Simplifying Big Data Analytics: Unifying Batch and Stream Processing

John Fanelli,
VP Product
In-Memory Compute Summit
June 30, 2015
FORRESTER

Streaming Analytics

Scale-up Database

Data And Compute Grid

Clustered Database

General-purpose data processing cluster

© 2015 DataTorrent Confidential – Do Not Distribute
DataTorrent enables enterprises to process data in motion and take action in real-time!

Faster Time to Insight
Faster Time to Action

Trend: Batch and streaming use cases
Data Processing Categories in Big Data Use Cases

Questions known?

- Known
  - Batch Processing
  - Stream Processing

- Unknown
  - Ad-hoc Query
  - N/A
Data Sources

Transactional Data
Web Click Stream
Mobile Devices
Operational Log Files
Public Data
Sensor Data
Customer Uses

Real-Time Advertising
Customer Service
Operational
Fraud Detection
Predictive Maintenance
Processing Data In Motion

Ingest Archive → Transform Normalize → Analyze Business Logic → Alert Action → Visualize Persist
Online advertising dynamic inventory purchases

- Ad Server 1
- Ad Server 800
- Fault-Tolerant Flume
- In-memory analytic cube
- Campaign Analysis
- Ad Placement Strategy
- Real-time Dashboard
- Oracle DB

High volume auto-scaling fault tolerant event stream. Dimensional computing to identify performing ads.
Smart Grid provider with many partners has heterogeneous network sources, provides analytics to utilities & customers and provide ISV platform.
Batch Data

Customer Information
Historical Sales
Support Data
Product Configuration
Corporate Info
Batch Processing Data

- Ingest Archive
- Transform Normalize
- Visualize Persist
- Analyze Business Logic
- Alert Action
The Enterprise Data Processing Problem

Ingest Archive ➔ Transform Normalize ➔ Analyze Business Logic ➔ Alert Action ➔ Visualize Persist

ETL

BI & Analytics Platform

ETL

Business Analytics

Complex Event/Event Streaming
The Enterprise Data Processing Problem

Ingest Archive → Transform Normalize → Analyze Business Logic → Alert Action → Visualize Persist

ETL

BI & Analytics Platform

ETL

Business Analytics

Complex Event/ Event Streaming
The Enterprise Data Processing Problem

- Transmission team - Credit, Debit, ACH over Secure FTP
- Transformation team - Parse, Dedup, Transform, Encrypt
- Distribution team - Hadoop, MPP, DB
- Reports team – Dashboards & Alerts & Analytics Platform
- Ingest Archive
- Transform Normalize
- Analyze Business Logic
- Alert Action
- Visualize Persist
- ETL

- Separate applications for each step in the end to end process.
  - 4 to 5 batch jobs to complete the process end to end
  - 1 to 2 runs a day. So typical time to value is around 12 hours

Business Analytics
Complex Event/Event Streaming
Secure, fault tolerant, data ingestion, formatting & archiving. Data access layer for application processing.
Automated, faster time to insight, driving accurate payment, auditing and business planning.
DataTorrent RTS Architectural Overview

Ingestion & Distribution for Hadoop

Graphical Application Assembly

Real-Time Data Visualization

Ingest Archive ➔ Transform Normalize ➔ Analyze Business Logic ➔ Alert Action ➔ Visualize Persist

Re-Usable Java Operator Library

Scalable, High Performance, Fault Tolerant In-Memory Data Processing Platform

Hadoop 2.0 — YARN + HDFS

Physical  Virtual  Cloud

© 2015 DataTorrent Confidential – Do Not Distribute
DataTorrent - Project Apex

- Industry’s first *open source enterprise-class* unified stream and batch processing platform
- DataTorrent RTS 3 Core engine
- Key features requested by open source developer community
  - Event processing guarantees
  - In-memory performance & scalability
  - Fault tolerance and state management
  - Native rolling and static window support
  - Hadoop-native YARN & HDFS implementation
- Apache 2.0 License
- Complemented by open source Malhar operator library

https://www.datatorrent.com/product/project-apex/
DataTorrent enables enterprises to process data in motion and take action in real-time!

Trend: Batch and streaming use cases

Faster Time to Insight
Faster Time to Action