

Amazon Kinesis Data Streams

Kinesis Data Streams face parte din platforma de data streaming Kinesis, împreună cu Kinesis Data Firehose, Kinesis Video Streams și Kinesis Data Analytics.

Utilizare

Kinesis Data Streams se folosește pentru procesarea și agregarea unui volum continuu și rapid de date. Tipurile de date pot varia, de la log-urile unor aplicații și rețele sociale la feed-uri de burse sau date despre activitățile unui utilizator pe o pagină web. Având în vedere faptul că timpul de răspuns pentru consumarea și procesarea fluxului de date se realizează în timp real, putem cataloga procesarea ca fiind una lightweight.

Scenarii clasice de utilizare

- **Consumul accelerat și procesarea unui feed-urilor de loguri și date** – se pot folosi mai mulți producători ce adaugă direct datele într-un flux. Datele provenite de la sisteme de tip push sau log-urile aplicațiilor pot fi procesate în câteva secunde. Acest lucru împiedică pierderea datelor de log, în cazul în care interfața sau server-ul nu mai funcționează. Kinesis Data Streams oferă consumul accelerat al feed-ului de date, nefiind nevoie de o batch-uire a datelor pe server înainte ca acestea să fie trimise în flux.
- **Măsurători și rapoarte în timp real** – datele colectate în Kinesis Data Streams pot fi folosite pentru o analiză simplă de date și raportare, în timp real. Spre exemplu, o aplicație destinată procesării de date, poate crea rapoartele necesare în timp real, la momentul primirii datelor, nefiind necesară așteptarea unor batch-uri de date.
- **Analize de date în timp real** – sunt combinate **puterea** procesării paralele cu **valoarea** datelor real-time. Spre exemplu, procesarea în timp real a activității unui utilizator pe un website și apoi, analiza interesului utilizatorilor folosind diferite aplicații Kinesis Data Streams, rulate în paralel.

Beneficii

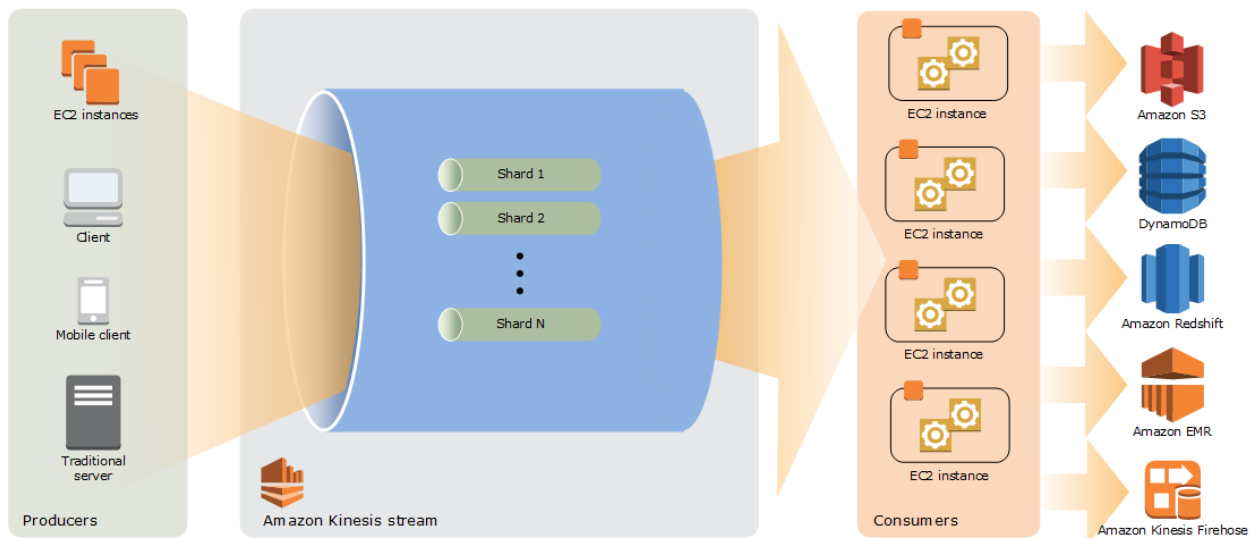
Deși Kinesis Data Streams poate fi folosit pentru a rezolva diverse probleme, o utilizare comună este cea de a agrega date în timp real, urmând ca rezultatul să fie încărcat într-un data warehouse sau într-un cluster de tip map-reduce.

Datele sunt introduse direct în Kinesis data streams, ceea ce asigură durabilitatea și elasticitatea. Întârzierea dintre momentul în care o înregistrare este introdusă în fluxul de date și momentul când aceasta poate fi extrasă din flux este, de obicei, mai mică de o secundă. Cu alte cuvinte, o aplicație Kinesis Data Streams poate începe consumarea datelor dintr-un flux, imediat după ce acestea sunt adăugate.

Elasticitatea framework-ului îți oferă posibilitatea de a scala fluxul de date cât este necesar, astfel încât datele să nu fie pierdute înaintea momentului expirării acestora.

Mai multe aplicații Kinesis Data Streams pot consuma date dintr-un flux, astfel încât mai multe acțiuni, precum arhivarea și procesarea, pot avea loc concurrent și independent. Spre exemplu, două aplicații pot citi date din același flux. Prima aplicație poate realiza agregări și actualizări pe un tabel Amazon DynamoDB iar cealaltă, să comprime și să arhiveze date în Amazon S3.

Arhitectură



În această diagramă, este ilustrată arhitectura high-level a Kinesis Data Streams:

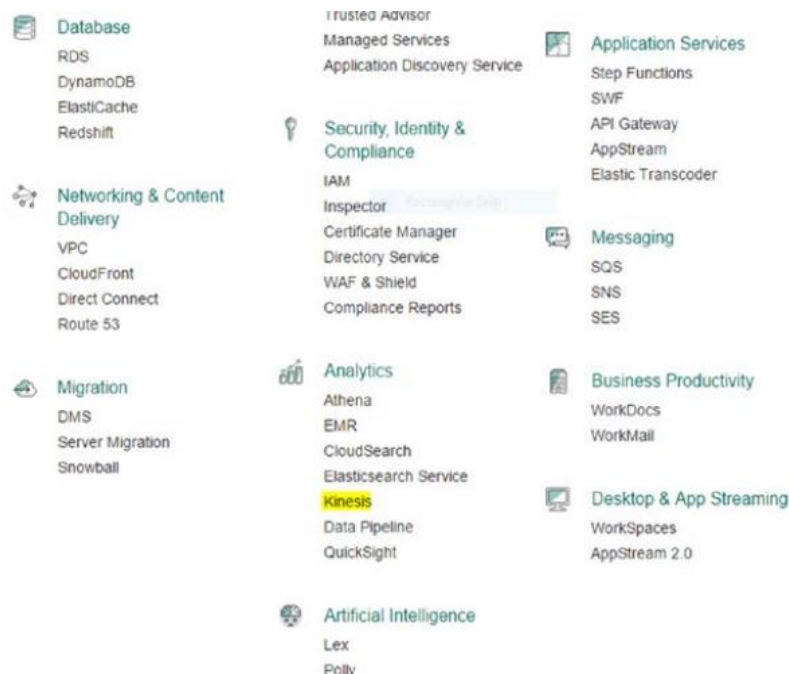
1. *Producer* – adaugă date în Amazon Kinesis Data Streams (ex: un server web ce trimite log-uri în cadrul unui flux)
2. *Consumer* – preia înregistrări din flux și le procesează, cunoscuți și sub numele de Aplicații Kinesis Data Streams
3. *Kinesis Data Stream* – un set de *shards*. Fiecare shard deține o secvență de înregistrări. Fiecare înregistrare are un *sequence number* asignat de către framework.
4. *Shard* – o secvență de înregistrări din cadrul unui flux, identificată în mod unic. Un flux este compus din unul sau mai multe shard-uri, fiecare furnizând o unitate fixă de capacitate. Fiecare shard poate suporta până la 5 tranzacții pe secundă, pentru citire, la o capacitate maximă de 2MB/s și până la 1000 de înregistrări pe secundă, la scriere, la o capacitatea maximă de 1MB/s. Capacitatea unui flux este stabilită în funcție de numărul de shard-uri alocate.
5. *Partition Key* – este folosit pentru a grupa datele în shard-uri, în cadrul unui flux.
6. *Sequence Number* – fiecare înregistrare deține un număr de secvență, unic pe partiție, în cadrul shard-ului său.

Kinesis Client Library – este compilat în cadrul aplicației create, pentru a oferi un consum de date fault-tolerant din cadrul fluxului.

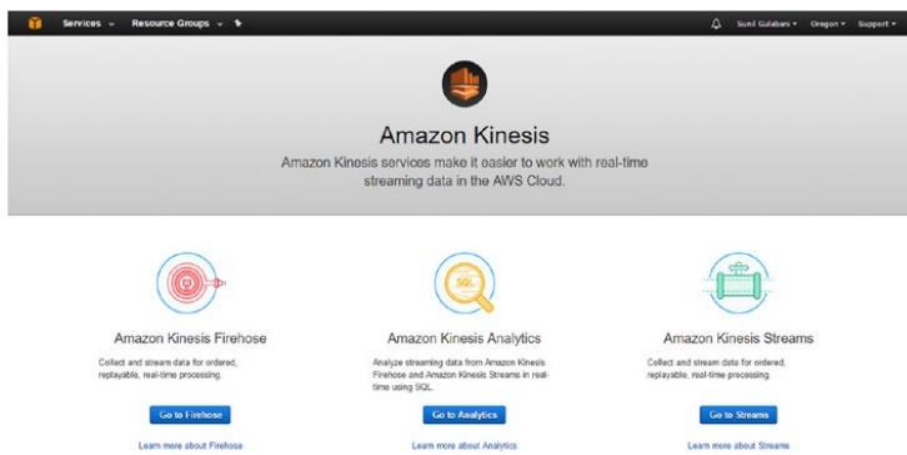
În continuare, voi prezenta un exemplu de utilizare al AWS Management Console pentru a gestiona o aplicație Kinesis Streams.

1. Crearea Fluxurilor

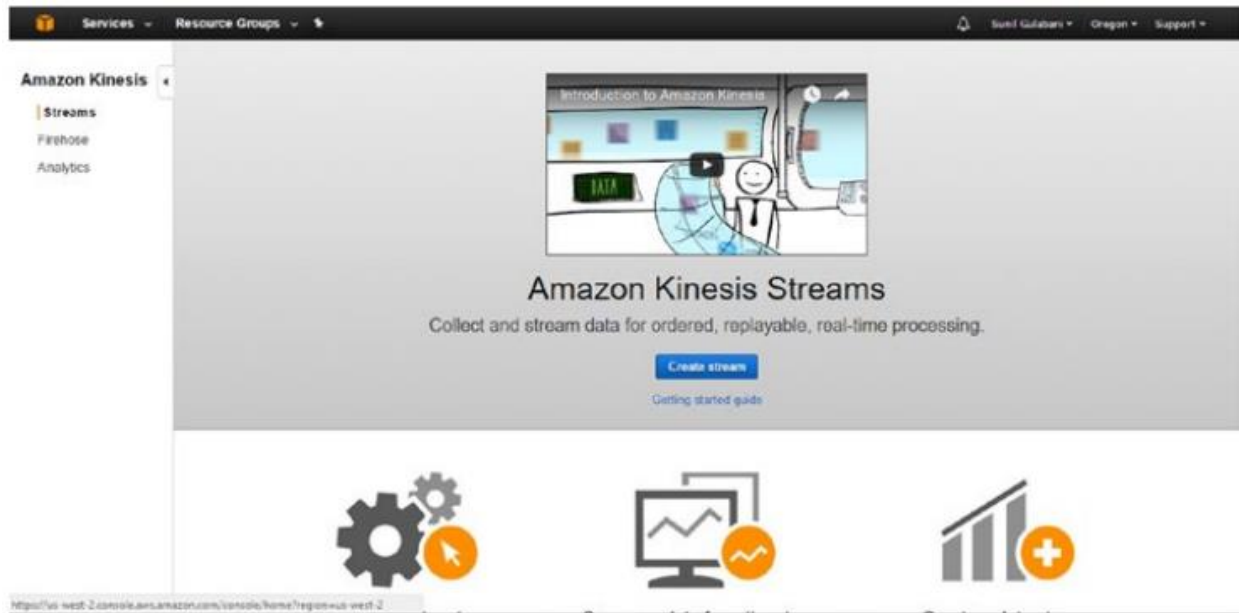
- Vă logați în AWS Console și apăsați pe Kinesis, din domeniul de Analytics



- Alegeți opțiunea Kinesis Streams



- Aici puteți observa Stream-urile deja create. Vom crea unul nou.



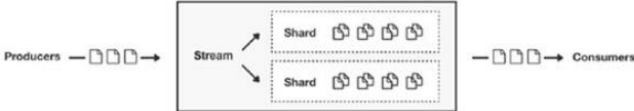
- Vom da nume Stream-ului și vom preciza numărul de shard-uri

Create stream ?

Stream name*

Shards

A shard is a unit of throughput capacity. Each shard ingests up to 1MB/sec and 1000 records/sec, and emits up to 2MB/sec. To accommodate for higher or lower throughput, the number of shards can be modified after the stream is created using the API. [Learn more](#)

Producers →  Consumers

► Estimate the number of shards you'll need

Number of shards*

The default shard limit for an account in this region is 50. [How can I increase this limit?](#)

Total stream capacity Values are calculated based on the number of shards entered above.

Write	<input type="text" value="5"/>	MB per second
	<input type="text" value="5000"/>	Records per second
Read	<input type="text" value="10"/>	MB per second

* Required Cancel Create stream

- Odată ce toate detaliile au fost completate, apăsați butonul Create stream. Veți putea observa că noul Stream a fost creat.

✓ Stream **Chapter4KinesisStream** has been successfully created.
[View details](#)

Create stream Actions

Filter or search by stream name << < Viewing 1 - 1 of 1 items > >>

	Stream name	Number of shards	Status
<input type="checkbox"/>	Chapter4KinesisStream	5	ACTIVE

<< < Viewing 1 - 1 of 1 items > >>

2. Configurarea Perioadei de retenție a datelor – apăsați butonul Edit, modificați valoarea și apăsați Save. Modificările vor deveni active după 30 de secunde.

Data retention period Cancel Save

The data retention period can be increased from 24 hours up to 168 hours for an additional cost. See [Kinesis Streams pricing](#).

Data retention period 24 hours ⓘ
Specify a data retention period between 24 and 168 hours

Data retention period Edit

The data retention period can be increased from 24 hours up to 168 hours for an additional cost. See [Kinesis Streams pricing](#).

✓ Successfully updated data retention period.

Data retention period 30 hours ⓘ

3. Configurarea Tag-urilor – nume și valori atribuite serviciilor AWS. Sunt folosite pentru a agrega resurse AWS și pentru a determina costurile, grupând tag-uri similare.

- Apăsați pe tab-ul de Tags
- Adăugați o cheie și o valoare și apăsați butonul de Save
- De asemenea, le puteți modifica pe cele existente, apăsând iconița de edit, din dreptul fiecărui tag.

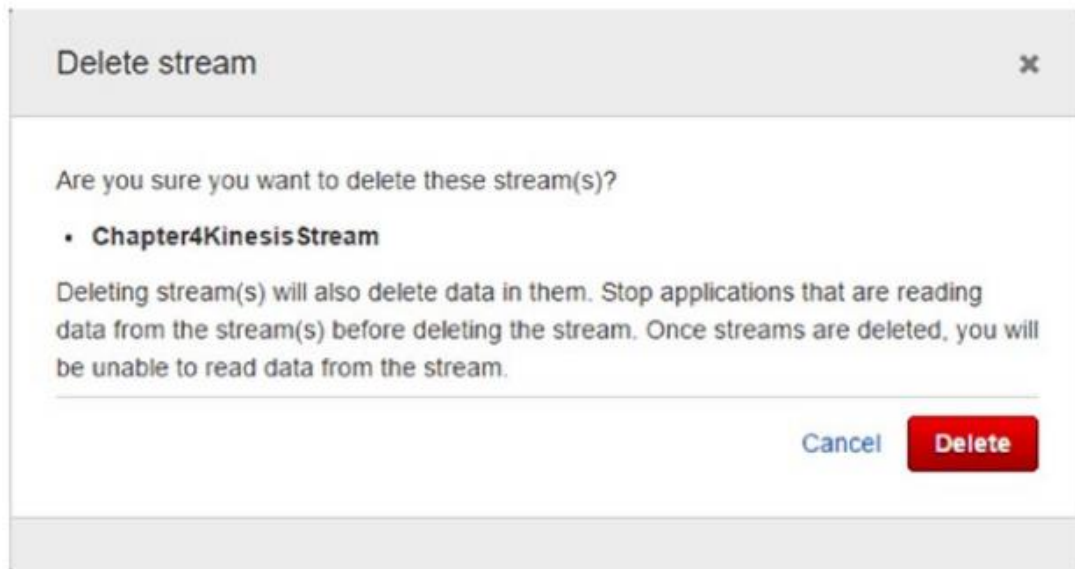
Details Monitoring **Tags**

Tags are used to help organize and identify your streams. A tag consists of a case-sensitive key-value pair. For example, you can define a tag with key=Department and with value=Marketing.

Key	Value	
Chapter	4	ⓘ
Environment	Production	ⓘ
Add a new key	Add a new value	

Save Cancel

4. **Ștergerea unui Stream** – se selectează Streamul ce se dorește a fi șters și se apasă butonul Delete



Referințe Bibliografice

- <https://docs.aws.amazon.com/streams/latest/dev/key-concepts.html>
- Practical Amazon EC2, SQS, Kinesis and S3, Sunil Gulabani, 2017