**SPA Assignment-1**

**User case: Social Media Analytics with Streaming Data**

**GROUP - 140 - Pothula Maruthi Chowdary ,Jayadhar Alla**

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](https://docs.microsoft.com/en-in/azure/stream-analytics/stream-analytics-twitter-sentiment-analysis-trends) blog. You can refer to this blog or other documentation provided by Microsoft team while interacting with the client.

**Question 1 and 2:**

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 strength and weaknesses.

* Narration should have
  + brief description of the social media analytics use case scenario
  + at least three different on-premise or cloud tools / solutions identified and reasoning for the same
  + short explanation about each tool / solution - how it can be used for social media analytics
  + justification about the comparison parameters and relevant detailing
  + a recommendation of the platform / tool for the media company use case

Q2. You are in a meeting with the firm’s management who are little bit concerned about the capabilities associated with social media analytics tools discussed in question 1. The client it 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
* Address how each of this key capability can be leveraged for the use case identified in part 1

**Scenario: Social media sentiment analysis in real time**

We have multiple Streaming tools serving the different kind of processings,

But here considering the Social media analytics — Requirement is to have the Site

Social Media sentiment — Data should be collected from all the relevant social sites and need to understand the End users priorities and serve them accordingly to their convenience.

In order to process without any overhead or latency and giving the best experience for the users, ADMAGIC site should incorporate the best possible streaming solution. We have Multiple Streaming tools available for the same.

Tool should have the required abilities — few of them are : Effective and Efficient processing of the data and provide insights, Highly available , Real time analytics , Inbuilt AI/Machine learning etc.

Few of the Tools were explained below with all of their characteristics.

Amazon Kinesis, Google Cloud Data flow and APACHE STORM along with Microsoft Azure Stream Analytics .

**Amazon Kinesis:**

Amazon Kinesis makes it easy to collect, process, and analyse real-time, streaming data such as video, audio, application logs, website clickstreams, and IoT telemetry, so you can get timely insights and react quickly to new information.

**Abilities and Advantages:**

>Kinesis has the ability to consume messages with the right sharing and a managed serviced and support large workload.

>It has sdk documentation, the usual speed, scale, availability etc. >Kinesis has the ability to integrate with various AWS services to consume and build easily.

>We can quickly set up an architecture that allows to consume live streaming data in order to do some processing and enrichment to make our customers more efficient with their data.

> It has distributed messaging queue.

> It has durability and ability to fire and forget.

**Drawbacks for this Use case:**

No support for Direct streaming and 1 megabyte limit, in our use case we may have files that go beyond 1 mg even after compression. some of the record/batch size limitations and need to manage it to prevent unexpected failures

**Google Cloud Data flow:**

Google Cloud Data flow is the Unified stream and batch data processing that's serverless, fast, and cost-effective.

It is Fully managed data processing service and provides Automated provisioning and management of processing resources

Google Cloud Dataflow frees you from operational tasks like resource management and performance optimization and it is a tool in the Real-time Data Processing category of a tech stack.

It has the feasibility of horizontal autoscaling of worker resources to maximize resource utilization and OSS community-driven innovation with Apache Beam SDK Reliable and consistent exactly once processing

**Key features:**

Autoscaling of resources and dynamic work rebalancing

Flexible scheduling and pricing for batch processing

Ready-to-use real-time AI patterns

Vertical autoscaling and Streaming Engine

**Drawbacks for this Use case:**

The major disadvantage of Google Cloud dataflow was it was bound to google technologies, though it was leading tool in the market, we can’t switch out easily to other technologies. And when compared to other competitors it was bit more price.

**APACHE STORM:**

Apache Storm is a distributed Realtime computation system and makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing.

Apache Storm is fast,a benchmark clocked it at over a million tuples processed per second per node. It is scalable, fault-tolerant, guarantees your data will be processed, and is easy to set up and operate. Apache Storm integrates with the queueing and database technologies you already use.

An Apache Storm topology consumes streams of data and processes those streams in arbitrarily complex ways, repartitioning the streams between each stage of the computation however needed

Apache Storm is simple, can be used with any programming language and has many use cases:

> Real Time analytics

>online machine learning,

>continuous computation, >distributed RPC,

>ETL, and more.

**Drawbacks for this Use case:**

Apache Storm can’t schedule the jobs, which can be an bottleneck while using real-time huge data streaming applications, It’s network highly depends on stream grouping and tuple arrival rate.

**AZURE STREAM ANALYTICS:**

Azure Stream Analytics is a real-time analytics and complex event processing engine that is designed to analyse and process high volumes of fast streaming data from multiple sources simultaneously.

Patterns and relationships can be identified in information extracted from a number of input sources including devices, sensors, clickstreams, social media feeds, and applications.

These patterns can be used to trigger actions and initiate workflows such as creating alerts, feeding information to a reporting tool, or storing transformed data for later use. Also, Stream Analytics is available on Azure IoT Edge runtime, enabling to process data on IoT devices.

The following scenarios are examples of when you can use Azure Stream Analytics:

>Analyse real-time telemetry streams from IoT devices

>Web logs/clickstream analytics

>Geospatial analytics for fleet management and driverless vehicles

>Remote monitoring and predictive maintenance of high value assets >Real-time analytics on Point of Sale data for inventory control and anomaly detection

**DECISION MADE FOR THE CLIENT:**

Considering the Requirement for the given use case of Social Media Analytics, Upon Comparing the features and abilities of multiple Streaming tools , We feel that AZURE STREAM ANALYTICS is apt and best suited platform which meets all the required functionalities of the ADMAGIC Site .

**Question 2:**

**Key Capabilities for Microsoft Azure:**

1. Ease to start and adopt: Azure stream tools are easy to start and adopt. It takes few clicks to minutes to get into that and connect different sources and sinks, creating end-to-end pipeline.
2. Programmer productivity: Azure Stream Analytics uses a SQL query language that has been augmented with powerful temporal constraints to analyze data in motion. You can also create jobs by using developer tools like Azure PowerShell, Azure CLI, Stream Analytics Visual Studio tools.
3. Fully managed: 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.
4. Run in cloud: Azure Stream Analytics can run in the cloud, for large-scale analytics, or run on IoT Edge or Azure Stack for ultra-low latency analytics. Azure Stream Analytics uses the same tools and query language on both cloud and the edge, enabling developers to build truly hybrid architectures for stream processing.
5. Low total cost of ownership: As a cloud service, Stream Analytics is optimized for cost. There are no upfront costs involved - 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.
6. Reliability: Azure Stream Analytics guarantees exactly once event processing and at-least-once delivery of events, so events are never lost. Exactly once processing is guaranteed with selected output as described in Event Delivery Guarantees.
7. Security: Azure Stream Analytics encrypts all incoming and outgoing communications and supports TLS 1.2. Built-in checkpoints are also encrypted. Stream Analytics doesn't store the incoming data since all processing is done in-memory. Stream Analytics also supports Azure Virtual Networks (VNET) when running a job in a Stream Analytics Cluster.
8. Performance: Stream Analytics 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.

For this use case, we need Fully managed, low total cost of Ownership, reliability, running in cloud, security and performance are important.

**Question 3**: The blog discusses use case 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.

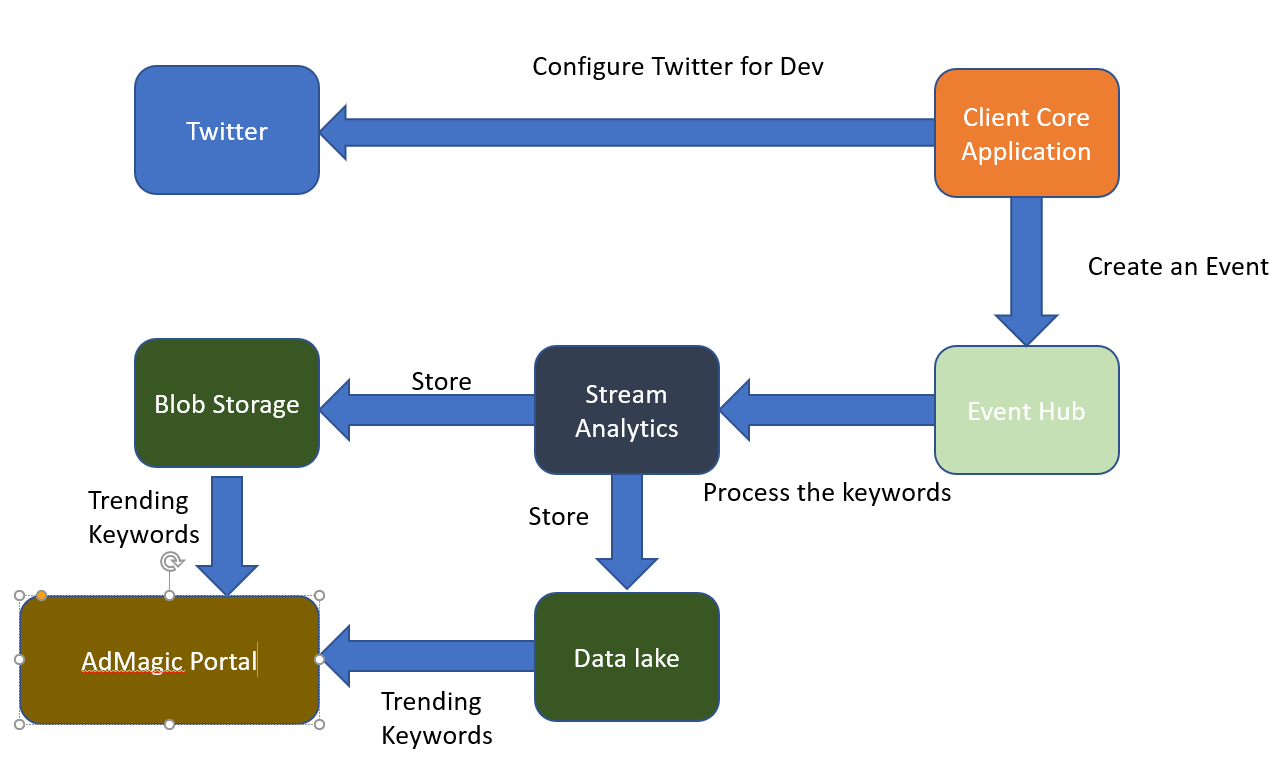
* Identify the architecture that can be fitted well for capturing the use cases
* Prepare an architecture diagram based upon your answer to earlier question
* Take care that use cases should be vividly coming out of the architecture diagram, if required add brief description about each flow

Use case: The company uses social media analysis on topics that are relevant to readers by doing real-time sentiment analysis of Twitter data. For that, need to identify the trending topics on twitter and we need to sentiment the key topics.

**Architecture with use case using Microsoft Azure**:

The Architecture consists of two blocks,

1. Configure Client core Application
2. Create an Event from Event Hub in Microsoft Azure
3. Process the huge amount of data through stream analysis.
4. Store it in some cloud or local server
5. Get the keywords trending in twitter
6. Search for news related to Keywords we got through twitter sentiment.
7. Display the Advertisements through AdMagic GUI.



**Question 4:**

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

There are few open source tools which can be used for real-time streaming analysis.

1. Apache Kafka
2. Apache flink
3. Apache Smaza
4. Apache Storm
5. IBM stream Analytics
6. Spark Streaming

Apache Storm is a real time computation system, Apache Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing. Apache Storm is simple, can be used with any programming language. Apache Storm has many use cases: realtime analytics, online machine learning, continuous computation, distributed RPC, ETL, and more. Apache Storm is fast: a benchmark clocked it at over **a million tuples processed per second per node**. It is scalable, fault-tolerant, guarantees your data will be processed, and is easy to set up and operate. Apache Storm integrates with the queueing and database technologies you already use. An Apache Storm topology consumes streams of data and processes those streams in arbitrarily complex ways, repartitioning the streams between each stage of the computation however needed. The foremost capability of Apache Stream is faster data processing. Apache Stream can carry out processes at the nodes with faster data processing than its competitors. Most important of all, you can integrate Apache Storm with Hadoop to improve its capability for larger throughputs.

The major advantage for this use case for Apache Storm was built by twitter and it aims at data streams.

