The Game of Big Data

Analytics Infrastructure at KIXEYE

Randy Shoup

@randyshoup

linkedin.com/in/randyshoup

QCon New York, June 13 2014

Free-to-Play Real-time Strategy Games

BATTLE PIRATES

WAR COMMANDER

VEGA CONFLICT

BACKYARD MONSTERS UNLEASHED

TOME IMMORTAL ARENA

- Web and mobile
- Strategy and tactics
- Really real-time ☺
- Deep relationships with players
- Constantly evolving gameplay, feature set, economy, balance

- >500 employees worldwide

Intro: Analytics at KIXEYE

User Acquisition

Game Analytics

Retention and Monetization

Analytic Requirements

User Acquisition

Goal: ELTV > acquisition cost

- *User's estimated lifetime value is more than it costs to acquire that user*

Mechanisms

- *Publisher Campaigns*
- *On-Platform Recommendations*

Game Analytics

Goal: Measure and Optimize “Fun”

- *Difficult to define*
- *Includes gameplay, feature set, performance, bugs*
- *All metrics are just proxies for fun (!)*

Mechanisms

- *Game balance*
- *Match balance*
- *Economy management*
- *Player typology*

Retention and Monetization

Goal: Sustainable Business

- *Monetization drivers*
- *Revenue recognition*

Mechanisms

- *Pricing and Bundling*
- *Tournament ("Event") Design*
- *Recommendations*

Analytic Requirements

- Data Integrity and Availability

- Cohorting
- Controlled experiments
- Deep ad-hoc analysis

“Deep Thought”

V1 Analytic System

Goals

Core Capabilities

Implementation

“Deep Thought”

V1 Analytic System

Goals

Core Capabilities

Implementation

V1 Analytic System

Grew Organically

- *Built originally for user acquisition*
- *Progressively grown to much more*

Idiosyncratic mix of languages, systems, tools

- *Log files -> Chukwa -> Hadoop -> Hive -> MySQL*
- *PHP for reports and ETL*
- *Single massive table with everything*

V1 Analytic System

Many Issues

- *Very slow to query*
- *No data standardization or validation*
- *Very difficult to add a new game, report, ETL*
- *Extremely difficult to backfill on error or outage*
- *Difficult for analysts to use; impossible for PMs, designers, etc.*

… but we survived (!)

“Deep Thought”

V1 Analytic System

Goals

Core Capabilities

Implementation

Goals of Deep Thought

Independent Scalability

- *Logically separate, independently scalable tiers*

Stability and Outage Recovery

- *Tiers can completely fail with no data loss*
- *Every step idempotent and replayable*

Standardization

- *Standardized event types, fields, queries, reports*

Goals of Deep Thought

In-Stream Event Processing

- *Sessionalization, Dimensionalization, Cohorting*

Queryability

- *Structures are simple to reason about*
- *Simple things are simple*
- *Analysts, Data Scientists, PMs, Game Designers, etc.*

Extensibility

- *Easy to add new games, events, fields, reports*

“Deep Thought”

V1 Analytic System

Goals

Core Capabilities

Implementation

Core Capabilities

- Sessionalization
- Dimensionalization
- Cohorting

Sessionalization

All events are part of a “session”

- *Explicit start event, optional stop event*
- *Game-defined semantics*

Event Batching

- *Events arrive in batch, associated with session*
- *Pipeline computes batch-level metrics, disaggregates events*
- *Can optionally attach batch-level metrics to each event*

Sessionalization

Time-Series Aggregations

- *Configurable metrics*
  - *1-day X, 7-day X, lifetime X*
  - *Total attacks, total time played*
- *Accumulated in-stream*
  - *V1 aggregate + batch delta*
- *Faster to calculate in-stream vs. Map-Reduce*

Dimensionalization

Pipeline assigns unique numeric id to string enums

- *E.g., “twigs” resource → id 1234*

Automatic mapping and assignment

- *Games log strings*
- *Pipeline generates and maps ids*
- *No configuration necessary*

Fast dimensional queries

- *Join on integers, not strings*

Dimensionalization

Metadata enumeration and manipulation

- *Easily enumerate all values for a field*
- *Merge multiple values*
  - *“TWIGS” == “Twigs” == “twigs”*

Metadata tagging

- *Can assign arbitrary tags to metadata*
  - *E.g., “Panzer 05” is {tank, mechanized infantry, event prize}*
- *Enables custom views*

Cohorting

Group players along any dimension / metric

- *Well beyond classic age-based cohorts*

Core analytical building block

- *Experiment groups*
- *User acquisition campaign tracking*
- *Prospective modeling*
- *Retrospective analysis*

Cohorting

Set-based

- *Overlapping groups: >100, >200, etc.*
- *Exclusive groups: (100-200), (200-500), etc.*

Time-based

- *E.g., people who played in last 3 days*
- *E.g., “whale” == ($$ > X) in last N days*
- *Autoexpire from a group without explicit intervention*

“Deep Thought”

V1 Analytic System

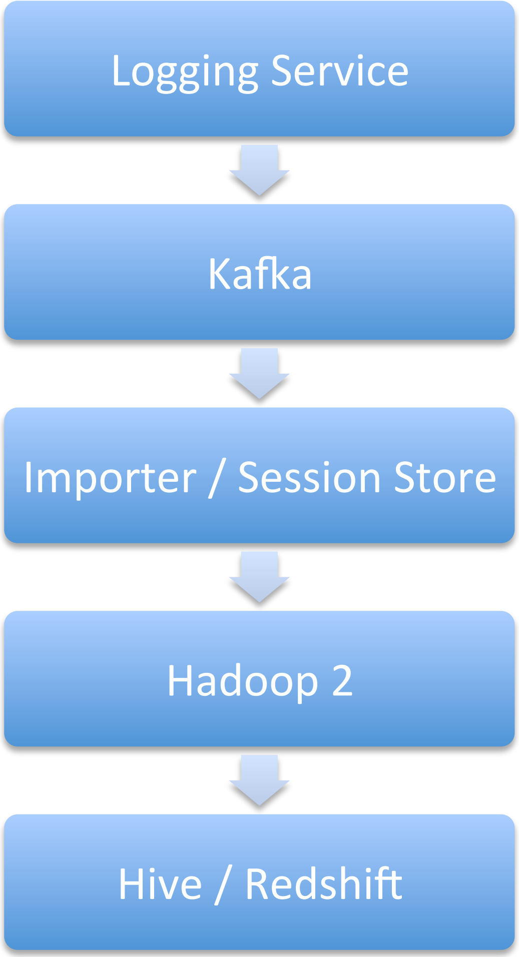
Goals

Core Capabilities

Implementation

Implementation of Pipeline

- Ingestion
- Event Log
- Transformation
- Data Storage
- Analysis and Visualization



Logging Service

Kafka

Importer / Session Store

Hadoop 2

Hive / Redshift

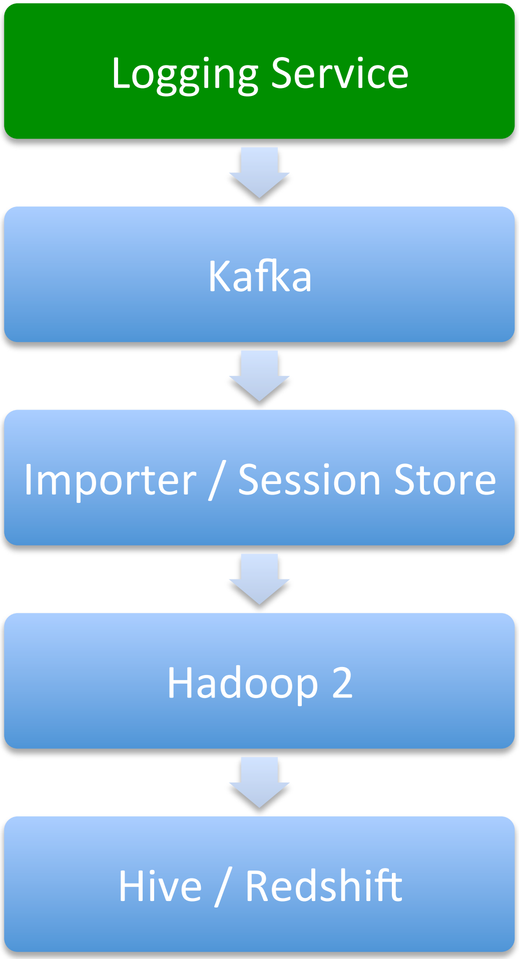
Ingestion: Logging Service

HTTP / JSON Endpoint

- *Play framework*
- *Non-blocking, event-driven*

Responsibilities

- *Message integrity via checksums*
- *Durability via local disk persistence*
- *Async batch writes to Kafka topics*
  - *{valid, invalid, unauth}*



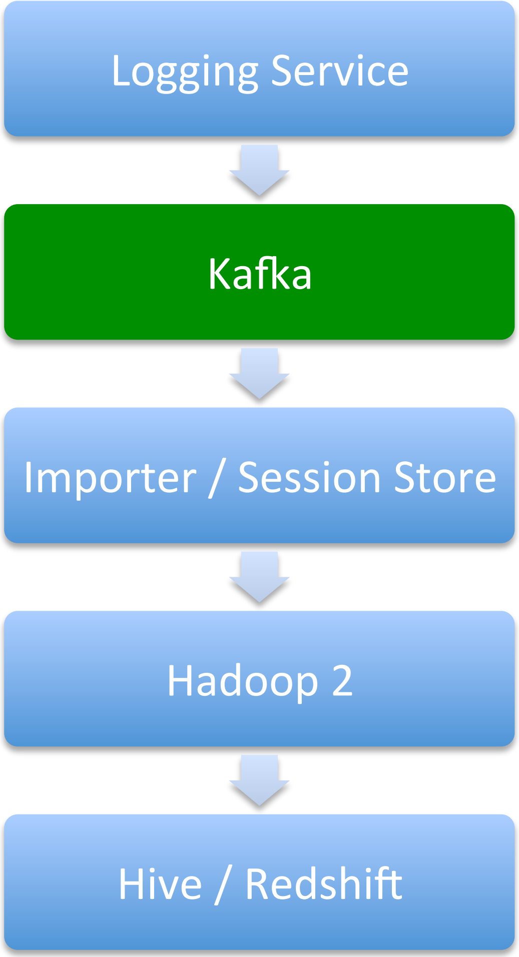
Event Log: Kafka

Persistent, replayable pipe of events

- *Events stored for 7 days*

Responsibilities

- *Durability via replication and local disk streaming*
- *Replayability via commit log*
- *Scalability via partitioned brokers*
- *Segment data for different types of processing*



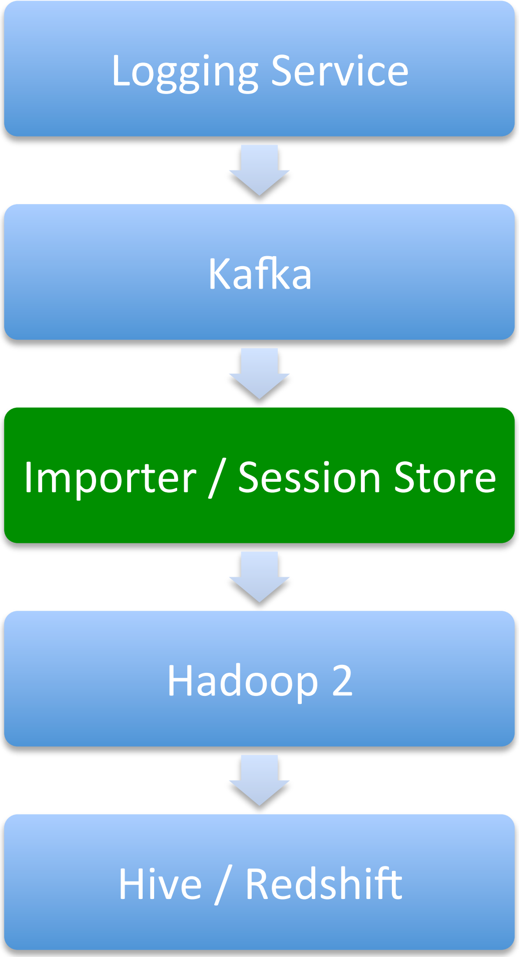
Transformation: Importer

Consume Kafka topics, rebroadcast

- *E.g., consume batches, rebroadcast events*

Responsibilities

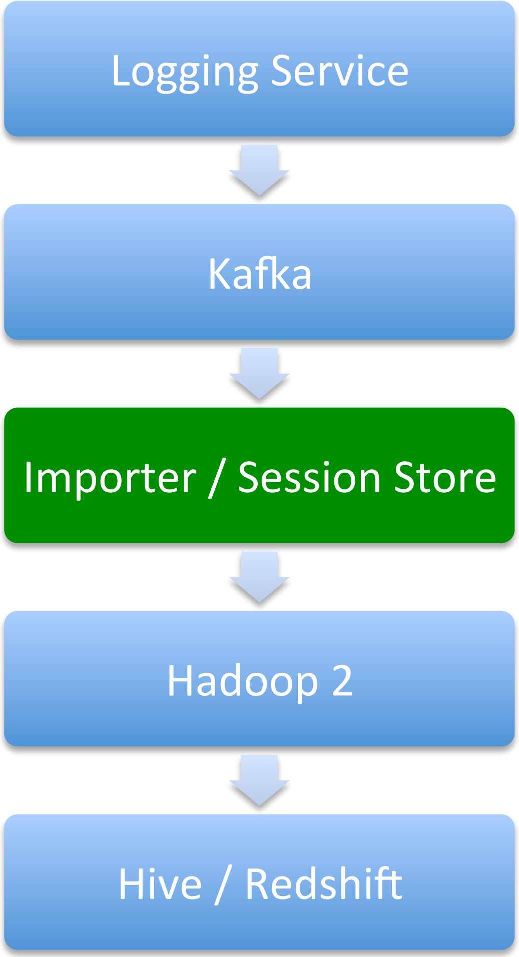
- *Batch validation against JSON schema*
  - *Syntactic validation*
  - *Semantic validation (is this event possible?)*
- *Batches -> events*



Transformation: Importer

Responsibilities (cont.)

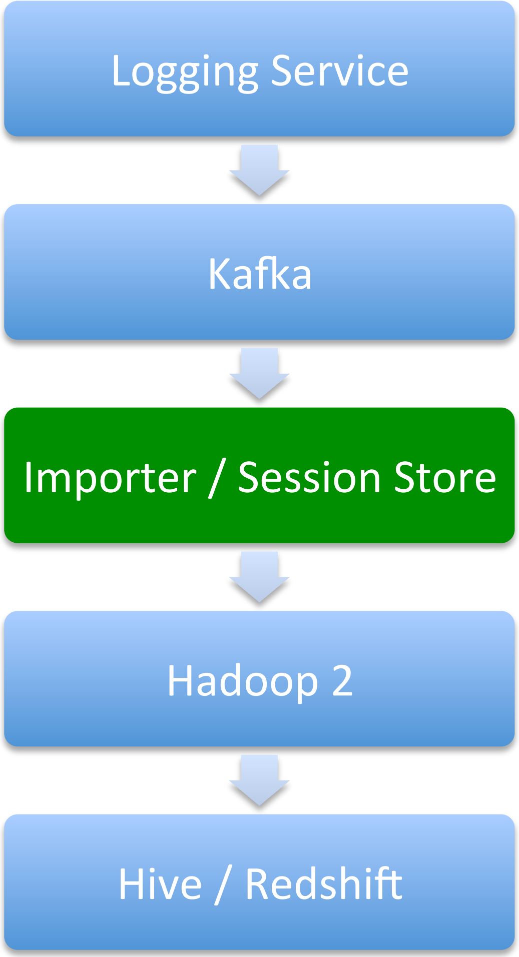
- *Sessionalization*
    - *Assign event to session*
    - *Calculate time-series aggregates*
- *Dimensionalization*
    - *String enum -> numeric id*
    - *Merge / coalesce different string representations into single id*
- *Player metadata*
    - *Join player metadata from session store*



Transformation: Importer

Responsibilities (cont.)

- *Cohorting*
  - *Process enter-cohort, exit-cohort events*
  - *Process A / B testing events*
  - *Evaluate cohort rules (e.g., spend thresholds)*
  - *Decorate events with cohort tags*



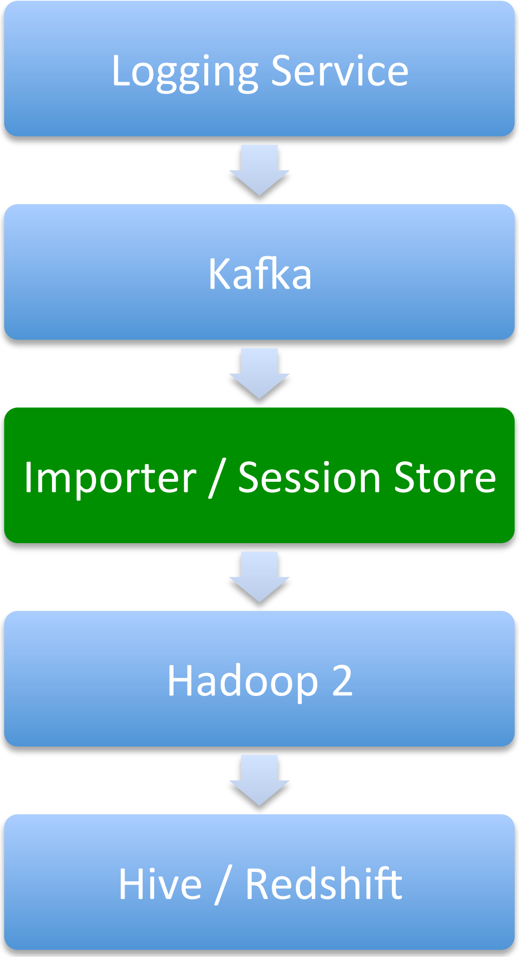
Transformation: Session Store

Key-value store (Couchbase)

- *Fast, constant-time access to sessions, players*

Responsibilities

- *Store Sessions, Players, Dimensions, Config*
  - *Lookup*
  - *Idempotent update*
- *Store accumulated session-level metrics*
- *Store player history*



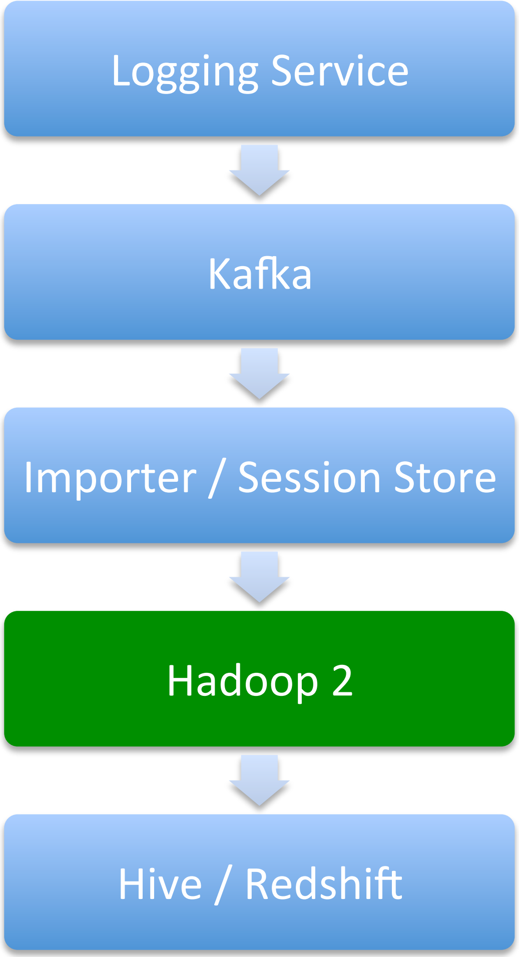
Storage: Hadoop 2

Camus MR

- *Kafka -> HDFS every 3 minutes*

append_events table

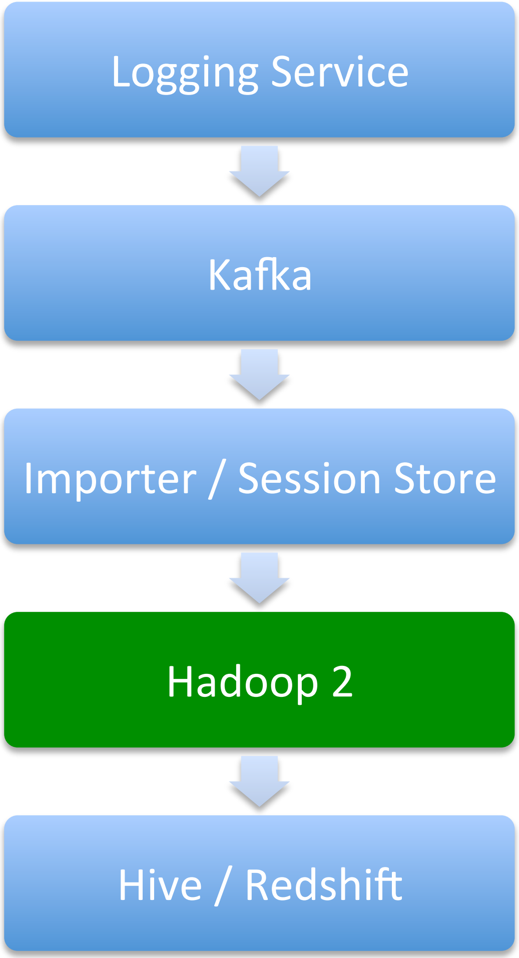
- *Append-only log of events*
- *Each event has session-version for deduplication*



Storage: Hadoop 2

append_events -> base_events MR

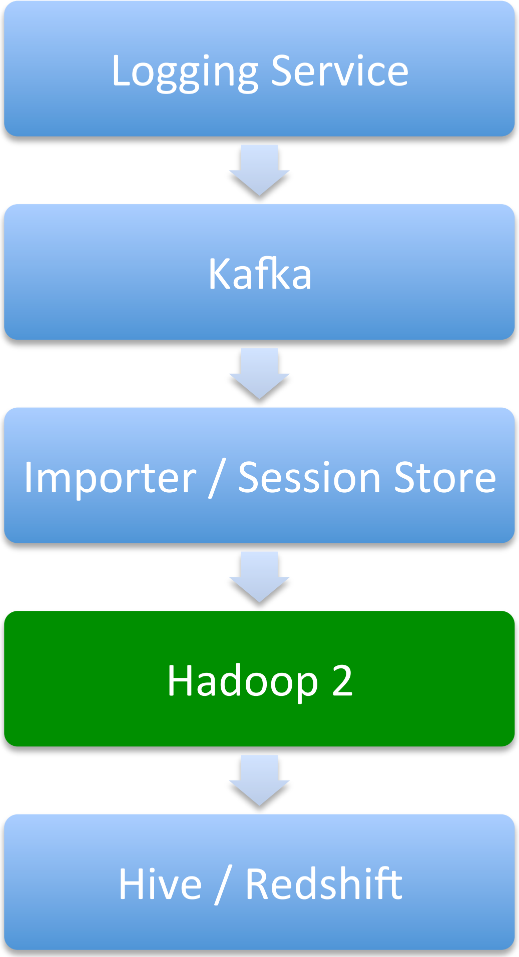
- *Logical update of base_events*
  - *Update events with new metadata*
  - *Swap old partition for new partition*
- *Replayable from beginning without duplication*



Storage: Hadoop 2

base_events table

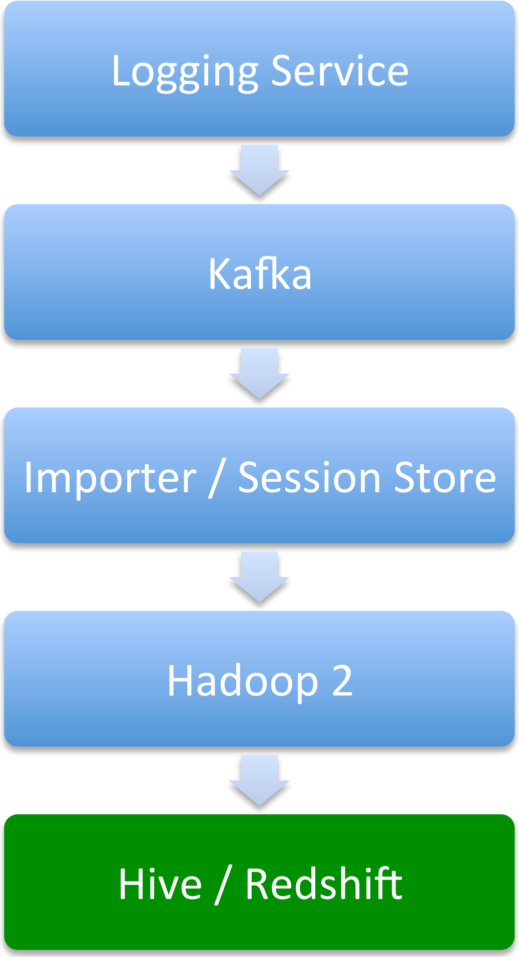
- *Denormalized table of all events*
- *Stores original JSON + decoration*
- *Custom Serdes to query / extract JSON fields without materializing entire rows*
- *Standardized event types → lots of functionality for free*



Analysis and Visualization

Hive Warehouse

- *Normalized event-specific, game-specific stores*
- *Aggregate metric data for reporting, analysis*
- *Maintained through custom ETL*
  - *MR*
  - *Hive queries*



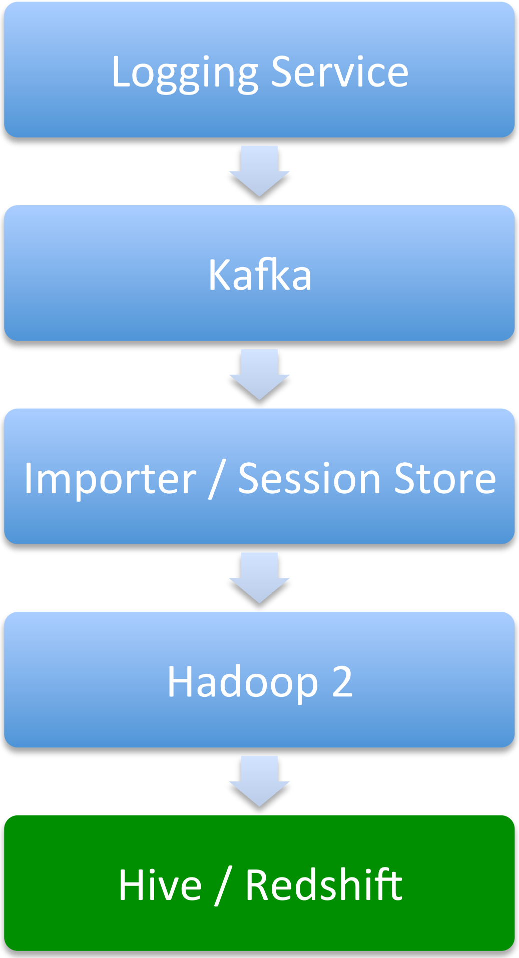
Analysis and Visualization

Amazon Redshift

- *Fast ad-hoc querying*

Tableau

- *Simple, powerful reporting*



Come Join Us!

KIXEYE is hiring in

SF, Seattle, Victoria,
Brisbane, Amsterdam

rshoup@kixeye.com

@randyshoup

Deep Thought Team:

- Mark Weaver
- Josh McDonald
- Ben Speakmon
- Snehal Nagmote
- Mark Roberts
- Kevin Lee
- Woo Chan Kim
- Tay Carpenter
- Tim Ellis
- Kazue Watanabe
- Erica Chan
- Jessica Cox
- Casey DeWitt
- Steve Morin
- Lih Chen
- Neha Kumari