Operational Hadoop and the Lambda Architecture for Streaming Data

Dale Kim, MapR Technologies

June 10, 2015
Topics

- From Batch to Operational Workloads on Hadoop
- Streaming Data Environments
- The Lambda Architecture – the What and Why
- Example Deployments
Review: Apache Hadoop Does This Big Compute Task
Scaling on Hadoop

Data volume, velocity

- Low: Scale out with commodity hardware
- High

Data variety

- Low: Use the right tool for unstructured, multi-structured, semi-structured, non-relational data
- High
Scaling on Traditional Technologies

Data volume, velocity
- Low: Scale up to bigger, faster machines
- High

Data variety
- Low: Extensive data modeling and ETL
- High
Batch versus Operational

Batch is…
a built-in delay.

You might have to wait here for the next window

Operational is…
an overloaded term.

Real-Time?

Business operations?

Interactive?

Industrial operations?

In this context: more this…

than this…
Examples of Operational Workloads

• Immediate reads/writes/updates to live data

• Immediate access to data

• Real-time analysis (operational reporting)

Let’s separate batch from operational in common Hadoop use cases…
Common Uses for Hadoop

- Enterprise Data Hub (or “Data Lake”)
- Anomaly Detection
- Internet of Things
- Customer 360 View
- Fraud Detection
- Recommendation Engine
Enterprise Data Hub Architecture

Load more data sources

LOAD

RELATIONAL, SAAS, MAINFRAME

DOCUMENTS, EMAILS

BLOGS, TWEETS, LINK DATA

STREAMING

Offload / Enrich / Reload

data marts

data warehouse

Analyze

BI REPORTS AND APPLICATIONS

LOAD

Graphical User Interfaces

PARSE, PROFILE, ETL

CLEANSE, MATCH

Enrich data in Hadoop

Batch Processing
MR, YARN, Hive, Pig, etc.

Interactive Querying
Drill, Impala, Presto, etc.

HBase, other data stores

Distributed File System

LOAD

Offload / Enrich / Reload
Example Data Flow within Hadoop

Hadoop

Canonical, immutable data → Transformed data → Snapshot data → Analyst data
Lambda Architecture

ALL DATA → PRECOMPUTE VIEWS

BATCH RECOMPUTE

NEW DATA STREAM → BATCH VIEWS

Partial aggregate → Partial aggregate → Partial aggregate

MERGE

REAL-TIME VIEWS → MERGED VIEW

REAL-TIME DATA

INCREMENT VIEWS → REAL-TIME INCREMENT

PROCESS STREAM
Why the Lambda Architecture?

• Explained in “How to Beat the CAP Theorem” blog by Nathan Marz

• What is the CAP Theorem?
  – You have to compromise with distributed systems
  – Consistency – data in is same as data out
  – Availability – the system can always respond with an answer
  – Partition Tolerance – the system still runs if nodes lose a connection
  – You get 2 out of 3 (actually, either C or A)

• The Lambda Architecture can be about more than beating CAP
Key Principles of the Lambda Architecture

• Store an immutable, constantly growing data set
• Only add new data points
• Recompute queries from raw data to avoid complexity
• Uses incremental algorithms to reduce query latency
Lambda Architecture Batch Layer

ALL DATA \rightarrow \text{PRECOMPUTE VIEWS}

\text{NEW DATA STREAM}

\text{BATCH LAYER}

\text{BATCH VIEWS}

\text{Partial aggregate} \rightarrow \text{Partial aggregate} \rightarrow \text{Partial aggregate}
Lambda Architecture Speed Layer
Lambda Architecture Serving Layer

- BATCH VIEWS
  - Partial aggregate
  - Partial aggregate
  - Partial aggregate

- REAL-TIME VIEWS
  - REAL-TIME DATA

- MERGED VIEW

- SERVING LAYER
  - MERGE
Lambda Architecture

ALL DATA (HDFS) → PRECOMPUTE VIEWS (Hive) → BATCH VIEWS → REAL-TIME VIEWS → REAL-TIME DATA → INCREMENT VIEWS

NEW DATA STREAM → PROCESS STREAM → REAL-TIME INCREMENT

KAFKA → MERGE VIEW (HBASE)

HADOOP

BATCH LAYER

SERVING LAYER

SPEED LAYER
Example Use Case – Predictive Maintenance

• Data and use case based on one of our customers
  – Time series sensor data from oil well pumps
  – Transactions and predictive analytics to predict imminent pump failure

• Data characteristics
  – Location, time stamp, electrical current, displacement, flow, sediment PPM, etc.
  – High volume, high velocity
  – Higher resolution (measurements at a higher frequency) enables more accurate analytics
Example Use Case – Predictive Maintenance (cont.)

- Sample queries and analysis
  - Dashboard of state of oil field
  - Failed pump measurement pattern
  - Imminently failing measurement pattern
  - Drill-down navigation
What about Historical Analytics on Streaming Data

• Need static view of data
  – Data can’t be moving as you’re building your model

• Need repeatable view of data
  – What if model is wrong? How can you fix it if data has changed?

• Need comparisons for better models
  – Alternate models should be run on same data

• Validate legitimacy for future data states
  – Compare to previous state

Check out the different snapshot implementations on Hadoop
Customer data, network security event data

Anomaly detection on large volumes of security event data, analytics on customer data to enable incremental sales
Telecommunications Company

Customer profile data, customer behavior data

Analytics on customer behavior for better recommendations