

# **BUILDING BETTER HUNT DATA**

**SANS THREAT HUNTING SUMMIT 2021, JOSH LIBURDI**

# AGENDA

## OR ... DATA: THE GOOD, THE BAD, AND THE UGLY

- » How should we evaluate data quality?
- » Why do we want high quality data?
- » What are signs of low quality data?
- » How can we improve data quality?



# BACKGROUND

- » Experience: 8+ years in detection & response, including hunting and systems engineering
- » Work: Security Engineer @ Brex
- » GitHub/Medium/Twitter: @jsh1brd





# GOALS FOR THIS TALK

» Threat Hunters

» "Do we have good data? Could it be better?"

» Security / Data Engineers

» "Do our systems provide the best data possible?"

» Security Leaders

» "I should ask about the quality of our data!"

# EVALUATING DATA QUALITY

# OR ... WHAT IS GOOD DATA?<sup>1</sup>

- » Accuracy
- » Completeness
- » Consistency
- » Timeliness
- » Uniqueness
- » Validity

<sup>1</sup> [https://threathunterplaybook.com/pre-hunt/data\\_quality.html](https://threathunterplaybook.com/pre-hunt/data_quality.html)

# **BENEFITS OF HIGH QUALITY DATA**

# INCREASED EFFICIENCY & IMPACT!

- » Reduces time and complexity of going from hypothesis to analysis
- » Improves trust in analysis
- » Increases impact hunt has on other groups, especially detection engineering
  - » Collaboratively share content
  - » Cooperatively improve data



# **SIGNS OF LOW QUALITY DATA**

# WARNING SIGNS

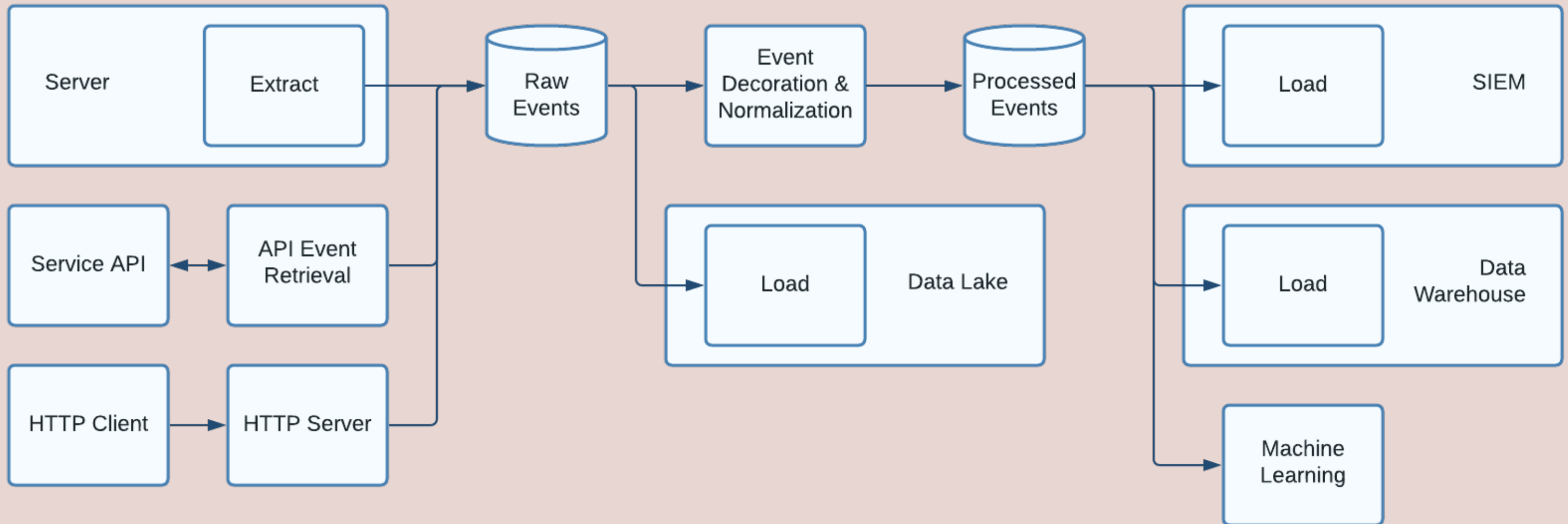
- » You look for data that doesn't exist
- » You can't find data that you know is there
- » You wait, and wait, and wait for data to arrive
- » You triple check your results
- » You spend more time in data prep than analysis

# AD HOC DATA PREPARATION

- » Annoyed with converting between data formats?
  - » CSVs haunt your dreams? Terrified of XML?
- » Tired of copy+pasting code to slice field values?
  - » Wasting time tinkering with regular expressions?
- » Sick of adding context?
  - » "Who is 8.8.8.8 anyway?"

# HOW CAN WE IMPROVE DATA QUALITY?

# DATA PIPELINES!

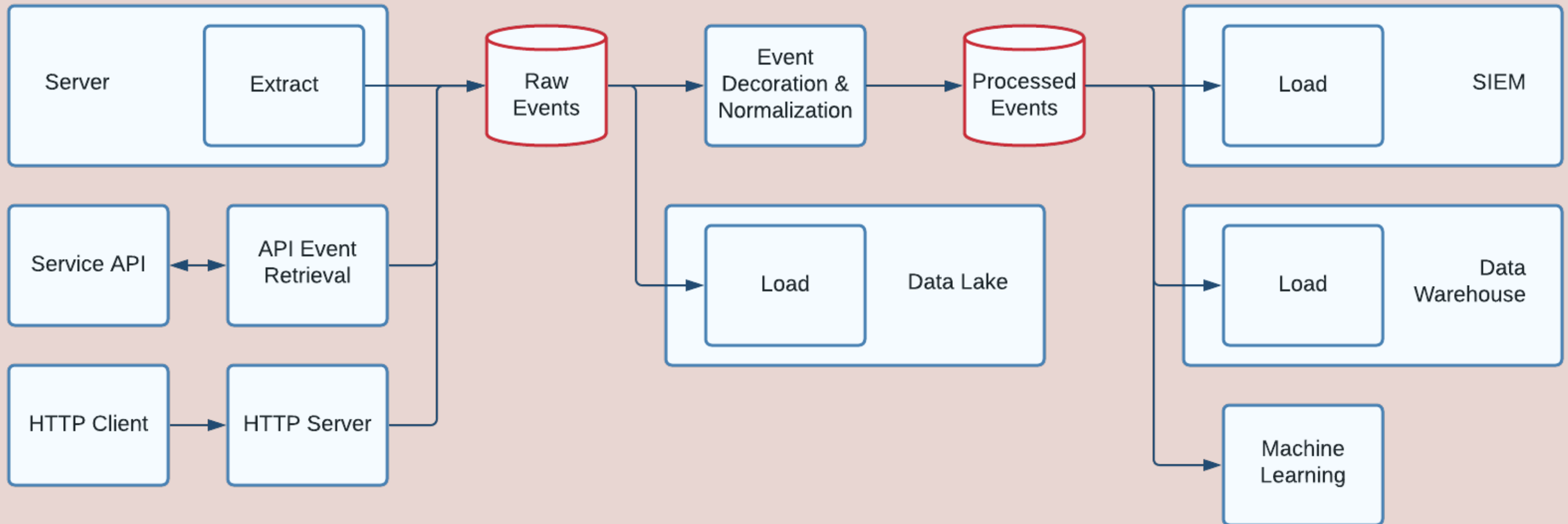


# FOCUS ON ...

- » Availability of data
- » Consistency of data
- » Timeliness of data
- » Completeness of data



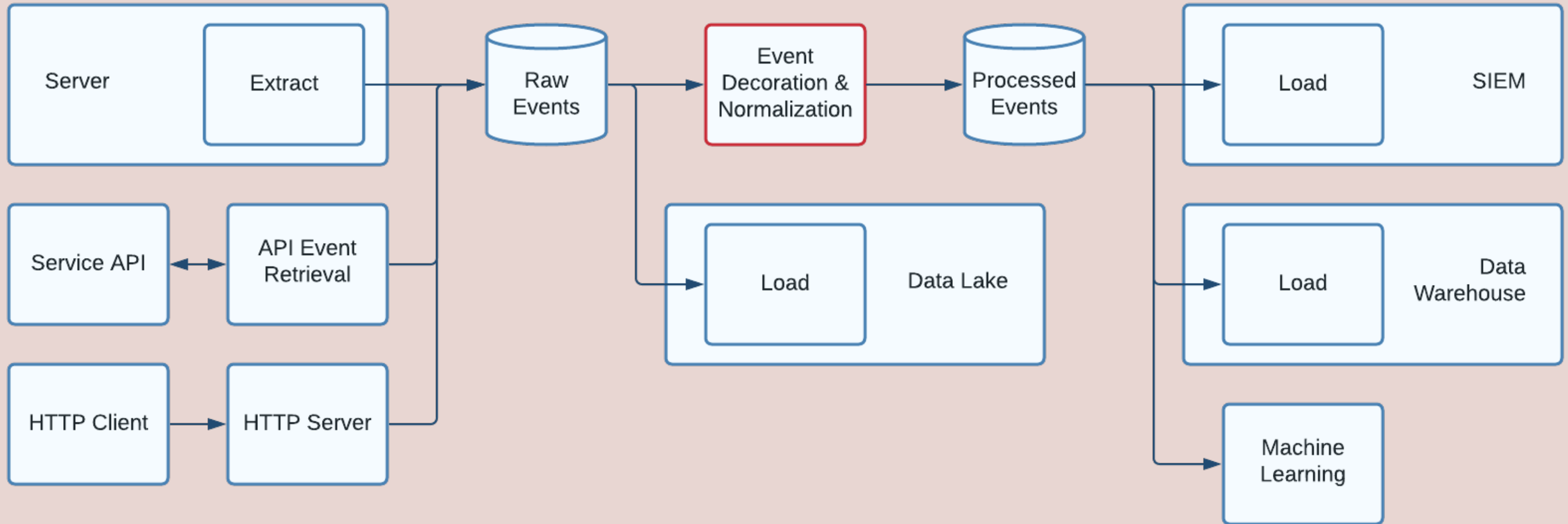
# DATA AVAILABILITY



# DATA AVAILABILITY

- » 2 event streams per dataset
  - » Raw: unmodified
  - » Processed: formatted, normalized, decorated
- » Supports concurrent downstream applications
  - » Filter, selectively load events into each app
  - » 50% into SIEM, 100% into warehouse, 5% into ML

# DATA CONSISTENCY



# DATA CONSISTENCY

- » Formatting

  - » Convert data between formats (e.g, CSV to JSON)

- » Normalizing (Common Information Models<sup>2</sup>)

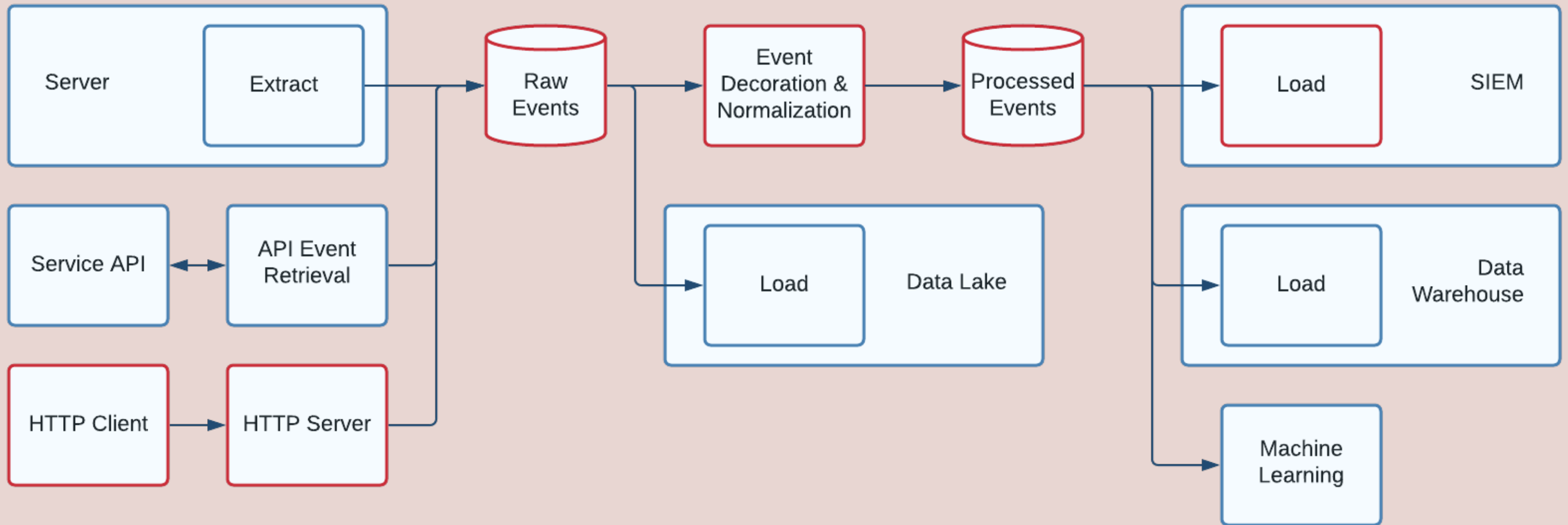
  - » Prefer unified, permissive schemas

- » Decorating

  - » Enrich data with external & internal context

<sup>2</sup> [https://threathunterplaybook.com/pre-hunt/data\\_standardization.html](https://threathunterplaybook.com/pre-hunt/data_standardization.html)

# DATA TIMELINESS



# DATA TIMELINESS

- » Retention

  - » How long should you keep your data?

- » Speed

  - » How soon does your data **need** to arrive?

- » Focus on what, how, who for determining timeliness

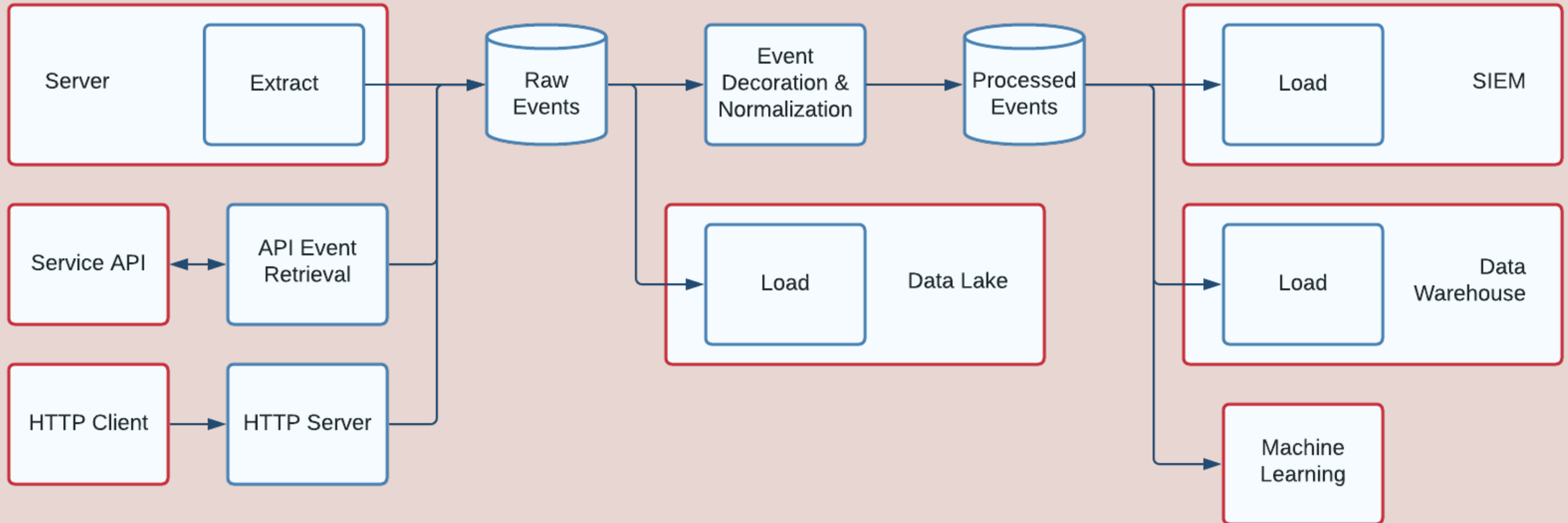
  - » Type of data (endpoint, network, service audit)

  - » Type of analysis (real-time, batch, ad hoc)

  - » End users, staffing model (24x7 vs 12x5)



# DATA COMPLETENESS



# DATA COMPLETENESS

## » Coverage

- » What % of systems delivered data?

- » Compare data against trusted sources

## » Reliability

- » What % of data was delivered? lost? malformed?

- » Test with labeled, scheduled data (e.g. tracers, simulated attack data)

# SUMMARY

# SUMMARY

- » Actively think about improving data quality
  - » Remember the signs of low quality data
- » Monitor & continuously improve data
  - » Measure & test for timeliness & completeness
  - » Use a unified, permissive CIM schema
- » Own your data with a self-managed data pipeline
  - » Focus on availability and consistency of data

# APPENDIX

# RESOURCES FOR DATA PIPELINES

» What Is a Data Pipeline?

» <https://hazelcast.com/glossary/data-pipeline/>

» Data Engineering and Its Main Concepts

» <https://www.altexsoft.com/blog/datascience/what-is-data-engineering-explaining-data-pipeline-data-warehouse-and-data-engineer-role/>