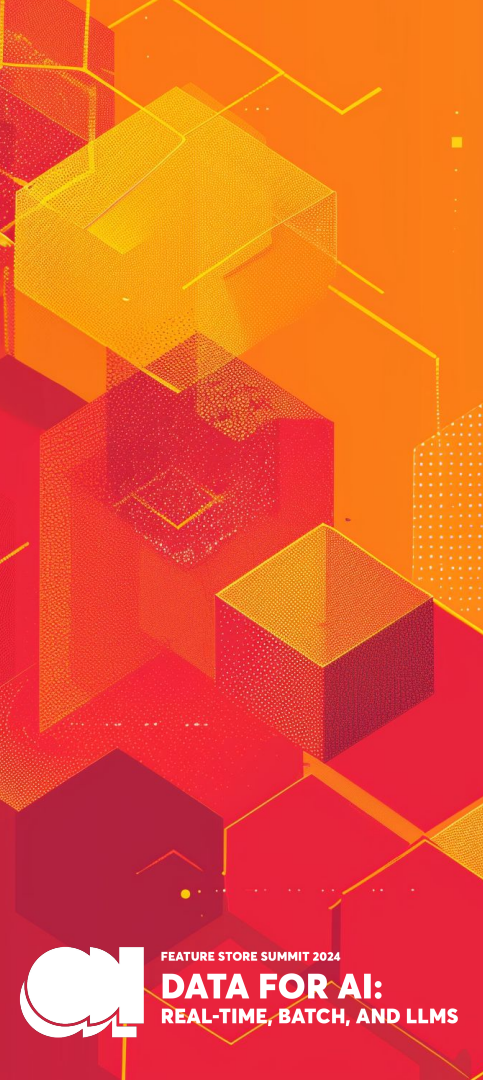


Enabling Low Latency Fraud Detection with Real-Time Feature Engineering

Tun Shwe, VP of Data @ Quix

Organized by HOPSWORKS

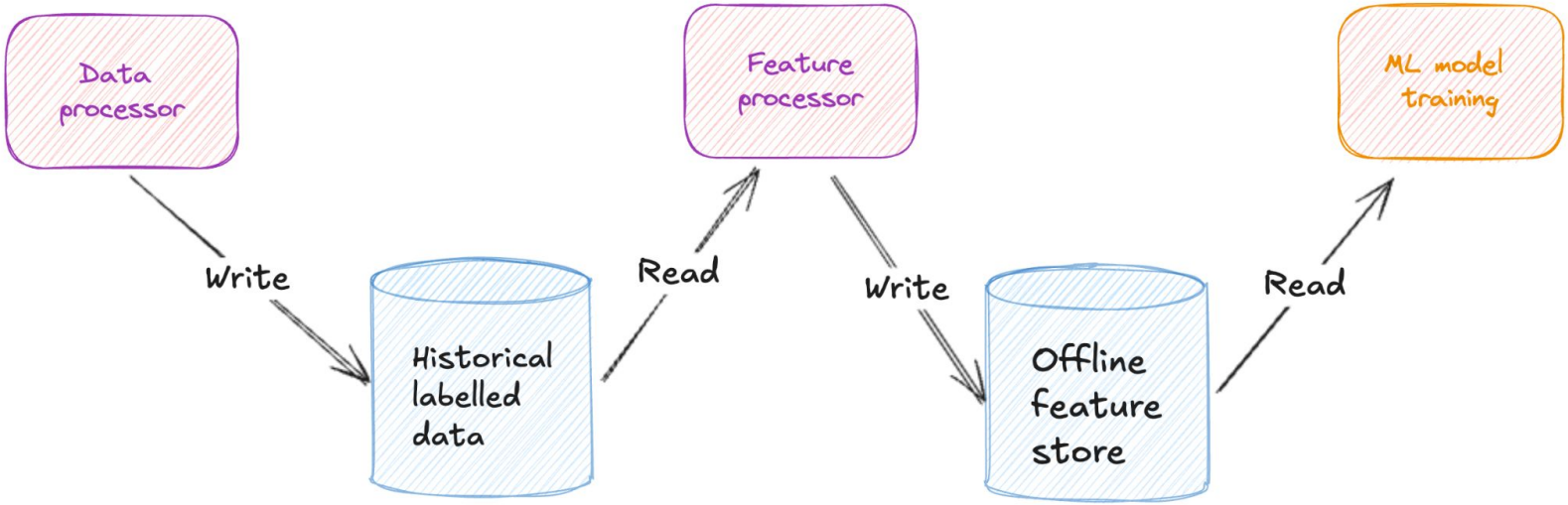
Agenda

- What is real-time feature engineering?
- How do I build a real-time feature pipeline?
- How do I achieve low latency?

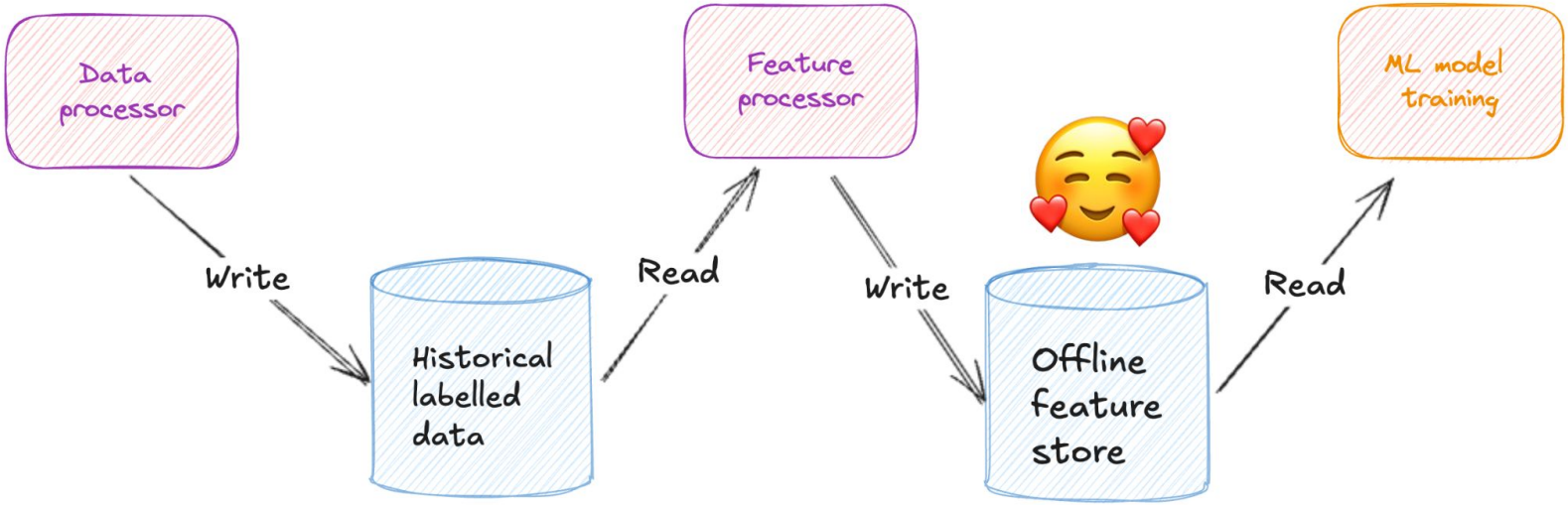
What is feature engineering?

- Features are properties that provide predictive power for machine learning models
- They are inputs for models during training
- They are inputs for models during inference
- They can be projected from data or computed, e.g. aggregations, vectors

Batch feature engineering



Batch feature engineering



Feature stores decouple feature engineering from models

Historical labelled data

Transactions


```
{
    "tid": "adb0a3cd4fd48e928bd61582978cfbb0",
    "datetime": "2024-07-15 19:41:03",
    "cc_num": "4561945063212434",
    "category": "Restaurant/Cafeteria",
    "amount": "60.78",
    "latitude": "33.7207",
    "longitude": "-116.21677",
    "city": "Indio",
    "country": "US",
    "fraud_label": 0
}
```

```
{
    "tid": "c9aa89860d0ecdab893f08f41785d0e7",
    "datetime": "2024-07-16 15:31:26",
    "cc_num": "4561945063212434",
    "category": "Cash Withdrawal",
    "amount": "200.00",
    "latitude": "-6.48167",
    "longitude": "106.85417",
    "city": "Cibinong",
    "country": "ID",
    "fraud_label": 1
}
```

Historical labelled data

Transactions and profiles

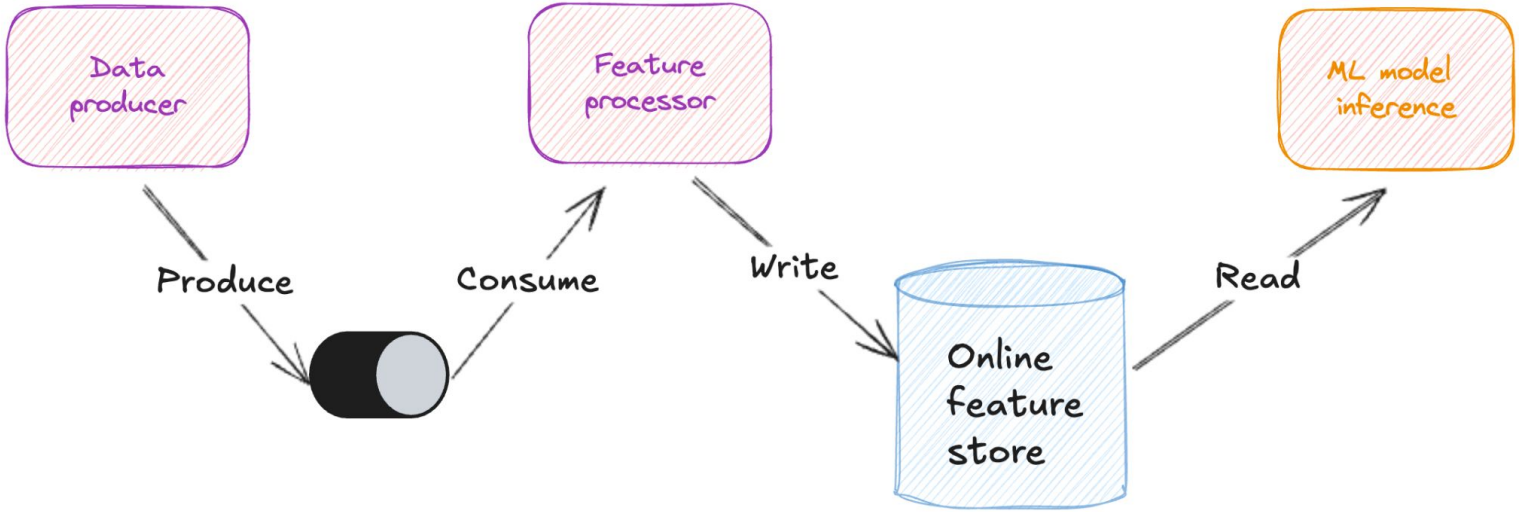

```
{
    "tid": "adb0a3cd4fd48e928bd61582978cfbb0",
    "datetime": "2024-07-15 19:41:03",
    "cc_num": "4561945063212434",
    "category": "Restaurant/Cafeteria",
    "amount": "60.78",
    "latitude": "33.7207",
    "longitude": "-116.21677",
    "city": "Indio",
    "country": "US",
    "fraud_label": 0
}
{
    "tid": "c9aa89860d0ecdab893f08f41785d0e7",
    "datetime": "2024-07-16 15:31:26",
    "cc_num": "4561945063212434",
    "category": "Cash Withdrawal",
    "amount": "200.00",
    "latitude": "-6.48167",
    "longitude": "106.85417",
    "city": "Cibinong",
    "country": "ID",
    "fraud_label": 1
}
```

```
{
    "cc_num": "4561945063212434"
    "cc_provider": "mastercard",
    "cc_type": "credit",
    "cc_expiration_date": "02/26",
    "name": "Andrea Watson",
    "birthdate": "1949-04-15",
    "age": "75",
    "city": "Collinwood",
    "country_of_residence": "US"
}
```

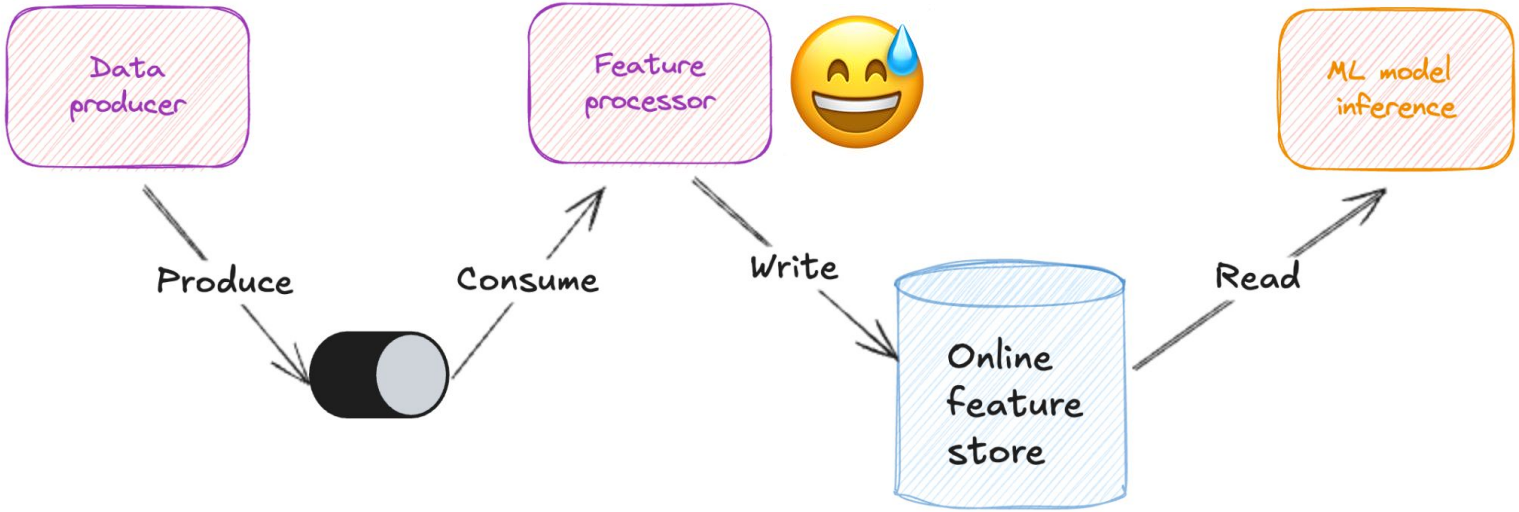
Features that could be computed

- Total transaction amount in the last month
- Total transaction amount per day of the week
- Average transaction amount per week
- Average spend per transaction category in the last month
- Count of transactions made outside of city of residence
- Count of transactions made in the past 24 hours

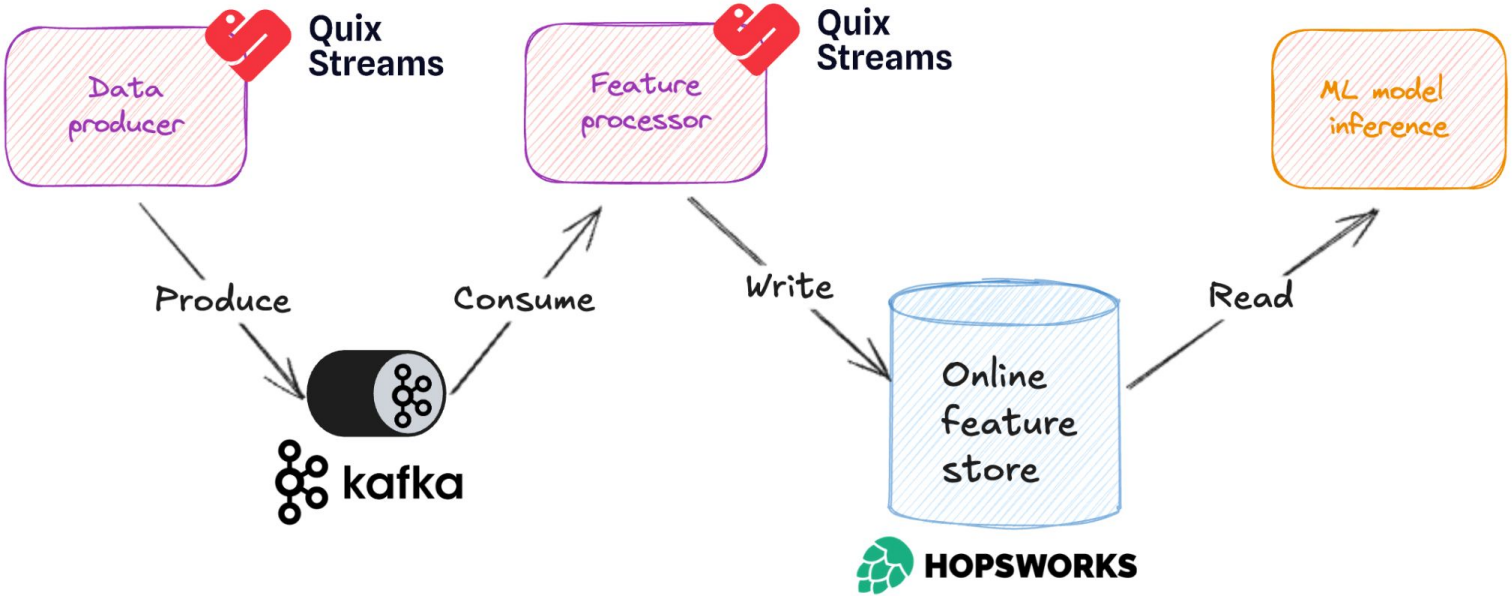
Real-time feature engineering

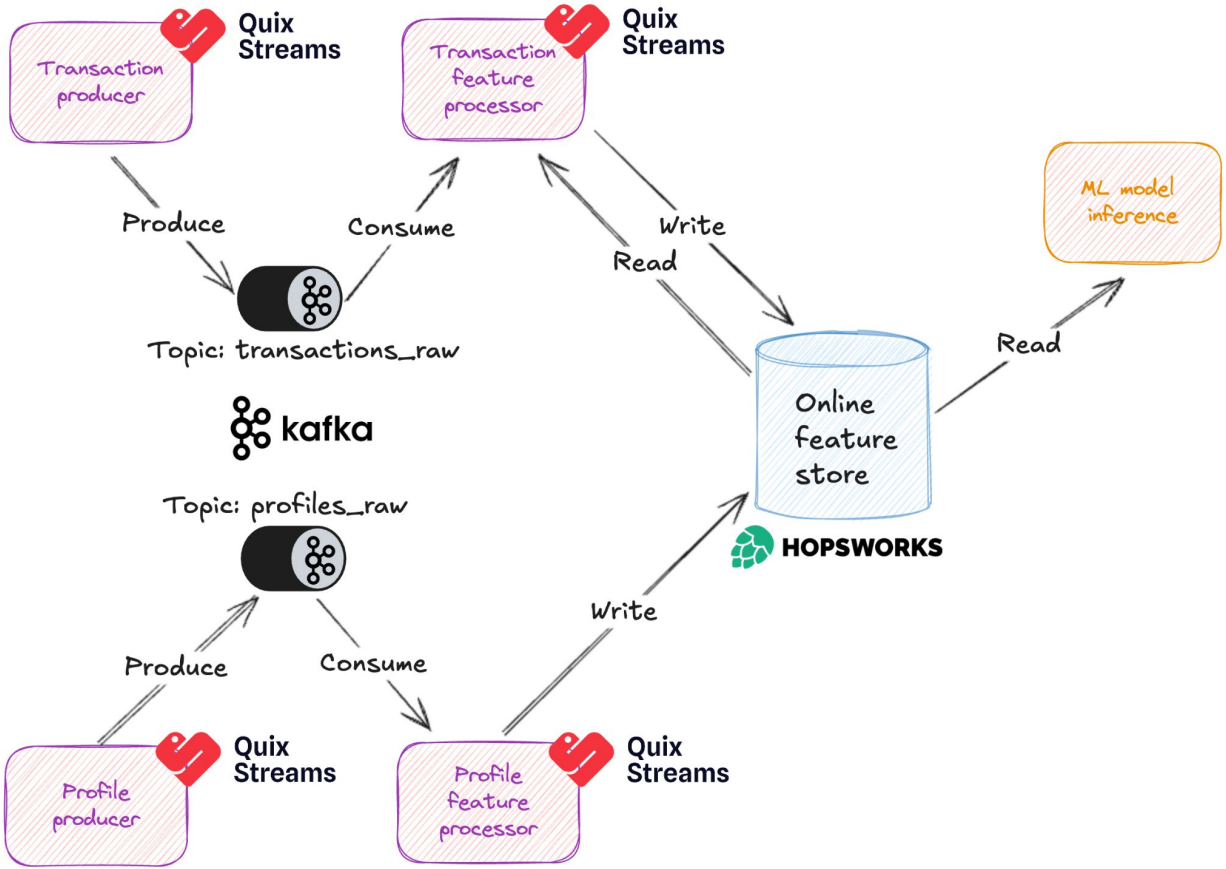


Real-time feature engineering (is challenging)



Real-time feature engineering (is easier these days)




Quix Streams
Transaction producer
Quix Streams
Transaction feature processor
ML model inference
Produce
Consume
Write
Read
Topic: transactions_raw
kafka
Online feature store
Read
HOPSWORKS
Topic: profiles_raw
Write
Produce
Consume
Quix Streams
Profile producer
Quix Streams
Profile feature processor

Enabling low latency through technology choices

Choose streaming data

- Use an online feature store for consistency and fast retrieval, e.g. Hopsworks
- Use a streaming message broker, e.g. Kafka, Redpanda, Pulsar
- Use stream processing for feature engineering, e.g. Spark, Flink, Quix Streams

Why is Kafka fast?

- Sequential I/O
- Zero copy principle

Sequential I/O

- Records have order guarantees
- Kafka is an append-only log
- Stored data is organised as contiguous blocks of memory
- Modern drives and SSDs are optimised for sequential I/O rather than random I/O
- Contrast that with databases that are optimised for random access

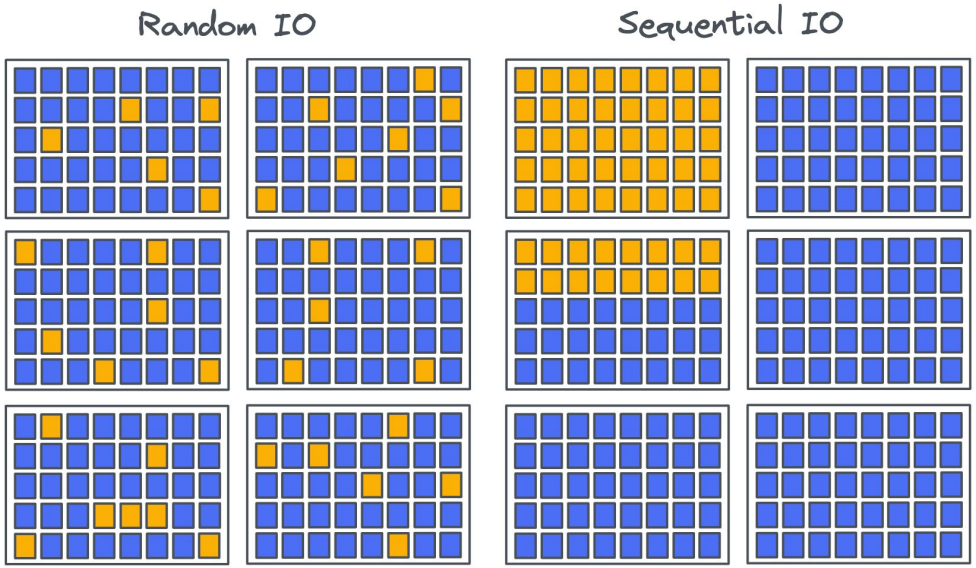


Image credit: Jack Vanlightly
https://jack-vanlightly.com/blog/2023/5/9/is-sequential-io-dead-in-the-era-of-the-nvme-drive

Zero copy principle

- Refers to the copying of data between kernel and application representations
- Does not mean making zero copies; actually means minimising number of copies
- Consumers read topic data directly from the log file using direct memory access (DMA)
- Doesn't apply when encryption/SSL/TLS is used

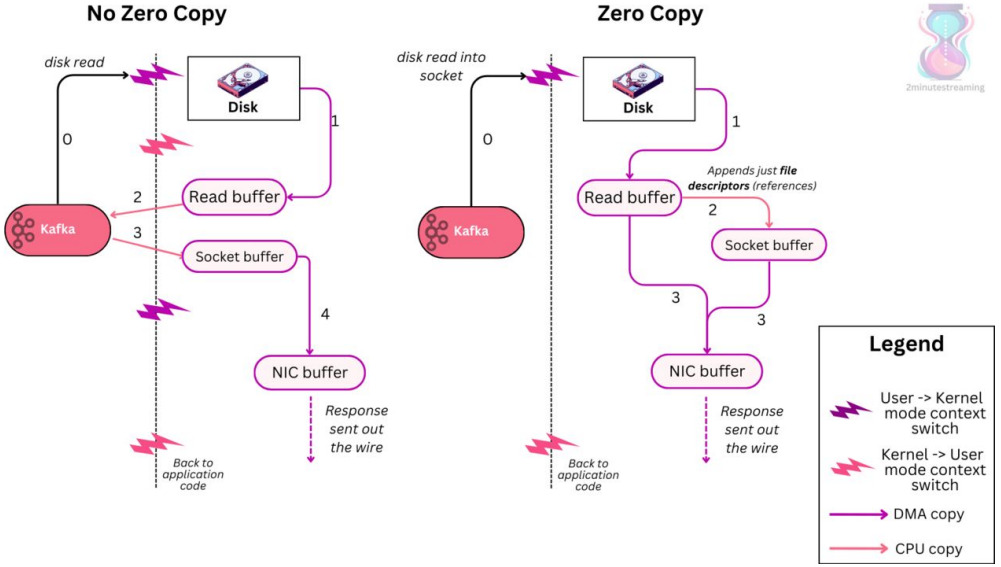


Image credit: Stanislav Kozlovski
https://2minutestreaming.beehiiv.com/p/apache-kafka-zero-copy-operating-system-optimization

Quix Streams

Streaming DataFrames

```python
from quixstreams import Application, State

app = Application(broker_address="localhost:9092")

input_topic = app.topic("my_input_topic")
output_topic = app.topic("my_output_topic")

# Create a Streaming DataFrame
sdf = app.dataframe(topic=input_topic)

sdf["field_C"] = sdf.apply(lambda value: value["field_A"] + value["field_B"])

sdf = sdf.to_topic(output_topic)

if __name__ == "__main__":
    app.run(sdf)
```

Quix Streams

Stateful operations

```python
from quixstreams import Application, State

app = Application(broker_address="localhost:9092")

input_topic = app.topic("my_input_topic")
output_topic = app.topic("my_output_topic")

def count(data: dict, state: State):
    total = state.get('total', default=0)
    total += 1
    state.set('total', total)
    data['total'] = total

sdf = app.dataframe(topic=input_topic)

sdf = sdf.update(count, stateful=True)

sdf = sdf.to_topic(output_topic)

if __name__ == "__main__":
    app.run(sdf)
```

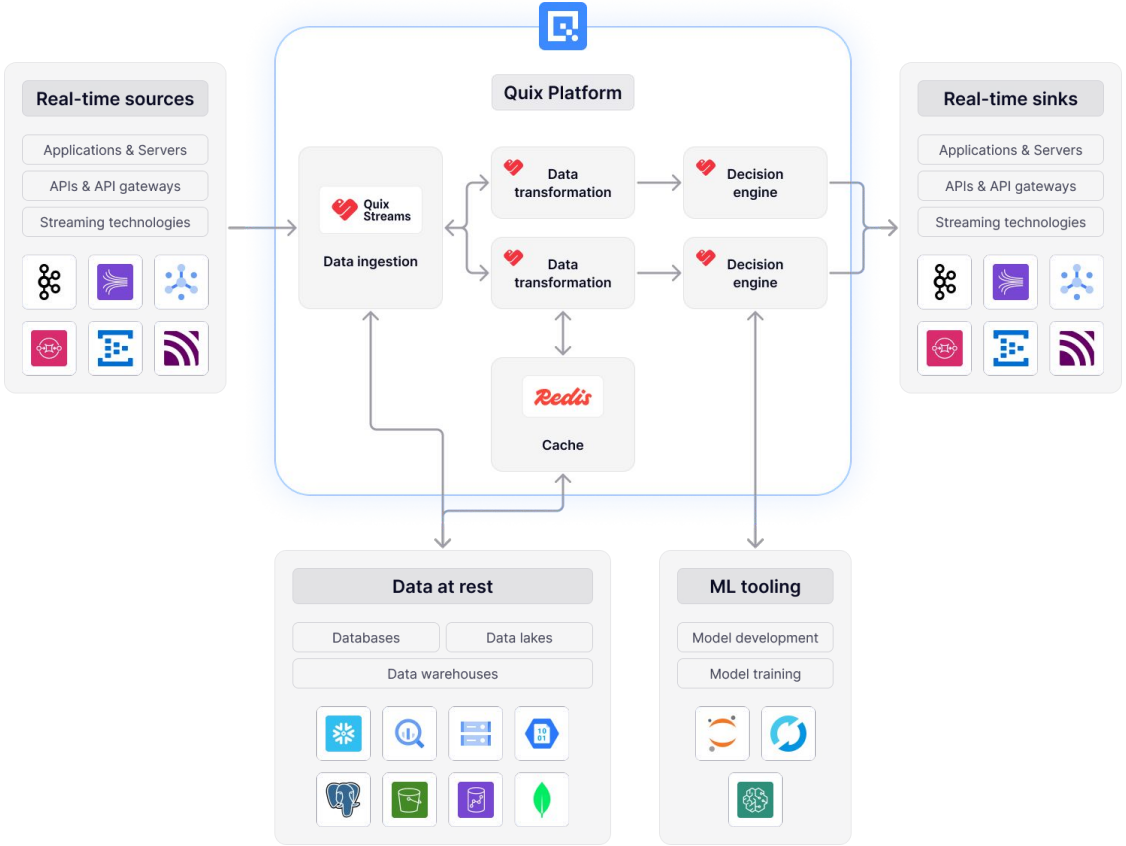
Quix Streams

Stateful window operations

```python
from quixstreams import Application
from datetime import timedelta

...

sdf = app.dataframe(input_topic)

sdf = (
    # Extract the "total" field from the record
    sdf.apply(lambda value: value["total"])

    # Define a tumbling window of 10 minutes with a 10 second grace period
    .tumbling_window(timedelta(minutes=10), grace_ms=timedelta(seconds=10))

    # Specify the "sum" aggregation function to apply to values of "total"
    .sum()

    # Emit results only when the 10 minute window has elapsed
    .final()
)
```

Quix

Quix Cloud and Quix Edge



Thank you!

github.com/quixio/quix-streams

linkedin.com/in/tunshwe

quix.io/slack-invite