Streaming & Kafka:
From Background to Use Case

Ben Polk
Logan Vining
Dalton Gomez

MONTANA STATE UNIVERSITY

CSCI-540: Advanced Databases
Fall 2021

Starting Discussion...

After going through NoSQL DB models all semester...

What makes a database a database?

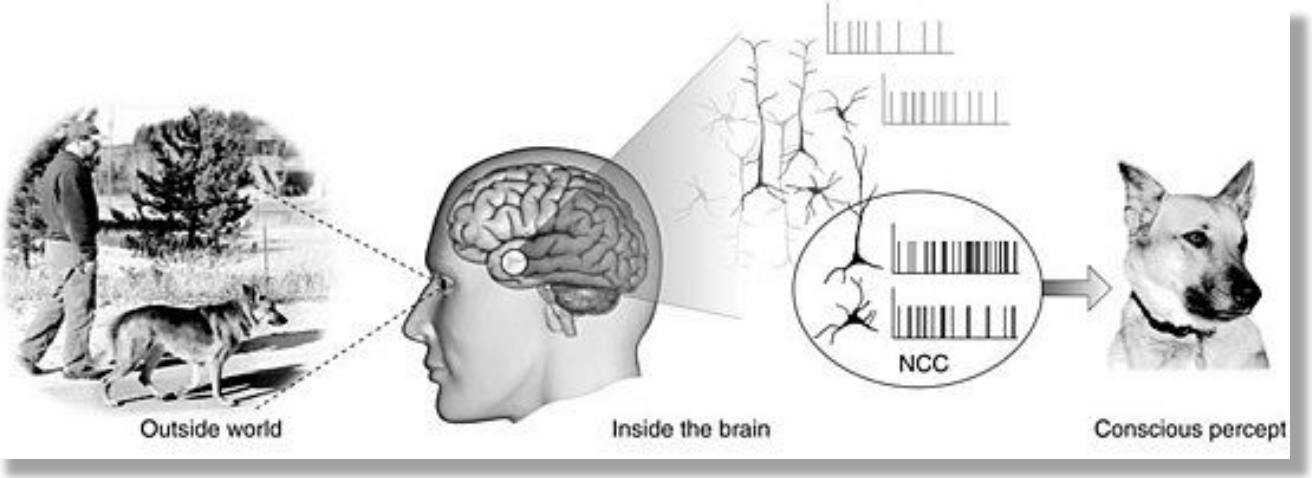
What are the defining properties or characteristics that it must have?

Learning Objectives

- Know what event streaming is and some common use cases of it in software systems
- Understand the architecture of Kafka and how it can be used as an event-streaming database
- Interact with a system that uses Kafka to visualize geospatial data in real-time

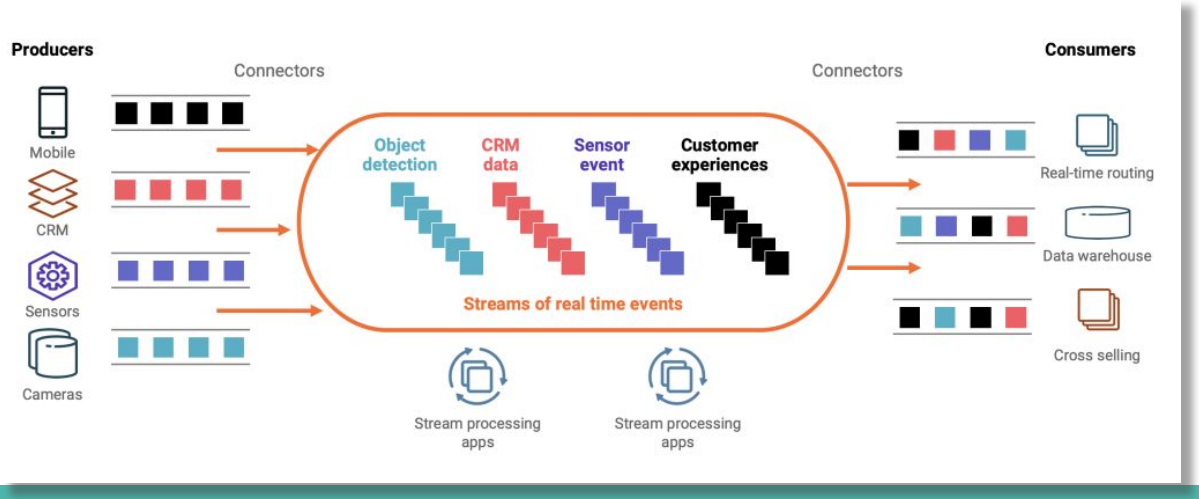
Building Intuition...

- **Consciousness ⊆ Event Streaming**
  - Our senses (vision, hearing, smell, etc.) ***produce*** a continuous ***stream*** of information into our brain.
  - Our brain organizes that information into ***events***, associates events to ***topics***, and ***logs*** it in chronological order.
  - We then ***consume*** that information ***in real-time*** to make inferences about what is happening around us.



What is event streaming?

- **A paradigm shift from “normal” databases:**
  - Time-oriented
  - Emphasis on events, not things
- **Event streaming is “data in motion”:**
  - Events are captured in real-time from data sources
  - Stored in chronological logs
  - Processed and routed to destinations as needed



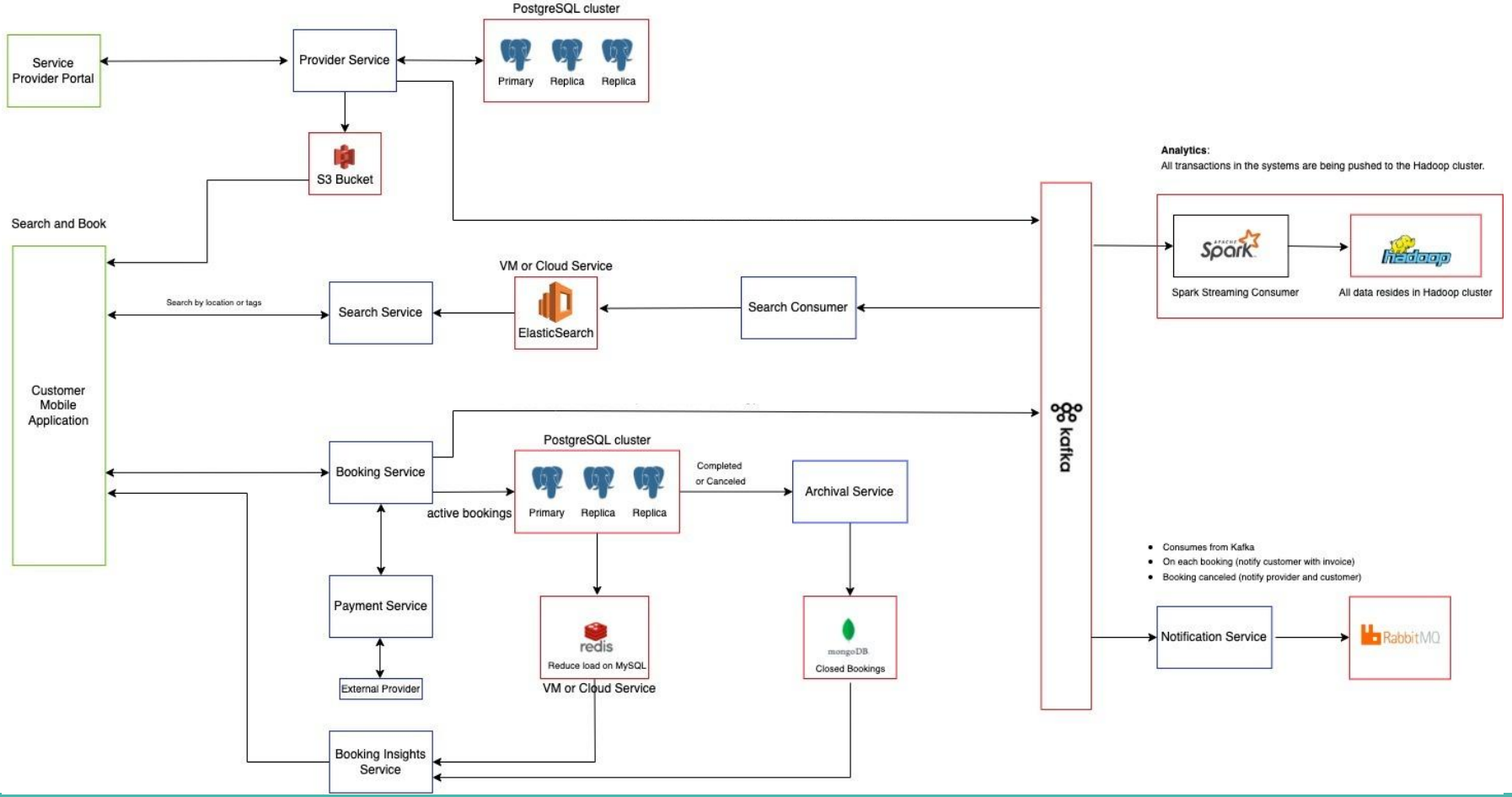
What is Kafka?

- Open-Source Software (Apache Project)
  - Originated at LinkedIn in 2011
- Built for Event Streaming
  - A pub/sub system that grew up to be a database
- Distributed System
- Durable, Fault-Tolerant Data Persistence
- Easily Scalable
- High Throughput and Low Latency
- ACID Properties (Sort Of...)
- 5 APIs:
  - Producer API
  - Consumer API
  - Connect API
  - Streams API
  - Admin API

Weaknesses???

Discussion Prompt

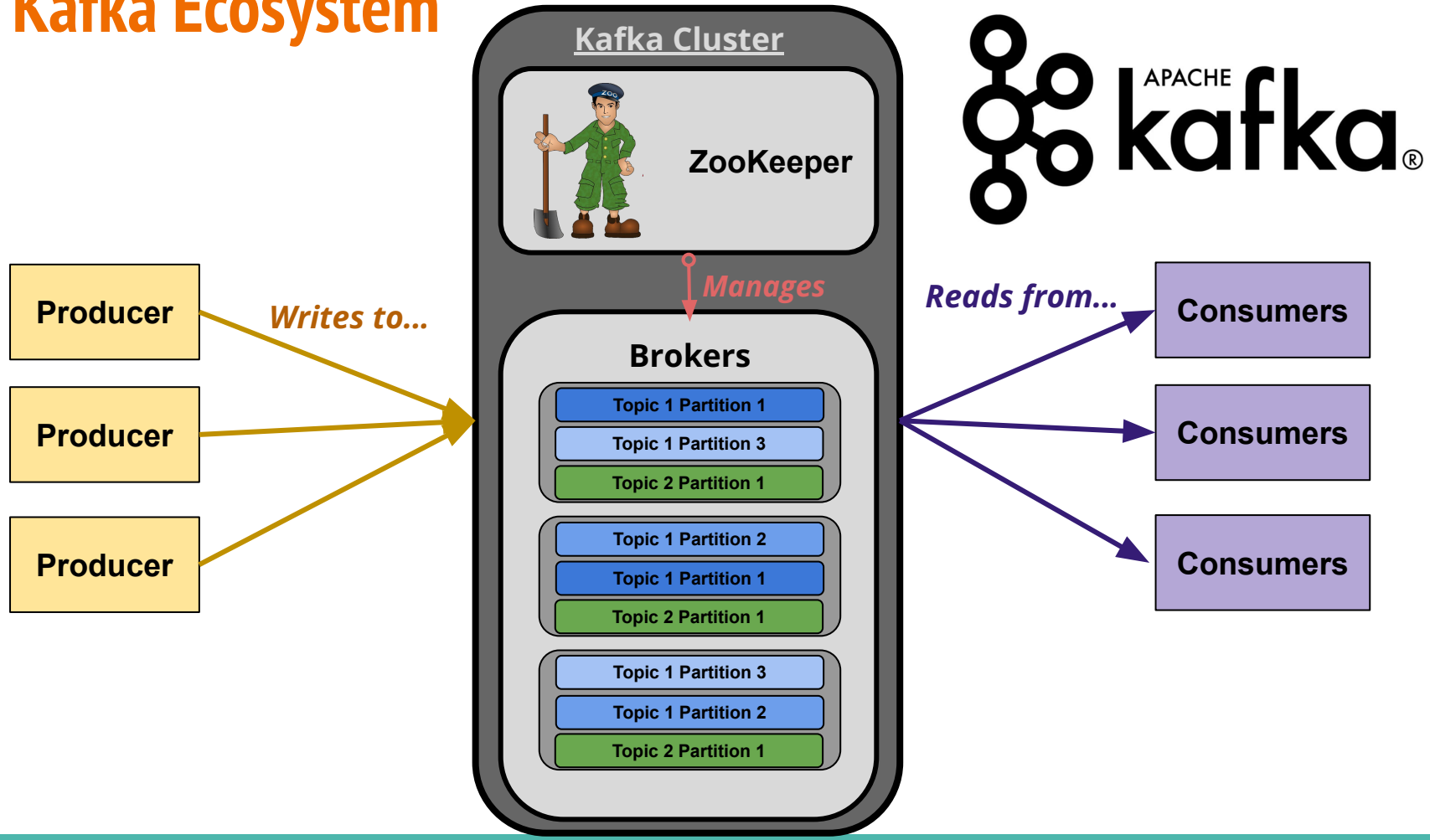
In Ibrahim's Booking System Architecture, he used Kafka and classified it as a database. What role did Kafka serve within that system?



Use Cases of Kafka

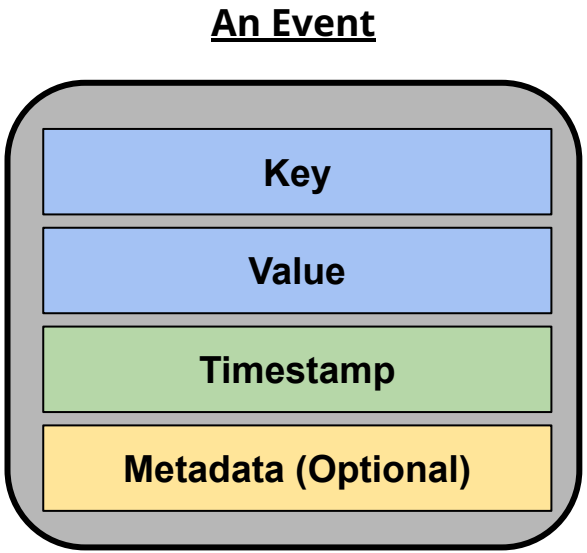
- Used in thousands of major companies
- Facilitates the user of software being more software
- Allows for the “always-on” nature of apps
- Goes beyond just a database or pub/sub system
- ***Often Kafka is a gatekeeper to the flow of information within a big system***

The Kafka Ecosystem



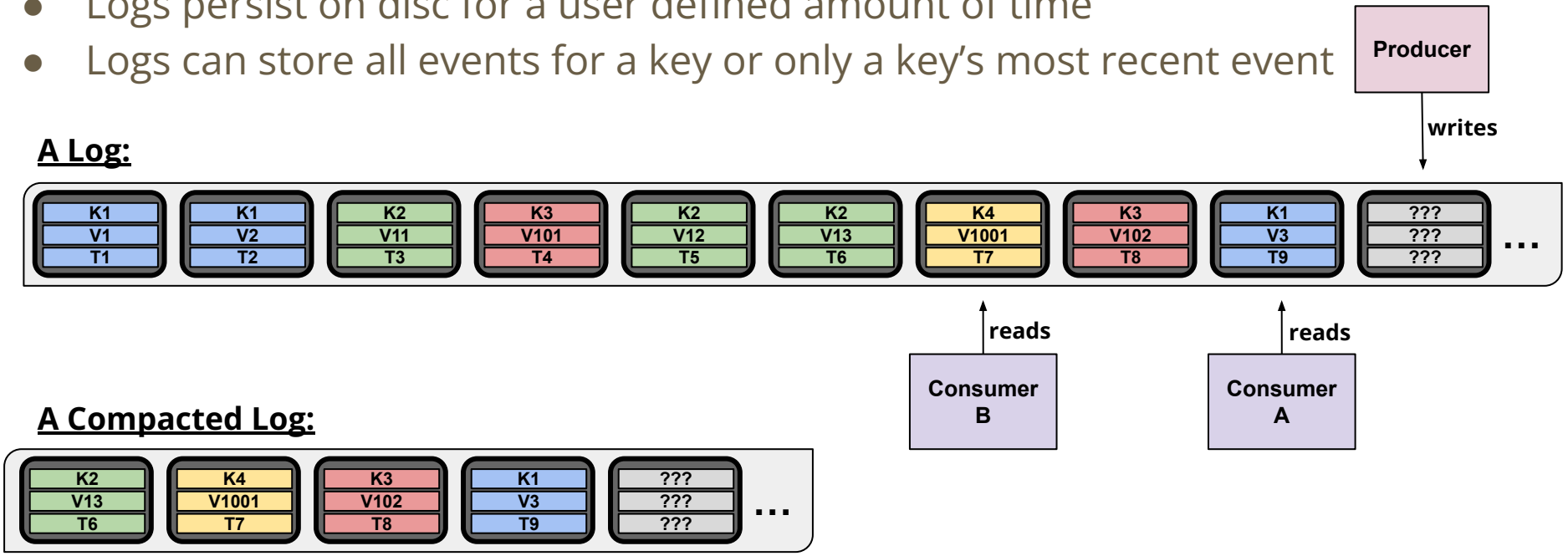
Events

- A “record” in a Kafka Database
- Things that happened, a description of them, and when they happened
- An event is comprised of:
  - Key
    - Not necessarily unique
    - Hashable
  - Value
    - Like JSON, XML, etc.
  - Timestamp
  - Optional Metadata
- Examples:
  - IoT Sensor Data
  - Business Change
  - User Interaction
  - Microservice Output



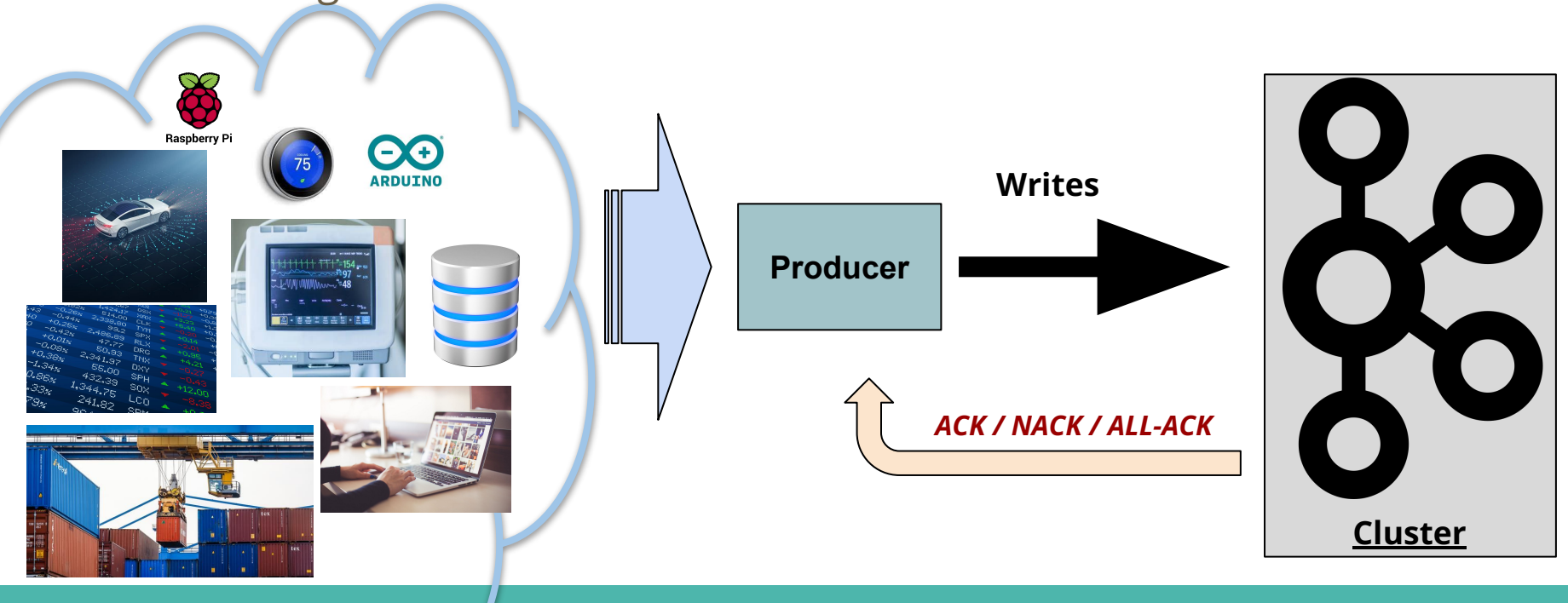
Topics and Logs

- Topics are relevant grouping of events
- Logs are the chronological stream of events
- Logs are immutable
- Logs persist on disc for a user defined amount of time
- Logs can store all events for a key or only a key's most recent event



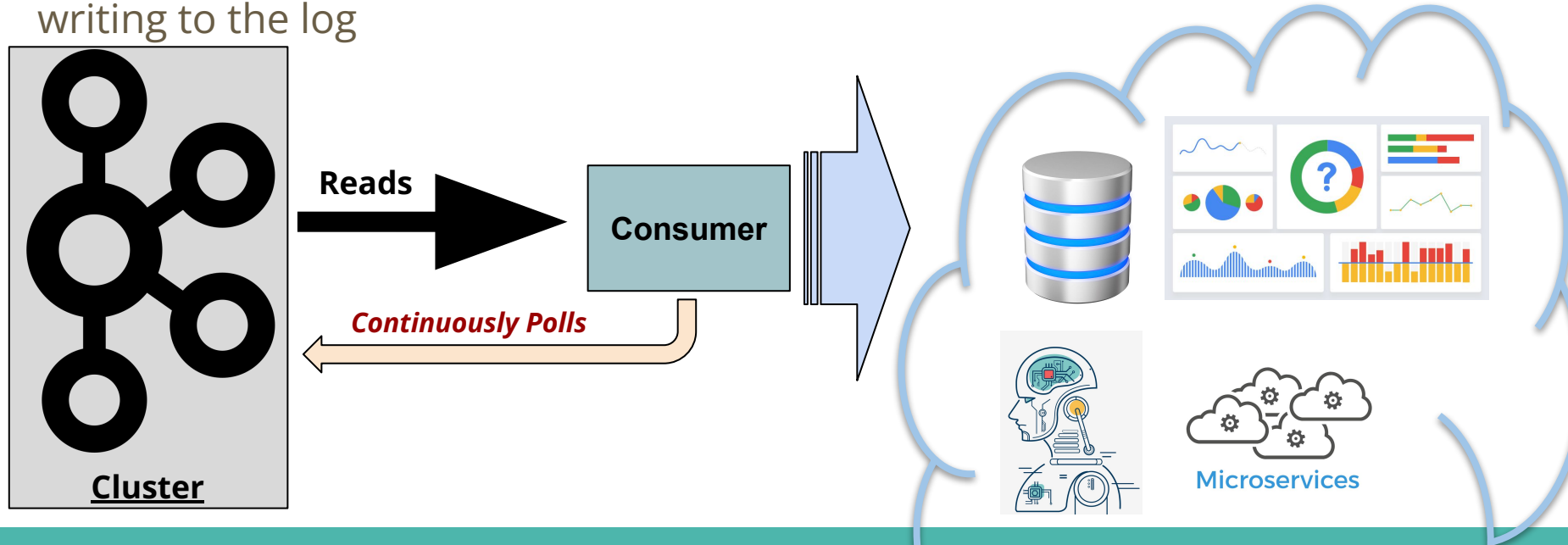
Producers

- A user-created program
- Publishes events to a Kafka topic or topics
- Can only publish at the end of a topic's log
- Can be designed to wait for an ACK or not



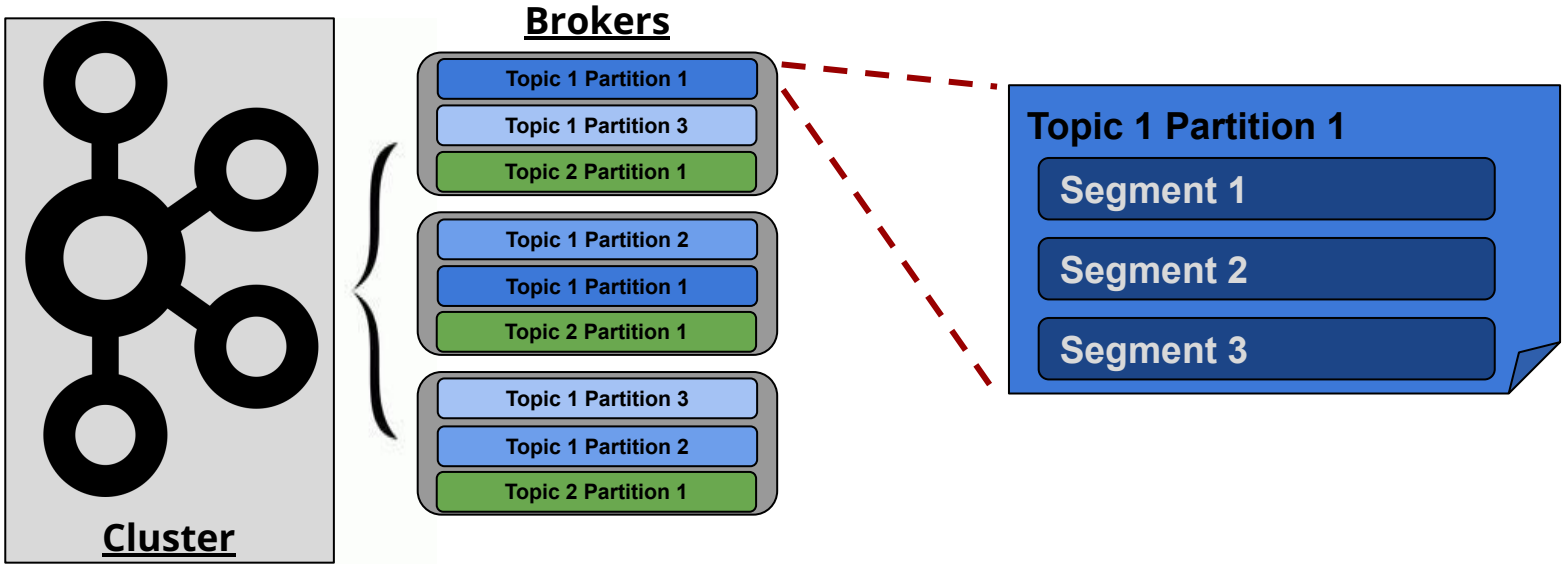
Consumers

- Also user-created
- Continuously poll a topic or topics for new events
- Stores an offset (in the Kafka cluster) that remembers the last read event in each topic's log
- Ideally a consumer is "caught-up" but can be reading behind the producer(s) writing to the log



Brokers and Partitions

- A broker is physical memory (e.g. a server) that is running in the Kafka cluster
- Each topic can be partitioned across many brokers
  - And then segmented into specific files, called segments
- Each topic is also replicated across many brokers
- Provides scalability



Apache ZooKeeper

- Another Apache Project
- A centralized service for synchronizing distributed systems
- In Kafka, ZooKeeper manages all the brokers
- Plan to phase out ZooKeeper and have Kafka natively coordinate brokers

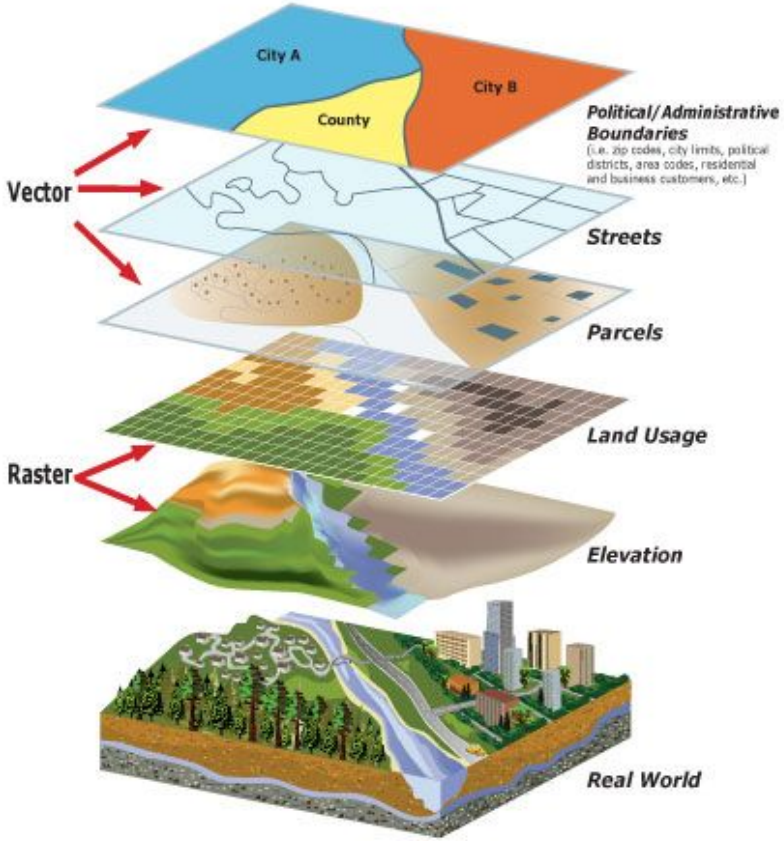
APACHE ZooKeeper™

What is Geospatial Data?

- Geospatial Data is Georeferenced
  - Any data has a reference to a place on planet Earth
- Vector-Based
  - Georeferenced points, lines, and polygons
- Raster-Based
  - Georeferenced bitmap where each cell contains a value

What apps can you think of that use geospatial data in realtime?

How might Kafka be used within those applications?



What is GeoMesa?

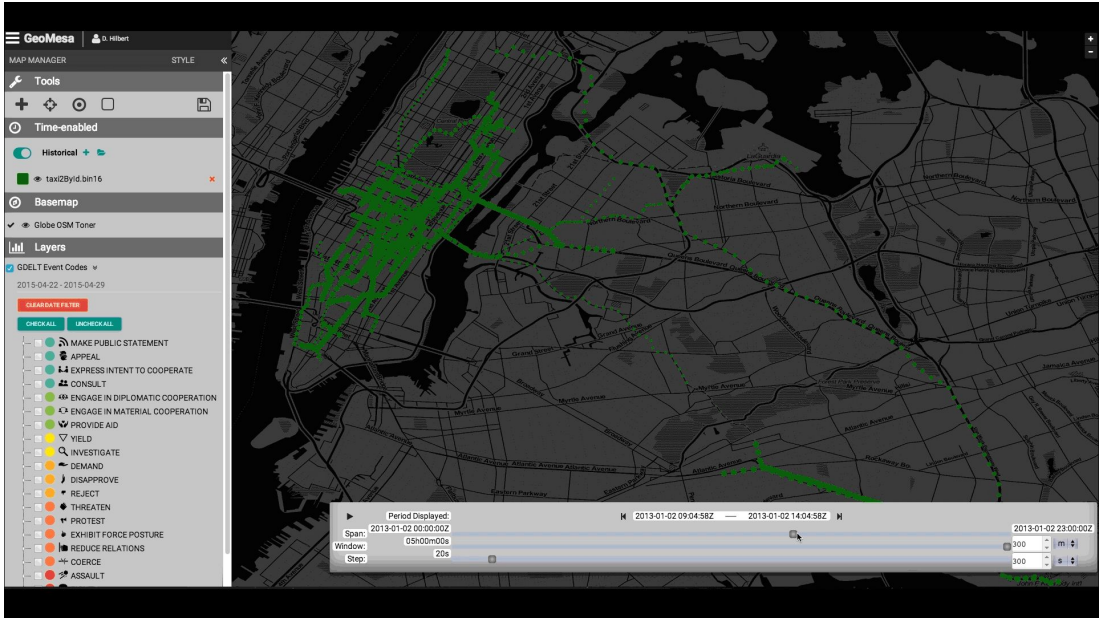
- Open-source software
- Provides storage and analysis of geospatial data
- Can be run with several underlying databases
  - HBase
  - Redis
  - Kafka!
- GeoMesa is a “suite of tools that enables large-scale geospatial analytics”

How does GeoMesa work with Kafka?

- GeoMesa's "suite of tools"...
  - They offer Scala and Java libraries as well as CLI's
  - These clients interact with data stores to facilitate analytics

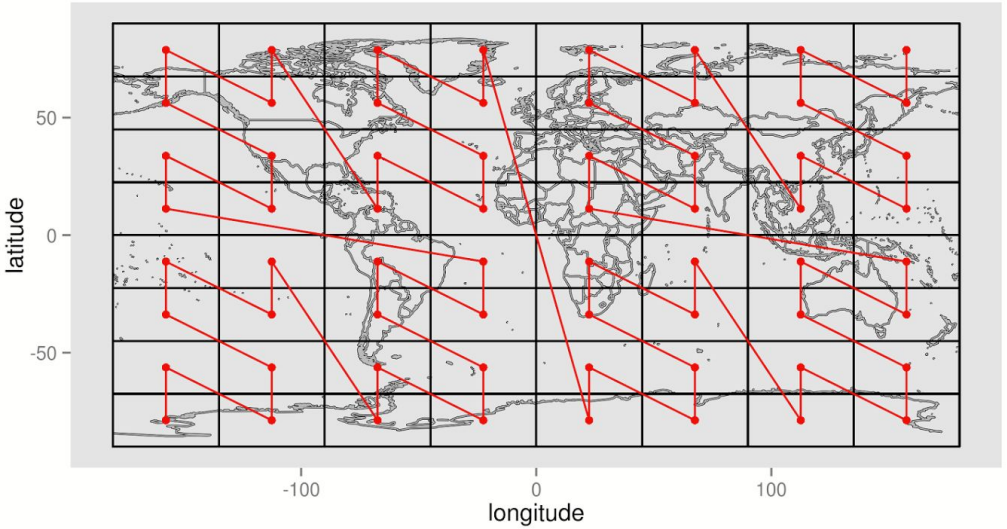
- What GeoMesa does:
  - Generate keys for key-value data stores
  - Keys allow *fast* analysis of geospatial data
  - Provide geospatial querying tools

- GeoMesa "sits on top of" the underlying datastore

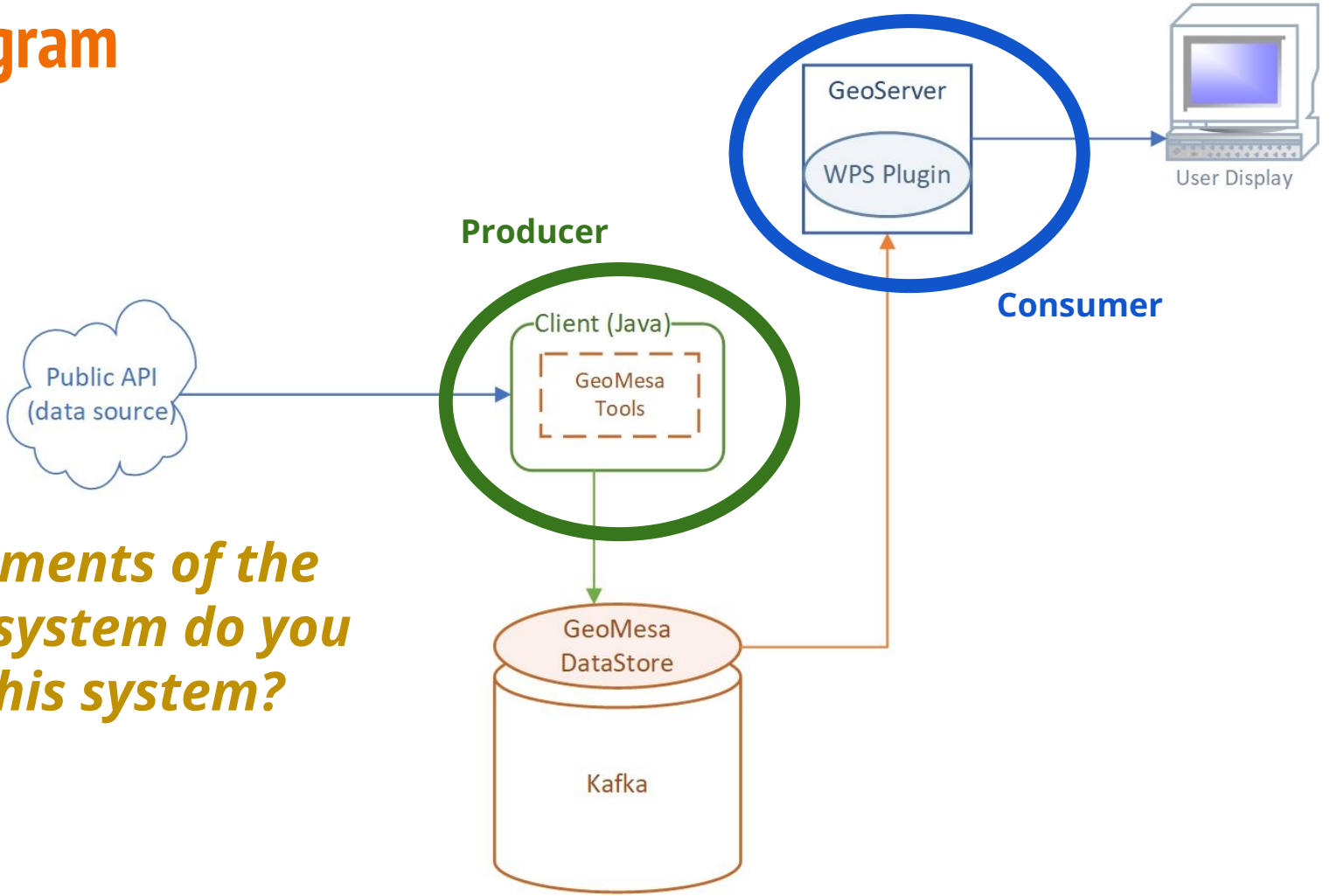


Indexing in GeoMesa

- Using smart indexes in a key-value store greatly improves database performance
- GeoMesa uses a Z-Curve to represent latitude, longitude, and time as a single key of the most commonly queried dimensions
- The curve visits each point only once to establish a unique order of points



System Diagram



What elements of the Kafka ecosystem do you see in this system?

References

- **Apache Kafka Documentation**
  - https://kafka.apache.org/
- **Confluent Documentation**
  - https://www.confluent.io/
- **GeoMesa Documentation**
  - https://www.geomesa.org/
- **GeoServer Documentation**
  - http://geoserver.org/
- **Our Github Repository**
  - https://github.com/Ncf4n1/GeoMesa-Streaming-Presentation
- **These Slides in Google Drive**
  - https://docs.google.com/presentation/d/1bSl7RDiruSq62-BCkrRX3V74uZhGicH_1MeVfAuC2L8/edit?usp=sharing