# Lab 2 - Streaming analytics with Amazon Managed Service for Apache Flink

Objective:

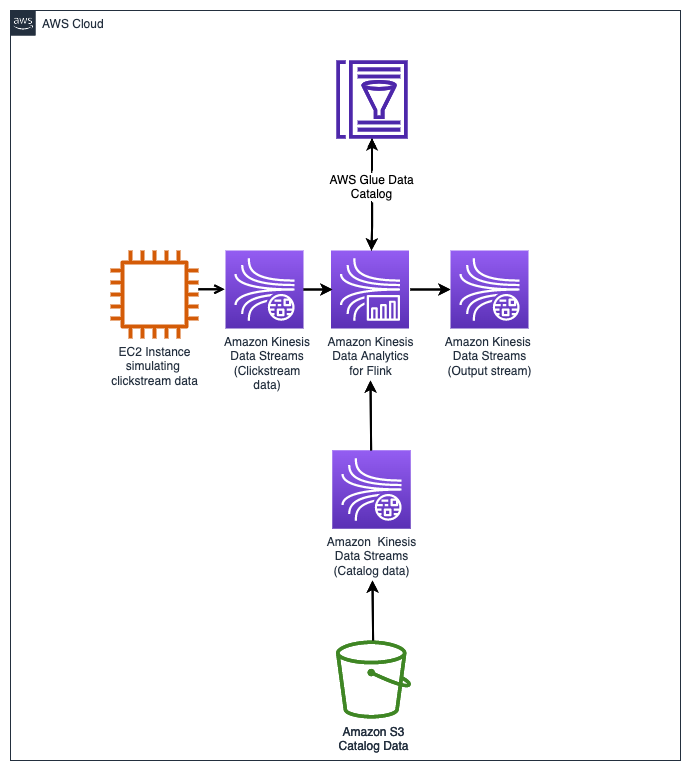
1. Build a real-time streaming analytics pipeline in Managed Apache Flink Studio using Apache Flink to ingest, enrich, and analyze the clickstream data.
2. Perform interactive data analytics and visualize using Apache Zeppelin notebooks with Managed Apache Flink Studio.

Imagine you have a river of data flowing in, and you want to collect and sort different things from it without stopping the flow. **Flink can do that.**

**It can handle data as it comes in, making it possible to get insights and take actions right away.**

**Process:**

1. Collects Data: Gather clickstream data from the website using Amazon Kinesis.
2. Processes Data: Use Amazon Managed Service for Apache Flink to process this data quickly.
3. Enriches Data: Combine the clickstream data with product information stored in Amazon S3 to add catalog details.
4. Analyze Data



Note - We saw how to send data from EC2 to S3 in the last lab.

Now, here:

Kinesis Data Stream (1) collects clicking data from EC2 and sends to Flink for analysis.

Also, KDS (2) transports catalog data from S3 and sends it to Flink for data enrichment.

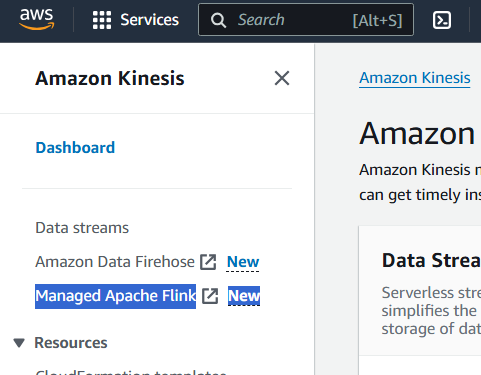
**The Flink application running in Amazon Kinesis Data Analytics combines the clickstream data with the catalog data.**

Output stream KDS(3) sends data for analysis and visualization.

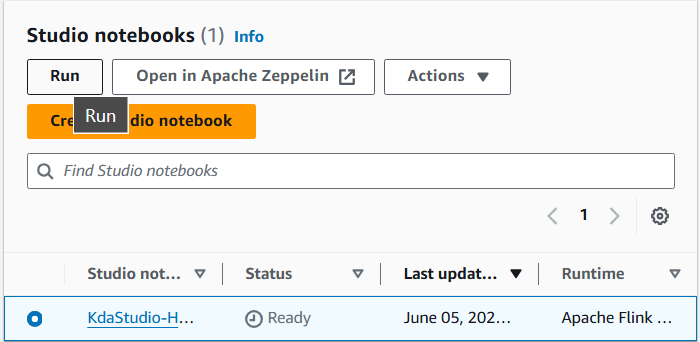
{AWS Glue Data Catalog is used to keep track of metadata.}

**Task 1: Setting up Zeppelin notebook environment**

1.1 Open Kinesis from console and click on Apache Flink



1.2 Open studio notebook option from left pane, and run this



1.3 Download the zeppelin file from the link



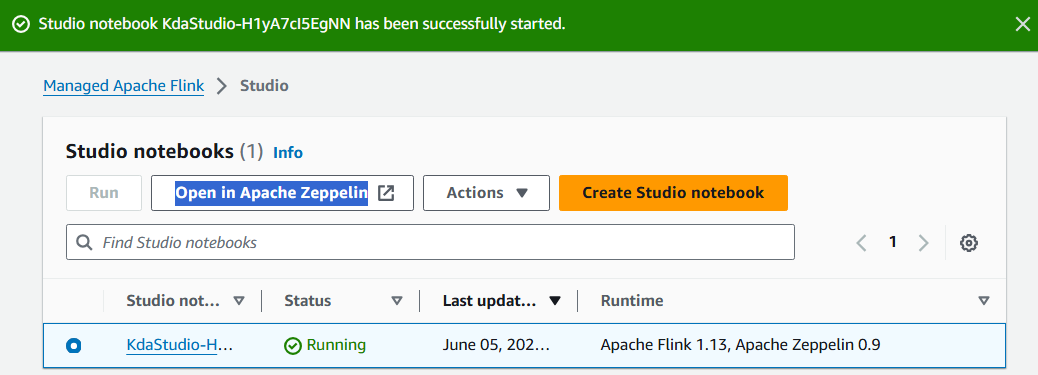
**Till then :**

**Task 2: Connect to the Amazon EC2 producer and start the clickstream generator**

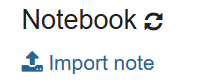
Skipping, same steps as lab 1

**Task 3: Import the Zeppelin notebook**

3.1 Click on open in zeppelin



3.2 Inside zeppelin UI, import the downloaded notebook



**Task 4: Analytics development in Managed Apache Flink Studio with the Zeppelin notebook**

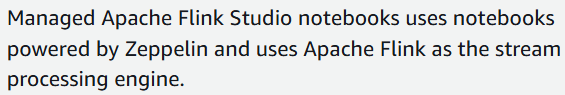
**The Managed Apache Flink application, which will be set up in the Flink Studio notebook, will consume the data from the Kinesis Data Stream, whose name was given in .py code.**

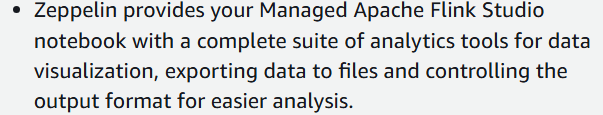
**Flink Studio Notebook:**

**This is where we "talk" to Managed Apache Flink. It's like a digital workspace where we can ask questions about the clickstream data and tell Flink what to do with it.**

**Interactively Query Data Streams:**

**Instead of just looking at the data all at once, we can ask Flink questions about the data in real-time**

****

****

4.0 Open the notebook

Task 4.1: Ingestion - from two sources:

1. From the Kinesis data stream with clickstream data (produced by the clickstream generator)
2. From the Kinesis data stream with catalog data in an Amazon Simple Storage Service (AmazonS3) bucket

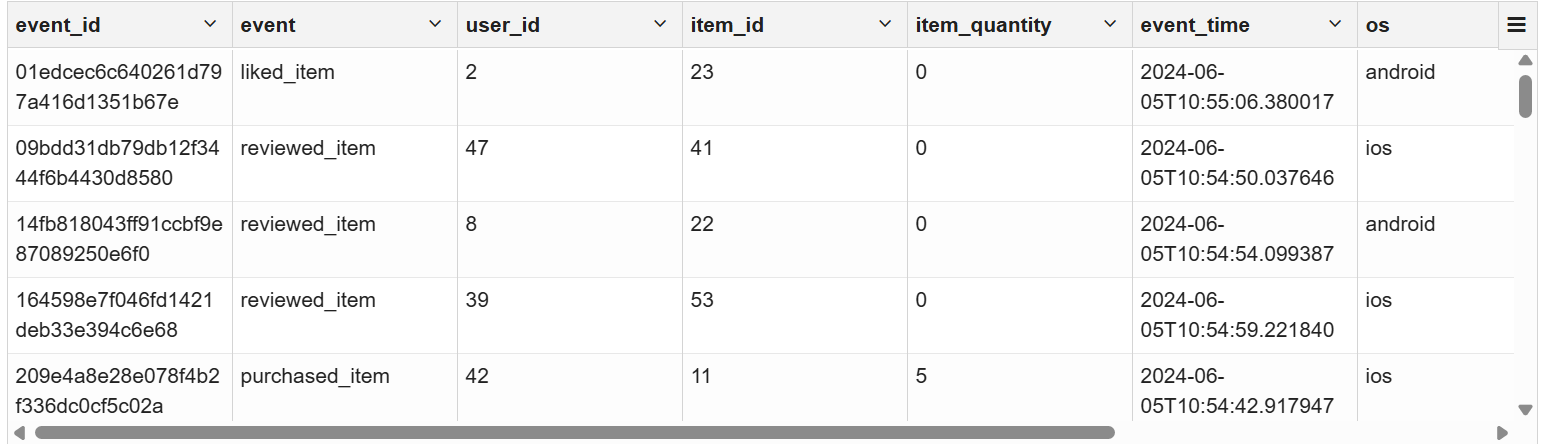
We create the in-memory table, clickstream\_events, using Kinesis connector





Here, we set stream value as the KDS (1) , which collected EC2 data.

Now: view the simulated clickstream data (skipped the code)

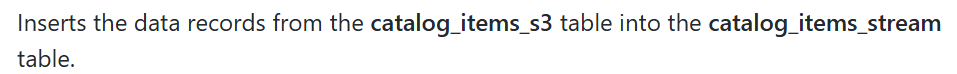


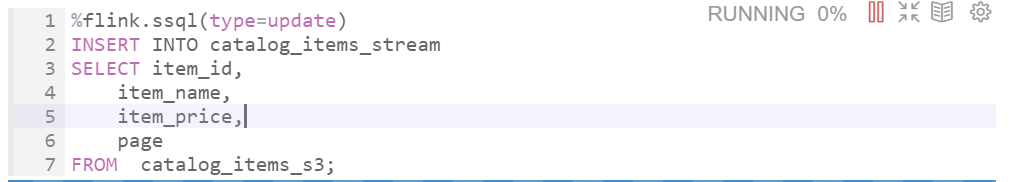
Now, ingest catalog data from S3:

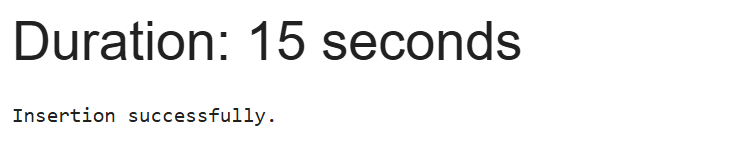
Steps: create table to read S3 data -> create table for KDS(2) -> insert S3 data into KDS(2)

In the above 3 steps:

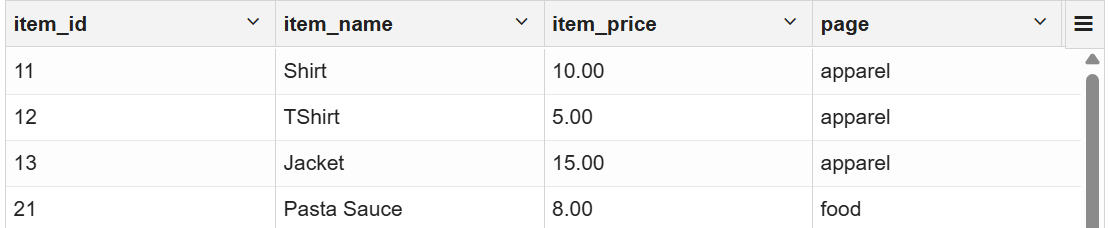
We directly go to last step of:





Output- 

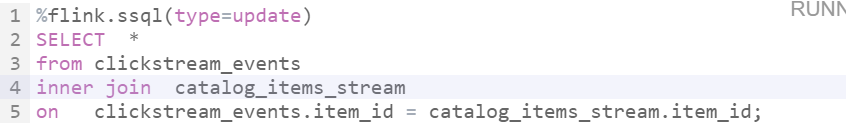
View the data from “catalog\_items\_stream”



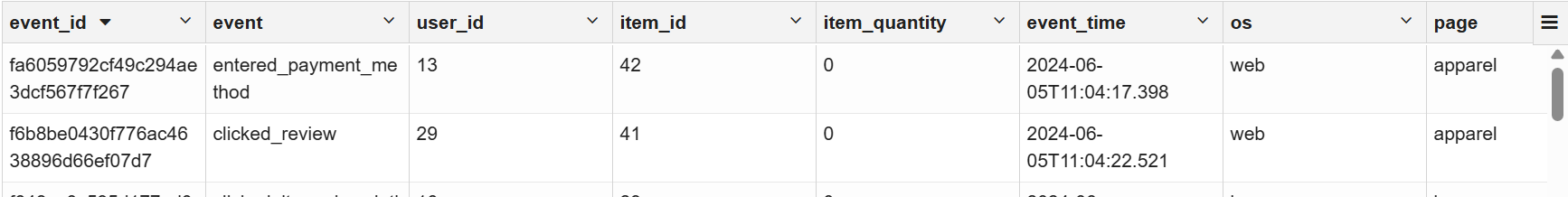
**Task 4.2: Data enrichment**

enrich the streaming clickstream data with the catalog data available in an S3 bucket.

Done using JOIN

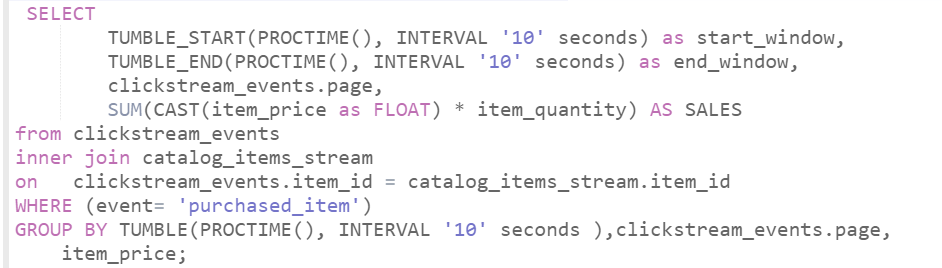


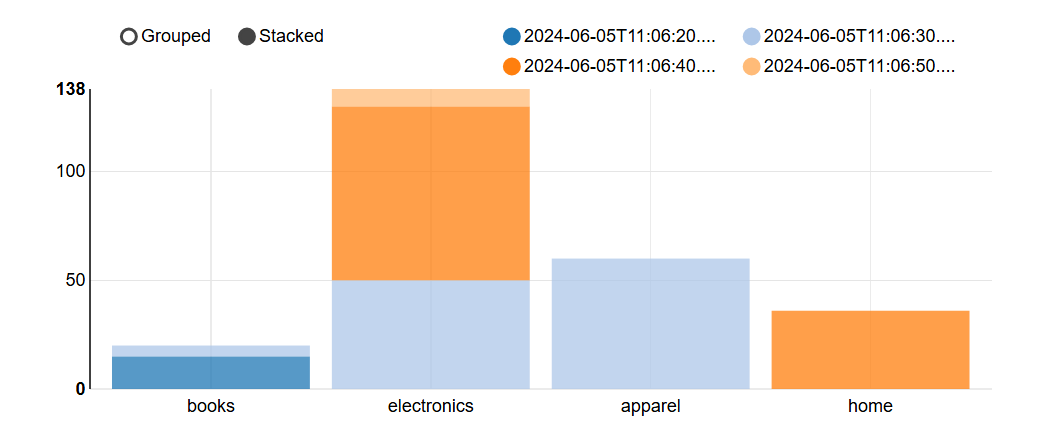
Output-



**Task 4.3: Analysis and visualization**

Task = analyze the data to determine the Sales per category in a given time interval.





**Task 4.4: Output to Kinesis data stream**

(From lab:)

you write the output of the analysis to a Kinesis data stream which will be used for further downstream processing.

Create table with output stream value

