenomena and it’s due to the broad availability of devices that are both connected and provided with digital sensors. This sensor allows the digitization while connection lets data be aggregated so it permits datafication. Cisco estimated that between 2008 and 2009 the number of connected devices may overtake the number of living people (Evans, 2011.) and another study mention that by 2020 there will be 26 billion devices on

[1] putting a phenomenon in a quantified format so that it can be tabulated and analyzed

earth, with a ratio of more than 3 devices per person (Gartner, 2014). The pervasive presence of a variety of objects (and that includes mobile phones, sensors, RFID - tags, actuators, etc.), those objects are able to interact with each other and cooperate with their neighbors to reach common goals under the concept of the ***IoT*** (Internet of Things) (Atzori, 2010) (Estrin, 2002). This increasing availability of sensor-enabled, connected devices is equipping companies with extensive information assets from which it is possible to create new business models, improve business processes and reduce costs and risks (Chui, 2010). In other words, IoT is one of the most promising fuels of Big Data expansion.

Today, increasing variety of Data types is another characteristic of the data generated. Structured data such as traditional text/numeric information is now combined by unstructured data like audio, video, images, text and human language) and semi-structured data, such as XML and RSS feeds (Russom, 2011.). The diversity of types is one of the challenges that organizations obliged to tackle in order to make value out of the extensive informational assets available today (Manyika, 2011).

Data could be found in three forms:

A) *STRUCTURED*

Any data that can be stored, accessed and processed in a fixed format form is defined as 'structured' data. Over the period of time, computer scientists have achieved greater success in developing techniques for working with this type of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes [6].

Nb: Data stored in a relational database management system is one example of a 'structured' data.

Examples of Structured Data:

An 'Employee' table in a database is an example of Structured Data.

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

FIGURE 1: STRUCTURED DATA

B) *UNSTRUCTURED*

This type includes any data where the structure is classified as unstructured data or with unknown form, and data with huge size, the un-structured data poses multiple challenges in terms of its processing for deriving value out of it. For example, the heterogeneous data source that contains a combination of simple text files, images, videos etc. Nowadays, organizations have wealth of data

available with them but unfortunately, they are not familiar with how to derive value out of it since this data classified as an unstructured format (Guru99, 2020).

Examples of Un-structured Data:

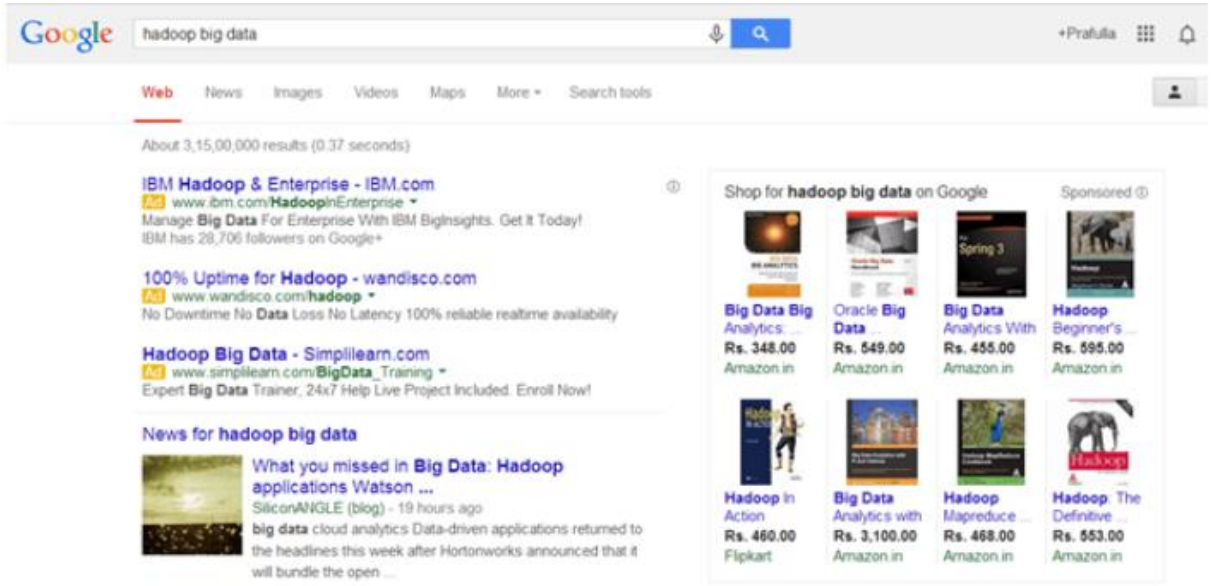


FIGURE 2: OUTPUT RETURNED BY 'GOOGLE SEARCH'

C) *SEMI-STRUCTURED*

Semi-structured data can contain both previous forms of data. It seems that semi-structured data as a structured in form but it is actually not defined with. It refers to the data that, although has not been classified under a database, still contains vital information or tags that segregate individual elements within the data.

The Personal data stored in an XML file is a good example of Semi-structured Data: (DeAngelis, 2020):

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

FIGURE 3: SEMI-STRUCTURED DATA

2.4 Source of Big Data

This concept "Big Data" is characterized by its sheer size. Its source includes both traditional and digital sources within and without a business. This concept includes all traditional sources where structured and more static sets of data come from for conventional data analytics. This kind of sources may include credit card application data, personal consumption data, retail sales data, etc. and it reveals some facets about the past of the consumer. These data set normally take data collectors quite a while along their business process to obtain, clean, structure and analyze. In fact, the traditional data still plays a critical role in many analyses, such as banking credit underwriting. The digital source, on the other hand, represents a wild array of websites and user interfaces [1].

2.5 Defining characteristics of Big Data

Big data is in reality so different from our old data analytics because of the letter V. Five words starting with V well defined this idea of Big Data, namely, volume, variety, velocity, veracity and value [5].

a) Volume

The name itself is related to a size which is enormous. Size of data plays a very important role in determining value out of data. In addition, whether a particular data can actually be considered as a Big Data or not is dependent upon tits volume. So, it's the is one characteristic needed to be considered while dealing with Big Data.

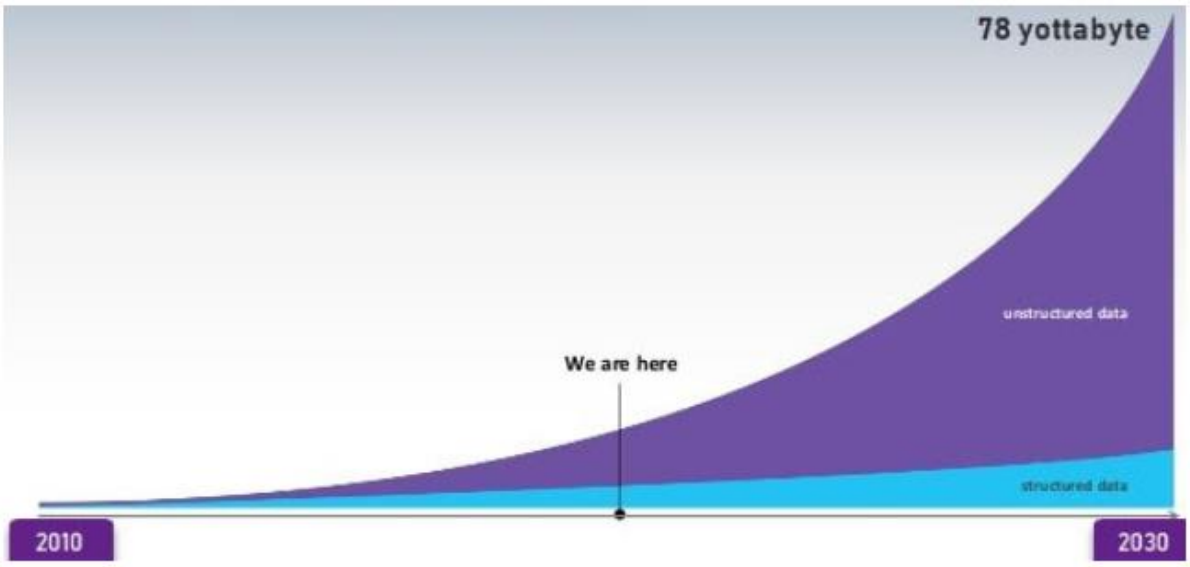


FIGURE 4: DATA GROWTH OVER THE YEARS

To put things into context. According to Zikopoulos, ***5PB*** of data, which is about 10,000 Blu-ray Discs, is generated every day from mobile devices. The size is often supported by modern social networks, which generate tons oftextual communications each day. Moreover, consumers carrying smart phones access of social networks as long as they are awake which make things worse. Local Storage is no longer a matter due to the cloud computing. It moves everything online with the capability of parallel processing, registering and result integrating. Actually, the volume factor unveils a different universe in almost every aspect of data analytics.

In the other hand, Business value of Big Data is highly related to the size of data, but it is not all that matters. Nobody should assume business superiority simply by possessing more data points than others do. It is what one does with the data and insights a business can truly get out of these data that really make a difference.

b) Variety

Variety is related to the net-centric life style. Traditional database systems don't handle unstructured data well. Numerical, logical and simple descriptive data are the majority of traditional data sets. Today's world is full of spontaneous communications in the format of text (Facebook posts and Twitter tweets, etc.). In addition, Internet offers a rich set of multimedia data (video, audio and pictures). Business environments further offer their own dimension of variety. For instance, a supply chain enabled by radio frequency tags constantly generates location and inventory data. As a result, we remark an augmentation in data variety in terms of format, terms of business content, and analytical needs. In the case of a call center, a sound classification system driven by Big Data analytics would identify customer voice change automatically, along with key words such as "outrage" and "unacceptable", trending to a churn event.

c) Velocity

Velocity which means time, the rate at which data pours into our system and the amount of time we have to understand the data and make decisions. Comparable to stock trading, picking up a trend ahead of others offers a superiority over all of your competitions. More importantly, propensity change and preference updates can be quickly captured by Big Data due to its velocity. Information updating speed borders real time. Velocity is a daunting concept that challenges technological and managerial limits. For example, the IBM report stated that: *"...one financial services sector (FSS) client analyzes and correlates more than 5 million market messages per second to execute algorithmic option trades with an average latency of 30 microseconds. Another client analyzes more than 500,000 Internet Protocol*

detail records (IPDRs) per second, which equals more than 6 billion IPDRs per day, on more than 4PB of data per year to understand trending and the current health of its networks.

Consider an enterprise network security problem. In this domain, threats come in microseconds, so you need technology that can respond and keep pace. However, you also need something that can capture lots of data quickly and analyze it to identify emerging signatures and patterns on the network packets as they flow across the network infrastructure..." Good technical solutions are not necessarily abundant in the market. Nonetheless, velocity has to be delivered in Big Data era.

d) Veracity

Veracity is overlooked among the characteristics of Big Data. It means useful signals being shifted out from tons of noises.

e) Value

It represents the amount of data a business possesses is spiking while result reliability and accountability is in decline. The way a system is able to remove noises and identity the true indicative piece of information, as quickly as possible, defines the inherent quality of a Big Data system.

3. Applications of Big Data

3.1 Finance

Big Data certainly means a lot more data at speeder velocity that drives deeper insights and expands operational echelon. On the other side, it challenges financial professionals for their decision making and compliances. For example, a financial firm will be able to predict the impact of tax and regulatory implications for a global expansion or business accounting change much better. Credit risk, for business borrowers in specific, has been challenging because of its inherent complexity of such risk and its correlation with many other aspects of a business [3].

3.2 Supply Chain

Big Data is about better processing data and analyzing data at a larger scale in a short time. Business and context knowledge lead to a deeper understanding of how Big Data and its applications can be beneficial for a business. Waller and Fawcett (2013) explain how supply chain specific analysis can deliver new perspectives using Big Data. The success of supply chain management hinges on two factors (precision and responsiveness). Big Data

allows recording of information at much lower and specific level (data can be recorded by SKU or by location).

3.3 Marketing

Marketing benefits greatly from Big Data, from consumer sentiment, advertisement to customer communications and more. Some consulting firms have done a straightforward analysis that went beyond traffic [4]. They correlated client's media spending, Web traffic, customer inquiries, and purchase data to discover the relationships between each of those metrics. They also tracked overall results to increases or decreases in advertising spending as well as the effects of lag time between media buys. That understanding has given the consulting forms the capacity to improve the performance of the company's media investment. For example, we have the story of a client who used software to look for patterns in call center conversations[2]. And the client that used speech analytics to analyze the conversations of customers who were terminating their accounts, then used key words and phrases to identify other at-risk customers[3]. [2].

4. Equipment for Working with Big DATA: Technology

Big Data is a term frequently associated with the particular technology that enables its utilization. The extent of the dataset size and the complexity of operations needed for its processing entail stringent memory storage and computational performance requirements. According to Google Trends, the most related query to "Big Data" is "Hadoop" that indeed is the most prominent technology associated with this topic. Hadoop is an open source framework that enables the distributed processing of big quantities of data by using a group of dispersed machines and specific computer programming models.

The distributed nature of information requires a specific technological effort for transmitting big quantities of data and for monitoring the overall system performance using special benchmarking techniques [1].

Another crucial technological element is to be able to store a bigger quantity of data on smaller physical devices. Although Moore's law suggests that storing capacity increases over time in an exponential manner [1], still it is required a continuous and expensive research and development effort to keep up with the pace at which data size increases

[2] The client discovered (to its horror) that it was unintentionally misleading some customers with its advertising.

[3] Reaching out to these at-risk customers, the client reported that it saved some 600 accounts and more than $12 million in revenue in the first year of the program

especially with the growing share of byte-hungry data types such as images, sounds and videos.

Big Data typically handle large volumes of data. Therefore, one of the most suitable environments to deploy Big Data solutions is the Cloud. Cloud systems are scalable, high-performant and fault tolerant environments where Big DBMS can be deployed.

Furthermore, cloud vendors allow its clients to hire on-demand resources in a pay-per use way, transferring all technical details from clients to itself, which allows clients focusing business needs rather than be concerned over IT problems such as software licenses, IT teams, infrastructure maintenance and so on. There are three cloud basic services:

a. Infrastructure as a Service (IaaS) is the basic service that allows cloud users to hire storage and processing power in order to serve their best purposes.
b. Platform as a Service (PaaS) such as Google App engine that offers a platform which can be used to develop and store online applications.
c. Software as a service (SaaS) such as Dropbox2, offers OnDemand software in the cloud. Big Data, as well as Analytics are often provided as SaaS, specifically Big Data as a Service (BDaaS) and Analytics as a Service (AaaS).

Regarding a cloud environment, a Big Data solution works at the application level, either as a BaaS or BDaaS hired from a vendor or a DBMS system – such as Hadoop, Cassandra, MongoDB and others – installed and maintained by the client itself.

Tools for Big Data analytics

There are various tools used in the big data namely Hadoop, HPCC, Storm, HBase, Grid Gain. It is used to improve the various factors in the development of big data and functionality of a computer system.

A. Hadoop is a project which is being developed by the Apache Software Foundation that supports huge data sets in a scattered location. This Hadoop Tool is a platform independent tool since it is developed in JAVA Framework which is used to process several applications or nodes at a same time with a speed ranging in terabytes.

B. HPCC (High Performance Computing Cluster) which is developed by LexisNexis Risk Solutions which is used to increase the performance factor of the System. It is used for parallel and batch-based processing of application using big data.

C. Storm is a tool which is mostly used in Real-time Systems that processes unbounded streams of data in a reliable manner which can be developed in any programming language.

This is a very fast processing tool which is used to process million tuples per second per node.

D. HBaseis a tool which mostly enhances the functionality of Hadoop and their Clusters based systems in high numbers. The main function of this tool is to map the data into different datasets, where each of the datasets are slatted into n- tuples and it is mapped into another task to reduce the output and group together to get back to its original form.

E. Grid gain is an open source tool developed in JAVA platform in real-time systems which is used as an alternative for Hadoop's Hbase in a distributed environment. This is an analysis tool which is highly used in commercially version.

The main interest of using these tools is to make the people skilled with more analytical initiative and to get more awareness in the areas of technologies and development. These tools are preferred in the area of Business intelligence to improve the existing business to larger extent and increase the economy of the corporate and also satisfy the business goals of the company. Those tools are implemented in many projects such as analytics-based projects and database projects to implement new ideas and thoughts of the peoples. It also chooses the optimal solution and provides the perfect result for the problem which is being subjected to analyses.

5. Conclusion:

In this chapter we attempted to define and clarify the concept of Big Data and presented its defining characteristics and its vital role in the context of three industries, namely, finance, supply chain and marketing. We put the light on Big data analytics many of the existing products are being modified with an additional feature. The main reason is that it assures the user or customer an accurate result and it also provides process and to manage complex data's and an ease to the client and server processors. In the last section we presented some of these tools to help the user or data analyst people make choices. The supply and demand of these tools are increasing in market day by day which creates high interest among the business peoples and corporate sectors.

II. CHAPTER 2: REAL TIME ANALYTICS

1. Introduction

With the emergence of terms like multi-structured data, Big Data, Internet of Things (IoT), and streaming data, business owners and operators are faced with the challenge of managing petabytes of data arriving at tremendous speed from various data pipelines across the enterprise. For that, Real-time Analytics has assumed supreme importance in recent times (Ghosh, on September 7, 2017).

In the past, competing businesses used advanced Business Intelligence (BI) capabilities to differentiate themselves from their competition. Now, Real-time Analytics can play the same role as a game-changer in the competitive Business Intelligence landscape. In today's business climate, the businesses which know how to extract insights from streaming data and use it judiciously for enhanced performance will win the race. Real-Time Analytics will play a crucial role in separating the winners from the followers or losers.

2. What are Real Time Analytics?

2.1 brief discussion and definition of Real Time Analytics

"*Real-time data analytics is a process that mainly focuses on the data produced or consumed or stored within a live environment*", In other words, Real-time data analytics is the process which allows analyzing the huge amount of data at the moment it is used or produced. Where we extract valuable information for the organization as soon as it's created or stored. Like in the case of analyzing a huge amount of data as it is produced within banks and branches, stock exchanges throughout the globe. The range of analytics can be from multiple sources. We can import the data and store it within a system and can execute data analysis algorithms over it. And these analytics data are delivered to the users/administrator through an analytics dashboard.

Or another short definition is the analysis of data as soon as that data becomes available. In other words, users get insights or conclusions immediately when the data enters their system. Take the example of customer experience management as a use case of real-time analytics. In customer relations managements and customer experience management, real-time analytics can provide up-to-the-minute information about an enterprise's customers and present it so that better and quicker business decisions can be made.

Here are some examples of how enterprises are tapping into real-time analytics:

- Fine-tuning features for customer-facing apps: Real-time analytics adds a level of sophistication to software rollouts and supports data-driven decisions for core feature management.
- Managing location data: Real-time analytics can determine what data sets are relevant to a specific geographic location and signal the appropriate updates.

- Detecting anomalies and frauds: Real-time analytics can be used to identify statistical outliers caused by security breaches, network outages or machine failures.
- Empowering advertising and marketing campaigns: Data gathered from ad inventory, web visits, demographics and customer behavior can be analyzed in real time to uncover insights that hopefully will improve audience targeting, pricing strategies and conversion rates.

Examples of real-time analytics include:

- Real-time credit scoring. Instant updates of individuals' credit scores allow financial institutions to immediately decide whether or not to extend the customer's credit.
- Financial trading. Real-time big data analytics is being used to support decision-making in financial trading. Institutions use financial databases, satellite weather stations and social media to instantaneously inform buying and selling decisions.
- Targeting promotions. Businesses can use real-time analytics to deliver promotions and incentives to customers while they are in the store and surrounded by the merchandise to increase the chances of a sale.
- Healthcare services. Real-time analytics is used in wearable devices -- such as smartwatches -- and has already proven to save lives through the ability to monitor statistics, such as heart rate, in real time.
- Emergency and humanitarian services. By attaching real-time analytical engines to edge devices -- such as drones-- incident responders can combine powerful information, including traffic, weather and geospatial data, to make better informed and more efficient decisions that can improve their abilities to respond to emergencies and other events.

2.2 Characteristics and challenges

a) Challenges

One major challenge faced in real-time analytics is the vague definition of real time and the inconsistent requirements that result from the various interpretations of the term. As a result, businesses must invest a significant amount of time and effort to collect specific and detailed requirements from all stakeholders in order to agree on a specific definition of real time, what is needed for it and what data sources should be used.

Once the company has unanimously decided on what real time means, it faces the challenge of creating an architecture with the ability to process data at high speeds. Unfortunately, data sources and applications can cause processing-speed requirements to vary from milliseconds to minutes, making creation of a capable architecture difficult. Furthermore, the architecture must also be capable of handling quick changes in data volume and should be able to scale up as the data grows [1].

The implementation of a real-time analytics system can also present a challenge to a business's internal processes. The technical tasks required to set up real-time analytics -- such as creation of the architecture -- often cause businesses to ignore changes that should be made to internal processes.

Enterprises should view real-time analytics as a tool and starting point for improving internal processes rather than as the ultimate goal of the business.

Finally, companies may find that their employees are resistant to the change when implementing real-time analytics. Therefore, businesses should focus on preparing their staff by providing appropriate training and fully communicating the reasons for the change to real-time analytics.

There for this the most important characteristics for a real-time analytics platform.

b) *Characteristics*

The term real-time analytics signifies the requirement for processing data as it generated and within a specified time interval. This time interval is typically in order of milli-, micro- or even nano-seconds, depending on the system in question. Real-time systems are often said to be the systems in which timeliness is essential to correctness.

Real-time property requires the collection of data associated with various data sources and processing them as they arrive. It often involves generating notifications to humans about significant occurrences in the environment, or invoking.

In order for the real-time data to be useful, the real-time analytics applications being used should have two critical characteristic a high availability and a low response time. These applications should also feasibly manage large amounts of data, up to terabytes. This should all be done while returning answers to queries within seconds.

The term REAL-TIME also includes managing changing data sources -- something that may arise as market and business factors change within a company. Consequently, the real-time analytics applications should have the ability to handle with big data. The exploitation of real-time big data analytics works on maximizing business returns, reducing costs and introduce an era where machines can interact over the IoT using real-time information to make decisions on their own which is known by industry 4.0.

Different technologies exist that have been designed to meet these demands, including the growing quantities and diversity of data. Some of these new technologies are based on specialized appliances -- such as hardware and software systems. Other technologies utilize a special processor and memory chip combination, or a database with analytics capabilities embedded in its design [1].

Some key requirements are:

⇒ *LOW LATENCY*

In general, latency can be defined as the time delay between the cause and the effect of some physical change in the system being observed.

In real-time analytics, latency refers to the time between a data generated in an environment and the start of its processing in the system, also returning answers to queries within seconds. This latency typically involves network latency, computer processing latency and the structure of data itself.

CHAPTER TWO: REAL TIME ANALYTICS

Real-time systems require low latency in order to respond to the events within specified time bounds.

A number of strategies can be adopted to support this requirement in analytics [8]. These include:

- In-memory processing should be minimized processing delay associated with the use of disks and I/O; this is increasingly viable due to the decreasing cost of memory (McCue, (Second Edition), 2015).
- Incremental evaluation that is updating calculations and query results for each new data item without re-evaluating the entire data set.
- Anticipatory fetching and processing, enabling faster access to data from multiple data streams.
- In-database analytics -- a technology that allows conducting data processing within the database by building analytic logic into the database itself.
- Data warehouse appliances -- a combination of hardware and software products designed specifically for analytical processing. An appliance allows the purchaser to deploy a high-performance data warehouse right out of the box.
- In-memory analytics an approach to querying data when it stored in random access memory, by opposition to querying data that is resides on physical disks.
- Massively parallel programming -- the coordinated processing of a program by multiple processors that work on different parts of the program, with each processor using its own operating system and memory.
- Use of flash technology to store all data that does not need to be in main memory; this approach increases access speed to data;

⇒ *HIGH AVAILABILITY*

Availability means an ability of a system to perform its function when required.

Real-time analytics systems require high availability because otherwise the events and data arriving from the outside world are not immediately processed and are difficult store or buffer for subsequent processing, especially with high volume, high velocity data streams.

many strategies can be adopted to support enable effective analytics including [8]:

- Distribute processing to multiple nodes so that if one machine fails another can take over its processing.
- Replication of data to several servers so that of one machine fails the data can still exist on another machine.
- Redundant processing of data, that is, having more than one node calculating a result for a data set (which implies both of the above).

⇒ ***HORIZONTAL SCALABILITY***

This term refers to the ability to add servers to an existing pool to increase performance and capacity. The ability to dynamically add new servers as data volume or processing workload requires is of high importance for real-time analytic systems to ensure that data is processed within specified time intervals.

Horizontal scalability is especially important if one cannot control the rate of data ingress. If a system is consuming a known, fixed-volume feed, then it can be sized to ensure the real-time requirements are met.

Note that horizontal scalability is to be contrasted to vertical scalability, which refers to the ability to add resources to a single server to improve performance and capacity [8].

2.3 Real-Time Analytics components

Real-time analytics platform consists of the following steps:

- Data Sources.
- Aggregator.
- Broker.
- Analytics Engine.

a) Data Sources

For real-time analytics, the first major need sources from where real-time data is obtained. There can be many sources of streaming data:

- Sensor Data: The sensor is the output of the device that measures a physical quantity and transforms it into a digital signal.
- Social Media Stream: Social media streaming like a Twitter feed, Facebook, Instagram, YouTube, Pinterest, etc.
- Clickstream: The stream contains the data about which pages the website visits and in what order.

b) Aggregator

It gathers or ingest the real time streaming data (and perhaps batch files) from many different data sources. Kafka or Kinesis can do the aggregation task like example.

⇒ ***KAFKA***

It is an open-source message broker project. It was developed by the Apache Software Foundation and written in Scala. Aiming to provide a platform particular unified, high throughput and low-latency to handle with real-time data feeds a single Kafka broker can handle hundreds of megabytes of reads and writes per second from thousands of clients. In order to support high

availability and horizontal scalability, data streams are partitioned and spread over a cluster of machines. Kafka depends on Zookeeper from the Hadoop ecosystem for coordination of processing nodes.

The main uses of Kafka are in situations when a very high throughput is demanded in applications for message processing, not to mention the other criterion as time, availability and scalability requirements.

⇒ *AMAZON KINESIS*

Amazon Kinesis is a cloud-based service for real-time data processing over large, distributed data streams. This service can continuously capture and store terabytes of data every hour from thousands of sources such as website clickstreams, financial transactions, social media feeds, IT logs, and location-tracking events.

Kinesis allows integration with Storm, as it provides a Kinesis Storm Spout that fetches data from a Kinesis stream and emits it as tuples. The inclusion of this movement component into a Storm topology provides a reliable and scalable stream capture, storage, and replay service.

c) Broker

it makes data available for consumption by querying it using languages like SQL. Like Hadoop, Spark, row-oriented databases (MySQL, PostgreSQL), column-oriented databases (Snowflake, Redshift, BigQuery), real-time databases, batch files, key/value databases (Memcached, Oracle NoSQL Database, Redis), Warehouses or any technology suits and available.

⇒ *HADOOP*

The Apache Hadoop software library represent a framework that grant for the distributed processing of large data sets across clusters of computers using simple programming models. This library is designed to scale up from any single servers to deferent thousands of machines, each offering local computation and storage. Rather than count on hardware to deliver high-availability, the software is designed to detect and manipulate failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

⇒ *SPARK*

Apache Spark, is more recent framework that combines an engine for distributing programs across clusters of machines with a model for writing programs on top of it. It is aimed at addressing the needs of data scientist community, in particular in support of Read-Evaluate-Print Loop (REPL) approach for playing with data interactively.

CHAPTER TWO: REAL TIME ANALYTICS

Spark maintains MapReduce's linear scalability and fault tolerance, but extends it in three important ways. First, rather than relying on a rigid map-then-reduce format, its engine can execute a more general directed acyclic graph (DAG) of operators. This means that, in situations where MapReduce intermediate results must be written to the distributed filesystem, Spark can pass them directly to the next step in the pipeline. Second, it complements this capability with a rich set of transformations that enable users to express computation more naturally. Third, this software supports in-memory processing across cluster of machines, thus not counting on the use of storage for recording intermediate data, as in MapReduce.

Spark supports integration with the variety of tools in the Hadoop ecosystem. It can read and write data in all of the data formats supported by MapReduce. It can read from and write to NoSQL databases like HBase and Cassandra.

Spark was written in Scala but it comes with libraries and wrappers that allow the use of R or Python.

⇒ *ROW ORIENTED DATABASES*

are databases that organize data by record, keeping all of the data associated with a record next to each other in memory. Row oriented databases are the traditional way of organizing data and still provide some key benefits for storing data quickly. They are optimized for reading and writing rows efficiently.

Common row-oriented databases:

- Postgres
- MySQL

⇒ *COLUMN ORIENTED DATABASES*

are databases that organize data by field, keeping all of the data associated with a field next to each other in memory. Columnar databases have grown in popularity and provide performance advantages to querying data. They are optimized for reading and computing on columns efficiently.

Common column-oriented databases:

- Redshift
- BigQuery
- Snowflake

⇒ *REAL-TIME DATABASES*

A real-time database is a database system which uses real-time processing to handle workloads whose state is constantly changing. This differs from traditional databases containing persistent data, mostly unaffected by time. Real-time processing means that a transaction is processed fast enough

for the result to come back and be acted on right away. This system is useful in many fields from accounting to banking, law and medical records, multi-media, process control, etc.

From the variety models of real-time database we have (Druid, RethinkDB, and MemSQL).

⇒ ***MEMCACHED***

This term presents a system of general-purpose distributed memory-caching. Memcached is often used to fast up dynamic database-driven websites by caching data and objects in RAM to ensure reduction of the number of times an external data source must be read. It is free and open-source software, licensed under the Revised BSD license which runs on Unix-like operating systems and on Microsoft Windows. It depends on the libevent library. Providing a large hash table distributed across multiple machines. Memcached has no internal mechanism to track misses which may happen ("Releases - memcached/memcached", s.d.).

⇒ ***REDIS***

This system is an open source (BSD licensed), in-memory data structure store, used as a database, cache and message broker. Deferent data structures such as bitmaps, strings, lists, sets, sorted sets with range queries, hashes, geospatial indexes with radius queries and streams are supported. Redis has fabricate replication and different levels of on-disk persistence, and provides high availability thanks to Redis Sentinel and automatic partitioning with Redis Cluster.

d) Analytics Engine

Correlates values and blends data streams together while analyzing the data like Hadoop, Spark, Pentaho, which can used to feed a front-end app (dashboards, real-time dashboards).

⇒ ***PENTAHO***

Pentaho’s data integration and analytics platform enables organizations to access, prepare, and analyze all data from any source, in any environment.

2.4 Benefits of Real-Time Analytics

Real-time analytics enables businesses to react without delay, quickly detect and respond to patterns in user behavior, take advantage of opportunities that could otherwise be missed and prevent problems before they arise.

Businesses that utilize real-time analytics greatly reduce risk throughout their company since the system uses data to predict outcomes and suggest alternatives rather than relying on the collection of speculations based on past events or recent scans -- as is the case with historical data analytics. Real-time analytics provides insights into what is going on in the moment.

Other benefits of real-time analytics include:

- Data visualization: Real-time data can be visualized and reflects occurrences throughout the company as they occur, whereas historical data can only be placed into a chart in order to communicate an overall idea.
- Improved competitiveness: Businesses that use real-time analytics can identify trends and benchmarks faster than their competitors who are still using historical data. Real-time analytics also allows businesses to evaluate their partners' and competitors' performance reports instantaneously.
- Precise information: Real-time analytics focuses on instant analyses that are consistently useful in the creation of focused outcomes, helping ensure time is not wasted on the collection of useless data.
- Lower costs: While real-time technologies can be expensive, their multiple and constant benefits make them more profitable when used long term. Furthermore, the technologies help avoid delays in using resources or receiving information.
- Faster results: The ability to instantly classify raw data allows queries to more efficiently collect the appropriate data and sort through it quickly. This, in turn, allows for faster and more efficient trend prediction and decision making.

2.5 The Disadvantages of Real-Time Analytics

Nothing is perfect and this includes streaming data analytics. And certain companies may not want to use it because of the reasons below (Scott, Real-Time Analytics: Streaming Big Data for Business Intelligence, Mar 13th, 2017) :

- Hadoop is incompatible. Hadoop is tool for historical big data analytics, but it is not designed to handle streaming, real-time data. Better options include Spark Streaming, Storm, Apache Flink, or Apache Samza. Most of that MongoDB represent an open source database able to be used in Big Data analysis.
- A new approach is required. If your company is used to receiving a single batch of insights one time every week, then a constant inflow of real-time Big Data can overwhelm business processes. If one person or automated system would normally direct information from the insights to the relevant parts on a weekly basis, he has to know what to do when he starts to receive insights every minute. Information workflows need to be reworked appropriately.
- Systems can fail. May be Big Data analytics looks like easy to implement. But if a business is not used to handling data at fast rates, it could lead to faulty analysis — and even system failure. Despite all of the fanfare over streaming data, smaller businesses might not actually need it — and they might not even be able to handle it.

3. Breaking Down Real Time Big Data Analytics

3.1 Understanding real time big data analytics

To understand the meaning of real time big data analytics better, we should decorticate the phrase into its component parts - "real-time", "big data" and "analytics" - and delve deeper into the nuances of each one (sumologic, 2014).

a) Real Time

In the context computing science, real-time data processing basically means that we are performing an operation on the data just milliseconds after it becomes available. When it comes to monitoring your security posture, detecting threats and initiating rapid quarantine responses, a real time response is necessary to mitigate cyber-attacks before hackers can damage systems or steal data. (Francis Xavier MCRORY, Warner Robins ; Rogelio SAUCEDO, Bonaire, Mar. 17, 2011)

In today's cyber security environment, it is no longer effective to analyze event logs after-the-fact to determine whether an attack occurred. Real-time big data analytics helps organizations mitigates attacks as they happen by analyzing event logs milliseconds after they are created.

b) Big Data

The term of Big Data is being used in a vast field, but the question is about the difference between data and big data. Throughout the digital age, the widespread use of software applications has resulted in the generation of massive amounts of data. The storage of this data has been enabled by the simultaneous evolution of hardware storage devices in increasing both cost and space-efficient (Paul Zikopoulos, Chris Eaton , October 2011).

When the world's leading data collectors generated data sets that included many cases and higher degrees of complexity, it’s clear that traditional data ways will no longer meet the requirements of these organizations. The increases in computer processing power led to the development of predictive analytics tools that could help these organizations to extract information and insights from their enormous data sets (sumologic, 2014).

IT organizations are able to leverage their big data through log management or SIEM tools that aggregate network, application and event log files into a centralized, normalized database.

c) Analytics

Analytics is a software capability that takes data input from various sources, searches it for patterns, interprets those patterns and ultimately communicates the results in a human readable format. Analytics software uses mathematics, statistics, probabilities, and predictive models to find hidden relationships in data sets that are too complex and varied to be efficiently analyzed manually.

The best analytics tools today combine advanced technologies like machine learning and pattern recognition with other software features to achieve a specified goal. In IT organizations, analytics

tools are used to review event logs and correlate events from across applications to identify Indicators of Compromise (IoCs) and respond to security incidents.

3.2 Historical Data Analysis

Historical data analysis, as the name implies, focuses on looking at the past. Analysts can export the relevant data from the prior day, month, quarter, or some other period of time and then perform at least one of three different types of analyses.

a) Descriptive Analytics

These analytics represent condenses historical data into a story that has an overall theme that is relevant and useful. Descriptive analytics is that type of business intelligence activity that most people are familiar with (Descriptive Analytics Insight into the past -"What happened"). Generally, it is used to understand what happened in the past as well as to predict or prescript analytical models.

b) Predictive Analytics

Predictive analytics is the process of using historical trends to make future predictions and present likely scenarios with the assistance of data mining, machine learning, and statistics.

One of the best examples is Amazon's well-known "recommendations" that present various products to customers based on what they have purchased in the past and therefore what may interest them in the future. Other examples include the now-defunct Google Flu Trends and The New York Times deciding what website articles to promote to which visitors.

c) Prescriptive Analytics

Descriptive analytics tells what happened. Predictive analytics tells what may happen if historical trends continue. Prescriptive analytics tells what to do the process takes the data and then prescribes real-world decisions. Let's take the example of an InterContinental Hotel Group. UPS centralizes and analyses information from hundreds of data sources to optimize the exact routes that trucks should take — thereby saving millions of dollars on fuel every year. The Aurora Health Care Centre combined and analyzed numerous data sets to lower readmission by 10% and save $6 million every year. (Scott, DZone , Mar. 29, 17)

So, as The ERP Channel notes, the typical use cases for historical data analysis and descriptive, predictive, or prescriptive analytics include (Scott, Real-Time Analytics: Streaming Big Data for Business Intelligence, Mar 13th, 2017):

- To create big-picture operational and business decisions out the immediate flow of production
- To construct or modify prescriptive models based on static, historical data
- To produce periodic reports, evaluate interactive data discovery, and make "what if" modelling.

3.3 Historical and Real-Time Data Analysis

Although predictive analytics is the newest type of batch analytics but steal historical data analysis is not that new. While streaming analytics is comparatively new (Scott, DZone , Mar. 29, 17). Dataversity summarizes:

"*Stream processing analyses and performs actions on real-time data through the use of continuous queries. Streaming analytics connects to external data sources, enabling applications to integrate certain data into the application flow, or to update an external database with processed information*" (Streaming Analytics 101: The What, Why, and How). In descriptive and prescriptive analytics, one exports a set of historical data for batch analysis. And in the other hand, for real-time analytics one analyses and visualizes data in real time (Sedkaoui, 2018).

4. Real Time Technology Stack

To building a real-time application we should start with connecting the pieces of our data pipeline (Smith, August 5, 2019).

To make rapid and informed decisions, organizations need to quickly ingest application data, transform it into a digestible format and store it, to make it available all at sub-second speed (MemSQL Blog The Ideal Stack for Real-Time Analytics).

A real-time data pipeline could be architected as follows (Hooten, Mar. 05, 17):

- Application data is ingested through a distributed messaging system to capture and publish feeds.
- To distil information, enrich data, and deliver the right formats we need to call a transformation tier.
- To ensure the persistence, an easy application development and analytics Data must be stored in an operational real-time data warehouse.
- From there, data can be queried with SQL to power real-time dashboards.

As new applications generate increased data complexity and volume, it is important to build an infrastructure for fast data analysis that enables benefits like real-time dashboards, predictive analytics, and machine learning.

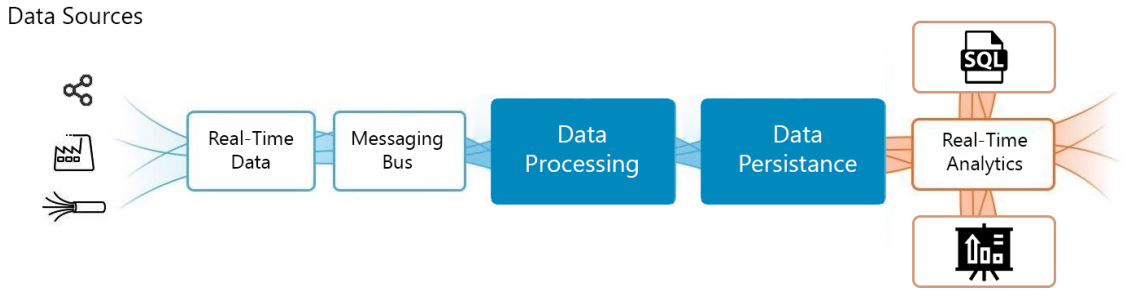


FIGURE 1: STANDARD REAL TIME ANALYTICS ARCHITECTURE

And here we have some examples of used real-time tech stack:

CHAPTER TWO: REAL TIME ANALYTICS

- Dream11, a fantasy sports platform based in India that aims to allow users to play fantasy sports such as cricket, hockey, football, and basketball. By 2019, this platform became the first Indian gaming company to enter the 'Unicorn Club'. And deals with more than 3 TB and plus than10 transactions per day. In addition to billions of clickstream events every day (Khan, August 18, 2020).

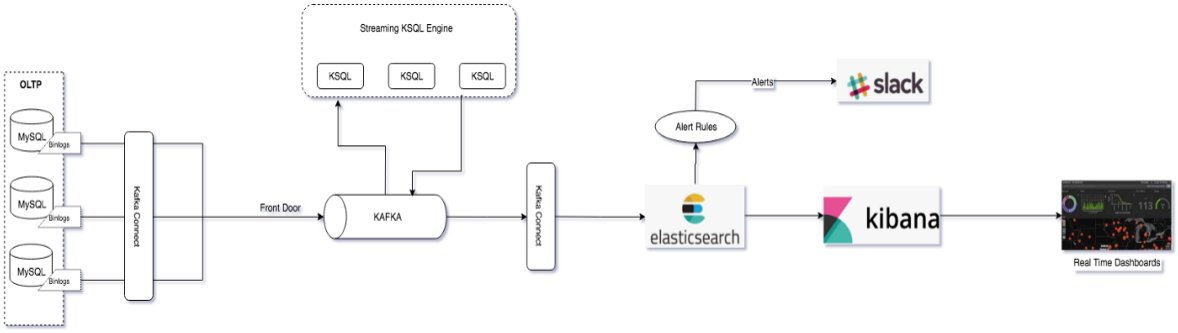


FIGURE 2: DREAM11 REAL TIME ANALYTICS STACK

- Sisense is a late-stage SaaS start-up and one of the leading providers of business analytics software, and was looking to improve its ability to analyze internal metrics derived from product usage - over 70bn events and growing.

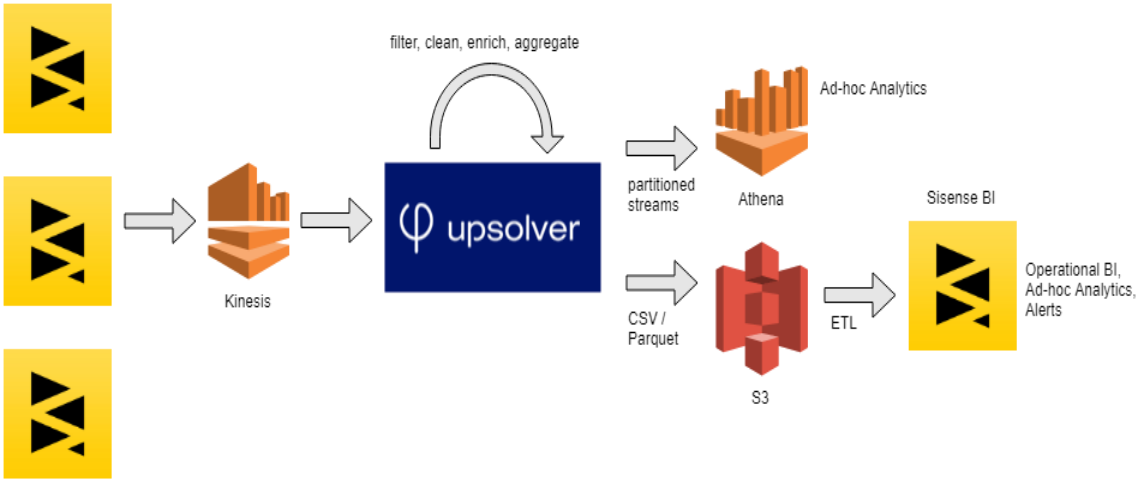


FIGURE 3: SISENSE REAL TIME ANALYTICS STACK

- Pinterest is an American social media service allows sharing images designed to insure saving and discovering of information on the World Wide Web using images and, on a smaller scale and so on in the form of pin boards. This site had about 300 million monthly active users (Presearch, s.d.).

FIGURE 4: PINTEREST REAL TIME ANALYTICS STACK

- One solution that is increasing in popularity is to use the open source ELK Stack, which centralizes any desired log and machine data with Logstash, stores it in Elasticsearch, and then displays it with real-time Kibana visualizations. ELK, for example, can be used to analyze Salesforce.

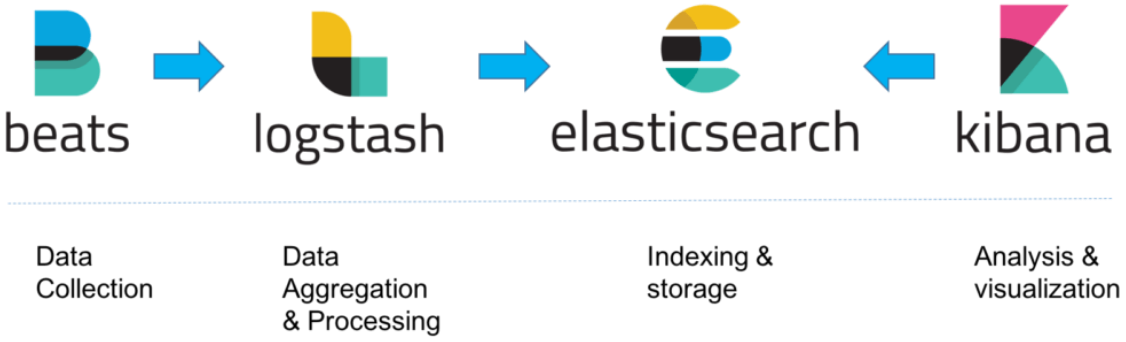


FIGURE 5: CLASSIQUE ARCHITECTURE

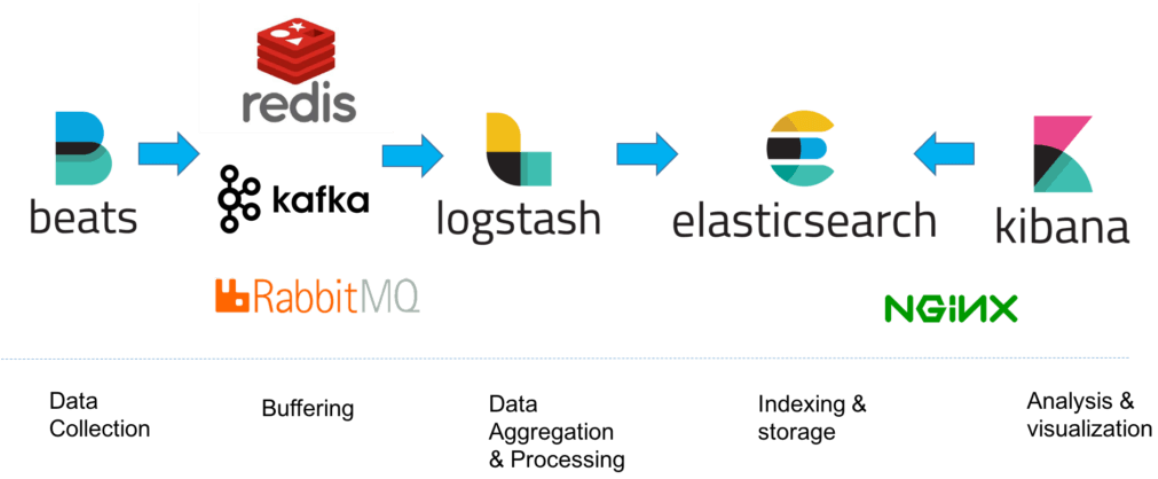


FIGURE 6: ADVANCED ARCHITECTURE

5. Conclusion

The benefits of real-time analytics, demand sensing, among others. Data streaming allows organizations to make the most out of data and enable them to gain operational efficiency. Companies need to implement these technologies and tools in their business processes and harness the power of data in every way possible.

III. CHAPTER 3: SOCIAL MEDIA ANALYTICS

1. Introduction

In the past decade social media has seen such a huge evolution becoming an important factor in getting and spreading information among different domains such as: business, entertainment, science, crisis management and politics. Among many reasons of social media being this famous is the possibility and the ability of creating and sharing messages in public at low costs and all around the globe. The increasing usage of social media led the data spread and shared in it to pill up enormously, which presented to us a new term "Social media big data". The data shared on social media comes in many forms textual, picture, videos, sounds and even geolocations. All of the stated types can be derived under two types of data structured and unstructured. The unstructured data can be seen in the format of text as a clear example and in the other hand we can see the followers/ friends as structured data.

This huge accumulation of data in social media opened up a new opportunity in the domains of data analyzing and patterns in communication. This collected data from social media can be used for example in gaining insight into issues, trends, influential actors and other kind of information. In 2011 Golder and Macy conducted an analytic study on Twitter data to monitor and study how people's mood can change over day time, weekday, and seasons. Information Systems (IS) is one of the many fields that use social media data in its studies to answer question such us how can the network position influence the information diffusion.

In this chapter we present social media and its data growth, in addition to its being a new source for big data and the relation between it and the big data. We also present the process of social media analytics: methods, types, and tools.

2. What is Social MEDIA?

2.1 Overview and definition

Earlier in the era pre internet media was limited to TV, newspapers, magazines, etc. Once the term internet was presented to the world it brought new means to spread media to the world and the media was no longer static.

Media before Internet existed was about television, newspapers, magazines, etc. after media became available through the World Wide Web it was not static anymore. According to **E. Gilbert & Karahalios** "*Social media (SM) is a set of Internet-based applications that is grounded by the idea of Web 2.0*"

A brief definition for social media would be: "*interactive server or machines and technologies that facilitates the creation and sharing of information, ideas, news …etc. via virtual communities and networks*".

2.2 Social Media data

Social network data are voluminous, even from a single social network site, mostly unstructured, refers to all of the raw insights and information collected from individual's social network activity – e.g. your prospects and customers. Social network data tracks how individuals engage with your content or channels like LinkedIn, Facebook, and Twitter. Then, basically it gathers numbers, percentages, and statistics that allow you to infer the performance of your social media strategy [24].

CHAPTER THREE:

Consumers spend increasingly of their total Internet time on social networking sites and forums. Statistics show that Internet users spend an average of 2 hours and 22 minutes per day on social networking in 2019 while the average daily time in 2018 was 142 minutes a day, Facebook has more than 1.4 billion active users on a daily basis. Twitter has 330 million monthly active users. These statistics, combined with the millions of blog posts on the Internet and discussions that occur by the minute on forums and social networking sites, for example, can provide a pool of data on market trends and such things. Still, data is one thing; analyzing it successfully to gain useful insights is quite another [25].

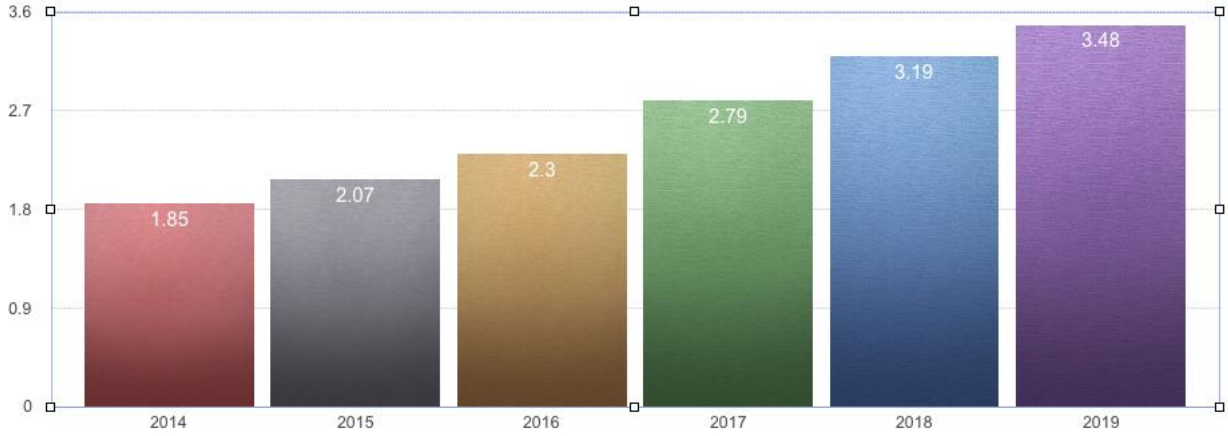


FIGURE 5 : SOCIAL MEDIA USAGE IN THE PAST 5 YEARS

2.3 The relation between big data and social media

In this technological and informational era Big Data proved its presence strongly in many fields and we can call it the Big Data era. Many large companies around the world like Microsoft, Amazon, and Google collects large amounts of data daily and we can imagine the huge amount of the collected data in last years which can be estimated by Hexabytes or larger. Billions of people are currently users on YouTube, Twitter, Facebook and other social networks. Therefor the social media was targeted by many business companies to promote their offered services and products in addition to staying in touch with their clients. Clients also keep an eye on social media and use it to get informed about interesting services and goods.

The huge increasing and growth of data published on social media by users provided an excellent chance to mine valuable insights and understand better users' behaviors. This has led to developing big data solutions to solve varied issues in real- life.

As we stated above the relation between big data and social media can be seen in collecting user's data to find real life issues solutions.

3 Social media analytics

3.1 Overview

The last decade has seen an increase in the use of social media, getting information from the crowd was a new source besides traditional media for people. Different platforms currently exist but they share many characteristics. The large amounts of data produced daily by users on different

platforms has prompted the organization to understand what issues and trends are evolving to identify risks and opportunities in communication and draw useful implications from them. In addition to the amount of content, it is important for organizations to know and understand who creates the content and which actors are the most influential drivers of communication [26].

All types of organizations seek to obtain data produced by the crowd in order to better understand mass communications. Influencers or opinion leaders, for example, can be identified through social media analysis, and by examining their network of followers, one can reveal the reach of such an individual. In addition, role behavior is examined in order to understand the causes of a key role in the network and the effects it has on the entire network. Companies such as media agencies have recognized the importance of influencers and use them. For product placement. In addition, the analysis of social media content has evolved in recent years to become one of the main objectives of research in information systems. A research objective could be to identify and analyze the dissemination of information journalism and political communication [26].

One example of an area where social media data has had an impact is crisis communication research. Social media is often used as a channel for emergency management agencies to inform people in an affected area about the current state of the respective crisis or how to behave. Social media data in the context of crisis communication can also be analyzed to gain additional, previously unknown, information whether volunteers e.g. take photos or videos and spread the information to the crowd. Data collected on social media can also be analyzed to detect a specific place or area where the crisis is occurring [26].

3.2 Benefits

Social media analytics has evolved from a simple tool for collecting customer requests and comments to a way to get critical business information and make quick and efficient decisions. By adding forecast functions to social media analysis, companies can predict more precisely what their customers are likely to do. Predictive analytics involves using regression models and advanced techniques like neural networks to get a comprehensive view of customers and their future actions based on their transactional data, social media and more [26]. Here are the areas where social media analytics can have a huge impact [26]:

- Innovation: Product development teams can use social media to understand what customers like and dislike about a brand, what product features a target audience wants, and what product features the brand has. Competitor. This information can be used to correct errors in the next iteration, spark new ideas, and review ideas and products in development. Most crowdfunding campaigns now use social media to promote ideas and contributions. Comments on new product demonstrations can also provide information about customer preferences in different markets.
- Marketing: Companies can no longer rely on analyzing yesterday's conversations with customers to shape today's marketing campaigns. With the help of social media analytics, marketers can use real-time marketing to deal with ever-changing customer preferences. By uncovering trending topics, marketers can quickly refine tweets and social media updates to keep them in line with current topics, stay relevant, and build customer loyalty. Companies like Dell and McDonald's use social media analytics to listen to customers in real time and adjust advertising campaigns and content on the fly to appeal to social media users. Based on feedback on social media, restaurants in Group 5 decided to tone down one of their Mexican chilly dishes, despite internal debates. Marketers can also use image recognition technology to see, for example, what images are shared by customers and what impact it has on have the sales.
- Distribution: Predictive analytics such as affinity or shopping cart analyzes provide details on which products are frequently bought together, as well as the right mix of products and services for

customers - such as a game and a movie based on the game. This information can be used for cross-selling, upselling, and personalization of products and services. Customer sentiment can be used to forecast sales and revenue and prepare ahead of time for a peak in demand.

- Customer Service: Social media channels allow companies to identify potential customer service issues before they escalate and damage a brand's reputation. By monitoring social media for real-time feedback on a new product release, the customer service team can identify issues and proactively contact customers to resolve issues. Customer service can also predict the type of problems customers may encounter at certain times and prepare accordingly.
- Competitive Intelligence: Nothing can be more valuable in business than solid Competitive Intelligence. With the help of social media analytics, companies can track competitor mentions on social media and understand how competitors use various social media platforms for branding and loyalty, for example. This information can be helpful in reviewing and strengthening current social media strategies. Tracking reviews and posts from bloggers and thought leaders on competing products can provide valuable information that can be used to improve various functions within the company.

4. Data analytics methods

With a very large dataset, the challenge is figuring out what data you have and how to analyze it. Social networks typically contain a huge amount of content and association data that can be used for analysis. These types can be subdivided into unstructured or structured data, depending on whether they are organized in a predefined way (structured data) or not (unstructured data). To illustrate this with an example, time-based events are structured, while event data based on tweets and "likes" is unstructured. Structured data in social networks is usually structured graphically. In their most basic form, they are modeled using a social network represented as a graph G = (V, E), where V is a set of nodes or entities (e.g., people, organizations and products) and E is a set of edges or relationships that connect nodes through interaction models. This type of data is measured by Social Media Analysis, a graphical analysis application that focuses on extracting information from this interrelated data. On the other hand, unstructured data is the content data shared in social networks, also known as user generated content (UGC). They are considered to be the cornerstone of SNS (Social Network Software) and include text, images, videos, tweets, product reviews, and other multimedia data, which are typically examined with content-based analytics, whose techniques include data structuring algorithms. Figure 6 summarizes the types of data and the corresponding analysis approaches and methods that have been carried out in ***OSN*** (Online Social Network) [27].

CHAPTER THREE:

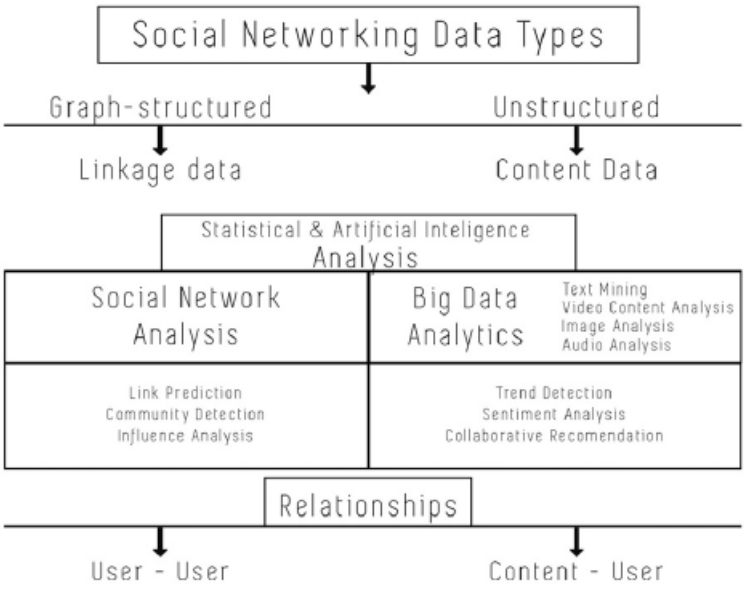


FIGURE 6: DATA TYPES AND ANALYSIS

Big data analytics is able to address various dominant research problems with the help of computer intelligence. Therefore, the analysis of big data on social media is divided into different classes to capture their characteristics. Figure 7 shows the many categories of big data analytics. The classification is based on four aspects: data sources, characteristics, computer intelligence and techniques. The proposed classification helps provide a systematic approach to understanding big data analysis techniques and technologies used in social media data. Figure 7 illustrates this.

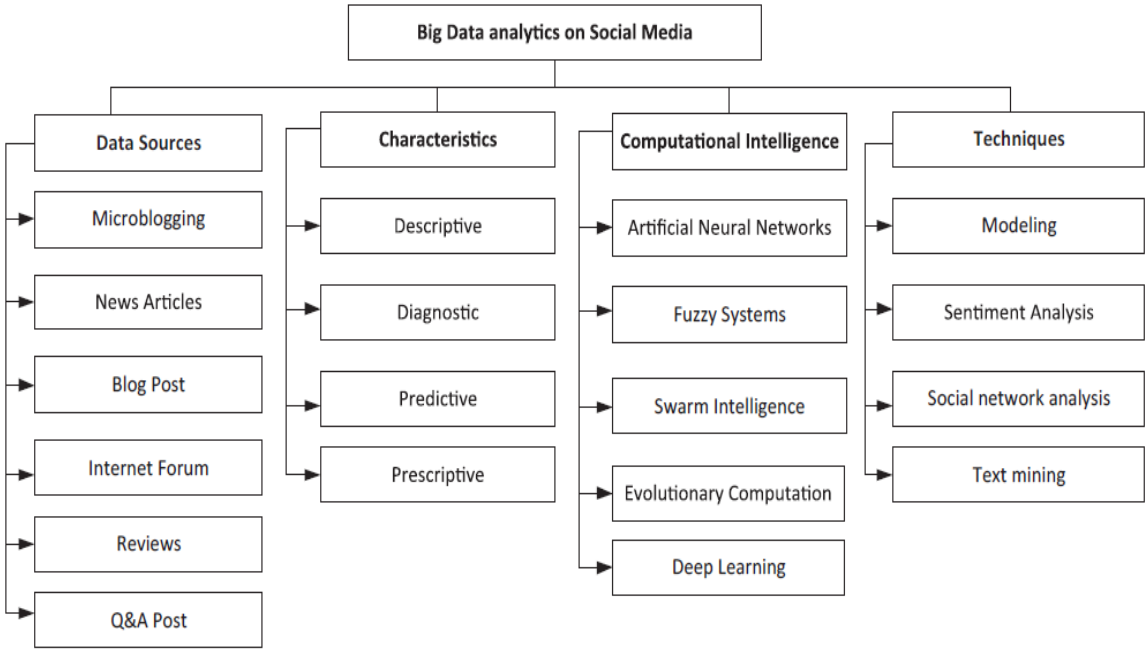


FIGURE 7: CLASSIFICATION OF BIG DATA ANALYSIS ON SOCIAL MEDIA

CHAPTER THREE:

5. Process of social media analytics

Social media analysis involves three main stages: identifying data, analyzing data, and interpreting information. To maximize the value derived at any point in the process, analysts can define a question to answer. Important questions for data analysis are: "Who? What? Where? When? Why? And how?" These questions will help determine the right data sources to assess, which may influence the type of analysis that can be performed [28].

5.1 Data identification

Data identification is the process of identifying available subsets of data on which to focus for analysis. Raw data is useful when interpreted. Once the data is analyzed, it can begin to transmit a message. All data that conveys a meaningful message becomes information. At a high level, unprocessed data takes the following forms to translate into an accurate message: noisy data; relevant and irrelevant data, filtered data; only relevant data and information; data that conveys a vague message, knowledge; data that conveys a precise message, wisdom; data that conveys the exact message and the reason behind it. To derive wisdom from unprocessed data, we need to start processing it, refine the data set to include the data we want to focus on, and organize the data to identify the information. In the context of social media analytics, identifying data means "what" content is interesting. In addition to the text of the content, we want to know: who wrote the text? Where was it found or on what social network did it appear? Are we interested in information from a specific location? When has someone said something on social media? [28].

The data attributes to be taken into account are as follows [28].

- Structure: Structured data is data that has been organized in a formatted repository - usually a database - so that its elements can be addressed for more efficient processing and analysis. Unstructured data, unlike structured data, is the least formatted data.
- Language: language becomes important if we want to know the sentiment of a message rather than the number of mentions.
- Region: It is important to ensure that the data included in the analysis comes only from the region of the world on which the analysis is focused. For example, if the goal is to identify drinking water issues in India, we would like to make sure that the data collected is only from India.
- Content type: data content can be text (written text that is easy to read and understand if you know the language), photos (drawings, simple sketches or photographs), audio (audio recordings of books, articles, presentations, or discussions), or Videos (recording, live broadcast).
- Location: Social media content is generated in a variety of places such as news sites and social networking sites (e.g. Facebook, Twitter). Depending on the type of project for which the data is collected, the location becomes very important.
- Time: It is important to collect the data displayed in the analyzed time period.
- Data ownership: is the data private or publicly available? Are there any copyrights on the data? These are the important questions to address before collecting data.

5.2 Data analysis

Data analysis is a series of activities that help turn raw data into information, which in turn leads to a new base of knowledge and business value. In other words, data analysis is the phase in which filtered data is used as input and turned into valuable information for analysts. There are many types of analysis that can be performed on social media data, including analyzing posts,

sentiments, sentiment factors, geography, demographics, and more. The data analysis step begins as soon as we know what problem we want to solve and we have enough data to get a meaningful result. How do we know if we have enough evidence to warrant a conclusion? The answer to this question is: we don't know. We can't know if we don't start analyzing the data. When analyzing whether the data is insufficient, repeat the first phase and change the question. If the data is deemed sufficient for analysis, we need to create a data model [28].

Data model development is a process or method by which we organize data elements and normalize the relationship between individual data elements. This step is important because we want to run a computer program on the data. We need to be able to tell the computer which words or topics are important and whether certain words relate to the topic we are studying.

When analyzing our data, it is helpful to have several tools in place to look at the discussions surrounding the topic from a different perspective. The aim is to configure the maximum number of tools to run for a given task. For example, when we think of a word cloud, we say "the IT architect" and create a word cloud, arguably the largest word in the world. Cloud would be an "architect". This analysis also applies to the use of the tools. Some tools are good at setting the mood while others are better at breaking down text into a grammatical form that allows us to better understand the meaning and usage of different words or phrases. When performing analytical analysis, it is difficult to list each step on an analytical journey. Rather, it is an iterative approach, as there is no prescribed procedure [28].

The taxonomy and the lessons learned from this analysis are as follows [28]:

Depth of Analysis: Simple descriptive statistics based on streaming data, ad hoc analysis of accumulated data or in-depth analysis of accumulated data. This dimension of analysis is really determined by the time that is available to produce the results of a project. This can be viewed as a broad continuum with analysis time varying from a few hours on one end to several months on the other end. This analysis can answer the following types of questions:

How many people mentioned Wikipedia in their tweets?

Which politician got the most likes during the debate?

Which competitor is mentioned most often in the context of social business?

Machine Capacity: The amount of CPU required to process records in a reasonable time. In addition to meeting the requirements of the processor, the capacity numbers must also meet the network capacity required to retrieve the data. This analysis could take the form of an ad hoc investigation and an in-depth analysis in real time, almost real time. Real-time analysis of social media is an important tool in understanding the audience's perception of a particular topic as it unfolds to allow for an immediate response or change of course. Real-time analysis assumes that data is ingested into the tool more slowly than real-time. Ad hoc analysis is a process of answering a single specific question. The product of an ad hoc analysis is usually a report or summary of the data. In-depth analysis is an analysis that extends over a long period of time and involves a large amount of data, which usually results in high CPU consumption.

- Analysis area: The analysis area is roughly divided into external social media and internal social media. When people use the term social media, they primarily mean external social media. This includes content generated by popular social media sites such as Twitter, Facebook, and LinkedIn. Internal social media includes the corporate social network, a private social network used to facilitate communication within the company.

- Data Speed: Data speed on social media can be divided into two categories: data at rest and data in motion. The dimensions of the speed of data in motion can answer questions such as: How does the mood of the population towards gamers change during the game? Does the crowd convey a positive vibe to the player who actually loses the game? In these cases, the analysis is performed on arrival. In this analysis, the amount of detail generated is directly correlated with the complexity of the analysis tool or system. A very complex tool creates more detail. The second type of analysis in the context of speed is the analysis of data at rest. This analysis is performed as soon as the data is fully recorded. Performing this analysis can provide information such as: Which of your company's products is mentioned most often compared to the others? How do your products feel relative to a competitor's product?

5.3 Information interpretation

The ideas derived from the analysis can be as different as the original question asked in the first step of the analysis. At this point, the way the data is presented becomes important since non-technical business users are the recipients of the information. How can the data have an effective meaning so that it can be used for good decisions? The visualization (graphic) of information is the answer to this question.

The best visualizations are those that reveal something new about the underlying models and the relationships that hold the data. Uncovering and understanding models plays a key role in the decision-making process. There are three main criteria to consider when visualizing data [28]:

- Understand the audience: Before creating the visualization, establish a goal that is to convey large amounts of information in a format that is easy for the information consumer to understand. It is important to answer, "Who is the audience?" And "Can you assume that the public is aware of the terminology used?" An expert audience has different expectations than a general audience. Hence, expectations must be taken into account.
- Establish a clear framework: The analyst must ensure that the visualization is syntactically and semantically correct. For example, if you use a symbol, the size, color and position of the element should be similar to the object represented and convey the meaning to the viewer.
- Tell a story: Analytical information is complex and difficult to digest. Therefore, the visualization serves to understand and understand the information. The storytelling helps the viewer understand the data better. The visualization should summarize information into a structure that is presented as a story and easy to remember. This is important in many scenarios where the analyst is not the same person as a decision maker.

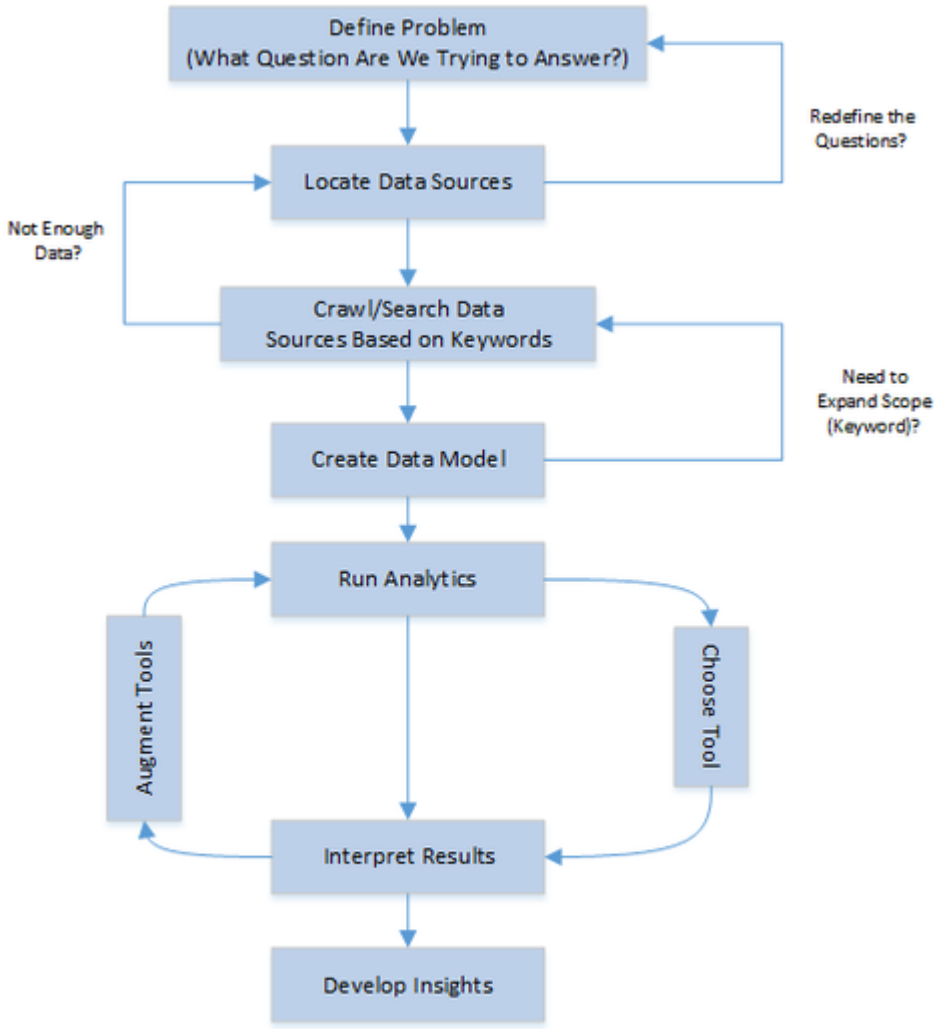


FIGURE 8

6. Key Areas for Social Media Analytics

Social media marketing can really benefit from analysis, this is a look into exactly where measuring need to be done, and why.

Here are the key areas:

6.1Audience Analytics

Communities are an important aspect of networks and are important for exploring a network and forecasting connections that have not yet been observed. Community discovery is essentially a data cluster issue where the goal is to reasonably assign each node to a community or cluster. What is the audience of this particular entity (brand, individuals…) [26]?

It all starts with the audience. Knowing who the target audience for the business is essential. It helps to develop an effective marketing strategy to the public that supports communities on the path to conversion.

53% of marketers agree that more personalized content should be served. It means knowing what customers want and how they will react to efforts to get them interested in the product / service. Hyper-personalized “smart” content becomes the norm [26].

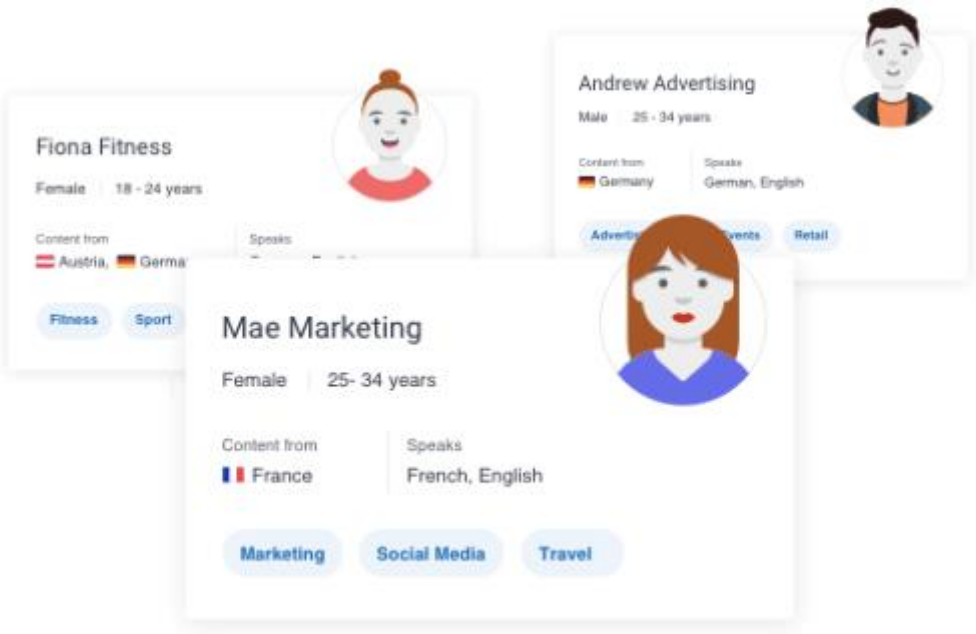


FIGURE 9

That way, marketers can instantly see who their digital audience is, their ideal customer, and ensure that their broader marketing strategy is aligned with the characters' traits, interests, and behaviors. You can also use this information to reach new audiences and create lots of business opportunities.

The most important thing is to keep track of your audience's evolution. This is a great way to keep your content and strategy updated and effective.

In fact, the right strategy has one immediate benefit: savings on the advertising budget. The more targeted and personalized the ad, the lower the cost. Knowing your audience is important to reduce your cost per click and run your ad more often.

These tactics are all useful, if not essential, but the bottom line is performance. A marketer needs to wonder what good social media performance looks like and whether their own performance is improving or deteriorating over time. The next question is, "Can performance levels be measured as they rise and fall?" [26].

6.2 Social Media Performance Analytics

Measuring performance is key to understanding where the strategy is working and getting a good ROI, and where a correction is needed (ROI).

When the social media marketing budget has grown dramatically over the past year. A higher ROI (return on investment) is expected, but the influence of performance on ROI (return on investment) needs to be proven [26].

Key performance indicators to monitor:

- Insight into interactions across platforms and over time to understand whether published content effectively engages audiences.
- View both the cumulative number of interactions and the number of interactions per 1,000 subscribers to see how well the community is responding to messages.
- Increase the number of clicks on your posts to see how effectively you are driving social media traffic to the web.
- Monitor subscriber growth over time to see if the audience is growing as a result of the team's efforts on social media.

Again, tracking all of these metrics over time is very important to spot larger trends, understand the results of your business's social media strategy, and see what return you're getting on your investment [26].

6.3 Competitive Analytics

To better understand performance metrics, they need to be viewed in a competitive context. This is where competitive analysis and benchmarking come into play.

FIGURE 10

Almost every CMO wants to improve their performance in two key areas: effectiveness (getting the result you want) and efficiency (reducing waste), but it's often not clear how much they need to improve?

Benchmarking is a way to evaluate business performance and understand changes.

Comparing the company's performance on social media to that of its competitors is the best way to gauge work effectiveness and team strategy. It also helps to find out whether the performance and return on investment are successful relative to the market. It can also be compared to multiple competitors.

It is best to conduct regular competitive scans to keep an eye on competitive news and advise teams on strategic actions they can take to help the company stay ahead [26].

With modern social media analytics, especially when based on AI, the benchmarking solution enables the performance of competitors based on industry, country and region. It helps to identify their strategy and easily be one step ahead of them [26].

6.4 Paid Social Media Analytics

Social media advertising budgets have doubled globally from $ 16 billion in 2014 to $ 31 billion in 2016. This trend will continue in 2020. A lot of money is being spent. Therefore, it is absolutely essential that businesses know the effectiveness of their social media ad spend [26].

It's too easy to waste time and resources promoting poorly performing content when the business owner has no idea what paid content their audience is responding to well.

Some companies use artificial intelligence-based tools that predict exactly what content should be under budget for the best results - and early adopters have great results with it.

Paid social media advertising can seem complex with so many moving parts, and advertising spend reporting can be complicated with so many channels, accounts, and profiles. To keep track of expenses, it is important to put all the conduits in the same place before starting the measurement.

CHAPTER THREE:

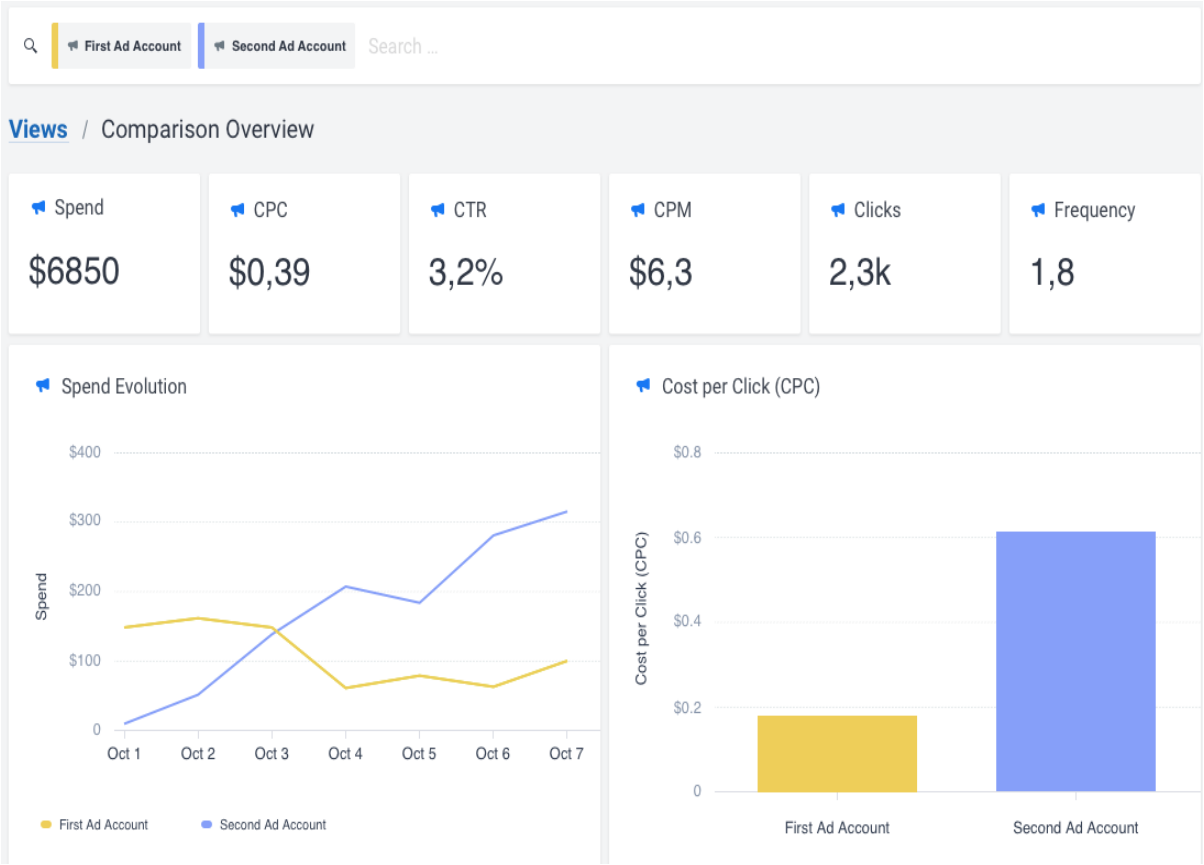


FIGURE 11

Here are some basic stats to keep in mind [26]:

- Number of announcements
- Total of expenses
- Clicks
- Click rate
- Cost per click
- Cost per commitment
- Cost per action
- Cost per purchase
- KING of the campaign

These KPIs show exactly where the money is going, how much is paid for various aspects of advertising campaigns, and how successful the effort is.
Monitoring paid social media metrics - both owned and competitive - is most effective and useful when done on a regular basis. After all, the company probably runs multiple paid campaigns, busy vacation times come and go, and expenses and results can fluctuate.
This helps identify key trends - such as overpayment for clicks - and adjust the budget accordingly.

He may even see opportunities to have a real impact on share of voice by spending more on targeted ads when he sees that the market is spending less.
For businesses working with agencies, paid analytics provides real visibility into where their budget is going and the results they're getting for their money. Agencies can easily transmit campaign data to clients in real time, with important data visualized.

6.5 Customer Service and Community Management Analytics

With 67% of consumers leveraging social media to seek resolution for issues, we simply can't underestimate the power of well-functioning social media customer service. That's why the insurance that your teams are doing an excellent job handling customer request is needed [26].
How? We need to look into the community management teams 'key performance metric - the average response time - and see what it looks like for particular team members and platforms.

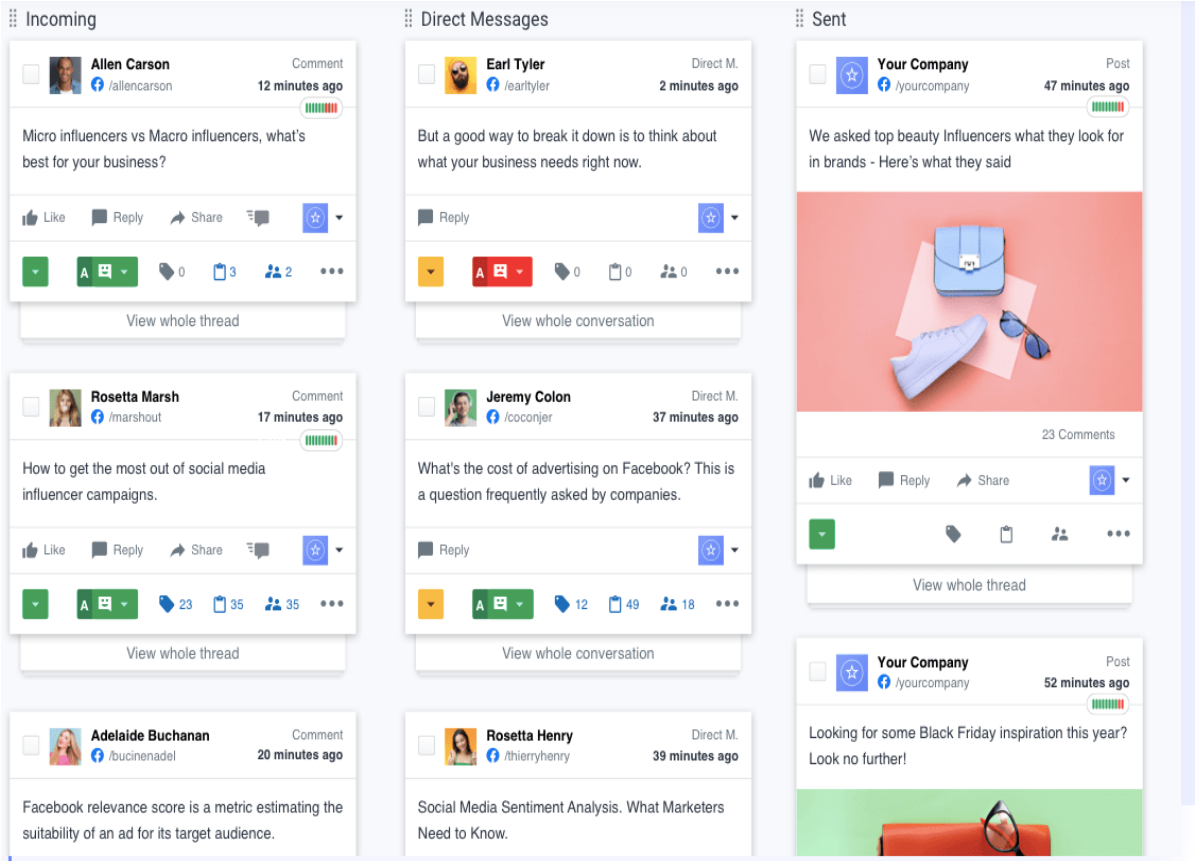


FIGURE 12

Monitoring your community management teams 'performance metrics, we can ensure that the business communication is appropriate and timely. This, in turn, translates to enhanced brand image and stronger, lasting relationships with the customers and prospects [26].
Without a supporting analytics platform, teams and managers will rapidly have no idea many customer queries they get, how long they take to answer, or indeed, if they have answered them all. Things can get messy fast.
The answer is to measure every metric that matters to your customer care and then monitor them - it's the only way to systematically improve.

It's so important to nurture the audience across all digital touchpoints which means keeping an eye on how fast the team answer and solve customer queries, and how audience sentiment fluctuates around the brand [26].
It's also crucial to track sentiment, as it's such an important aspect of customer care, but there's more about that in the Sentiment Analysis section a bit further down.

6.6 Influencer Performance Analytics

In the graph community, centrality metrics deal with the nodes 'positions in the network and are typically used for measuring the dominance of nodes, quantifying the strength of connections and uncovering the patterns of influence diffusion. In OSN (Online Social Network) a critical research topic is to identify 'experienced' or 'trusted' users that may be trendsetters since their opinionated posts are the ones that can rapidly spread far and wide in the network enabling them to influence other users. An interesting fact regarding trendsetting, is that, how much credence another person gives to a post may depend on how many times they hear it from different sources (Flow) and not how soon they hear it (Geodesic Distance). Identification of influential users and of whether individuals would still propagate information in the absence of social signals about that information are two elements required to be studied in order to study information flow in OSNs [24].

Hiring influencers can be pricey, so we need to make sure that we spend budget collaborating with the right partners on the right campaigns. We need data to verify influencer choice, and only work with influencers whose performance metrics are high around your key topics. Tracking influencers 'key performance metrics will help make the right decision and make the most of influencer marketing.

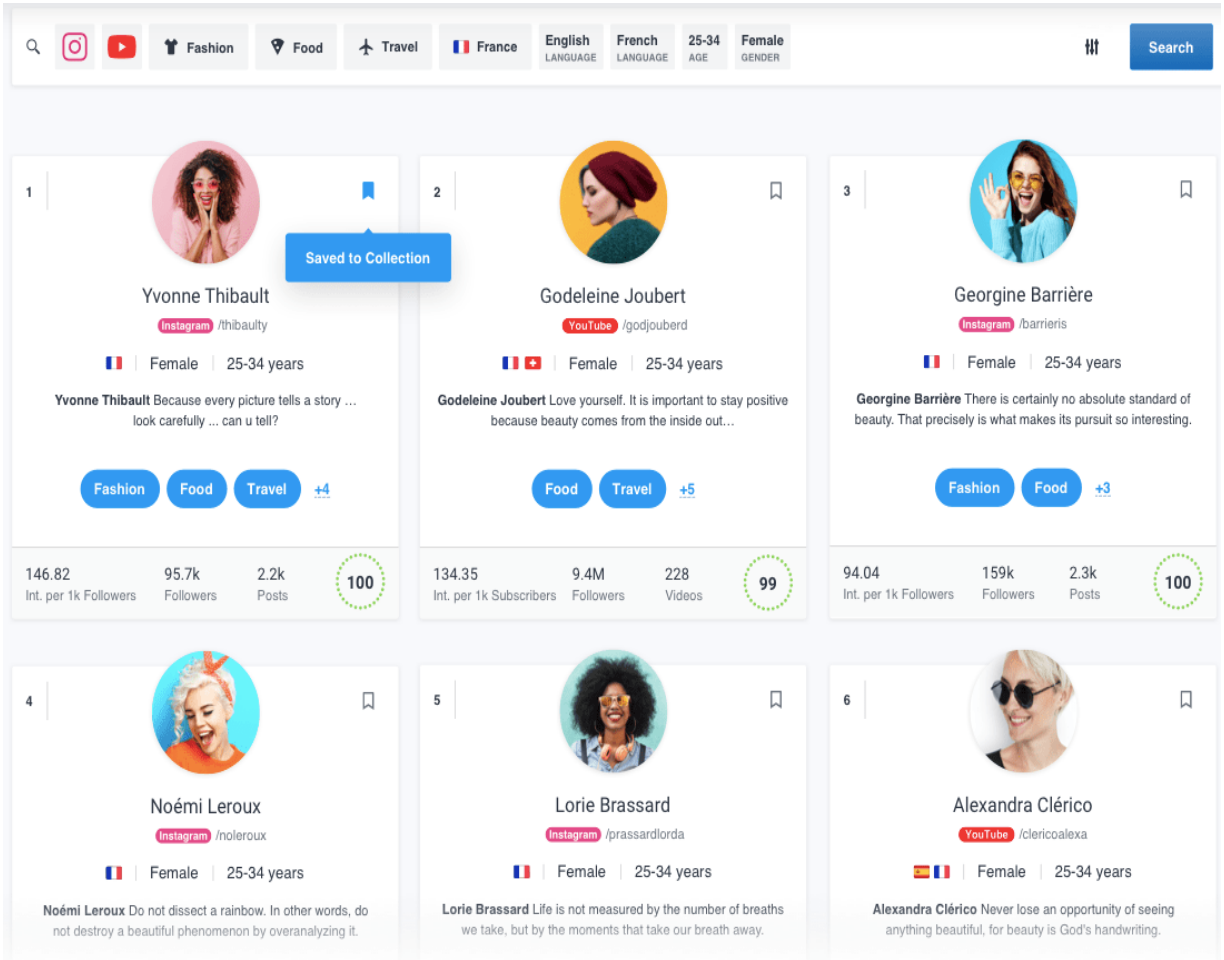


FIGURE 13

Influencer marketing should be measured like any part of marketing is. Analytics help you organize every part of the influencer campaigns - from finding the right influencers to measuring how much their campaigns helped the business. But which influencer metrics should we track? Here's some key metrics [26]:

• Interactions per 1,000 followers will help understand if they're effectively generating engagement.

• Audience size to estimate reach. Reach alone is not enough if the influencer doesn't have the right audience, which is why it's crucial to measure their key metrics before deciding to collaborate.

• Number of posts to see how active they are. We should also check the brands the influencer worked with in the past to verify their authentic and get a better understanding of the verticals they're working in.

• Have a close look at any past collaborations the influencer has done, and verify that they have a business license. Influencer fraud is on the rise and canny businesses protect themselves by doing their due diligence.

6.7 Sentiment Analysis

Sentiment Analysis (SA) is an ongoing area of research in data analysis that determines what others think about entities, individuals, issues, events, topics. It refers to the recognition of polarity as positive or negative for a particular entity or in general [24].

The idea is to find the feel of the text by classifying it as positive, negative or neutral. This analysis is generally used for binary decisions; H. Users like or dislike something, or the product is good or bad (Ohbe, Ozono & Shintani, 2017). Sentiment analysis, also known as opinion extraction, is the categorization of consumer attitudes, emotions, and opinions about a company's products, brands, or services. Sentiment analysis has a variety of uses on social media. For example, this analysis can be used to identify the sentiments of consumers in a marketing and customer service department, which can be used to determine whether consumers are satisfied or dissatisfied with a product (Povoda, Burget, Dutta & Sengar, 2017) [25].

If sentiment is not measured, businesses run the risk of alienating their social media audience and undermining the perception of their brand in the eyes of customers and prospects.

Most businesses post content on their social media profiles every day or every few days. Analyzing audience sentiment is essential for sustaining positive fan growth, interactions, engagement rates, and later conversions.

Sentiment Analysis helps track mentions online in real time, making it easy to see if a potential PR crisis is looming. Organizations can quickly spot an increase in negative sentiment, investigate it immediately, and take action to alleviate it.

On social media, news travels fast, and negative comments get the most attention and spread the fastest. Unless you are dealing with dissatisfied customers, it is risky. They can share their anger more widely on their social media accounts. We absolutely want to avoid this.

It all starts with the audience, and sentiment analysis is the most basic social media analysis that can be done. Once the brand perception has been damaged, it is very difficult to repair it. That's why it's important to know whether or not the audience is happy with the brand and the content.

CHAPTER THREE:

7. Tracking Unique Metrics from Each Platform

Regardless of the platform, a solid understanding of the performance of any network is required. Social networks return the favor by offering their own native analytics.
Let's take a look at the most popular native tools below [26]:

7.1 Facebook Insights

For those who have a Facebook business page, we can analyze certain KPIs (Key Performance Indicators) within the social network. The most important Facebook metrics are:

- Engagement: This statistic shows the number of clicks, likes, comments, and approvals over the past seven days. In addition, the data is compared to the previous week.
- Impressions: The number of times the Facebook page is viewed, including those who click and do not click on the content or page.
- Organic Likes: The number of people who liked the page without coming from an ad campaign.
- Page likes: This metric shows the total number of pages likes and new likes pages with weekly data comparisons.

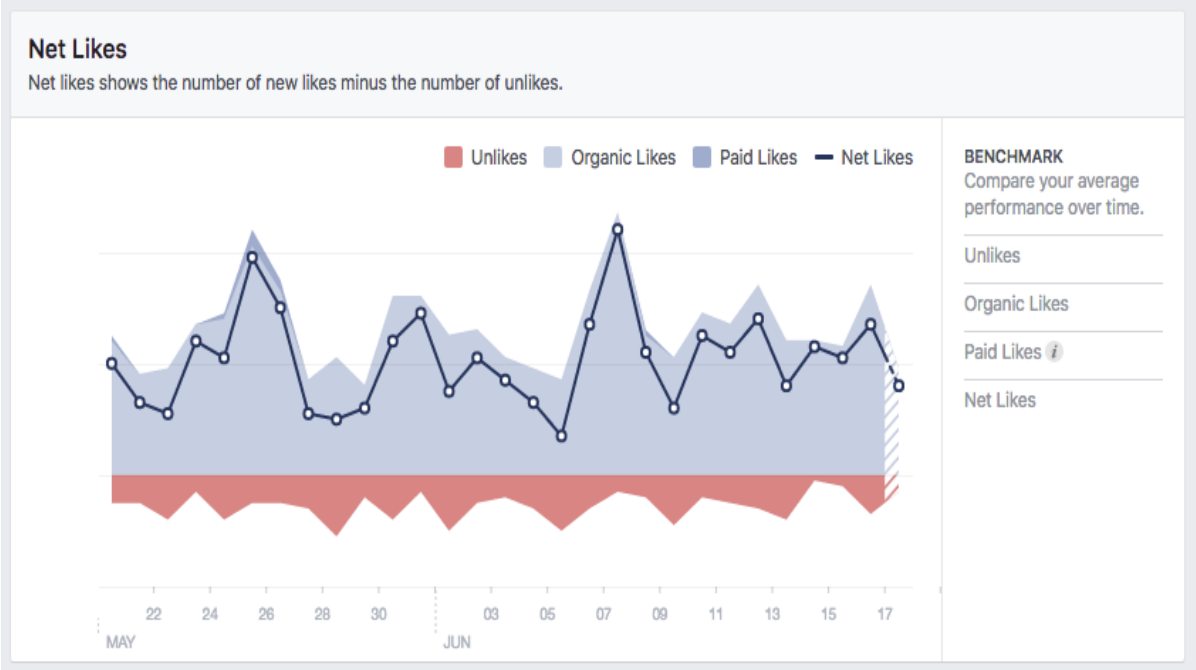


FIGURE 14

• Paid likes: the number of people who liked the page directly from a paid advertising campaign.
• Publish audience: this statistic indicates the total audience. This is the number of people who saw content or ads associated with the page. There is also the page reach, which is the number of views (impressions) for the page contributions.
• Reactions: This metric shows the different reactions that users have posted to the post, including Like, Love, Haha, Wow, Sad, and Angry.
• Dislikes: The number of people who dislike the Facebook page.

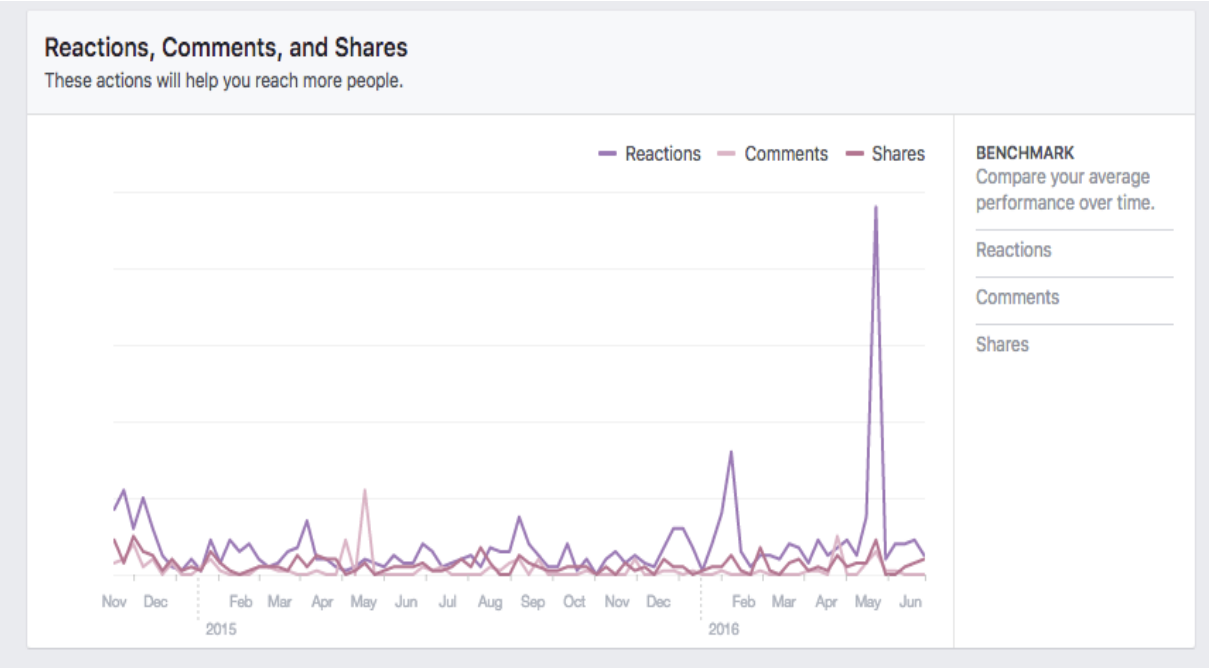


FIGURE 15

7.2 Instagram Insights

Instagram's native information is split into two sections, with metrics focused on individual posts and profile data:

- Account Impressions: The number of times the posts and stories have been viewed.
- Total Audience: This statistic tracks the number of unique accounts that viewed articles and stories.
- Website Clicks: This metric tracks the number of clicks on the bio link.
- Profile visits: the number of clicks on the account page.

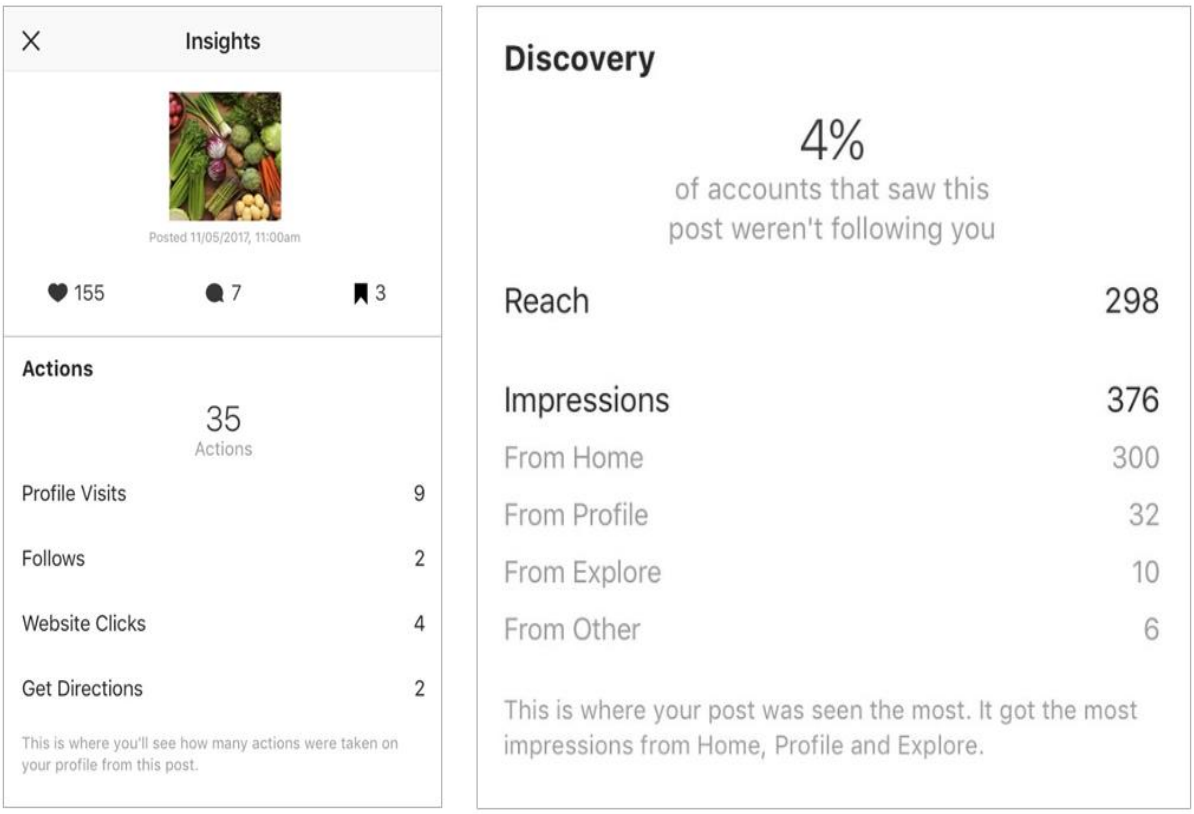


FIGURE 16

• Post-Likes: the number of likes that a particular post receives.
• Comments on posts: the number of comments collected on a particular article.
• Posts Saved: This statistic records the number of unique accounts saved in the post.
• Tracking: allows you to track the number of accounts that tracked the account over a given period.

7.3 Twitter Analytics

Whether you use Twitter for business or personal use, you have access to their analytics just by having an account. The dashboard gives a 28-day summary of the content and other key Twitter data. Here are some of the most important Twitter metrics to follow:

- Engagement Rate: Total link clicks, Retweets, favorites and replies on the Tweet divided by total impressions.
- Followers: Total number of Twitter followers.
- Link Clicks: Total number of URL and hashtag links clicked.
- Mentions: How many times the @username was mentioned by others.

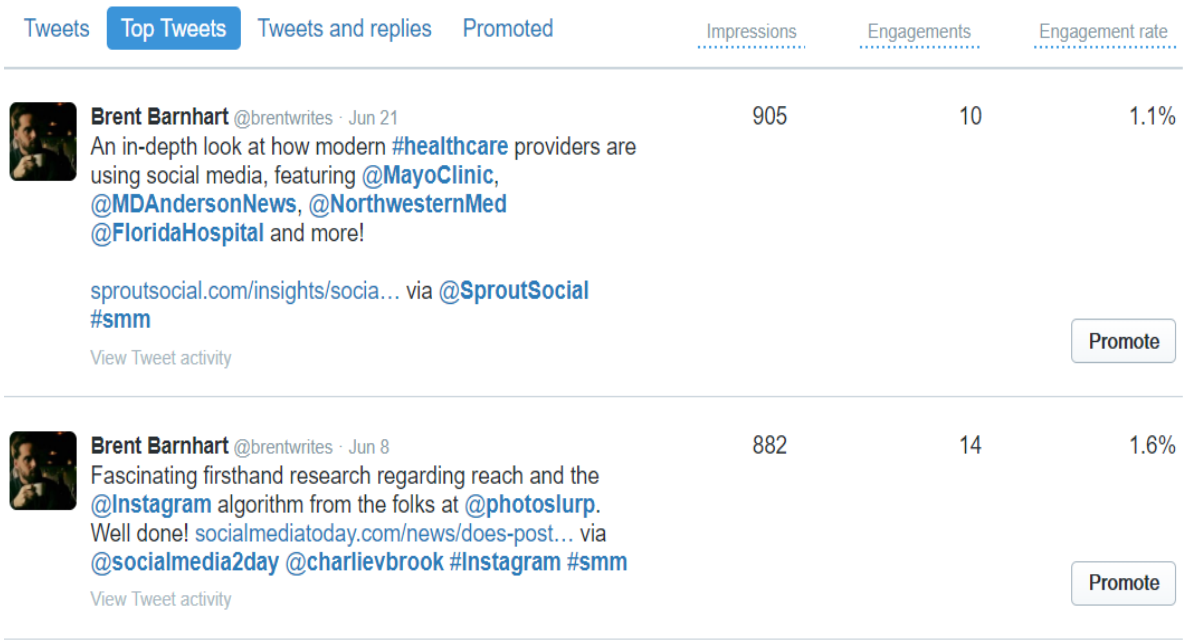


FIGURE 17

- Profile Visits: Total Twitter profile visits.
- Replies: How many times people replied to the Tweets.
- Retweets: Total retweets received by others.
- Tweet Impressions: Total of times the Tweet has been viewed whether it was clicked or not.
- Tweets: How many Tweets and account posted.

FIGURE 18

7.4 LinkedIn Analytics

Another popular social media platform with built-in analytics tools is LinkedIn. You can access LinkedIn Analytics through the Company Page. This shows all the social media data going into the LinkedIn Page. Here are the top LinkedIn metrics:

- Clicks: Total clicks on a post, company name or logo.
- Engagement: Total interactions divided by the number of impressions.

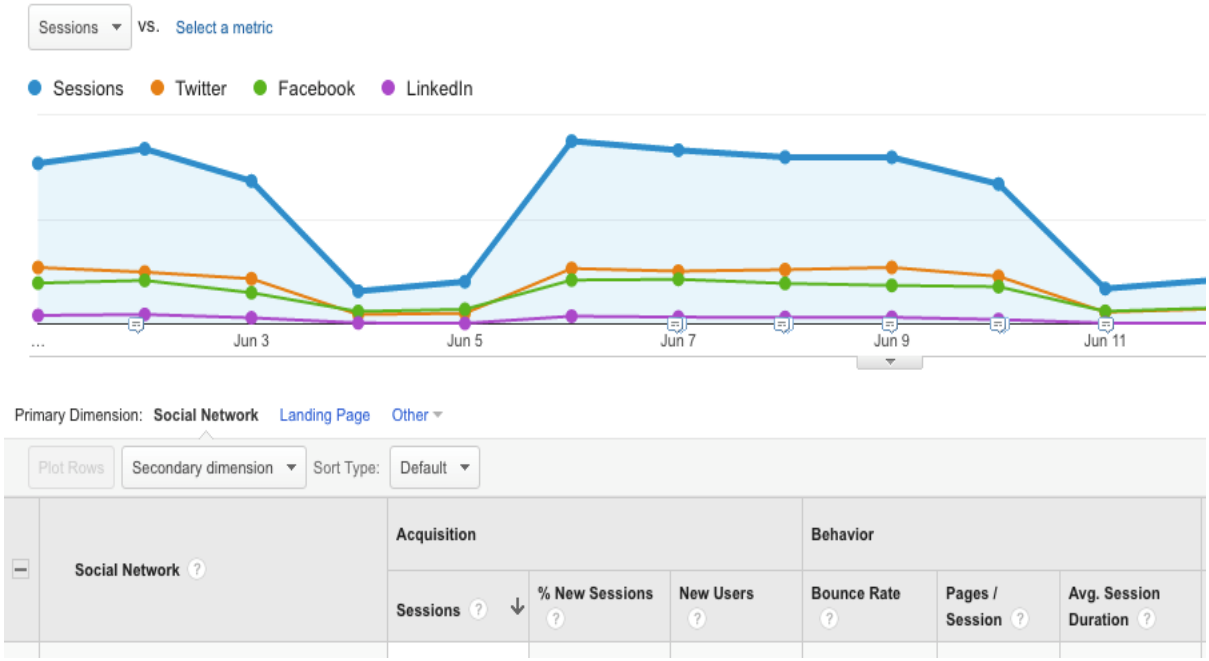
- Followers: Total number of new followers through a sponsored update.
- Impressions: Total times the update was visible to other users.
- Interactions: Total number of comments, likes, comments and shares.

FIGURE 19

7.5 Google Analytics

While the other platforms provide helpful insights, Google Analytics steps up data game. Here you can learn about product sales, leads, guide downloads, duration times and much more. When it comes to social media data, there are a few metrics to note in Google Analytics:

- Average Session Duration: Average session times users spend on your site.
- Bounce Rate: Percentage of users leaving the site after a single view.
- New Users: Total number of new users coming to the site for the first time.



- Pages / Session: Average number of pages a user views each session.
- Page views: Number of pages loaded or reloaded in a browser.
- Sessions: Total times when users are active on the site.

	Social Network	Acquisition			Behavior			Conversions	
		Sessions ↓	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration	Goal Conversion Rate	Goal Completions
		180 % of Total: 0.55% (32,509)	86.11% Avg for View: 90.85% (-5.22%)	155 % of Total: 0.52% (29,536)	92.78% Avg for View: 91.13% (1.81%)	1.12 Avg for View: 1.14 (-1.43%)	00:00:23 Avg for View: 00:00:39 (-41.16%)	0.00% Avg for View: 0.00% (0.00%)	0 % of Total: 0.00% (0)
☐	1. Facebook	140 (77.78%)	85.71%	120 (77.42%)	92.86%	1.12	00:00:17	0.00%	0 (0.00%)
☐	2. Pinterest	22 (12.22%)	86.36%	19 (12.26%)	95.45%	1.14	00:00:15	0.00%	0 (0.00%)
☐	3. Twitter	9 (5.00%)	88.89%	8 (5.16%)	77.78%	1.22	00:02:27	0.00%	0 (0.00%)
☐	4. StumbleUpon	6 (3.33%)	100.00%	6 (3.87%)	100.00%	1.00	00:00:00	0.00%	0 (0.00%)
☐	5. Blogger	1 (0.56%)	0.00%	0 (0.00%)	100.00%	1.00	00:00:00	0.00%	0 (0.00%)
☐	6. Instagram	1 (0.56%)	100.00%	1 (0.65%)	100.00%	1.00	00:00:00	0.00%	0 (0.00%)
☐	7. LinkedIn	1 (0.56%)	100.00%	1 (0.65%)	100.00%	1.00	00:00:00	0.00%	0 (0.00%)

FIGURE 20

8. Conclusion

Social media big data along with the progress in analytics tools have emerged as the key to crucial insights into human behavior and are continually stored and processed by corporations, individuals, and governments. First, we defined the social media and the analytics process putting light on some of its benefits. Second, we classify the analytics based on important aspects, such as data types, characteristics, and techniques then we described the principal analytics process steps. Third, we provided some key areas for social media analytics with their purposes and examples, then bring forth some tools and use cases examples (we focused on the native tools to clarify the difference of each social media and their quality attributes)

IV. CHAPTER 04: REAL TIME SOCIAL MEDIA ANALYTICS

1. Introduction

This chapter present the result of this work, we tried to explore the deferent parts that this chapter relies on (Big Data, real time analytics, social media analytics) from the previous chapters and present their combination in a technology stack that handle the analysis of the Social Media Big Data in Real Time and putting the light on its use cases (analyzing Social Media Data like Twitter) advantages by providing some performance ,functionalities and features analysis and disadvantages so the reader will have easy time understanding, choosing and even making improvements.

The purpose of real time analysis is to give real time insights which lead to quick decisions that help containing a dangerous spreading problem or help for fast plan adaptation so a fault tolerant and very low latency data ingestion is required (see chapter 2), and by adding the social media to the equation the challenges rise duo the structure of the data that the social media issues (see chapter 3) which hardens the ingestion and the analyzing process , let's say that everything is good for now, the data get analyzed in real time and the alerts pop up in time and dangers situations got avoided or plans putted to execution, the biggest data size get the more the insights and the decisions become reliable (see chapter 3), and here the third variable equation pops up , the size of data, by time the data will become bigger and non-stopping bigger regardless the source of data , the data stored becomes Big Data (see chapter 1), and if we are talking about a social media data source that's an even bigger data and very faster growing, so we need a very robust, scalable fast data exploration to handle this enormous and fast growing data. All these problems and more are considered by using this stack.

We also approach an architecture used by Twitter company and clarified what's the differences and suggest some improvements by using this stack or some parts of it.

2. What we need from this stack?

First of all, the analytics must be real time and real time means **very low latency** from the beginning of the data generation to the dashboard of the user, this is the main feature that need to be fulfilled by this analysis technology stack. (Chapter 2)

After the first criteria clarified, the next most important criteria's can be delivered which is the **scalability**, the stack need to handle the large number of producers or few but with enormous data set that can be generated and delivered over long or short time, from the data ingesting process to the data visualization layer.

After those two criteria's we can mention some other criteria's like the **reliability** which is well needed also to not miss the picks of data specially in social media data case which is like we mention earlier about its data specifications are very fast evolving, we can imagine a scenario of a 60 minute shutdown the system can't feed the data visualization layer so the dashboards won't be updated, in this time with the spreading speed of the information in the social media, the damage can be significant for the brand, company or even individuals, if the alert was in time we can say the damage could be avoided or the sure thing is it could be reduced to a good amount duo to the in-time handling.

Also, we need to have useful and effective functions to help exploring the social media data.

So, we need from our stack to be:

- Very low latency.
- Scalable.

- Reliable.
- Adequate functions for social media (Twitter)

3. The Tech-Stack (Kafka, DRUID, METATRON discovery)

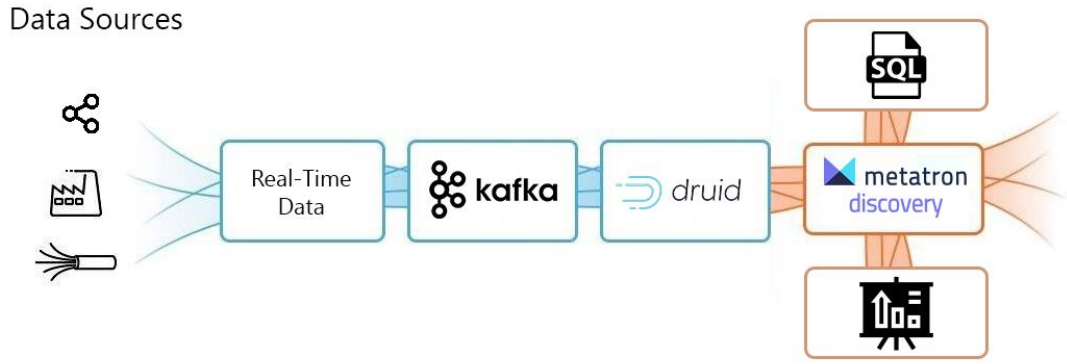


FIGURE 21: KDM REAL TIME ANALYTICS STACK

3.1 Why Apache Kafka?

Apache Kafka is currently the most popular framework used to ingest the data streams into the processing platforms. More than 30% of 500 successful company which are considered as leaders in their fields, travel companies, banks, insurance companies, telecom companies, and much more use Kafka. Let's take LinkedIn or Microsoft for example they process 10^{12} messages a day with Kafka [27].

From the available messaging systems **Kafka is performance-oriented,** it has **low latency**, **scalable**, durable, and fault-tolerant publish-subscribe messaging system, and has higher throughput, **reliability**, and replication characteristics, which makes it applicable for things like tracking service calls (tracks every call) or tracking IoT sensor data where a traditional Messaging Oriented Middleware might not be considered.

It will match our needs for a real time analytics very well [28] [29].

3.2 Why Apache Druid?

Apache Druid is considered as a real-time analytics database who is designed in the first aim for fast slice-and-dice analytics[4] on large data sets [30]. It can be used as a database for powering real time analytics use cases, it was designed to reach the most performance and scalability that can be reached around ingesting and exploring large quantities of transactional events by combining various techniques, ideas and architectures from various dedicated software's to solve their particular problems.

In other explanation Druid was designed to satisfy the use cases where the ability to rapidly slice and dice and drill into the data effectively without any restrictions along with very low latency queries (complicated or not), which needed for an interactive data exploration, also its design satisfied the ability to make the ingested data explorative almost in real time after their occurrence, this is a critical feature to be able to call it real time analytics.

By analyzing its key features, we can group them by criteria's, the key features that help solving **low latency** criteria are [30]:

- **Columnar storage format.** Using column-oriented storage, that's mean it only needs a particular query to load the exact columns. This boosts the queries that only hit a few

[4] Known as well by OLAP queries

columns. Moreover, each column is stored optimized for its particular data type, and that insure a fast scans and aggregations.

- **Massively parallel processing.** Druid can process the queries across the entire cluster in parallel way.
- **Real time or batch ingestion.** Druid can ingest data either real-time (ingested data is immediately available for querying) or in batches.
- **Indexes for quick filtering.** Druid uses roaring or CONCISE compressed bitmap indexes to create indexes that power fast filtering and searching across multiple columns.
- **Automatic summarization at ingest time.** Druid optionally supports data summarization at ingestion time. This summarization partially pre-aggregates your data, and can lead to big costs savings and performance boosts.
- **Time-based partitioning.** First the Druid divides data by time, and it can additionally divide it according to other fields. This means time-based queries will only access the partitions that match the time range of the query. Which leads to significant performance improvements for time-based data.
- **Approximate algorithms.** Histograms and quantiles of computation of approximate, approximate ranking, and algorithms for approximate count-distinct are included in Druid. Those algorithms offer bounded memory usage and they are faster than exact computations. In addition, Druid offers exact count-distinct and exact ranking when accuracy is more important than speed.

The key features that help solving the **scalability** criteria are:

- **Scalable distributed system.** Druid is typically deployed in clusters of tens to hundreds of servers, and can offer ingest rates of millions of records/secs, retention of trillions of records, and query latencies of sub-second to a few seconds.

The key features that help solving the **reliability** criteria are:

- **Self-healing, self-balancing, easy to operate.** To scale the cluster out or in, simply add or remove servers and the cluster will rebalance itself automatically, in the background, without any downtime. The system will automatically route around the damage if any Druid servers fail until the server can be replaced. It is designed to run 24/24 hour for 7/7 day with no need for planned downtimes for any reason, including all configuration changes and software updates needed.
- **Cloud-native, fault-tolerant architecture that won't lose data.** When Druid has ingested your data, a copy will be stored safely in deep storage[5]. Your data is recoverable from the deep storage even if all Druid servers fails. Replication ensures that queries are still available while the system recovers for more limited failures affecting just a few Druid servers.

Druid's quantitative assessments are focused on Query latency and Ingestion latency, but we may ask how fast Druid is, see (Figure 22, Figure 23, Figure 24 for more details see [31], [32], [33])

[5] (typically cloud storage, HDFS, or a shared filesystem)

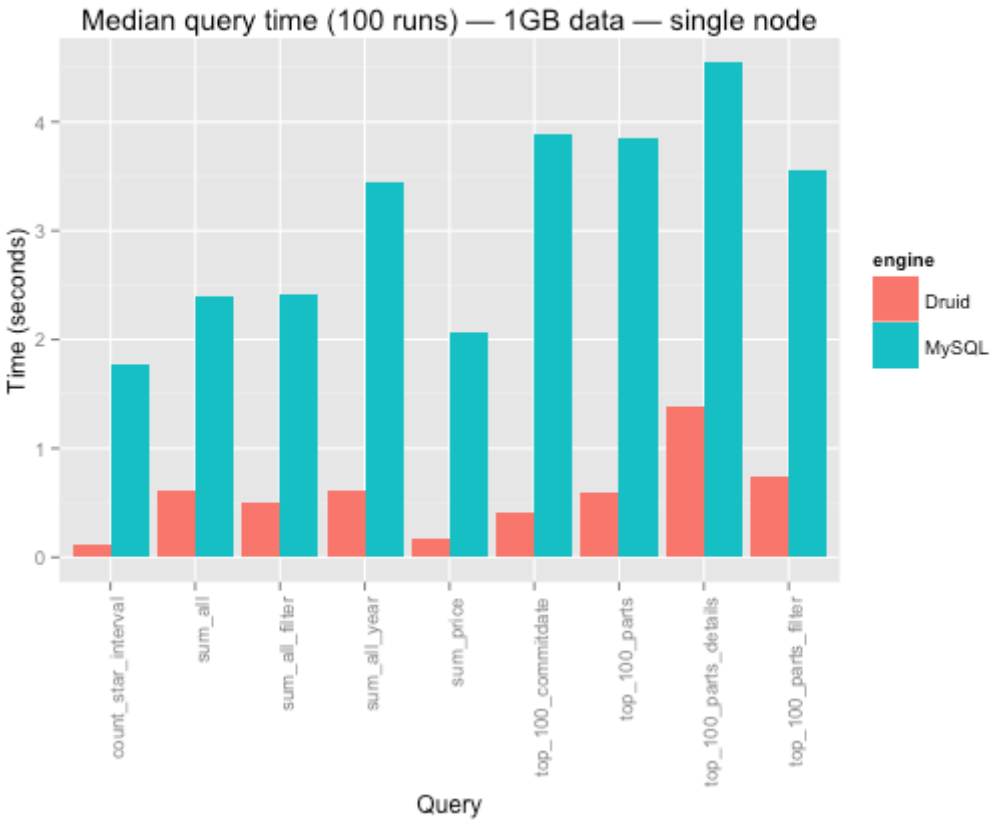


FIGURE 22: DRUID VS MYSQL

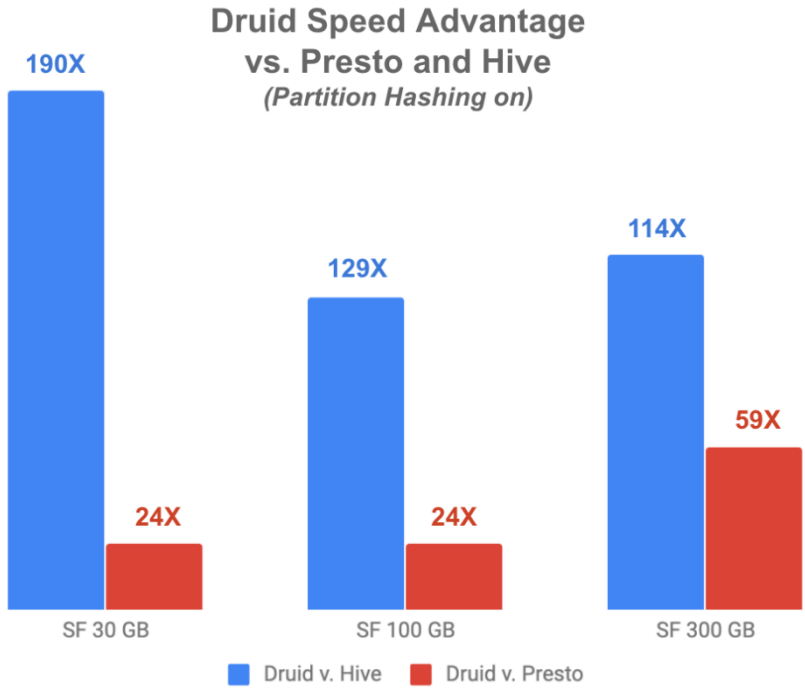


FIGURE 23: DRUID VS HIVE VS PRESTO

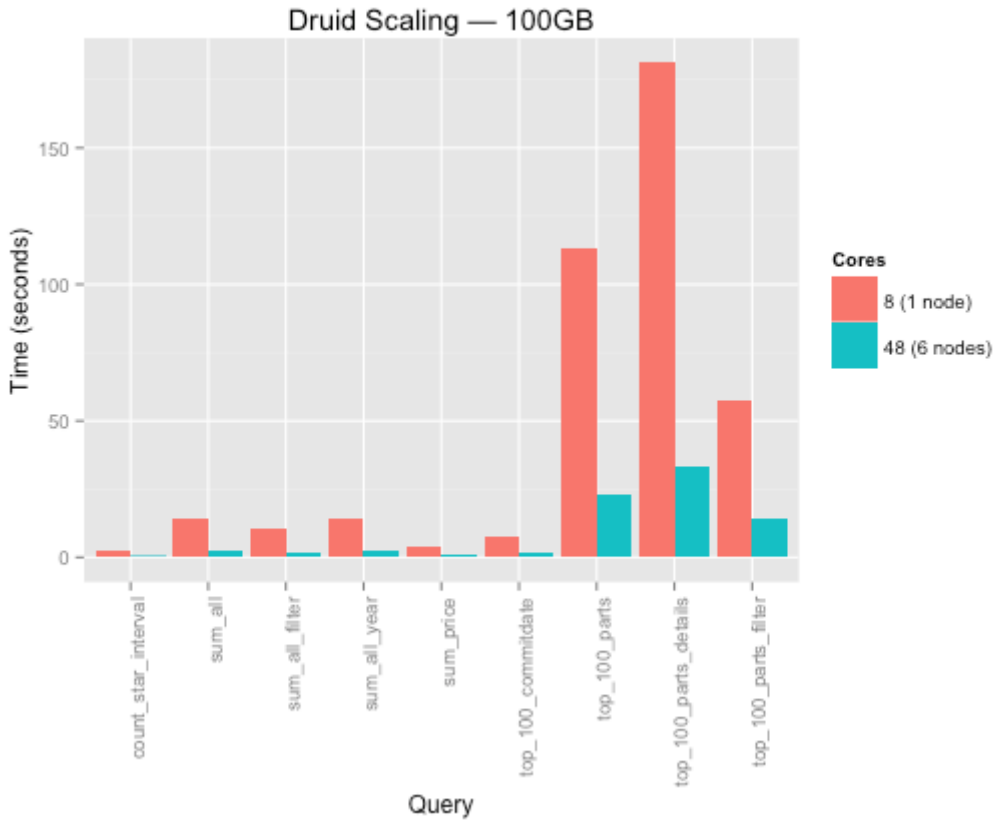


FIGURE 24: DRUID SCALING

TABLE 1: COMPARISON OF BIG DATA TOOLS [37]

BIG DATA TOOLS	HADOOP	STORM	HPCC	HBASE	GRIDGAIN	DRUID
CHARECTERISTICS						
DEVELOPER	Apache Software Foundation	Canonical Ltd	HPCC Systems, Lexis Nexis Risk Solutions	Hadoop Framework	Apache Ignite	**Apache Druid**
PROGRAMMING LANGUAGE	JAVA	Clojure & JAVA	C++, ECL	Many Languages	JAVA Based	**JAVA**
CURRENT VERSION	2.7.0	0.9.3	4.22	V4	3.0	**0.19.0**
COMMUNITY SUPPORT	Distributed File System	Distributed Stream Processing	Data Processing (Batch(/Parallel)	Framework/Cluster/File System)	Hadoop Distributed File System	**Distributed Data Store**
Operating System	Cross Platform	Cross Platform	Linux	Platform Independent	Windows Linux, MAC OS	**Cross Platform**
ORGANIZATIONS	Facebook, eBay, Adobe, Twitter,	Backtype, Twitter	Google, HBase Platforms	Twitter, Webinar	Social Media	**Alibaba, Airbnb, Cisco, Deep.BI,**

	Amazon, Yahoo					**eBay, Lyft, Netflix, PayPal, Pinterest, Twitter, Walmart, Wikimedia Fondation, Yahoo**
SIMPLICITY	Easy to Implement	Difficult to Handle	Flexible to use	Uncomplicated	Effortless	**Easy and Flexible to implement**
PERFORMANCE	Boosted Performance	Accustomed Performance	High Performance	Efficient Performance	Vary Performance	**Interactive (Real Time) Performance**
SOFTWARE TYPE	Open-Source	Open-Source	Open-Source	Freeware	Trail ware	**Open-Source**
EXTENSIBILITY	Hadoop DFS to Teradata	Online Analytic Applications	Extensible to ML Platform	Processing of Big Data	Extensible to Application Scaling	**Real Time Distributed Data Store, Powering a user-facing application**
VOLUM	250PB	One Million 100bytes	Many Petabytes	Petabytes	Terabytes to Petabytes	**trillions of events, Petabytes of Data**
VARIETY	Structured, Semi and Unstructured Data	Cluster of Data	Data Centric and Query Processing	Task Processing and Node Processing	Memory Processing	**event-driven data**
VELOCITY	Fast Collection, Processing and Consumption	High Velocity Data	More Than Hadoop	Complex Processing of Dataset	High Speed and Fast Processing	**Fast Aggregation and query performance**

One more thing about choosing Druid is that Druid is designed to excel at processing(ingesting, aggregating, ,querying) time series event data and The data delivered from social medias (Twitter) can be considered as an events type specially at the real time context, because every Tweet represent an event which is the Tweeting action its self and the creation time of the Tweet is the timestamp of the Tweeting event and the data included in the Tweet (text, hashtags …..) are the event data, see more details on [30] which make Druid the optimal engine and the best choice for what the stack needs.

3.3 Why Metatron Discovery?

Metatron Discovery is the data visualization layer in this stack, it's the tool that will use the previous stack blocks (Kafka, Druid) and harvest the results and introduce it as dashboards ,graphs and insights in other terms help derive value from the data, Metatron Discovery is OLAP-based business intelligence (BI) solution that combines OLAP, visualization, and machine learning technologies and introduced with an easy and clear user interface that even the non-experts users will derive value from their data swiftly.

What make Metatron also as a good choice is that it was built based on Apache Druid engine which make it more compatible with the stack, it was developed by SK Telecom, a telecommunications service provider with the greatest number of subscribers in South Korea to manage the significant amount of data generated by their users, because the lack of an existing IT solution for processing this amount of data, built its own Hadoop infrastructure to store mass amounts of data at low cost but they faced some limitations like the data couldn't be analyzed in real time and having a different solutions and different managers support each stage of data analytics costs a lot of time and resources and produced a poor data accessibility [34].

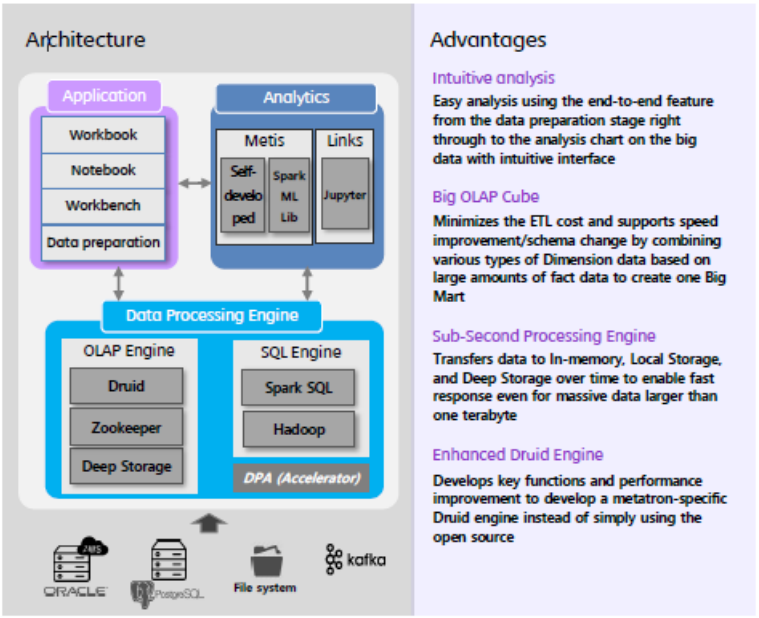


FIGURE 25: ARCHITECTURE AND ADVANTAGES OF THE METATRON DISCOVERY

The open-source Druid has the following limitations [34]:

- Till now, Metatron uses another SQL engine for data preparation since it does not yet have full support for joins;
- Druid supports only a subset of SQL queries;
- For a data lake, a traditional SQL engine is more appropriate;
- Except some unusual cases Druid cannot append to or update already indexed segments;
- Nulls are not allowed;
- Filtering is not supported for metric columns.
- Linear scalability is not ensured. Increasing the number of servers doesn't improve the performance as much.
- Only a few data types are supported and it is difficult to add a new one.
- The management and monitoring tools are not powerful enough.

Druid and Spark are complementary solutions as Druid can be used to accelerate OLAP queries in Spark. which they used it in building Metatron (Discovery Spark Engine), Druid and Spark create a great synergy, Druid is focused on data querying ,ingestion, filtering in very low latency while Spark enable riche API's and data processing, Spark ability to process query data through Spark SQL provided much wider API for data querying in Metatron [3]. Druid improvements results in richer querying API, metrics, functions (expression, aggregation …) and better performance (results can be stored, Bit-slice indexing …) and flexibility (data exportation …) which give wider reception floor for the varied social media

data, and more capabilities to get deeper insights on the data mine of the social media, and without mentioning the performance (protect the real time analytics criteria) and the flexibility as well. So, we can easily understand the background choice of Metatron Discovery.

4. The advantages

By taking Twitter as a social media use case, we divided the hole data analytics process into two main parts for clear understanding, data ingestion and data serving, the first one cover the data processing (transforming data …) and the ingestion process of course, the second part cover the data requests and the data visualization, in each part we describe:

- Performance: which are the main criteria because we are talking about the real time analytics here.
- Functions: options of data processing included.
- UI/UX: The User interface and the User Experience.

4.1 Data ingestion

Data ingestion is the first step of analyzing data, and have crucial impact for the second part especially in the context of the real time analytics.

a) Performance

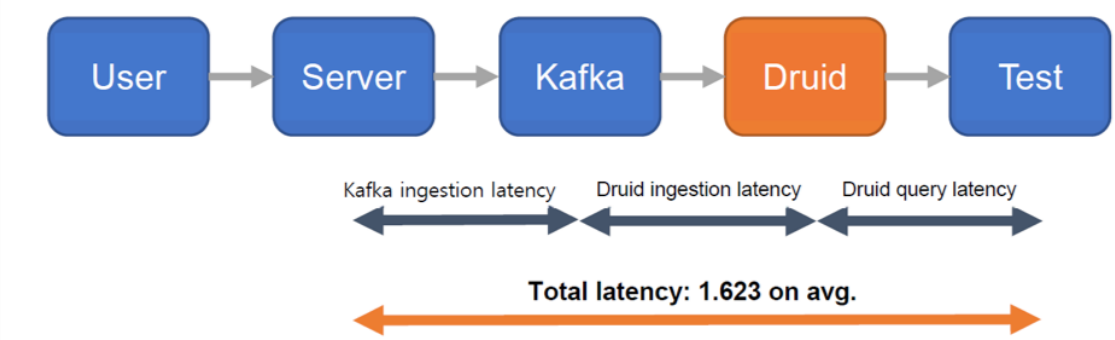


FIGURE 26: TOTAL LATENCY INVALID SOURCE SPECIFIED.

The Data ingestion speed tends to be affected by various factors such as complexity and the real time stream rate of social media data like Twitter. However, the results show that it mostly suits the stack goals see Ref for more details [35].

b) Functions

After connecting to the data stream (Kafka) successfully (Figure 27), we have the option to use the default schema of the data coming in from the stream or provide a custom one by uploading a CSV or JSON files (Fig 5), this function will help reduce good deal of data transformation process (performance and configuration) from the beginning.

Concerning the data transformation functions (Figure 29), we can divide them into two main functions [36] [37]:

1. Changing:

1. changing columns data types which include a decent set of data types:

- String
- Boolean
- Integer
- Decimal

- Date/Time
- Array
- Latitude
- Longitude
- Point
- Polygon
- Line

2. changing columns values by applying a rule if the data is missing, the options are if the rule is applied:

- Discard the column.
- Replace its value with a static value.

2. Creating:

Which is the ability to create a new column by providing two methods:

- Point: take a Latitude and Longitude column and convert to a new Point type column.
- Expression: using a bunch of math, time, string, value, logical functions on the existing columns to create an expression for calculating the new column value.

Also defining a timestamp column is required because the next setting depends on it (Query Granularity, Segment Granularity), by selecting an existing Date/Time/Timestamp type column or consider the default one which is the current time (Timestamp) of the data ingestion.

Next there is the data ingestion settings that include data granularity process that depend on the timestamp column in the first place, it includes two types (Figure 30):

1. Query granularity: the minimum granularity to query the data at (the data time precision), considering the social media use case, the most convenient is by seconds because the data generation speed and changes rate [38].
2. Segment granularity: data will be divided into segments by date size or number of rows, considering the Twitter use case it is advised to be used at hourly rate [39].

And finally, there is the Rollup function, it's an operation that aggregates some columns to reduce stored data size see https://druid.apache.org/docs/latest/ingestion/index.html#rollup) for more details, it can be useful in some use cases where the stream data rate is very fast that we can sacrifice some time precision to gain a boost of performance (Figure 30).

c) UI/UX

It's pretty simple, effective, straight forward and easy to memorize. (Figure 27, Figure 28, Figure 29, Figure 30)

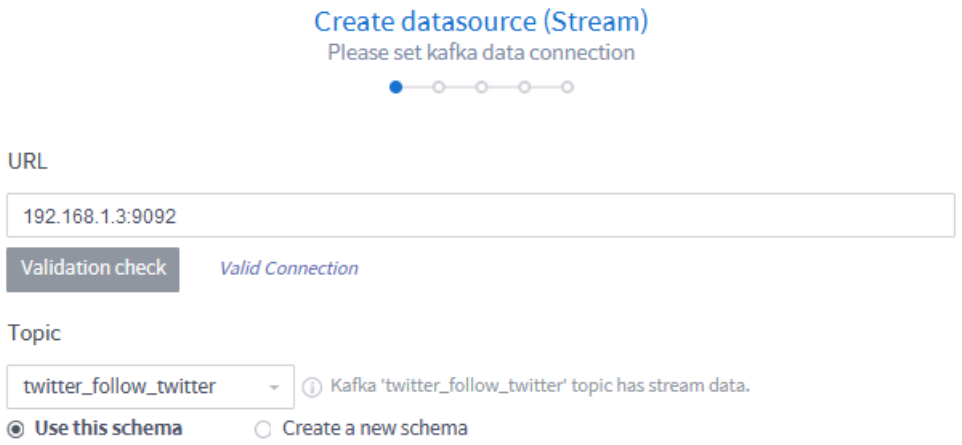


FIGURE 27: CONNECT TO DATA STREAM (STEP1)

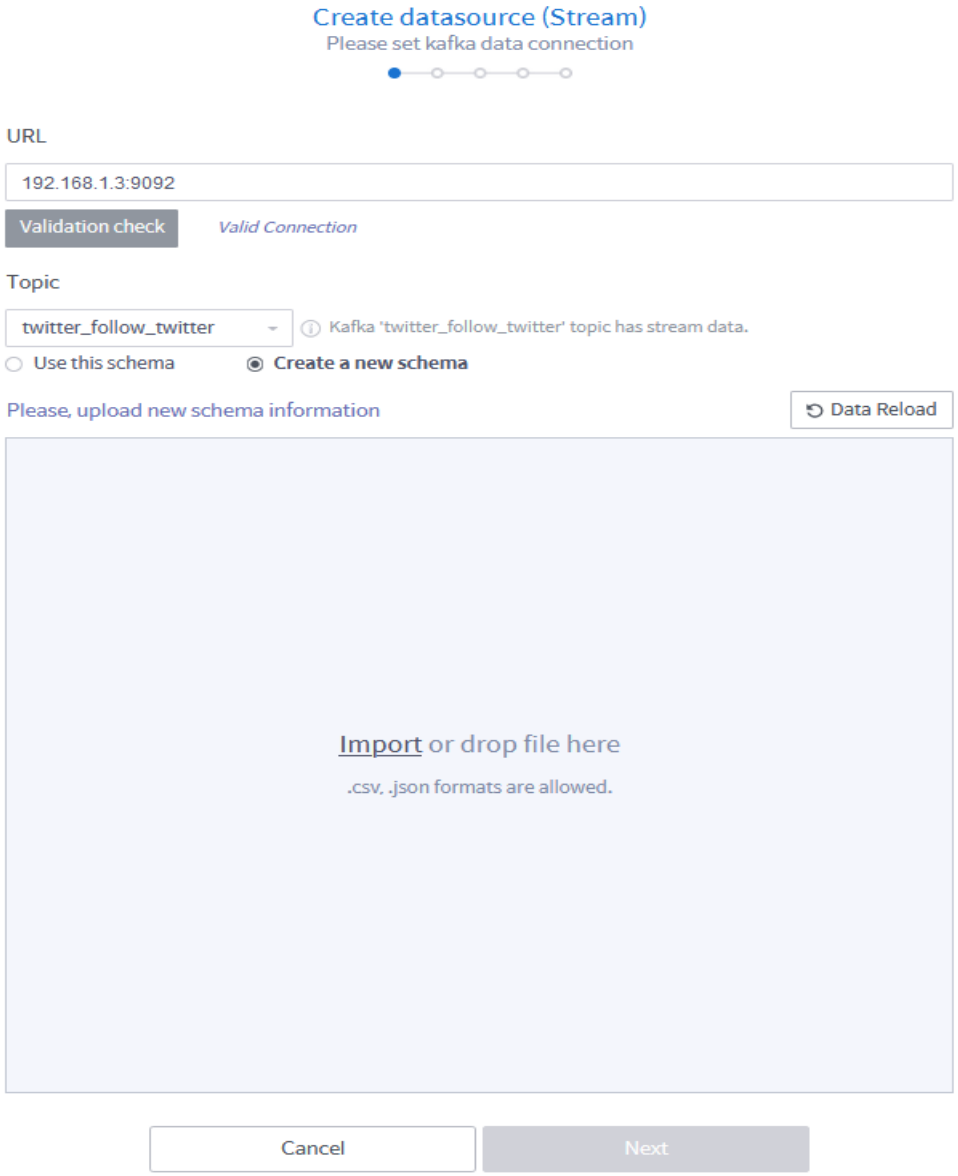


FIGURE 28: SCHEMA DEFINING (STEP1)

CHAPTER FOUR:

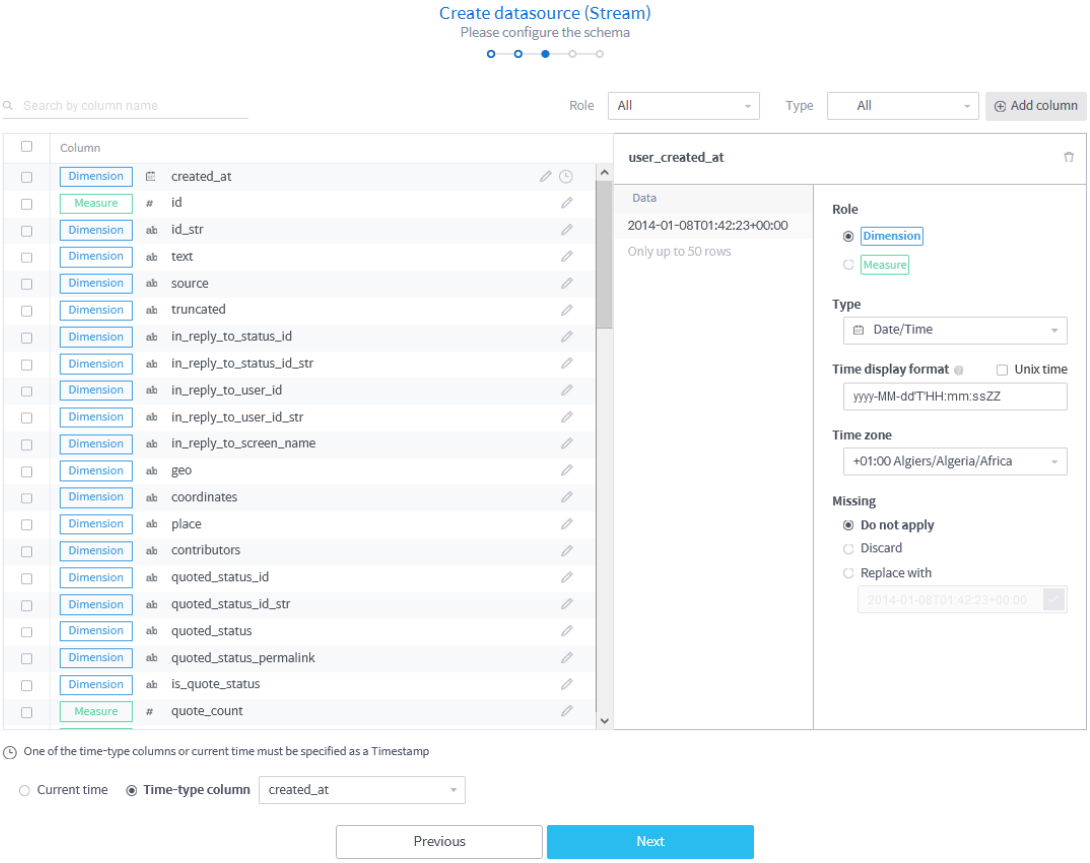


FIGURE 29: DATA TRANSFORMATION (STEP2)

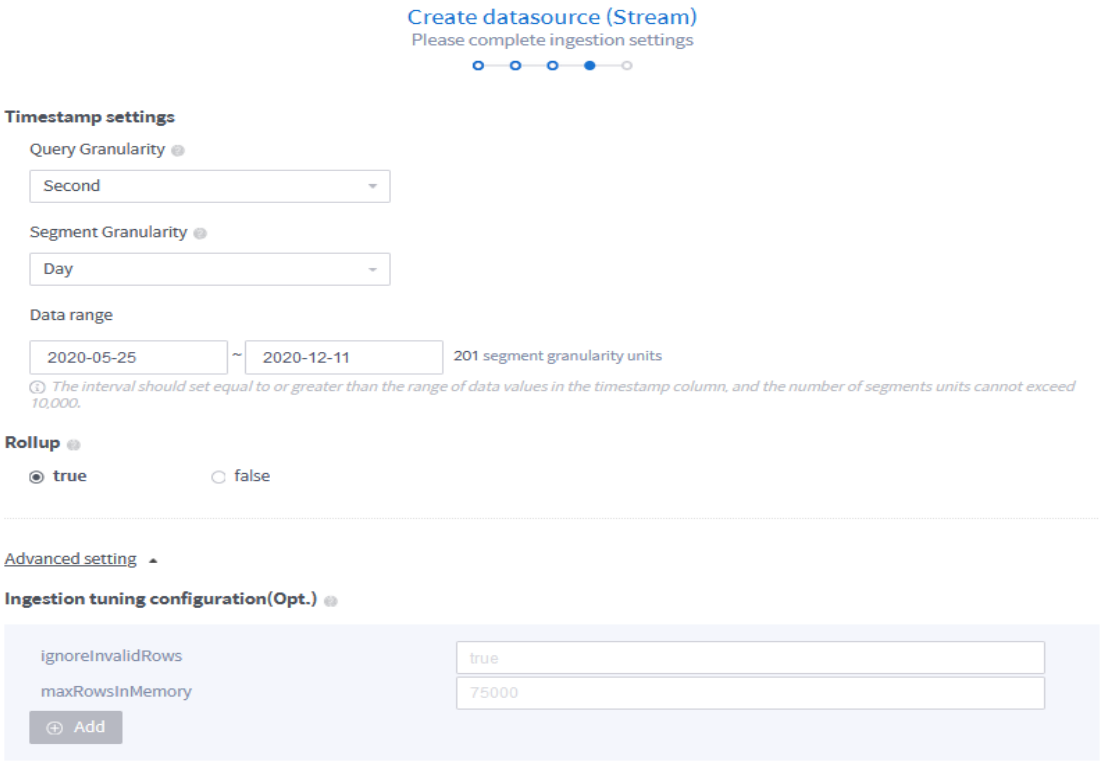


FIGURE 30: INGESTION SETTINGS (STEP 3)

4.2 Data serving

Data serving is the last step of analyzing data, it begins with querying the data and ends with its visualization on dashboards.

a) Performance

Because of Druid and Spark SQL synergy outcome the query time is remarkable. Figure 31

FIGURE 31: DRUID VS HIVE

Also, Metatron provide its own tool for monitoring the performance of its engine. Figure 32

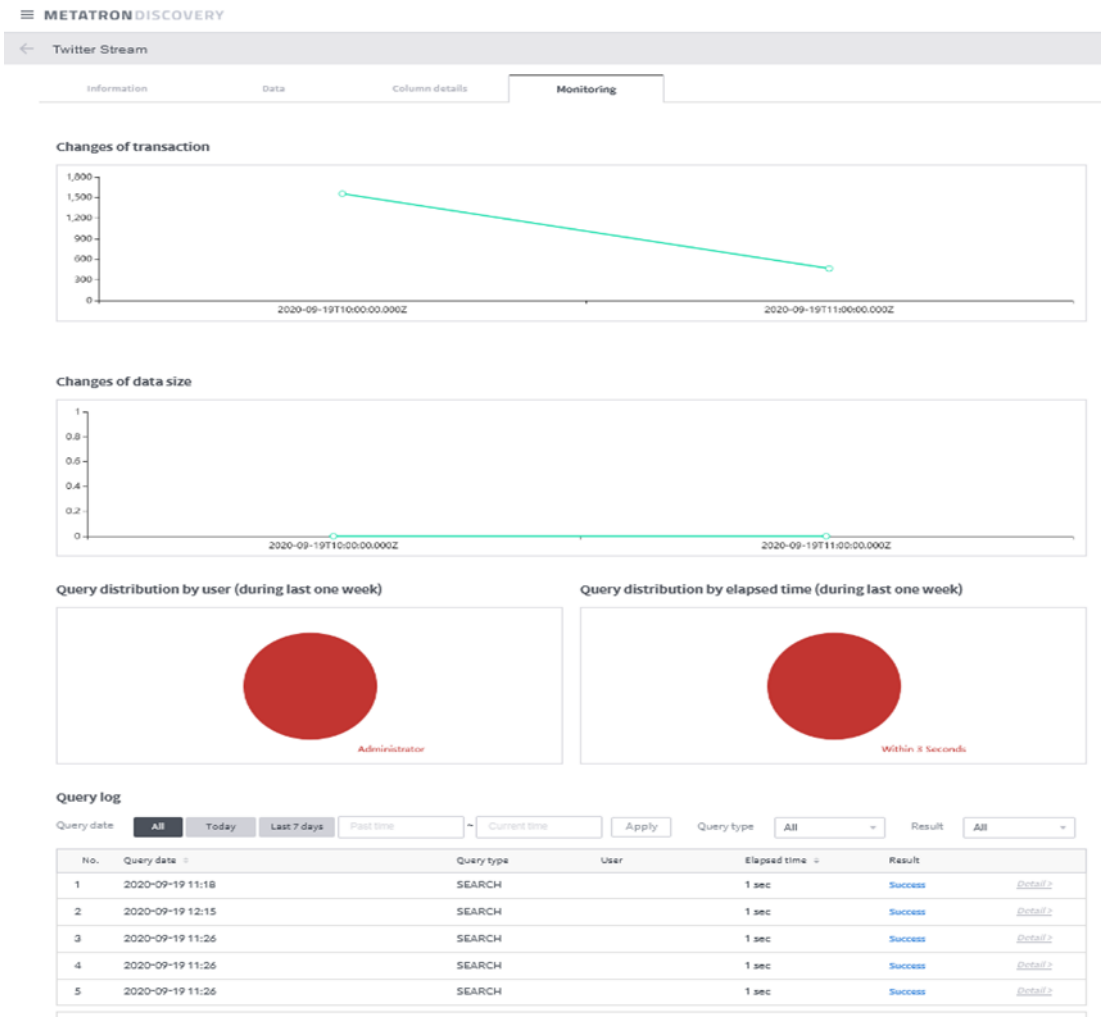


FIGURE 32: METATRON DATA SOURCE MONITORING

we didn't provide any precise calculations at the graphs point like we did at the requests performance, we concluded based on our visual judgment in our test machine as the graphs must be interactive (less than 1 second wait time) the graphs processing has a very low latency, so as a result we conclude that the real time criteria is still conserved.

b) Functions

The stack provides decent Filter and Aggregation functions, the filtering can be used on different levels, the Dashboard level, the Chart level and finally at the Data Source level (Dimension, Measure), it's very satisfying and effective specially in the social media context, you can filter your data by selecting specific values from the already read values if the column was dimension type or filter by selecting a range if the column was measure type.

The available aggregations functions are (Figure 39): Sum, Average, Count, Median, Min, Max, Percentile (1/4, ¾) and they cover our needs in our Social Media (Twitter) analytics use case.

along with the previous functions there is other functionalities like:

- Setting the rate of refreshing the data, it can be set from 1 second to 60 seconds (useful because some charts has not high variation rate like words Map (Figure 47, Figure 48) or the Maps (Figure 51, Figure 52).
- Create Aliases for the column's names and values.

- Align in Ascending, Descending or Data order.
- Create custom columns by creating a formula to calculate their values using wide collection of functions (Aggregation, String, Time, Type conversion ...). (Figure 39)

c) UI/UX

Still pretty simple, effective, straight forward and easy to memorize, this is some of the based points:

- Creamy Chart collection (Word Cloud, Map View, Heat map, Network Diagram, …) with the feature of suggesting a chart if the Dimension and Measures were selected first and opposite as well which save a lot of time.
- Nice Dashboard customization (Text, Layout ….).
- Decent Chart customization (Colors, Number Format, Data Labels …).
- Very good Workflow preservation because of a good exploitation of the popup windows.

Workbook is a place that contain a collection of Dashboards and a description for them and as a nice future you can share it in the future. (Figure 34, Figure 35)

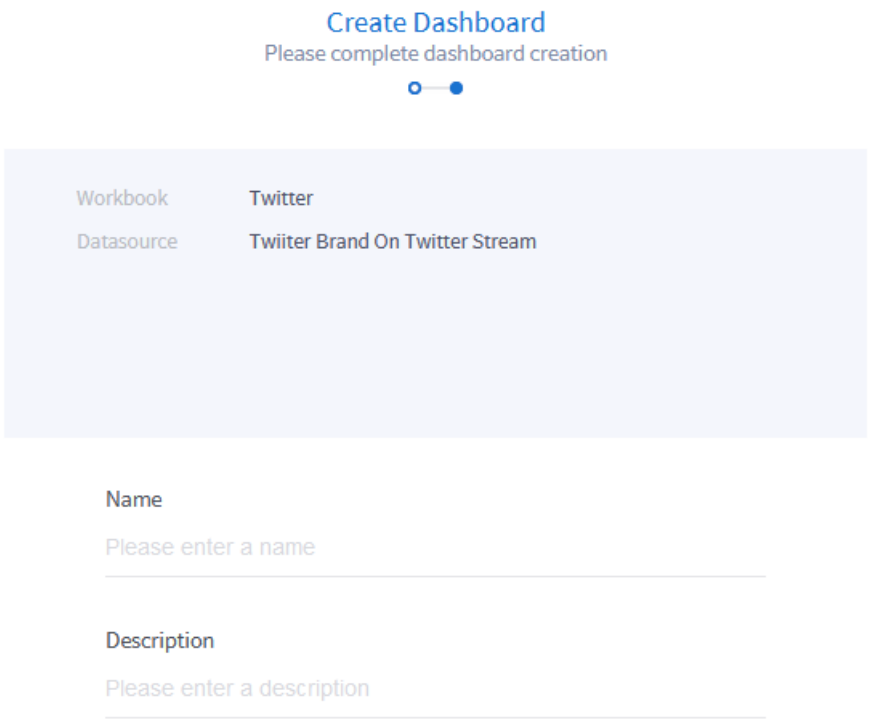


FIGURE 33: DASHBOARD CREATION

CHAPTER FOUR:

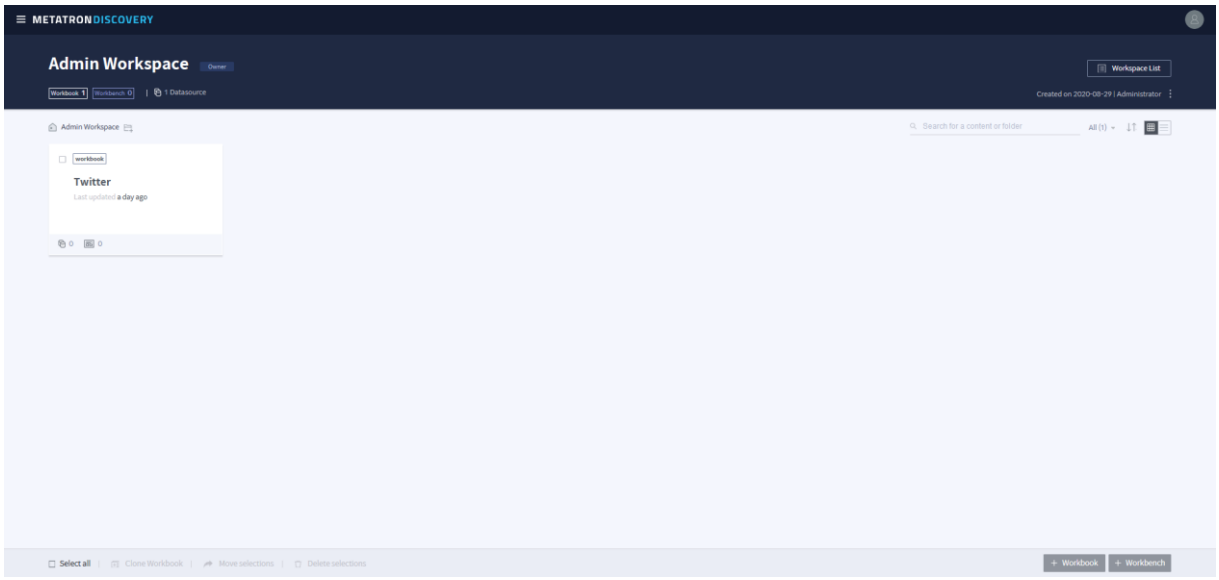


FIGURE 34: WORKSPACE

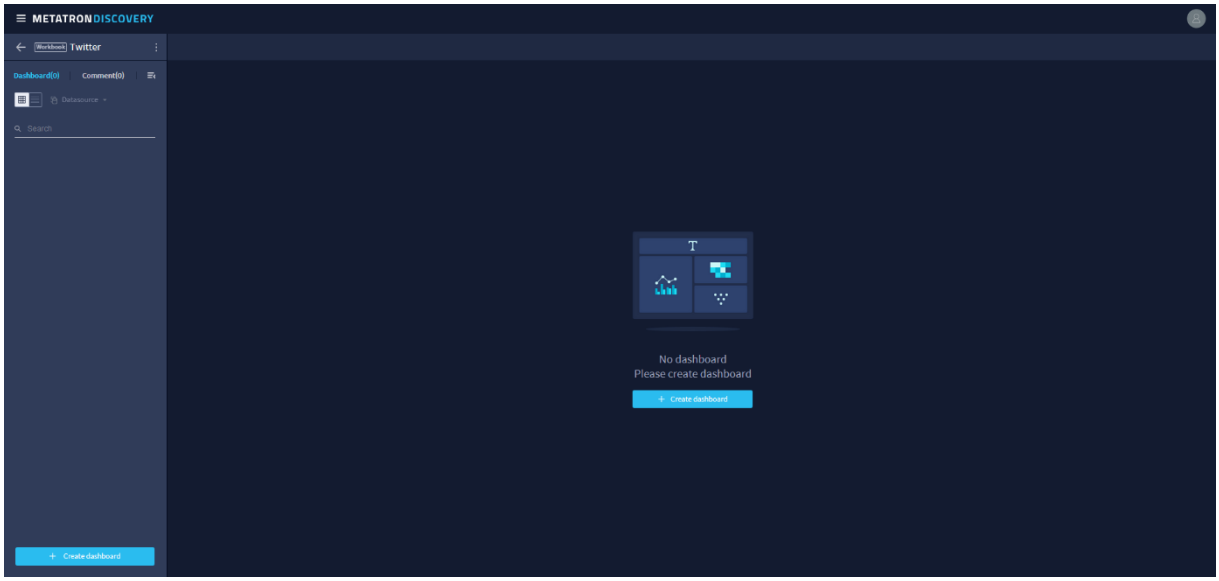


FIGURE 35: THE WORKBOOK

CHAPTER FOUR:

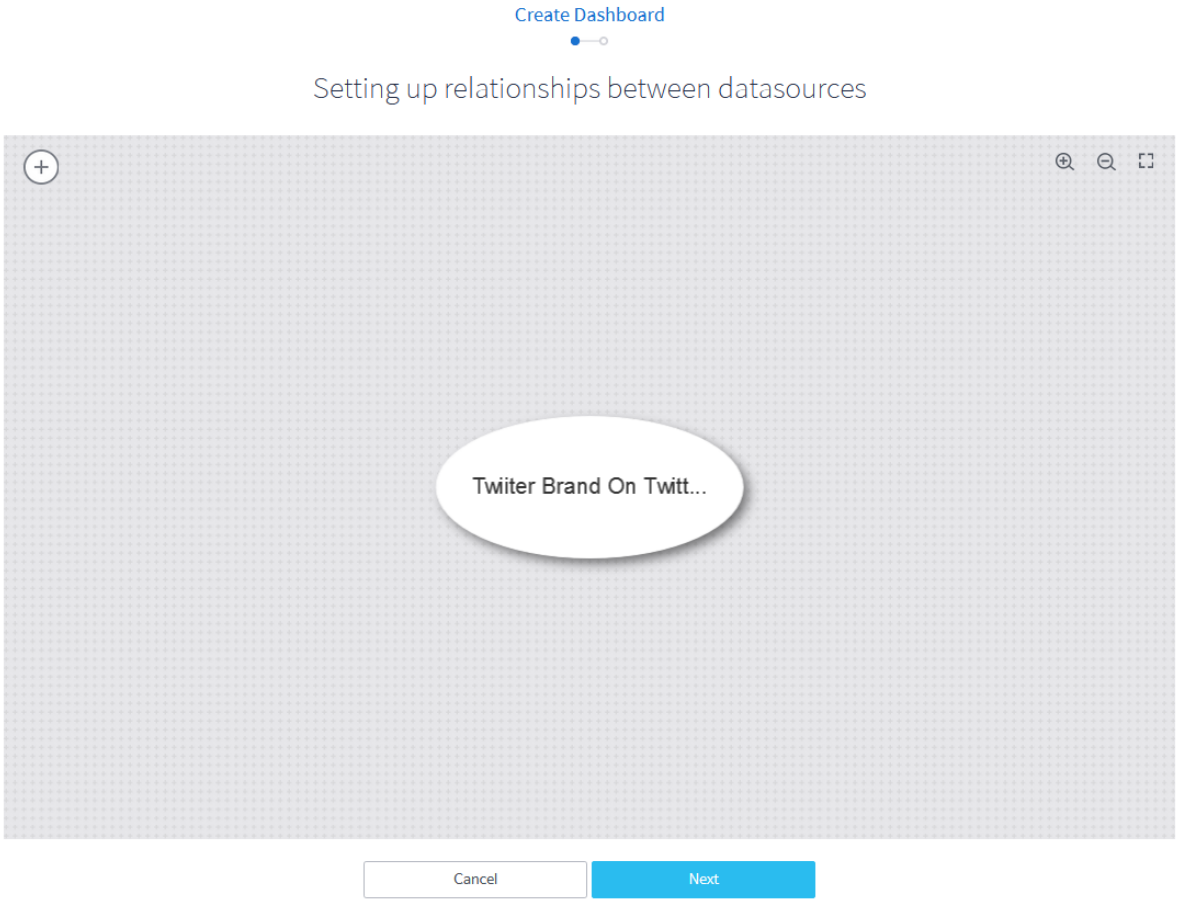


FIGURE 36: SETTING DATA SOURCE FOR THE DASHBOARD

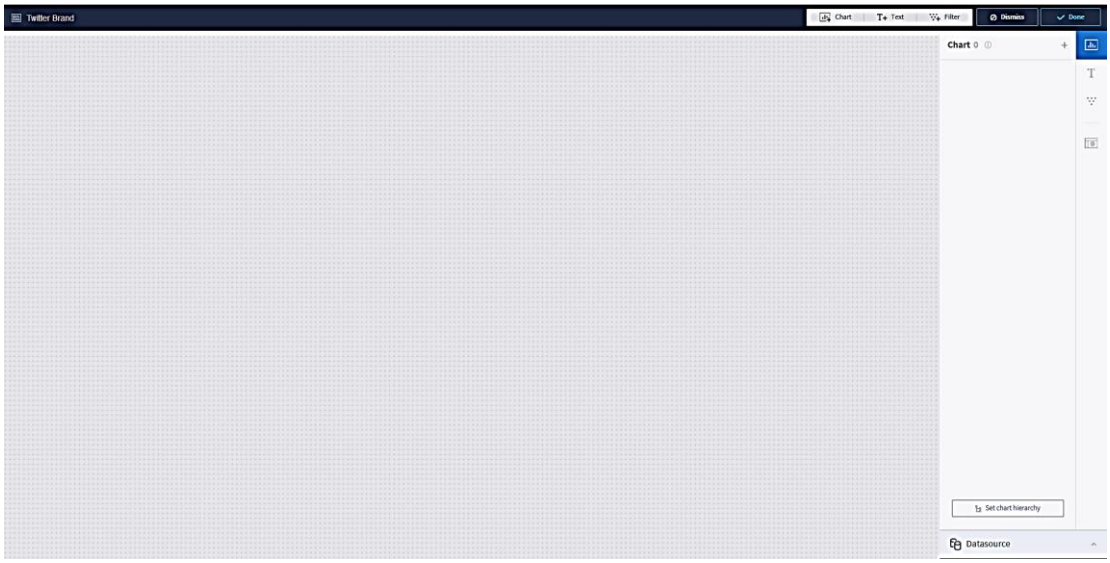


FIGURE 37: EDITING DASHBOARD

CHAPTER FOUR:

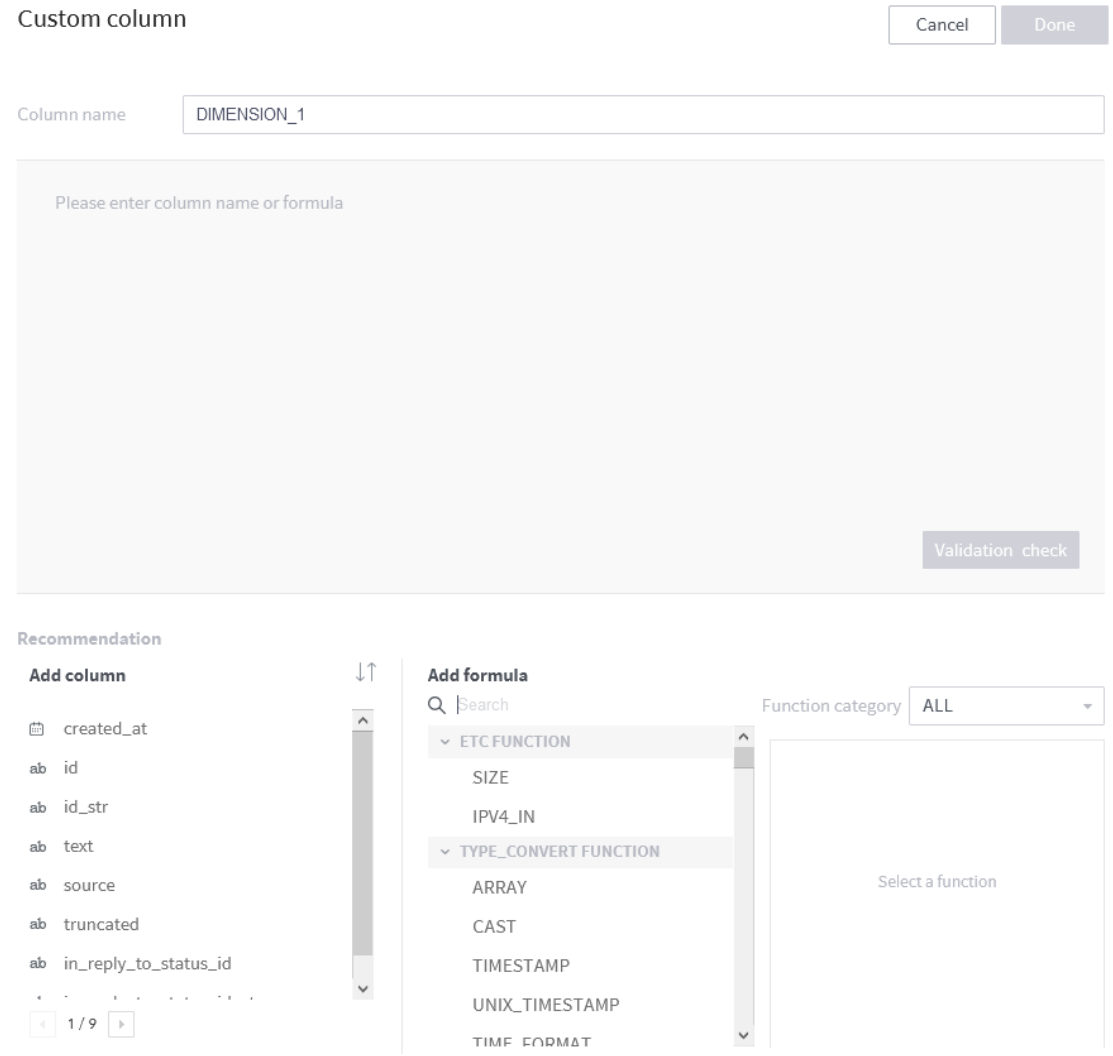


FIGURE 38: CREATE CUSTOM COLUMN

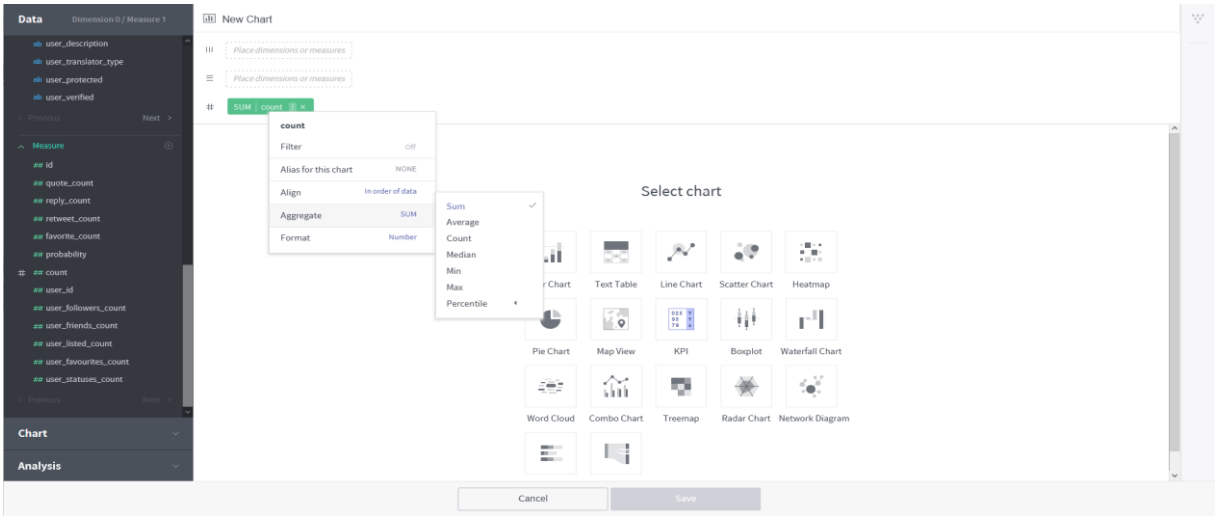


FIGURE 39: AGGREGATION FUNCTIONS

CHAPTER FOUR:

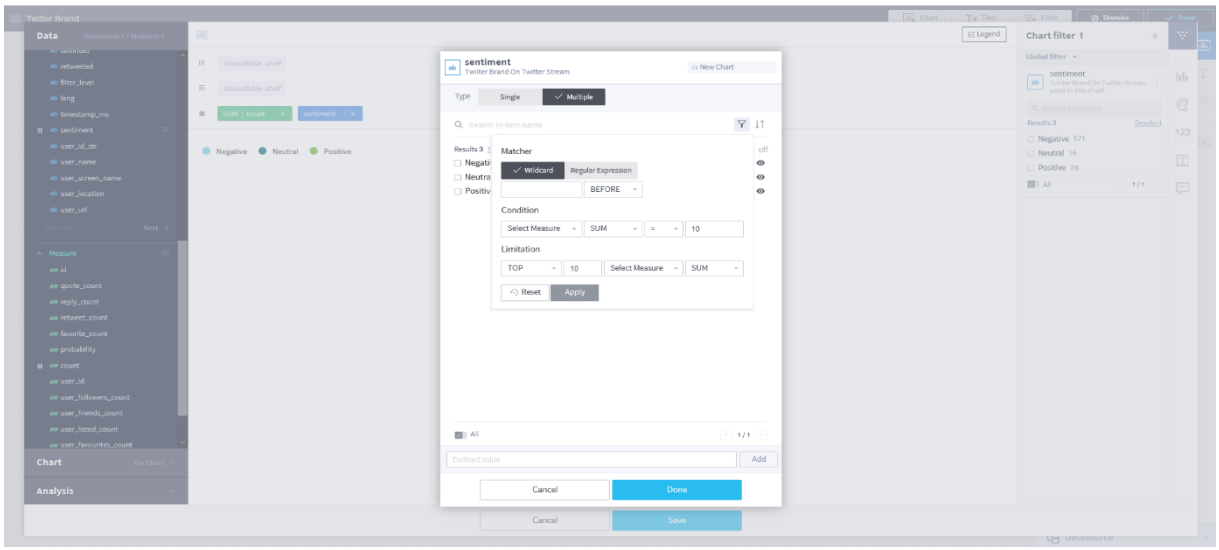


FIGURE 40: DATA FILTERING (CHART LEVEL)

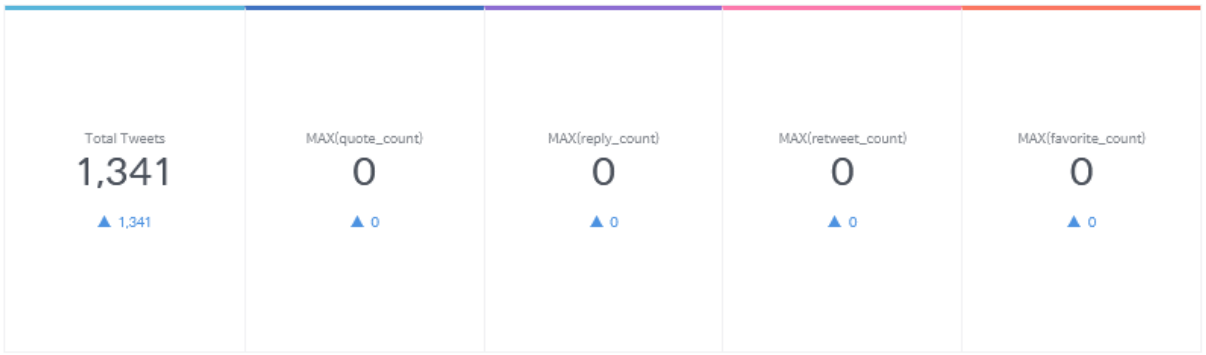


FIGURE 41: TWEETS TRENDS

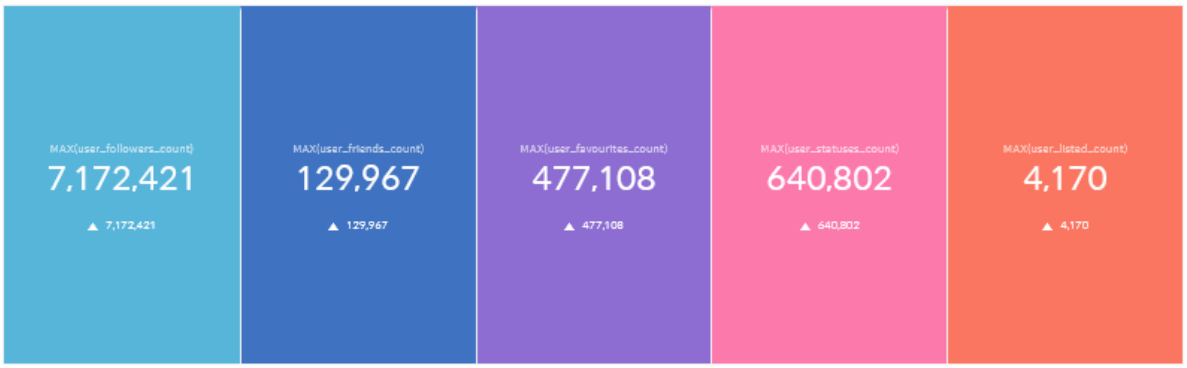


FIGURE 42: FOLLOWERS TRENDS

CHAPTER FOUR:

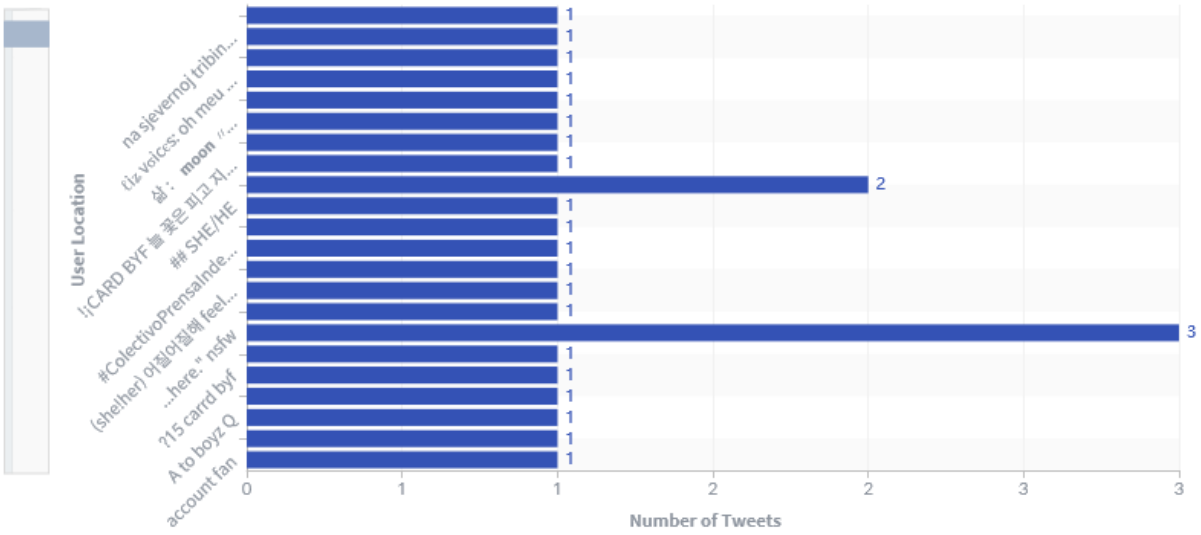


FIGURE 43: NUMBER OF TWEETS PER USER LOCATION

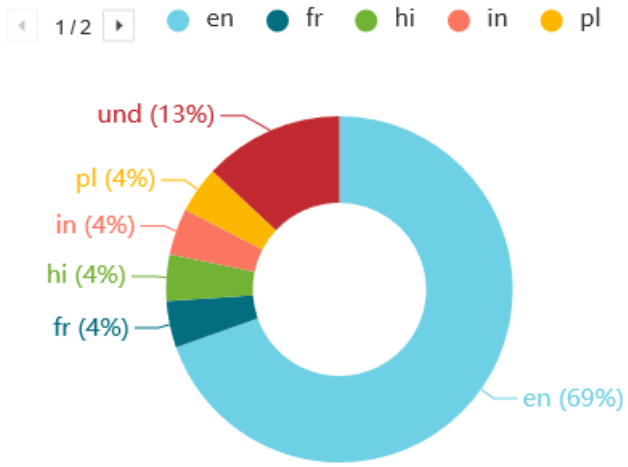


FIGURE 44: TWEETS LANGUAGE %

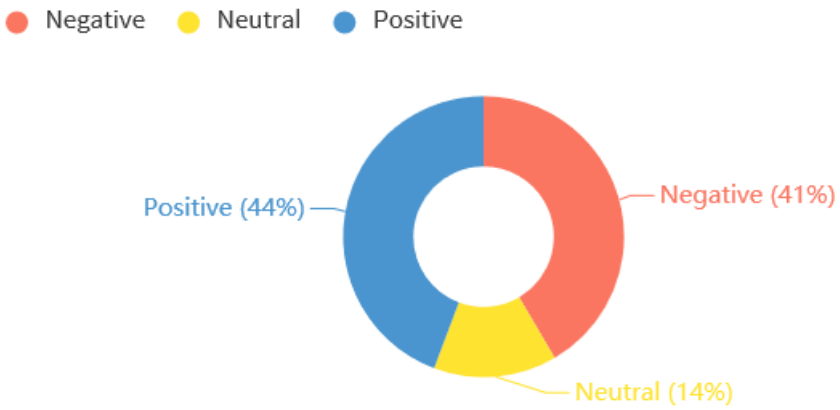


FIGURE 45: TWEETS SENTIMENT %

CHAPTER FOUR:

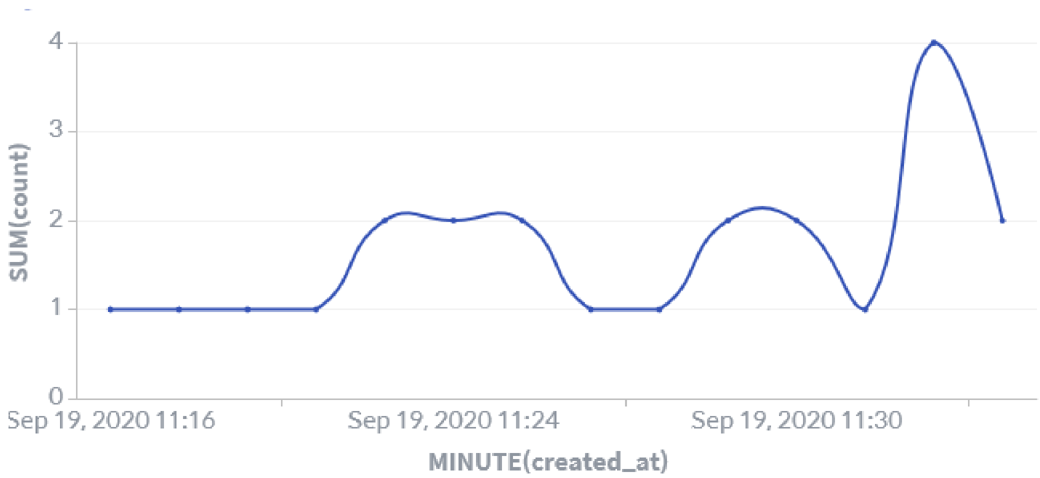


FIGURE 46: NUMBER OF TWEETS PER MINUTE

FIGURE 47: TEXT WORDS MAP (TEXT TRENDS)

FIGURE 48: HASHTAGS WORDS MAP (HASHTAGS TRENDS)

CHAPTER FOUR:

FIGURE 49: SENTIMENT % PER LANGUAGE % (1)

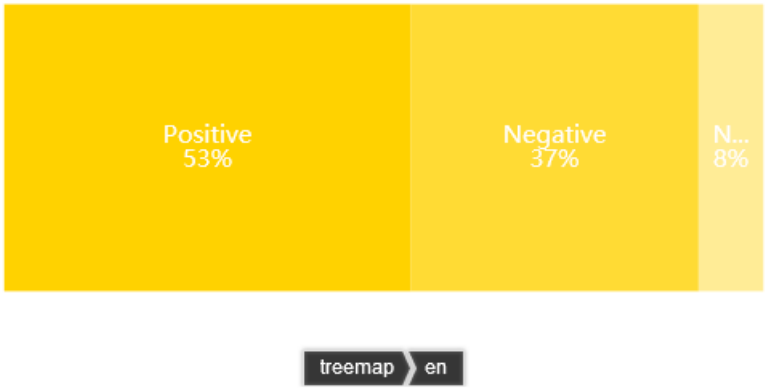


FIGURE 50: SENTIMENT % PER LANGUAGE % (2)

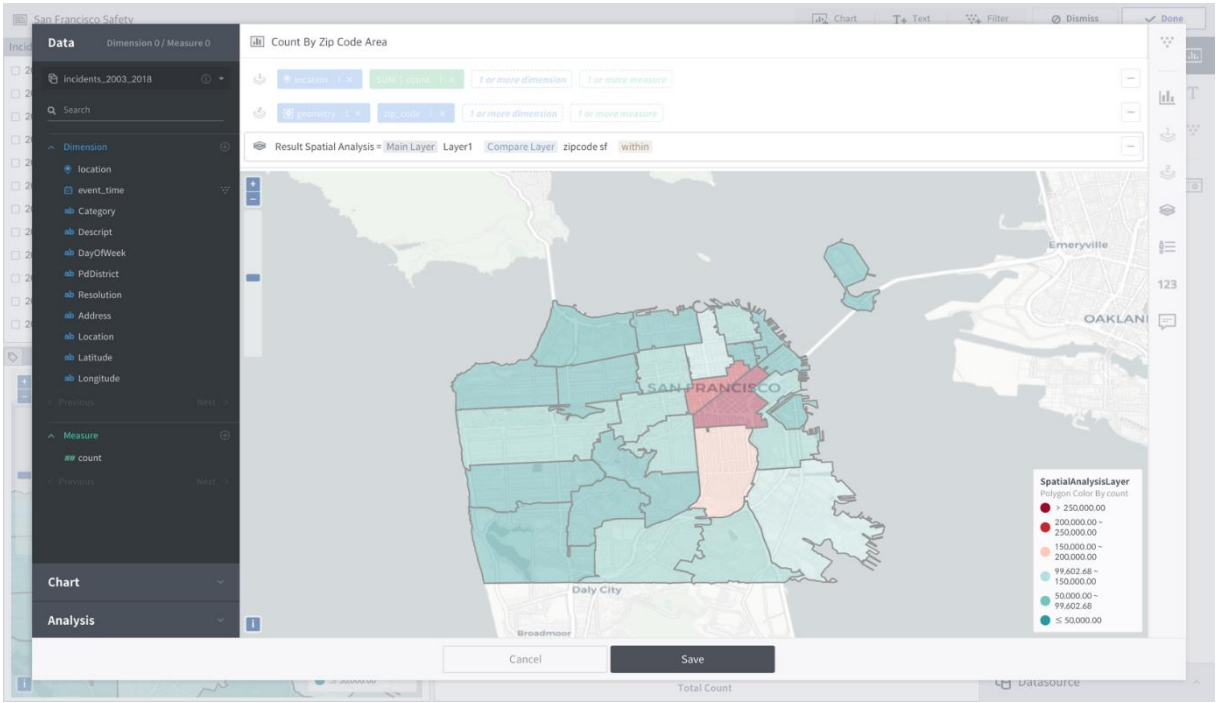


FIGURE 51: MAP ANALYSIS 1

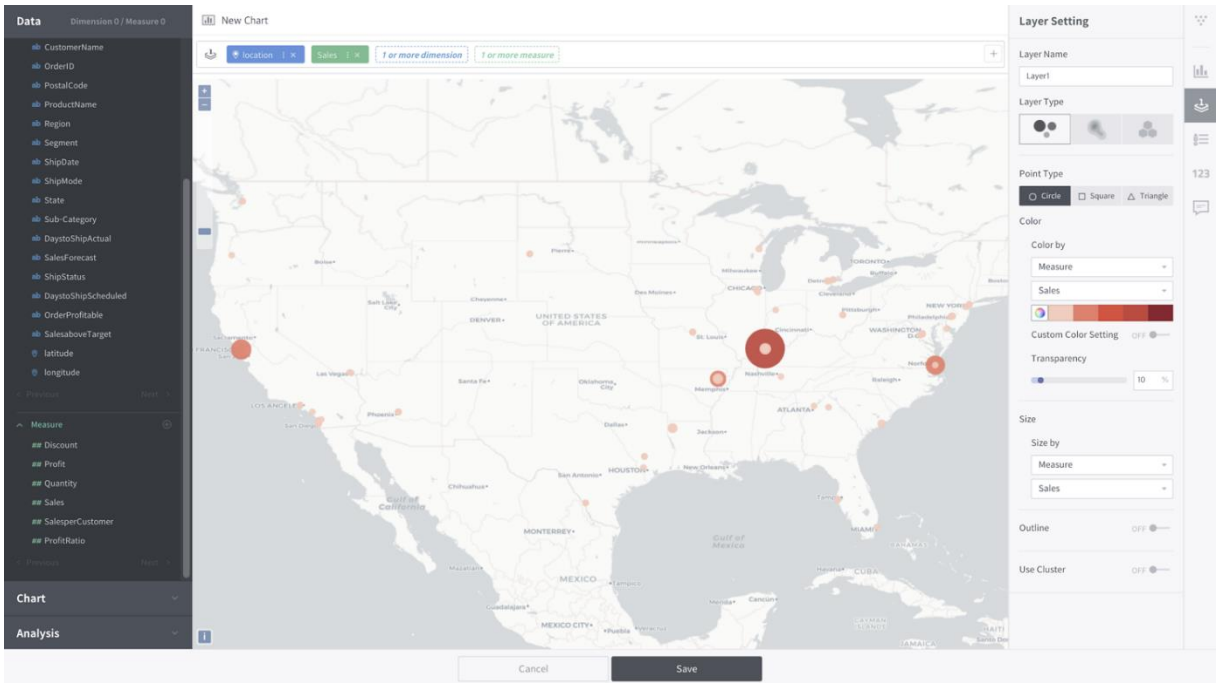


FIGURE 52: MAP ANALYSIS 2

4.3 Scalability

The scalability of this stack is something remarkable because it depends on the microservices paradigm, a set of components work together to achieve a goal, the most advantage of this paradigm is the high scalability and also the reliability, besides that, each component (Kafka, Apache Druid, Metatron Discovery) it was designed to be scalable on its own, so, as a result, we had a very scalable real-time analytic tool.

4.4 Reliability

Because of the paradigm of microservices and the dependency on a set of components (Kafka, Apache Druid, and Metatron Discovery) that are reliable on their design, as a result, we had a very reliable, fault tolerance analytic tool.

4.5 Cover social media analytics key areas

As we mention in the 3rd chapter (social media analytics) the analytics area keys are covered using the Twitter API features and the data exploration capabilities of the stack.

The audience, sentiment, competitive analytics can be done with the help of the filtering feature of the Twitter API stream, filter it by topic (Twitter account in this example) or filtering in the stack level so that we collect the Tweets that are directed to that specific topic (account) in real-time, our account or a competitive account, a detailed user object is present with every Tweet we receive, we may need extra step concerning the sentiment analysis because we need to process every Tweet to reveal the sentiment behind it so an extra process after collecting the data from the API and before the ingestion in our opining is needed to achieve this, after confirming the presence of the information about the audience (user data, sentiment…) , the only step left is to explore the data collected with our tool using the aggregation functions and the effective and very expressive charts to get clear insights (Figure 53, Figure 54).

Some charts examples:

- Figure 41, Figure 42 are Key Performance Indicators, Figure 41 show the total Tweets from the beginning of the stream and how much did it develop from the previous day using the count aggregation function, also using the Max aggregation function it indicates the top retweeted, replied, quoted and favorited Tweets from the beginning of the stream, in the other hand Figure 42 shows indicators concerning the Followers, By using the Max Function we indicated the top Followers in number of Followers, Friends, Favorited, Status and Listed who Twitted the Tweet, low data sync interval is favorable in this kind of charts.
- We did some Sentiment along with Language Analysis using the Pie and Treemap charts, Pie Charts like Figure 44 shows the percentage of the Tweets languages and Figure 45 shows the percentage of the Tweets sentiments (don't look pretty good for you Twitter!) while Figure 49 and Figure 50 indicates the percentage of sentiments in every Language like we observe 69% of the Tweets are in English language and 53% of them are Positive, 37% Negative, 14% Native, you can put the data sync interval a bit high because the insights here are Percentage based not count based.
- Figure 46 shows a simple line chart that indicates the number of Tweets over time, the data sync for this chart need to be as the time unite, if the sync was by seconds and the chart time unite is in minutes performance is wasted, the chart refresh every second with no effect, in the opposite side we lose precision if the sync time is higher than chart time unite, the unite for the time used here is minute and that depends on the first data source configuration (Figure 27 to Figure 30) exactly the Query granularity but not the inverse, if we make the data granularity by seconds, we can than choose the granularity in our chart from seconds to years but if the query granularity was like by hour we can't choose less than an hour as data granularity in the chart level because it's impossible.
- Concerning the Word Map charts they are good for analyzing anything written but not structured (Unstructured data), we used them for analyzing the texts and Hashtags of the Tweets (Figure 47, Figure 48) and they show us the most used words (Tweets text or Hashtags), the more the word is bigger the more it is used widely by the followers (audience), it is very useful for knowing the audience trends and ideas without the pain of direct contact or feedback, the data sync interval need to be 60 seconds (Max) in this type of charts because it take some time (at least minutes) for the Twitter users to spread a Hashtag or a Tweet so make it like 1 second based data sync is useless.
- The lack of the geo and place information in Twitter data prevent us from creating Map views to get some important sights like how much tweets are coming and from where, which make us understand the geo concentration points of our audience, Figure 51, Figure 52 are just sample examples to present the Map chart, and concerning the Figure 43 is a horizontal bar chart that shows instead of the Tweets place the user's place because the user place information is more present than the Tweets place information, it's not very precise information and has a lot of noise but at least we can have a clue of the audience location.

CHAPTER FOUR:

FIGURE 53: TWITTER DASHBOARD

FIGURE 54: TWITTER DASHBOARD 2

The performance analytics can be achieved easily using the existing aggregation and filtering functions from the stack and the performance metrics provided by the Twitter API in their API responses like the number of retweets, number of followers ….

The influencer analytics depend on the search end point of the Twitter API to get what data exactly you need like Tweets or Users profiles and explore it in the stack easily, the stack actually can easily replace this end point by giving it enough time to accumulate data or just ingest a significant amount of data and keep it up dated by the real time stream so that any search will be done will be reliable almost like searching in the native data source.

The API also provide stream and search access to the messages so customer and community management analytics it's not impossible.

CHAPTER FOUR:

5. Challenges

First challenge was the test machine, we hadn't the luxury to test the stack in real servers thought the results were significant using our very humble testing machine which promise with great results if more powerful machines or clusters were used instead.

There is also the Twitter API challenge, which is the privilege to use the premium features like the real time stream of all the Tweets generated which will push the ingestion process of the stack to its limits then we can observe the impact on the real time analytics.

The Twitter Objects (Data Stream) doesn't contain nor the geo or the place data in most cases, which has hardened our job in creating the Map View Chart (Figure 51: Map Analysis 1, Figure 52) to complete the social media analysis.

The Metatron Discovery (Data Visualization layer) still in beta version while we are writing these lines, so talking about bugs won't be beneficial as much as suggesting some features like the ability to connect, disconnect, enable, disable and modify the data sources (schema, granularity…) including the streams.

6. Perspectives

The first thing we suggest for further improvements is a user interface (UI) to connect, transform, manage and control the data sources like the Twitter API in this example.

Apache Druid has very promising effectiveness in feeding IA algorithms as much as feeding an end user dashboard.

Apache Druid can improve mostly the performance of the Twitter TSAR architecture see ("Robust, Scalable, Real-Time Event Time Series Aggregation at Twitter"), the general idea is to replace the data serving components (RDMS, Manhattan, Nighthawk …) which are considered in the first place for responding queries (ad hoc queries, feeding dashboards…) and Apache Druid excels at handling this kind of queries much better because the low latency is required along with some other criteria like the scalability and the reliability.

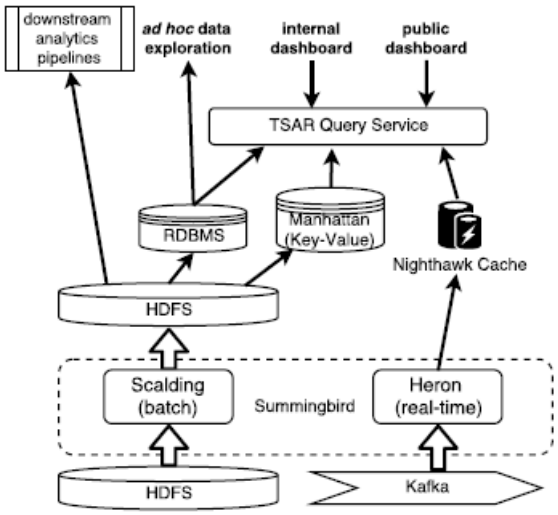


FIGURE 55: OVERALL ARCHITECTURE OF TSAR

7. Conclusion

The Data analysis is a very demanded domain specially in this period where the data is massively generated from various sources (mobiles, social medias, services….), the need of analyzing it and using it has emerged because this data is like a gold mine for who wants to create a solid work ground to create, improve and maintain his decisions or plans in any domain(Business, Political, Science…), because of the world speed today and the crucial impact that can happen if the decisions were late, the analyzing process has to evolve to the point of giving results in real time, in other side the social media become a very important source of data to get those results, the KDM stack (Apache Kafka, Apache Druid, Metatron Discovery) emerge as an excellent tool for analyzing the social media data (Twitter use case) because of the real time ingestion, aggregation and the low latency data serving it can provide, along with other criteria like the good reliability and the scalability, the functions (aggregation, filters…) provided to analyze the data and the decent UI/UX that help simplify, fast create the dashboards with clear and expressive charts.

ABSTRACT

GENERAL CONCLUSION

Big Data analyzing impact has been proved countless times and in many areas (Business, Science…) but world of today has been more faster than ever, an Immense volume of data is being generated every hour from many sources where the social media sources are the majority, using this wealth of social media data is well needed and necessary for many domains so an automated scalable reliable real time analysis is required, the KDM stack suite well this use case and others and fulfill all it needs and more.

ABSTRACT

ABSTRACT

The development of social media creates a multitude of new Real-Time Analytics (RTA) application possibilities. However, already the topic of big data forced the use of analytical solutions and among them, there are some decent near real-time solutions but using them in the social media analytics domain will reveal some other flaws especially in the context of the real-time analytics where the social media analytics native tools can't be challenged in general and in real-time analytics specifically.
This work aims to address this gap by exposing a decent tool (Technology stack) excels at handling real-time social media analytics by providing abilities like fast processing of data, real-time data aggregation, ingestion, interactive, and effective data exploration, and visualization.
We describe our experiments for the architecture in Twitter use case and evaluate the functional and non-functional tests such as real-time update performance and time taken for data flow among components. All components were able to handle their functionalities properly.

RESUME

Le développement des médias sociaux crée une multitude de nouvelles possibilités d'application d'analyse en temps réel (RTA). Cependant, déjà le sujet du big data a forcé l'utilisation de solutions analytiques et parmi elles, il existe des solutions décentes en temps quasi réel, mais leur utilisation dans le domaine de l'analyse des médias sociaux révélera certaines d'autre failles, en particulier dans le contexte de l'analyse en temps réel. Où les outils natifs d'analyse des médias sociaux ne peuvent pas être remis en question en général et en analyse en temps réel en particulier.
Ce travail vise à combler cette lacune en exposant un outil décent (pile technologique) excelle dans la gestion de l'analyse des médias sociaux en temps réel en fournissant des capacités telles que le traitement rapide des données, l'agrégation de données en temps réel, l'ingestion, l'exploration de données interactive et efficace, et visualisation.
Nous décrivons nos expériences pour l'architecture dans le cas d'utilisation de Twitter et évaluons les tests fonctionnels et non fonctionnels tels que les performances de mise à jour en temps réel et le temps nécessaire au flux de données entre les composants. Tous les composants ont pu gérer correctement leurs fonctionnalités.

ملخص

خلق تطور وسائل التواصل الاجتماعي العديد من إمكانيات تطبيق Real Time Analytics (RTA) الجديدة، ومع ذلك ، فقد أجبر موضوع البيانات الضخمة (Big data) بالفعل على استخدام الحلول التحليلية ومن بينها يوجد بعض الحلول اللائقة القريبة من خاصية الوقت الحقيقي للتحليل ولكن استخدامها في مجال تحليلات الوسائط الاجتماعية يكشف عن بعض العيوب الأخرى خاصة في سياق تحليلات الوقت الحقيقي حيث لا يمكن تحدي أدوات التحليل الأصلية لمنصات التواصل الاجتماعي بشكل عام وفي تحليلات الوقت الحقيقي على وجه التحديد.
يهدف هذا العمل إلى معالجة هذه الفجوة من خلال الكشف عن أداة لائقة تتفوق في التعامل مع تحليلات الوسائط الاجتماعية في الوقت الفعلي من خلال توفير قدرات مثل المعالجة السريعة للبيانات ، وتجميع البيانات في الوقت الفعلي ، والابتلاع ، وفعالية استكشاف البيانات بتفاعلية وعرضها.
نصف تجاربنا لهذه الأداة في حالة استخدامها مع منصة Twitter ونقيم الاختبارات الوظيفية وغير الوظيفية مثل أداء التحديث في الوقت الفعلي والوقت المستغرق لتدفق البيانات بين المكونات. كانت جميع المكونات قادرة على التعامل مع وظائفها الخاصة بشكل صحيح

ABSTRACT

REFERENCES

References

[1] M. J. C. a. C. L. Vinayak R. Borkar, «Big Data Platforms: What's Next?,» *RIGHTLINK,* vol. 19, n° %11, p. 6, 2 0 1 2.

[2] M. C. T. P. J.-H. P. Ming Zhou, «Clarifying Big Data: The Concept and Its Applications,» *RIGHTSLINK,* p. 4.

[3] J. B. Pedro Neves, «Big Data Issues,» p. 2, july 2015.

[4] D. S. J. D. U. Alfredo Cuzzocrea, «Big Data: A Research Agenda,» *RIGHTSLINK,* p. 6, 2013.

[5] J. H. a. S. Kandel, «Interactive Analysis of Big Data,» *RIGHTSLINK,* vol. 19, n° %11, p. 5, 2012.

[6] M. G. a. M. G. Andrea De Mauro, «What is Big Data? A Consensual Definition and a Review of Key Research Topics,» chez *International Conference on Integrated Information (IC-ININFO 2014)*, 2015.

[7] B. P. (. Ghosh, «Fundamentals of Real-Time Analytics,» on September 7, 2017. [En ligne]. Available: https://www.dataversity.net/fundamentals-real-time-analytics/.

[8] F. A. R. Zoran Milosevic, «Real-Time Analytics,» *ResearchGate,* p. 25, December 2016.

[9] C. McCue, «in Data Mining and Predictive Analysis,» *sciencedirect,* (Second Edition), 2015.

[10] « "Releases - memcached/memcached",» [En ligne]. Available: https://github.com/memcached/memcached/releases. [Accès le 3 November 2018].

[11] S. Scott, «Real-Time Analytics: Streaming Big Data for Business Intelligence,» Mar 13th, 2017. [En ligne]. Available: https://logz.io/blog/real-time-analytics/.

[12] «sumologic,» 2014. [En ligne]. Available: https://www.sumologic.com/glossary/real-time-big-data-analytics/.

[13] Francis Xavier MCRORY, Warner Robins ; Rogelio SAUCEDO, Bonaire, Real-time data-processing systems : a methodology for design and cost/performance analysis, Patent Application Publication , Mar. 17, 2011 .

[14] Paul Zikopoulos, Chris Eaton , Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media, October 2011.

ABSTRACT

[1 «Descriptive Analytics Insight into the past -“What happened”,» [En ligne]. Available:
5] http://www.analythica.com/Descriptive%20Analytics.html.

[1 S. Scott, «DZone,» Mar. 29, 17. [En ligne]. Available: https://dzone.com/articles/real-time-
6] analytics-streaming-big-data-for-busines.

[1 «Streaming Analytics 101: The What, Why, and How».
7]

[1 S. Sedkaoui, Data Analytics and Big Data, John Wiley & Sons., 2018.
8]

[1 F. Smith, The Ideal Stack for Real-Time Analytics, August 5, 2019.
9]

[2 MemSQL Blog The Ideal Stack for Real-Time Analytics, MemSQL Blog.
0]

[2 M. Hooten, Building the Ideal Stack for Real-Time Analytics, Mar. 05, 17 .
1]

[2 A. Khan, BCCI Grants IPL 2020 Title Sponsorship To Fantasy Cricket App Dream 11, Controversy
2] Over Decision To Forgo Tata’s Offer, August 18, 2020.

[2 «Presearch,» [En ligne]. Available: https://www.presearch.org/results?q=PINTEREST.
3]

[2 AndronikiSapountzi Kostas E.Psannis, «Social networking data analysis tools & challenges☆,»
4] September 2018.

[2 C. R. C. Vinaya Kumar Mylavarapu, «Social Media Analytics: Enabling Intelligent, Real-Time
5] Decision Making,» Cognizant, August 2013.

[2 «social media analytics the complete guide,» socialbakers.
6]

[2 « understanding social media analytics.,» businessnewsdaily.
7]

[2 «wikipedia,» [En ligne]. Available: https://en.wikipedia.org/wiki/Social_media_analytics.
8]

[2 «Social media big data analytics: A survey».
9]

[3 W. SMITH, «Understanding Apache Kafka and How it Can Make Your Business More Efficient:
0] Part 1,» 19 MAY 2020. [En ligne]. Available: https://aeccloud.com/blog/understanding-apache-kafka-and-how-it-can-make-your-business-more-efficient-part-1#:~:text=More%20than%20one-third%20of,Netflix%2C%20to%20name%20a%20few..

ABSTRACT

[3 Bhole Rahul Hiraman ; Chapte Viresh M. ; Karve Abhijeet C, «A Study of Apache Kafka in Big
1] Data Stream Processing,» chez *2018 International Conference on Information , Communication, Engineering and Technology (ICICET)*, Pune, India, 29-31 Aug. 2018.

[3 Paul Le Noac'H, Alexandru Costan, Luc Bougé, «A Performance Evaluation of Apache Kafka in
2] Support of Big Data Streaming Applications,» 24 Nov 2017.

[3 «Introduction to Apache Druid,» Druid, [En ligne]. Available:
3] https://druid.apache.org/docs/latest/design/index.html. [Accès le 30 04 2020].

[3 [En ligne]. Available: https://druid.apache.org/blog/2014/03/17/benchmarking-druid.html .
4]

[3 [En ligne]. Available: https://imply.io/post/performance-benchmark-druid-presto-hive.
5]

[3 [En ligne]. Available: https://imply.io/post/apache-druid-google-bigquery-benchmark .
6]

[3 R. S. P. V. P. R. R. J. Vijayaraj, «A COMPREHENSIVE SURVEY ON BIG DATA ANALYTICS TOOLS,»
7] chez *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, 2016.

[3 «Metatron development background and Druid integration».
8]

[3 «Apache Druid vs Spark».
9]

[4 «Evaluation on Druid,» September 6, 2018 .
0]

[4 «DATA TRANSFORMATION FUNCTIONS,» [En ligne]. Available: https://www.etl-
1] tools.com/articles/data-transformation-functions.html. [Accès le 19 05 2020].

[4 «Data Transformation Defined,» [En ligne]. Available:
2] https://www.talend.com/resources/data-transformation-defined/. [Accès le 04 06 2020].

[4 [En ligne]. Available: https://druid.apache.org/docs/latest/querying/granularities.html.
3]

[4 [En ligne]. Available: https://druid.apache.org/docs/latest/design/segments.html.
4]

[4 K. Coyle, «Mass Digitization of Books.,» *Journal of Academic Librarianship,* pp. 32(6), pp.641–
5] 645., 2006..

[4 Mayer-Schönberger, V. & Cukier, K., Big Data: A Revolution That Will Transform How We Live.
6] Work and Think, London: John Murray. , 2013..

ABSTRACT

[4 Michel, J.-B. et al, «Quantitative analysis of culture using millions of digitized books.,» *Science ,* 7] p. 176, 2011..

[4 D. Evans, The Internet of Things - How the Next Evolution of the Internet is Changing 8] Everything., CISCO white paper, 2011..

[4 Gartner, Gartner Says the Internet of Things Will Transform the Data Center., 2014. 9]

[5 L. I. A. &. M. G. Atzori, The Internet of Things: A survey. Computer Networks, 2010. 0]

[5 D. e. a. Estrin, Connecting the physical world with pervasive networks, IEEE Pervasive 1] Computing, 2002.

[5 M. L. M. &. R. R. Chui, The Internet of things., McKinsey Quarterly, 2010. 2]

[5 P. Russom, Big data analytics. TDWI Best Practices Report, Fourth Quarter., 2011.. 3]

[5 J. e. a. Manyika, Big data: The next frontier for innovation, competition, and productivity, 4] McKinsey Global institute, 2011.

[5 «Guru99,» 12 04 2020. [En ligne]. Available: https://www.guru99.com/what-is-big-5] data.html#:~:text=550000-,Unstructured,is%20classified%20as%20unstructured%20data.&text=A%20typical%20example%20of%20unstructured,files%2C%20images%2C%20videos%20etc.
.

[5 S. DeAngelis, «Understanding Big Data Beyond Its Size,» 14 07 2020. [En ligne]. Available: 6] https://www.enterrasolutions.com/blog/understanding-big-data-beyond-its-size/#:~:text="Semi-structured%20data%20can%20contain,that%20you%20have%20set%20up..

[5 «Filter Aggregation,» [En ligne]. Available: 7] https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-filter-aggregation.html#search-aggregations-bucket-filter-aggregation. [Accès le 17 06 2020].

[5 «SQL Server Aggregate Functions,» [En ligne]. Available: 8] https://www.sqlservertutorial.net/sql-server-aggregate-functions/. [Accès le 04 06 2020].

[5 «microstrategy,» [En ligne]. Available: https://doc-9] archives.microstrategy.com/producthelp/10.7/ReportDesigner/WebHelp/Lang_1033/Content/ReportDesigner/list_of_available_aggregation_functions.htm. [Accès le 04 06 2020].

[6 S. Subotin, «Dashboard Design - Considerations and Best Practices,» Designers, [En ligne]. 0] Available: https://www.toptal.com/designers/data-visualization/dashboard-design-best-practices. [Accès le 13 06 2020].

ABSTRACT

[6 «Mobile App Analytics Best Practices: How to Improve UI, UX, and Performance,»
1] seventablets, [En ligne]. Available: https://seventablets.com/blog/mobile-app-analytics-best-practices-how-to-improve-ui-ux-and-performance/. [Accès le 09 06 2020].

[6 ngdata, [En ligne]. Available: https://www.ngdata.com/the-importance-of-scalability-in-big-
2] data-processing/. [Accès le 07 06 2020].

[6 M. Narula, «An Approach to Achieve Scalability and Availability of Data Stores,» ebay, [En
3] ligne]. Available: https://tech.ebayinc.com/engineering/an-approach-to-achieve-scalability-and-availability-of-data-stores/. [Accès le Availability of Data Stores].

[6 Niklas Ekstedt ; Carl Johan Wallnerström : Sajeesh Babu : Patrik Hilber : , «Reliability Data – A
4] Review of Importance, Use, and Availability,» September 2014.

[6 C. Crawford, 20 Jan, 2020. [En ligne]. Available: https://www.socialbakers.com/blog/social-
5] media-analytics-the-complete-guide. [Accès le 28 05 2020].

[6 Krish Krishnan, Shawn P. Rogers, «Social Analytics in the Enterprise,» *Social Data Analytics,*
6] 2015.

[6 Raghunath Nambiar ; Adhiraaj Sethi ; Ruchie Bhardwaj ; Rajesh Vargheese, «A Look at
7] Challenges and Opportunities of Big Data Analytics in Healthcare,» chez *IEEE International Conference on Big Data*, 2013.

[6 Zhi-Hua Zhou, Nitesh V. Chawla, Yaochu Jin, and Graham J. Williams, «Big Data Opportunities
8] and Challenges: Discussions from Data Analytics Perspectives».*IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE.*

[6 David J. Pauleen, William Y.C. Wang, «Does big data mean big knowledge? KM perspectives on
9] big data and analytics,» Emerald , 13 February 2017.

[7 Chao Shang ; Feng You, «Data Analytics and Machine Learning for Smart Process
0] Manufacturing: Recent Advances and Perspectives in the Big Data Era,» *Engineering,* pp. Pages 1010-1016, December 2019.

[7 S. Patnaik, «Social media data analysis,» 2018.
1]

[7 «review42,» [En ligne]. Available: 17. https://review42.com/how-much-time-do-people-
2] spend-on-social-media/. [Accès le 16 02 2020].

[7 Uthayasankar Sivarajaha; Zahir Irania ; Suraksha Guptab ; Kamran Mahroof, «Role of big data
3] and social media analytics for business to business sustainability: A participatory web context,» *Industrial Marketing Management,* April 2020.

[7 U. Ruhi, «Social Media Analytics as a Business Intelligence Practice:Current Landscape &
4] Future Prospects,» chez *Journal of Internet Social Networking and Virtual Communities*, August 2014.

ABSTRACT

[7 J. H. a. S. Kandel, «Interactive Analysis of Big Data,» *RIGHTSLINK,* vol. 19, n° %11, p. 5, 2012. 5]