AdBench: A Complete Benchmark for Modern Data Pipelines

MILIND BHANDARKAR (milind@ampool.io)

FOUNDER & CEO, AMPOOL, INC.

Agenda

- Rationale
  - New Use Cases
  - New Architectures
- Benchmark Scenario
- AdBench Description
- Prototype Implementation
- Demo
- Q & A

About Ampool

- Stealth mode startup, founded 2015
- Based in Santa Clara, CA & Pune, India
- Building next generation data infrastructure
- Targeting modern data pipeline workloads
- Utilizing modern commodity hardware, e.g. Storage Class Memory, low-latency network & RDMA
- We are hiring!

New Use Cases

- Analytics on the Internet of Things
  - Edge (micro-batch) & Cloud (large batch) Computing
  - Event-at-a-time CEP
  - Ad-Hoc Real-Time Queries
- AI & Deep Learning
  - Train/Re-Train Models
  - Incremental Updates to Models
  - Serving Models
- Conversational User Interfaces
  - Interpret within Context (State Transitions within Session)
  - Cross-context correlation (Path Analysis)
  - Integration with transactions

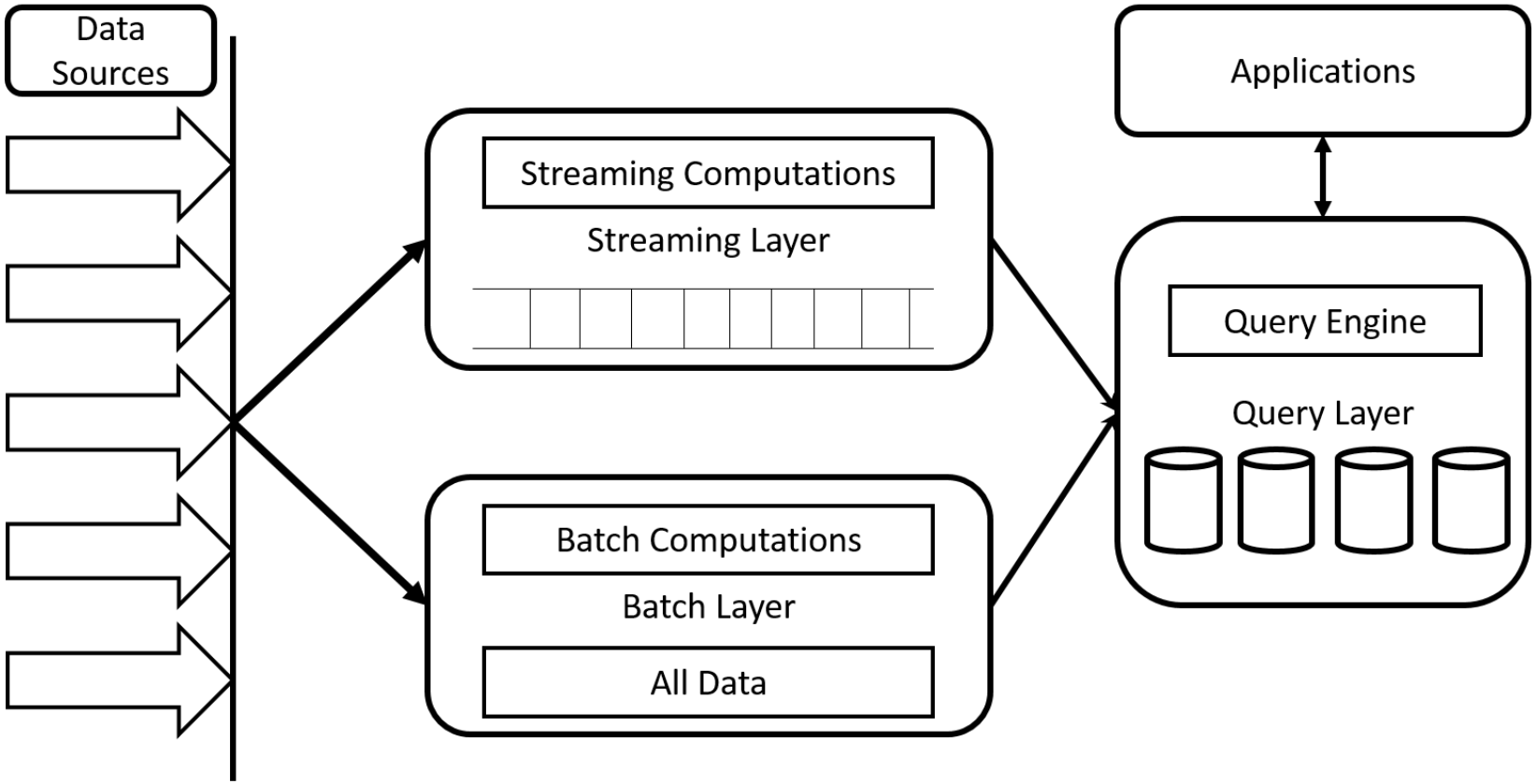
New Technologies

- Scale-Out On-Demand Compute Infrastructure
  - Public & Private Clouds
- Fine-Grained Virtualization & Microservices
  - Containerization & Orchestration

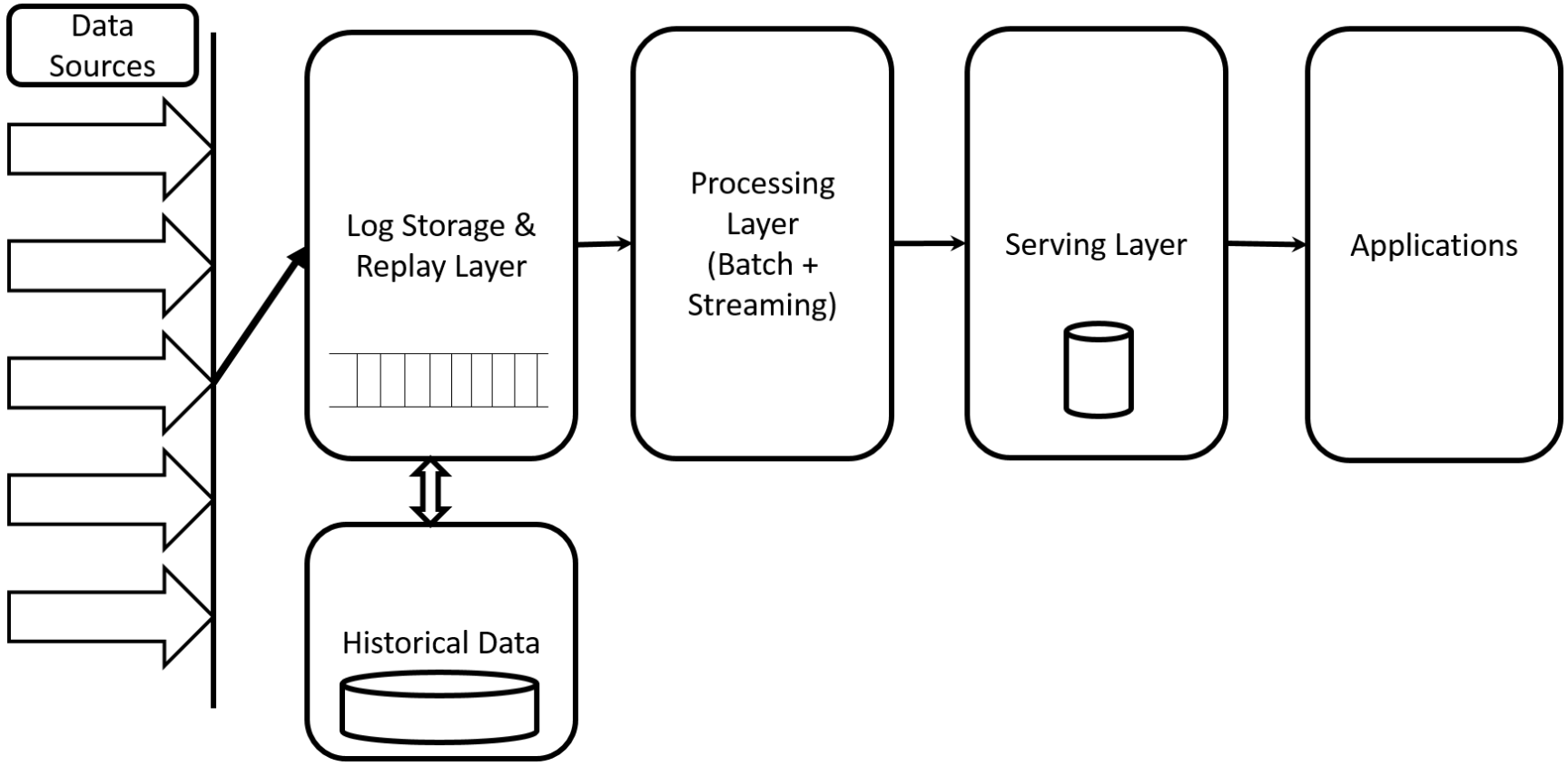
- Huge (and rapidly growing) gap between memory and I/O bandwidths
- Rapidly increasing Network Bandwidth (~1000x in ~15 years)

- Plummeting costs of Solid State Storage (Comparable to HDD by 2019)
- NVDIMMs supported by major OSs, 10x density, 1/5th $/GB compared to DRAM
- Emergence of Storage Class Memory
  - 3D XPoint, PCM, MRAM, Memristor etc.

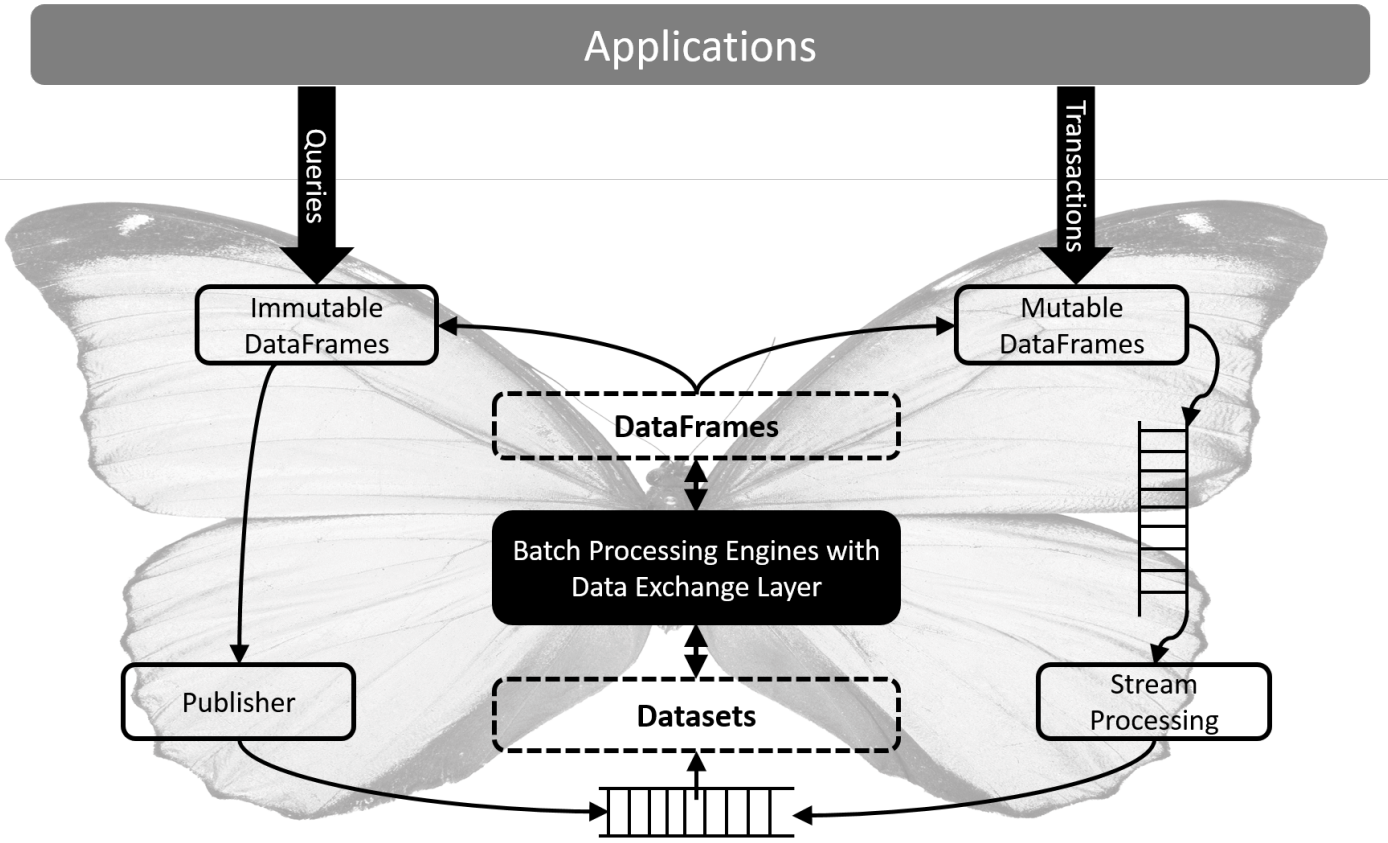
Lambda Architecture



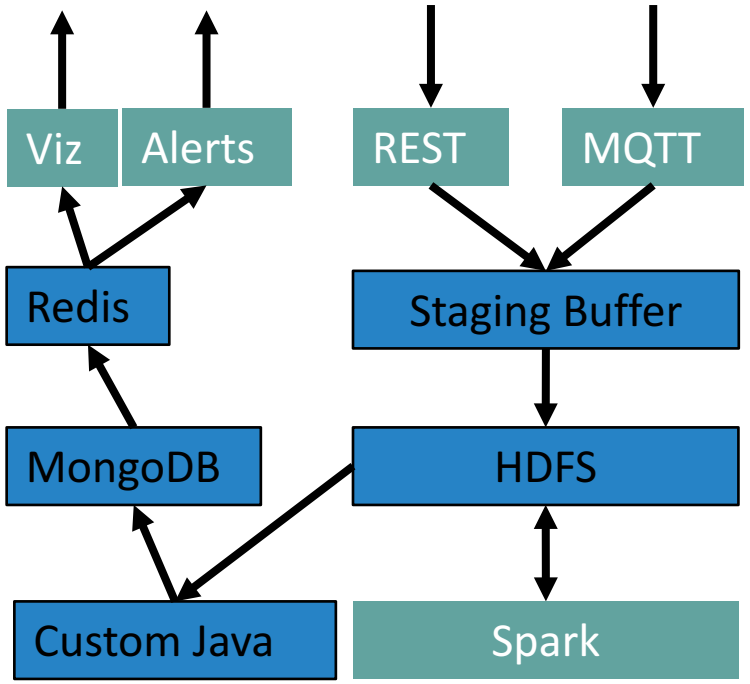
Kappa Architecture



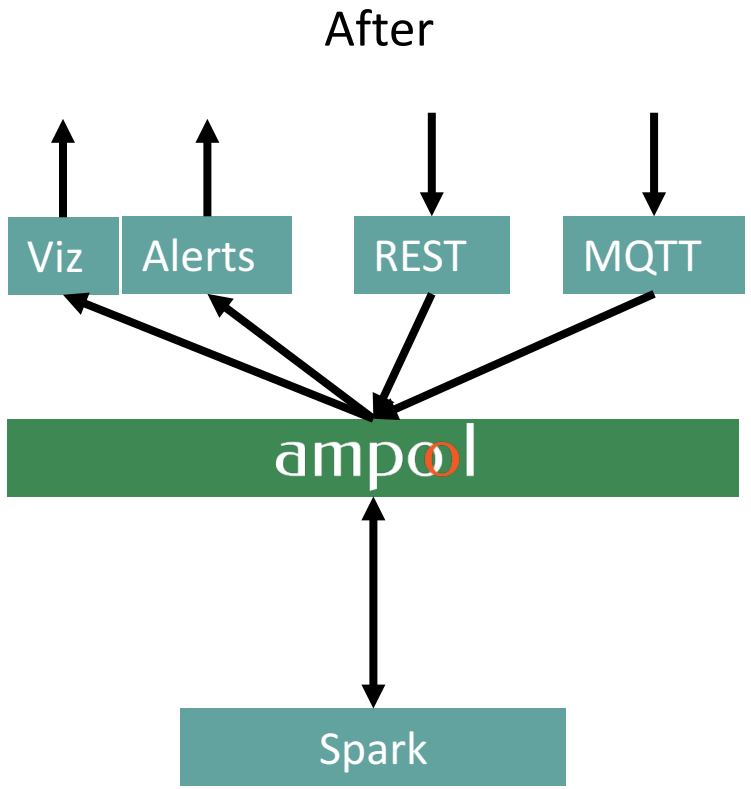
Butterfly Architecture



IoT Analytics Example: Lambda



IoT Analytics Example: Butterfly



Benchmark Scenario

- Acme is an Ad-Tech Company
- Three entities
  - Consumers (Users)
  - Advertisers
  - Content Publishers
- Goals
  - Deliver personalized content to Consumers
  - Maximize Content Relevance for Consumers
  - Maximize Ad Relevance for Consumers & Content

Acme Corp in Numbers

- 100 Million registered users, with 50 Million daily unique users
- 100,000 advertisements across 10,000 advertisements campaigns
- 10 Million pieces of content (News, Photos, Audio, Video)
- 50,000 keywords in 50 topics & 500 subtopics as user interests, content topics, and for ad targeting

Datasets - Tables

User Profiles					
UserID	Age	Sex	Location	User-Since	Interests
UUID	0..255	M/F/Unknown	Top3(LatLong)	DateTime	List(topic:subtopic:kw)

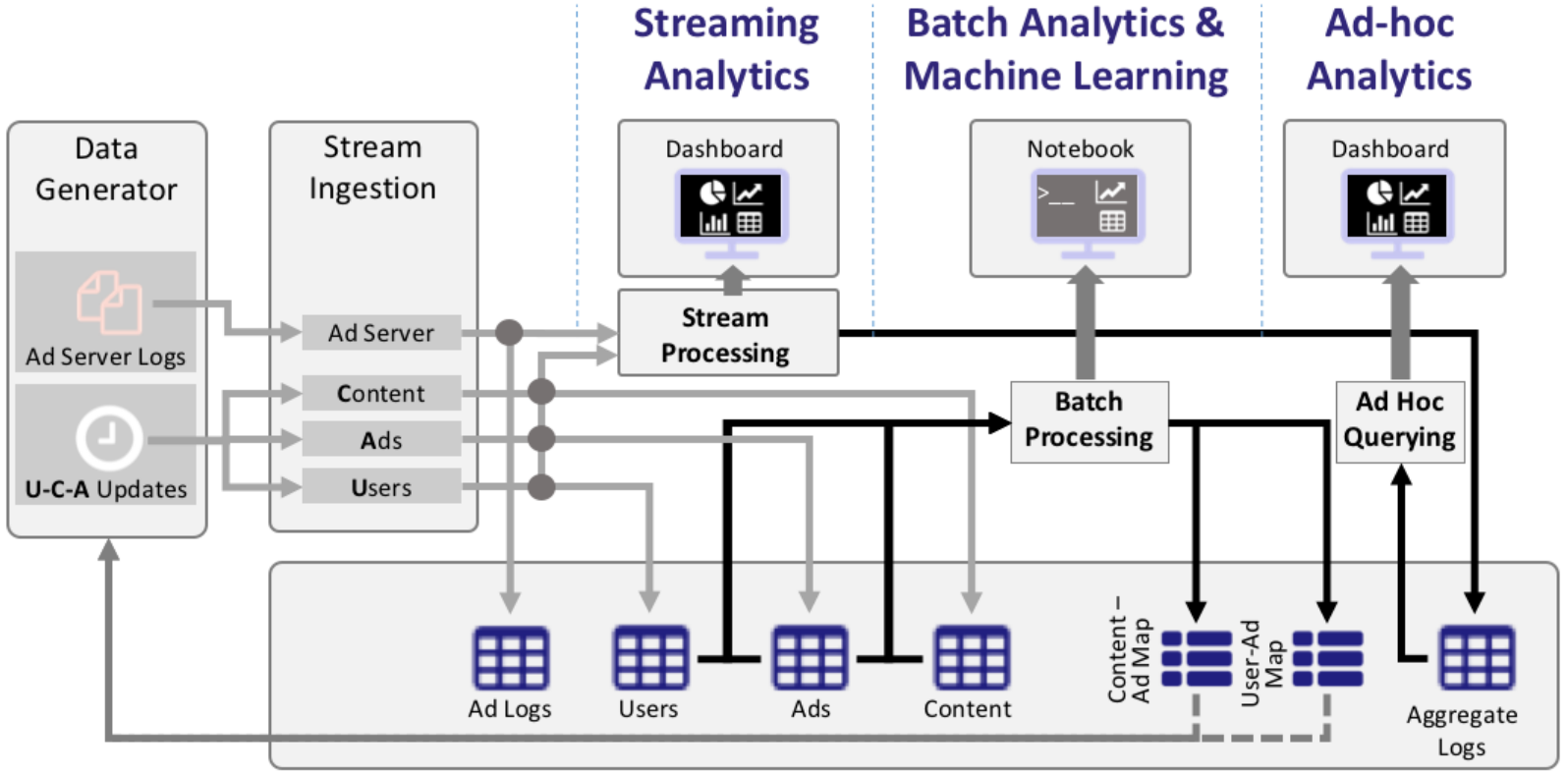
Advertisements								
AdID	Campaign ID	Customer ID	AdType	Platform	KW	PPC	PPM	PPB
UUID	UUID	UUID	Category	Category	List(topic:subtopic:kw)	$ Float	$ Float	$ Float

Content		
ContentID	Content-Type	Keywords
UUID	0..255	List(topic:subtopic:kw)

Dataset - Streaming

Timestamp	DateTime
IP Address	IPv4 / IPv6
User ID	UUID
Ad ID	UUID
Content ID	UUID
Ad Type	{Banner|Modal|Search|Video}
Ad Platform	{Web|Mobile}
EventType	{View|Click|Conversion}

Computations & Dataflow



Computations: Streaming

- Based on Yahoo! Streaming Analytics Benchmark
- Parse the event record
- Extract Timestamp, AdID, EventType and AdType
- Look up CampaignID from AdID
- Windowed aggregation of event types for each AdID, and CampaignID
- Store these aggregates in an aggregate dataset
- Prepare these aggregations for a streaming visualization dashboard for a CampaignID, and all Ads in that Campaign
- Output:
  - (AdID, Window, nViews, nClicks, nCon, ∑ PPV, ∑ PPC, ∑ PPCon)
  - (CmpgnID, Window, nViews, nClicks, nCon, ∑ PPV, ∑ PPC, ∑ PPCon)

Computation: Update User, Ad, Campaign Profiles

- Ingest a {user|campaign|ad} {update|insert} event from message queue
- Parse the event to determine which dataset is to be updated.
- Update respective dataset.
- Keep track of total number of updates for each dataset.
- When 1% of the records are either new or updated, launch the batch computation stage and reset update counters.

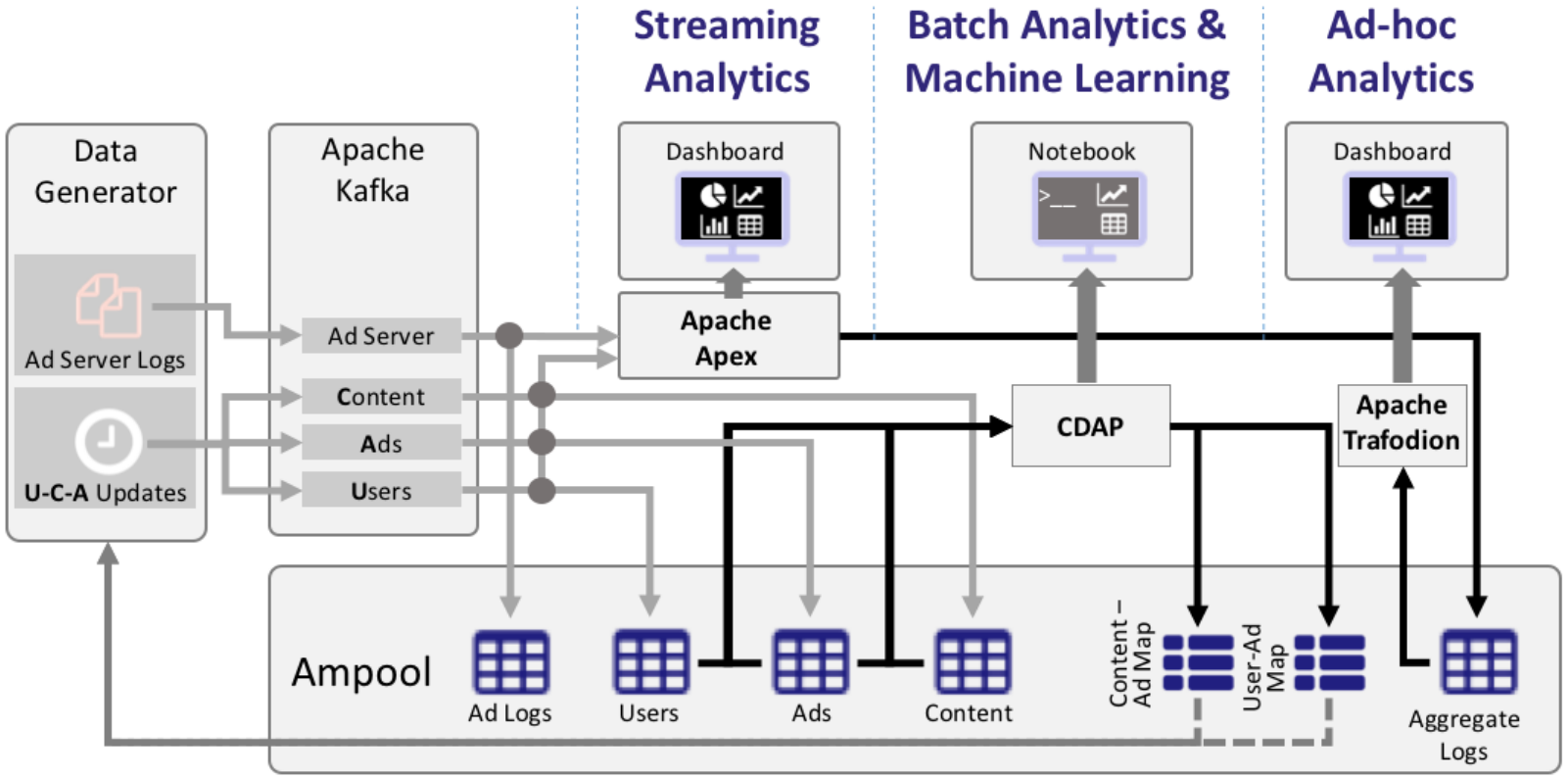
Batch Computation: Ad Relevance

- Based on User Interests, Ad Keywords, and Content Keywords, determine the top 3 Ads to be targeted for each user and each content
- Weighted Keyword Match
  - Exact Keyword = 1.0
  - Exact Subtopic = 0.1
  - Exact Topic = 0.01
- Relevance (Ad, User) = Cosine (Ad Keywords, User Interests)
- Relevance (Ad, Content) = Cosine (Ad Keywords, Content Keywords)
- Relevance (Ad, User, Content) = 0.7 * Relevance(Ad, User) + 0.3*Relevance(Ad, Content)

Interactive & Ad-Hoc Queries

- What was the {per-minute, hourly, daily} conversion rate:
  - For an Ad?
  - For a campaign?
- How many Ads were clicked on as a percentage of viewed, per hour for a campaign?
- How much money does a campaign owe to Acme for the whole day?
- What are the most clicked ads & campaigns per hour?
- How many male users does Acme have aged 0-21, 21-40?

Prototype Implementation



Scale Factors

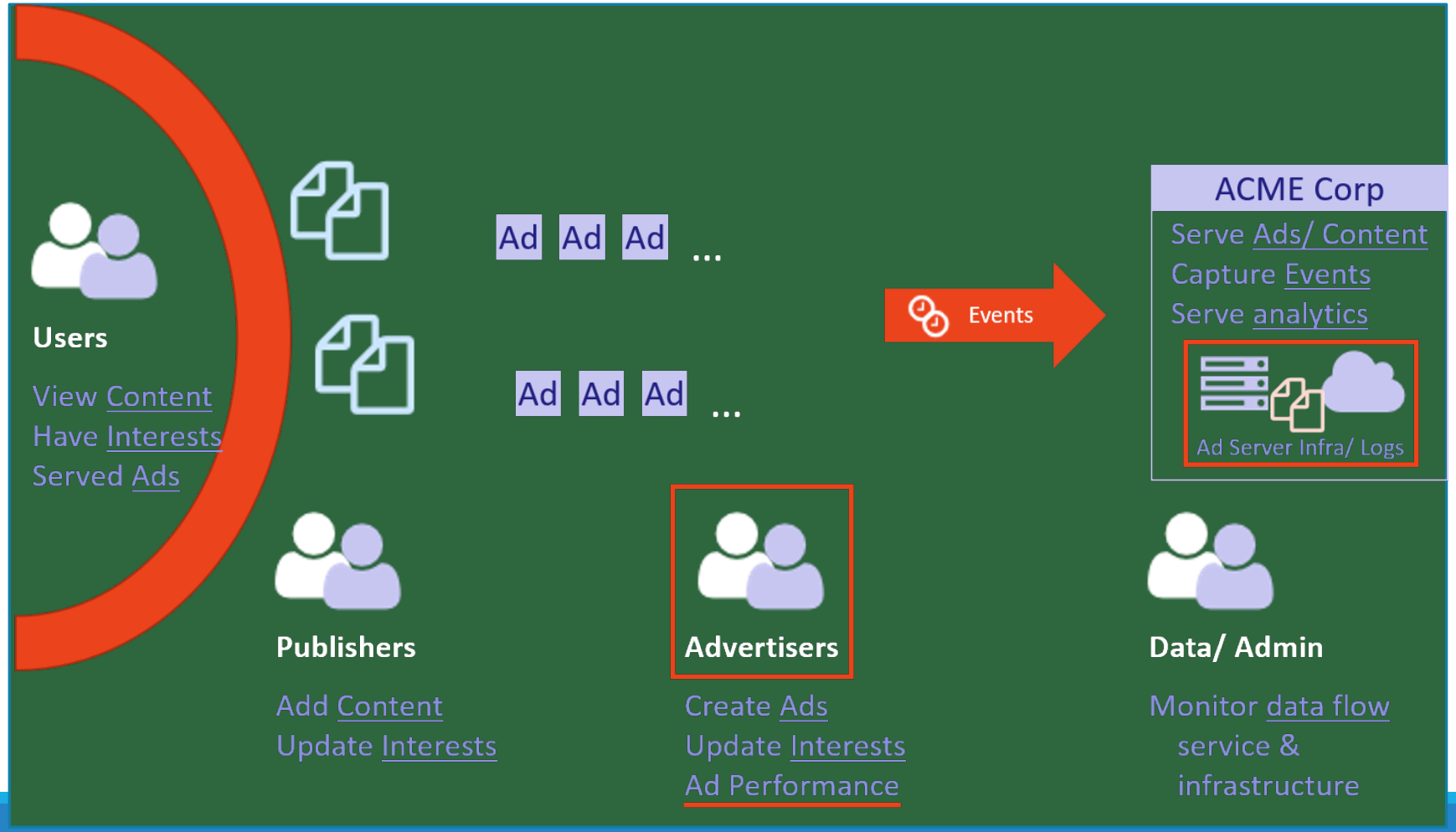
Class	Users	Ads	Contents	Events/Second	Typical Industry
Tiny	100,000	10	10	1,000	None: Test
Small	1,000,000	100	100	10,000	Banking, Healthcare
Medium	10,000,000	1,000	1,000	100,000	Media, Gaming
Large	100,000,000	10,000	10,000	1,000,000	Telco, Web-Scale, Viral Apps
Huge	1,000,000,000	100,000	100,000	10,000,000	Huge Web-Scale, e.g. FB, Google

Metrics

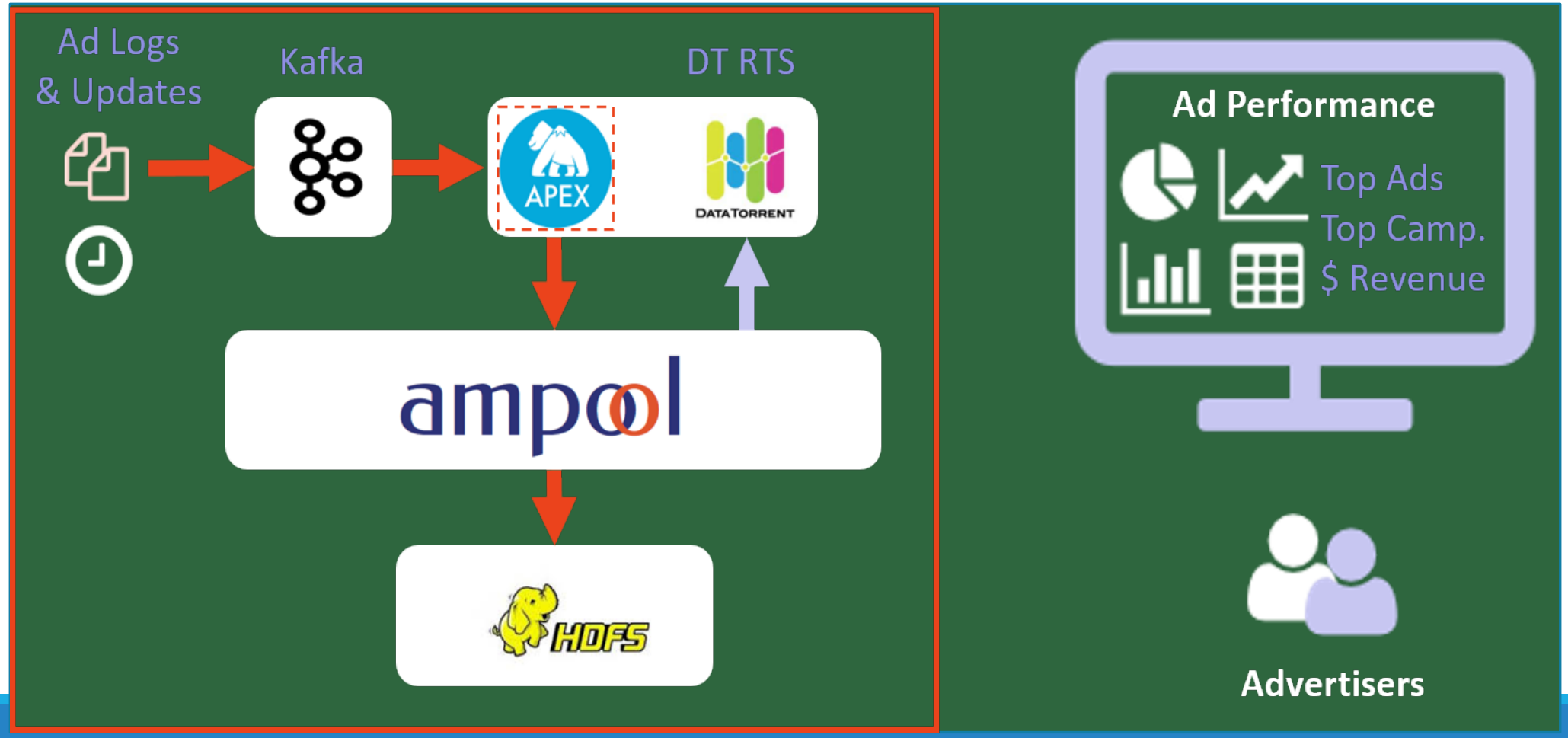
- Different Metrics across different stages
  - Number of events processed per second
  - Time needed for batch computation & Ad-Hoc Queries
  - Query Concurrency
- Combined Metrics
  - Latency between Event Generation to Event Processing
- Cost to meet SLAs
- Operational Complexity

Demo

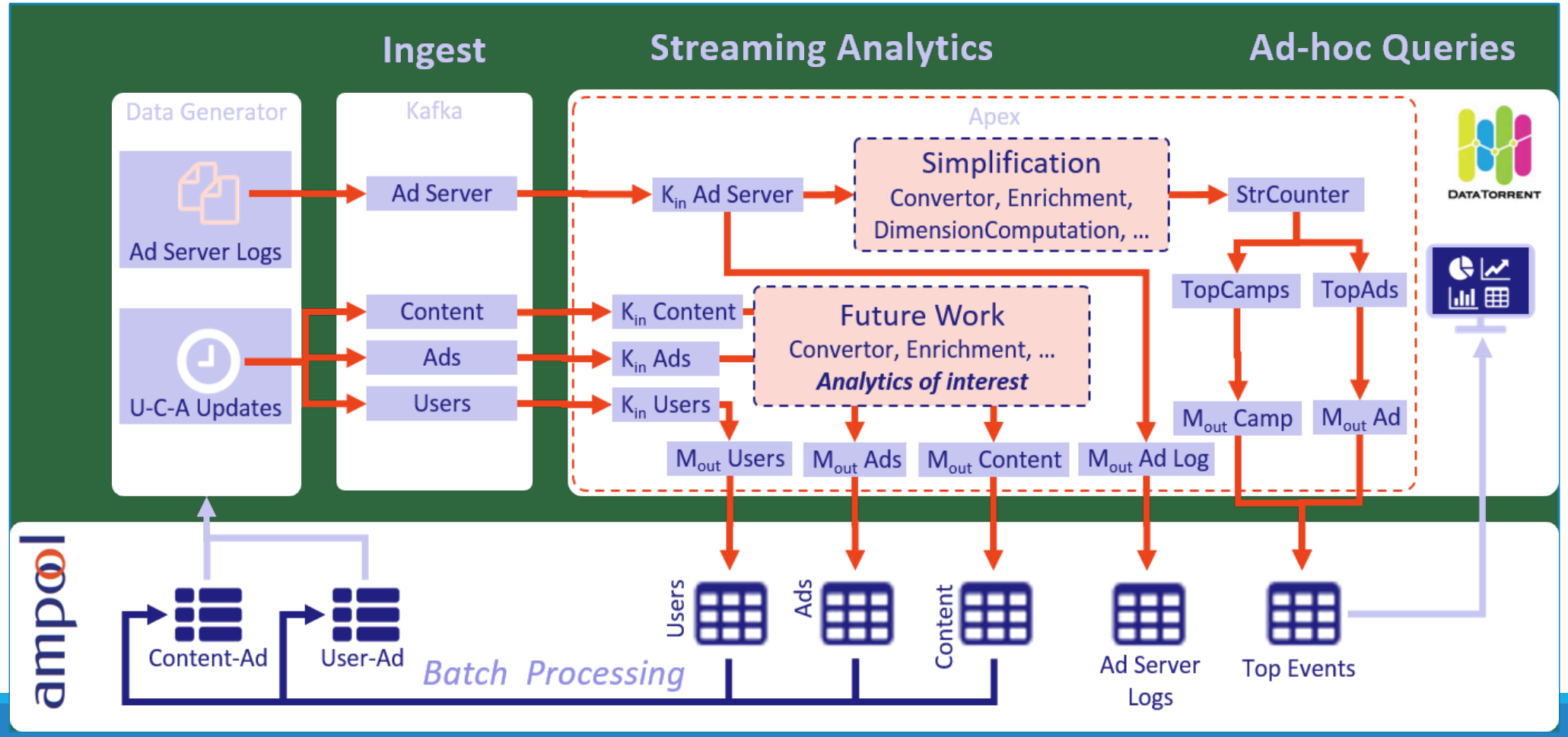
Illustrative Use Case in Ad-Tech

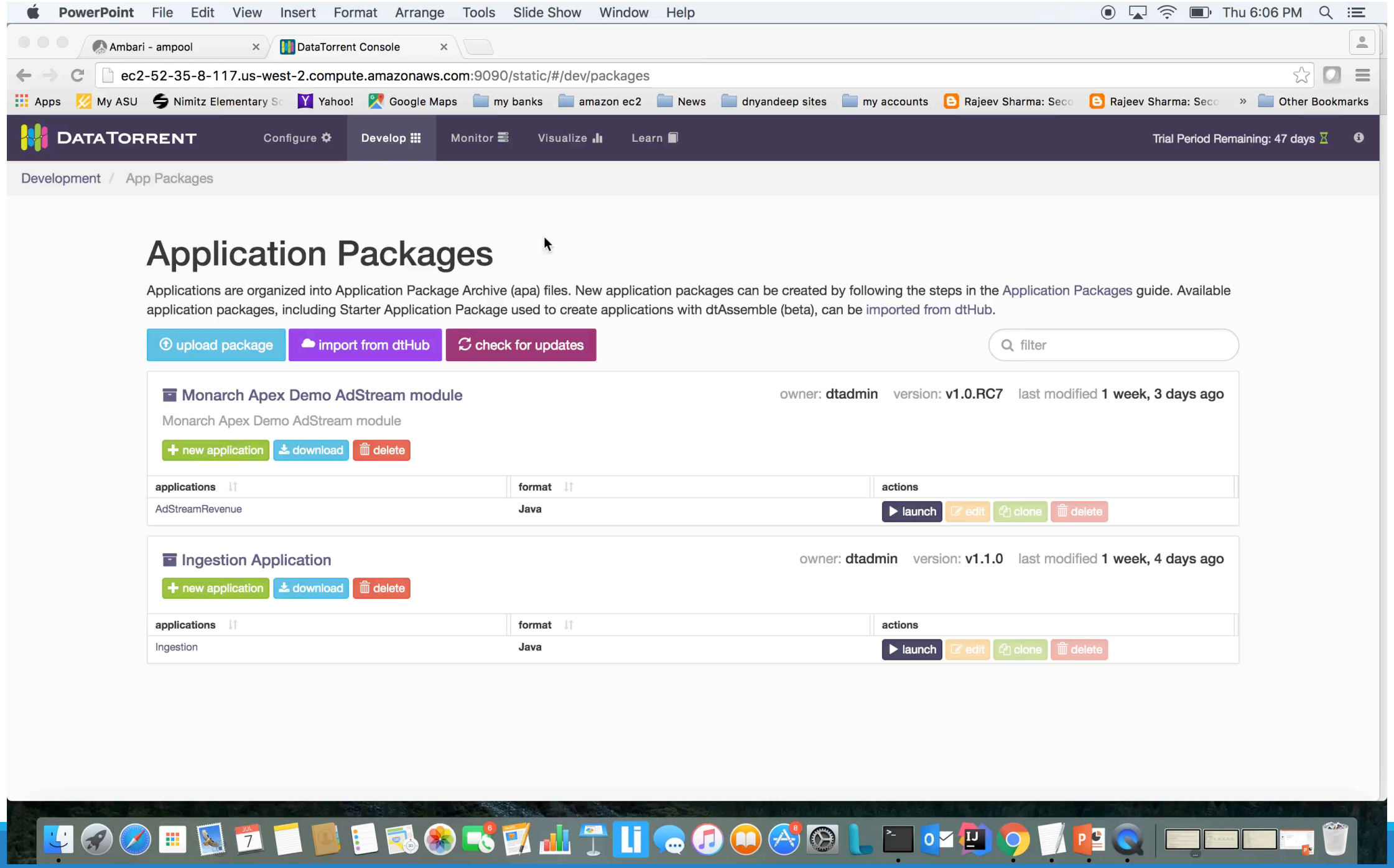


Ad Analytics Pipeline with Kafka-Datatorrent-Ampool

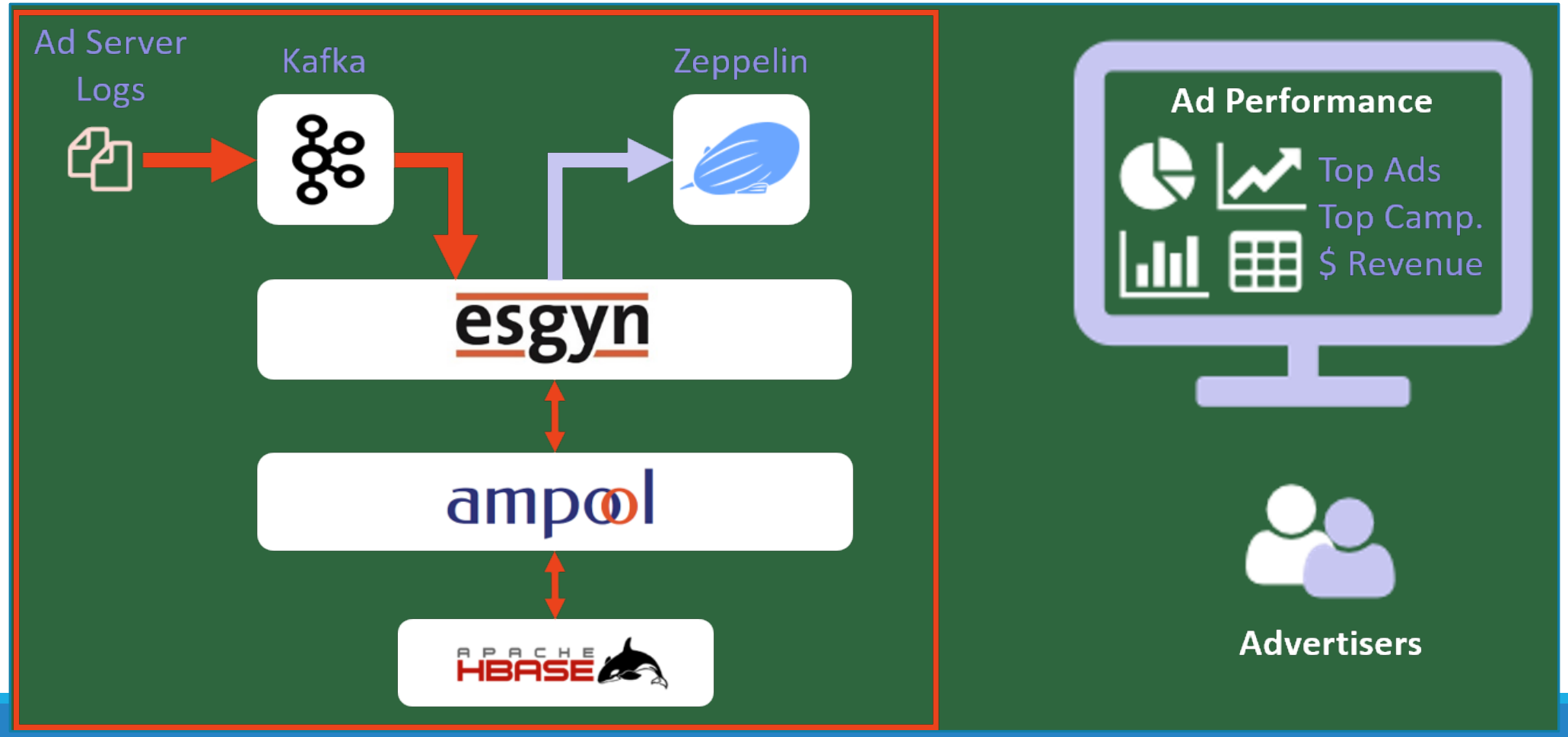


Streaming Ad Analytics

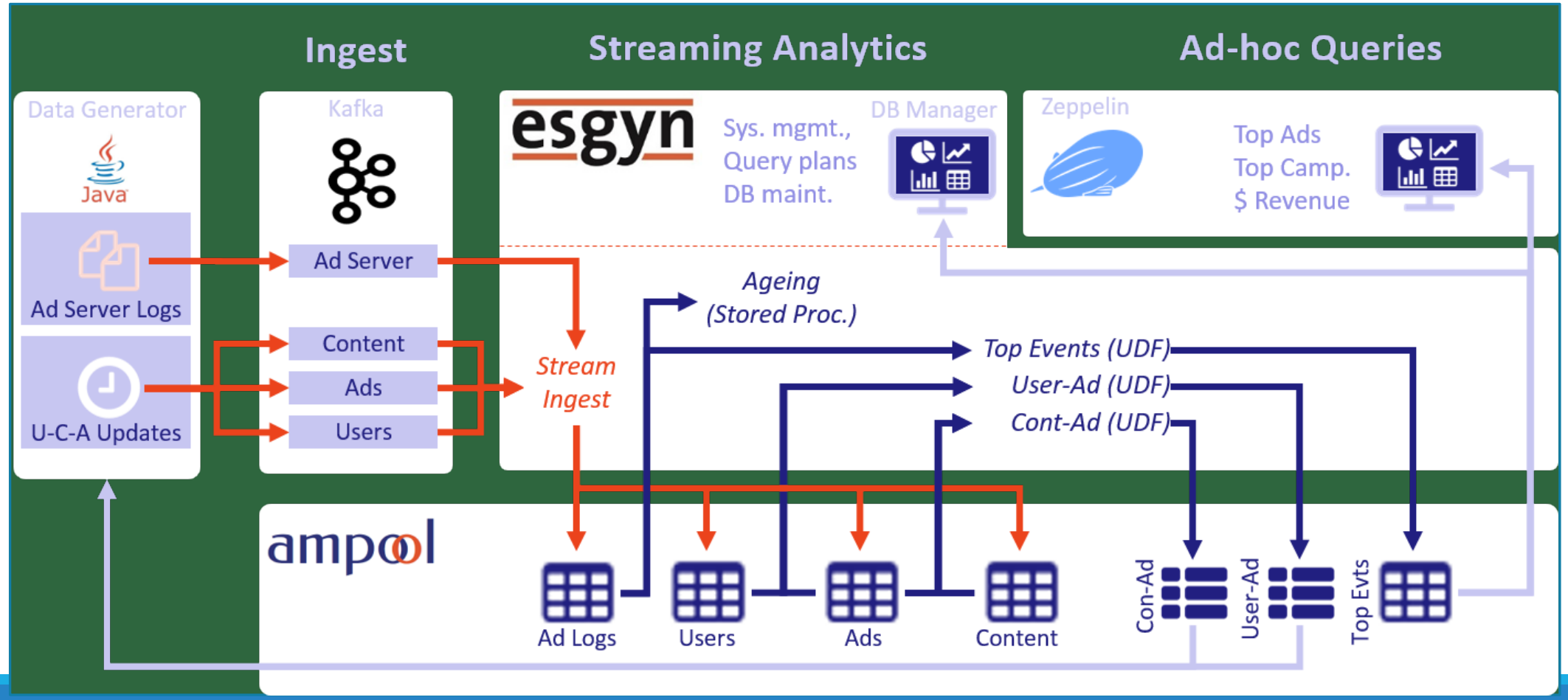



PowerPoint File Edit View Insert Format Arrange Tools Slide Show Window Help
Thu 6:06 PM
Ambari - ampool
DataTorrent Console
ec2-52-35-8-117.us-west-2.compute.amazonaws.com:9090/static/#/dev/packages
Apps My ASU Nimitz Elementary Sc Yahoo! Google Maps my banks amazon ec2 News dnyandeep sites my accounts Rajeev Sharma: Seco Rajeev Sharma: Seco Other Bookmarks
DATATORRENT
Configure
Develop
Monitor
Visualize
Learn
Trial Period Remaining: 47 days
Development / App Packages
Application Packages
Applications are organized into Application Package Archive (apa) files. New application packages can be created by following the steps in the Application Packages guide. Available application packages, including Starter Application Package used to create applications with dtAssemble (beta), can be imported from dtHub.
upload package
import from dtHub
check for updates
filter
Monarch Apex Demo AdStream module
owner: dtadmin version: v1.0.RC7 last modified 1 week, 3 days ago
Monarch Apex Demo AdStream module
new application download delete
applications format actions
AdStreamRevenue Java launch edit clone delete
Ingestion Application
owner: dtadmin version: v1.1.0 last modified 1 week, 4 days ago
new application download delete
applications format actions
Ingestion Java launch edit clone delete

Ad Analytics Pipeline with Kafka-EsgynDB-Ampool



Ad Analytics with Real-Time SQL



10.1.10.126:9995/#/notebook/2BEZGSUWU

Zeppelin Notebook Interpreter Connected

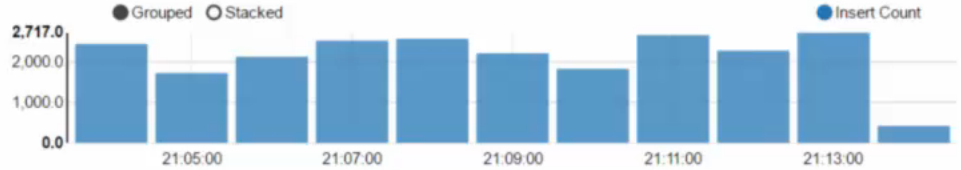
Acme Kafka To EsgynDB Ingestion

default

Kafka Message Ingestion Per Minute

FINISHED

settings




```
%sql
--upsert into adtechnew.AdServerLog
--select converttimestamp((210866760000009 + Impression_ts)*1000),
--    ip_address, UserID, AdID, ContentID, event_type
--from udf(trafka.kafka('localhost:2181', -- zookeeper connection
--    0,                          -- Kafka group id
--    'ad-server-log-topic',      -- Kafka topic
--    'LC16C36C36C36C1',          -- int, and two char output cols
--    '|',                        -- field delimiter
--    10,                         -- max. rows to read
--    10000))                     -- timeout 10 seconds
--KafkaResult(Impression_ts, ip_address, UserID,
--    AdID, ContentID, event_type);-- name the output columns
```

ERROR

Took 0 seconds. (outdated)

trafodion@sandbox:~/adtech/generator

```
ad-server-log-topic [1459286085892|192.168.65.18|69
     |77                                  |11979be3-bf0c-41c2-8e51-15ecc6b0aca7|
V]
ad-server-log-topic [1459286086194|192.168.64.71|256
     |892                                 |4f880175-a850-48ae-afba-47661354aa01|
C]
ad-server-log-topic [1459286086498|192.168.12.12|359
     |459                                 |78621243-bebc-4578-bbd6-58e54bc12e75|
K]
ad-server-log-topic [1459286086802|192.168.88.251|48
      |506                                 |9d150dc6-a332-41b3-9f65-74092f718521
|V]
users-topic [598                                  |18|M|60644|1459088779082|LG 25
 CF Refrigerator 3 Door White Tall



           ]
ad-server-log-topic [1459286087106|192.168.144.199|701
       |118                                 |8cf97b34-fb19-4635-9f50-d58377d2784
3|K]
ad-server-log-topic [1459286087411|192.168.107.41|215
     |322                                 |36f0d3e9-fa16-4eb7-a199-b1c8dacc302a
|C]
```

trafodion@sandbox:~/trafka

```
Total ad-server-log-topic processed : 16,090
Total users-topic processed : 4,910
Total ad-server-log-topic processed : 16,100
Total ad-server-log-topic processed : 16,110
Total ad-server-log-topic processed : 16,120
Total ad-server-log-topic processed : 16,130
Total users-topic processed : 4,920
Total ad-server-log-topic processed : 16,140
Total ad-server-log-topic processed : 16,150
Total ad-server-log-topic processed : 16,160
Total users-topic processed : 4,930
Total ad-server-log-topic processed : 16,170
Total ad-server-log-topic processed : 16,180
Total ad-server-log-topic processed : 16,190
Total users-topic processed : 4,940
Total ad-server-log-topic processed : 16,200
Total ad-server-log-topic processed : 16,210
Total ad-server-log-topic processed : 16,220
Total users-topic processed : 4,950
Total ad-server-log-topic processed : 16,230
Total ad-server-log-topic processed : 16,240
Total ad-server-log-topic processed : 16,250
Total ad-server-log-topic processed : 16,260
Total users-topic processed : 4,960
Total ad-server-log-topic processed : 16,270
```

Future Work

- Parallel Data Generation
  - Generate 10 M events per second
  - Ad Serving events cannot be generated before User, Content, Ad inserts
  - Client-side caching, updated after micro-batch model builds
- Open Source the Prototype Implementation
  - After ongoing validation at 2-3 Ad Tech customers
- Submit proposal for TPC-x
- One Benchmark to Rule Them All !

milind@ampool.io

www.ampool.io @AmpoolIO /company/ampool-inc- /AmpoolIO