Amazon Kinesis

Real-Time Streaming Data Processing

Content Prepared By: Chandra Lingam, Cotton Cola Designs LLC

amazon web services

Kinesis Platform

- Continuous capture, store, analyze
- Fully Managed
- Scales automatically – TBs per hour
- Capabilities
  - Kinesis Streams
  - Kinesis Firehose
  - Kinesis Analytics

Figure: Pipeline - Clickstream Analytics

Streaming Data Examples

- Generated Continuously and typically processed in-order, over a sliding time window
- Batch versus Stream Processing
- Examples:
  - Location based promotion
  - Log and Event Data
  - Service Metering
  - Tracking movement of goods
  - Industrial Equipment monitoring
  - Website Clicks/Customer interaction
  - Social Media Trending topics

Kinesis Streams

Kinesis Streams Concepts

- Stream is divided into Shards
- Data is stored in Shards
- One Shard provides: 1 MB/s WRITE, 2 MB/s READ, and up to 1,000 PUT operations
- Add or remove Shards dynamically depending on need

Kinesis Streams Concepts

- Data Record – unit of data stored in streams
  - Sequence Number
  - Partition Key
  - Data Blob (stored in Base64 encoding)
- Max Size per Data Record 1MB

Kinesis Streams Concepts

- Partition Key is used to route Data Record to different Shards.
- Partition Key is specified by the producer

Kinesis Streams Concepts

- Sequence Number is an unique identifier for every data record
- Assigned by Kinesis Streams
- Sequence number for a partition key generally increase over time

Kinesis Streams Pricing

Kinesis Streams Pricing is based on:

- Shard Hours
- PUT Payload Size in Chunks of 25KB
- Data Retention - Default is 1 day up to a max of 7 days
- GET calls are free
- Data Transfer is free
- NO FREE TIER! *All Demos in this lecture incur minimal charges. Make sure you delete the resources after you are done with the demo!*

Kinesis Streams Demo

Demo – Create Stream

Describe Streams – list shards that are available

Put Records – note the Shard ID where data record is stored

Get Shard Iterator – Specify where you want to read from

Get Record – in Base64 encoding. (https://www.base64decode.org/)

Kinesis Split Shard Demo

- Split Shard
- Existing Shard is split into two child shards. We need to specify the hash key range for the children.
- Parent shard would still have data. Any new data is distributed among the two child shards
- You would have to finish consuming parent shard first and then start consuming data from both children
- Use BigInt Calculator to find the new hash values

Libraries

- Kinesis Producer Library
- Kinesis Client Library (KCL)
  - Consume data from streams
  - Uses DynamoDB to checkpoint processed records, worker to shard mapping
  - Handles resharding
  - Handles autoscaling (new EC2 instances)
  - Automatically keeps track of shards and shard iterator

Kinesis Firehose

Kinesis Firehose

Continuously Load Streaming Data into:

- S3
- Kinesis Analytics
- Redshift
- Elasticsearch Service

Use Existing dashboards and business intelligence tools

Kinesis Firehose

Before loading to destination, Firehose can:

- Batch Data
- Transform Data (with AWS Lambda functions)
- Compress Data
- Encrypt Data

Firehose Pricing

Firehose Pricing is based on

- Data Ingested into Firehose
  - Number of data records
  - Size of the data record rounded up to nearest 5KB. (i.e 38 KB payload is rounded to 40KB)
- Data is retained in firehose for 24 hours

Demo - Firehose

- Create Firehose Destination Stream
- Configure to send the events to S3 Bucket
- Put one record
- Skeleton JSON Payload for Firehouse CLI calls
- Put a single json record
- Put a batch of json records
- Verify S3 bucket

Kinesis Analytics

Kinesis Analytics

- Analyze Streaming data using SQL!
- Supports Querying Kinesis Streams and Firehose
- Write the results back to another Firehose or Kinesis Stream destination
- Usage:
    - Time Series Analytics – Trend over period of time
    - Feed Real Time Dashboards
    - Alarms and Notifications based on threshold
    - SQL Based Anomaly Detection, Top-K, Distinct Items and so forth

Kinesis Analytics

- Map Incoming Streaming Data to a Schema
  - Kinesis Analytics understands CSV, JSON, TSV data payload and automatically creates a baseline schema
  - If data is unstructured, you can define your own schema
- Kinesis Service applies the schema to streaming data and presents the data like a SQL table
- Write SQL Queries against the table
- Store SQL Query Results in Results Stream
- Optional: Persist Query Results to Kinesis Streams, Kinesis Firehose

Kinesis Analytics Concepts

- Application - Refers to Kinesis Analytics Application that continuously process streaming data
- In-Application Stream - Your application specific “table” that you can query using SQL. Continuously added/updated
- Pump - A continuous query that inserts data from one in-application stream to another in-application stream

Kinesis Analytics

Two timestamps provided automatically for every data record for Time series analysis

ROWTIME = Timestamp when kinesis analytics inserts a row in the first in-application stream after reading from streaming source. This is then maintained throughout your application

ApproximateArrivalTimeStamp = Timestamp when event was added to the streaming source. Server-side time.

Kinesis Analytics Pricing

Kinesis Analytics Pricing is based on

- KPUs (Kinesis Processing Unit)
  - Single KPU consists of 4GB RAM and 1 vCPU
  - Automatically allocates required KPUs to complete your analysis

Kinesis Analytics Demo

Create Kinesis Stream Data Source

Populate Data

Create Kinesis Analytics Application

Configure Input/Schema

Define SQL

Store results in Analytics Application Stream