Social Media Analytics with Streaming Data

Social media analytics solutions help organizations understand trending topics. Trending topics are nothing but subjects and attitudes that have a high volume of posts on social media. Sentiment analysis, also termed as opinion mining, makes utilization of social media analytics tools to determine attitudes toward a product or idea. Because of the hashtag subscription model, Real-time Twitter trend analysis is a great example of an analytics tool that enables organizations to listen to specific keywords (hashtags) and develop sentiment analysis of the feed.

“AdMagic” is a company that has a news media website is interested in gaining an advantage over its competitors by featuring site content that is immediately relevant to its readers. The company wants to explore the social media analytics tools on the topics that are relevant to readers by doing real-time sentiment analysis of Twitter data. To identify trending topics in real time on Twitter, the company needs real-time analytics about the tweet volume and sentiment for key topics.

You are appointed as a Streaming Analytics expert for this firm which is looking for utilizing the solutions / platforms available from the Streaming Analytics space. As the firms maturity level in the social media data analytics space is at very nascent stage, you need to help them to understand how Streaming Analytics is helpful in their several use cases and also further on identifying the various options of tools and platforms those can be leveraged for this activity.

Microsoft Azure is leading player in the field of streaming analytics. Under the umbrella term “Streaming analytics”, they have developed a several cloud services to handle various streaming analytics use cases in very simpler manner. One of the solution for streaming media analytics is described at this blog. You can refer to this blog or other documentation provided by Microsoft team while interacting with the client.

Q1. You need to introduce the client with other streaming analytics tools available for streaming analytics which are suitable for the use case of social media analytics. For that purpose, you need to formulate a comparison that describes the available tools / solutions along with their strengths and weaknesses.
Narration should have

- Brief description of the social media analytics use case scenario.

Social media analytics is the ability to gather and find meaning in data gathered from social channels (Websites and channels: Facebook, YouTube, Instagram, Twitter, LinkedIn, Reddit and many others) to support business decisions — and measure the performance of actions based on those decisions through social media.

Social media analytics is broader than metrics such as likes, follows, retweets, previews, clicks, and impressions gathered from individual channels.

Social media analytics can be leveraged for a wide range of activities

a. Customer Experience
b. Brand maintenance
c. New trends
d. Sentiment analysis

AdMagic could be benefitted by knowing the above-mentioned parameters through social media analysis.

Customer experience:

The client base for the company can be increased and the loyalty of the existing customers can be maintained by analysing the customer attitudes (Key words)

Brand maintenance:

It measures the people's attitude and behaviour towards the services provided, it makes use of conversation and sentiment drivers, time, location and impact of an article, influencing and issue identification.

New trends:

Real time social media analytics can help in understanding the present trends that are going on in the society ranging from politics to health and many more which could be used for advantage in creating new news articles and feeds.

Sentiment analysis:

This could be used to understand the people's reactions and their perception towards a problem or issue and the news could be generated in such a way to attract a larger group of audience.

o At least three different on-premise or cloud tools / solutions identified and reasoning for the same

- **Microsoft Azure Stream Analytics** - It is designed to be easy to use, flexible, reliable, and scalable to any job size. It is available across multiple Azure regions, and runs on IoT Edge or Azure Stack.
- **Amazon Kinesis Data Analytics** - It takes care of everything required to run streaming applications continuously, and scales automatically to match the volume and throughput of your incoming data. With Amazon Kinesis Data Analytics, there are no servers to manage, no minimum fee or setup cost, and you only pay for the resources your streaming applications consume.

- **Google Cloud DataFlow** - Ingest, process, and analyze event streams in real time. Google Cloud's stream analytics solutions make data more organized, useful, and accessible from the instant it's generated.

o short explanation about each tool / solution - how it can be used for social media analytics

Microsoft Azure Stream Analytics - An Azure Stream Analytics job consists of an input, query, and an output. Stream Analytics ingests data from Azure Event Hubs (including Azure Event Hubs from Apache Kafka), Azure IoT Hub, or Azure Blob Storage.

The query, which is based on SQL query language, can be used to easily filter, sort, aggregate, and join streaming data over a period of time. You can also extend this SQL language with JavaScript and C# user-defined functions (UDFs). You can easily adjust the event ordering options and duration of time windows when performing aggregation operations through simple language constructs and/or configurations.

Each job has one or several outputs for the transformed data, and you can control what happens in response to the information you've analyzed. For example, you can:

- Send data to services such as Azure Functions, Service Bus Topics or Queues to trigger communications or custom workflows downstream.
- Send data to a Power BI dashboard for real-time dashboarding.
- Store data in other Azure storage services (for example, Azure Data Lake, Azure Synapse Analytics, etc.) to train a machine learning model based on historical data or perform batch analytics.

Amazon Kinesis Data Analytics - Kinesis Data Analytics enables you to quickly build end-to-end stream processing applications for log analytics, clickstream analytics, Internet of Things (IoT), ad tech, gaming, and more. The four most common use cases are streaming extract-transform-load (ETL), continuous metric generation, responsive real-time analytics, and interactive querying of data streams.

Google Cloud DataFlow - The Google Cloud Dataflow model works by using abstraction information that decouples implementation processes from application code in storage databases and runtime environments. In simpler terms, it works to break down the walls so that analyzing big sets of data and Realtime information becomes easier.

Dataflow runs on the same serverless, fully-managed model as many of the features on the GCP, and the idea behind this is that it means that developers in an organization have more freedom to keep their focus on developing innovative code, while the management and provisioning of computing needs can be left in the hands of the Dataflow service. The high level of abstraction that data scientists can tap into means that they can work at a more productive and efficient level.

o justification about the comparison parameters and relevant detailing

	Microsoft Azure Stream Analytics	**Amazon Kinesis Data Analytics**	**Google Cloud DataFlow**
Ease of getting started	takes a few clicks to connect to multiple sources and sinks, creating an end-to-end pipeline	Easy. setup streaming data sources, write your queries or streaming applications, and setup destination for processed data	Easy. Ingest and analyze hundreds of millions of events per second from applications or devices virtually anywhere on the globe with Pub/Sub.
Programmer productivity	SQL query language that has been augmented with powerful temporal constraints to analyze data in motion. Jobs by using developer tools like Azure PowerShell, Azure CLI, Stream Analytics Visual Studio tools, the Stream Analytics Visual Studio Code extension, or Azure Resource Manager templates, also to develop transformation queries offline and use the CI/CD pipeline to submit jobs to Azure.	Streaming ETL. Real-time analytics. Stateful event processing.	Allow teams to focus on programming instead of managing server clusters as Dataflow's serverless approach removes operational overhead from data engineering workloads.
Fully managed	Azure Stream Analytics is a fully	Amazon Kinesis is fully managed and	Fully managed data processing service,

	managed (PaaS) offering on Azure. You don't have to provision any hardware or infrastructure, update OS or software. Azure Stream Analytics fully manages your job, so you can focus on your business logic and not on the infrastructure.	runs your streaming applications without requiring you to manage any infrastructure.	automated provisioning and management of processing resources
Total cost of ownership	No upfront costs involved - you only pay for the usage.	No upfront costs involved - you only pay for the usage.	No upfront costs involved - you only pay for the usage.
Mission-critical ready	Available across multiple regions worldwide	Available across multiple regions worldwide	Available across multiple regions worldwide
Performance	Can process millions of events every second and it can deliver results with ultra low latencies. It allows you to scale-up and scale-out to handle large real-time and complex event processing applications. Stream Analytics supports higher performance by partitioning, allowing complex queries to be parallelized and executed on multiple streaming nodes. Azure Stream	The Amazon Kinesis Client Library (KCL) is a pre-built library that helps you build consumer applications for reading and processing data from an Amazon Kinesis data stream. The KCL handles complex issues such as adapting to changes in data stream volume, load balancing streaming data, coordinating distributed services, and processing data with fault-tolerance. The KCL enables you	Automated Resource Management: Minimize latency and boost performance with the automated management and provisioning of extra processing resources within the cloud structure. Auto-Scaling (Horizontal): Google Cloud Dataflow allows companies to horizontally scale their worker resources for very best performance throughout the enterprise.

	Analytics is built on Trill, a high-performance in-memory streaming analytics engine developed in collaboration with Microsoft Research	to focus on business logic while building applications.	
Input	Event Hubs IoT Hub Blob Storage	Kinesis Data Streams Kinesis Firehose	Pub/Sub
Output	Event Hubs Data Lake Storage Azure Stream Analytics SQL Database Blob storage PowerBI Table Storage Service Bus Queue Service Bus Topic Cosmos DB	Kinesis Data Streams Kinesis Firehose Lambda Function	Cloud Datalab Tableau QLikView

o a recommendation of the platform / tool for the media company use case

As per the above analysis, **Microsoft Azure** can be used for analysing the real-time Twitter trend analysis enabling **"AdMagic"** to listen to specific keywords (hashtags) and develop sentiment analysis of the feed.

Q2. You are in a meeting with the firm's management who are a little bit concerned about the capabilities associated with social media analytics tools discussed in question 1. The client is a bit hesitant to rely on the tools for this analytics. In order to assist the client

• Briefly narrate the at least five key capabilities of the tool / solution that you have recommended

- ➔ It is easy to set up with a few clicks to connect to multiple sources and sinks, creating an end-to-end pipeline.
- ➔ For developer's productivity, the SQL query language has been augmented with powerful temporal constraints to analyze data in motion. Jobs by using developer tools

like Azure PowerShell, Azure CLI, Stream Analytics Visual Studio tools, the Stream Analytics Visual Studio Code extension, or Azure Resource Manager templates, also to develop transformation queries offline and use the CI/CD pipeline to submit jobs to Azure.

- ➔ Azure Stream Analytics is a fully managed (PaaS) offering on Azure. You don't have to provision any hardware or infrastructure, update OS or software. Azure Stream Analytics fully manages your job, so you can focus on your business logic and not on the infrastructure.
- ➔ Cost wise, no upfront costs involved, you only pay for the usage.
- ➔ Available across multiple regions worldwide, for any critical implementation.
- ➔ From a performance perspective, it can process millions of events every second and it can deliver results with ultra low latencies. It allows you to scale-up and scale-out to handle large real-time and complex event processing applications. Stream Analytics supports higher performance by partitioning, allowing complex queries to be parallelized and executed on multiple streaming nodes.
- ➔ Azure Stream Analytics is built on **Trill**, a high-performance in-memory streaming analytics engine developed in collaboration with Microsoft Research.
- ➔ Input sources - Event Hubs, IoT Hub, Blob Storage
- ➔ Output - Event Hubs, Data Lake Storage, Azure Stream Analytics, SQL Database, Blob storage, PowerBI, Table Storage, Service Bus Queue, Service Bus Topic, Cosmos DB

• Address how each of this key capability can be leveraged for the use case identified in part 1

Ease of getting started

As it is a new setup and being introduced for the first time it is easy to deploy Azure stream analytics as it can be set up just with a few clicks.

Programmer productivity

As the data can be analysed using simple SQL query language no new language or programming needs to be learnt and can be handled using the existing knowledge and the usage of SQL query language is high and complexity is low making it ideal solution to train the resources.

Fully managed

There is no requirement to procure or maintain any hardware all the hardware related configuration and maintenance is managed by Azure itself making it easy and economically viable solution

Low total cost of ownership

you only pay for the streaming units you consume. There is no commitment or cluster provisioning required, and you can scale the job up or down based on your business needs.Streaming Units (SUs) represents the computing resources that are allocated to execute

a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. This capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream Analytics job in a timely manner.

Reliability

As a managed service, Stream Analytics guarantees event processing with a 99.9% availability at a minute level of granularity. Azure Stream Analytics has built-in recovery capabilities in case the delivery of an event fails, it also has a built-in checkpoint to maintain the state of your job and provides repeatable results.

Performance:

The latency is extremely low and can process millions of events every second, scaling up or down is possible with real time events or complex events, it allows parallel execution of queries on multiple streaming nodes.

Q3. The blog discusses use cases which the media company has addressed in space of social media analytics. But the solution is described in terms of various cloud services offered by Microsoft Azure. The client does not have the knowledge about the cloud computing and Azure. In fact all the use cases can be very well addressed with a general architecture used in the big data analytics and streaming analytics. You need to work upon helping client to understand those common architectures.

- Identify the architecture that can be fitted well for capturing the use cases

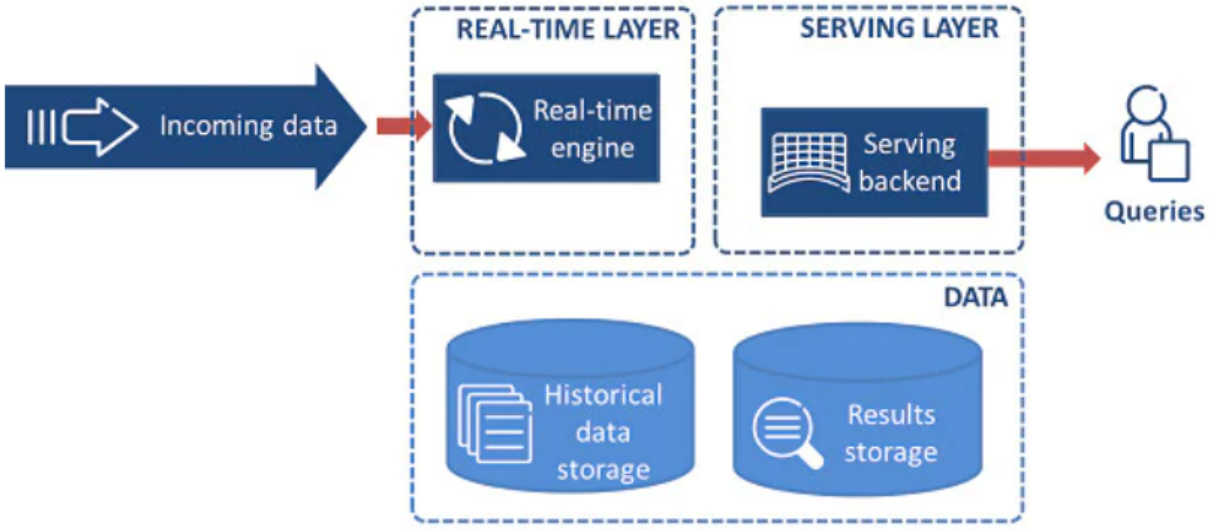
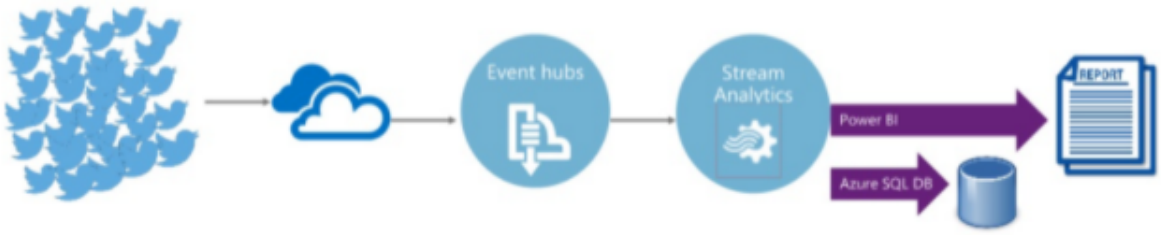


Figure 2 Kappa architecture

- Prepare an architecture diagram based upon your answer to earlier question



- Take care that use cases should be vividly coming out of the architecture diagram, if required add brief description about each flow

- Using Twitter API, we can tap into public conversation to extract the data.
- This high volume data is handled by the Azure Events Hub, which is a big data streaming platform and event ingestion service will be transformed and stored
- The data is then sent to Full managed Azure Stream Analytics, which is real-time analytics designed to help you analyze and process fast moving streams of data
- The processed data is then used to get insights, build reports or trigger alerts and actions.

Q4. The client is now impressed with the capabilities of the Microsoft Azure and how it's streamlining the application development and deployment. But they also want to discover more on the open source tools / platforms that can be leveraged. As a result, you need to work upon identifying the open source tools for the use case.

- Identify the tools / platforms that can be used to solve it

1. **Socioboard**

Socioboard is a free and open source social media management software that has a powerful social media suite for all businesses. It generates reports and analyzes data to help social media managers make informed business decisions.

Features:

- Live chat, a news feed, timeline, profiles, events, notifications
- Highly customizable and scalable open-source interface.
- Scheduling & publishing is easy with Socioboard
- It generates tickets to reply to negative comments
- RSS Feed to distribute a list of headlines, notices, and content to the audience automatically
- Google Analytics integration to track website traffic.
- The innovative feature of Sentiment analysis conducts language processing, text analysis, and computational linguistics.
- Sophisticated analytics and interactive social discovery tools
- Team collaboration tools and 24/7 Technical Support
- Social CRM tools and Helpdesk integration

2. Apache Kafka

To handle a high volume of data and enables us to pass messages from one end-point to another, Apache Kafka is a distributed publish-subscribe messaging system. It is suitable for both offline and online message consumption.

Moreover, in order to prevent data loss, Kafka messages are persisted on the disk and replicated within the cluster. In addition, it is built on top of the ZooKeeper synchronization service.

While it comes to real-time streaming data analysis, it can also integrate very well with Apache Storm and Spark. There are many more features of Apache Kafka. Let's discuss them in detail.

Features :

a. **Scalability**
Apache Kafka can handle scalability in all the four dimensions, i.e. event producers, event processors, event consumers and event connectors. In other words, Kafka scales easily without downtime.

b. **High-Volume**
Kafka can work with the huge volume of data streams, easily.

c. **Data Transformations**
Kafka offers provision for deriving new data streams using the data streams from producers.

d. **Fault Tolerance**
The Kafka cluster can handle failures with the masters and databases.

e. **Reliability**
Since Kafka is distributed, partitioned, replicated and fault tolerant, it is very Reliable.

f. **Durability**
It is durable because Kafka uses Distributed commit log, that means messages persists on disk as fast as possible.

g. **Performance**
For both publishing and subscribing messages, Kafka has high throughput. Even if many TB of messages is stored, it maintains stable performance.

h. **Zero Downtime**
Kafka is very fast and guarantees zero downtime and zero data loss.

i. **Extensibility**
There are as many ways by which applications can plug in and make use of Kafka. In addition, offers ways by which to write new connectors as needed.

j. **Replication**
By using ingest pipelines, it can replicate the events.

- Draw a solution diagram using the tools identified in earlier question the flow should come out clearly from the solution diagram