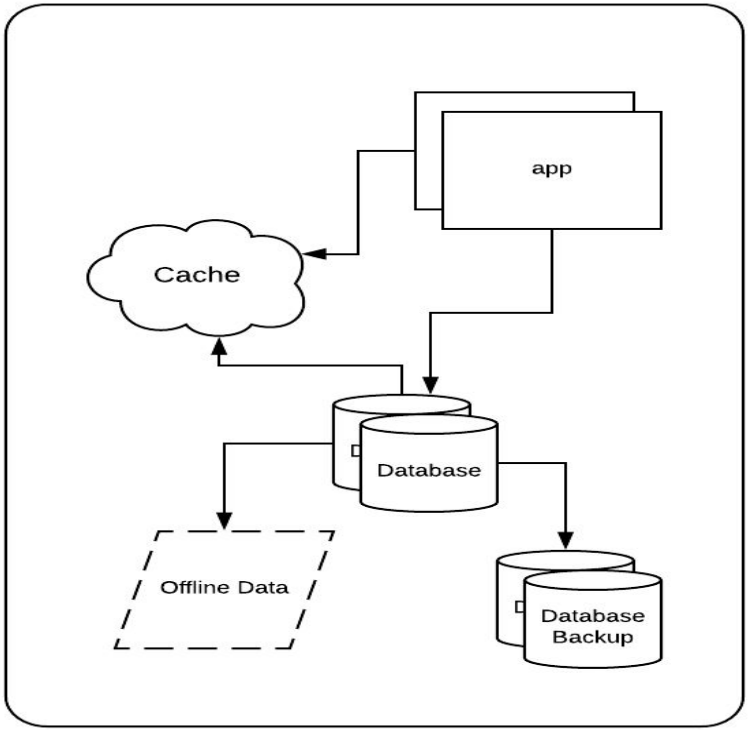
Track that Clone: Near-realtime data audit for distributed data replication

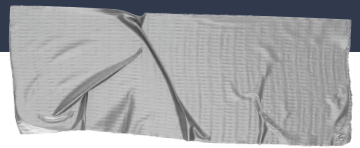
Janardh Bantupalli
Senior Staff Engineer

Srivathsan Vijaya Raghavan
Senior Engineer

Cute and Happy!

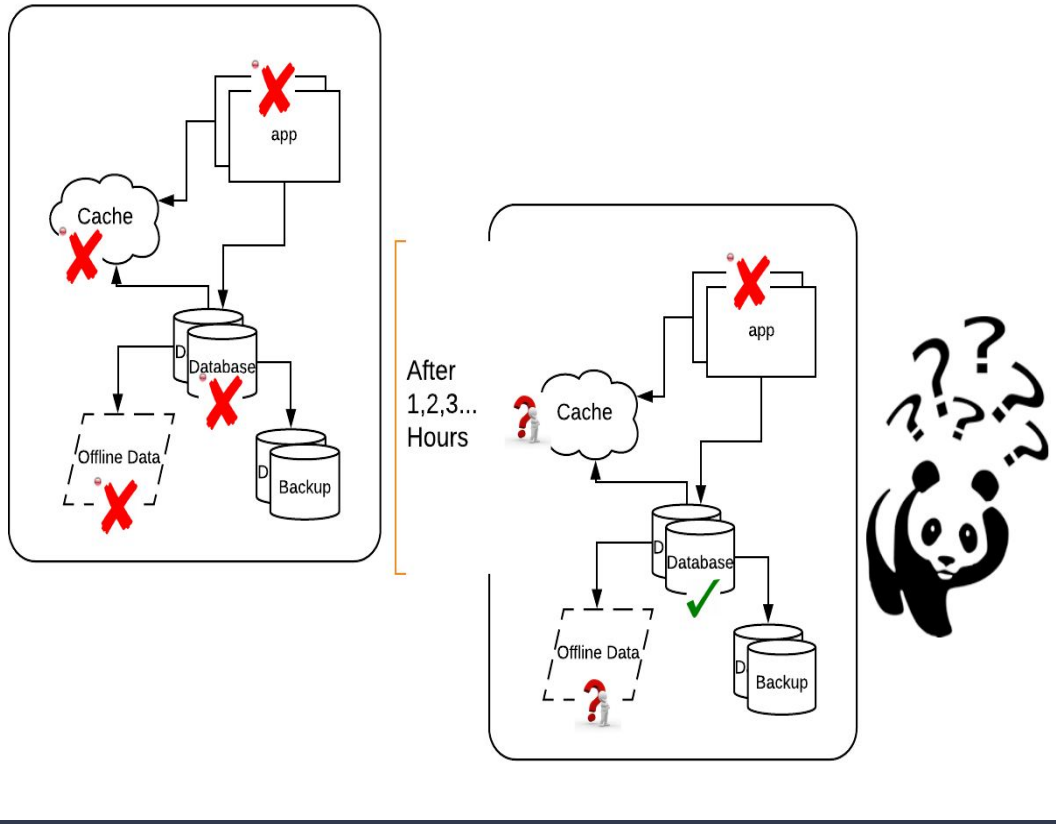
➔ **Everything on single data center!**



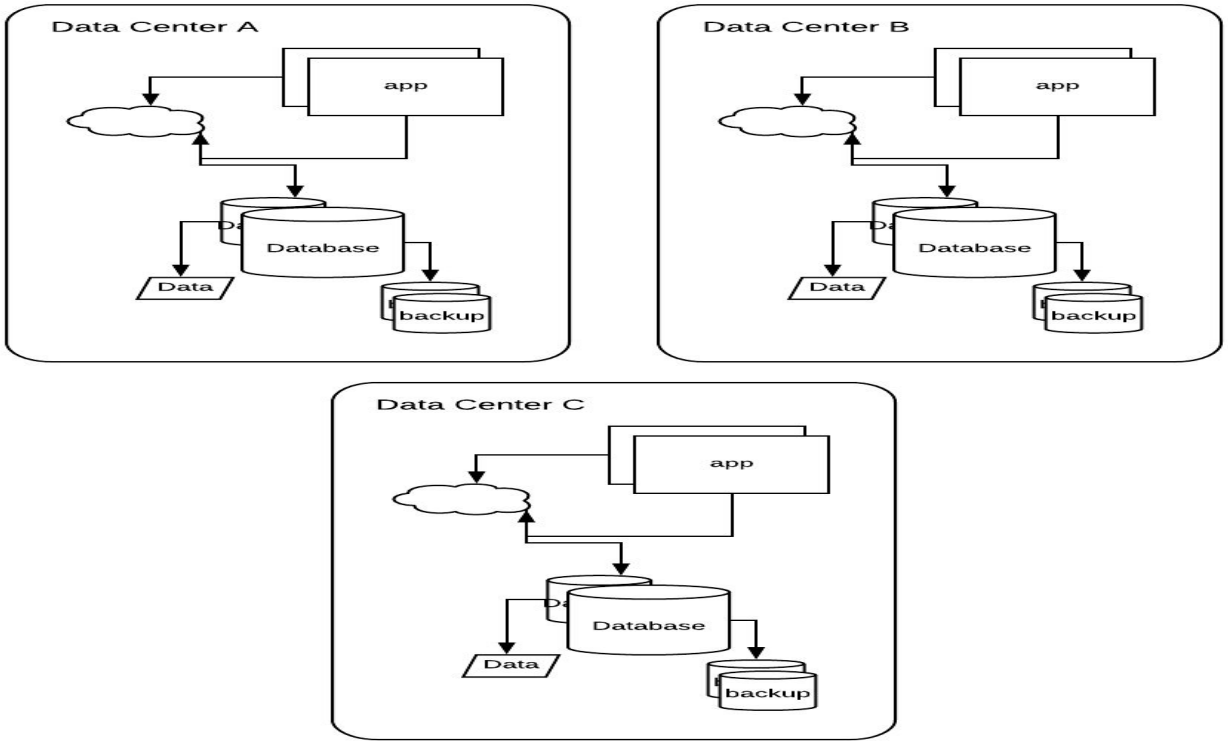


Not-So-Happy!

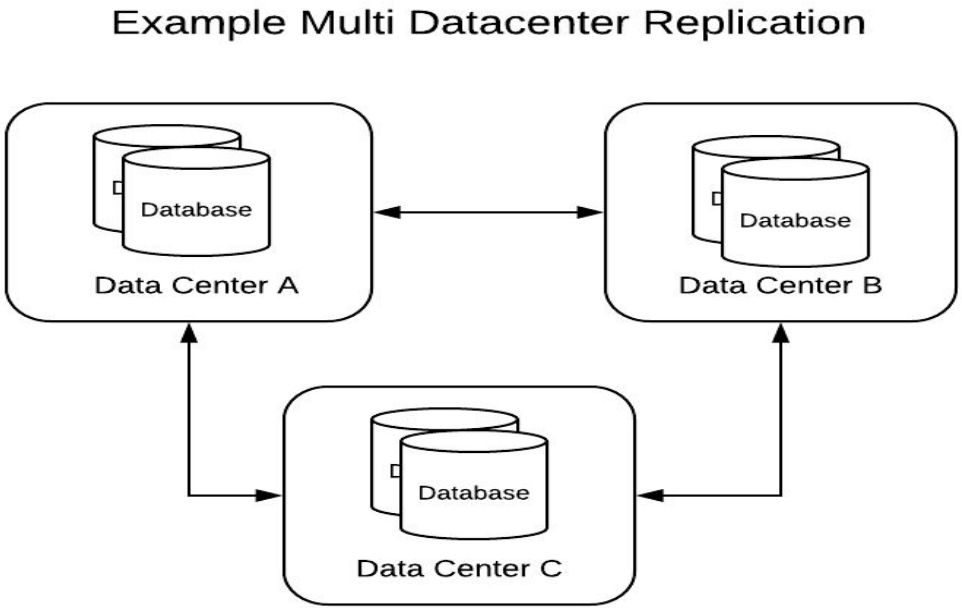
➔ **$hit happens!**



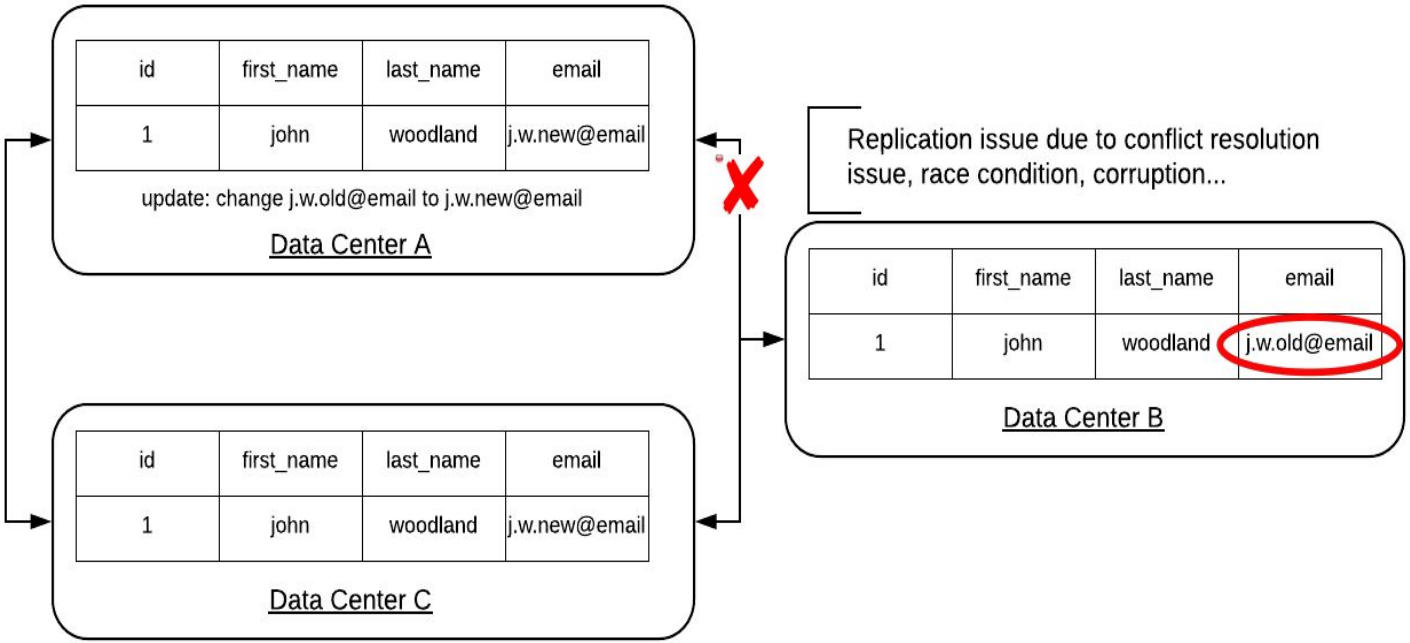
Hurray! High Availability!



Data High Availability!

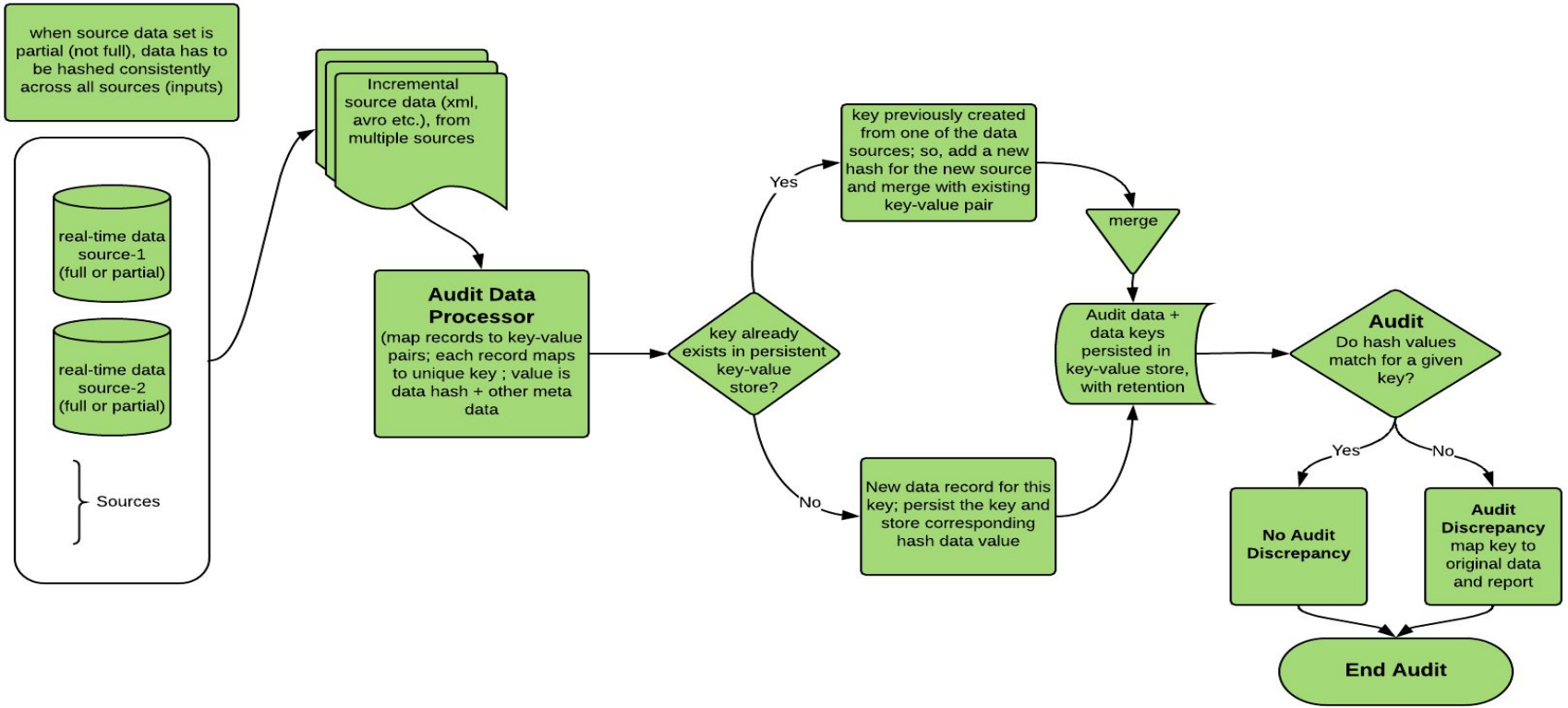


High Availability?!

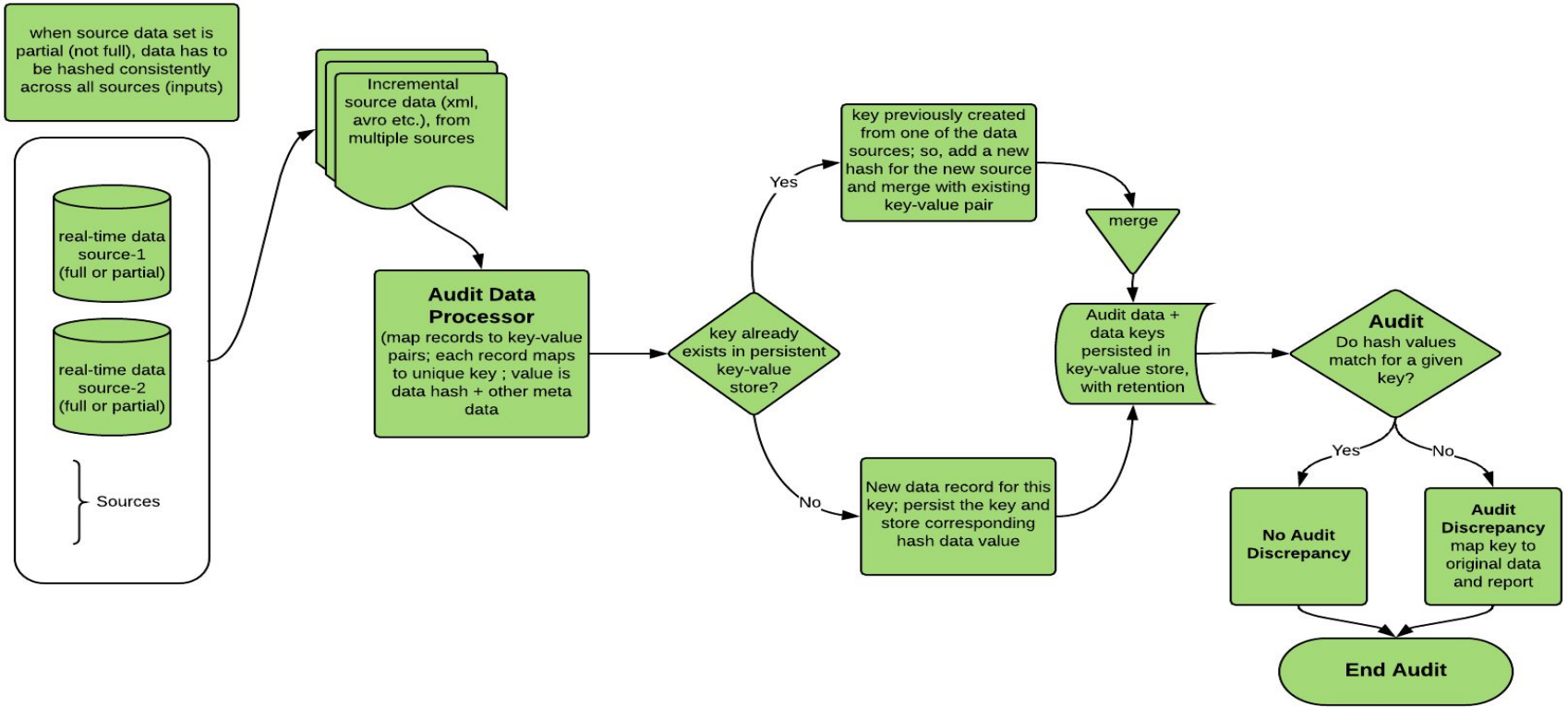


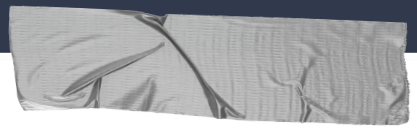
- ★ Near real-time data validation across multiple locations
- ★ Proactive discrepancy detection and reporting
- ★ Better data quality for downstream consumption

Real-time Data Audit Framework



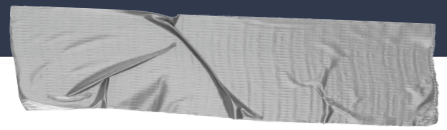
Real-time Data Audit Framework





What's Required?

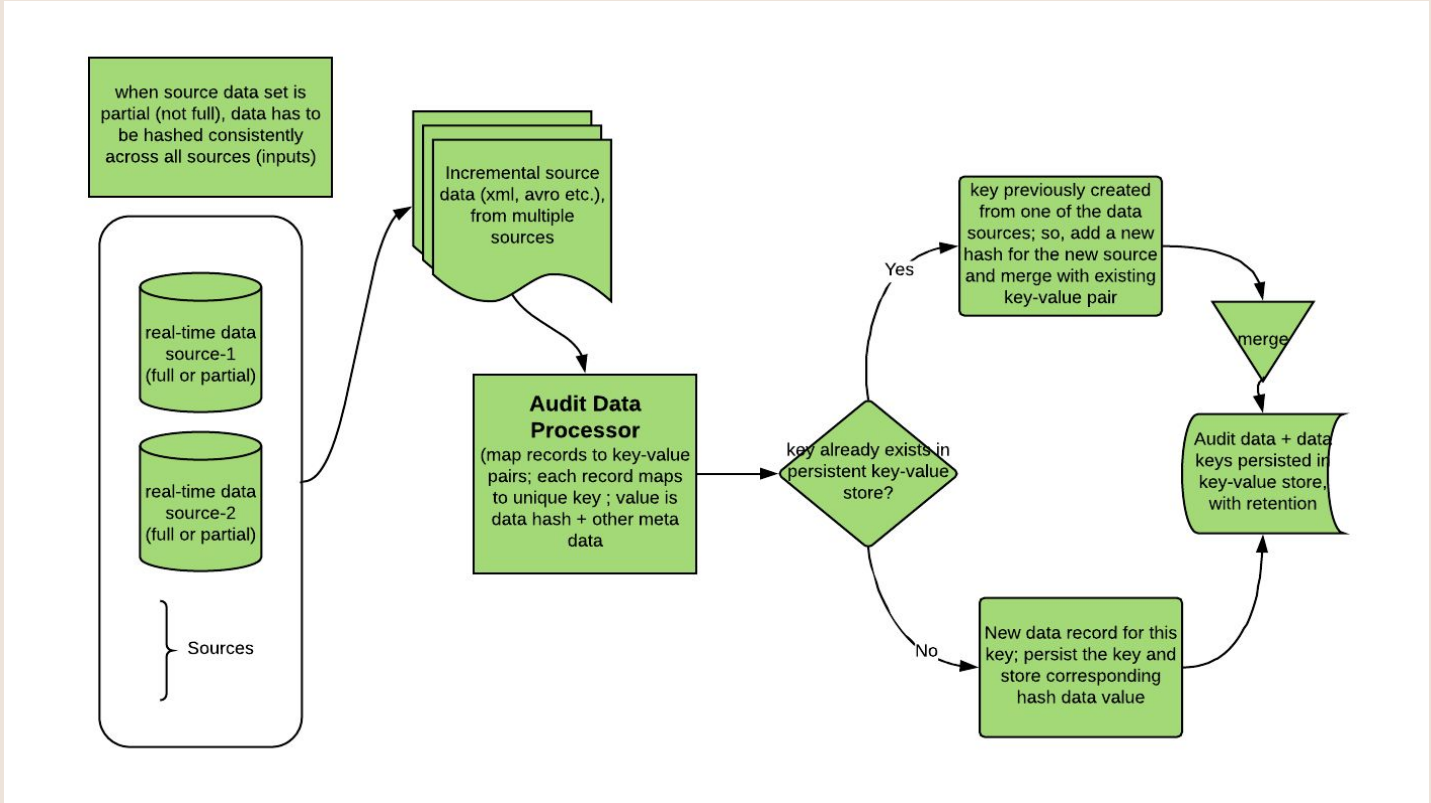
- Incremental data input from sources
  - full (or) partial data set
- Last modified timestamp value that stays consistent across all sources
  - this field would typically be used for conflict resolution
- Persistence store to facilitate isolation of data processing and audit



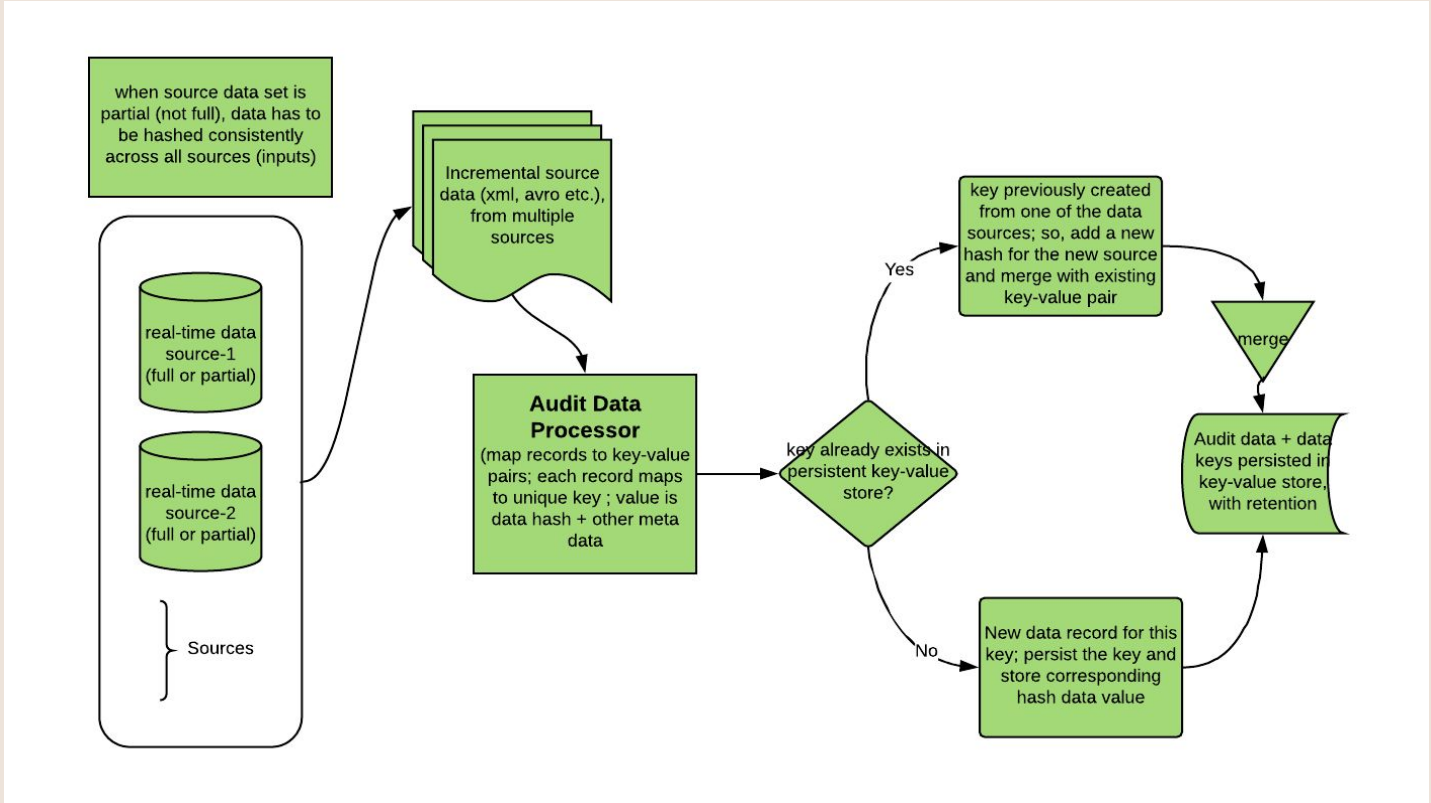
What's Required?

- Incremental data input from sources
  - full (or) partial data set
- Last modified timestamp value that stays consistent across all sources
  - this field would typically be used for conflict resolution
- Persistence store to facilitate isolation of data processing and audit

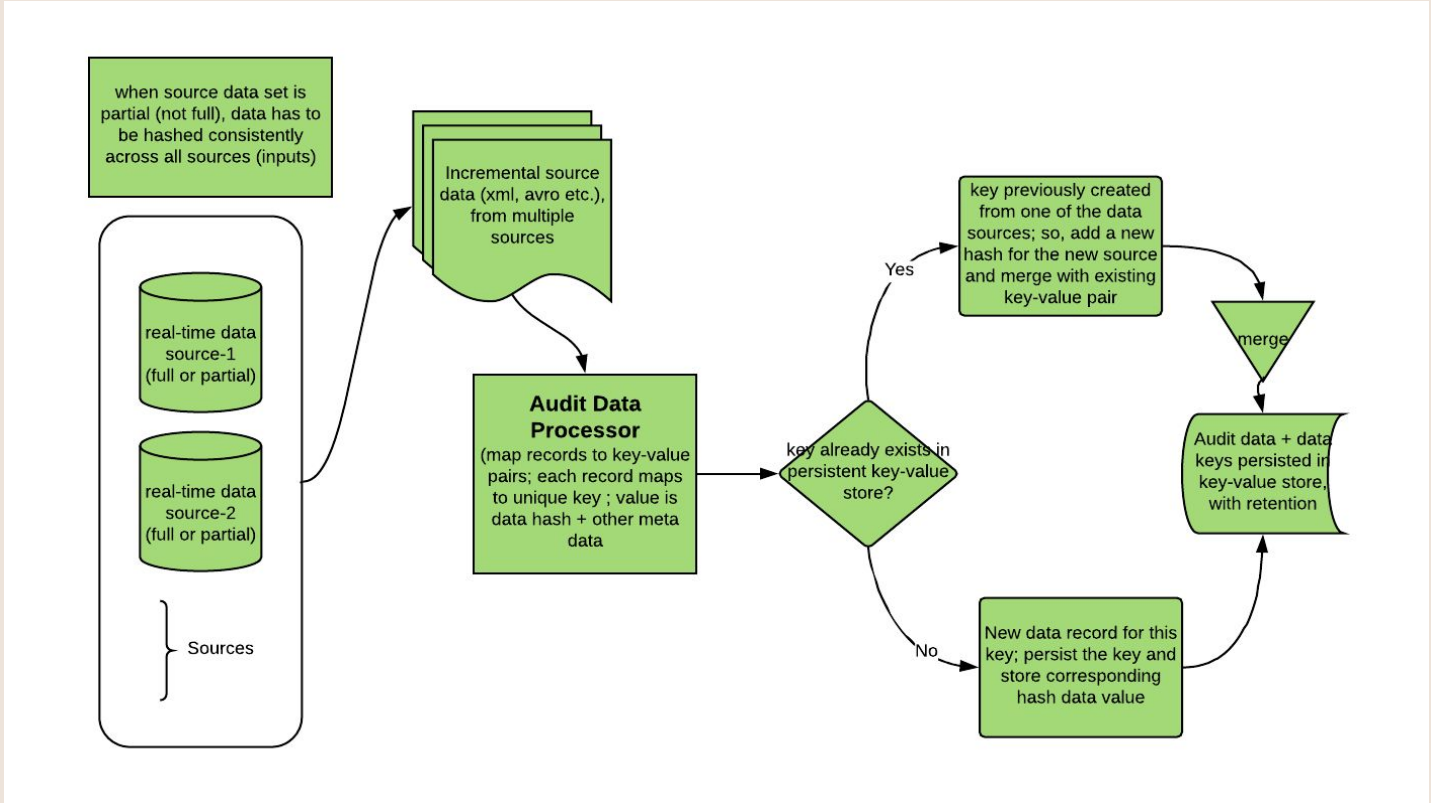
Data Processor



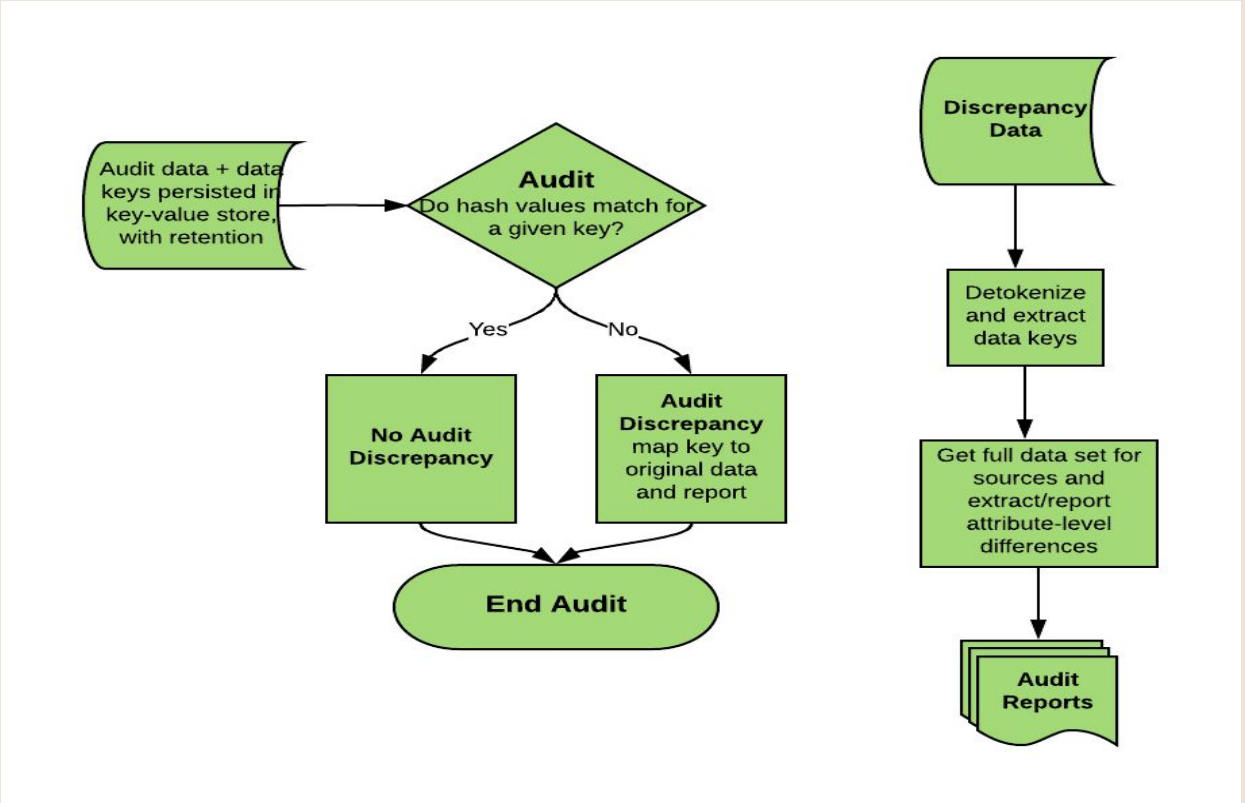
Data Processor



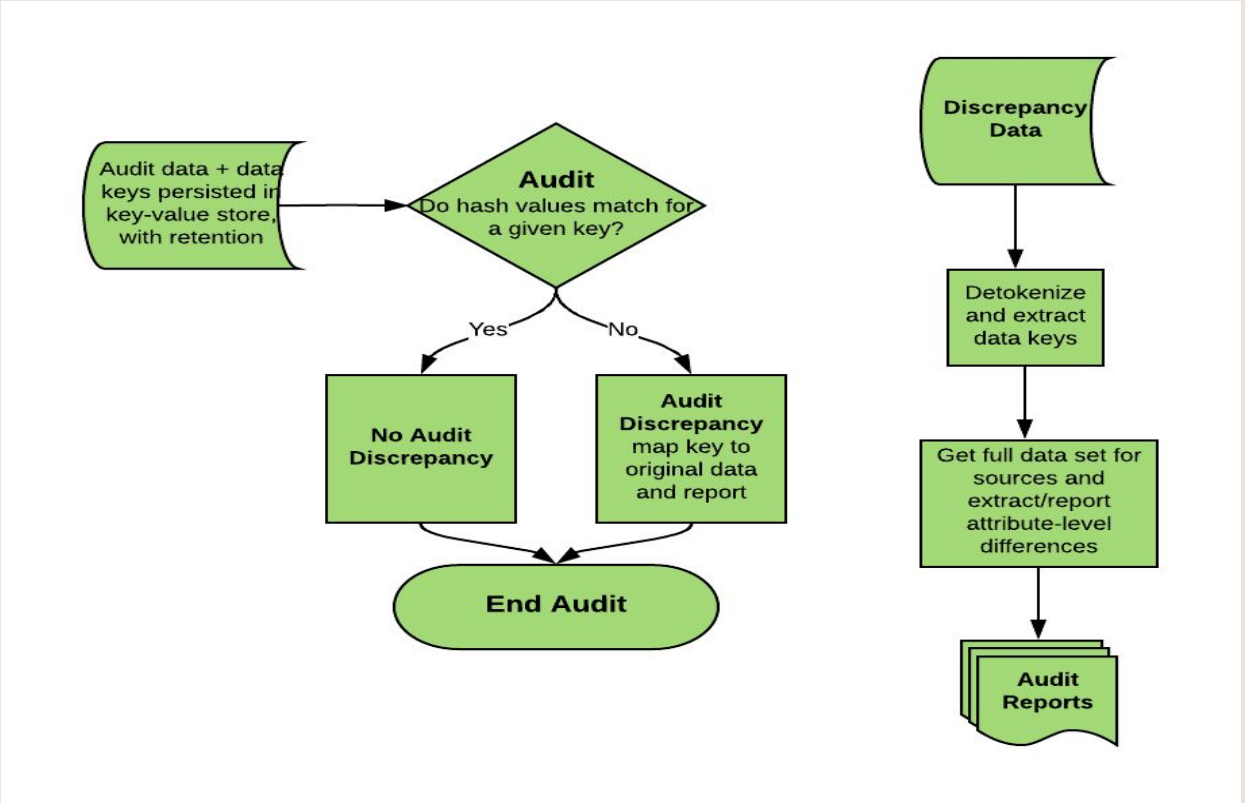
Data Processor



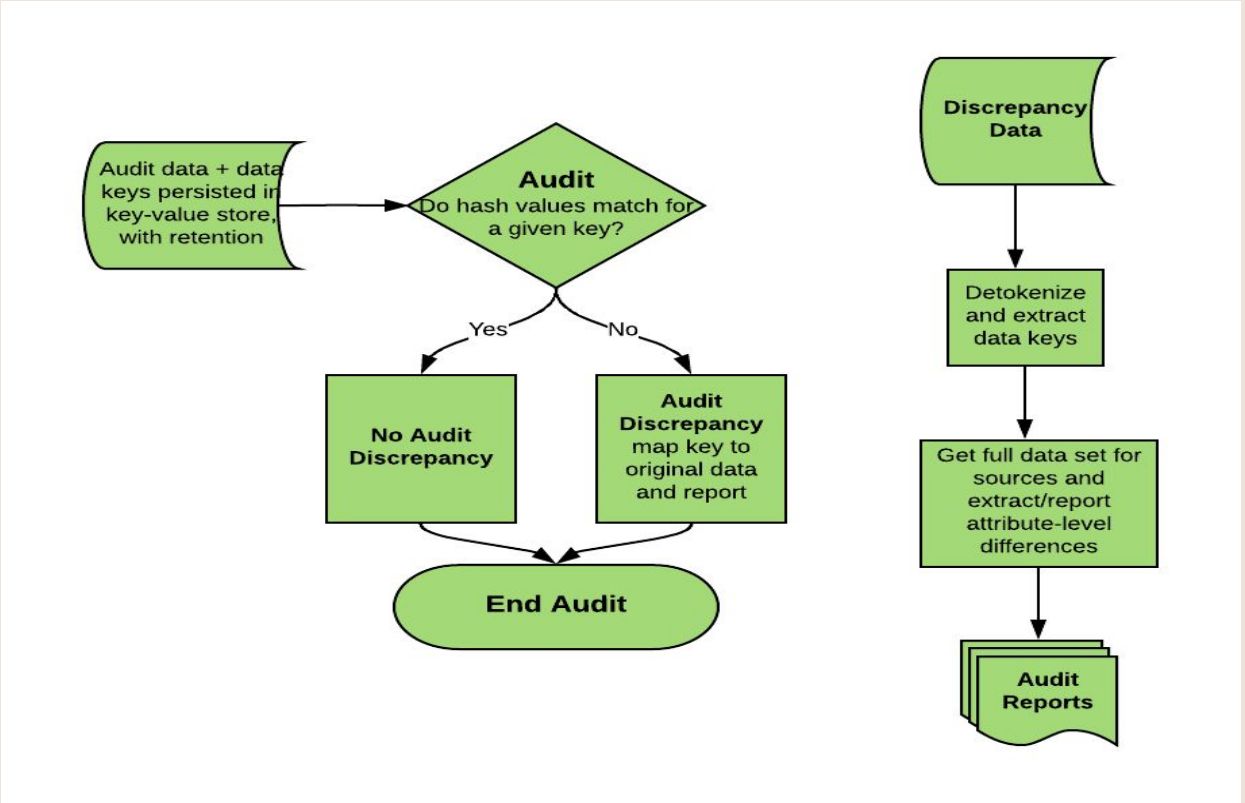
Audit Processor



Audit Processor



Audit Processor



Features

- ➔ Deduplication and Checkpointing
- ➔ Recovery from failures and Replay
- ➔ Data bucketing window customization to allow more deduplication
- ➔ Synchronization with replication and source data watermarks

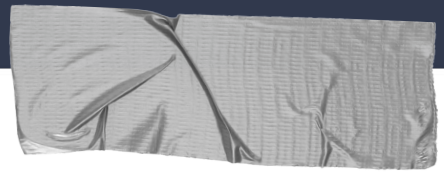
Features

- ➔ Deduplication and Checkpointing
- ➔ Recovery from failures and Replay
- ➔ Data bucketing window customization to allow more deduplication
- ➔ Synchronization with replication and source data watermarks

Live at [in]

Active-active multi data-center replication for online relational data

Exploring similar data validation for our nosql data stores and other distributed data flows



Thank You!

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

https://www.linkedin.com/in/srivathsan-vijayaraghavan