Machine Learning on Streaming Data

with Apache Kafka, Apache Beam, & TensorFlow

About Us

Mikhail Chrestkha
Machine Learning Specialist
Google Cloud

linkedin.com/in/mchrestkha

Stéphane Maarek
CEO & Kafka Instructor
DataCumulus

linkedin.com/in/stephanemaarek

Big Thanks to:

Julianne Cuneo
Big Data Specialist, Google Cloud

Kai Waehner
Technology Evangelist, Confluent

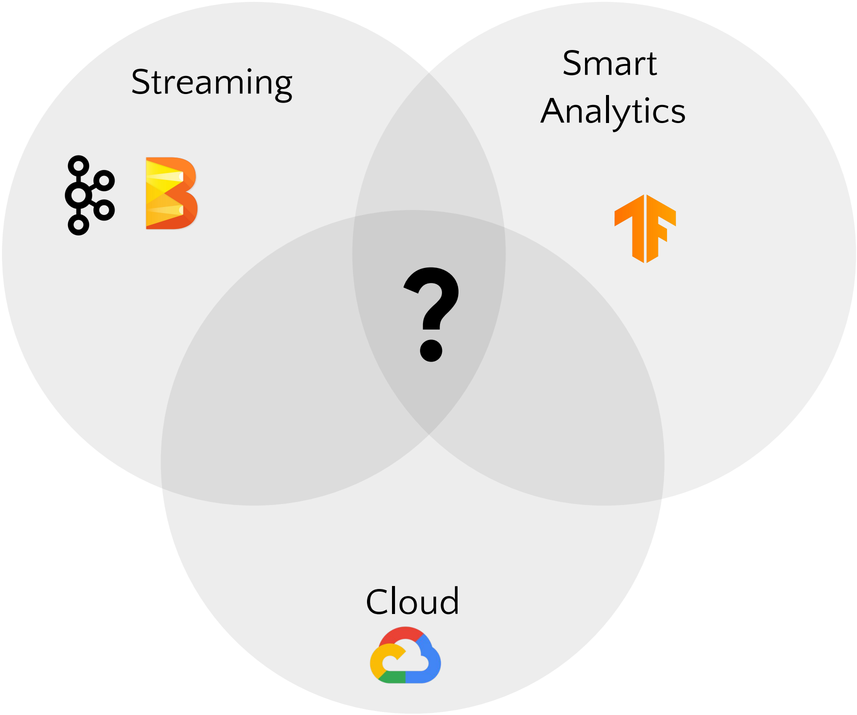
Agenda

1. Motivation
2. Architecture
3. **Use Case Walk-Through w/ Demo**
4. Summary

1

Motivation

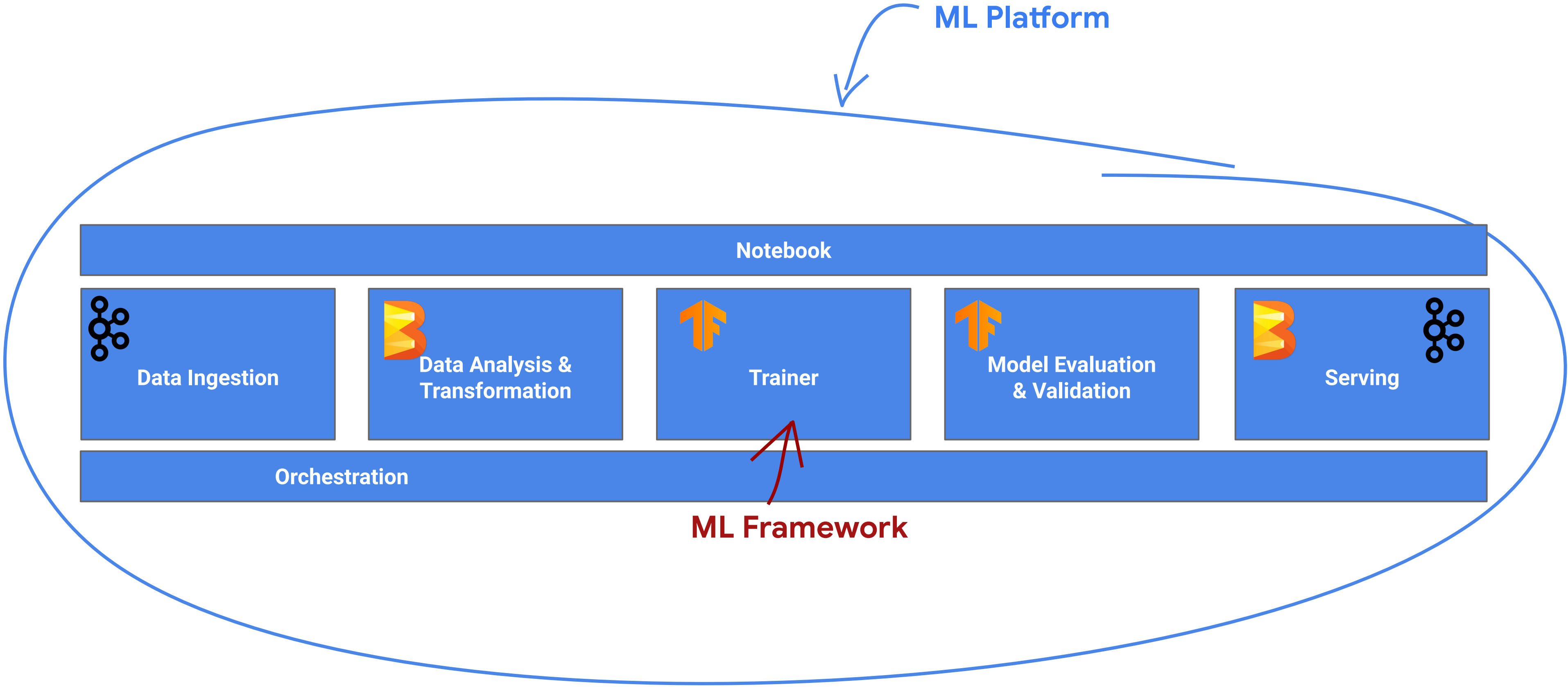
Technology Landscape



InfoWorld's 2019 Technology of the Year Award Winners:

- **Apache Beam**
- **Apache Kafka**
- Elastic Stack
- DataStax Enterprise
- Firebase
- Horovod
- H2O Driverless AI
- Keras
- Kubernetes
- LLVM
- .Net Core
- PyTorch
- Redis
- **TensorFlow**
- Visual Studio Code
- XGBoost

https://www.globenewswire.com/news-release/2019/01/30/1707685/0/en/InfoWorld-Announces-2019-Technology-of-the-Year-Award-Winners.html



ML Platform

Notebook

Data Ingestion

Data Analysis & Transformation

Trainer

Model Evaluation & Validation

Serving

Orchestration

ML Framework

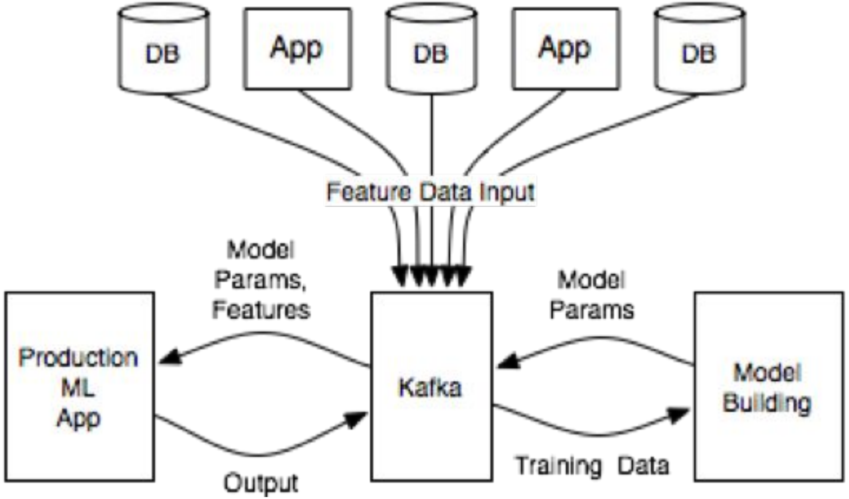
OSS	Managed Service
Apache Kafka Event streaming platform	**Confluent Cloud** Monitoring, Replication, Data Balancing
Apache Beam Data processing pipelines Unified batch & streaming	**Dataflow** Automated resource management of workers
TensorFlow Robust foundation for machine and deep learning	**Cloud Machine Learning Engine** ● Training: Distributed training infrastructure that supports CPUs, GPUs, and TPUs ● Serving: Host models for batch & online prediction

2

Architecture

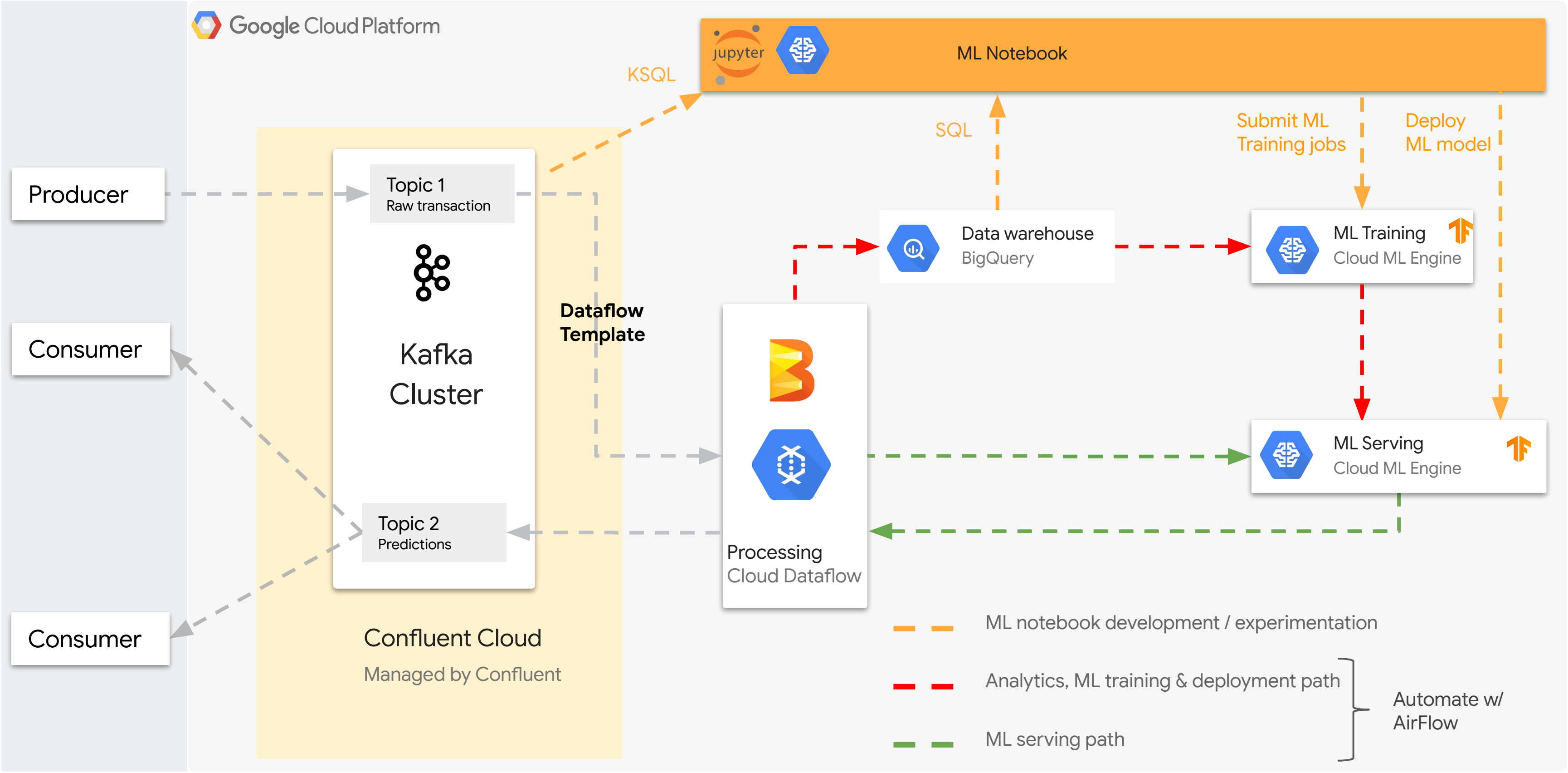
Reference Kafka ML Architecture

- Data pipelines are simplified
- Building analytic modules is decoupled from servicing them
- Usage of real time or batch as needed
- Analytic models can be deployed in a performant, scalable and mission-critical environment



Kai Waehner
Technology Evangelist, Confluent
https://www.confluent.io/blog/build-deploy-scalable-machine-learning-production-apache-kafka/

Leverage managed services to simplify & focus on code not infrastructure



Google Cloud Platform

Producer

Consumer

Consumer

Topic 1
Raw transaction

Kafka Cluster

Topic 2
Predictions

Confluent Cloud
Managed by Confluent

KSQL

Dataflow Template

jupyter

ML Notebook

SQL

Submit ML Training jobs

Deploy ML model

Data warehouse
BigQuery

ML Training
Cloud ML Engine

ML Serving
Cloud ML Engine

Processing
Cloud Dataflow

ML notebook development / experimentation

Analytics, ML training & deployment path

ML serving path

Automate w/ AirFlow

3

Use Case Walk-Through

Kaggle Case Study
Fraud Detection of Credit Card Transactions

284,807

transactions

492

Fraud (0.172%)

- Collect transaction data
- Analyze historical data
- Train model on historic sample
- Evaluate model based on precision & recall
- Predict fraud on new streaming data

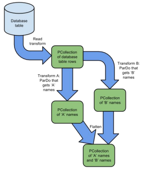
- Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015
- Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon
- Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE
    - Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)
- Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier
- Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing

DEMO 1 - **5 min**

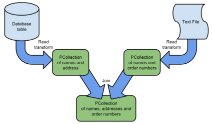
Sending our credit card data

Confluent Cloud, Creating a Topic, Python Script, Security

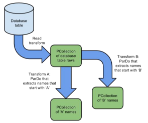
Kafka to BigQuery



Database table
Read transform
PCollection of database table rows
Transform A: ParDo that gets 'A' names
Transform B: ParDo that gets 'B' names
PCollection of 'A' names
PCollection of 'B' names
Flatten
PCollection of 'A' names and 'B' names

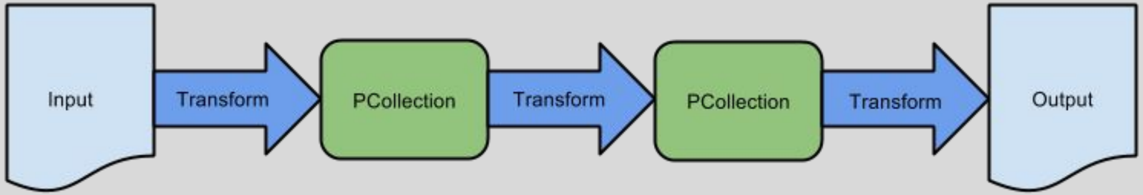


Database table
Text File
Read transform
Read transform
PCollection of names and address
PCollection of names and order numbers
Join
PCollection of names, addresses and order numbers



Database table
Read transform
PCollection of database table rows
Transform A: ParDo that extracts names that start with 'A'
Transform B: ParDo that extracts names that start with 'B'
PCollection of 'A' names
PCollection of 'B' names

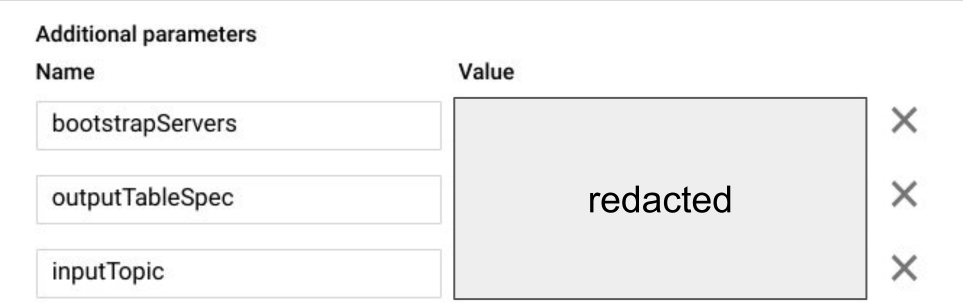
Java Code



```
KafkaIO.<String, String>read()                    BigQueryIO.writeTableRows()
```

Create a template for easy re-usability by an analyst

Dataflow Template



Additional parameters

Name	Value
bootstrapServers	redacted
outputTableSpec	
inputTopic	

Images from https://beam.apache.org/documentation/pipelines/design-your-pipeline/

Explore data & train ML model

Query directly from topic

```
from ksql import KSQLAPI
```

redacted

Query petabytes of data

```
%%bigquery
```

redacted

Submit ML training job

```
gcloud ml-engine jobs
submit training
```

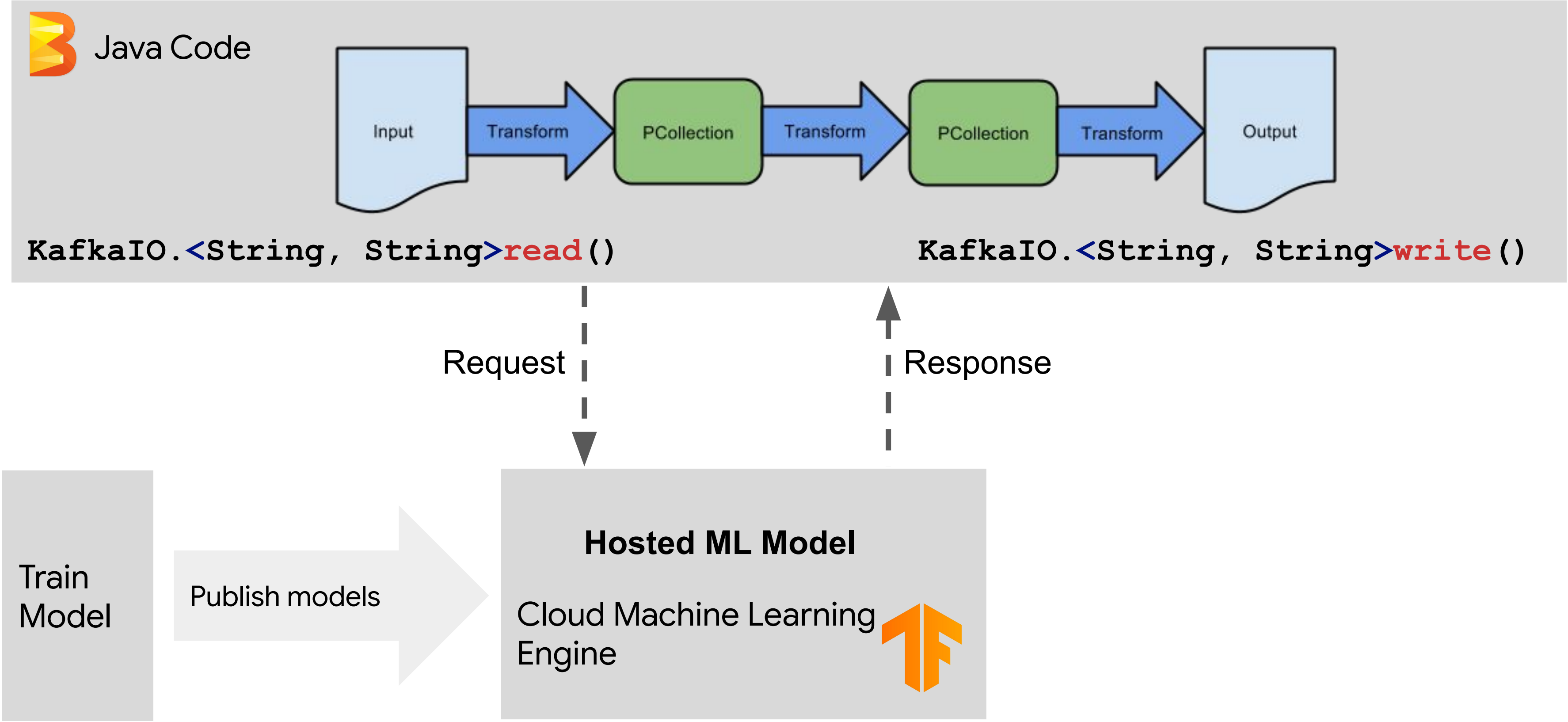
redacted

DEMO 2 - **5 min**

Dataflow template & job

Jupyter: KSQL, BQML, TensorFlow CMLE job

Send Predictions back to Kafka



Java Code

```
KafkaIO.<String, String>read()
```

```
KafkaIO.<String, String>write()
```

Request

Response

Train Model

Publish models

Hosted ML Model

Cloud Machine Learning Engine

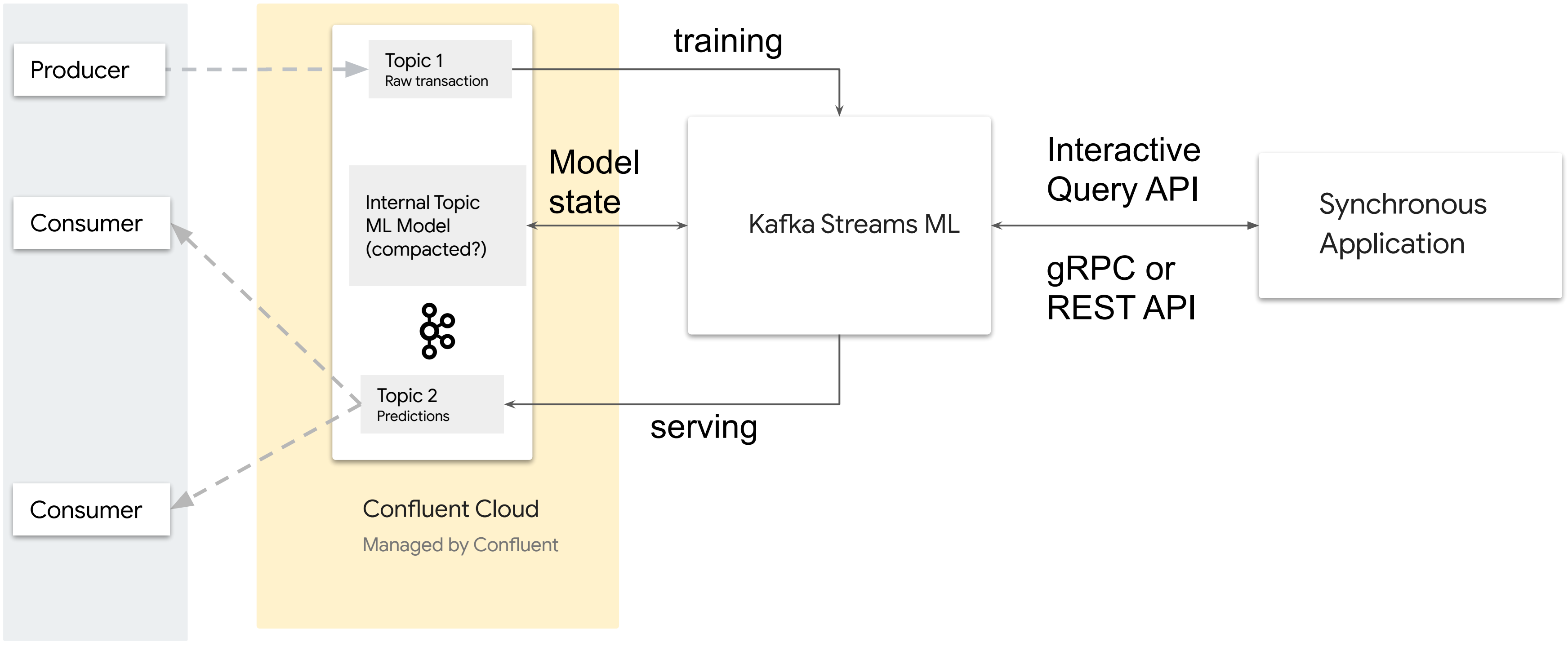
Image from https://beam.apache.org/documentation/pipelines/design-your-pipeline/

DEMO 3 - **5 min**

(1) Deploy model as an end point
(2) Prediction sent to Kafka topic to be consumed
(3) Track models & monitor predictions in CMLE UI

Futuristic Architecture: Pure Kafka-based ML

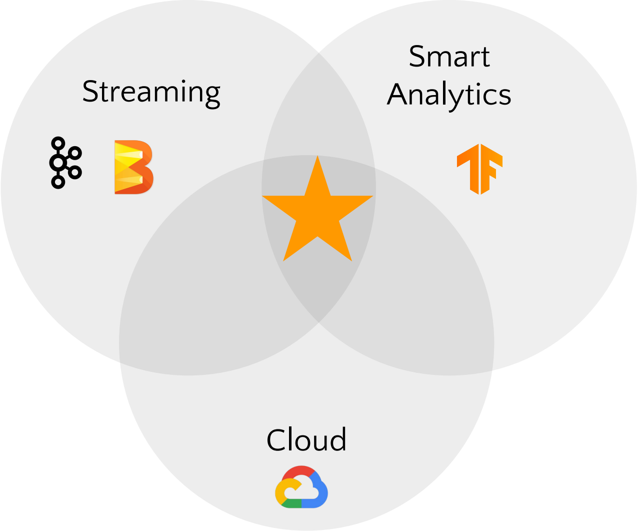
Resilient, highly available, sync & async



4 Summary

Summary

- Kafka + Beam + TensorFlow = Great foundation for future
  - Batch today ⟶ streaming tomorrow
  - Small data ⟶ big data tomorrow
  - Shallow learning today ⟶ deep learning tomorrow
- Make data & ML easier for yourself by using managed services
- Build for many other use cases:
  - Predictive maintenance
  - Logistics routing
  - Image search & recommendations in e-commerce



Talk to Google Cloud

K1

Booth

Learn More

Blog: Enabling connected transformation with Apache Kafka and TensorFlow on Google Cloud Platform
bit.ly/2CHERol

KafkaIO on Beam
bit.ly/2YwL3Jc

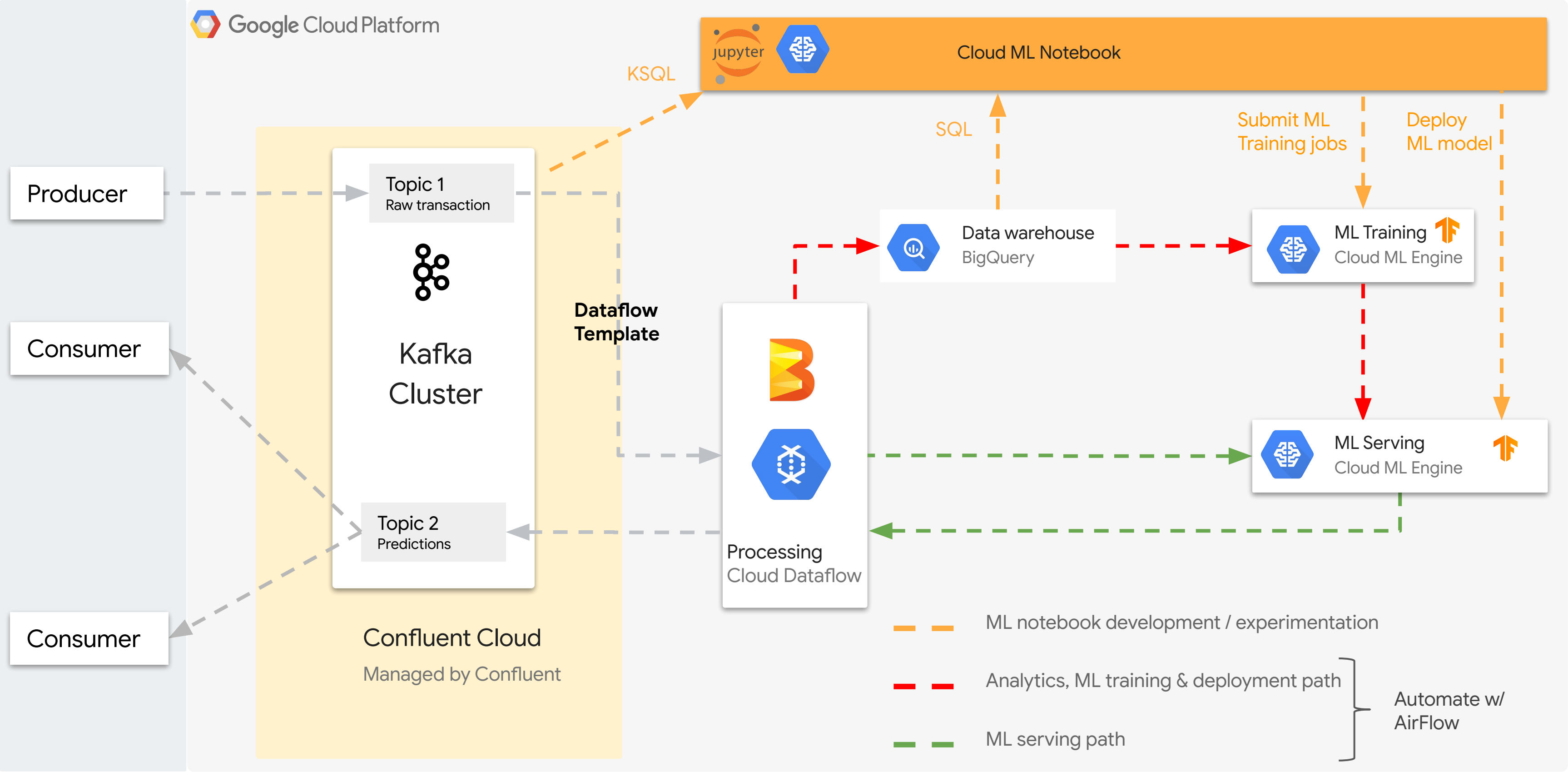
KafkaToBigQuery Dataflow Template Example
bit.ly/2HQqVN0

Contact us

linkedin.com/in/mchrestkha

linkedin.com/in/stephanemaarek

Questions


Google Cloud Platform
jupyter
Cloud ML Notebook
KSQL
SQL
Submit ML
Training jobs
Deploy
ML model
Producer
Topic 1
Raw transaction
Kafka
Cluster
Topic 2
Predictions
Consumer
Consumer
Dataflow
Template
Data warehouse
BigQuery
ML Training
Cloud ML Engine
ML Serving
Cloud ML Engine
Processing
Cloud Dataflow
Confluent Cloud
Managed by Confluent
ML notebook development / experimentation
Analytics, ML training & deployment path
ML serving path
Automate w/
AirFlow