Junaid A Girkar

60004190057

TE Comps A-4

EXPERIMENT – 1

1) **What database have you selected?**

ANS 1) The database I have selected is Amazon Kinesis which is a part of the Amazon Web Services (AWS). Amazon Kinesis makes it easy to collect, process, and analyse real-time, streaming data so you can get timely insights and react quickly to new information. Amazon Kinesis offers key capabilities to cost-effectively process streaming data at any scale, along with the flexibility to choose the tools that best suit the requirements of your application. With Amazon Kinesis, you can ingest real-time data such as video, audio, application logs, website clickstreams, and IoT telemetry data for machine learning, analytics, and other applications. Amazon Kinesis enables you to process and analyse data as it arrives and respond instantly instead of having to wait until all your data is collected before the processing can begin.

2) **Technical Specifications of the database?**

ANS 2)There are four main components of Kinesis that can be used to accomplish different tasks using their AWS services.

- Kinesis Firehose
- Kinesis Data Analytics
- Kinesis Data Streams
- Kinesis Video Streams

To help ingest real-time data or streaming data at large scales, AWS customers turn to Amazon Kinesis Data Streams. Kinesis Data Streams can continuously capture gigabytes of data per second from hundreds of thousands of sources. The data collected is available in milliseconds, enabling real-time analytics.

To provide this massively scalable throughput, Kinesis Data Streams relies on shards, which are units of throughput and represent a parallelism. One shard provides an ingest throughput of 1 MB/second or 1000 records/second. A shard also has an outbound throughput of 2 MB/sec. As you ingest more data, Kinesis Data Streams can add more shards. Customers often ingest thousands of shards in a single stream.

3) **What application is using this and for what purposes?**

ANS 3) The major application using this database is Netflix.

Netflix is the world's leading internet television network, with more than 200 million members in more than 190 countries enjoying 125 million hours of TV shows and movies each day. Netflix uses AWS for nearly all its computing and storage needs, including databases, analytics, recommendation engines, video transcoding, and more—hundreds of functions that in total use more than 100,000 server instances on AWS.

Moreover, Netflix, a leading content producer, has used AWS to build a studio in the cloud. This virtual studio enables Netflix to engage top artistic talent, no matter the location, and Netflix artists and partners have the freedom to collaborate without technological or geographical barriers.

Another database that can be used as an alternative is Apache Kafka

Apache Kafka is an open-source "event streaming platform" — a platform that writes and reads event streams. Kafka handles data streams in real-time (like Kinesis.) It's used to read, store, and analyse streaming data and provides organizations with valuable data insights. Uber, for example, uses Kafka for business metrics related to ridesharing trips. The big difference between Kinesis and Kafka lies in the architecture. Kafka "decouples" applications that produce streaming data (called "producers") in the platform's data store *from* applications that consume streaming data (called "consumers") in the platform's data store. Kafka has more of a scattered nature than Kinesis, making it useful for node failures.

According to Apache, over 80 percent of Fortune 100 companies use (and trust) Kafka.

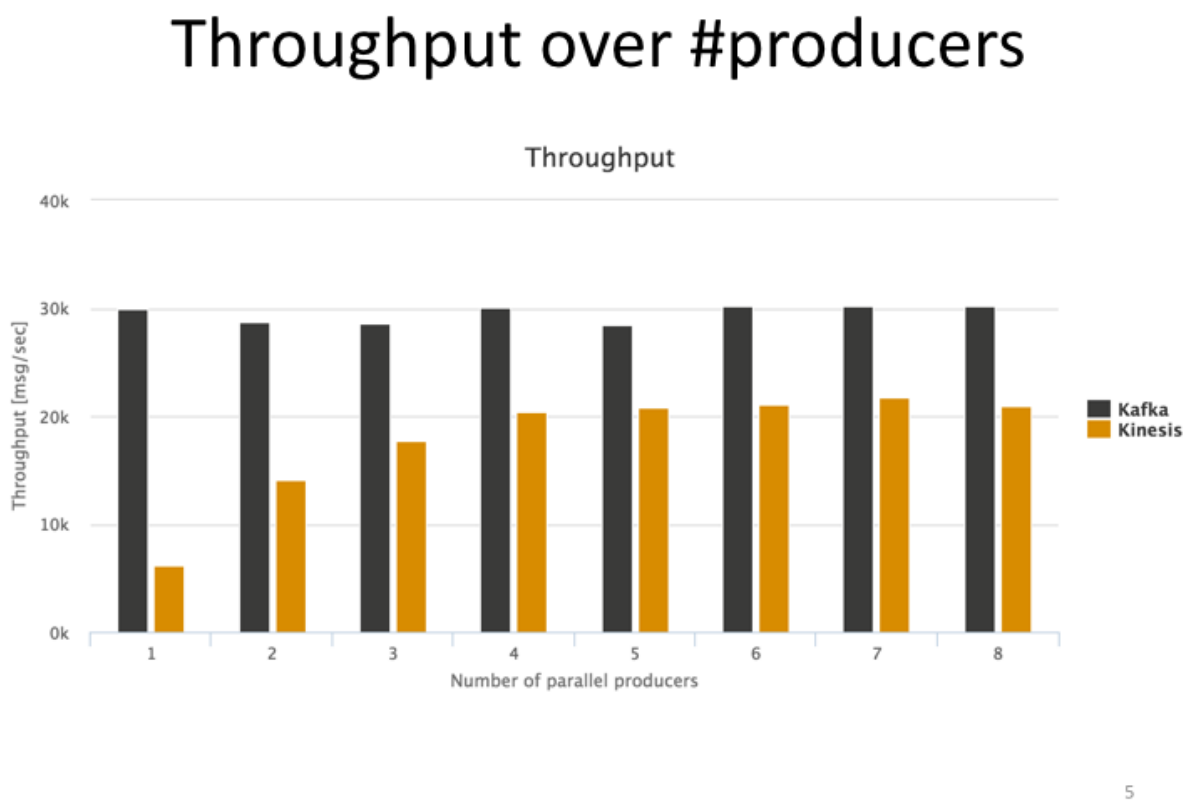
	KINESIS	KAFKA
Where is data stored?	Kinesis Shard	Kafka Partition
Support for SDK?	Java, Android, .NET, Go	Java
Data retention period	7 days	Longer (Users configure retention periods)
Skill level required	Basic	Advanced
Customization	Yes	Yes
Performance limitations	Write synchronously to 3 machines at a time	Fewer limitations
Store	Dynamo db	Zookeeper
Price	Based on resources used, with no upfront costs	Free (open-source) but consider hardware/s costs
Support	Developer center, tutorials, and more	Tutorials, meetups, videos, and more

Kinesis and Kafka have several unique features.

- Kinesis supports Java, Android, .NET, and Go; Kafka only supports Java.
- Kinesis lets users write synchronously to three machines or data centers; Kafka users have more configurations.
- Kinesis stores data in shards; Kafka stores data in partitions

Regarding performance, Kinesis reaches a throughput of thousands of messages every second. Kafka, however, reaches a throughput of around 30,000 messages every second, making it the clear winner.

There are limitations to both platforms. Kinesis has a 7-day data retention period. Kafka takes a lot of effort to set-up and run. (Kafka requires distributed engineering and cluster management experience.)



5

4) What is your recommendation

ANS 4) For Netflix, Amazon Kinesis is doing a good enough job. Apache Kafka would give a slightly better performance but in the case of Netflix, scalability and reliability is more important and for that Amazon is the better choice. Plus using Kafka would require them to hire more people, this would increase their costs and also Kafka requires higher maintenance time which would ensure Netflix losing more money that what they would have had to spend on using Kinesis.

5) **What is the conclusion?**

ANS 5) Both Apache Kafka and AWS Kinesis Data Streams are good choices for real-time data streaming platforms. If you need to keep messages for more than 7 days with no limitation on message size per blob, Apache Kafka should be your choice. However, Apache Kafka requires extra effort to set up, manage, and support. If your organization lacks Apache Kafka experts and/or human support, then choosing a fully-managed AWS Kinesis service will let you focus on the development. AWS Kinesis is catching up in terms of overall performance regarding throughput and events processing.