MYTOYS GROUP

Exploring the alchemy of Streaming and Solr cloud Bbuzz 21

By
Atita Arora

https://www.linkedin.com/in/atitaarora/

https://twitter.com/atitaarora

Agenda

- Introduction
- Statistics
- Problem Area
- Proposed Solution
- Improvements
- Questions/Suggestions?

Who am I ?

MYTOYS GROUP

Engineer / Architect / Mother of 2

Active member of Search community / Avid traveller / Love for Nature

Solr consultant helped teams to build/improve the index pipeline , optimizing search experience , build / improve search model , migration *upgrade search platform*.

About MyToys GmbH

MYTOYS GROUP

Founded in 1999 , is a subsidiary of Otto Group

Germany's best online shop in the child and baby category.

Various other shops dealing with products from pregnancy , everything for kids (new born – high school) , wide range of shoes and home furnishing.

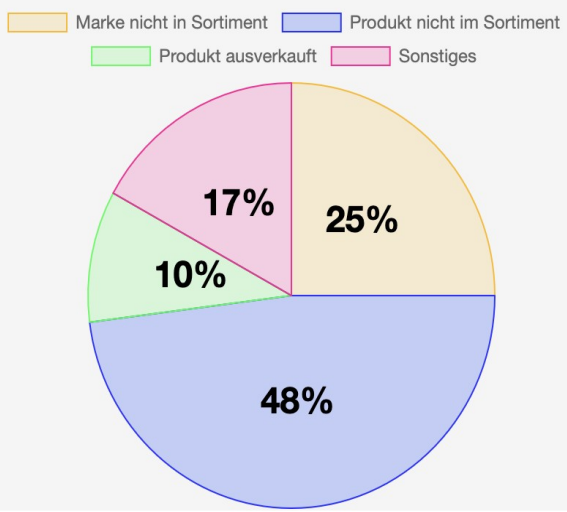
Over 4 million active users

Statistically speaking ..

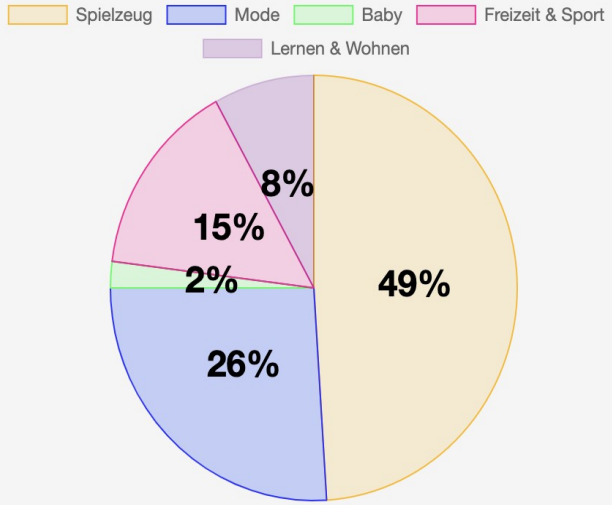
MYTOYS GROUP

over 450 million Search queries	over 220,000 Searches for the "Playstation 5"	51% PSP2PDP (Click-through rate)	321 million sales with the search
2.68% Zero hit rate (50% less than FY 2019/20)	1.23 million ⌀ searches per day	over 286 million sales with the top 500 searches	8.57% Search conversion rate

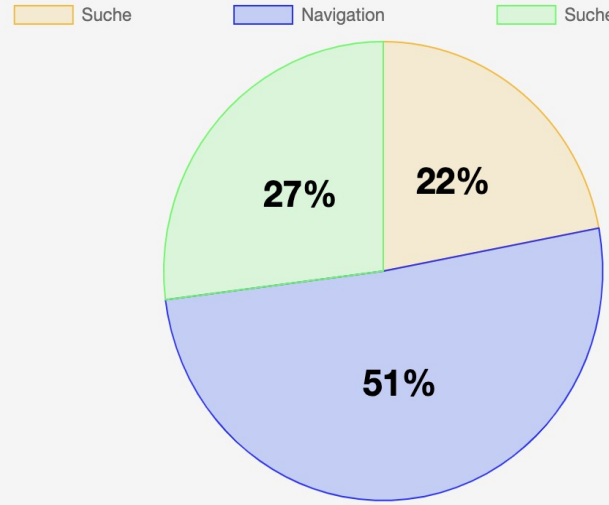
Zero hit reasons



Our customers are looking for ...



Customers who browse through myToys use ...



MYTOYS GROUP

The most searchable days in 2020

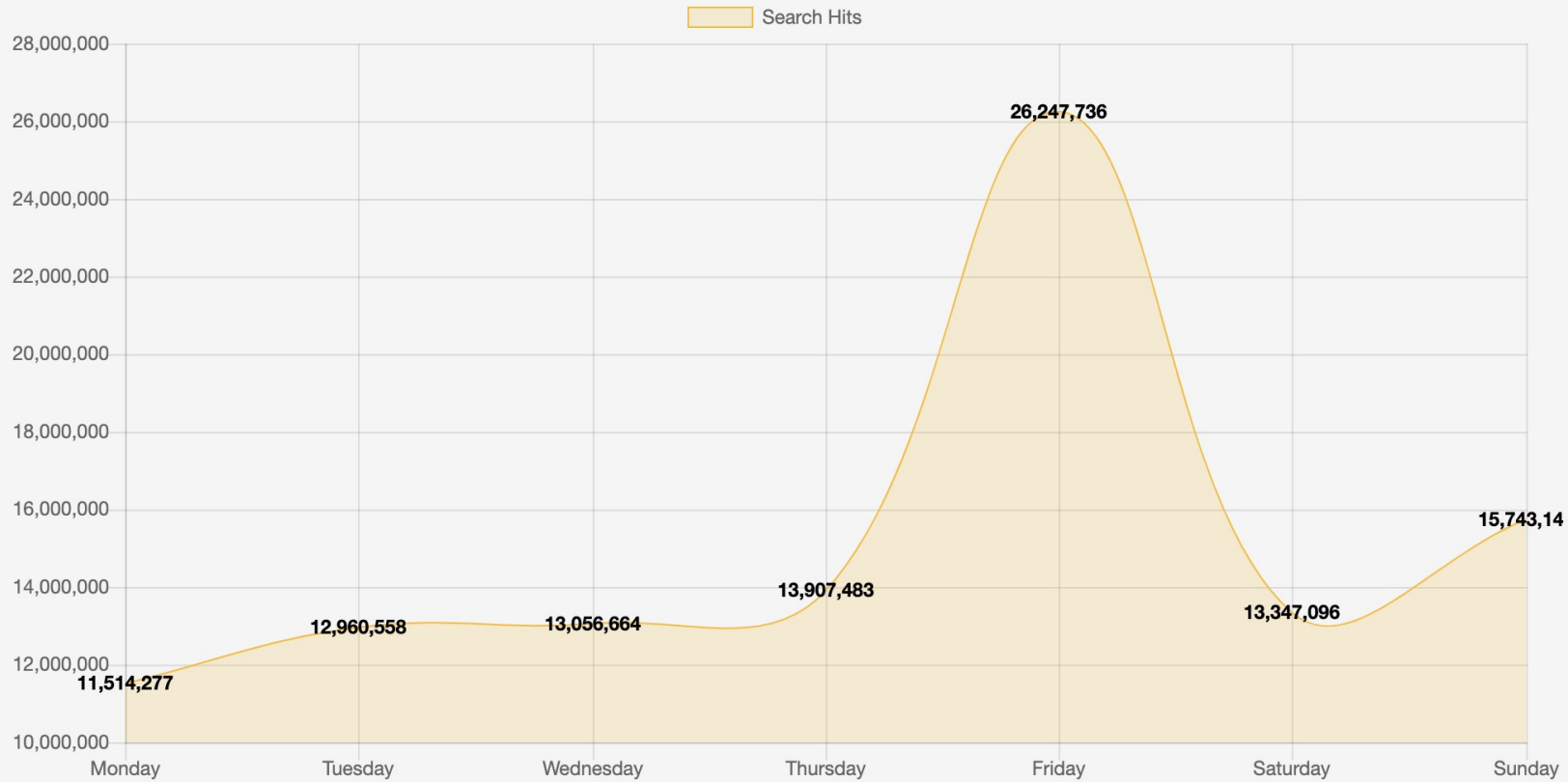
by number of searches

1 Black Friday 11/27/2020 26.2 million

2 Singles Day 11/11/2020 15.9 million

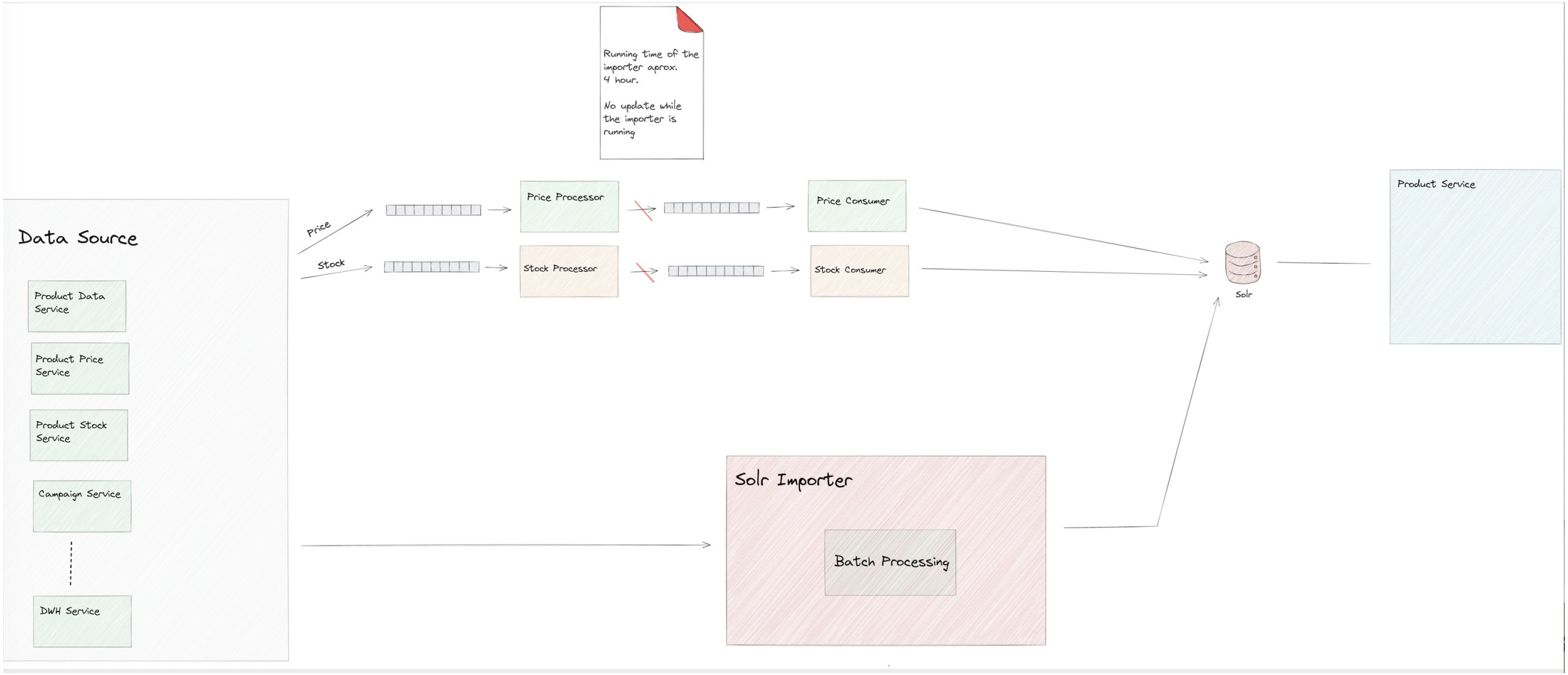
3 Black Friday Weekend 11/29/2020 15.7 million

Black Friday search history



MYTOYS GROUP

About the platform



Shortcomings

1. Limited number of products due to limited resources

2. Complete catalog index – 2 x day (taking close to 4 hrs)

3. Price / Stock updates every 5 mins (ONLY for existing products)

4. No product updates while catalog reindex

5. Scaling is always worrisome

6. No disaster recovery mechanism

7. Bulky document size (~ 550 fields) - Data served from Solr

8. Solr 6.2 Master -Slave

The Plan

1. Keep only the searchable data in Solr
   Redis to support Solr

2. Reduce Pipeline time
   4 hrs → NRT

3. Manage index footprint
   548 fields → 78 fields

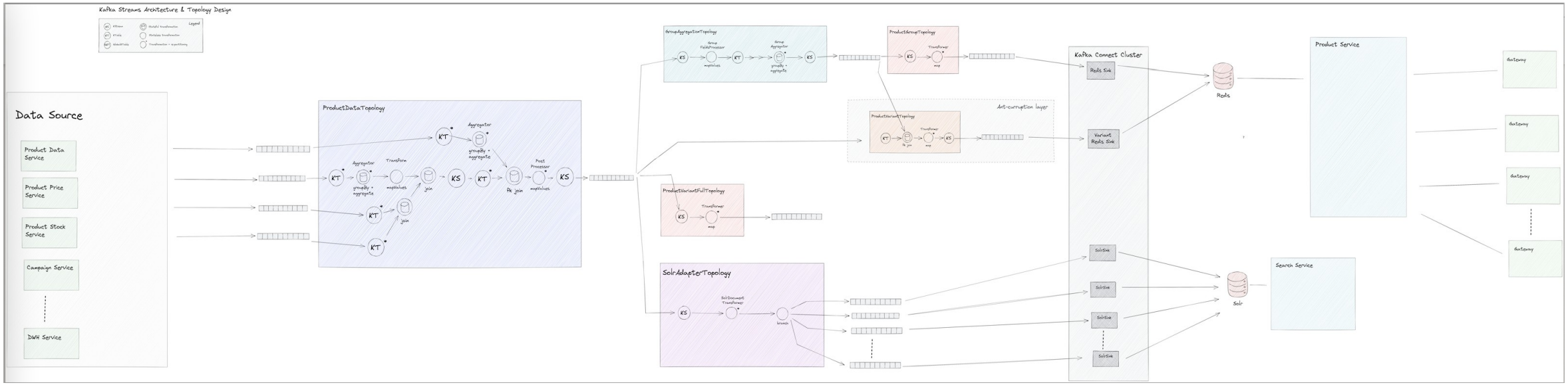
4. Disaster recovery achieved through
   Kafka replay using Kafka-Connect

5. New Article/Article-Product-Data-Update processed in Near Real Time
   No wait for the full index

6. Infra scalability Management through K8s

Architectural Design

Core Streaming Platform Breakdown & Progress



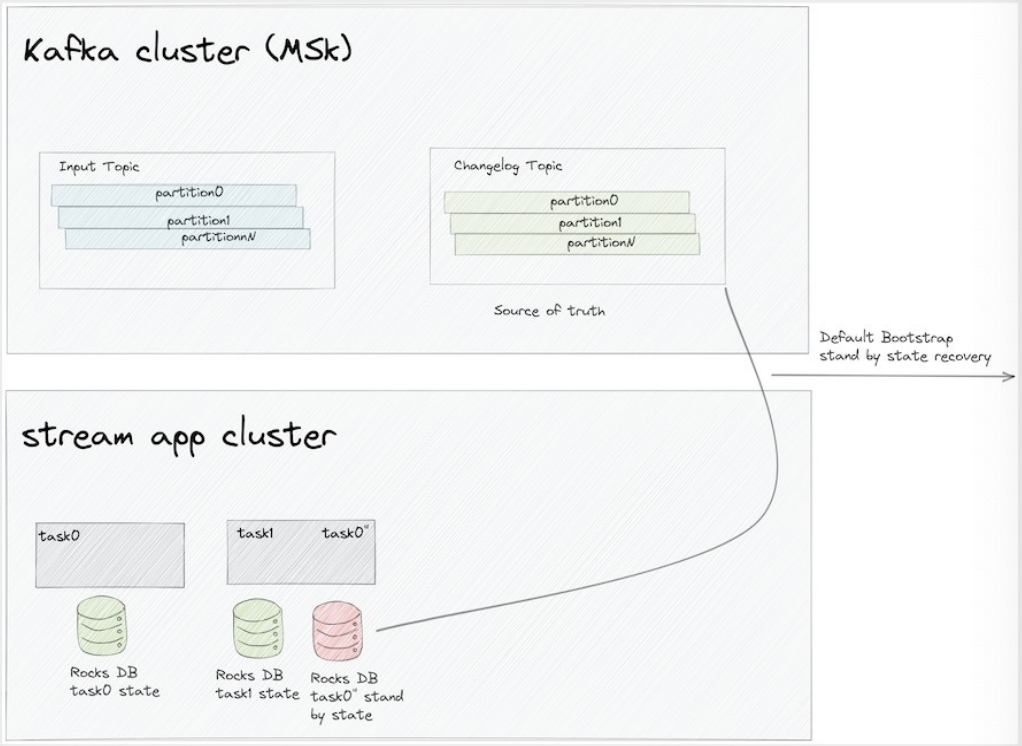
Kafka Streams' Challenges

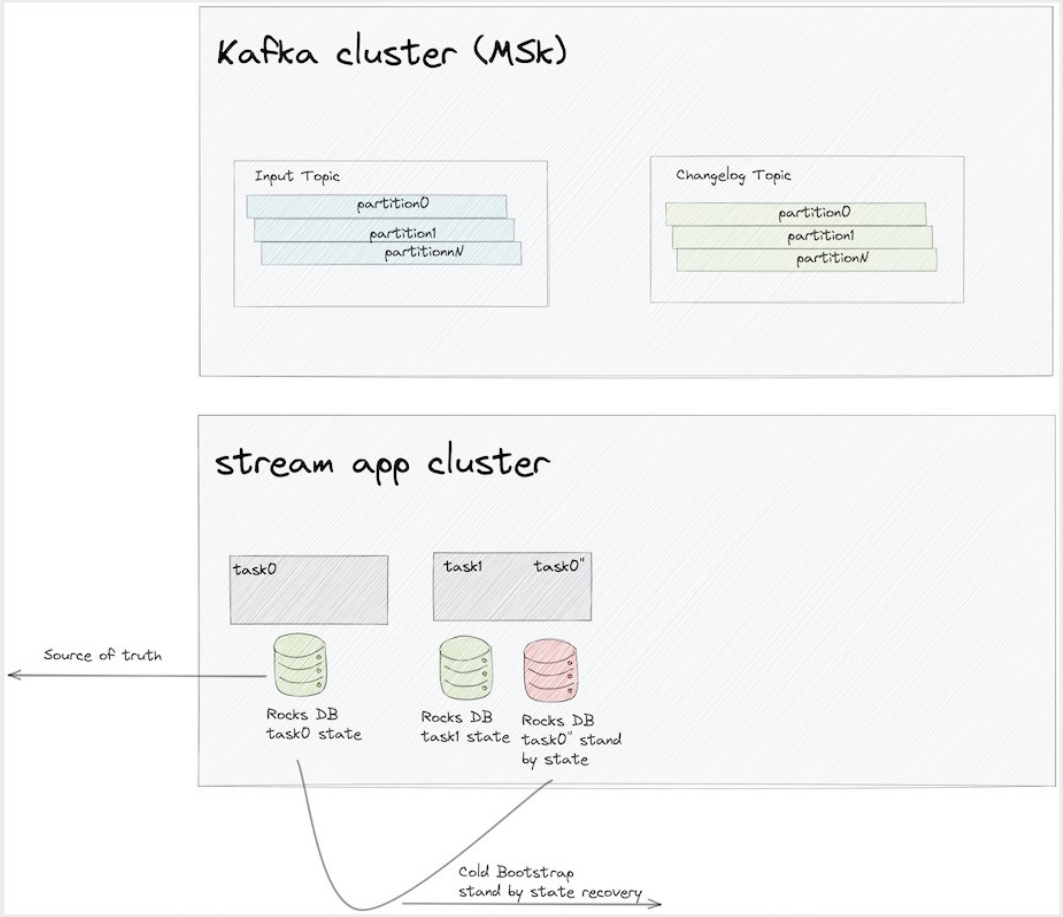
MYTOYS GROUP

1. Slow and expensive kafka state restoration (Cold bootstrapping)
2. Kafka-Solr Connect (dealing with deletes)
3. Horizontal Scalability → based on metrics (consumer lags)
4. RocksDB – Stores the stream state (default rocksdb) with custom config
5. Large Clusters - Maintaining 4 high volume topologies (managing data ca 100GB)
6. Migration from Self Managed Kafka → MSK (Mirrormaker)

Kafka Default State Recovery Vs Cold Bootstrapping

MYTOYS GROUP





And what did that improve ?

MYTOYS GROUP

Product updates in real time through switching from batch → streaming

Index Pipeline Time from 4 hrs → Near Real Time

Upgraded to Solr 8.8.2

Moving from Master-slave to Cloud

Switched to managed kafka

Using Redis to support Solr with non-changing fields

MYTOYS GROUP

Future work

Planning to open up to community for respective services :

1. Solr client for Kafka

2. State recovery lib

3. Kafka connect cutomisations

Any questions?

References

1. https://github.com/jcustenborder/kafka-connect-solr

2. https://kafka.apache.org/documentation/streams/

3. https://www.confluent.io/kafka-summit-ny19/kafka-streams-at-scale/

4. https://atitaarora.medium.com/solr-upgrade-from-version-6-x-8-x-a-road-from-master-slave-to-solr-cloud-fuss-free-fb9ac4c1b38f