SPA Assignment 1

Group Number 97

No.	Member	Student ID
1	Shreysi Kalra	2021fc04586
2	Vinayak Nayak	2021fc04135
3	Ajith Praveen R	2021fc04329

Streaming Analytics

You are appointed as a Streaming Analytics expert for a firm which is looking for utilizing the solutions / platforms available from the Streaming Analytics space. As the firms maturity level in the big data 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.

Amazon Web Service (AWS) is leading player in the space of cloud computing. They have developed a special cloud service named “Amazon Kinesis” exclusively for handling various streaming analytics use cases in very simpler manner. To introduce these services to the world, they also have prepared a nice documentation – part of which also contains the white paper. You can refer to this white paper which can help you while interacting with the client.

Question

You need to introduce the client with several examples where streaming analytics has already been used. For that purpose, you need to formulate one example of each type of Real-time application scenarios mentioned in the white paper.

- The example should be different from the ones discussed in the document
- Narration should have
  - brief description of the use case scenario
  - short explanation about how it can leverage streaming analytics solutions / platforms
  - justification about how it falls under the particular category

Client Pitch

In today's fast paced real world, data is the new oil. Lots of data is being generated at an ever increasing pace and it holds the keys to unlock a lot of opportunities for business. Previously, CoEs would munge over the data for days/weeks and come up with reasonable meaningful insights which could help the business/organization identify new opportunities and decide their action plan accordingly. That still happens in the today's world, but over the course of last decade or so, attention spans of people have reduced and technological advances have been huge which has fuelled the need for instant gratification much more. *So, we need to draw* **actionable insights** *from* **lots of data** *over both* **smaller and bigger timespans** *to best suit the needs of end users/businesses*.

To elaborate, there are two main processing patterns – `Stream processing` which expects real/near-real time analysis and responses to events in data streams; And `Batch processing` where some analysis is done periodically over hours/days/fortnight/mmonth and the results of this analysis are used to drive the underlying process/phenomenon forward.

Let us understand this further with an example or a real time application in Auto Industry. Streaming data and real time analytics are disrupting the rental car industry.

These days, renting a car for family vacation is commonplace. Safety is a central aspect which no user would compromise on. Can the rental company predict breakdown and alert their drivers to take the necessary action in earnest? Can they provide incentives to drivers based on their driving behaviors like speeding or fuel efficiency? Could they alert the driver whether or not s/he is authorized to drive into/out of a particular geographical region? This is possible using a network of vehicles/cars connected amongst themselves over the web with the use of `Stream-Processing` and `Batch-Processing` architectures.

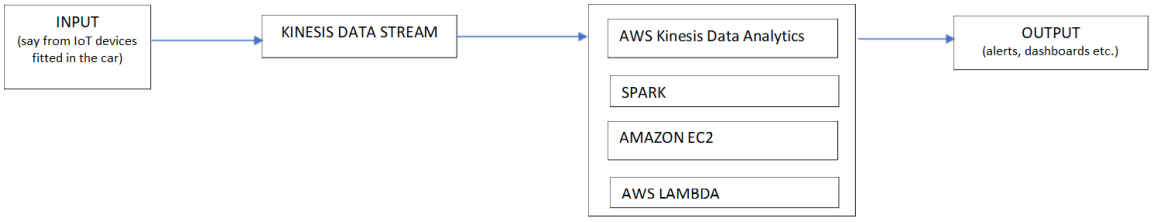
Here, data collected can be categorized as :

- Behavioral Data – Track driver's use of vehicle using indicators like speed, steering, braking and fuel efficient driving. Data can be used to award incentives or recognitions for good driving behavior and/or provide constructive feedback in case of rash/irresponsible driving.
- Diagnostic Data – Track the health of vehicle using indicators like engine temperature, tyre pressure, fuel tank etc. and notify drivers when a checkup/maintainence/service is required

Requirement 1 – Predict breakdown and alert drivers (stream processing architecture)

Diagnostic data enables car rental agencies to assess the heath of a vehicle and notify drivers when a service is required with in car voice communication. This is a classic example of building a real time analytics solution. This will comprise of two components –

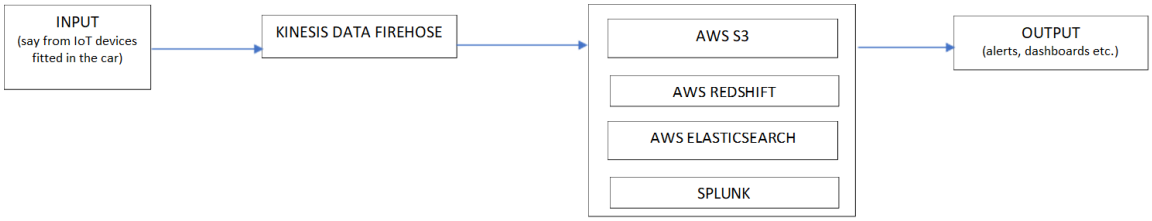
- First, stream processors continuously collect & parse data from sources such as sensors fitted within the car as it occurs and then delivers data to a streaming transport system
- Secondly, streaming analytics solution consume data from streaming transport systems over a limited time window that allows for data manipulation, enrichment and analysis
- Ultimately data is delivered for a variety of uses such as alerting, real time visualization or persisting event data for historical analysis later.



Requirement 2 – Recognize good driving behavior (batch or micro batch processing architecture)

Behavioral data like the speed of the vehicle, braking information and fuel efficiency can be sent to Kinesis Firehose so that they are stored within an S3 bucket. The analytics application can later connect to S3 to perform analysis and identify drivers with good, average or bad driving behaviors. People with good driving behaviors can be provided with some personalized offers for any subsequent car hires.

Since this doesn't require any real time analytics, therefore the application can consume the events data stored within S3 later in the day to calculate driver efficiency scores based on which the recommendations or future incentives are provided. And since it doesn't involve real time stream analytics, therefore this can be categorized under batch or micro-batch processing architecture.



Amazon Kinesis Data firehose is the easiest way to reliably load streaming data into data lakes, data stores and analytic tools. It can capture, transform and load streaming data into S3, Redshift or Elastic Search. Splunk enables near real time analytics with existing business intelligence tools and dashboards

---

Question

You are in a meeting with the firm’s management who are little bit concerned about the challenges associated with streaming analytics. The white paper describes few challenges faced while adapting the streaming analytics. In order to assist the client

Briefly narrate the four critical challenges in your own words

Critical Challenges

Some critical challenges faced with real time stream processing analytics include the following

- Developing and maintaining custom streaming data pipelines is cumbersome and resource intensive as it involves collecting, preparing, and transmitting data being generated simultaneously plethora of data sources.
- Storage and compute come at a cost; they must be efficiently utilized to retrieve & transmit data efficiently for maximum performance, low latency, high throughput and so on. A high number of servers or compute capacity might be required to accommodate varying volumes and speeds of incoming data.
- Continual monitoring the system and recovering from any server or network failures gracefully without being unresponsive/unavailable will be essential.
- Load Balancing/Scaling/Elasicity is critical as request traffic would not be uniform; there would be high traffic in the holiday season like New Year's Eve or Christmas/Diwali etc.
- Security/Governance of data is crucial. Since customers/drivers entrust us with some of their PII information, it is critical that we keep such information secure and out of reach of potential spammers/scammers etc.

Identify the different tools that can be used to resolve / mitigate those challenges

Using a cloud service provider like AWS/GCP who provide infrastructure/software as a service apparently resolves a lot of the above concerns. AWS real time data streaming services (like Kinesis) enable us to collect, process and analyze continuous streaming data at scale and take necessary action as and when needed. We can build real time applications and leverage secure, highly available, durable and scalable managed services provided by AWS. Some tools / services that can be used to address the highlighted challenges are as follows.

1. Stream Ingestion – Services AWS IoT Core could integrate with continuous data produced from various data sources in a durable and secure manner. AWS Cloudwatch stores log streams for a lot of services which could also be used at a later point in time for analysing anomalies, or service breakdowns/failures etc.
2. Stream Storage – We can choose between Amazon Kinesis Data Streams, Kinesis Data Firehose and Managed Streaming for Apache Kafka (MSK) that meets our storage needs based on scaling, latency and processing requirements.
3. Stream Processing – We can choose a selection of services ranging from solutions to transform and deliver data continuously to a destination like Kinesis Data Firehose to real time applications and analytics integrations like Amazon Kinesis Data Analytics and Apache Druid etc.
4. Destination – Deliver streaming data to data lakes, data warehouses and analytics services like S3, Redshift, Elastic Search Services and Amazon EMR
5. Security - Our services could be provided as APIs managed through AWS API gateway which could be used for creating, publishing, maintaining, monitoring, and securing REST, HTTP, and WebSocket APIs at any scale.

Address how each of the challenge is resolved with the tools / platforms identified

- Using the managed services provided by AWS, we will be able to create the required server infrastructure within few clicks and therefore the set-up time to have an environment up and running is drastically reduced.
- Storage systems like S3 is most cost effective and also processing elements like Lambda are serverless and therefore they are the quite efficient and cost-effective ways for storing & computation respectively.
- Auto scaling is handled by the managed service and therefore the number of shards / servers involved in Kinesis data stream is increased or reduced based on the data volume that it needs to handle at that given point in time. Same is the case with any other processing service as well like lambda. It supports horizontal and vertical scaling based on request traffic and task severity.
- Cloudwatch and Cloudtrail could help in monitoring and maintainence and API gateways keep the communication between the client and server secure.

Question

The white paper discusses three different use cases which the toll station company has addressed using streaming data. But the solution is described in terms of various cloud services offered by AWS. The client does not have the knowledge about the cloud computing and AWS. In fact all the three 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 all three use cases

Explanation

Let's cover each of these points one at a time.

For our usecase we have to use both batch and stream processes for different subtasks and Lambda architecture was built to address this very issue where we want the best of both worlds. Let's dive into specific use-cases where these two paradigms could be used.

Batch Processing

Every car/vehicle would be generating data that's getting logged to some data store. Such data persisting in databases like dynamo and S3 can be used by business analysts to analyse/identify

- The number of drivers with least `Repair/Maintain/Check Car` requests.
- `Average number of miles` travelled before a breakdown happens due to engine failure
- `Mean time to breakdown` due to pressure drop in tyres etc. which is not duly attended to by the driver
- Average number of highway `road miles` vs village road miles; Mean time to breakdown on highway vs village roads etc.

This sort of aggregation could be done by vehicle managers/ business analysts and other stakeholders to gauge vehicle performance, driver performance etc. and it need not be real-time in nature. We could use all the historical records and whenever needed we could perform aggregations to get answers/perform analysis for the above questions or similar questions.

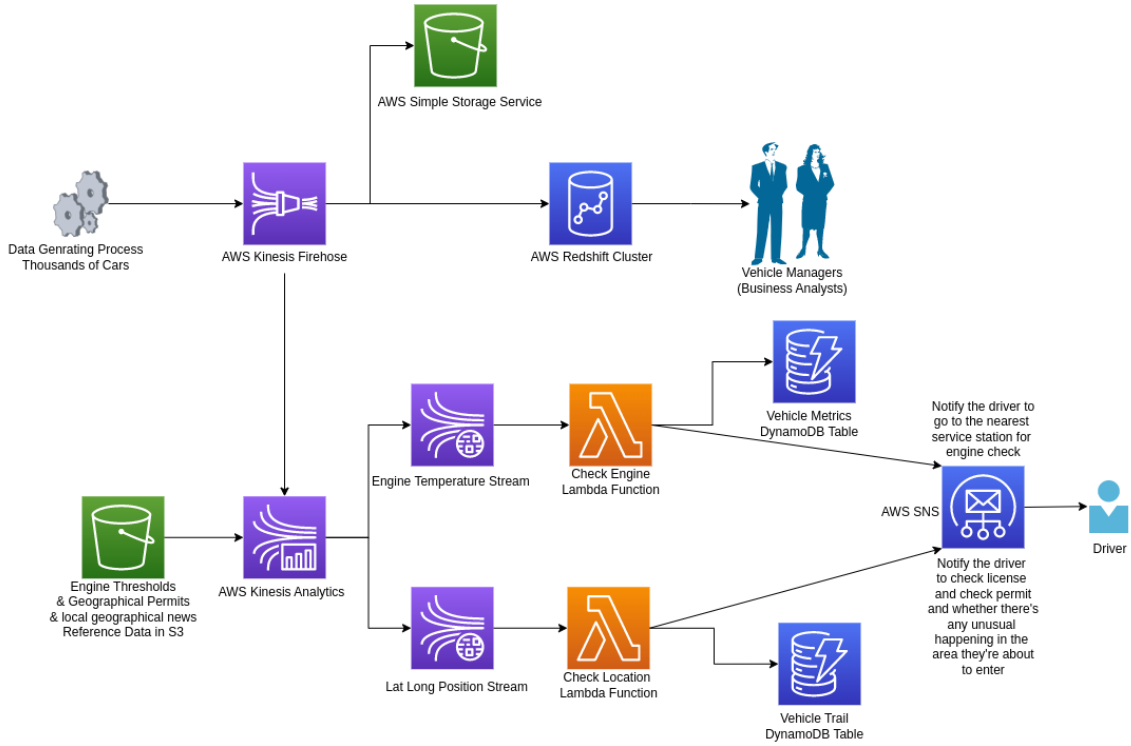
Stream Processing

Drivers on the run need to be on their toes to give the customers a pleasant experience during their ride and safety is of the utmost importance here. Any indications of breakdown/failures should be caught preemptively and must be corrected/acted upon. CHECK ENGINE signal is one such important indicator and the driver must be notified in real time if there's a possibility of imminent failure.

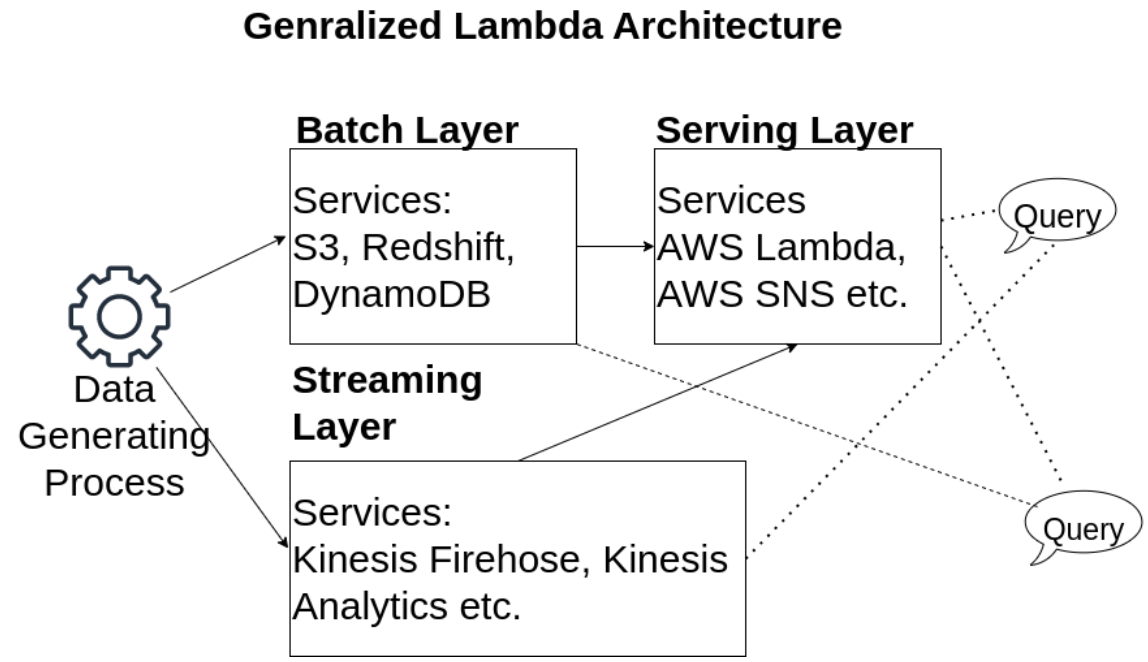
Similarly, if there is a border that shouldn't be crossed, or a region which has an announcement of curfew etc., the driver must be intimated well in advance based on the direction that he's headed so that he could slow down, comprehend the notification and then hit the brakes to avoid accidentally stepping in prohibited areas. Here we need a stream processing capability to be present in our architecture.

Hence, the Lambda Architecture is the best suited for our use case as it allows us to have the best of both worlds with a slight downside of increased complexity and maintenance but eventually it's worth the pain to support our use-case.

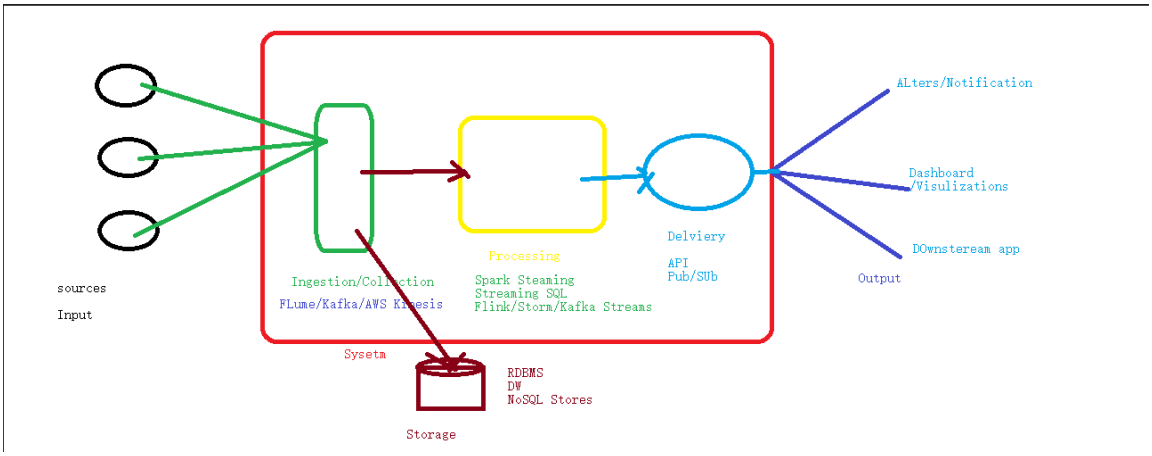
Convert the final architecture diagram provided by AWS team into an architecture diagram based upon your answer to earlier question. Take care that all three cases should be vividly coming out of the architecture diagram, if required add brief description about each flow



In a more generalized sense the following diagram explains the components from a high level perspective



This diagram could be thought of wrt what we had discussed in class as follows



- We have a data ingestion mechanism like kinesis meant to consume sensor readings via network through multiple sources
- We have a data store where we collect and persist necessary information periodically
- There is a server that is performing some aggregations/rule based/ML based computation and again generating insights in the form of data
- There are client side applications to which these results and the corresponding decisions/meanings are fed so that they could take the right set of actions in real time.

Question

The client is now impressed with the capabilities of the AWS 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 each of the use case.

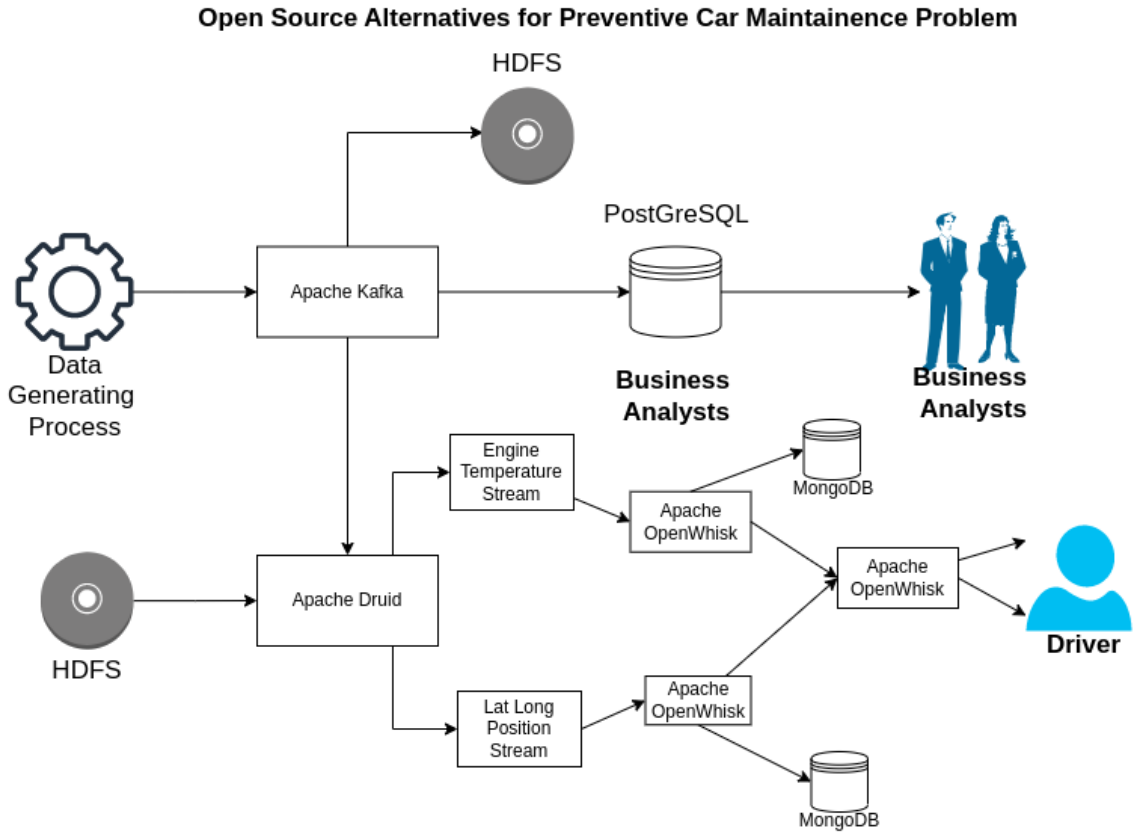
Identify the tools / platforms that can be used to solve it

Open Source Alternatives to the services used in the above architecture diagram are as follows

AWS Service	Opensource Service
AWS S3	HDFS
AWS Kinesis	Apache Kafka/Apache Flume
AWS Kinesis Data Analytics	Apache Druid
DynamoDB	MongoDB/MongoDB Atlas
AWS Lambda	Apache OpenWhisk
AWS Redshift	PostGreSQL

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

The final solution diagram after incorporating the above services looks as follows:



PS: SNS's job would be done by OpenWhisk. We like to think of it as we didn't have SNS but a notification AWS Lambda in the corresponding AWS architecture above.