Bringing ML Online: Defining ML Features in Real-Time with Bytewax

Zander Matheson, Founder & CEO @ bytewax

FEATURE STORE SUMMIT 2023

Organized by HOPSWORKS

Recommendations	Addictive TikTok
Ads	Annoying or not, at least less so with real-time relevance
Personalization	Mind reading in 2023, or at least it feels like it
Fraud Detection	Alerting us of our suspect train ticket purchases in foreign countries 😄

I’m Zander!

Working on Bytewax -> github.com/bytewax/bytewax

Proud human and dog dad

This photo is me trying to look cool 😆

You can find me in the Bytewax slack, or if you are in Santa Cruz, send me a LinkedIn and we can grab a coffee.

Today's Agenda

1. What and why real-time ML?
2. Building a pipeline to analyze streaming data
3. Building a real-time feature pipeline

1. What and why real-time ML?
2. Building a pipeline to analyze streaming data
3. Building a real-time feature pipeline

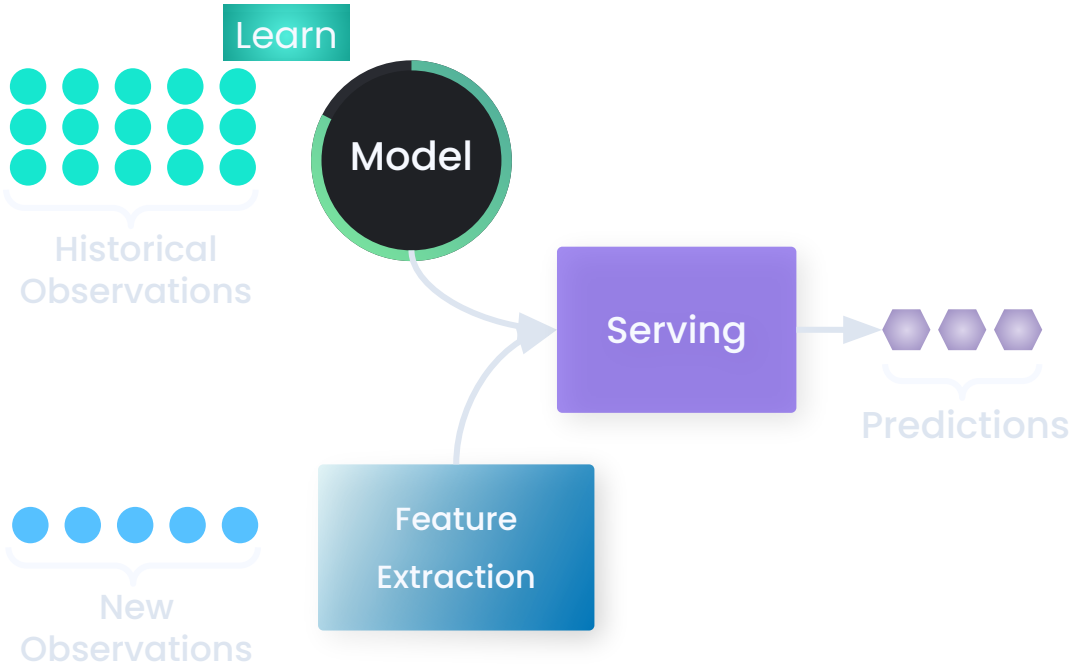
Google	For every 100 ms of latency -> drop in search traffic of ~0.2%
Akamai	For every 100 ms of latency -> a 7% drop in conversion rates
Walmart	For every 1s of load time improvement -> a 2% increase in conversion rates
Staples	For every 1s of load time improvement -> a 10% increase in conversion rate

Real-Time Machine Learning

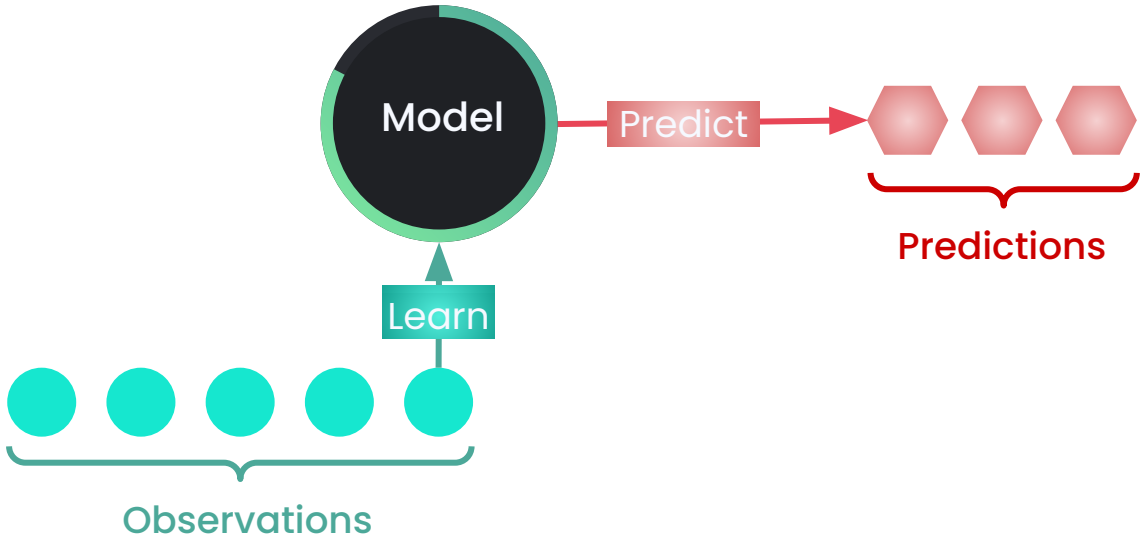
Real-Time ML can be classified into to types:

1. Offline training with online serving
2. Online training **and** serving

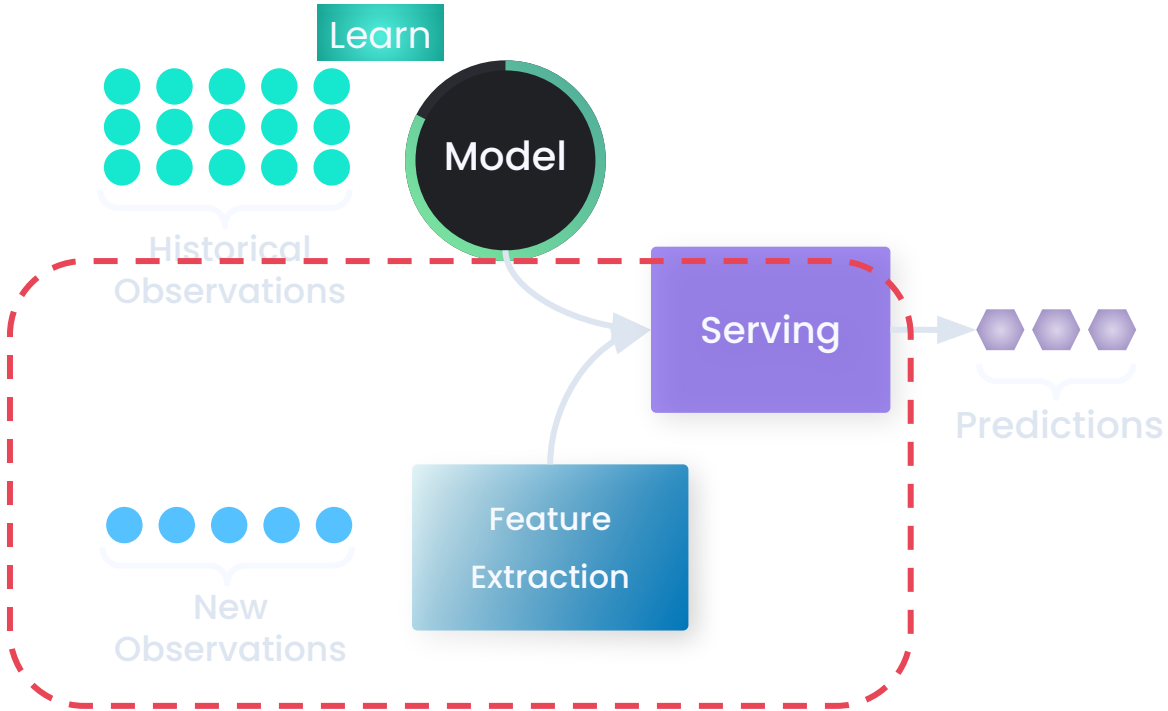
Offline Training, Online Serving



Online Training and Serving



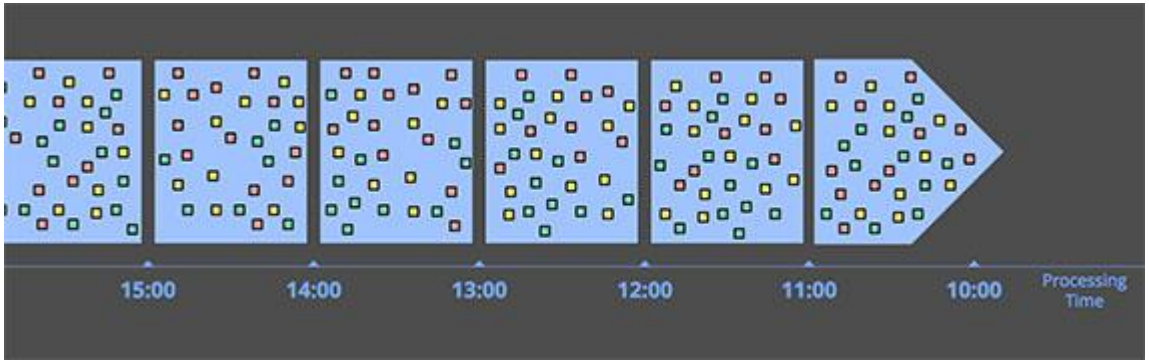
Offline Training, Online Serving

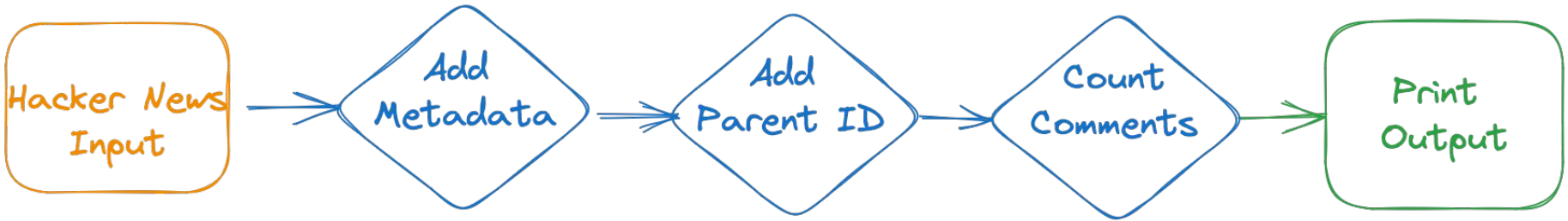


Outline:

1. What and why real-time ML?
2. Building a pipeline to analyze streaming data
3. Building a real-time feature pipeline

Streaming Data

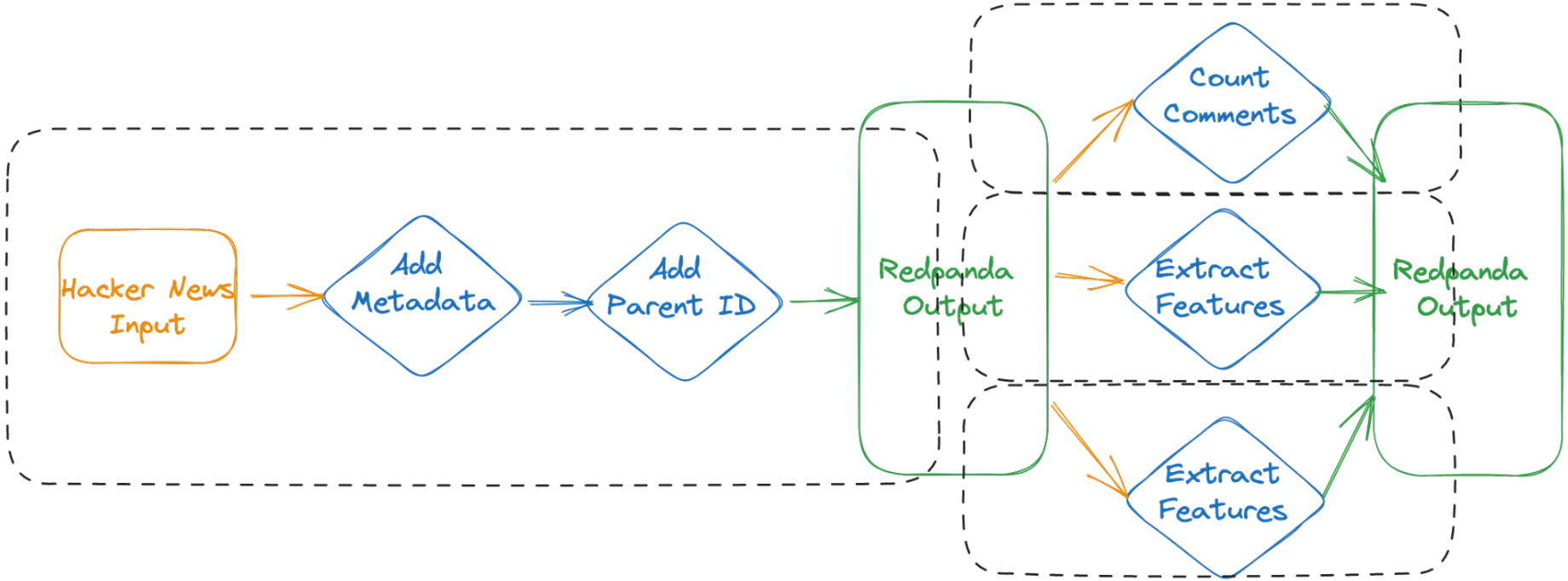

15:00
14:00
13:00
12:00
11:00
10:00
Processing Time


Hacker News Input
Add Metadata
Add Parent ID
Count Comments
Print Output

```python
import requests
from datetime import datetime, timedelta
import time
from typing import Any, Optional

from bytewax.connectors.periodic import SimplePollingInput
from bytewax.connectors.stdio import StdOutput
from bytewax.dataflow import Dataflow

import logging

logging.basicConfig(level=logging.INFO)
logger = logging.getLogger(__name__)


class HNInput(SimplePollingInput):
    def __init__(self, interval: timedelta, align_to: Optional[datetime] = None, init_item: Optional[int] = None):
        super().__init__(interval, align_to)
        logger.info(f"received starting id: {init_item}")
        self.max_id = init_item

    def next_item(self):
        '''
        Get all the items from hacker news API between
        the last max id and the current max id
        '''
        if not self.max_id:
```

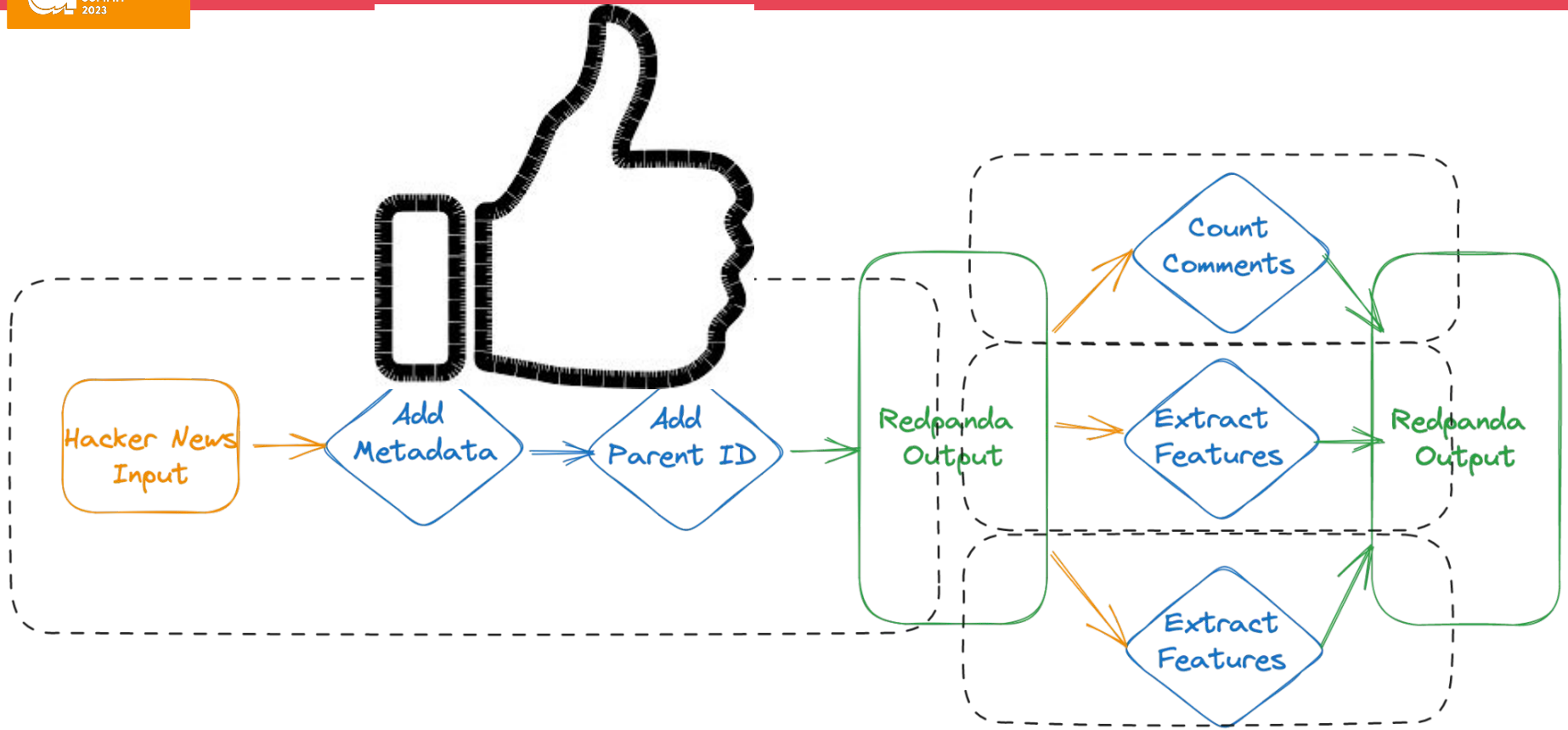
PROBLEMS 118 OUTPUT DEBUG CONSOLE TERMINAL

```
efronts in two main areas, Russian-Ukraine and Israel-Palestine, seems to echo the situation before WWI and WWII. Not to mention coming at a time of int
ensifying geopolitical &amp;quot;chicken&amp;quot;, and increased international distrust.<p>Add that to a world suffering under the stress of economic crises, g
as shortages and price hikes, and just out of a global pandemic--it seems as if we could be on a precipice.<p>While there are some surface-level similar
ities between now and the pre-histories of both world wars of last century, there are also substantial differences. History doesn&amp;#x27;t repeat itself e
xactly, though it can sometimes rhyme.<p>Differences include the presence of global institutions, nuclear deterrence, interconnected economies, the infl
uence of 24&amp;#x2F;7 media, and a less ideological drive behind conflicts. Similarities feature regional conflicts that could draw in larger powers, econo
mic crises and resource scarcity, a general atmosphere of global unrest, and rising nationalism.<p>Alarmism is easy, and apocalyptic doomsayers are comm
onplace, so it&amp;#x27;s wise to be cautious and often sensible to dismiss such talk as baseless catastrophizing.<p>However, it is possible that this time
is different. What say you?', 'time': 1696764219, 'title': 'Ask HN: How close are we to a World War situation?', 'type': 'story'}}
<dataflow.Story object at 0x7f9f38162f80>
('37805009', 1)
('37809516', 4)
('37809582', 2)
INFO:dataflow:current id: 37813690, new id: 37813690
```


Hacker News Input
Add Metadata
Add Parent ID
Redpanda Output
Count Comments
Extract Features
Extract Features
Redpanda Output

Outline:

1. What and why real-time ML?
2. Building a pipeline to analyze streaming data
3. Building a real-time feature pipeline


Hacker News Input
Add Metadata
Add Parent ID
Redpanda Output
Count Comments
Extract Features
Extract Features
Redpanda Output

What is the next step to integrate this with feature stores for real-time ML?

With hopsworks you can register the features in the kafka topic and the ingestion into the hsfs will happen automatically. You can check out the tutorial in the logical clocks repo.

https://github.com/logicalclocks/hopsworks-tutorials/tree/master/integrations/bytewax

You can find the code used in the videos - https://github.com/awmatheson/guide-periodic-hackernews

bytewax.io
github.com/bytewax/bytewax