Welcome

BEAM SUMMIT Austin, 2022

Welcome to Beam Summit

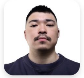
Pablo Estrada & Danielle Syse

BEAM SUMMIT Austin, 2022

Going since 2018...

Thanks to our speakers!

Adam Najman
Software Engineer at BlueVoyant


Alex Chan
Trustpilot, Senior Machine Learning Engineer

Alex Van Boxel
Alex Van Boxel, Principal Architect at Collibra

Alexander Falca
Adobe, Project Lead

Chandan Prasad
Sr. SDE @ Amazon

Charles Adetiloye
MLOps Engineer @MavenCode

Claire McGinty
Data Engineer @ Spotify

Constantin Scacun
Senior Software Engineer at Adobe

Kamaljeet Singh
Lead Architect - Wipro

Kenn Knowles
ASF member & Chair of Apache Beam PMC

Kerry Donny-Clark
Manager Apache Beam @ Google Cloud

Konstantin Buschmeier
Machine Learning Engineer at ML6

Nilabhra Roy Chowdhury
NLP Engineer @ Carted

Omkar Deshpande
Software Engineer at Intuit

Pablo Estrada
Software Engineer, Google. PMC member, Apache Beam.

Parag Ghosh
Data Architect

Shailesh Mangal
VP Engineering

Shangjin Zhang
Staff Software Engineer at LinkedIn

Subham Rakshit
Specialist Solution Architect, AWS

Svetak Sundhar
Technical Solutions Engineer, Google

Alexander Lerma
MLOps Engineer @MavenCode

Alexey Romanenko
Principal Software Engineer in Talend

Alfredo Gimenez
Data Engineering Manager at BlueVoyant

Amar Surjit
Amazon AWS, Sr. Solution Architect

Danny McCormick
Software Engineer @ Google

Daria Malkova
Senior Software Engineer at Akvelon

Debasish Das
Sr Manager, Machine Learning, Credit Karma

Devon Peticolas
Principal Engineer at Oden Technologies

Lak Lakshmanan
Operating Executive at Silver Lake

Le Chang
Sr Software Engineer at Google

Lohit Vijayarenu
Software Engineer at Twitter

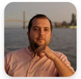
Marco Robles
Software Engineer at Wizeline

Patrick Ames
Sr. SDE @ Amazon

Paul Balm
Strategic Cloud Engineer, Google Cloud

Peter Coyle
Head of Risk Technology Engineering Excellence , HSBC

Pragalbh Srivastava
Sr. Manager at Wayfair


Talat Uyarer
Senior Principal Software Engineer at Palo Alto Networks

Thomas Sauvagnat
Data Engineer @ Orange

Tobias Kaymak
Senior Data Engineer, Ricardo

Tucker Leavitt
Data Engineer at BlueBoyant

Amruta Deshmukh
Amruta Deshmukh - Senior Software Engineer at Strivr

Andrew Pilloud
Apache Beam Committer, Software Engineer at Google

Andy Ye
Software Engineer at Google

Angus Neilson
Trustpilot, Senior Data Engineer

Dunja Panic
Product Manager, Intuit - AI+Data

Haikuo Liu
Google, Software Engineer

Israel Herraiz
Strategic Cloud Engineer, Google Cloud

Iñigo San Jose Visiers
Google, Software Engineer

Matt Casters
Neo4j Chief Solutions Architect, Hop project lead and Kettle project founder

Mazlum Tosun
Tech lead GCP and data / FP and GCP enthusiast / Craftsman

Michel Davit
Data Engineer @ Spotify

Miren Esnaola
Google, Cloud Consultant

Praveen Killamsetti
Staff Engineer at Twitter

Ragy Abraham
Co-Founder and CTO at Monita

Raj Subramani
Flumaion

Rickard Zwahlen
Data Engineer at Spotify

Tyler Flach
Bluevoyant software engineer

Vishnu Venkataraman
VP of Data Science and Engineering at Credit Karma

Wei Hsia
Developer Advocate At Google Cloud

Yuhong Cheng
Software Engineer at Linkedin

Annica Ivert
Data Engineer @ Spotify

Austin Bennett
Principal at Bamboozoology

Beam Summit Team

Beam Summit Team

Brian Hulette
Software Engineer at Google,

Jack McCluskey
Software Engineer @ Google

Jan Lukavský
Freelance data engineer, Apache Beam committer

Jiajun Yao
Anyscale Software Engineer

Jérémie Gomez
Data Cloud Consultant @ Google

Nagaraja Tantry
Principal Engineer at Intuit

Nancy Xu
SWE at Google

Navin Viswanath
Staff Software Engineer at Twitter

Nick Hwang
Engineering Manager, Stream Processing Platform

Robert Burke
Distributed Gopher Wrangler

Roy Arsan
Solutions Architect @ Google Cloud

Sayak Paul
Machine Learning Engineer at Carted

Sergei Lilichenko
Solutions Architect, Google Cloud

Zeeshan
Cloud Data Engineer

Zhenzhao Wang
Staff SWE Twitter

Thanks to our Sponsors

Google Cloud

GOLD

Spotify

aws

MAVENCODE

SILVER

Adobe

Thanks to our partners

PARTNERS

THE APACHE SOFTWARE FOUNDATION

AKVELON
We make BIG software happen

Packt

MANNING

monita

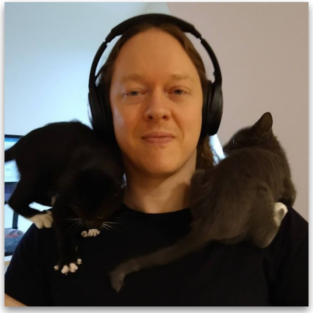
AICamp

D Data Council

WOMEN WHO CODE

ChickTECH

KEYNOTE SESSIONS



Kerry Donny-Clark

Manager Google Cloud
Dataflow

GOOGLE'S INVESTMENT ON BEAM AND ITS INTERNAL USE

10:00 - 10:25 AM

Lak Lakshmanan

Operating Executive
Silver Lake

MACHINE LEARNING DESIGN PATTERNS: BETWEEN BEAM AND A HARD PLACE

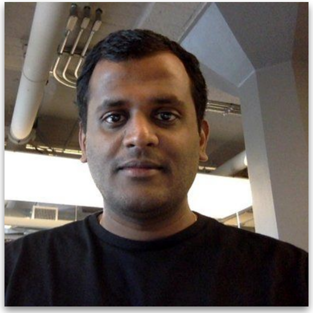
10:25 - 10:50 AM

Rickard Zwahlen

Data Engineer
Spotify

TAILORING PIPELINES AT SPOTIFY

10:50 - 11:15 AM



Lohit Vijayarenu

Principal Software Engineer
Twitter

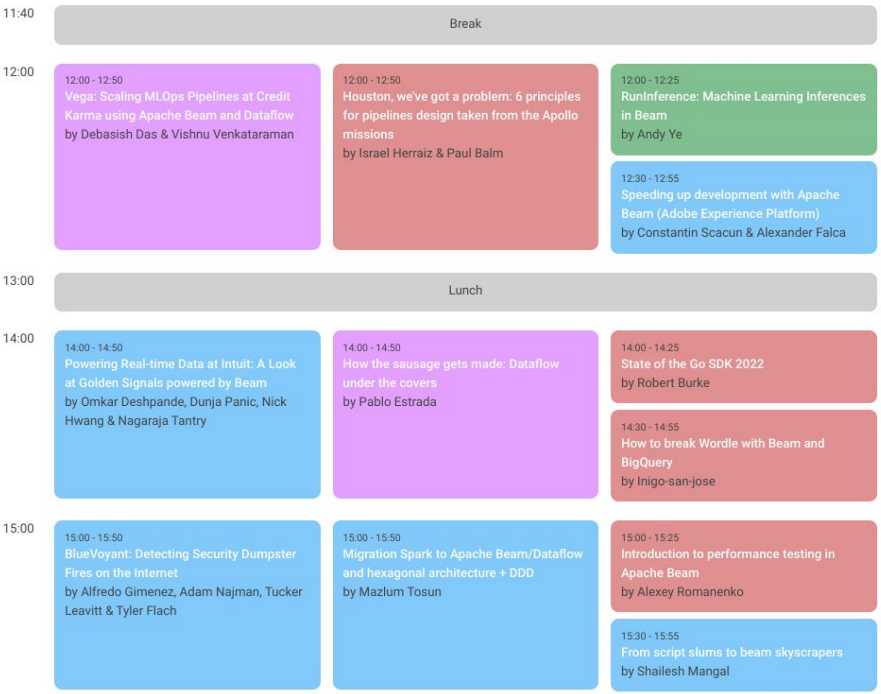
THE ADOPTION, CURRENT STATE, AND FUTURE OF APACHE BEAM

11:15 - 11:40 AM

Before anything..!

Please, PLEASE fill our survey:

Monday Schedule



11:40 Break

12:00

- 12:00 - 12:50 Vega: Scaling MLOps Pipelines at Credit Karma using Apache Beam and Dataflow by Debasish Das & Vishnu Venkataraman
- 12:00 - 12:50 Houston, we've got a problem: 6 principles for pipelines design taken from the Apollo missions by Israel Herraiz & Paul Balm
- 12:00 - 12:25 RunInference: Machine Learning Inferences in Beam by Andy Ye
- 12:30 - 12:55 Speeding up development with Apache Beam (Adobe Experience Platform) by Constantin Scacun & Alexander Falca

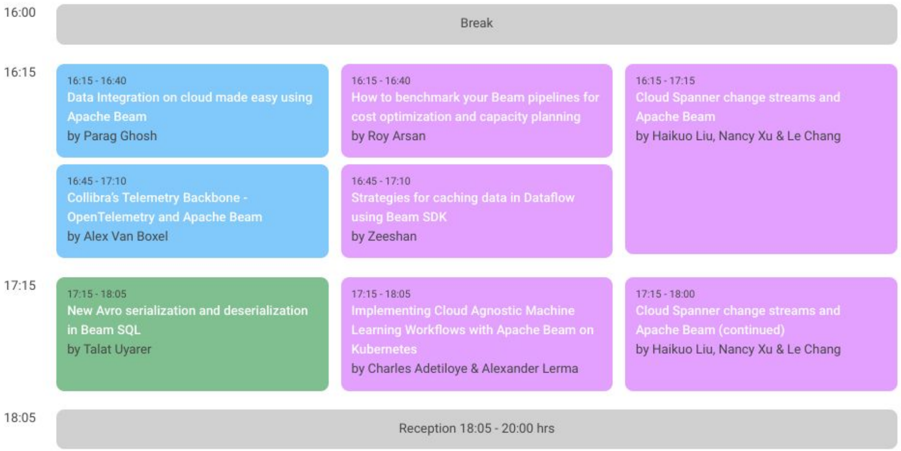
13:00 Lunch

14:00

- 14:00 - 14:50 Powering Real-time Data at Intuit: A Look at Golden Signals powered by Beam by Omkar Deshpande, Dunja Panic, Nick Hwang & Nagaraja Tantry
- 14:00 - 14:50 How the sausage gets made: Dataflow under the covers by Pablo Estrada
- 14:00 - 14:25 State of the Go SDK 2022 by Robert Burke
- 14:30 - 14:55 How to break Wordle with Beam and BigQuery by Inigo-san-jose

15:00

- 15:00 - 15:50 BlueVoyant: Detecting Security Dumpster Fires on the Internet by Alfredo Gimenez, Adam Najman, Tucker Leavitt & Tyler Flach
- 15:00 - 15:50 Migration Spark to Apache Beam/Dataflow and hexagonal architecture + DDD by Mazlum Tosun
- 15:00 - 15:25 Introduction to performance testing in Apache Beam by Alexey Romanenko
- 15:30 - 15:55 From script slums to beam skyscrapers by Shailesh Mangal



16:00 Break

16:15

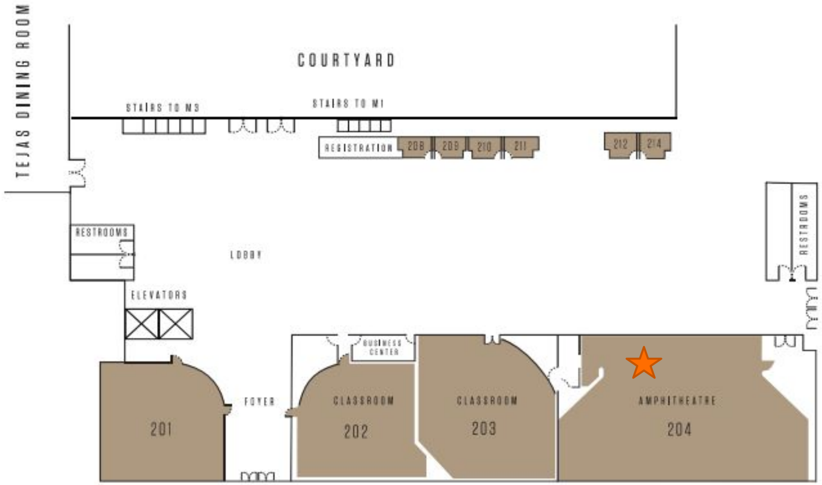
- 16:15 - 16:40 Data Integration on cloud made easy using Apache Beam by Parag Ghosh
- 16:15 - 16:40 How to benchmark your Beam pipelines for cost optimization and capacity planning by Roy Arsan
- 16:15 - 17:15 Cloud Spanner change streams and Apache Beam by Haikuo Liu, Nancy Xu & Le Chang
- 16:45 - 17:10 Collibra's Telemetry Backbone - OpenTelemetry and Apache Beam by Alex Van Boxel
- 16:45 - 17:10 Strategies for caching data in Dataflow using Beam SDK by Zeeshan

17:15

- 17:15 - 18:05 New Avro serialization and deserialization in Beam SQL by Talat Uyarer
- 17:15 - 18:05 Implementing Cloud Agnostic Machine Learning Workflows with Apache Beam on Kubernetes by Charles Adetiloye & Alexander Lerma
- 17:15 - 18:00 Cloud Spanner change streams and Apache Beam (continued) by Haikuo Liu, Nancy Xu & Le Chang

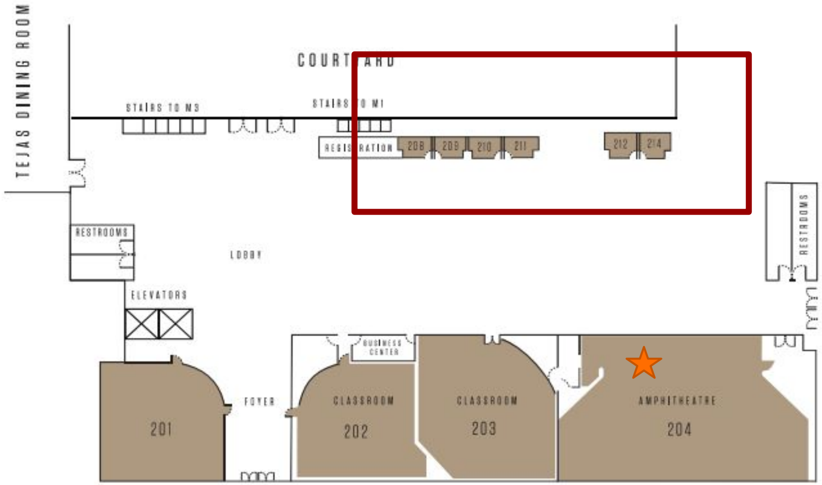
18:05 Reception 18:05 - 20:00 hrs

Where to Go Next?



- All sessions will take place on this floor across 202-204
    - Keynotes will be held in the Amphitheatre only
- Lunch will take place from 1-2 PM in the Tejas Dining Room
    - Lunch box options include roast beef on ciabatta, chicken salad croissant and falafel fritter wraps
- Session rooms will be noted outside each door as well on each calendar invite/Beam Summit page
- Restrooms located at each end of the hall with elevators to our left next to the Dining Room

About the space…



- We have rooms with whiteboards across the hall. Feel free to use the whiteboards.
  - We also have easel pads in the presentation rooms. Feel free to use in technical convos.

BEAM SUMMIT

Networking Opportunities

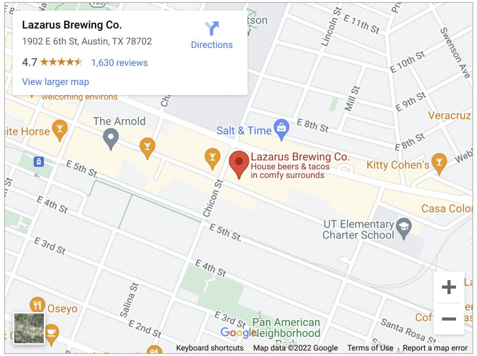
Please join us for networking opportunities while you’re with us:

Reception tonight!

Join us for drinks after the event from **6:00 - 8:00 pm** at the **AT&T Conference Center Courtyard.**

After Party Tuesday

Tuesday at 6:30pm at Lazarus Brewing Co, where beer on the house will be waiting for you! Send the directions to your phone by scanning the following QR code.



Lazarus Brewing Co.
1902 E 6th St, Austin, TX 78702
4.7 ★★★★★ 1,630 reviews
View larger map
Directions

Networking Opportunities

Job Openings

Reminder to take a look at the current job openings gathered by our sponsors:

Speakers!

- Please arrive **a little early** to your room for setting up

Thank You

Google's investments in Beam

By Dr. Kerry Donny-Clark,
Google Engineering Manager for Beam

BEAM SUMMIT Austin, 2022

Hello!
I’m Kerry

Me at work

Me at home

My old job

Apache Beam

- Unified Model:
  - Batch and Streaming
- Many SDKs
  - Java, Python, Go, TS*
- Portability
  - Dataflow, Flink, Spark, Hazelcast, Ray*, Dask*, etc

*Experimental or in progress

Apache Beam used in Google

DeepMind

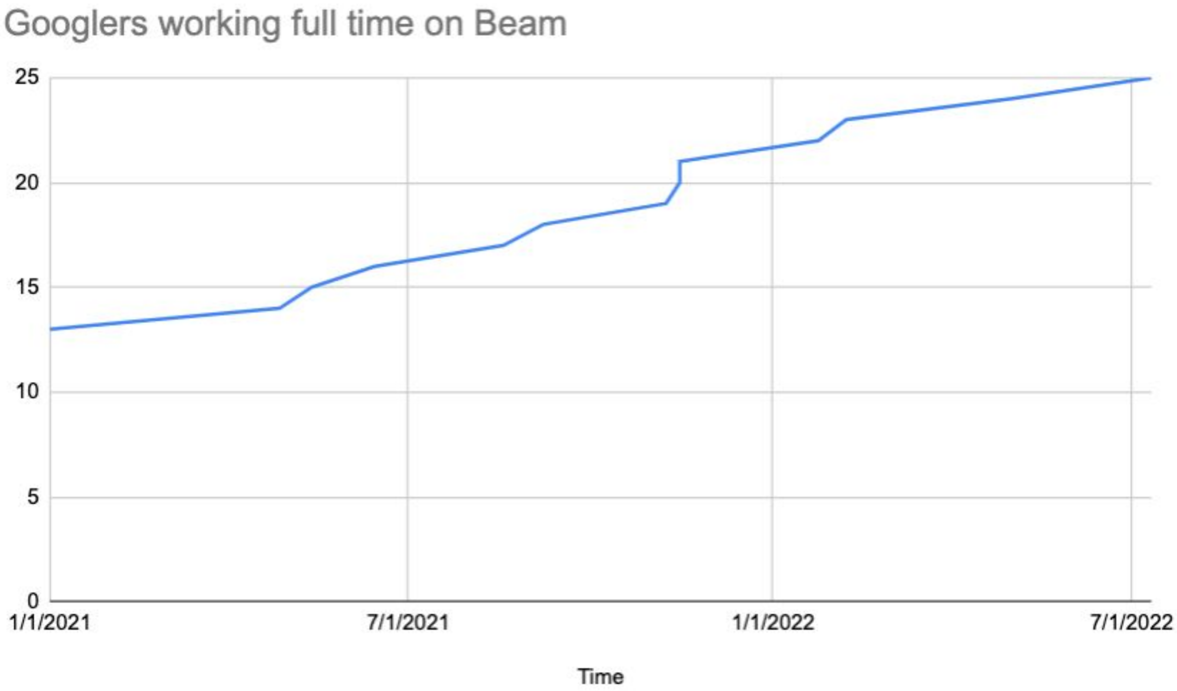
vertex.ai

Dataplex

YouTube

waze

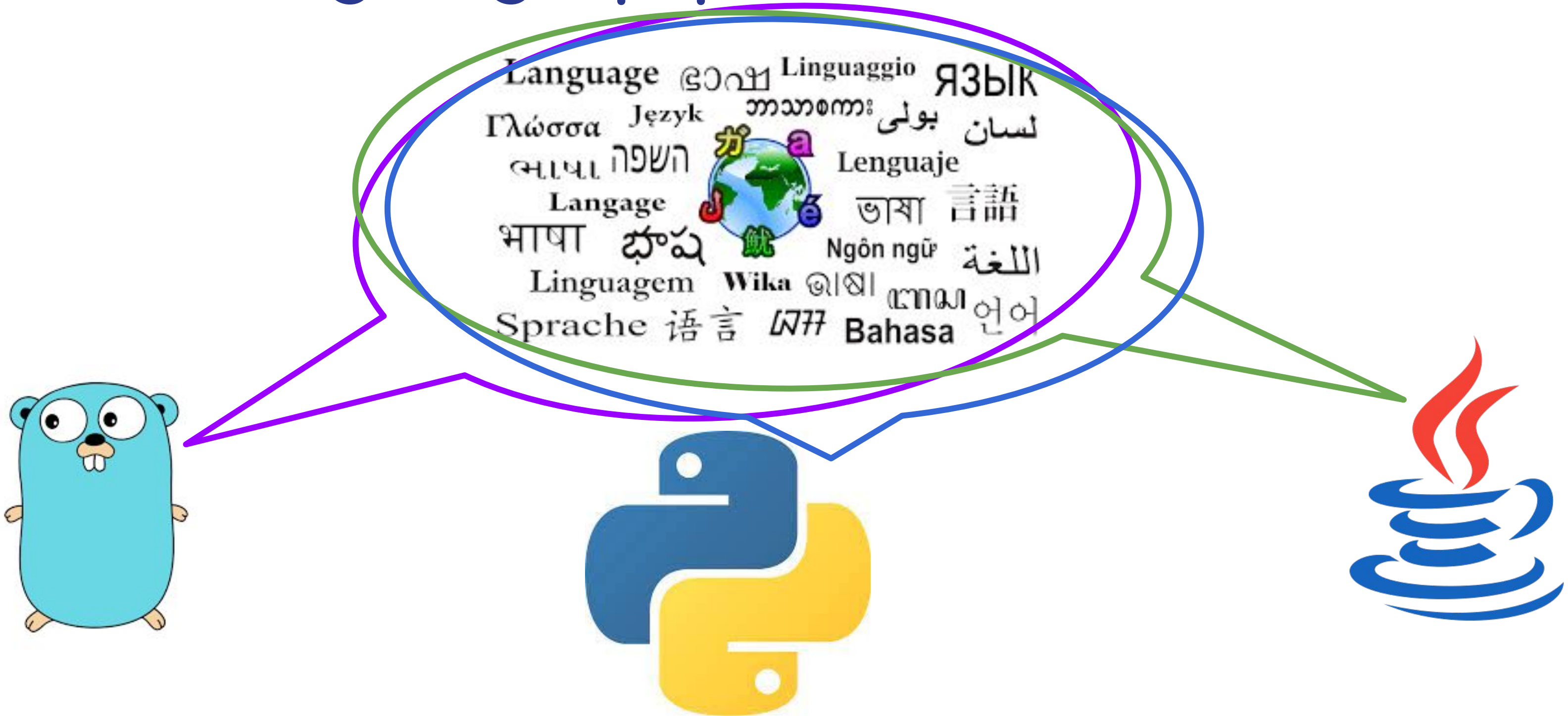
The Beam Team at Google



The Beam Team at Google

And more!

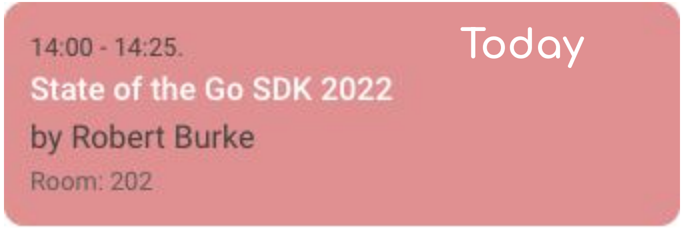
Beam going in new directions: Multi Language pipelines



A Rebus Riddle

Beam 2.40, Dataflow GA 7/20

Beam Going in new directions



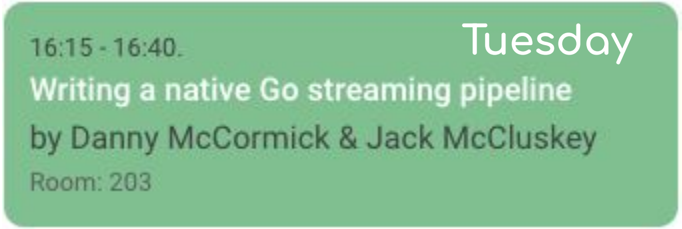
Today

14:00 - 14:25.

State of the Go SDK 2022

by Robert Burke

Room: 202



Tuesday

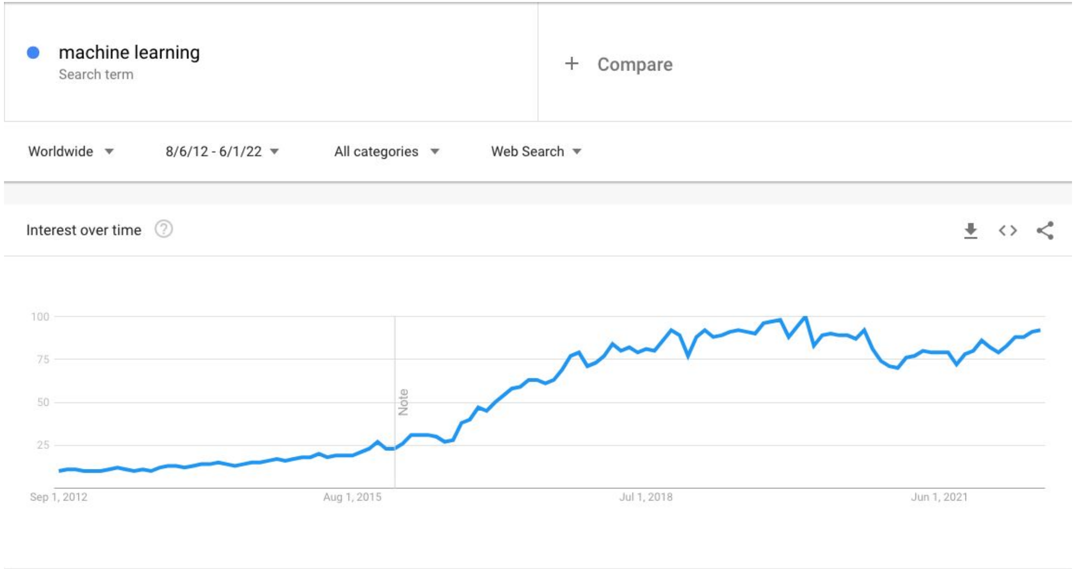
16:15 - 16:40.

Writing a native Go streaming pipeline

by Danny McCormick & Jack McCluskey

Room: 203

Beam going in new directions: RunInference in Beam Python



machine learning
Search term

Worldwide ▾ 8/6/12 - 6/1/22 ▾ All categories ▾ Web Search ▾

Interest over time

Beam going in new directions: RunInference in Beam Python

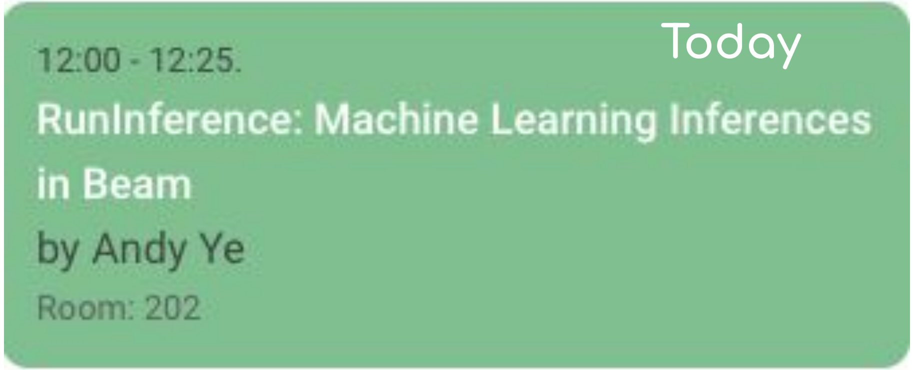
PYTORCH

TensorRT

TensorFlow

scikit learn

Beam going in new directions: RunInference in Beam Python



Today

12:00 - 12:25.

RunInference: Machine Learning Inferences in Beam

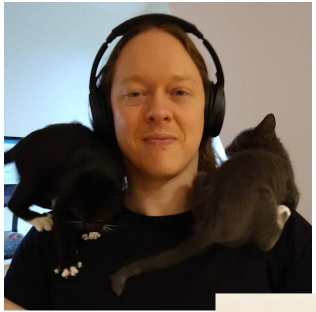
by Andy Ye

Room: 202

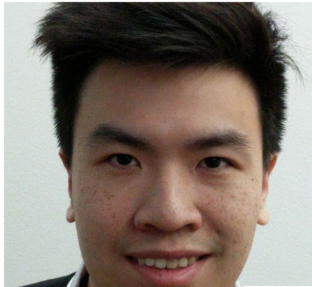
RunInference in Beam 2.40, GA on Dataflow 7/20

https://beam.apache.org/documentation/sdks/python-machine-learning/

Beam going in new directions: Typescript SDK

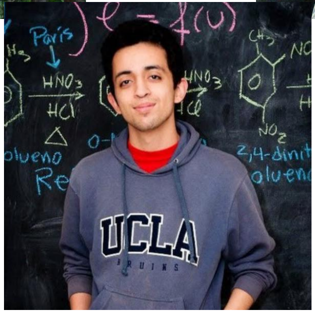

Kerry Donny-Clark

Jack McCluskey


Jonathan Lui

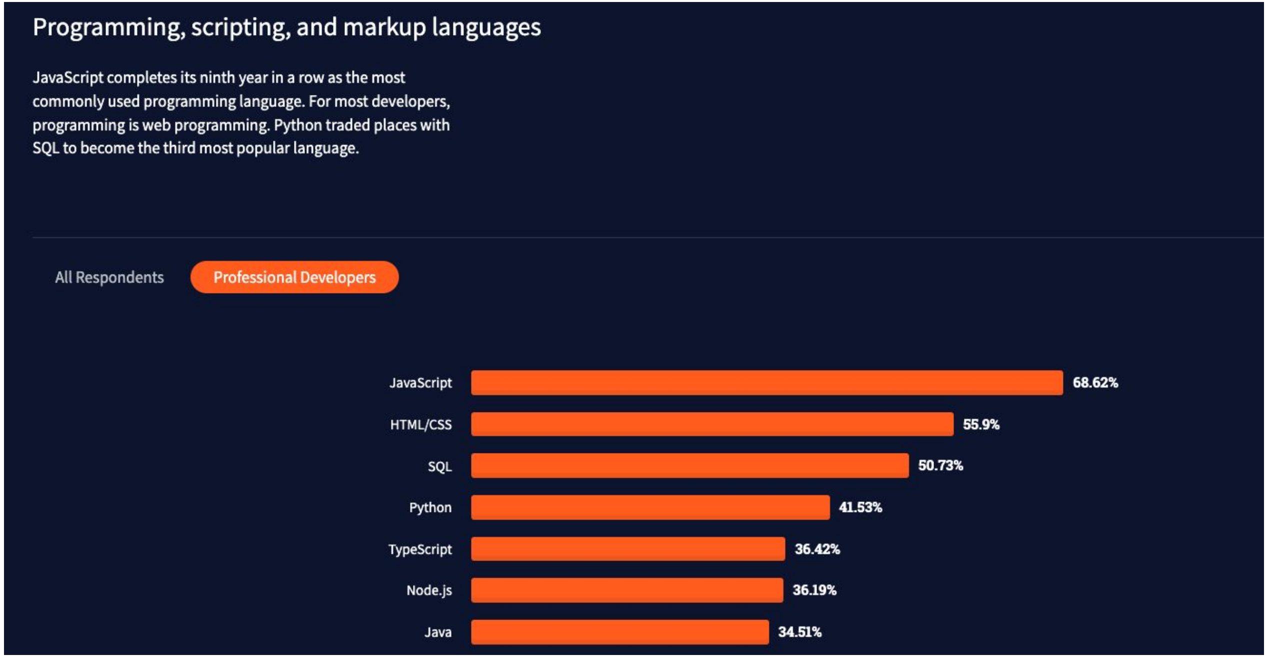
Kenneth Knowles

Kevin Puthusseri


Pablo Estrada

Robert Bradshaw

Beam going in new directions: Typescript SDK



Programming, scripting, and markup languages

JavaScript completes its ninth year in a row as the most commonly used programming language. For most developers, programming is web programming. Python traded places with SQL to become the third most popular language.

JavaScript	68.62%
HTML/CSS	55.9%
SQL	50.73%
Python	41.53%
TypeScript	36.42%
Node.js	36.19%
Java	34.51%

Beam going in new directions: Typescript SDK

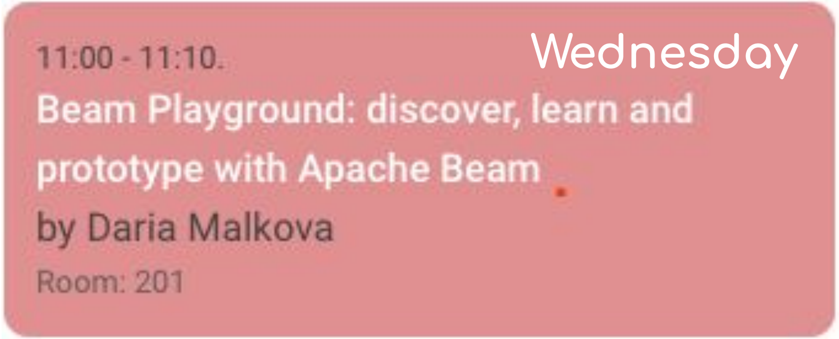
… to contribute!

https://github.com/apache/beam/tree/master/sdks/typescript

A better way to learn Beam: Beam Playground

https://play.beam.apache.org/

A better way to learn Beam: Beam Playground



11:00 - 11:10.

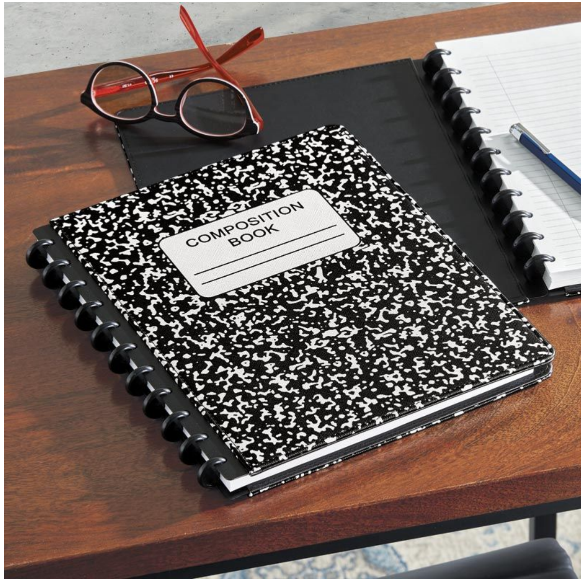
Wednesday

Beam Playground: discover, learn and prototype with Apache Beam

by Daria Malkova

Room: 201

A better way to learn Beam: Cloud notebooks



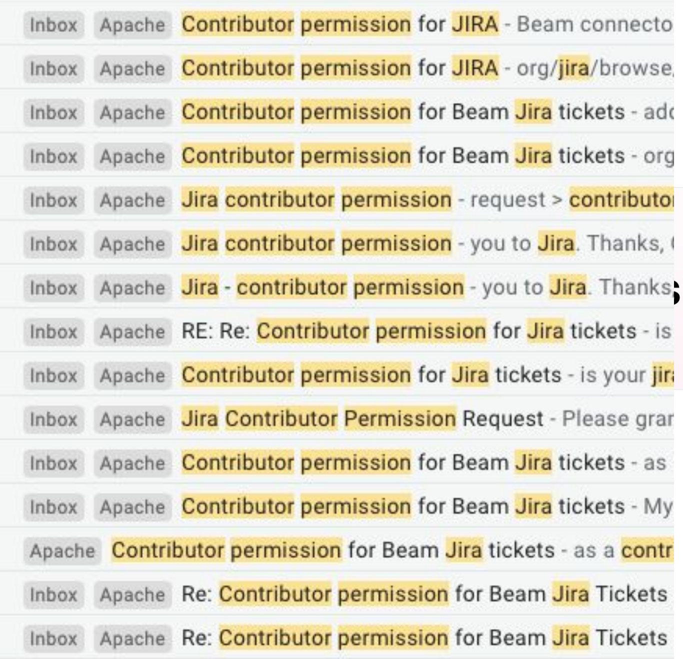
https://cloud.google.com/dataflow/docs/guides/interactive-pipeline-development

A better way to learn Beam:
A Tour of Beam

Apache Beam

Coming in late 2022!

Contributing to Beam has never been easier: Github Issues



Inbox Apache Contributor permission for JIRA - Beam connecto

Inbox Apache Contributor permission for JIRA - org/jira/browse

Inbox Apache Contributor permission for Beam Jira tickets - adc

Inbox Apache Contributor permission for Beam Jira tickets - org

Inbox Apache Jira contributor permission - request > contributor

Inbox Apache Jira contributor permission - you to Jira. Thanks,

Inbox Apache Jira - contributor permission - you to Jira. Thanks

Inbox Apache RE: Re: Contributor permission for Jira tickets - is

Inbox Apache Contributor permission for Jira tickets - is your jira

Inbox Apache Jira Contributor Permission Request - Please gran

Inbox Apache Contributor permission for Beam Jira tickets - as

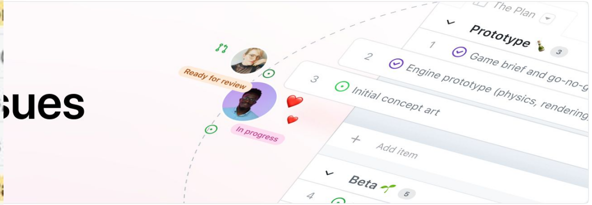
Inbox Apache Contributor permission for Beam Jira tickets - My

Apache Contributor permission for Beam Jira tickets - as a contr

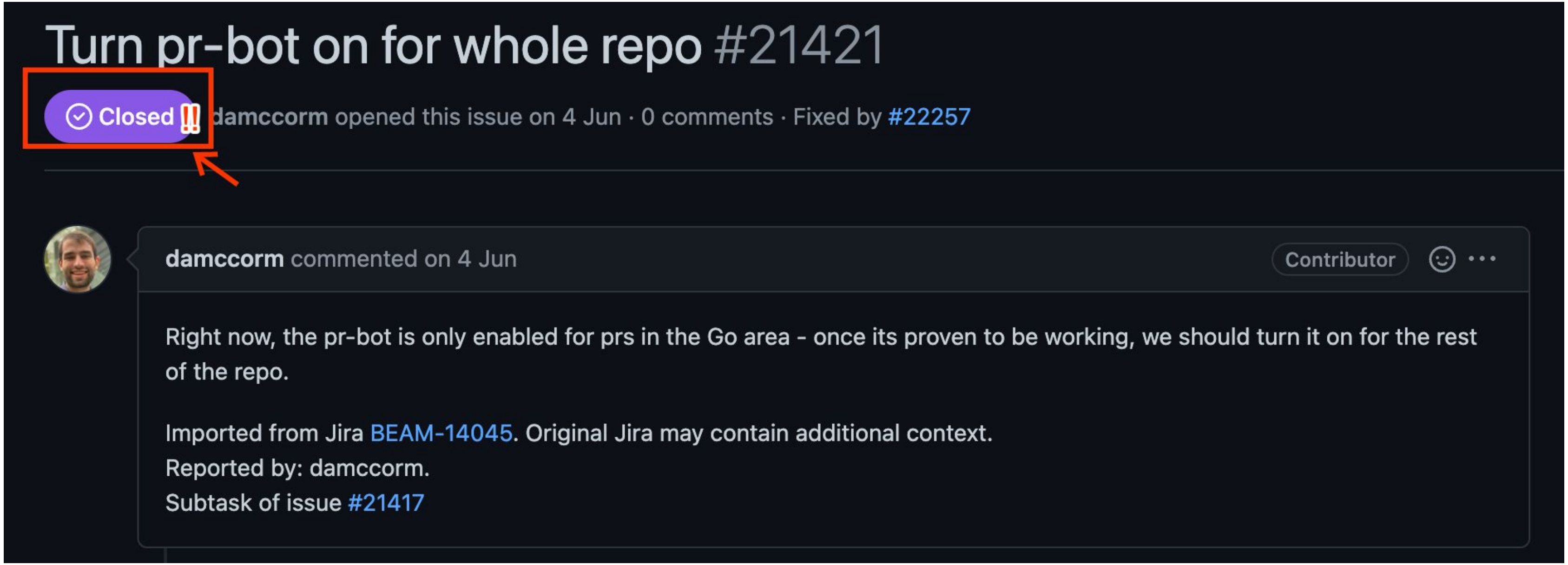
Inbox Apache Re: Contributor permission for Beam Jira Tickets

Inbox Apache Re: Contributor permission for Beam Jira Tickets

sues



Contributing to Beam has never been easier: PR-bot



Turn pr-bot on for whole repo #21421

damccorm opened this issue on 4 Jun · 0 comments · Fixed by #22257

damccorm commented on 4 Jun

Right now, the pr-bot is only enabled for prs in the Go area - once its proven to be working, we should turn it on for the rest of the repo.

Imported from Jira BEAM-14045. Original Jira may contain additional context.
Reported by: damccorm.
Subtask of issue #21417

Conclusion

Beam is growing

- Multi Language
- Beam Go SDK
- RunInference in Python
- TypeScript SDK

Learn Beam

- Beam Playground
- Beam Notebooks
- A Tour of Beam

Contribute to Beam

- Github Issues
- PR-bot

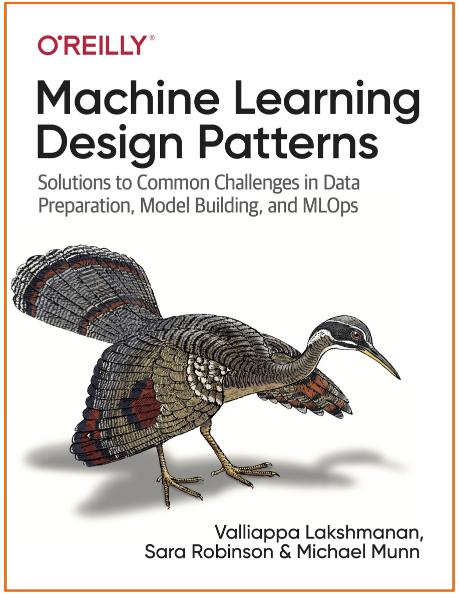
Machine Learning Design Patterns: Between Beam and a Hard Place

Lak Lakshmanan

@lak_luster

BEAM SUMMIT Austin, 2022

Formalized best practices to solve common problems

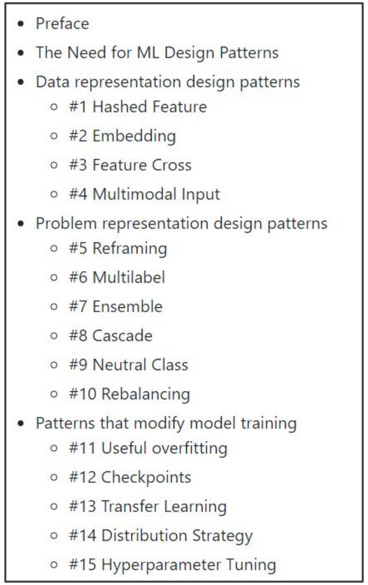


O'REILLY®

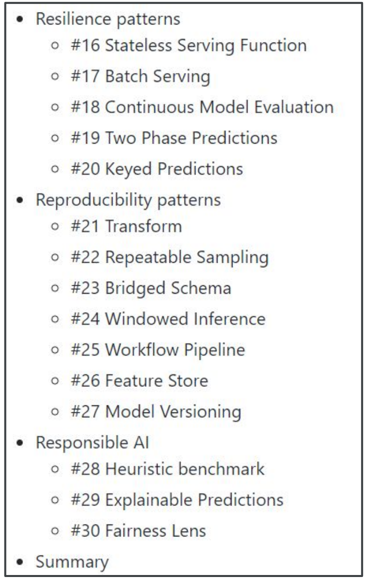
Machine Learning Design Patterns

Solutions to Common Challenges in Data Preparation, Model Building, and MLOps

Valliappa Lakshmanan, Sara Robinson & Michael Munn

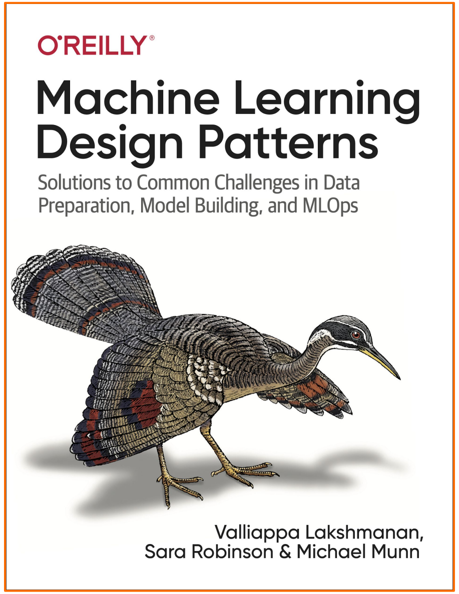


- Preface
- The Need for ML Design Patterns
- Data representation design patterns
  - #1 Hashed Feature
  - #2 Embedding
  - #3 Feature Cross
  - #4 Multimodal Input
- Problem representation design patterns
  - #5 Reframing
  - #6 Multilabel
  - #7 Ensemble
  - #8 Cascade
  - #9 Neutral Class
  - #10 Rebalancing
- Patterns that modify model training
  - #11 Useful overfitting
  - #12 Checkpoints
  - #13 Transfer Learning
  - #14 Distribution Strategy
  - #15 Hyperparameter Tuning



- Resilience patterns
  - #16 Stateless Serving Function
  - #17 Batch Serving
  - #18 Continuous Model Evaluation
  - #19 Two Phase Predictions
  - #20 Keyed Predictions
- Reproducibility patterns
  - #21 Transform
  - #22 Repeatable Sampling
  - #23 Bridged Schema
  - #24 Windowed Inference
  - #25 Workflow Pipeline
  - #26 Feature Store
  - #27 Model Versioning
- Responsible AI
  - #28 Heuristic benchmark
  - #29 Explainable Predictions
  - #30 Fairness Lens
- Summary

ML flavors of the same problems that arise in all software



O'REILLY®

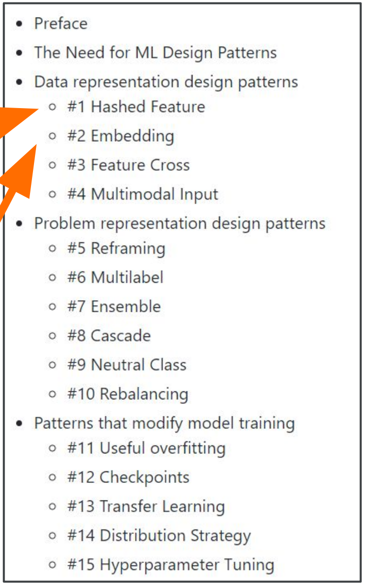
Machine Learning Design Patterns

Solutions to Common Challenges in Data Preparation, Model Building, and MLOps

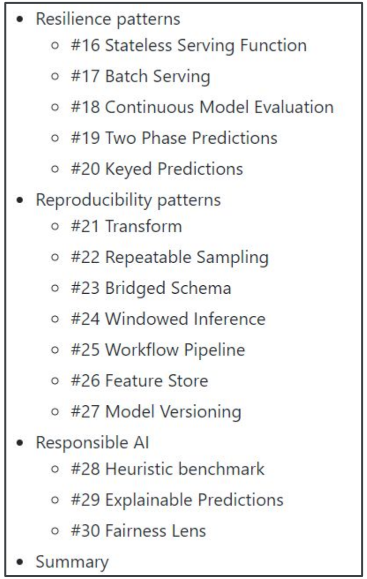
Valliappa Lakshmanan,
Sara Robinson & Michael Munn

Maintainability
How do you represent categorical data when the vocabulary increases over time?

Reusability
How do you avoid having to relearn relationships between categorical variables used in related ML problems?

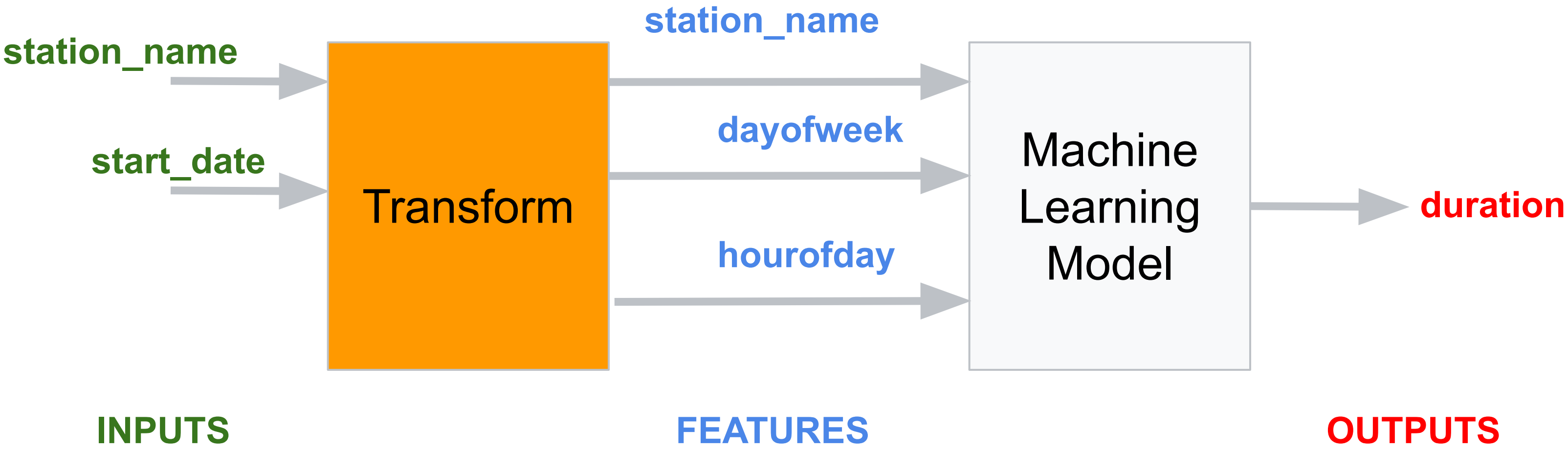


- Preface
- The Need for ML Design Patterns
- Data representation design patterns
  - #1 Hashed Feature
  - #2 Embedding
  - #3 Feature Cross
  - #4 Multimodal Input
- Problem representation design patterns
  - #5 Reframing
  - #6 Multilabel
  - #7 Ensemble
  - #8 Cascade
  - #9 Neutral Class
  - #10 Rebalancing
- Patterns that modify model training
  - #11 Useful overfitting
  - #12 Checkpoints
  - #13 Transfer Learning
  - #14 Distribution Strategy
  - #15 Hyperparameter Tuning



- Resilience patterns
  - #16 Stateless Serving Function
  - #17 Batch Serving
  - #18 Continuous Model Evaluation
  - #19 Two Phase Predictions
  - #20 Keyed Predictions
- Reproducibility patterns
  - #21 Transform
  - #22 Repeatable Sampling
  - #23 Bridged Schema
  - #24 Windowed Inference
  - #25 Workflow Pipeline
  - #26 Feature Store
  - #27 Model Versioning
- Responsible AI
  - #28 Heuristic benchmark
  - #29 Explainable Predictions
  - #30 Fairness Lens
- Summary

Beam is widely used in a few design patterns



Ideally, client code does not have to know about all the transformations that were carried out

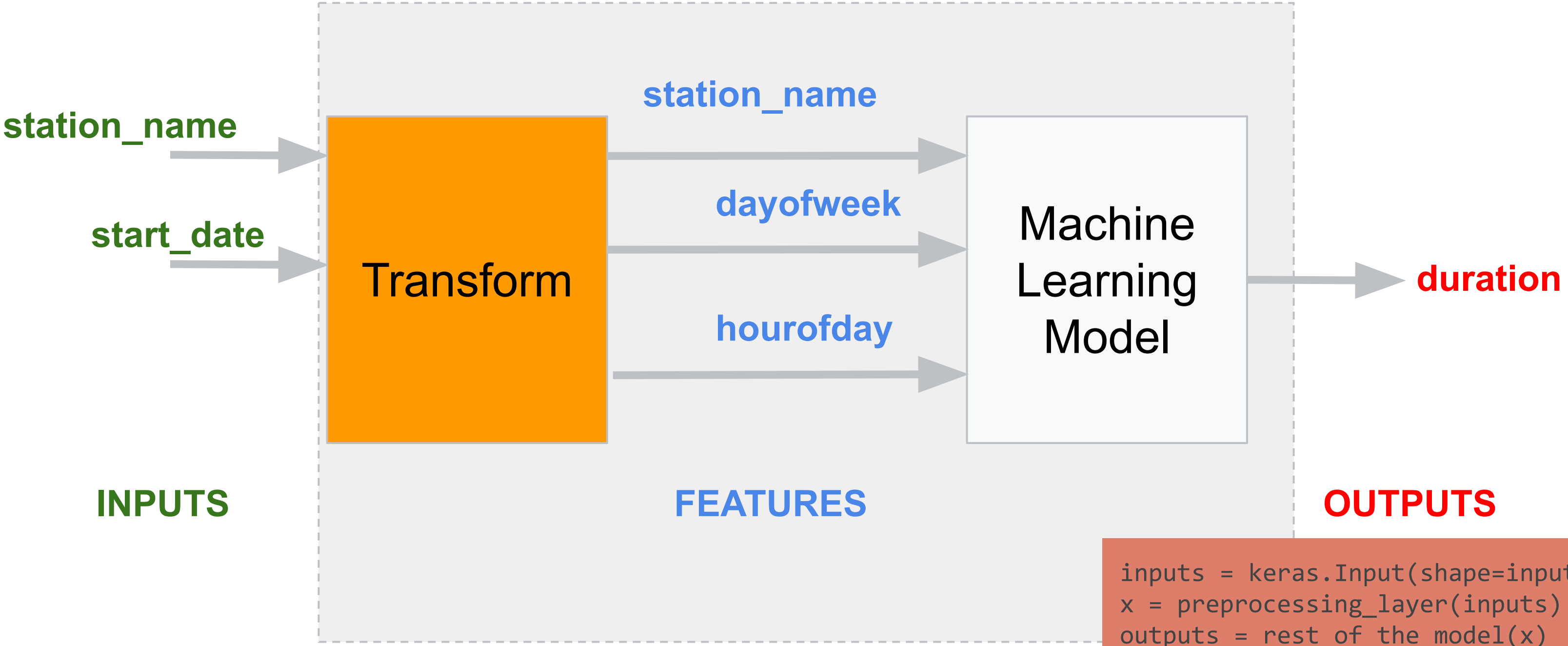
```
CREATE OR REPLACE MODEL ch09edu.bicycle_model
OPTIONS(input_label_cols=['duration'],
        model_type='linear_reg')
AS

SELECT
  duration
  , start_station_name
  , CAST(EXTRACT(dayofweek from start_date) AS STRING)
         as dayofweek
  , CAST(EXTRACT(hour from start_date) AS STRING)
         as hourofday
FROM
  `bigquery-public-data.london_bicycles.cycle_hire`
```

```
SELECT * FROM ML.PREDICT(MODEL ch09edu.bicycle_model,(
  350 AS duration
  , 'Kings Cross' AS start_station_name
  , '3' as dayofweek
  , '18' as hourofday
))
```

```
SELECT * FROM ML.PREDICT(MODEL ch09edu.bicycle_model,(
  350 AS duration
  , 'Kings Cross' AS start_station_name
  , CURRENT_TIMESTAMP() as start_date
))
```

Leading cause of training-serving skew

The Transform pattern: the model graph should include transformations



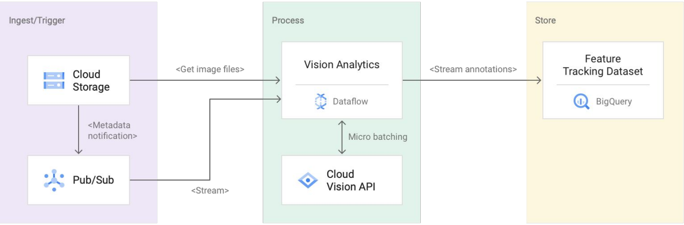
```
inputs = keras.Input(shape=input_shape)
x = preprocessing_layer(inputs)
outputs = rest_of_the_model(x)
model = keras.Model(inputs, outputs)
```

tf.transform provides reuse *and* efficiency

```
def main(output_dir):
  with tft_beam.Context(temp_dir=tempfile.mkdtemp()):
    transformed_dataset, transform_fn = (
        (raw_data, raw_data_metadata) | tft_beam.AnalyzeAndTransformDataset(
            preprocessing_fn))
  transformed_data, transformed_metadata = transformed_dataset

  # Save the transform_fn to the output_dir
  _ = (
      transform_fn
      | 'WriteTransformFn' >> tft_beam.WriteTransformFn(output_dir)

  return transformed_data, transformed_metadata
```

```
class ExportModel(tf.Module):
  def __init__(self, trained_model, input_transform):
    self.trained_model = trained_model
    self.input_transform = input_transform

  @tf.function
  def __call__(self, inputs, training=None):
    x = self.input_transform(inputs)
    return self.trained_model(x)
```

Other patterns that Beam supports well

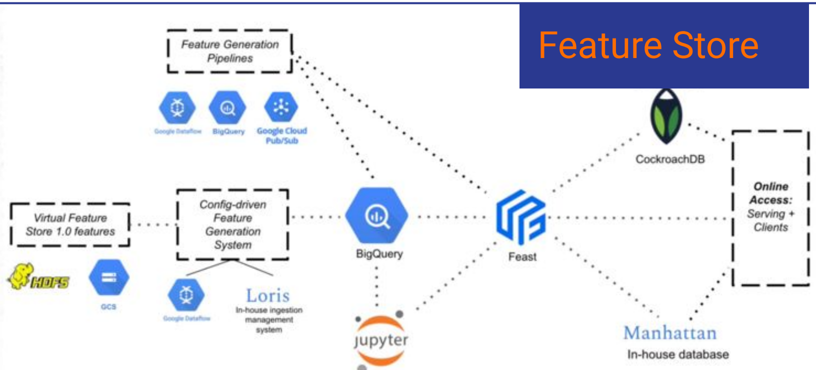


Batch Serving

```
## Time window into 2 hour windows, triggered every minute
WINDOW_INTERVAL = 2 * 60 * 60. # 2 hours, in seconds
PANE_INTERVAL = 10*60 # 10 minutes, in seconds
windowed = (data
        | 'window' >> beam.WindowInto(
                beam.window.SlidingWindows(WINDOW_INTERVAL, PANE_INTERVAL),
                accumulation_mode=beam.trigger.AccumulationMode.DISCARDING))
model_state = (windowed
        | 'model' >> beam.transforms.CombineGlobally(ModelFn()).without_defaults())
anomalies = (windowed
        | 'latest_slice' >> beam.FlatMap(is_latest_slice)
        | 'find_anomaly' >> beam.Map(is_anomaly, beam.pvalue.AsSingleton(model_state)))
```

Windowed Inference

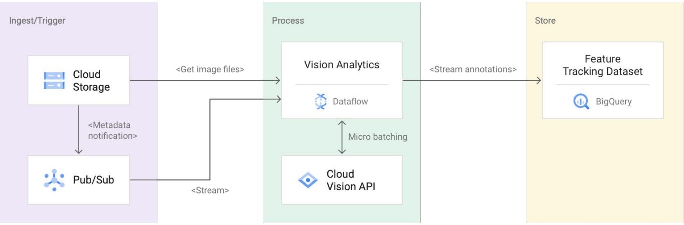
Feature Store


Feature Generation Pipelines
Google Dataflow
BigQuery
Google Cloud Pub/Sub
CockroachDB
Online Access: Serving + Clients
Virtual Feature Store 1.0 features
Config-driven Feature Generation System
BigQuery
Feast
HDFS
GCS
Google Dataflow
Loris
In-house ingestion management system
jupyter
Manhattan
In-house database

Feature Store at Twitter. Source: https://twitter.com/ttzeneva

Why?

Other patterns that Beam supports well

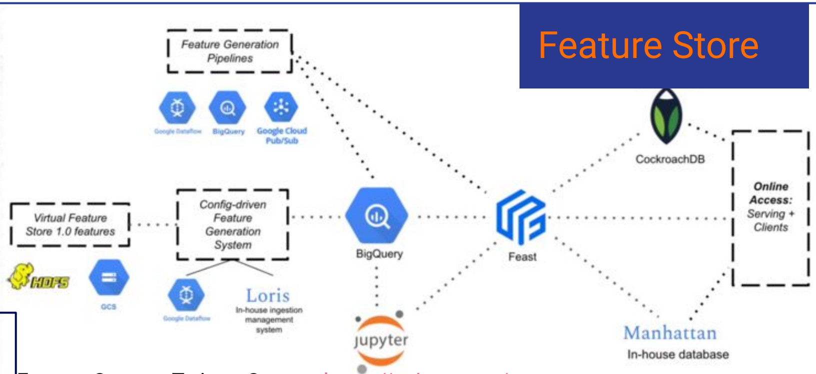


Batch Serving

```
## Time window into 2 hour windows, triggered every minute
WINDOW_INTERVAL = 2 * 60 * 60. # 2 hours, in seconds
PANE_INTERVAL = 10*60 # 10 minutes, in seconds
windowed = (data
         | 'window' >> beam.WindowInto(
                   beam.window.SlidingWindows(WINDOW_INTERVAL, PANE_INTERVAL),
                   accumulation_mode=beam.trigger.AccumulationMode.DISCARDING))
model_state = (windowed
         | 'model' >> beam.transforms.CombineGlobally(ModelFn()).without_defaults())
anomalies = (windowed
         | 'latest_slice' >> beam.FlatMap(is_latest_slice)
         | 'find_anomaly' >> beam.Map(is_anomaly, beam.pvalue.AsSingleton(model_state)))
```

Windowed Inference

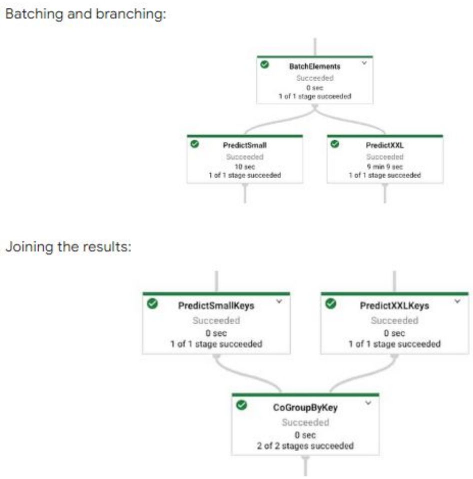
Feature Store


Feature Generation Pipelines
Google Dataflow
BigQuery
Google Cloud Pub/Sub
Virtual Feature Store 1.0 features
HDFS
GCS
Config-driven Feature Generation System
Google Dataflow
Loris
In-house ingestion management system
BigQuery
jupyter
Feast
CockroachDB
Manhattan
In-house database
Online Access: Serving + Clients

Feature Store at Twitter. Source: https://twitter.com/ttzeneva

- *Easy way to run in parallel*
- *Side inputs and windowing*
- *Unified Batch and Stream*
- *Much harder without Beam*

There are other patterns where Beam could be used, but isn’t



Batching and branching:

Joining the results:

Cascade
Transfer Learning
Continuous Evaluation
Two Phase Predictions
Multimodal Input
Workflow Pipeline

https://cloud.google.com/blog/products/data-analytics/ml-inference-in-dataflow-pipelines

What’s common to these?

Cascade
Transfer Learning
Continuous Evaluation
Two Phase Predictions
Multimodal Input
Workflow Pipeline

Why?

- *Training, evaluation: One-off, rare task*
- *Online serving: On-demand to millions*
- *Artifact Management among multiple ML models: Orchestration*

What if Beam could:

- scale from zero to millions of QPS
- consume/produce HTTP, cloud events
- be GPU-accelerated
- be run on-demand (start instantaneously)?

Imagine …

A Beam Runner that runs on Cloud Run

Portable way to run Java/Go/Python across serverless container options on AWS, GCP, Azure

Scales to zero, suitable for rare ETL, scales on-demand code

Portable ML code across training, inference, evaluation

Thank you!

BEAM SUMMIT Austin, 2022

Tailoring pipelines at Spotify

By Rickard Zwahlen

BEAM SUMMIT Austin, 2022

Rickard Zwahlen

Data engineer @ Spotify

in @rickardzwahlen

BEAM SUMMIT

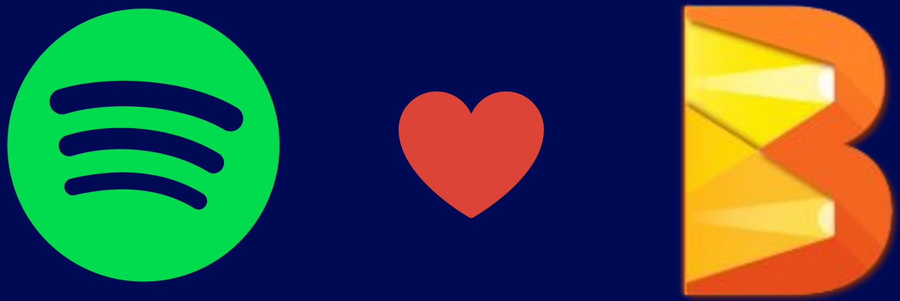
(Mel: ABBA - Super Trouper)
Super Deduper runs as fast as lightning
Handles massive skew
Gets events to you
But only one of each, not two

Rickard Zwahlen

Data engineer @ Spotify

in @rickardzwahlen

Smörgåsbord of data



How it started

How it’s going

Why Scala?

- Productivity + performance
- Functional & type safe
- Large software ecosystem for data

The love triangle

Word count

```
val sc = ScioContext()
sc
  .textFile("shakespeare.txt")
  .flatMap { _ .split("[^a-zA-Z']+")
  .filter(_.nonEmpty) }
  .countByValue
  .saveAsTextFile("wordcount.txt")

sc.run()
```

Joins

```
val sc = ScioContext()
sc
   .avroFile[Artist](args("artists"))
   .keyBy(_.getArtistId)
   .hashJoin(musicLabels)
   .map { case (artistId, (artist, label)) =>
     (
       artistId,
       businessLogic(artist, label)
     )
    }

sc.run()
```

Bread & butter pipelines

A large majority of pipelines are written in Scio

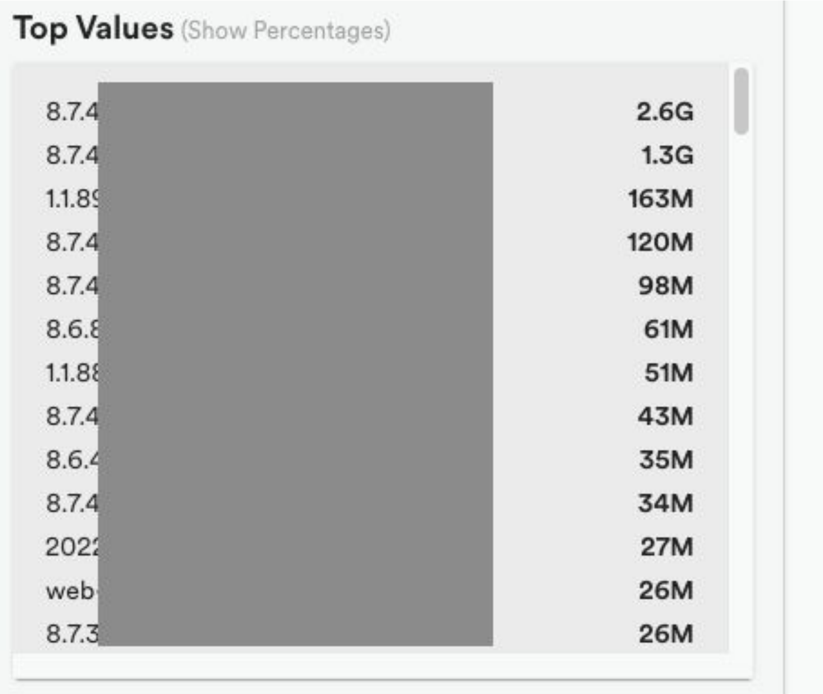
Cake mix pipelines

Just add water

Data profiling

```
schedule: hourly
docker_image: grc.io/data-profiling/1.2.3@sha256:foo
docker_args:
        - wrap-luigi
        - --module
        - luigi_tasks
        - ProfileRunner
        - --input-dataset
        - Impressions.gcs
        - --partitioning
        - hours
        - --project
        - my-cloud-project
```

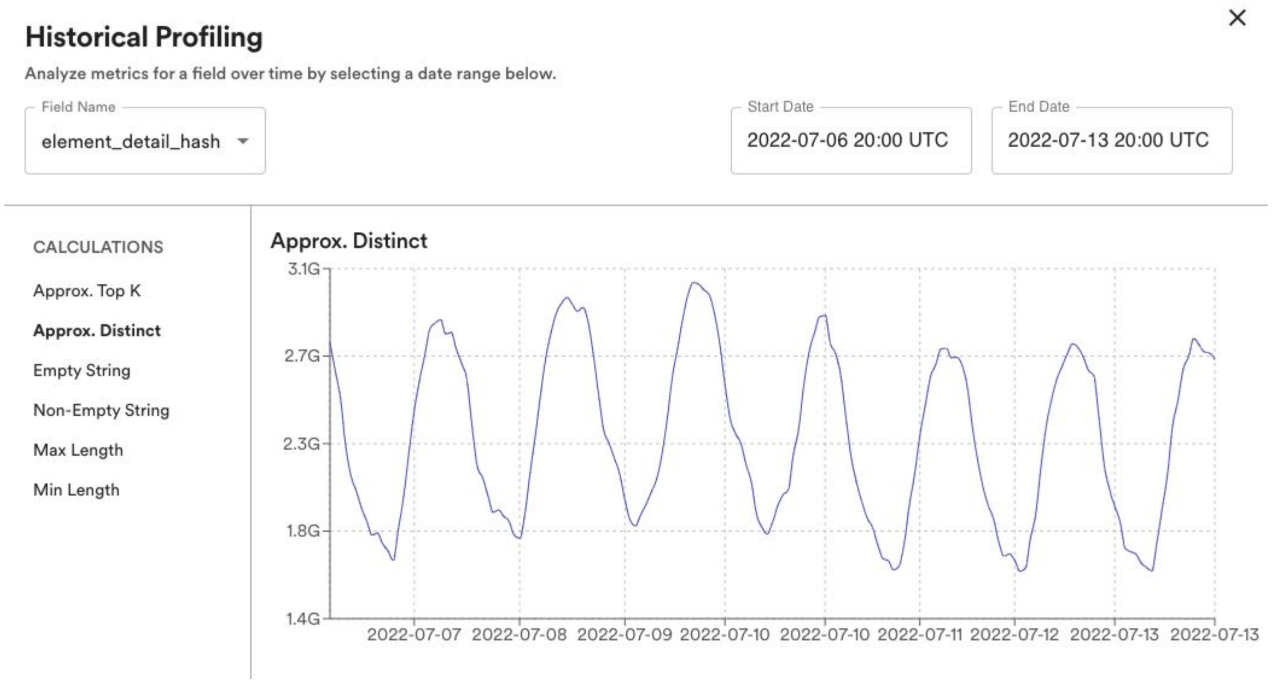
Data profiling (pt 2)



Top Values (Show Percentages)

8.7.4 [redacted]	2.6G
8.7.4 [redacted]	1.3G
1.1.8 [redacted]	163M
8.7.4 [redacted]	120M
8.7.4 [redacted]	98M
8.6. [redacted]	61M
1.1.8 [redacted]	51M
8.7.4 [redacted]	43M
8.6. [redacted]	35M
8.7.4 [redacted]	34M
202 [redacted]	27M
web [redacted]	26M
8.7.3 [redacted]	26M

Data profiling (pt 3)



Historical Profiling

Analyze metrics for a field over time by selecting a date range below.

Field Name: element_detail_hash

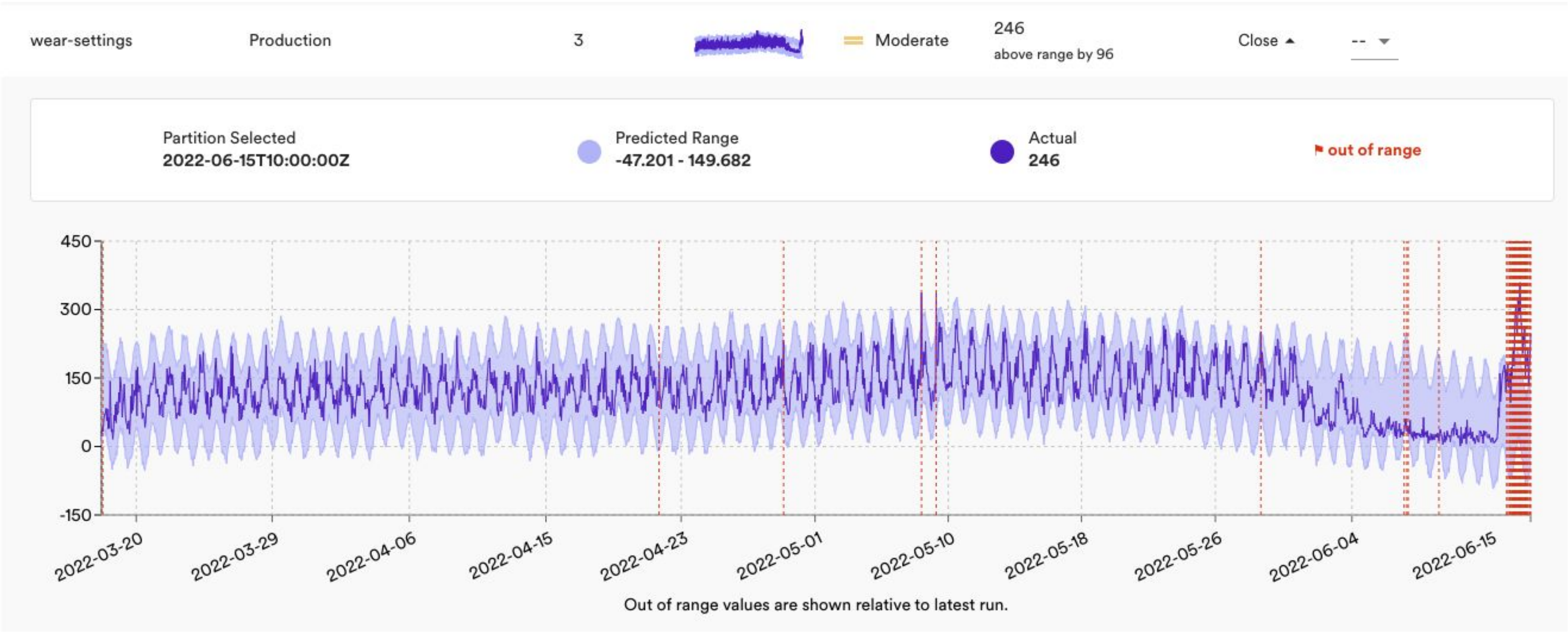
Start Date: 2022-07-06 20:00 UTC

End Date: 2022-07-13 20:00 UTC

CALCULATIONS

- Approx. Top K
- **Approx. Distinct**
- Empty String
- Non-Empty String
- Max Length
- Min Length

Anomaly detection



wear-settings | Production | 3 | ▬ Moderate | 246 above range by 96 | Close ▴ | -- ▾

Partition Selected 2022-06-15T10:00:00Z

● Predicted Range -47.201 - 149.682

● Actual 246

⚑ out of range

Out of range values are shown relative to latest run.

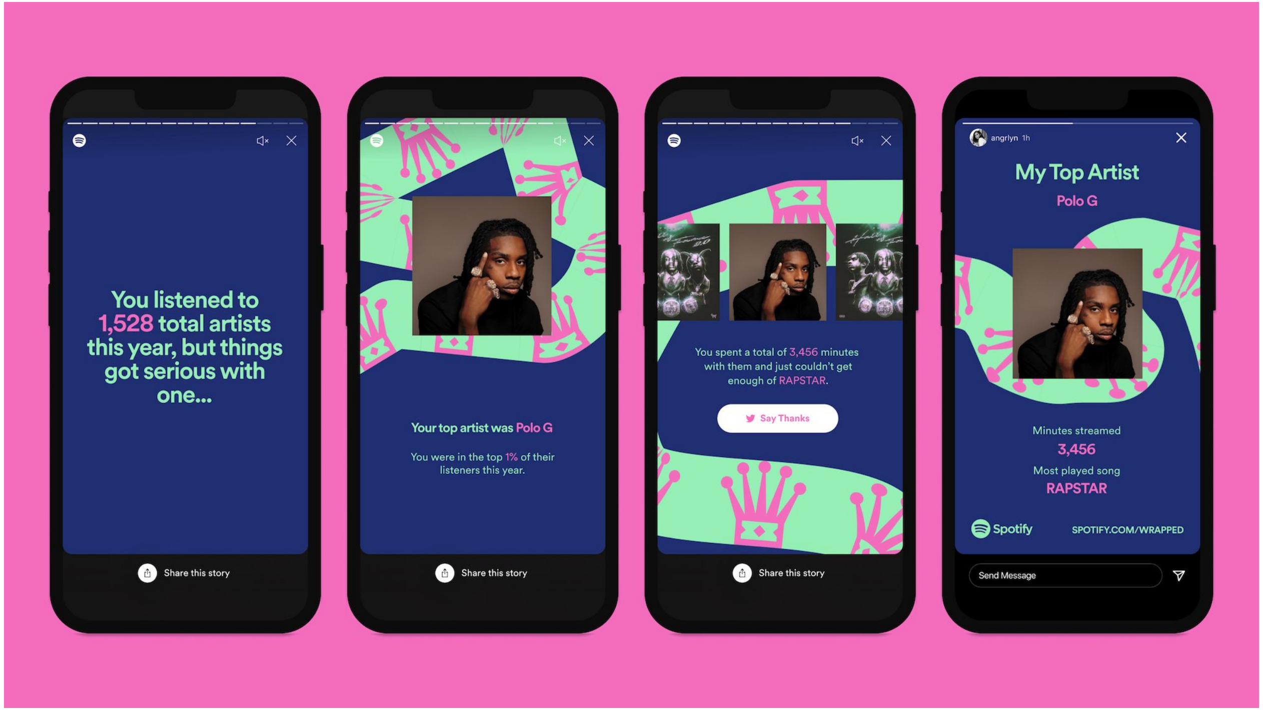
The difficult stuff

Scale, complexity, edge cases

Case by case

or

Spotify Wrapped



You listened to 1,528 total artists this year, but things got serious with one...

Share this story

Your top artist was Polo G

You were in the top 1% of their listeners this year.

Share this story

You spent a total of 3,456 minutes with them and just couldn't get enough of RAPSTAR.

Say Thanks

Share this story

angrlyn 1h

My Top Artist

Polo G

Minutes streamed

3,456

Most played song

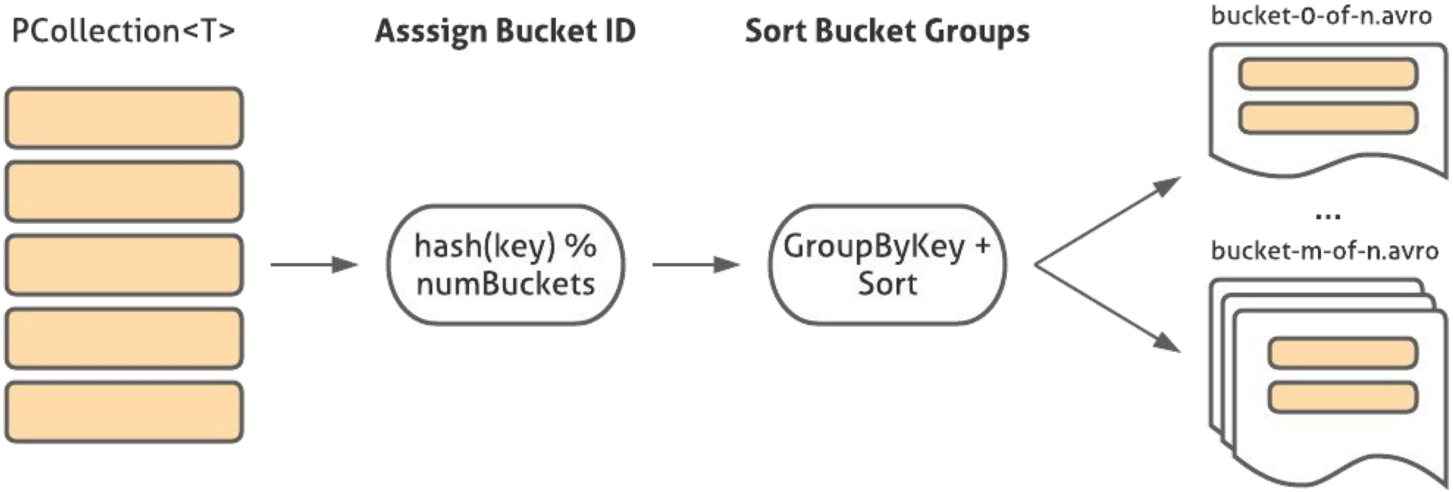
RAPSTAR

Spotify

SPOTIFY.COM/WRAPPED

Send Message

Scio + Sort Merge Bucket



TL;DR

We use Beam at the highest level of abstraction that fits the use case

- Beam SDK
- Scio Scala API
- Plug-and-play images

Thanks!

Check out the Scio workshop on Wednesday

Beam @Twitter Evaluation, Adoption, Migration and future.

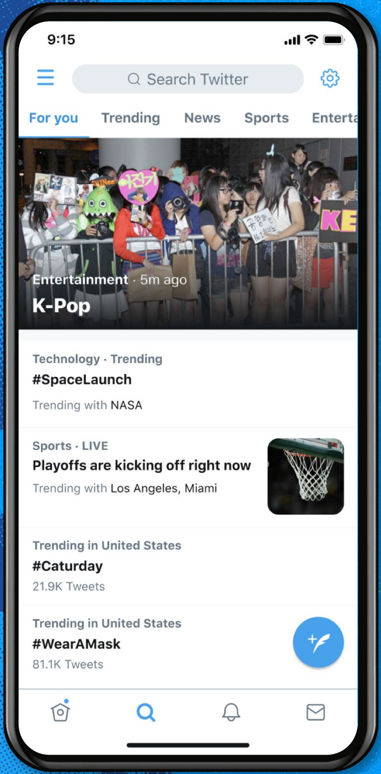
Beam : Adoption, Current state and future @Twitter

Lohit VijayaRenu
@lohitvijayarenu

BEAM SUMMIT Austin, 2022

Data Processing @Twitter

Twitter Timelines
Recommendations
Analytics
Ad products
Trends, Search, Explore
Many more... or
Everything



9:15

Search Twitter

For you | Trending | News | Sports | Enterta

Entertainment · 5m ago
K-Pop

Technology · Trending
#SpaceLaunch
Trending with NASA

Sports · LIVE
Playoffs are kicking off right now
Trending with Los Angeles, Miami

Trending in United States
#Caturday
21.9K Tweets

Trending in United States
#WearAMask
81.1K Tweets

Twitter
@Twitter

The only thing we want going viral is this Tweet

#WearAMask😷

12:56 PM · Sep 21, 2020 · Sprinklr

36.2K Retweets **3.4K** Quote Tweets

163.7K Likes

Technology

Streaming : Apache Storm, Apache Heron

Batch : Scalding, Spark, Apache Tez, Apache Hadoop

SQL : Presto, HIVE

Cloud : Google Cloud Platform (BQ, DF, GCS…)

Open Source

Twitter initiated projects : 🐦 TwitterOSS

Contributions & Adoption : Apache Software, Linux Foundation, Python Foundation, Scala Center…

Every day challenges

Data pipelines	50k+
Data Volume processed	200+PB
Data across storage systems	1+EB
Events processed	7+Tri

Continuous improvement

Data Processing requirements

- Stream vs Batch (Unification)
- Modern execution framework
- Newer technologies (Spark, Tez, Flink, Beam)
- Newer API (Scio, Beam, Spark, SQL, Streaming SQL)
- Easier adoption (Metrics, configuration, debugging tools, deployment and support)

Data Processing Evaluation

Dimensions

API
Unified and modern API, API Support, Language Support, Conversion tools from existing to new API.

Platform Offering
Platform availability, support and stability. Evaluation of different runners.

Platform Integration
Integrate with other tools, SQL, tabular, Data formats, Industry adoption.

Twitter Integration
Security, Orchestration, Deployment, Workbook integration, Chargeback, Monitoring, Cost.

Model use cases
Production vs Ad hoc stream and batch processing, ML workloads, Analytics. Right tool for customer use case.

Why Beam is attractive

- **Unified API**, Modern Execution frameworks
- **Flexibility of different runners** and how it affects company strategy
- Attractive for **multi cloud support**
- Different **programming languages**
- Strong open source **community and support**

BEAM SUMMIT

Streaming Adoption
Ad Engagement Analytic Platform

- Ad Engagement pipelines built on **lambda architecture**
- Stream processing **millions of events per second**
- Migrate **Apache Heron pipelines to Apache Beam**
- Utilize **same API** for both batch and streaming components
- Increased **developer velocity** and cleaner abstraction

Ad Engagement Analytics Platform

Batch Adoption
Experimentation Pipelines

- **Modernizing** Twitter Experimentation Pipelines
- Scalding based hard to **maintain, debug and scale**
- Easier **programming paradigm**
- Increase **developer productivity**
- Pipeline runtime from **days to hours**

Challenges

- **Language** :SCIO, Java, Python
- **Migration** : Variety and Scale
- **Custom libraries** : Use case specific logic
- **Long term support** : Compare against other APIs
- **Twitter Integration** : Metadata, deployment, monitoring…

Current Use cases

Source goes here.

- Machine Learning & Feature Engineering pipelines
- Curated data and metrics calculation
- Data Replication and Ingestion framework
- Real Time Analytics and Monitoring
- Ad Analytics platform
- Twitter Health monitoring pipelines
- Product learning platform

Future for Beam @Twitter

- Migration of all pipelines to **Apache Beam**
- **Unifying streaming/batch** and increase streaming use cases
- Integration tooling for **data delivery, metadata and monitoring**
- **Self serve** deployment and management
- Excited about **community engagement and contributions**

More at Beam Summit

- Talk to us about **opportunities**
- Tuesday, 19 14:00
  Log ingestion and data replication at Twitter by Praveen Killamsetti & Zhenzhao Wang
- Tuesday, 19 17:15
  Apache Beam backend for open source Scalding by Navin Viswanath

Thank you!

Twitter Career
Twitter Engineer Blog
Twitter Open Source