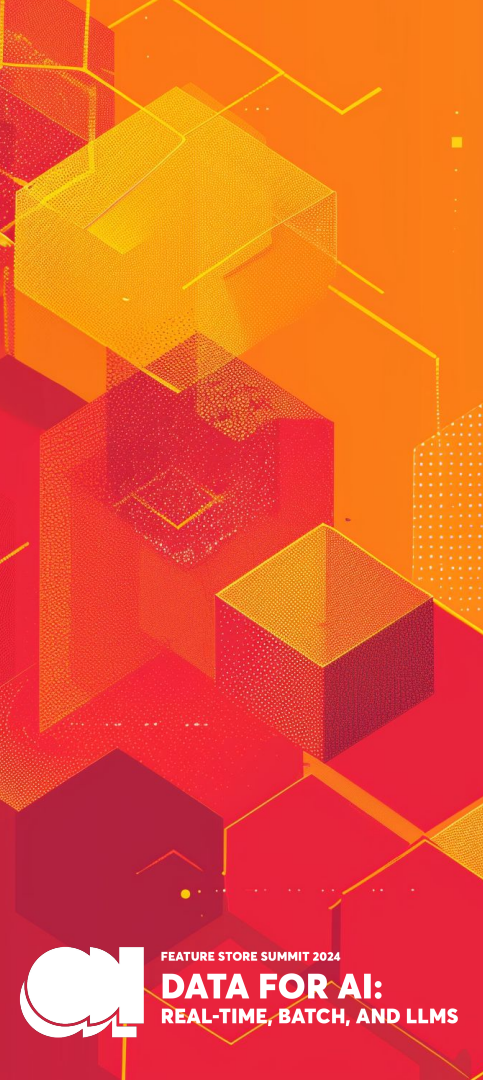


Etsy

Serving Real-Time Features at Etsy

Tianne Cui, Senior Software Engineer, Etsy

Prachi Poddar, Engineering Manager, Etsy

Organized by HOPSWORKS

Overview

- Understanding Etsy scale
- Features at Etsy
- Rivulet: Real-Time Feature System
  - History
  - Overview
  - Feature Workflows
- Challenges and Ideas

Etsy.com Stats

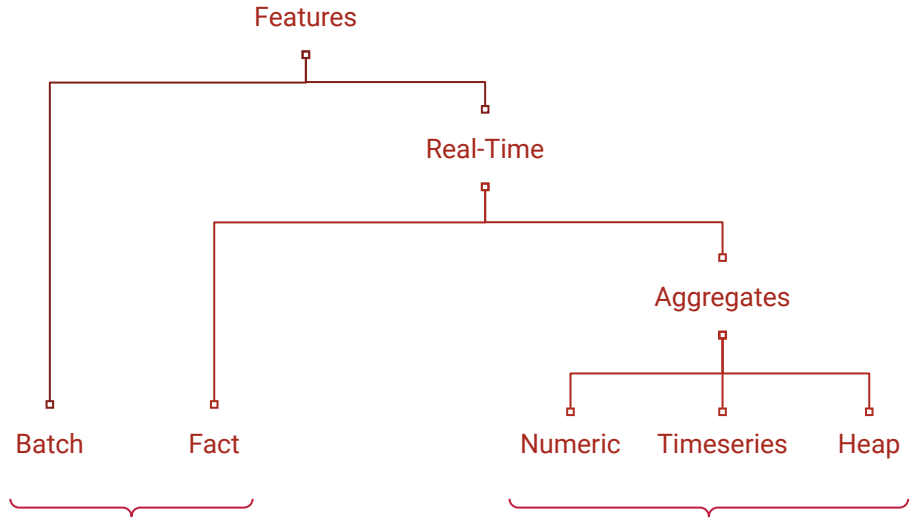
Active Sellers: 6.6 Million
Active Buyers: 91.5 Million
Active Listings: More than 100 Million

ML Feature System Stats

Batch Feature Requests: ~200 Thousands / second
Batch Feature Retrieval Latency [p99]: ~40 msec

Real-Time Feature Requests: ~160 Thousands / second
Real-Time Feature Retrieval Latency [p99]: ~50 msec
Real-Time Feature Data Freshness: < 1 seconds (timeseries) ~ 15 seconds (numeric)

Features at Etsy



key	feature_family_1		feature_family_2
	feature_1	feature_2	feature_3
user_1	value_1	value_2	value_3

key	default_column_family
	default_qualifier
<hash>#<feature_name>#<feature_version>#user_1#<timestamp>	value_1

Features at Etsy

Search Ranking Use Case

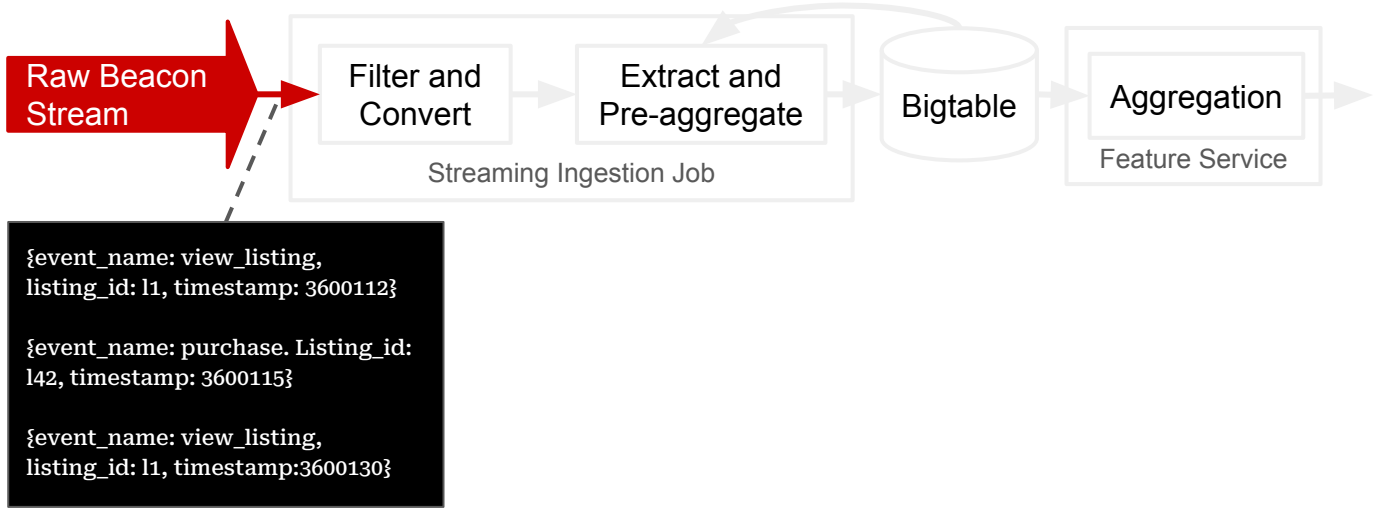
Goal: Find the most relevant listings for a search query.

Listing's total view counts in the past 24 hours(Popularity)

Count Feature

Timeseries Feature

Heap Feature



Features at Etsy

Search Ranking Use Case

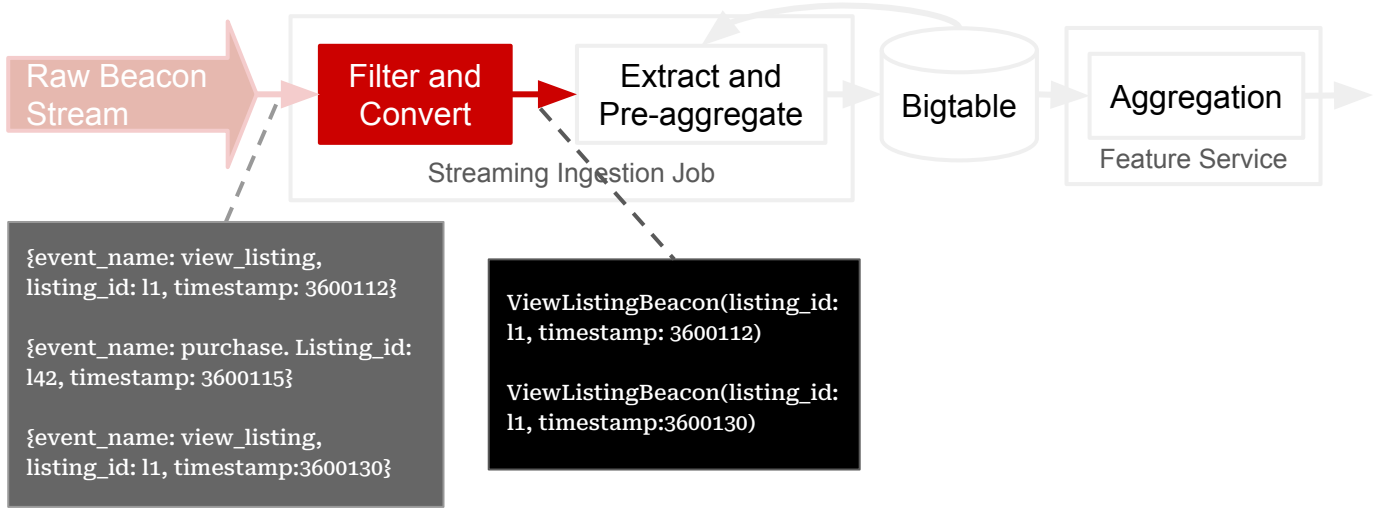
Goal: Find the most relevant listings for a search query.

Listing's total view counts in the past 24 hours(Popularity)

Count Feature

Timeseries Feature

Heap Feature



Features at Etsy

Search Ranking Use Case

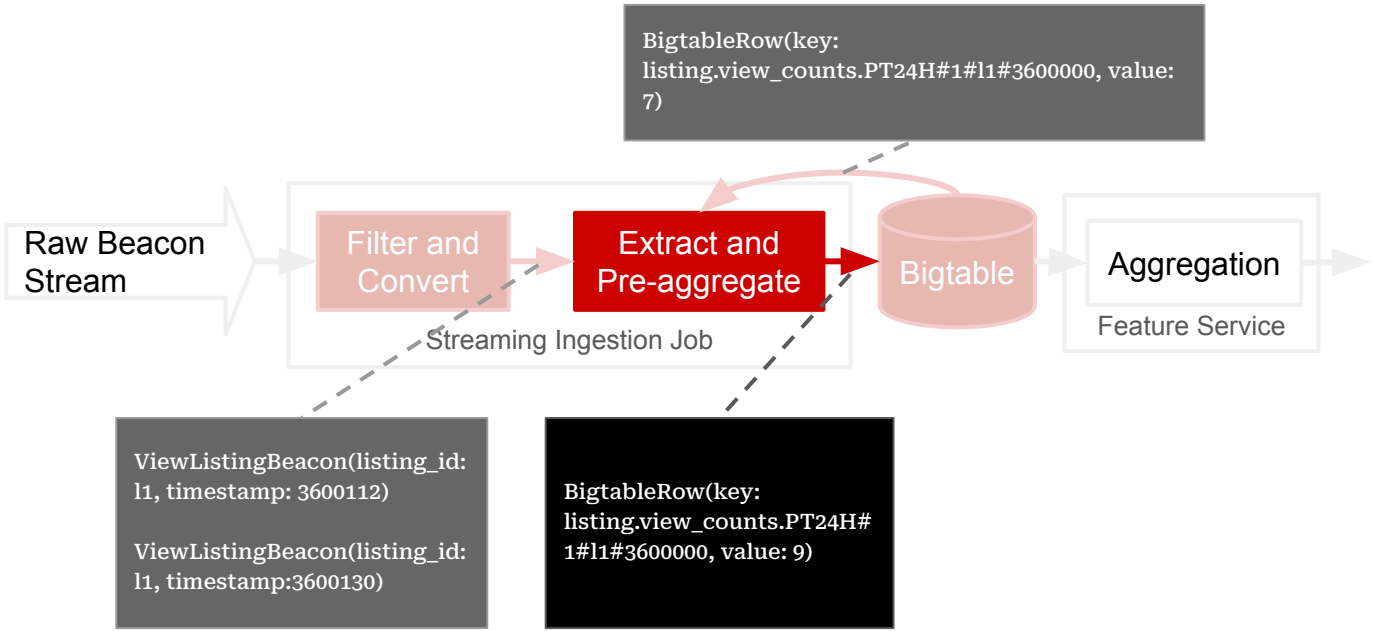
Goal: Find the most relevant listings for a search query.

Listing's total view counts in the past 24 hours(Popularity)

Count Feature

Timeseries Feature

Heap Feature



Features at Etsy

Count Feature

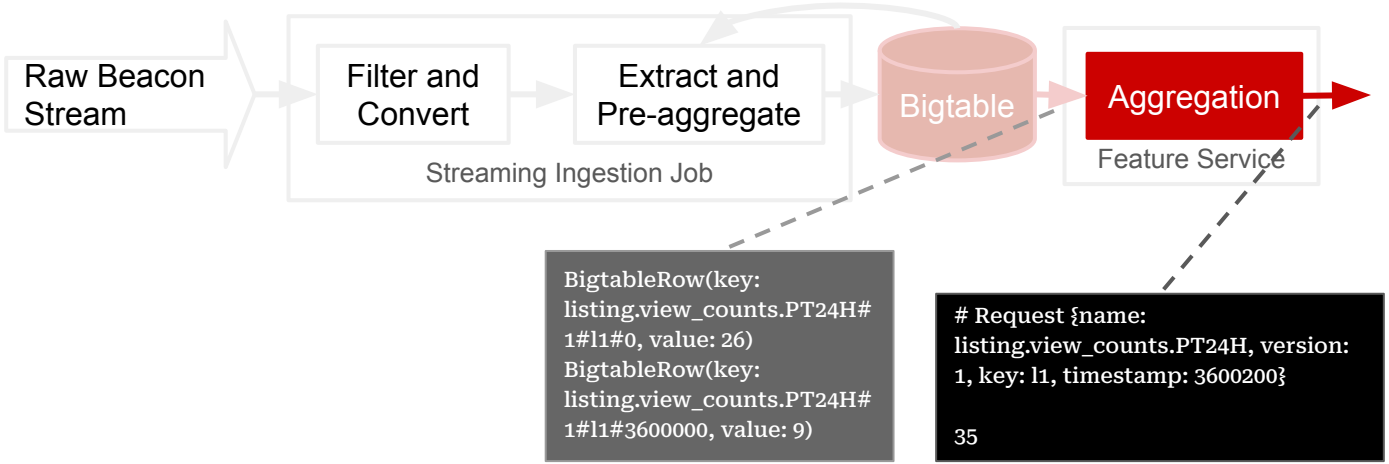
Timeseries Feature

Heap Feature

Search Ranking Use Case

Goal: Find the most relevant listings for a search query.

Listing's total view counts in the past 24 hours(Popularity)

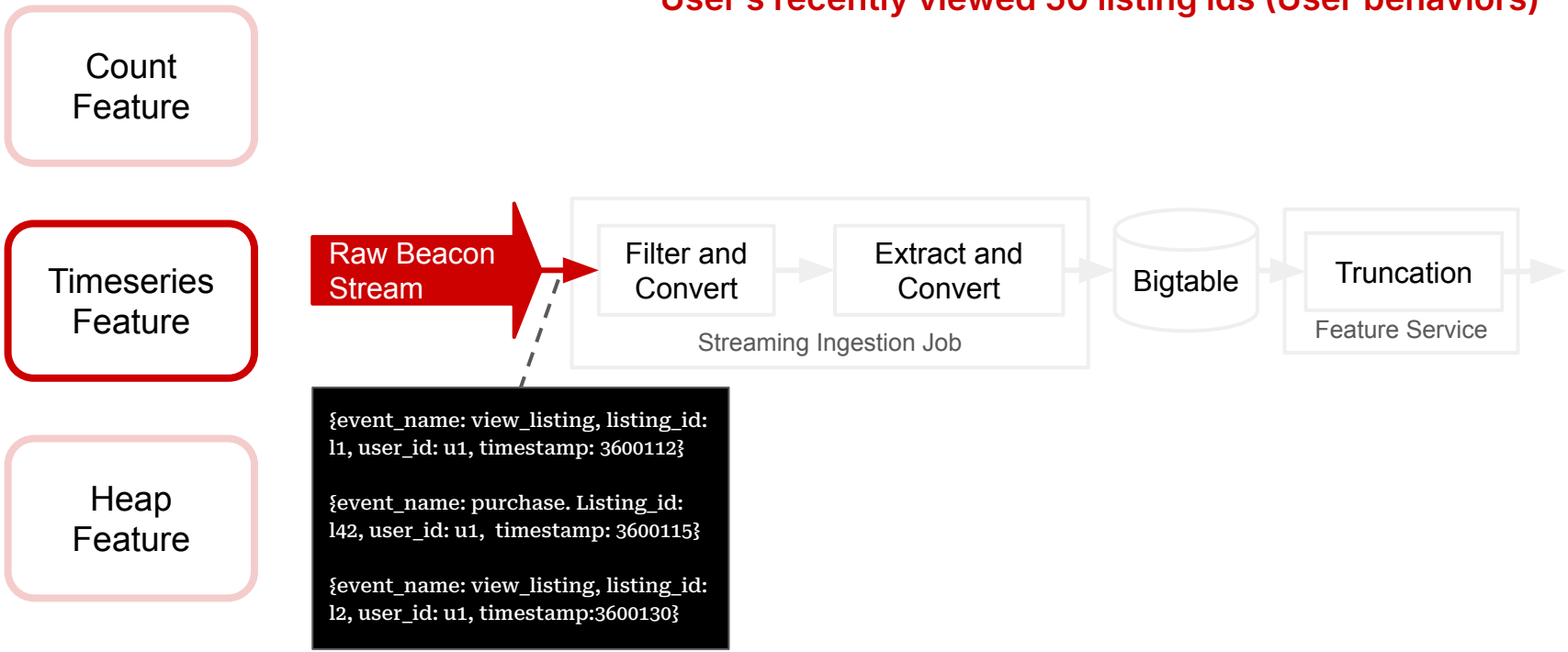


Features at Etsy

Search Ranking Use Case

Goal: Find the most relevant listings for a search query.

User's recently viewed 50 listing ids (User behaviors)



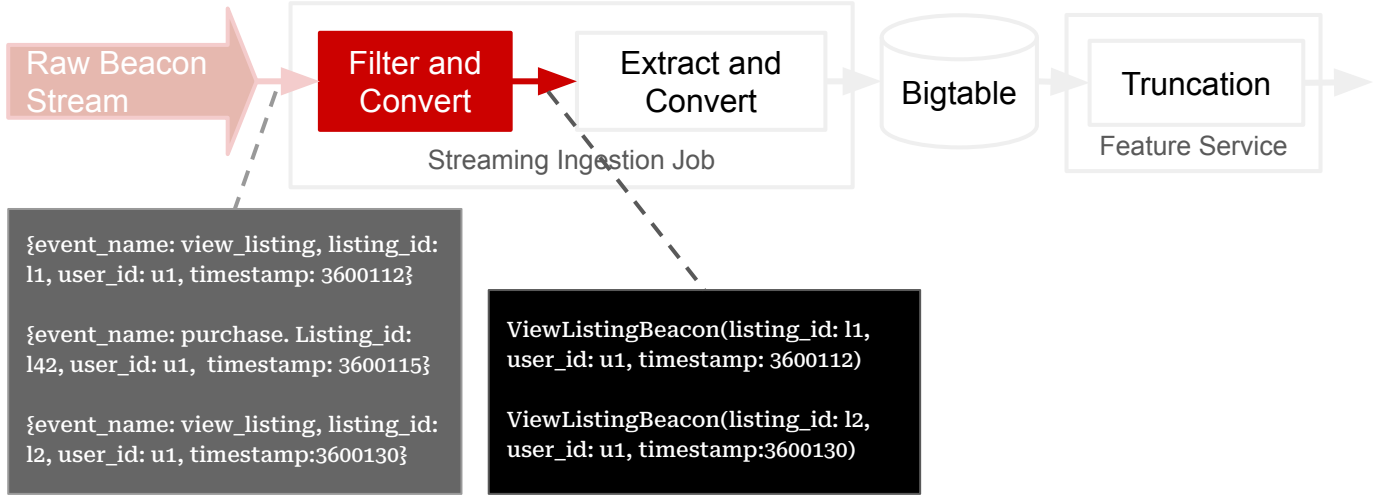
Search Ranking Use Case

Goal: Find the most relevant listings for a search query.

User's recently viewed 50 listing ids (User behaviors)

Features at Etsy

- Count Feature
- Timeseries Feature
- Heap Feature



Features at Etsy

Count Feature

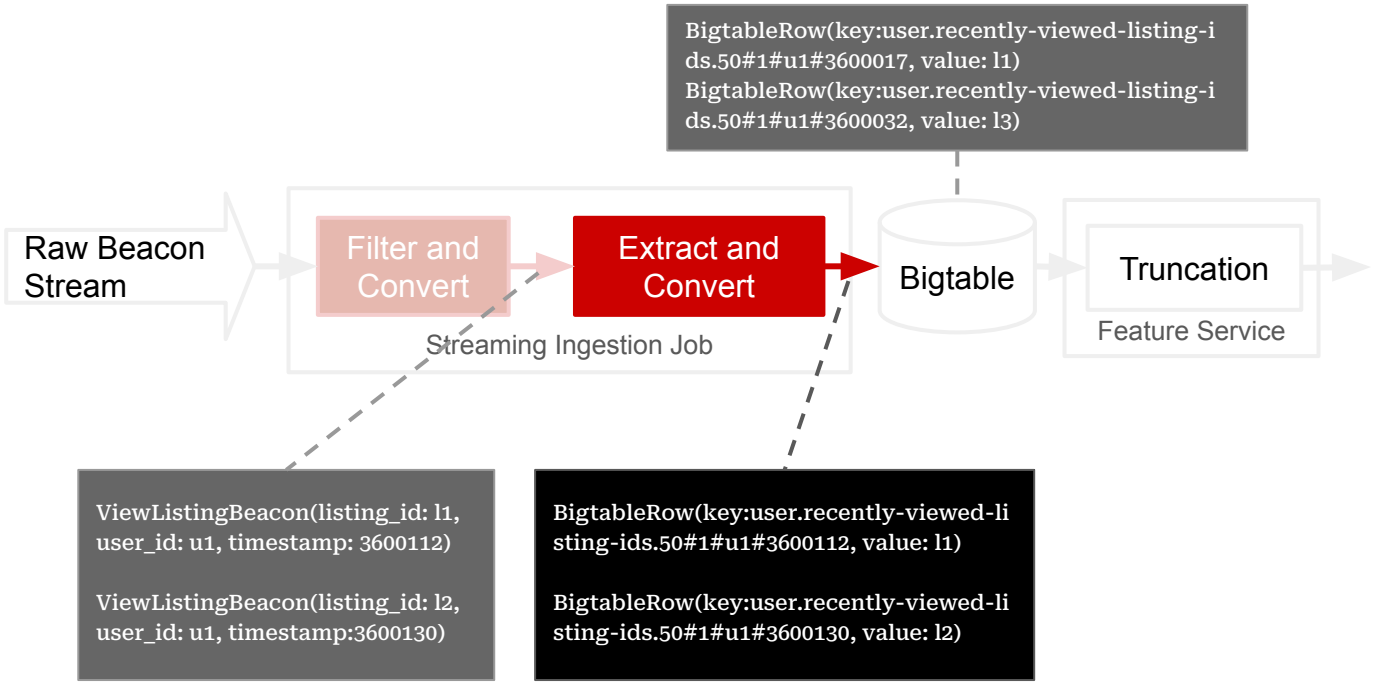
Timeseries Feature

Heap Feature

Search Ranking Use Case

Goal: Find the most relevant listings for a search query.

User's recently viewed 50 listing ids (User behaviors)



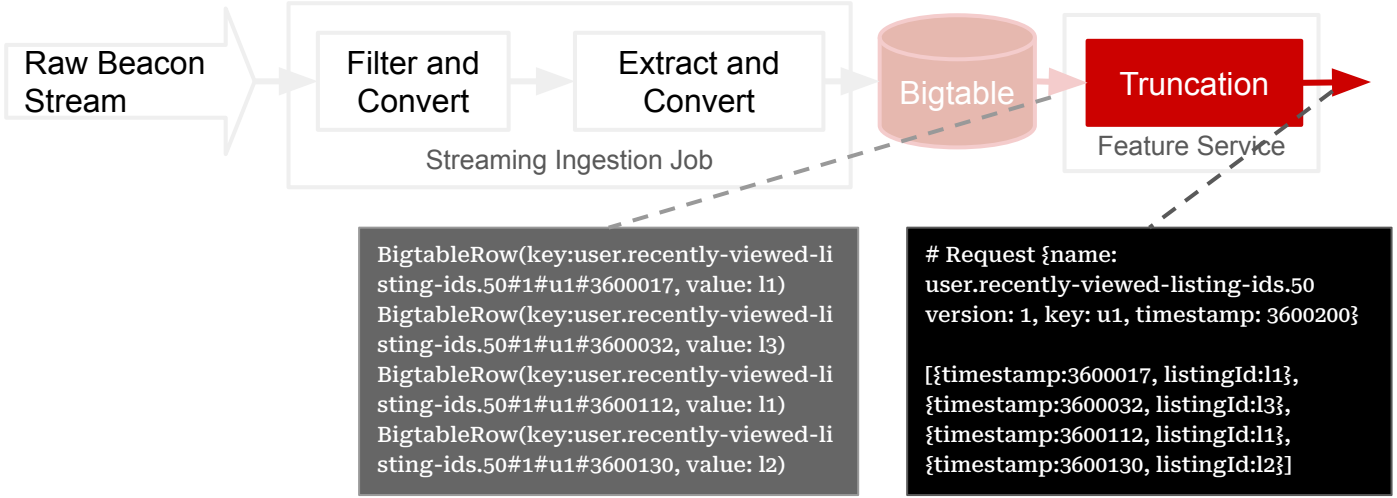
Features at Etsy

- Count Feature
- **Timeseries Feature**
- Heap Feature

Search Ranking Use Case

Goal: Find the most relevant listings for a search query.

User's recently viewed 50 listing ids (User behaviors)



Features at Etsy

Count Feature

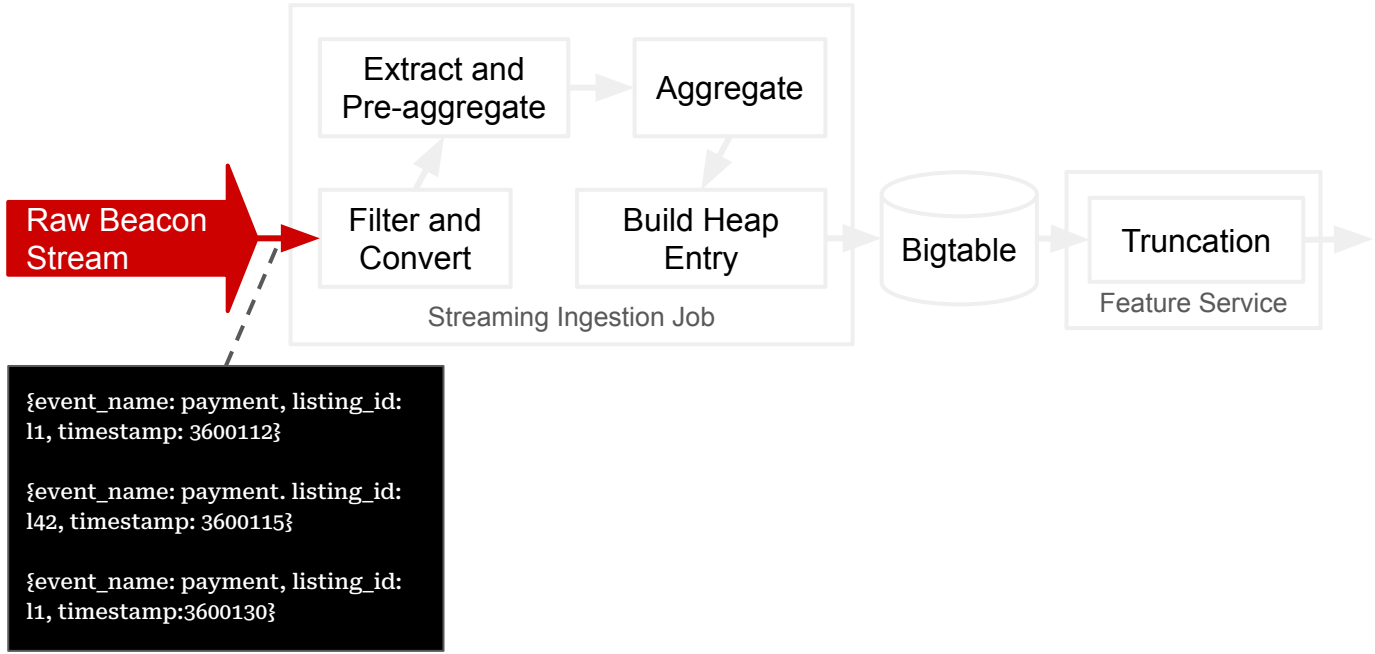
Timeseries Feature

Heap Feature

Use Case

Goal: Determine top items.

Top purchased listing ids in the past 4 hours



Features at Etsy

Count Feature

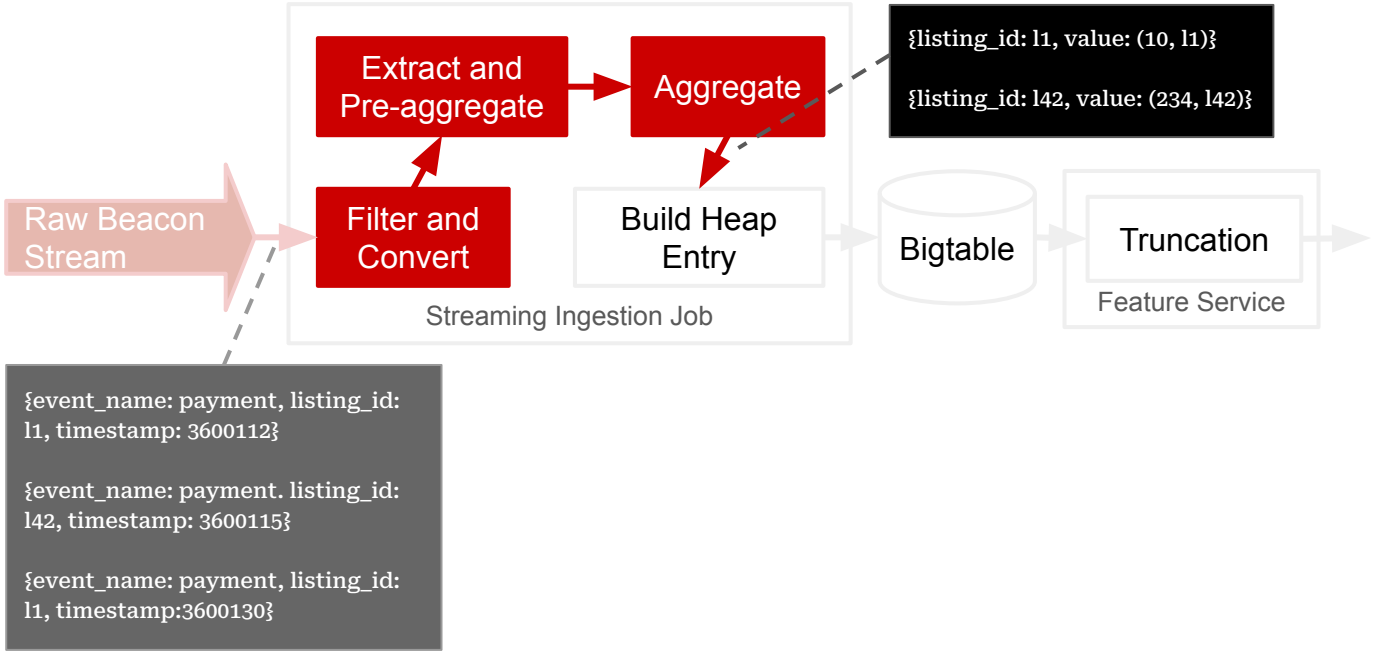
Timeseries Feature

Heap Feature

Use Case

Goal: Determine top items.

Top purchased listing ids in the past 4 hours



Features at Etsy

Count Feature

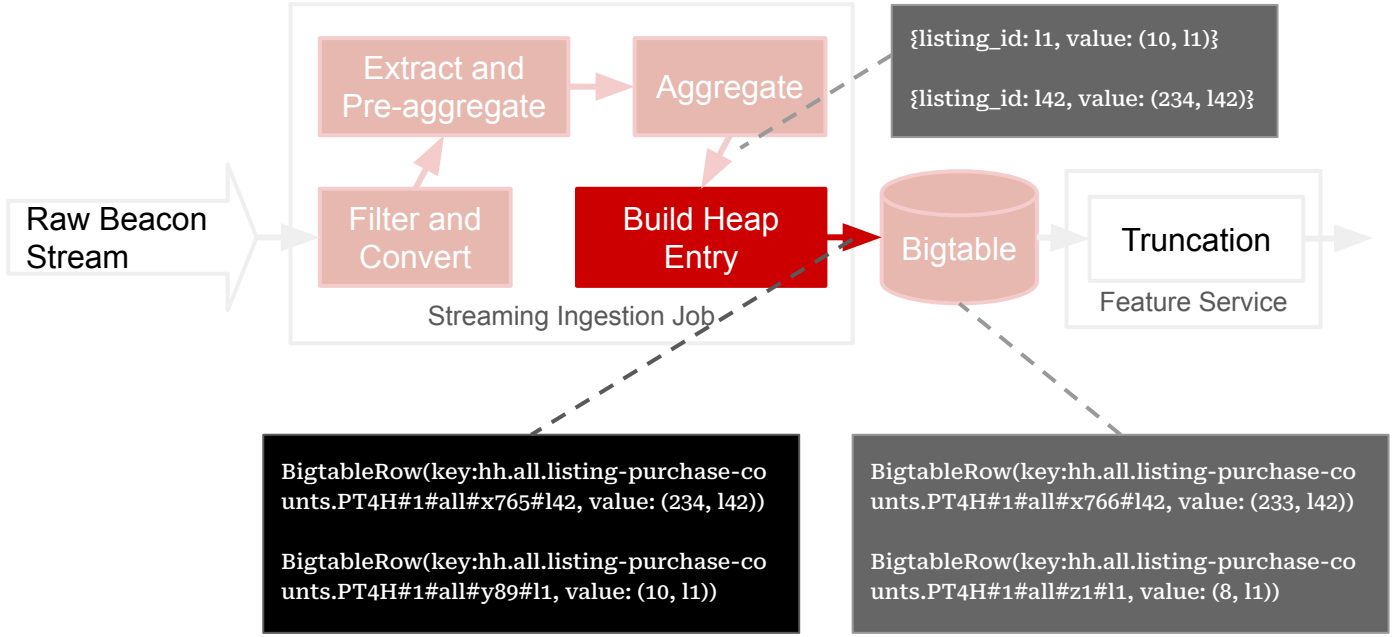
Timeseries Feature

Heap Feature

Use Case

Goal: Determine top items.

Top purchased listing ids in the past 4 hours



Features at Etsy

Count Feature

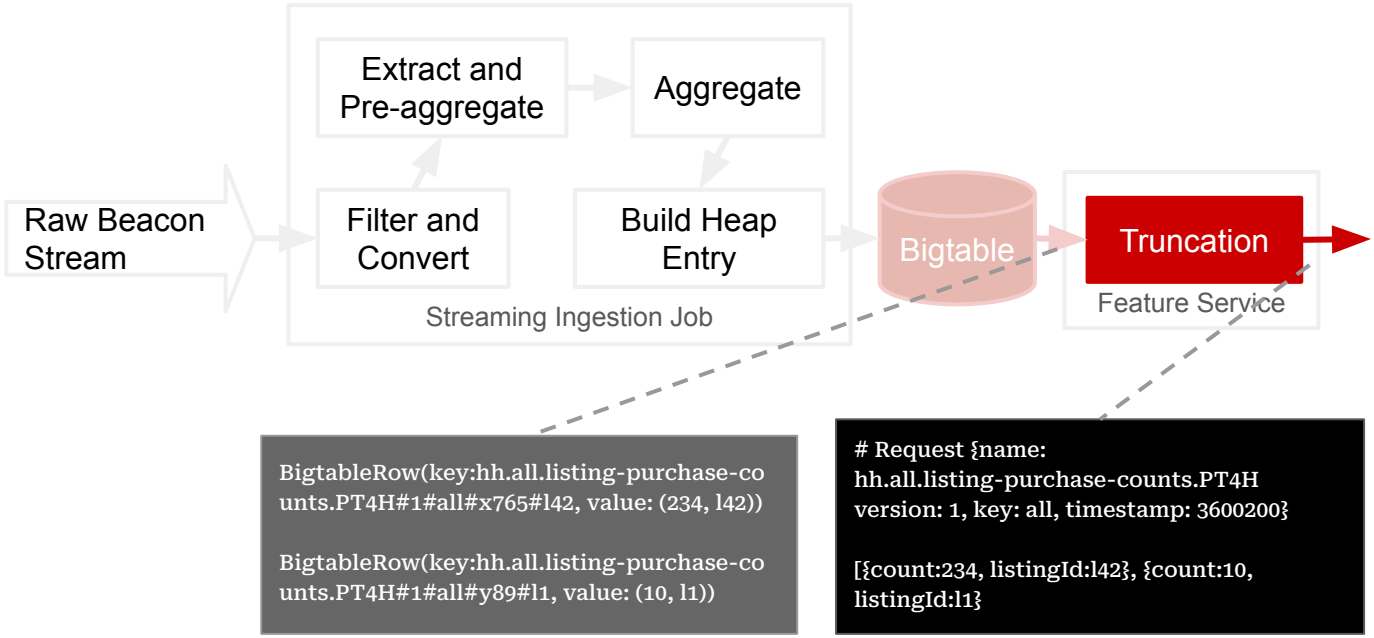
Timeseries Feature

Heap Feature

Use Case

Goal: Determine top items.

Top purchased listing ids in the past 4 hours



Rivulet: Real-Time Feature System

FEATURE STORE SUMMIT 2024
DATA FOR AI:
REAL-TIME, BATCH, AND LLMS

Rivulet - The History

2019 - Rivulet Launch

- First centralized Feature Store at Etsy
- Feature Definitions in Scala Code
- Adopted nearly 100% by ML-mature teams
- Built specifically for the cloud environment (GCP)

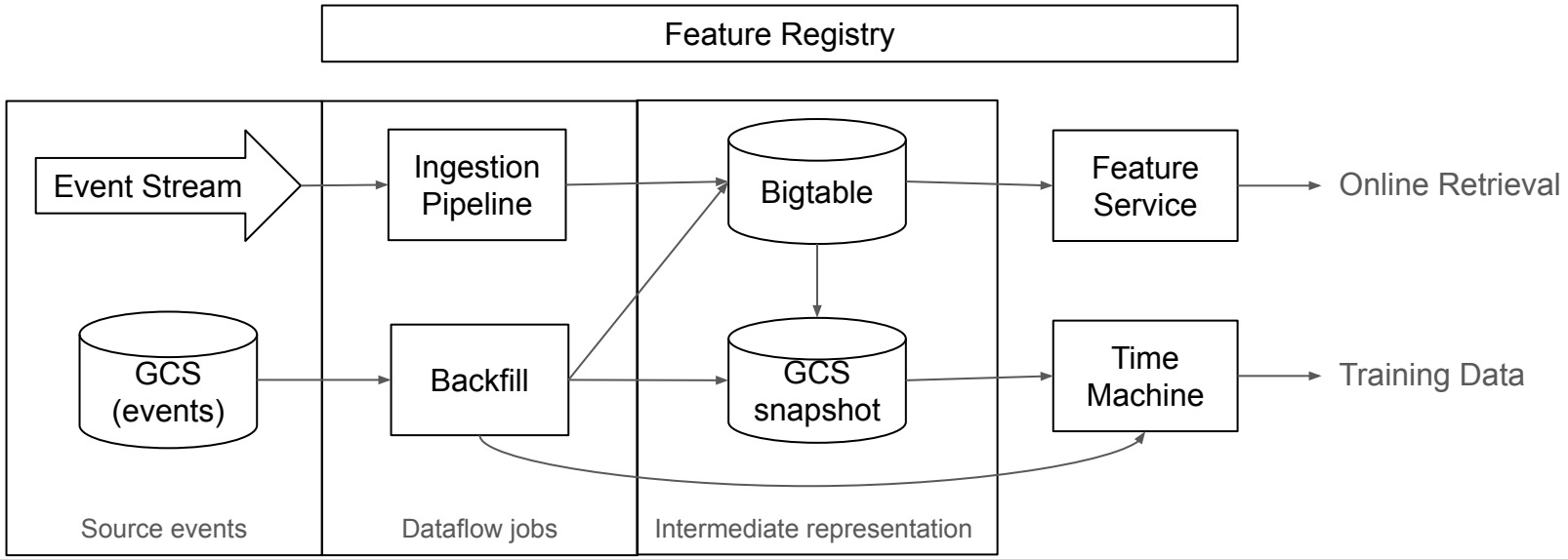
- Resilience to uncertainty
- Designed for small volume online requests from the non-ML systems
- Designed as a unified model of streaming and batch

2024 - Today

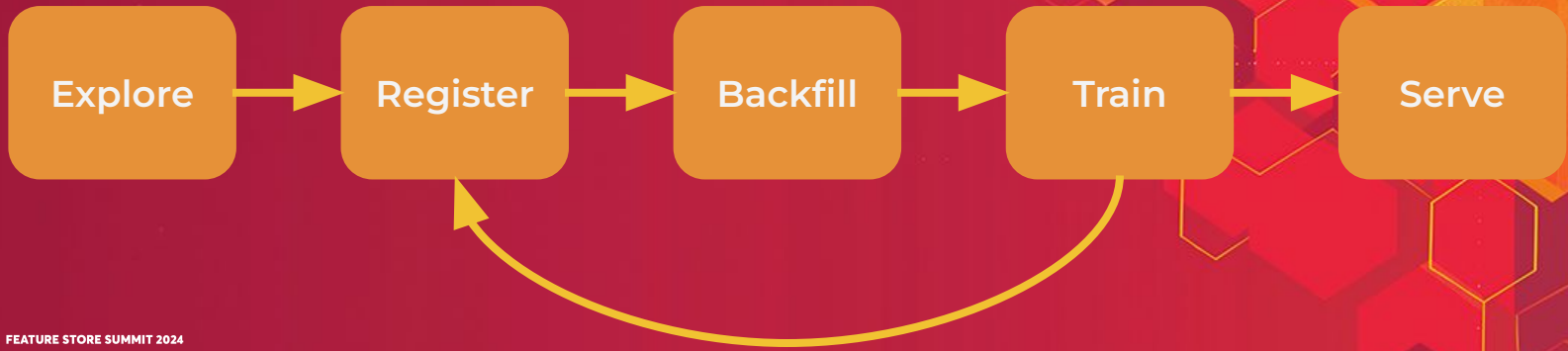
- Accumulators (2022)
- Multi-Window Features (2023)
- Feature Backfill Workflow (2023)
- Time Machine and Backfill Workflow (2023)

- Uncertainties have been eliminated
- Majority of requests consist of large batches for many features to serve ML model inference
- Most Use cases become ‘slicing and aggregating time series in real time’

Rivulet - Today's Overview



Rivulet Feature Workflow



Rivulet Feature Workflow

Feature Exploration

Google BigQuery

Hydrology

powered by rivulet

Rivulet Feature Workflow

Feature Exploration

Google BigQuery

Rivulet Feature Workflow

Feature Registration

Via Scala Code

```scala
// ==== [start] Example listing.view-counts Feature ====
val viewListingSource = new View[Beacon, ViewListingBeacon](
    rawBeaconSource,
//    prefilter = ???,
    postfilter = CommonExtractors.isValid
  )

val listingViewCounts =
  new CountFeature[ViewListingBeacon](
    name =  "count.listing.view-counts.P30D",
    version = 1,
    doc = "For each listing, the number of times it was viewed over the past month",
    source = viewListingSource,
    batchSource = None,
    selector = CommonExtractors.isValid,
    keyExtractor = ToString(ViewListingBeacon.listingIdExtractor),
    timestampExtractor = CommonExtractors.timestampExtractor,
    maybeAggregationWindow = Some(Duration.ofHours(CountAggregationHoursThirtyDay)),
    maybeBinSizeSec = Some(BinSizeSecondOneHour)
  )
// ==== [end] Example listing.view-counts Feature ====
```

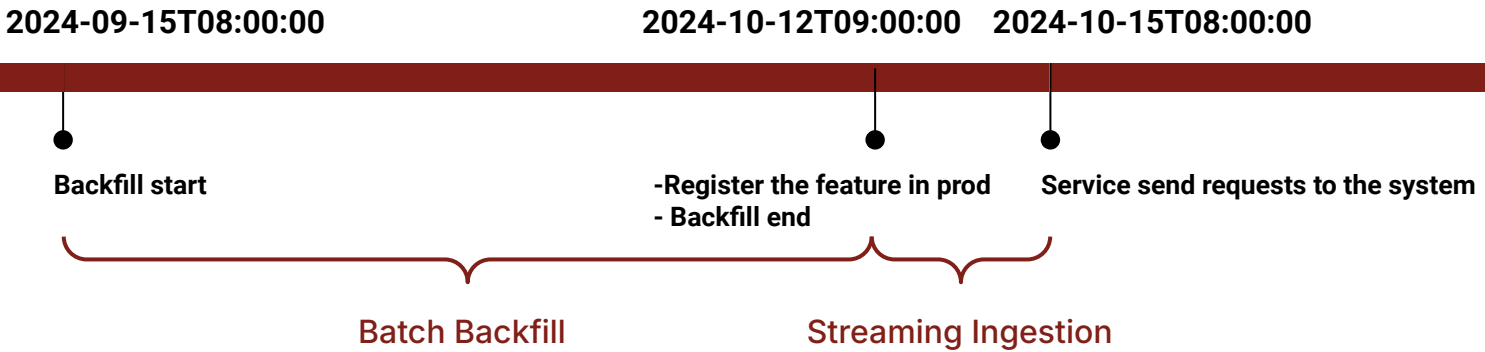
Via Yaml (WIP)

```yaml
familyName: counts
description: Count features.
entityId: ListingId
ttl_ms: 604800000
features:
  - name: viewCountP30DFV1
    description: For each listing, the number of times it was viewed ov
    dataType:
      raw: IntNumeric
    createdAt: '2024-10-14T80:00:00Z'
    owner:
      org: ML Enablement
      team: Feature Store
      slackChannel: '#feature-system'
    additionalInfo:
      source_service: rivulet_feature_service
      rivulet_feature_name: count.listing.view-counts.P30D
      rivulet_feature_version: 1
      rivulet_api_version: v1
      rivulet_aggregation_window: 30d
```

Rivulet Feature Workflow

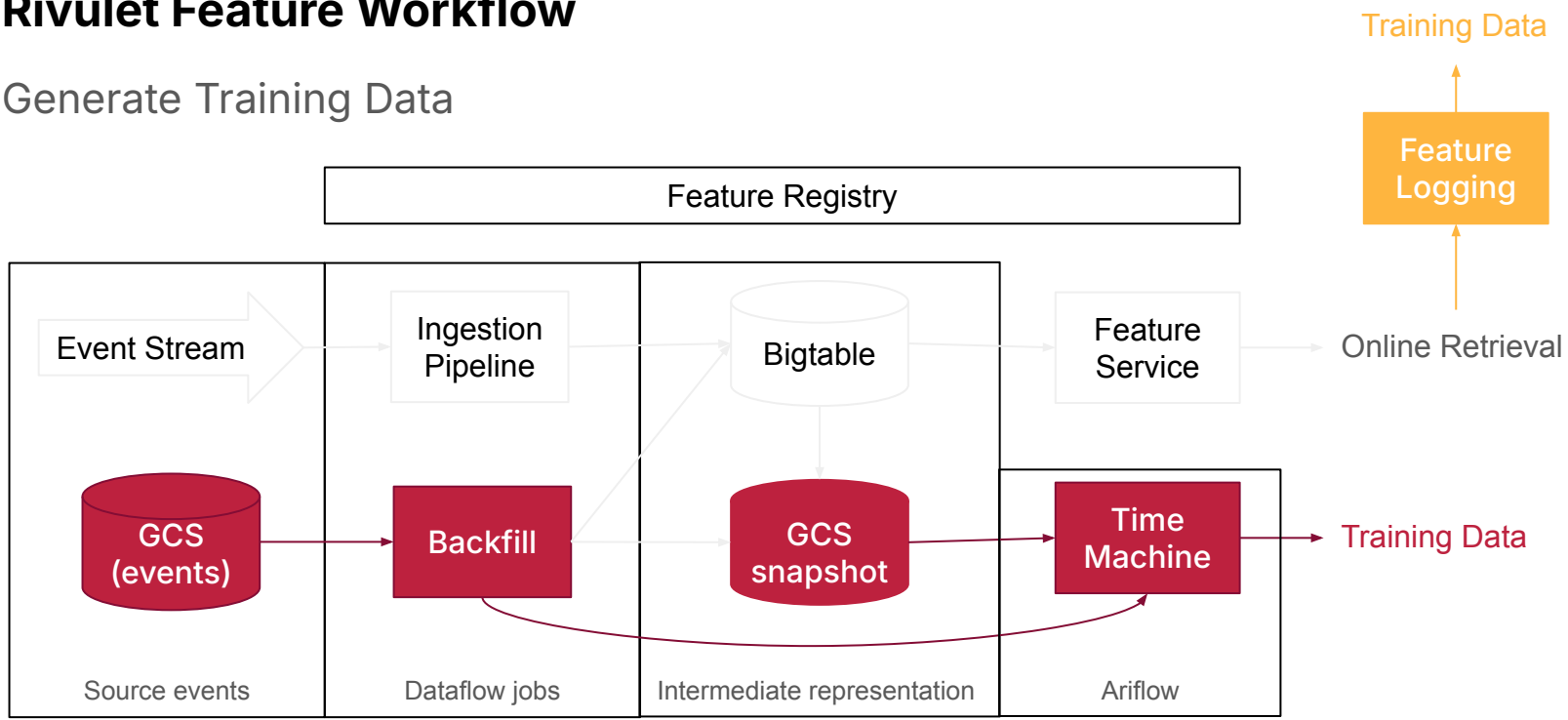
Backfill historic data

Listing's total view counts in the past 30 days(Popularity)



Rivulet Feature Workflow

Generate Training Data



Rivulet Feature Workflow

Online retrieval - Feature Service APIs

Single Requests

GET: /api/v1/feature/{name}/{version}/{key}	

Parameters	• [Required] name: Feature Name
	• [Required] version: Feature Version
	• [Required] key: Key for value to return
	• [Optional] numResults: Maximum number of results to return
	• [Optional] minTimestamp, maxTimestamp: min and max timestamp of the return data
	• [Optional] accumulator: streaming scan
	  ○ Default: return all fetched results matching the conditions
	  ○ Dedupe: return only the unique records matching the conditions
	  ○ Session(accumulatorParams={“accumulatorParams”: “isoDuration”}): return records within the provided session window ending at the most recent record.

Response JSON Object	{"name": "string", "version": 0, "key": "string", "value": {}}
Status Codes	• 404: Feature for name/version does not exist or nothing found for key

Rivulet Feature Workflow

Online retrieval - Feature Service APIs

Single Requests - Example

Request:
GET /api/v1/feature/{name}/{version}/{key}?numResults=<num>&accumulator=dedupe

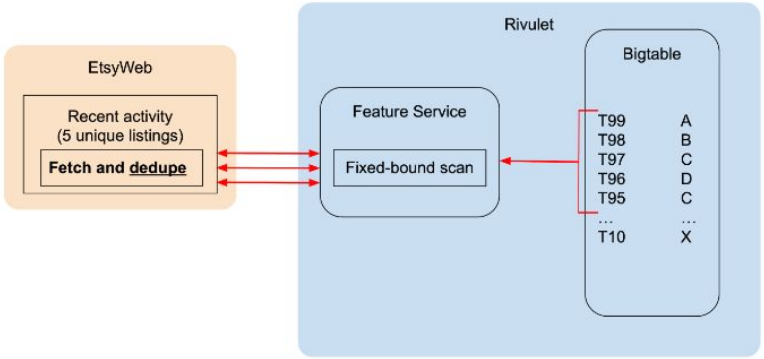
Response:

```
{
  "name": "<name>",
  "version": <version>,
  "key": "<key>",
  "value": [
    {
      "timestamp": "<timestamp_iso>",
      "<timeseries_item_name>": <timeseries_item>
    }
  ]
}
```

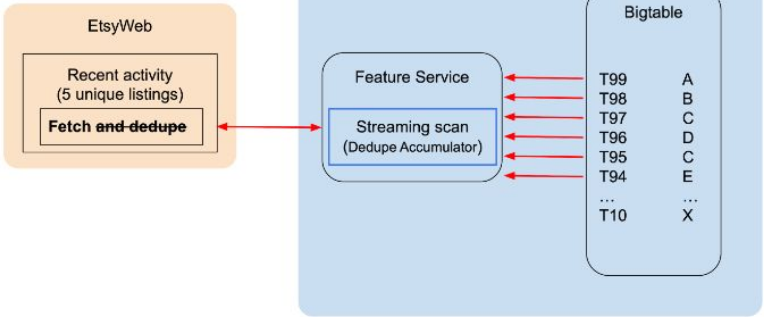
Rivulet Feature Workflow

Online retrieval - Feature Service APIs

Accumulators



Fixed-bound Scan



Streaming Scan

Rivulet Feature Workflow

Online retrieval - Feature Service APIs

Batch Requests

POST: /api/v1/feature/batch-lookup	

Request Json Object	{
	"featureIdentifiers": [
	{
	"name": "string", "version": 0, "key": "string",
	"numResults": 0,
	"minTimestampMillis": 0, "minTimestampIso": "string",
	"maxTimestampMillis": 0, "maxTimestampIso": "string",
	"accumulator": "string", "accumulatorParamsStr": "string"
	}
	]
	}

Response JSON Object	[{"name": "string", "version": 0, "key": "string", "value": {}}]
Status Codes	• 404: Nothing was found for any of the lookups requested

Challenges and Ideas

Challenges and Ideas

Difficulty adding features	A config based Feature Registry with automatic validation and deployment.
Service API Performance	Redesign the service to handle batch request more efficiently. Optimize the service API latency with other framework (e.g. gRPC).
Code and Maintenance Complexity	Redesign the ingestion pipelines to remove unnecessary abstractions

Thank you! Questions?

And many thanks to the amazing documentation and knowledge sharing from

- Kevin McHale, Principal Engineer, Etsy
- Nick Sawyer, Staff Software Engineer, Etsy
- Lucia Yu, Senior Applied Scientist, Etsy
- Sheila Hu, Staff Software Engineer, Etsy