Hadoop Application Architectures: Architecting a Next Generation Data Platform

Strata Data Conference, New York 2018
Ted Malaska | @tedmalaska
Jonathan Seidman | @jseidman
Logistics

- Break at 3:00 – 3:30 PM
- Questions at the end of each section
- Slides at tiny.cloudera.com/app-arch-ny2018
- Code at https://github.com/hadooparchitecturebook/Taxi360
About the book

- @hadooparchbook
- hadooparchitecturebook.com
- github.com/hadooparchitecturebook
- slideshare.com/hadooparchbook
About the presenters

Ted Malaska

- **Capital One**: Director of Enterprise Architecture (Starting Next Week)
- **Blizzard Ent**: Director of Engineering of Global Insights
- **Cloudera**: Principal Solution Architect
- **FINRA**: Lead Architect
- **Contributor**: Apache Spark, Hadoop, Hive, Sqoop, Yarn, Flume, others
About the presenters

Jonathan Seidman

▪ Software Engineer at Cloudera
▪ Previously Technical Lead on the big data team at Orbitz
▪ Co-founder of the Chicago Hadoop User Group and Chicago Big Data
Case Study Overview
Internet of Things and Entity 360
Customer 360

Questions?
tiny.cloudera.com/ny2018questions
Connected Cars
Entity (Taxi) 360 View

- Streaming Vehicle Data
- Geo-location/Traffic Data
- Customer Data
- Maintenance Data
- Other Data Sources

Questions?
tiny.cloudera.com/ny2018questions
In the old days

Data Warehouse → Extract → Transform (SQL) → Load → Data Mart
At the time of our Book

Questions?
tiny.cloudera.com/ny2018questions
Enabling a Range of New Use Cases…

- Fraud Detection
- Market Transactions
- Internet of Things
- Network Security
Challenges

And so many more

Questions?
tiny.cloudera.com/ny2018questions
Challenges – Architectural Considerations

- Reliable and scalable ingress of multiple data types and sources:
  - High volume event data? Batch data?

- Reliable and scalable storage to support multiple workloads and access patterns
  - Historical data? Real-time search? Analytics?

- Processing engines (for background processing):
  - Stream processing? Batch processing?

- Data Modeling
  - Modeling data for real-time random access? Analytic access? Batch access?

- Deployment Considerations
  - On-premise, cloud?
Case Study Requirements

Overview
Requirements

- Allow users (technical and non-technical) to analyze and visualize data…
Requirements

- Provide analysts with query capabilities via a standard interface…
Requirements

- Provide developers the ability to perform batch processing on historical data…
Requirements

- To support all this, we need:
  - Reliable ingestion of streaming and batch data.
  - Ability to perform transformations on streaming data in flight.
  - Ability to perform sophisticated processing of historical data.
  - Reliable and scalable storage to support modeling and processing of multiple data formats.
High level architecture
Walkthrough
High level architecture

- Data Sources
  - Code
  - Agents
  - Log Aggregators

- Producers
  - Data Governance/Metadata
  - Enrichment
  - Stream Processing
  - Routing
  - Transport
  - Replication

- Streaming Pipes
  - Data Governance/Metadata
  - Enrichment
  - Stream Processing
  - Routing
  - Transport
  - Replication

- Storage
  - File/Object
  - RDBMS/MPP
  - Time Series
  - Reverse Indexed
  - Memory
  - Stream

- Access
  - SQL
  - Machine Learning
  - Request Response
  - Batch Processing

Questions?
tiny.cloudera.com/ny2018questions
Data Sources
And Data Acquisition
Considerations
High level architecture

Questions?
tiny.cloudera.com/ny2018questions
Key to Customer 360 Success

Your project is only as good as the quality and variety of data sources

Questions?
tiny.cloudera.com/ny2018questions
High Level Architecture

Data Sources
- Code
- Agents
- Log Aggregators

Producers
- Data Governance/Metadata
- Enrichment
- Stream Processing
- Routing
- Transport
- Replication

Streaming Pipes

Storage
- File/Object
- RDBMS/MPP
- Time Series
- Reverse Indexed
- Memory
- Stream

Access
- SQL
- Machine Learning
- Request Response
- Batch Processing

Questions?
tiny.cloudera.com/ny2018questions
Producers

- Embedded code
- Agents
- Flume
- Kafka
Producers

- JDBC
- DI Tools
Producers

- REST
- MQ
- Etc.

Questions?
tiny.cloudera.com/ny2018questions
Producers

- StreamSets
- CDAP
- streamlio
- Informatica

Questions?
tiny.cloudera.com/ny2018questions
Producers

- Set SLAs and make sure you can meet them.
- Set delivery guarantees.
- Be defensive.
- Operationalize – repeatable deployments, monitoring, etc.
Kafka Clients

Apache Kafka Clients

Ecosystem Clients
REST Proxy
Talking to Non-native Kafka Apps and Outside the Firewall

- Provides a RESTful interface to a Kafka cluster
- Simplifies message creation and consumption
- Simplifies administrative actions

Non-Java Applications

REST Proxy

Native Kafka Java Applications

REST / HTTP
Kafka Connect
Streaming Data Capture

- Fault tolerant
- Manage hundreds of data sources and sinks
- Preserves data schema
- Part of Apache Kafka project
- Includes simple transformations

Kafka Connect API

- Connector
- Connector
- Connector
- Connector

Sources

- JDBC
- Logs
- MQTT

Kafka

Sinks

- RDBMS
- Key/Value
- HDFS

Questions?
tiny.cloudera.com/ny2018questions
# Ecosystem of Connectors

## Databases
- DB2
- MongoDB
- Cassandra
- Giraph
- Impala
- Couchbase
- DynamoDB
- InfluxDB
- HBase
- RethinkDB
- Kudu

## Datastore/File Store
- Hadoop
- Ignite
- FTP
- syslog-ng
- Hazelcast

## Analytics
- Vertica
- Elastic
- Mixpanel
- Solr
- Kudu

## Applications / Other
- Amazon S3
- Syncsort
- Attunity
- Bloomberg
- Twitter
- MQTT

[Questions?](tiny.cloudera.com/ny2018questions)
High level architecture

- Data Sources
  - Code
  - Agents
  - Log Aggregators

- Producers
  - Data Governance/Metadata
  - Enrichment
  - Stream Processing
  - Routing
  - Transport
  - Replication

- Streaming Pipes
  - Data Governance/Metadata
  - Enrichment
  - Stream Processing
  - Routing
  - Transport
  - Replication

- Storage
  - File/Object
  - RDBMS/MPP
  - Time Series
  - Reverse Indexed
  - Memory
  - Stream

- Access
  - SQL
  - Machine Learning
  - Request Response
  - Batch Processing

Questions?
tiny.cloudera.com/ny2018questions
Why Data Governance?

- Regulation (for example GDPR).
- Data management.
- Data quality.
- Data models/schemas will change.
Implementing Data Governance

- Vendor tools
- Custom implementation
High level architecture

Data Sources
- Code
- Agents
- Log Aggregators

Producers
- Data Governance/Metadata
- Enrichment
- Stream Processing
- Routing
- Transport
- Replication

Streaming Pipes
- Long Term SQL
- Speed Layer SQL
- Time Series
- Reverse Indexed
- Stream

Storage
- SQL
- Machine Learning
- Request Response
- Batch Processing

Access

Questions?
tiny.cloudera.com/ny2018questions
Goals for our Transport Layer

- Buffering
- Partitioning
- Speed
- Replayability
- High throughput
- Reliable
Goals for our Transport Layer

- To meet these goals we want some kind of publish-subscribe queue:
  - Kafka
  - Kinesis
  - RabbitMQ
  - Azure Queues
  - Azure Service Bus
  - Google Pub/Sub
  - etc…
Goals for our Transport Layer

- In general you have two decisions:
  - Cloud service or your own managed component
  - Guarantee Levels
Buffering Data

- What do we mean by “buffering” and why do we need it?

- Network partitions happen
- Producers and Consumers work at different rates
- Reliable storage is hard
  Stream processing is hard
- Lets do one at a time

This is bad!
Buffering Data – Message Brokers

Publisher

Publisher

Publisher

Message Queue

Subscriber

Subscriber

Subscriber
What is Kafka?

- It’s like a message queue, right?
  - Actually, it’s a “distributed commit log”
  - Or “streaming data platform”

![Diagram of Kafka data flow]

Questions?
tiny.cloudera.com/ny2018questions
# Topics and Partitions

- Messages are organized into topics, and each topic is split into partitions.
  - Each partition is an immutable, time-sequenced log of messages on disk.
  - Note that time ordering is guaranteed within, but not across, partitions.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Partition 0</th>
<th>Partition 1</th>
<th>Partition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 3 4 5 6 7 8</td>
<td>0 1 2 3 4 5 6 7 8</td>
<td>0 1 2 3 4 5 6 7 8</td>
</tr>
</tbody>
</table>

Data Source
In Our Architecture

Producer

Taxi Trip Data

Kafka

Topic
taxi-trip-input

Consumers

Stream Processing (Analytic)

Stream Processing (Lookup)

Stream Processing (Search)

Stream Processing (Long Term)
Input Events

vendor_name,Trip_Pickup_DateTime,Trip_Dropoff_DateTime,Passenger_Count,Trip_Distance,Start_Lon,Start_Lat,Rate_Code,store_and_forward,End_Lon,End_Lat,Payment_Type,Fare_Amt,surcharge,mta_tax,Tip_Amt,Tolls_Amt,Total_Amt

CMT,2009-01-05 08:31:55,2009-01-05 8:37:50,1,0.90000000000000002,-73.977936999999997,40.745919000000001,,,-73.983609000000001,40.755051000000002,Credit,5.2999999999999998,0,,0.79000000000000004,0,6.0899999999999999

Questions?
tiny.cloudera.com/ny2018questions
Kafka Considerations – Reliability

▪ Several knobs we can turn
▪ But remember there are tradeoffs…
Kafka Considerations – Reliability

- Different reliability levels for topics:

  - 100% – dups are ok (“At least once”)
  - <=100% (“At most once”)

Exactly once is now supported (if you really need it).
Kafka Reliability – Replication

Producer

Broker

Leader

Partition 1

Partition 2

Partition 3

Questions?
tiny.cloudera.com/ny2018questions
Kafka Reliability – Replication

Producer

Partition 1

Partition 2

Partition 3
Kafka Reliability – Replication
Kafka Reliability – Replication

Questions?
tiny.cloudera.com/ny2018questions
Kafka Reliability – Replication

Questions?
tiny.cloudera.com/ny2018questions
View from the Consumer

Kafka Broker Cluster

Partition A

Partition B

Partition C

Consumer Group A

Consumer Group B

Questions?

tiny.cloudera.com/ny2018questions
View from the Consumer

Kafka Broker Cluster

Partition A
- B
- A

Partition B
- B
- A

Partition C
- B
- A

Consumer Group A
- Consumer

Consumer Group B
- Consumer
- Consumer
- Consumer

Questions?
tiny.cloudera.com/ny2018questions
Kafka Reliability – Replication

- So how does this relate to our application?

```bash
kafka-topics --zookeeper ZKHOST:ZKPORT --partition 2 --replication-factor 3 \   --create --topic taxi-trip-input

kafka-topics --zookeeper ZKHOST:ZKPORT --partition 2 --replication-factor 1 \   --create --topic customer-sentiment
```
Kafka Reliability – Producers

Producer

Taxi Trip Data

Message failure?

Resend message

acks=all

Kafka

taxi_trip_input

Partition 1
Partition 2
Partition 3

Topic B

Partition 1
Partition 2
Partition 3

Questions?

tiny.cloudera.com/ny2018questions
Kafka Reliability – Producers

- What about duplicates?

<table>
<thead>
<tr>
<th>ID</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>2009-01-04 03:02:00,1,2.629,...</td>
</tr>
<tr>
<td>1001</td>
<td>2009-01-04 03:03:00,3,4.549,...</td>
</tr>
<tr>
<td>1001</td>
<td>2009-01-04 03:38:00,3,4.549,...</td>
</tr>
</tbody>
</table>
Kafka Scaling – Partitions

Producer → Kafka (taxi-trip-input) → Partition 1 → Consumer
Producer → Kafka (taxi-trip-input) → Partition 2 → Consumer
Producer → Kafka (taxi-trip-input) → Partition 3 → Consumer

Consumer Group
Kafka Scaling – Partitions

Producer

Kafka

taxi-trip-input

Partition 1
Partition 2
Partition 3
Partition 4
Partition 5

Consumer Group

Consumer
Consumer
Consumer
Consumer
Consumer

Higher throughput

More resources (memory)

Higher throughput

More resources (file handles)

Questions?
tiny.cloudera.com/ny2018questions
How many partitions?

- Adding partitions late in the game is painful
- Basic formula:
  total desired throughput / throughput of slowest consumer or producer
- Or ~25GB disk space
- Not too many because:
  - Each partition takes broker heap memory and file handles
  - Each partition slows down node shutdown / recovery
  - 1000 – 4000 partitions per broker max
  - Producers will produce smaller batches – lower throughput
Kafka Scaling – Producers

Producers

Kafka

Kafka topic: taxi-trip-input

Consumer Group

Partitions 1 to 5

Consumers
Guarding Against Message Loss

- **Producer** – What happens if the producer loses connection to Kafka and the buffer overflows?
  - You get an exception. You can choose to… block? Write to file?

- **Source** – What happens if events are lost before getting sent to producer?
  - Once again use some kind of buffer to provide sufficient retention of data.
Stream Processing
Considerations
High level architecture

Data Sources
- Code
- Agents
- Log Aggregators

Producers
- Data Governance/Metadata
- Enrichment
- Stream Processing
- Routing
- Transport
- Replication

Streaming Pipes
- Code
- Agents
- Log Aggregators

Storage
- File/Object
- RDBMS/MPP
- Time Series
- Reverse Indexed
- Memory
- Stream

Access
- SQL
- Machine Learning
- Request Response
- Batch Processing

Questions?
tiny.cloudera.com/ny2018questions
Streaming agenda

▪ What do we mean by streaming?
▪ Streaming use-cases
▪ Streaming semantics
▪ Which streaming engine to choose?
▪ Streaming in our use-case
What do we mean by streaming?
What do we mean by streaming?

- Constant low milliseconds & under
- Low milliseconds to seconds, delay in case of failures
- 10s of seconds or more, re-run in case of failures

- Real-time
- Near real-time
- Batch

Questions?
tiny.cloudera.com/ny2018questions
What do we mean by streaming?

- Constant low milliseconds & under
- Low milliseconds to seconds, delay in case of failures
- 10s of seconds or more, re-run in case of failures

Real-time

Near real-time

Batch
But, there’s no free lunch

- Constant low milliseconds & under
- Low milliseconds to seconds, delay in case of failures
- 10s of seconds or more, re-run in case of failures

- Real-time
- Near real-time
- Batch

- “Difficult” architectures, lower latency
- “Easier” architectures, higher latency

Questions?
tiny.cloudera.com/ny2018questions
Streaming use-cases
What are Streaming Use Cases?

- Atomic Enrichment
- Notifications
- Joining
- Partitioning
- Windowing
- Ingestion
- Aggregation & Counting
- Incremental Machine Learning
- Progressive Analytics
Atomic Enrichment

- Isolated Transforms
- Transform with context
- FlatMap & Filtering & Routing
Atomic Enrichment

- Isolated Transforms
Atomic Enrichment

- Transform with context
Atomic Enrichment
- Flat Map & Filtering
Notifications

- Making a decision on a event
- Send out a notification to an action taker
Notifications
Joining

- Joining big to small table
- Broadcast join
- Joining two large tables
- Getting latest on same key from different feeds
- Windowing
Windowing

- A table view of a stream
Windowing Lead and Lag
Windowing Lead and Lag

Questions?
tiny.cloudera.com/ny2018questions
Windowing Lead and Lag

N<B and N<A => Valley
Windowing Lead and Lag

N>B and N>A => Peak
Tumbling Window
Tumbling Window
Sliding Window

Questions?
tiny.cloudera.com/ny2018questions
Sliding Window

Questions?
tiny.cloudera.com/ny2018questions
Sliding Window

Step
Length
Now

Questions?
tiny.cloudera.com/ny2018questions
What to do with Sessions?

- Window of events defined by a gap

- What to do with them
  - Length of sessions
  - Number of sessions within a day
  - Average separation between events in a session
  - Common orders within sessions
  - Counts of types of events within sessions
What to do with Sessions?

Questions?
tiny.cloudera.com/ny2018questions
What to do with Sessions?

- **Start**
- **Length**
- **Min Session Gap Length**
- **New Event**
- **Start**
- **Length**
- **Close session**
- **New Start**
Respect to Inflection Points

- Think about
  - Churn
  - Heart attacks
  - Death

- Inflection point type
- Length before inflection point
Inflection Point Windowing

Questions?
tiny.cloudera.com/ny2018questions
Windowing

- Processing Time
  - Time the event reaches your system
- Event Time
  - Time the event was created
Ingestion

- Taking values from a stream into a data store
  - File System
  - Object Store
  - NoSQL
  - NoSQL Time Series
  - Document Store
  - Lucene based System
  - Spanner System
Ingestion

- File Systems & Object Store
  - Normally you want larger files
  - Normally you want high compression
  - Normally you want deduping
    - Window of deduping
    - 100% deduping in this case is difficult
      - Think about sequence numbering from source
Ingestion

- NoSQL, Lucene, and Spanner Systems
  - If you use the key deduping is given for free
  - Different levels of latency and through put
    - NoSQL
    - Document
    - Spanner
    - Lucene
  - Partitioning on the way in
Ingestion

- **NoSQL Time Series**
  - Dumb inserts
  - You may need Aggregation across metrics
  - To increase throughput you may want to buffer writes

- **Context**
  - Row is a key time and data points There
    - Row1 Time1, DataPoint1
    - Row1 Time2, DataPoint2
  - Is slower then
    - Row1 Time1, DataPoint1, Time2, DataPoint2
Aggregation & Counting

- Counting value with respect to an entity ID
- Counting totals and with respect to window times
Aggregation & Counting

- This is where we talk about Lambda
- There are many definitions
  - Common but not correct: Jobs that involve both Batch & Streaming
  - Correct: Perfect count is not possible with streaming so we use a combination of streaming and batch to show the right value
Aggregation & Counting

- Why is streaming not perfect

Incrementing Speed Layer
- Get Event to Process
  - Increment
  - Acknowledge Event
  - Get Event Again
  - Increment
  - Acknowledge Event

NoSQL
- Value: 10
- Value: 12
- Value: 12
- Value: 14
- Value: 14

Questions?
tiny.cloudera.com/ny2018questions
Aggregation & Counting

- Why is streaming not perfect (more)
- Duping in general

Questions?
tiny.cloudera.com/ny2018questions
Aggregation & Counting

- What is Lambda

Time Window of Events
Streaming Results

So over time lots of errors
Aggregation & Counting
- What is Lambda

Time Window of Events
Streaming Results
Batch Corrections
Batch Corrections
Batch Corrections
Batch Corrections

Less room for error. So in theory more correct

Questions?
tiny.cloudera.com/ny2018questions
Aggregation & Counting

- How is lambda Resolved with out batch?
- Failure problem
- Deduping problem
Aggregation & Counting

- Failure Problem Solved by Adding internal State

Micro Batch Layer

Result as Batch X
Value = 10

Result as Batch X
Value = 12

Micro Batch Layer

Result as Batch X
Value = 10

Result as Batch X
Value = 12

Micro Batch Layer

Result as Batch X
Value = 10

Result as Batch X
Value = 12

Micro Batch Layer

Get Batch Y
Count Batch Y
Update StateByKey
Foreach Value Put
Acknowledge Batch

Reset State to Start of X Batch

Get Batch Y
Count Batch Y Value
Update StateByKey
Foreach Value Put
Acknowledge Batch

Put(14)

Put(14)

NoSQL

Value: 12
Value: 12
Value: 12
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14
Value: 14

Questions?
tiny.cloudera.com/ny2018questions
Aggregation & Counting

- Deduping Problem Solved by
  - Some deduping can happen with in a batch
  - But this is not good enough
- Source Sequence Numbers
  - With Order, Partitioning, and Internal State
Aggregation & Counting

- Deduping with Sequence Numbers

<table>
<thead>
<tr>
<th>Source</th>
<th>Sequence</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Sequence</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>200</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Seq of A</th>
<th>Value of A</th>
<th>Seq of B</th>
<th>Value of B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>3</td>
<td>300</td>
</tr>
</tbody>
</table>
Aggregation & Counting

- Closing Thoughts on Lambda
Progressive Analytics

- Running SQL on a stream
- The idea of streams and tables being one
Progressive Analytics

- Table and Streaming
  - Tables are a set of columns and rows
  - They represent the current state of things
- Stream is a set of change longs
  - The combined total of the changed logs equals a table
Windowing – Example

- Trigger => every 4 messages
- Evict => every 6 message
- Action => Group by Sum

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
</tbody>
</table>

Current State

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
</tbody>
</table>
Windowing – Example

- Trigger => every 4 messages
- Evict => every 6 message
- Action => Group by Sum

**WAL**

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
</tbody>
</table>

**Current State**

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
</tbody>
</table>
Windowing – Example

- Trigger => every 4 messages
- Evict => every 6 message
- Action => Group by Sum

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
</tr>
</tbody>
</table>
Windowing – Example

- Trigger => every 4 messages
- Evict => every 6 message
- Action => Group by Sum

### WAL

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
</tbody>
</table>

### Current State

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
</tbody>
</table>
## Windowing – Example

- Trigger => every 4 messages
- Evict => every 6 message
- Action => Group by Sum

### WAL

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
</tbody>
</table>

### Current State

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
</tbody>
</table>

### Evicted

<table>
<thead>
<tr>
<th>Trigger Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:3</td>
</tr>
<tr>
<td>B:6</td>
</tr>
<tr>
<td>C:3</td>
</tr>
</tbody>
</table>
Streaming semantics
Delivery Types

▪ At most once
  - Not good for many cases
  - Only where performance/SLA is more important than accuracy

▪ Exactly once
  - Expensive to achieve but desirable

▪ At least once
  - Easiest to achieve
Semantics of our architecture

Source System 1
Ingest
Message broker
Extract
Streaming engine
Push
Destination system

Source System 2
Ingest
At least once
Ordered
Partitioned

Source System 3
Ingest
At least once

Questions?
tiny.cloudera.com/ny2018questions
Classification of storage systems

- **File based**
  - S3
  - HDFS

- **NoSQL**
  - HBase
  - Cassandra
  - MongoDB
  - Duid.io

- **Document based**
  - ES & Solr

- **NoSQL-SQL**
  - Kudu, CockroachDB
Classification of storage systems

- **File based**
  - S3
  - HDFS
- **NoSQL**
  - HBase
  - Cassandra
  - MongoDB
  - Duid.io
- **Document based**
  - ES & Solr
- **NoSQL-SQL**
  - Kudu, CockroachDB

De-duplication at file level

Semantics at key/record level
Which streaming engine to choose?
Requirements

- Fault-tolerant and distributed
- Effectively once semantics
- Handle processing time vs. event time
- Allow stateful transformations
Spark Streaming

- Micro batch based architecture
- Allows stateful transformations
- Feature rich
  - Windowing
  - Sessionization
  - ML
  - SQL (Structured Streaming)
Spark Streaming - Gaps

▪ Not as low of a latency
  - Efforts towards reducing latency e.g. RISElab’s Drizzle

▪ Global consistent execution state
  - Stop overall execution of distributed computation
  - Eagerly persist records in transit meaning larger snapshots
Flink

- True “streaming” system, but not as feature rich as Spark
- Much better event time handling
- Good built-in backpressure support
- Allows stateful transformations
- Cleaner APIs
  - for (Triggering, Evicting, State management)
- Lower Latency
  - No Micro Batching
  - Asynchronous Barrier Snapshotting (ABS)
Flink - ABS
Questions?
tiny.cloudera.com/ny2018questions
Flink - ABS

Operator

Barrier 1A Hit

Buffer

Barrier 1B Still Behind

Operator

Both Barriers Hit

Buffer

Check Point

Questions?
tiny.cloudera.com/ny2018questions
Flink - ABS

Operator

Buffer

Both Barriers Hit

Check Point

Operator

Buffer

Barrier is combined and can move on

Buffer can be flushed out

Questions?
tiny.cloudera.com/ny2018questions
Storm

- Old school
- Didn’t manage state – had to use Trident
- No good support for batch processing
Samza

- Good integration with Kafka
- Doesn’t support batch
- Forked by Kafka Streams
Flume

- Well integrated with the Hadoop ecosystem
- Allowed interceptors (for simple transformations)
- Supports buffering
  - Memory
  - File
  - Kafka
- But no real fault-tolerance
- No state management
Kafka Streams

- Good integration with Kafka
- Light-weight library (not a framework)
- No micro-batching, uses Kafka as internal messaging layer
- Maintains local state per node (in RocksDB, or in memory hash map)
- Handles late events
- Stream-to-stream joins
Kafka Streams architecture
Storage Layer
Considerations
High level architecture

Questions?
tiny.cloudera.com/ny2018questions
Data Modeling
Structured Landing Zones

- Long Term Storage
  - Local - Remote
  - Styles
    - Relational
    - Nested
- Speed Layer
- Time Series
- Reverse Index
- Graph
HDFS
Basic of GFS => HDFS

- NameNode
  - Metadata of all the files/blocks
  - Which data node they are assigned too
  - Replication management
- Data Nodes
  - Metadata for each block location on disk
Basic of GFS => HDFS

Write Path
A. Ask Name Node for Location to Write
B. Write to DataNode with NN Instructions
C. DataNode does replication
D. Confirms file is persisted to client
Basic of GFS => HDFS

- File are immutable
- File can be of any type
- Files are block up into Blocks (128MB -> 1GB)
  - Metadata cost is at the Block not the data size
- File may be splittable or may not be when reading
Object Stores
Object Store (Like and not like HDFS)

- Like HDFS
  - Contains files
  - Break up large files

- Not like HDFS
  - Not really a file system and is more Key value like a NoSQL
  - Doesn’t have any metadata limit problem
  - Traversing Folder directories is more work
  - There is no rename, only copy and delete
  - Eventually consist issues with listing files
  - (seen with things like MR and Spark)
  - Can be mostly addressed with EMRFS
Object Store (Many Ideas Continued)

- Because it is like a File System
  - Almost everything you learned from HDFS stores true
  - Hive works the same
  - Files and file formats are the same
Object Store (Thinking Remote)

- Unlike HDFS the storage is always remote
  - Not on the same nodes as the execution
- Which allow you to save money in the cloud
  - Execution nodes are expensive vs storage only
- Network will be used to Read and Write
  - In fact you are normally throttled well before the network limit of your node
- You will want the highest rates of compression possible
  - To save money on storage
  - To read and write faster
Background Information
Compression Styles and Entropy
Compression Styles and Entropy

Questions?
tiny.cloudera.com/ny2018questions
Compression Codecs

- **Snappy**: 2x-3x : Fast Read, Fast Write
- **Lzo**: 2x-3x : Fast Read, Fast Write
- **Gzip**: ~8x: ~Fast Read, Normal Write
- **Default**: ~8x: ~Fast Read, Normal Write
- **BZip2**: ~10x ~Fast Read, Slow Write
- Others ..

- **Always be skeptical**
  - All data compresses differently
  - Use your own data
Introducing the Hive Metastore

- **Hive Metastore**
- Adds a table like metadata layer over a file system, block store, NoSql, or other
- Allows for SQL access
- Allows for greater security options
- Allows for external metadata
- Allows for partitioning
Typical Hive Table

- ParentFolder
  - TableFolder
    - Date=20171212
      - DataFiles
      - DataFiles
    - Date=20171211
      - DataFiles
      - DataFiles
Access Patterns

- Partitioning
- Filter push down
- Indexing should be considered poor
- Ideal for large scans
- Bucketing and Sorting
Relational Storage
Thinking about Object/Tables

1. Let's start off easy

1. Use Case: We are a Netflix type company and we have a log of users and movies watched that looks something like this:

<table>
<thead>
<tr>
<th>User ID</th>
<th>Age</th>
<th>Account Start Date</th>
<th>Category Of User</th>
<th>Movie Watched</th>
<th>Movie Category</th>
<th>Start Time</th>
<th>Events List</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>42</td>
<td>12/12/2012</td>
<td>Basic</td>
<td>Die Hard</td>
<td>Action</td>
<td>5/4/2016 12:00</td>
<td>Play 0, pause at 15, FF at 40 to 55, E at 90</td>
</tr>
<tr>
<td>Kat</td>
<td>31</td>
<td>12/12/2012</td>
<td>Platum</td>
<td>Beauty and the Beast</td>
<td>Family</td>
<td>5/4/2016 12:00</td>
<td>Play 0, pause at 15, FF at 40 to 55, E at 90</td>
</tr>
</tbody>
</table>
Thinking about Object/Tables

1. To make this into objects we need to do some separation
Query Considerations

- Data is normally big so
- Partition respectively to access patterns
- Join with care
- Consider sampling or local testing before experimenting
- Data is files
- Latency to accessibility is high – seconds, minutes or more.
Look for big tables
Mutation Patterns

- File is written once and can not be mutated
- Fine for append or snapshot use cases
- Mutation will require a compaction
## Compaction Recap

<table>
<thead>
<tr>
<th>Key</th>
<th>Time</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>101</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key</th>
<th>Time</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>103</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>103</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key</th>
<th>Time</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>103</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>103</td>
</tr>
</tbody>
</table>

Questions?

[tiny.cloudera.com/ny2018questions](tiny.cloudera.com/ny2018questions)
View Strategies

- Hive Relational Model
- Hive Nested Model
- Hive Normal Views
- Hive Materialized Table Views

Use in the cases where the view requires a join that is done through a shuffle

Use only for tables that filter records/columns or use for marking fields

Questions?
tiny.cloudera.com/ny2018questions
Nested Structures
Nested

- Less Space than Denormalization
- Still have tables but the cost of joins is all but gone
- Also great for cartesian joins
  - $N \times M$ vs $N + M$
- Not really supported yet with Kudu or HBase with SQL
Nested Example

CREATE TABLE fact_contacts (id BIGINT, name STRING, address STRING) STORED AS PARQUET;
CREATE TABLE dim_phones
(
    contact_id BIGINT,
    category STRING,
    international_code STRING,
    area_code STRING,
    exchange STRING,
    extension STRING,
    mobile BOOLEAN,
    carrier STRING,
    current BOOLEAN,
    service_start_date TIMESTAMP,
    service_end_date TIMESTAMP
)
Nested Example

CREATE TABLE contacts_detailed_phones (
    id BIGINT, name STRING, address STRING,
    phone ARRAY < STRUCT <
        category: STRING,
        international_code: STRING,
        area_code: STRING,
        exchange: STRING,
        extension: STRING,
        mobile: BOOLEAN,
        carrier: STRING,
        current: BOOLEAN,
        service_start_date: TIMESTAMP,
        service_end_date: TIMESTAMP
    >>
) STORED AS PARQUET;

https://www.cloudera.com/documentation/enterprise/latest/topics/impala_complex_types.html
De-normalized vs Nested

- **Nested Pros**
  - Co-location
  - Faster to group by
  - Faster to window
  - Joins are free
  - Less data
  - Better compression
  - Tables and Columns can be read without penalty from one not read
  - Great for limiting the effort are Cartesian Joins

- **Nested Cons**
  - Size limitation of parent row
  - Adding child requires the re-write the whole parent record
Options for appending Nested

- It is all about the parent record
- We can add more than one Partition key for the parent
- In our use case
  - User & watch month or day
Storage and In Memory

- Also don’t limit the idea of nested to just tables
- In Spark they can be used as in memory constructs to
  - conserve on networking
  - In memory cost
Nested Writing Example in Spark

{
  "id": "0001",
  "type": "donut",
  "name": "Cake",
  "ppu": 0.55,
  "batters":
  {
    "batter":
    [
    { "id": "1001", "type": "Regular" },
    { "id": "1002", "type": "Chocolate" },
    { "id": "1003", "type": "Blueberry" },
    { "id": "1004", "type": "Devil's Food" }
    ]
  },
  "topping":
  [
  { "id": "5001", "type": "None" },
  { "id": "5002", "type": "Glazed" },
  { "id": "5003", "type": "Sugar" },
  { "id": "5004", "type": "Chocolate with Sprinkles" }
  ]
}
Nested Writing Example in Spark

val jsonDF = hiveContext.read.json(jsonRDD)

jsonDF.write.parquet("./parquet")

hiveContext.createExternalTable("jsonNestedTable", "/parquet")
Time Series
Time Series Options

- HBase
- Cassandra
- InfluxDB
- Data Dog
- DruidDB
- Many More
Entity Centric Time Series

- Partition by Entity ID
- Order by Time
- Allows for free windowing
- Allows for fetching of single time window of single entity at web scale
HBase Entity Time Series

Questions?

tiny.cloudera.com/ny2018questions
HBase Entity Time Series

Rest Call → Short Scan

- Cust-A, 10
- Cust-A, 20
- Cust-A, 40
- Cust-C, 10
- Cust-C, 20
- Cust-C, 30
- Cust-C, 40

- Cust-B, 10
- Cust-B, 20
- Cust-B, 30
- Cust-B, 40
- Cust-F, 20
- Cust-F, 30
- Cust-F, 40

- Cust-D, 10
- Cust-D, 20
- Cust-D, 40
- Cust-G, 10
- Cust-G, 20
- Cust-G, 30
- Cust-G, 40

Questions?
tiny.cloudera.com/ny2018questions
HBase Entity Time Series

Mapper

Cust-A, 10
Cust-A, 20
Cust-A, 40
Cust-C, 10
Cust-C, 20
Cust-C, 30
Cust-C, 40

Mapper

Cust-B, 10
Cust-B, 20
Cust-B, 30
Cust-B, 40
Cust-F, 20
Cust-F, 30
Cust-F, 40

Mapper

Cust-D, 10
Cust-D, 20
Cust-D, 40
Cust-G, 10
Cust-G, 20
Cust-G, 30
Cust-G, 40
HBase Entity Time Series

- Cust-A, 10
- Cust-A, 20
- Cust-A, 40
- Cust-C, 10
- Cust-C, 20
- Cust-C, 30
- Cust-C, 40

- Cust-B, 10
- Cust-B, 20
- Cust-B, 30
- Cust-B, 40
- Cust-F, 20
- Cust-F, 30
- Cust-F, 40

- Cust-D, 10
- Cust-D, 20
- Cust-D, 40
- Cust-G, 10
- Cust-G, 20
- Cust-G, 30
- Cust-G, 40
What is meant by Bucketing and Sorting

- Partitioning on a Key
- Then sorting on that key + another field(s)
- Example
  - User_id + Watch Event Time
Example of Bucketed Sorted

- Cust-A, 10
- Cust-A, 20
- Cust-A, 40
- Cust-C, 10
- Cust-C, 20
- Cust-C, 30
- Cust-C, 40
- Cust-B, 10
- Cust-B, 20
- Cust-B, 30
- Cust-B, 40
- Cust-F, 10
- Cust-F, 20
- Cust-F, 40
- Cust-D, 10
- Cust-D, 20
- Cust-D, 40
- Cust-G, 10
- Cust-G, 20
- Cust-G, 30
- Cust-G, 40
Good for Appending Nested

New Data

| Cust-A, 50 |
| Cust-A, 60 |
| Cust-B, 50 |
| Cust-B, 60 |
| Cust-C, 50 |
| Cust-D, 50 |
| Cust-G, 50 |

Existing Data

| Cust-A, 10 |
| Cust-A, 20 |
| Cust-A, 40 |
| Cust-C, 10 |
| Cust-C, 20 |
| Cust-C, 30 |
| Cust-C, 40 |

| Cust-B, 10 |
| Cust-B, 20 |
| Cust-B, 30 |
| Cust-B, 40 |
| Cust-F, 10 |
| Cust-F, 20 |
| Cust-F, 40 |

| Cust-D, 10 |
| Cust-D, 20 |
| Cust-D, 40 |
| Cust-G, 10 |
| Cust-G, 20 |
| Cust-G, 30 |
| Cust-G, 40 |

Questions?
tiny.cloudera.com/ny2018questions
Good for Appending Nested

<table>
<thead>
<tr>
<th>New Data</th>
<th>Existing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cust-A, 50</td>
<td>Cust-A, 10</td>
</tr>
<tr>
<td>Cust-A, 60</td>
<td>Cust-A, 20</td>
</tr>
<tr>
<td>Cust-B, 50</td>
<td>Cust-A, 40</td>
</tr>
<tr>
<td>Cust-B, 60</td>
<td>Cust-B, 10</td>
</tr>
<tr>
<td>Cust-C, 50</td>
<td>Cust-B, 20</td>
</tr>
<tr>
<td>Cust-D, 50</td>
<td>Cust-B, 30</td>
</tr>
<tr>
<td>Cust-G, 50</td>
<td>Cust-B, 40</td>
</tr>
<tr>
<td></td>
<td>Cust-C, 10</td>
</tr>
<tr>
<td></td>
<td>Cust-B, 10</td>
</tr>
<tr>
<td></td>
<td>Cust-C, 20</td>
</tr>
<tr>
<td></td>
<td>Cust-B, 20</td>
</tr>
<tr>
<td></td>
<td>Cust-C, 30</td>
</tr>
<tr>
<td></td>
<td>Cust-B, 30</td>
</tr>
<tr>
<td></td>
<td>Cust-C, 40</td>
</tr>
<tr>
<td></td>
<td>Cust-B, 40</td>
</tr>
<tr>
<td></td>
<td>Cust-F, 10</td>
</tr>
<tr>
<td></td>
<td>Cust-C, 40</td>
</tr>
<tr>
<td></td>
<td>Cust-F, 20</td>
</tr>
<tr>
<td></td>
<td>Cust-C, 50</td>
</tr>
<tr>
<td></td>
<td>Cust-G, 10</td>
</tr>
<tr>
<td></td>
<td>Cust-D, 10</td>
</tr>
<tr>
<td></td>
<td>Cust-D, 20</td>
</tr>
<tr>
<td></td>
<td>Cust-D, 40</td>
</tr>
<tr>
<td></td>
<td>Cust-G, 10</td>
</tr>
<tr>
<td></td>
<td>Cust-D, 20</td>
</tr>
<tr>
<td></td>
<td>Cust-G, 20</td>
</tr>
<tr>
<td></td>
<td>Cust-G, 30</td>
</tr>
<tr>
<td></td>
<td>Cust-F, 10</td>
</tr>
<tr>
<td></td>
<td>Cust-G, 40</td>
</tr>
<tr>
<td></td>
<td>Cust-F, 20</td>
</tr>
<tr>
<td></td>
<td>Cust-G, 40</td>
</tr>
</tbody>
</table>
Good for Appending Nested

New Data

Cust-A, 50
Cust-A, 60
Cust-C, 50
Cust-B, 50
Cust-B, 60
Cust-D, 50
Cust-G, 50

Existing Data

Cust-A, 10
Cust-A, 20
Cust-A, 40
Cust-C, 10
Cust-C, 20
Cust-C, 30
Cust-C, 40
Cust-B, 10
Cust-B, 20
Cust-B, 30
Cust-B, 40
Cust-F, 10
Cust-F, 20
Cust-F, 40
Cust-D, 10
Cust-D, 20
Cust-D, 40
Cust-G, 10
Cust-G, 20
Cust-G, 30
Cust-G, 40

Questions?
tiny.cloudera.com/ny2018questions
Good for Appending Nested

Merge Join

Order Retained

Cust-A, 50
Cust-A, 60
Cust-C, 50

Cust-A, 10
Cust-A, 20
Cust-A, 40

Cust-C, 10
Cust-C, 20
Cust-C, 30
Cust-C, 40

Cust-A, 50
Cust-A, 60
Cust-C, 50

Cust-A, 10
Cust-A, 20
Cust-A, 40

Cust-C, 10
Cust-C, 20
Cust-C, 30
Cust-C, 40
NoSQL
What is a NoSQL

- It’s not NO SQL
- It’s not a Database
- Think of it more like a
  - HashMap
  - Log
  - Bucketed and Ordered
Hash Map

- There is a Key and a Value
- It is really fast to grab a key/value
- It is really fast to add a key/value
- Iteration is also possible

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
</tr>
</tbody>
</table>

Client
Log with Compactions

- When new records come in they don’t rewrite the old
- They compact in

<table>
<thead>
<tr>
<th>Key</th>
<th>Time</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>101</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key</th>
<th>Time</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>103</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>103</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key</th>
<th>Time</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>102</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>103</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>103</td>
</tr>
</tbody>
</table>

Questions?
tiny.cloudera.com/ny2018questions
Log with Compactions

- **Write Path**
  - Get Local for Record (Cached)
  - First to WAL
  - Then to Memstore
  - Sorting & batching
  - Flush to New Hfile
  - Later Hfiles will be compacted
Ordered

- All Records Columns are ordered
- Ordering allows for simpler indexing
- Ordering allows for simpler compactions

- We will also use this ordering
  - Windowing
  - Time series
  - Local scanning
Bucketing or Partitions

- HBase
  - Out of the Box:
    - Range
    - Desired:
      - Salt
- Cassandra
  - Out of the Box:
    - HashMod
    - Bucketed HashMod
So what about SQL

- Well SQL could totally work
- CQL for cassandra
- Hive and SparkSQL on Hbase & Cassandra

- Why is it not the best idea
- Built more for point look ups
- Scans are not as fast as parquet
  - However the mutability may be more important than speed
- Partitioning is not simple
  - It must be put into the key
Let’s talk about CAP for a Minute

- Strong Consistency
- HBase & Kudu
- Variable Consistency
- Cassandra
HBase Model

- Region Server owns range splits
- Region Server 1 fails
- Master needs to figure that out
- Master needs to assign new Region Server to own splits
- Region Server 2 has to get organized
- Region Server 2 is read to server reads and writes
Cassandra Model

Questions?
tiny.cloudera.com/ny2018questions
Cassandra Model
Cassandra Model (Common Models)

- **3 Write - 1 Read**
  - Client
  - Replica Node (Has Replica)
  - Client
  - Replica Node (Has Replica)
  - Client
  - Replica Node (Has Replica)

- **1 Write - 3 Read**
  - Client
  - Replica Node (Has Replica)
  - Client
  - Replica Node (Has Replica)
  - Client
  - Replica Node (Has Replica)

- **1 Write - 1 Read**
  - Client
  - Replica Node (Has Replica)
  - Client
  - Replica Node (Has Replica)
  - Client
  - Replica Node (Has Replica)
What do they share (Head Types)

- Master and a HA

- Leader Election
What do they share (Headless)
What do they share (CAP theorem)
What do they share (CAP theorem)
What do they share (CAP theorem)

- Cheating the CAP Theorem
- Cassandra is a good model
- Where they expand the definition of failure with variable consistency
- CAP still holds but …
Indexed Search
Lucene Indexing (Features)

- We don’t have enough time in this whole class
- Ordering logic
- NGrams
- Weights
- Text Indexing
- Translations
- Facets *
Lucene Indexing (Facets)

- Facets are a side effect of our wonderful indexes
- It allows us to count all the documents that belong to given indexes to produce
  - Grouped Counts
  - Charts and Graphs (Kibana or Banana)
- People will also call this access pattern cubing a dataset
Lucene Indexing (Kibana & Banana)

Questions?
tiny.cloudera.com/ny2018questions
## Lucene Indexing (Facets Example)

### Time Series Example

<table>
<thead>
<tr>
<th>Document ID</th>
<th>Hour of Day</th>
<th>User</th>
<th>State</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>4201</td>
<td>MD</td>
<td>click</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>4202</td>
<td>VA</td>
<td>click</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>4203</td>
<td>VA</td>
<td>click</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4201</td>
<td>MD</td>
<td>click</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>4202</td>
<td>VA</td>
<td>view</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>4204</td>
<td>CA</td>
<td>click</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>4205</td>
<td>VA</td>
<td>view</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4201</td>
<td>MD</td>
<td>click</td>
</tr>
</tbody>
</table>
## Lucene Indexing (Facets Example)

### Document ID | Hour of Day | User | State | Event
---|---|---|---|---
1 | 12 | 4201 | MD | click
2 | 12 | 4202 | VA | click
3 | 12 | 4203 | VA | click
4 | 1 | 4201 | MD | click
5 | 1 | 4202 | VA | view
6 | 2 | 4204 | CA | click
7 | 2 | 4205 | VA | view
8 | 2 | 4201 | MD | click
9 | 2 | 4204 | CA | click

### Hour of Day

<table>
<thead>
<tr>
<th>Hour of Day</th>
<th>12</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
</table>

### User

<table>
<thead>
<tr>
<th>User</th>
<th>4201</th>
<th>4202</th>
<th>4203</th>
<th>4204</th>
<th>4205</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>click</td>
<td>view</td>
<td>view</td>
<td>click</td>
<td>view</td>
</tr>
</tbody>
</table>

### State

<table>
<thead>
<tr>
<th>State</th>
<th>MD</th>
<th>VA</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>1</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

### Event

<table>
<thead>
<tr>
<th>Event</th>
<th>click</th>
<th>view</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hour of Day</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Lucene Indexing (Facets Example)

- Events per hour
- Simple array count

<table>
<thead>
<tr>
<th>Hour of Day</th>
<th>12</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

Events Per Hour

Questions?
tiny.cloudera.com/ny2018questions
Lucene Indexing (Facets Example)

- Events per hour by State
- Simple array count

<table>
<thead>
<tr>
<th>State</th>
<th>Hour of Day</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>MD</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>VA</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td>CA</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

Events per hour by State

Questions?

tiny.cloudera.com/ny2018questions
Lucene Indexing (Facets Example)

- Note the bucketing and ordered pattern

<table>
<thead>
<tr>
<th>Hour of Day</th>
<th>12</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State</th>
<th>MD</th>
<th>VA</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>1</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>VA</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>CA</td>
<td>6</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hour of Day 2</th>
<th>State MD</th>
<th>State VA</th>
<th>State CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>7</td>
</tr>
</tbody>
</table>
Lucene Indexing (Facets Example)

- Note the bucketing and ordered pattern

<table>
<thead>
<tr>
<th>Hour of Day 2</th>
<th>State MD</th>
<th>State VA</th>
<th>State CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hour of Day 2</th>
<th>State MD</th>
<th>State VA</th>
<th>State CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

+1 CA
Lucene Indexing (Facets Example)

- Note the bucketing and ordered pattern

<table>
<thead>
<tr>
<th>Hour of Day 2</th>
<th>State MD</th>
<th>State VA</th>
<th>State CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

+1 VA

<table>
<thead>
<tr>
<th>Hour of Day 2</th>
<th>State MD</th>
<th>State VA</th>
<th>State CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>7</td>
</tr>
</tbody>
</table>

+1 MD

<table>
<thead>
<tr>
<th>Hour of Day 2</th>
<th>State MD</th>
<th>State VA</th>
<th>State CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

+1 CA

Questions?
tiny.cloudera.com/ny2018questions
Partitioning

- Solr and Elastic Search partition the document o land on all nodes
- This means
  - You have the power of the cluster when querying
  - This mean you are accessing the cluster when querying
Writing Latency

- Lucene Indexing is more expensive than NoSQL work
- Think of it as micro batching
  - Larger batches ≈ better throughput
- Compaction is also invalid
- Deletes impact storage and performance until they are compacted
Storage Cost

- TTL is your friend
- Think of Lucene based systems as great if
  - Your dataset is manageable in size
  - You have a good TTL strategy
  - You have a boat load of money
Graphs
Thinking in terms of Graphs

- Nodes and Edges
Thinking in terms of Graphs

- Use cases
- Querying
  - Cassandra with Sparkle
  - Neo4j
- Batch operations
  - Giraph
  - GraphX
  - GraphLab
BSP Bulk Synchronous Parallel

- Process every Node Atomically
  - Node gets all messages sent to it
  - Nodes can mutate themselves and their edges
  - Nodes can send messages to other nodes
    - But nothing is received yet
- BSP waits until all the Node processing is done
- Then send messages to the right partition
- Repeat
Honorable Mentions

Kudu
CockroachDB
Druid.io
Why Honorable Mentions
Kudu

1. Replace the Region Servers with Tablet Servers
2. Replace block format HFile files with a parquet like TFiles
3. Replace the byte array focused HBase API with one that is more JDBC friendly
4. Tight integration with Spark SQL and Impala for SQL
5. Completely rewrite the compaction process to make for perfectly sized files with our having major compactions but always manageable micro compactions.
Kudu

Questions?
tiny.cloudera.com/ny2018questions
CoackroachDB

1. Think about Kudu but
   1. Nested tables
   2. Transitions
CockroachDB

Questions?
tiny.cloudera.com/ny2018questions
Druid.IO

1. Not a general SQL type store
2. More focused on metric
   1. Where larger aggregations may be required
   2. Many variables for same metric tag combinations
Druid.IO

Client

Broker Cluster

Broker Node  Broker Node  Broker Node

Query Planning and Response Preparation

Try to optimize read path based on time request of the query

Real Time Cluster

Real Time Node  Real Time Node  Real Time Node

Batch Ingestion

History Cluster

History Node  History Node  History Node

Pluggable Storage

ZooKeeper

Metadata Storage

House keeping services

Short TTL for hot memory cache

Main Ingestion Path

Streaming Ingestion

Async large batching

Questions?
tiny.cloudera.com/ny2018questions
Data Access
High level architecture

Data Sources
- Code
- Agents
- Log Aggregators

Producers
- Data Governance/Metadata
- Enrichment
- Stream Processing
- Routing
- Transport
- Replication

Streaming Pipes
- Data Governance/Metadata
- Enrichment
- Stream Processing
- Routing
- Transport
- Replication

Storage
- Long Term SQL
- Speed Layer SQL
- Time Series
- Reverse Indexed
- Stream

Access
- SQL
- Machine Learning
- Request Response
- Batch Processing

Questions?
tiny.cloudera.com/ny2018questions
Data Access

Batch Processing
Why have batch processing?

- When you need a larger context
  - Say, to train a model
- Complex periodic job that does something
  - Convert data to a nested structure for reduced number of shuffles
- For example,
  - Kudu -> HDFS Nested is batch processing
  - KMeans calculation, etc.
Types of Processing

- ETL
- Analytics
- Graph Analytics
- Machine Learning
- Deep Learning
Batch processing options

- Spark (+ MLlib)
- MapReduce (+ Mahout)
- Flink (+ Flink ML)
- Distributed query engines
Spark

- Pretty popular
- Much faster than MapReduce
- Thriving community
MapReduce

- S.looo.oow
Flink

- Pretty popular
- Batch is a special case of Streaming
- Developing community
Just SQL

- Apache Spark SQL
- Hive
- MR, Spark, Tez
- Impala
- Presto
Distributed Machine Learning

- Apache Spark MLlib
- Apache Spark Custom
- H2O
- Data Robot
- Google Tensorflow
- Microsoft Cognitive
Data Access

Interactive Access

Strata Data Conference
Types of data access

- REST server/APIs for querying entities and aggregates
- UI for displaying search facets
- SQL engine
REST servers
REST Servers

```java
import org.mortbay.jetty.Server
import org.mortbay.jetty.servlet.{Context, ServletHolder}

val server = new Server(port)
val sh = new ServletHolder(classOf[ServletContainer])
sh.setInitParameter("com.sun.jersey.config.property.resourceConfigClass", "com.sun.jersey.api.core.PackagesResourceConfig")
sh.setInitParameter("com.sun.jersey.config.property.packages", "com.hadooparchitecturebook.taxi360.server.hbase")
sh.setInitParameter("com.sun.jersey.api.json.POJOMappingFeature", "true")
val context = new Context(server, "/", Context.SESSIONS)
context.addServlet(sh, "/*
server.start()
server.join()
```
UI
Considerations
UI requirements

Something that can

- Represent search results really well
- Integrates with Apache Solr, query engines, etc.
UI options

- Hue
- Banana
- Kibana
Hue In Our Use Case

- Because it’s included
- Please look at the others
Data Access

SQL Engines
SQL engine criteria

- Low latency SQL access
- Allows for high concurrency
- JDBC/ODBC integration
- Capable of large scale aggregation
- Optionally integrates with Multiple Storage Systems for real-time updates to SQL tables
Apache Hive

- Good JDBC integration
- Not really low latency, even when using Tez
- Doesn’t integrate with Kudu
- Can run with MapReduce, Spark, or Tez
Apache Drill

- Similar in architecture to Impala
- Provides JDBC/ODBC access
- Doesn’t integrate with Kudu
Spark SQL

- Builds on top of Spark
- JDBC/ODBC access only via Spark Thrift Server
  - Doesn’t scale well with larger number of concurrent users
  - Doesn’t fully provide secure access.
Presto

- Low latency SQL engine from Facebook
- Provides JDBC/ODBC access
- Is only in-memory, large aggregations can lead to OOM errors
- Doesn’t integrate with Kudu
Apache Impala

- Low latency SQL access
- Provides JDBC/ODBC access
- Excellent concurrency support
- Integrates with Kudu for real-time SQL
Where else to find us?
Other Sessions

- Executive Briefing: Managing successful data projects – Wednesday, 2:55
Thank you!

tiny.cloudera.com/app-arch-ny2018
Ted Malaska | @tedmalaska
Jonathan Seidman | @jseidman