

Realtime Risk Management

USING KAFKA, PYTHON, AND SPARK STREAMING





UNDERWRITING CREDIT CARD TRANSACTIONS IS RISKY





WE NEED TO BE QUICK AT
UNDERWRITING





WE ALSO NEED TO AVOID
LOSING MONEY



Some Numbers

\$12Bn

CUMULATIVE
PROCESSED

200k+

MERCHANTS

14k

EVENTS / SECOND

6

RISK ANALYSTS

Risk Analysts

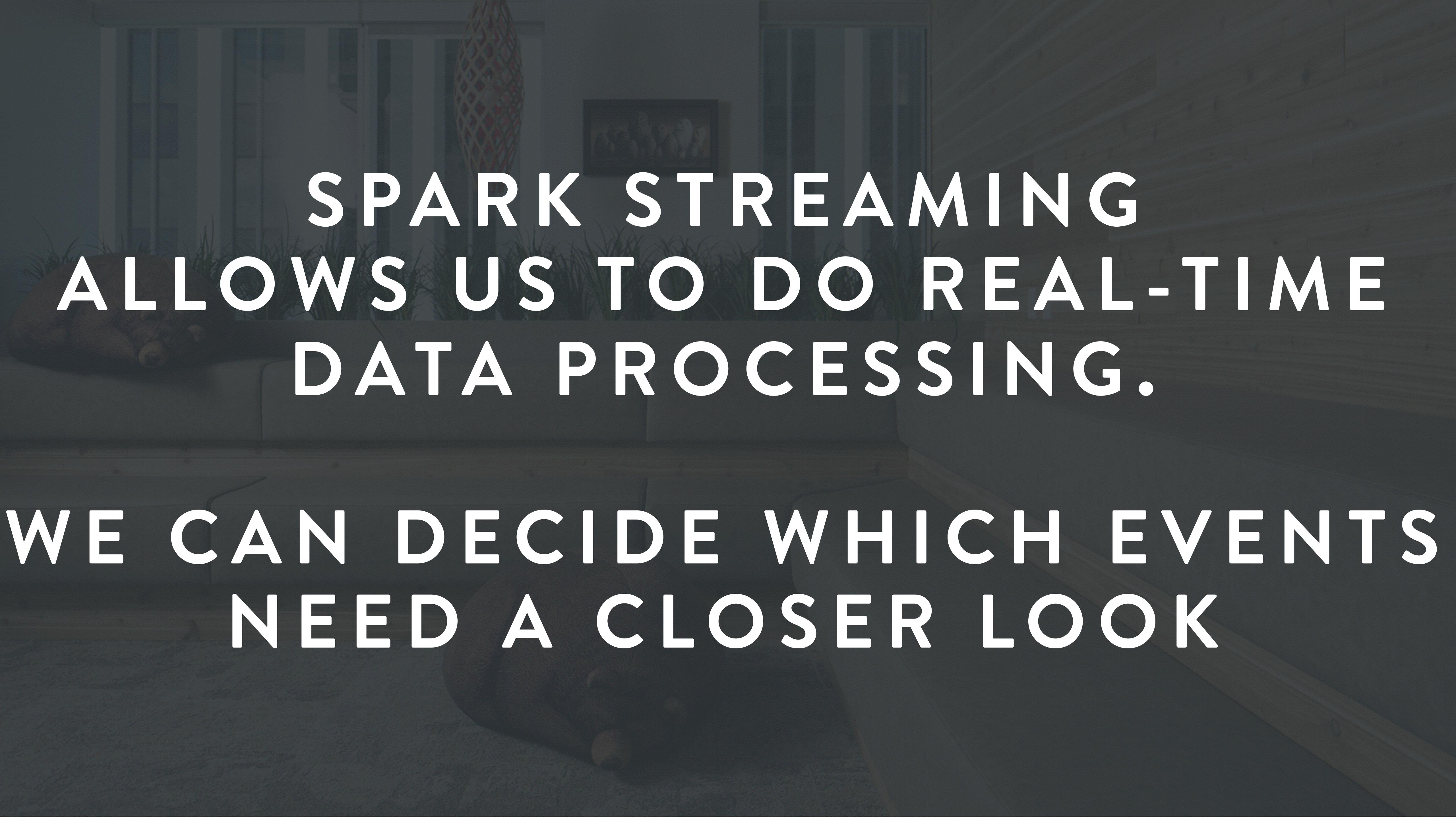




WE NEEDED TO BUILD
SOMETHING THAT STOPS THE
HOPE FACTOR







SPARK STREAMING
ALLOWS US TO DO REAL-TIME
DATA PROCESSING.

WE CAN DECIDE WHICH EVENTS
NEED A CLOSER LOOK

Intro to Kafka, Zookeeper, and Spark Streaming

Apache Kafka

Anatomy of a Topic

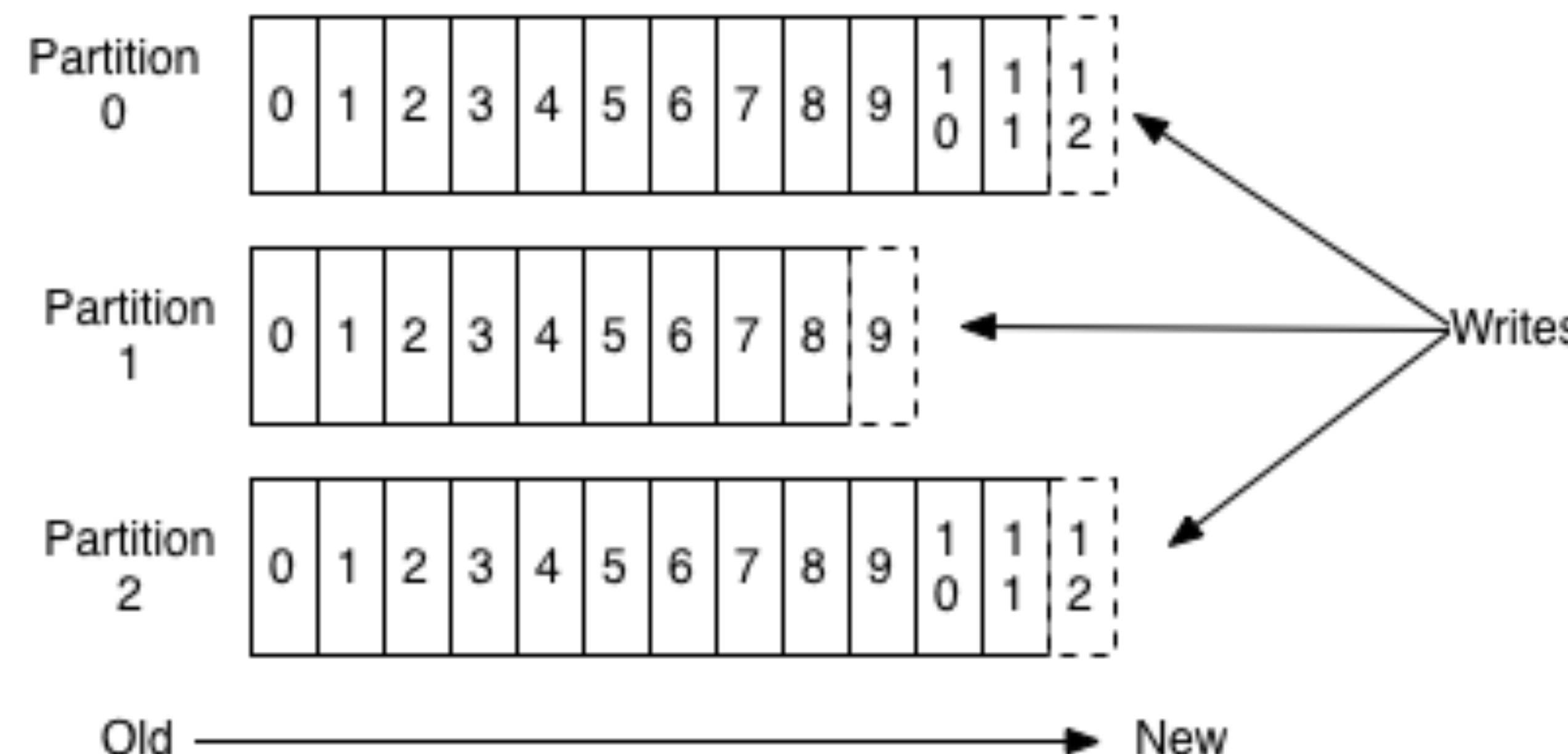


Image taken from Kafka docs

Apache Zookeeper

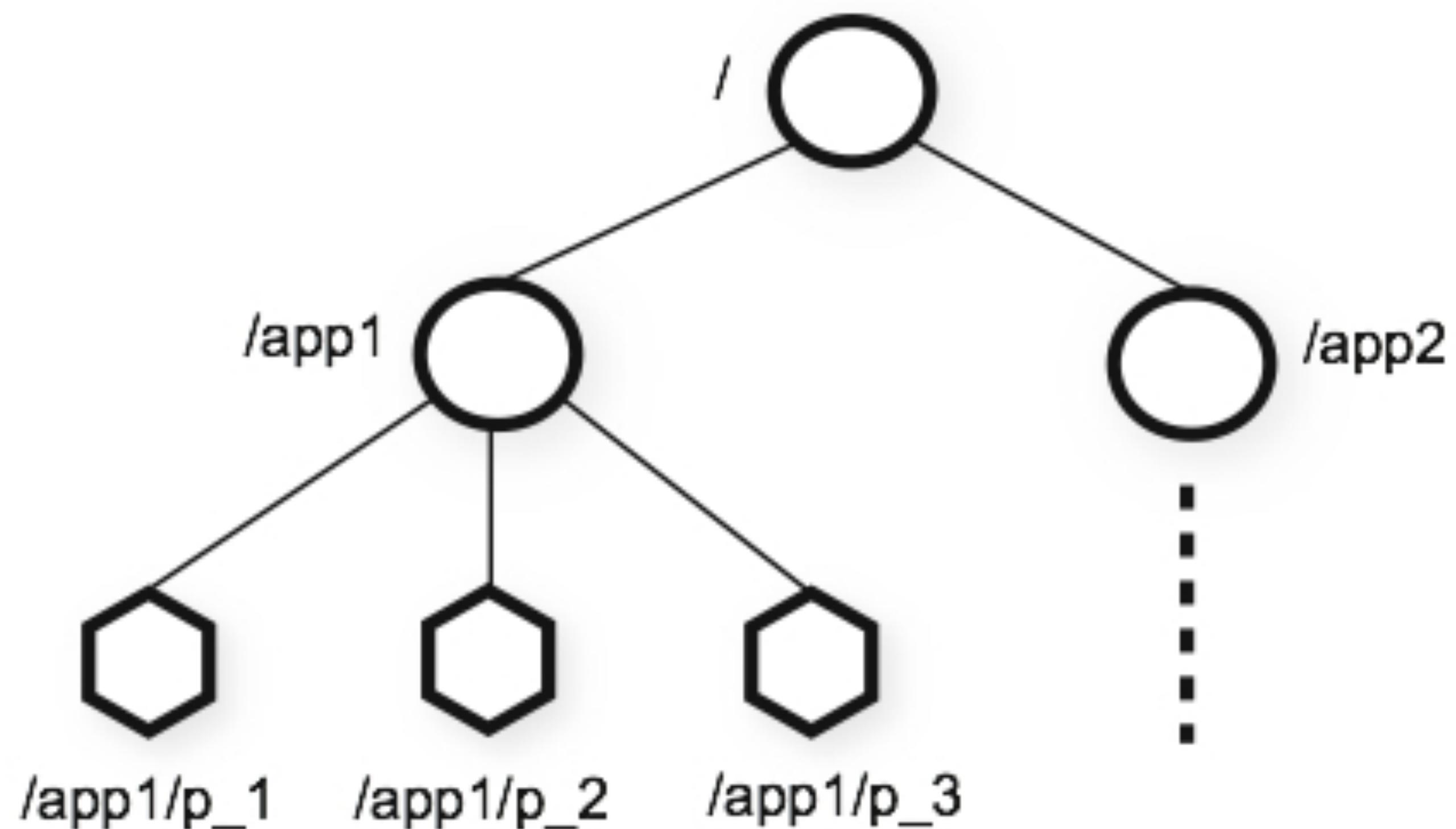


Image taken from Zookeeper docs

Apache Spark Streaming

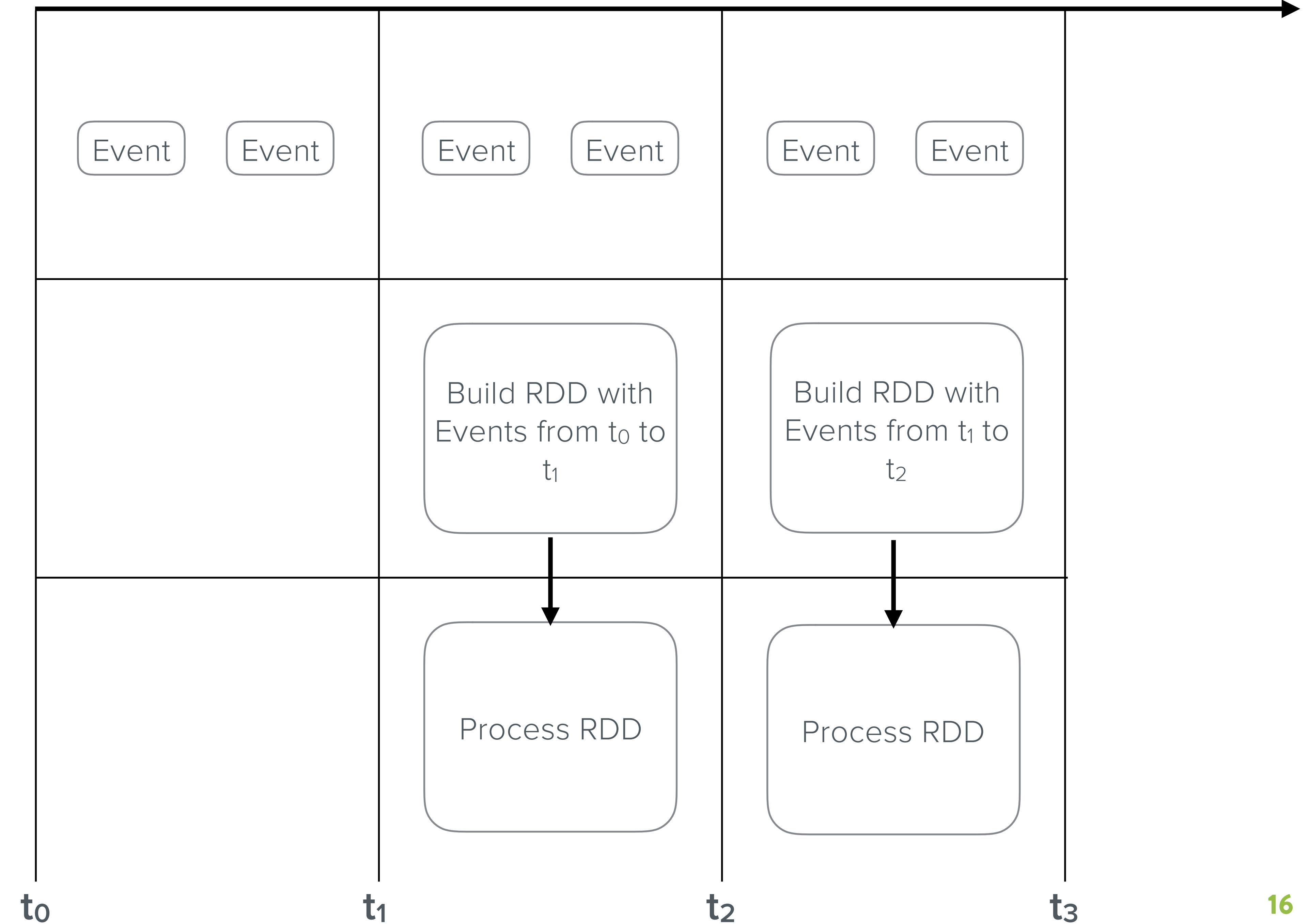


Image taken from mapr.com



OLD WAY: RECEIVERS

Kafka Receiver Spark Engine



Problems w/ Receivers

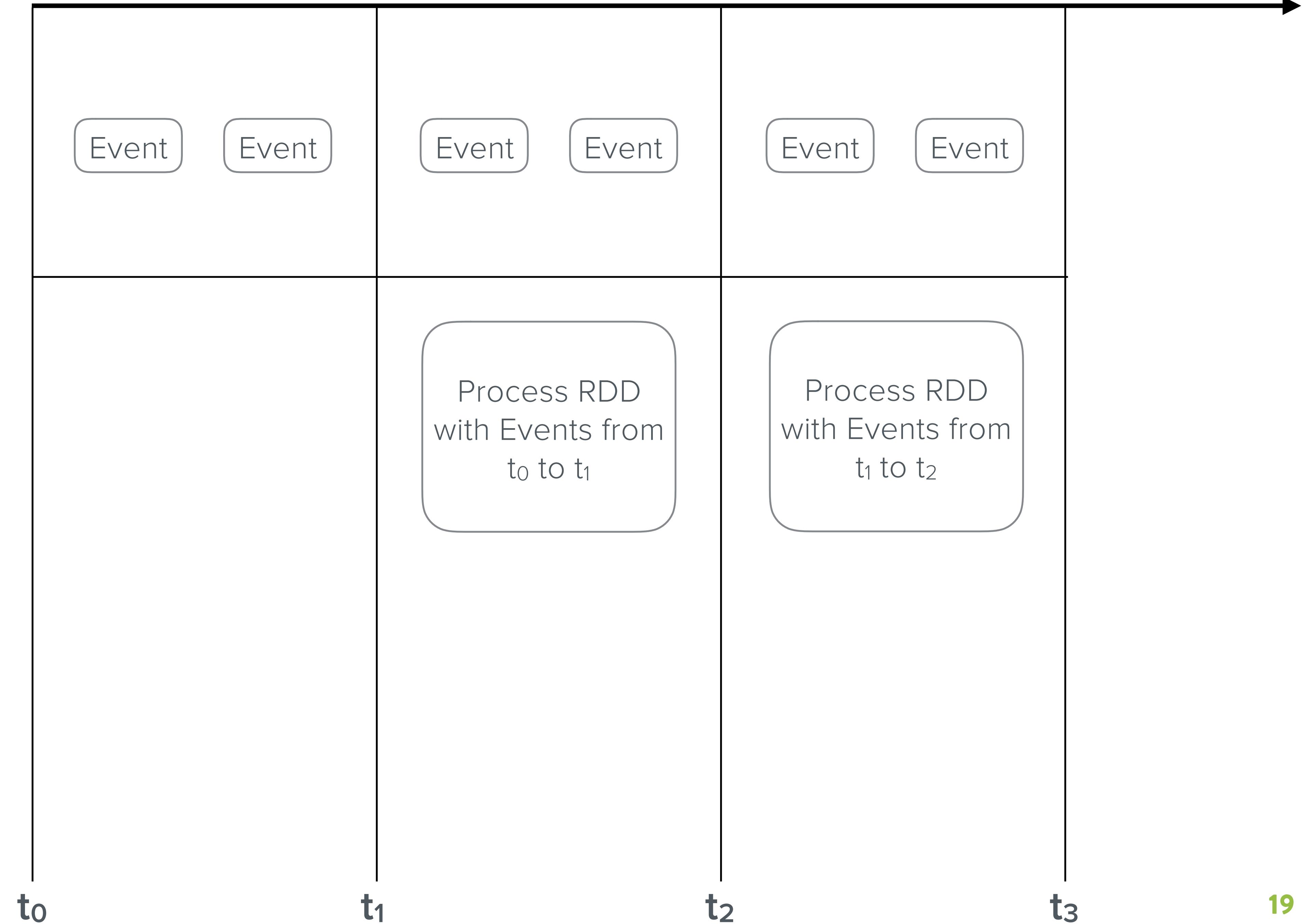
- The only way to get at-least once delivery makes it hard to deploy new code
- Zookeeper is updated with which offsets to start from when data is received, not when it is processed
- We're actually duplicating Kafka



NEW WAY: RECEIVERLESS

Kafka

Spark Engine



General Structure

- Load Kafka offsets from Zookeeper
- Tell Spark Streaming to create a DStream that consumes from Kafka, starting at the specified offsets
- Define your processing step (ie. filter out non-risky events)
- Define your output step (ie. POST the data to the case management software)
- Save Kafka offset of most recently processed event to Zookeeper
- Start your streaming application, and grab some popcorn!

Example Filtering: Risky Products

hair extensions

gucci

vaporizer

wifi pineapple

pharmacy

travel package

iPhone

gateway card

cannabis

Risky Products

RDD for Time 0

hair extensions
nice shoes
sweet bag
cannabis
taylor swift t-shirt
gucci

Filter

hair extensions
cannabis
gucci

Map

{"title": "hair exten..."}
{"title": "cannab..."}
{"title": "gucci"..."}

HTTPS Post

Case
Management
Software

```
return products\  
    .map(self.coalesce_product_body_html)\  
    .filter(self.shop_on_shopify_payments)\  
    .filter(self.search_for_high_risk_words_in_products)\  
    .filter(self.remove_blacklisted_shops)\  
    .filter(self.is_not_verified_shop(verified_shops_bc))\  
    .map(self.format_record)\  
    .map(self.format_triggers)\  
    .map(self.OUTPUT.project_row)
```

TIME-WINDED AGGREGATIONS

The Future

- **Time-Windowed Functions** – A necessity for most of the non-trivial jobs
- **Performance Tweaks** – We haven't spent any time on this, so lots of potential for gains
- **Machine Learning** – We could use the Risk Analyst decisions to build a ML model
- **Improved Monitoring** – We are only monitoring the basics right now
- **Apache Cassandra** – Others use it as a fast key/value store for their jobs
- **Improved Receiverless API** – An API to access Kafka / Zookeeper without hard work

Icon Credits

Credit Card by Rediffusion from the Noun Project

Money Bag by icon 54 from the Noun Project



WE'RE HIRING
[SHOPIFY.COM/CAREERS](https://shopify.com/careers)





@n_e_evans

THANK YOU!

