Unified Batch & Stream Processing Platform

Himanshu Bari
Director Product Management
Most Big Data Use Cases Are About

Improving/Re-write EXISTING solutions To KNOWN problems…
Current Solutions Were Built On

A. Imperfect information
B. Expensive s/w & h/w infrastructure
C. Relational data stores
Inevitable Course of the Re-write

Specialized solutions

Near perfect information

Next gen data management platform

In Memory Processing

Commodity hardware

Open source software

NoSQL

Graph

Hadoop

Search

EDW + RDBMS

© 2015 DataTorrent Confidential – Do Not Distribute
Every Batch Process *Could* Have Been A Stream Process

- Every ‘Static’ data point was ‘Streaming’ at some point
- We choose to wait and collect a bunch of data points and then process them at once in ‘Batch Mode’
- Move processing time closer to the data generation or ‘Event’ occurrence time
- Reduces time to insight and allows you to be proactive with timely actions
But We Still Need Batch

- Historical data analysis
- What-if analysis
- Experimentation
- Data re-statement
- Transaction processing and reconciliation
- Audit
- Machine learning model training
- ……..
Need A Unified Platform

Unified Batch & Stream Processing Engine to execute data pipelines

- Streaming only pipelines
- Batch processing only pipelines
- Dual Batch & Stream processing pipelines

YARN

HDFS

Enterprise Repositories

RDBMS
EDW
Search
NoSQL
Graph

Any Notification Engine
Any BPM Engine

© 2015 DataTorrent Confidential – Do Not Distribute
DataTorrent RTS Provides A Unified Batch & Streaming Platform

Transform Normalize ➔ Descriptive & Predictive Analytics ➔ Alert Action

Unified Batch & Streaming Data Ingestion & Distribution Pipeline for Hadoop

Graphical Application Assembly ➔ Real-Time Data Visualization

Re-Usable Java Operator Library
Apache 2.0 licensed – Project Malhar

Scalable, High Performance, Fault Tolerant In-Memory Data Processing Platform
Apache 2.0 licensed - Project Apex

Hadoop 2.0 —YARN + HDFS

Physical      Virtual      Cloud

Management & Monitoring
DataTorrent Project Apex - Unified Batch & Stream Processing Engine

1. In-memory compute engine
2. Hadoop native (YARN + HDFS) architecture
3. Sub-second event processing
4. Event order guarantees
5. Flexible stream partitioning mechanism
6. Stateful fault tolerance with NO data loss
7. Automatic & incremental recovery
8. Rolling & tumbling windows
9. Processing guarantees
10. Ability to update application without downtime
Unified Data Ingestion & Distribution Pipeline

**Input & Output Variety**
- FTP, S3 etc.
- Kafka & JMS
- Change data capture

**Tackle data size & speed fluctuations**
- Overcome HDFS small file problem
- Skew management

**Hadoop Native**
- Runs within the Hadoop cluster over YARN & HDFS

**Easily customizable**
- Easily extend and insert operations for data preparation

**Run-time updates**
- Parameters like filtering criteria, bandwidth utilization & polling interval should be updateable at runtime

**Simple to build & Operate**
- Graphical UI & API
- End to end metrics
Hadoop Data Ingestion & Distribution Application

Launch Ingestion

- Use a config file
  If you have saved XML configuration files in this app package, you may launch the app using one of them by enabling this option.
  No configuration files found in this app package.

  Input data source
  ➔ select ➔

  Output data source
  ➔ select ➔

- Compression
  - None
  - LZO
  - GZIP
  This option enables compression of the data before it lands in the target directory, to save space.

- Encryption
  - None
  - AES
  - PKI
  This option enables encryption of the data before it lands in the target directory.

- Bandwidth to use
  MB/sec

Option to set the maximum bandwidth used when copying files from the source directory/directories to the target directory. If not specified, the maximum possible bandwidth will be used.

Launch  Cancel
Data Prep & Analytics Layer Requirements

• Truly scale horizontally across the Hadoop cluster

• Pre-built operators
  ◦ Re-ordering
  ◦ Normalization
  ◦ Transpose
  ◦ De-duplication
  ◦ Tagging
  ◦ Filtering
  ◦ Enrichment

• Operators work seamlessly in both streaming & batch mode
  ◦ Data local HDFS read & process
  ◦ Ability to do computations on time window as well as file boundaries

• Ability to re-use existing business logic

• Simple workflow & scheduling capabilities
  ◦ Built-in or integrations with Oozie or other schedulers
Development API Requirements

- Consistent between Streaming & Batch pipelines
- No mapreduce
- No exposure of low level processing engine concepts
- Easily extendible
Visual Application Assembly

- **Easy to Use**
  - Web based drag-n-drop development environment
  - Automatic port compatibility validation
  - Simple schema management
  - Generic property configurator

- **Easy to Operate**
  - No external component dependency - Runs natively in Hadoop
  - Integrated with DataTorrent management platform

- **Simple to extend**
  - Simple API to enable any existing DataTorrent operator
  - Ability to plug any business logic using a custom operator
Streaming or Batch Data Processing Visualization

- Intuitive user interface
  - Auto-generate or custom create
  - One dashboard for multiple apps
  - Supports bar, line, pie, area charts & data tables

- Easy to Operate
  - No external component dependency - Runs natively in Hadoop
  - Integrated with DataTorrent management platform

- Simple to extend
  - Any DAG operator can be made a real-time datasource
Summary

- World is moving from ‘Batch’ to ‘Streaming’ BUT both are required
- Need a Hadoop native in memory compute engine that is scalable & fault tolerant in BOTH batch & streaming modes
- With out -of-the box data prep & analytics operators
- Using a consistent & functional development API
- Operationalized through a common management & monitoring layer

Project Apex
https://www.datatorrent.com/product/project-apex/

DataTorrent RTS Sandbox
https://www.datatorrent.com/download/

Graphical Application Assembly
Real-Time Data Visualization

Management & Monitoring

Malhar Apache 2.0 license Re-Usable Java Operator Library

Scalable, High Performance, Fault Tolerant In-Memory Data Processing Platform
Apache 2.0 licensed - Project Apex

Hadoop 2.0 — YARN + HDFS

Physical Virtual Cloud

Transform Normalize → Descriptive & Predictive Analytics → Alert Action

Unified Batch & Streaming Data Ingestion & Distribution Pipeline for Hadoop

Unified Batch & Streaming Data Ingestion & Distribution Pipeline for Hadoop
## Some Verticals & Use Cases

<table>
<thead>
<tr>
<th>Ad-Tech</th>
<th>Telco &amp; Cable</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Real-time customer facing dashboards on key performance indicators</td>
<td></td>
</tr>
<tr>
<td>• Click fraud detection</td>
<td>• Call detail record (CDR) &amp; extended data record (XDR) analysis for</td>
</tr>
<tr>
<td>• Billing optimization</td>
<td>• Service quality improvement</td>
</tr>
<tr>
<td></td>
<td>• Capacity planning/optimization</td>
</tr>
<tr>
<td></td>
<td>• Understanding customer behavior AND context</td>
</tr>
<tr>
<td></td>
<td>• Packaging and selling anonymous customer data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Financial Services</th>
<th>IoT</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fraud &amp; risk monitoring</td>
<td>• Process optimization</td>
</tr>
<tr>
<td>• Sentiment based trading strategies</td>
<td></td>
</tr>
<tr>
<td>• Usage based insurance</td>
<td>• Proactive maintenance prediction</td>
</tr>
<tr>
<td>• Improved credit risk assessment</td>
<td>• Remote monitoring &amp; diagnostics</td>
</tr>
<tr>
<td>• Improving turn around time of trade settlement processes</td>
<td></td>
</tr>
</tbody>
</table>