Real-Time Analytics with MicroStrategy

Tony Maresco, Solutions Architect
MicroStrategy World - 2016
Agenda

- Motivation
- Real-time Architecture
- Integration with MicroStrategy
- Demo and Questions
- Enhanced Implementations
- Q&A
Motivation for Real-time Analytics
Motivation

- Streaming analytics are central to the use of BI in front-line operations
  - Big data velocity and volume + real-time analytics to identify opportunities and threats with low-latency
  - Operational Intelligence, environmental, fraud, recommendations, cybersecurity, grid, patient monitoring, risk, financial
- Organizations are using streaming platforms to build their own stream processing applications and want visualization and alerts
  - Using Storm, Spark, Streams, StreamBase, Amazon Kinesis, Hortonworks, Cloudera, MapR, Sybase/SAP Event Stream Processor
- Technology Trends include IoT, M2M with need to analyze machine logs, sensor data and wearables data
  - business, government, healthcare, automobiles, aircraft, power grids, manufacturing, environment, etc.
  - HealthKit, HomeKit, Samsung, Thermostats, SmartThings, Geofencing, Beacons, Environmental sensors
## MicroStrategy’s View of the 3 V’s + Value

<table>
<thead>
<tr>
<th>VOLUME</th>
<th>VARIETY</th>
<th>VELOCITY</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petabytes generated by a surveillance camera&lt;br&gt;Thousands of call records per second</td>
<td>Social Media&lt;br&gt;Clickstream data&lt;br&gt;Web logs&lt;br&gt;Machine data</td>
<td>Data in Motion Time sensitive streaming data</td>
<td>How fast/best the data can be analyzed and acted upon to provide business VALUE?</td>
</tr>
</tbody>
</table>

### VOLUME
- Large Scale of Data ranging in petabytes or more

### VARIETY
- Structured, Semi-structured and Unstructured Data

### VELOCITY
- Data in Motion Time sensitive streaming data
Use Cases

- Healthcare
- Security
- Fraud
- Churn
- Real-time recommendations
- Manufacturing
- Individual customer and Aggregate Sentiment
- Traffic Control
- IT Management
- Trading and Risk analytics
- Finance
- Real-time Telecommunications Operations and Customer Management
Use Cases - continued

- Sentiment analysis
- Cross-sell, up-sell, campaign management
- e-Commerce retail and web behavior and personalization
- Credit card fraud
- Aerospace Monitoring
- Retail store monitoring
- Smart Shelves
- Smart Grid
- Smart Cities
- Pizza and Blackouts
- Log analytics and operational failure prevention and recovery
- Real-time scoring needs to merge with historical context to present understandable options
- Role of streams in this equation
Use Cases - continued

- Patient Monitoring
- Manufacturing
- Energy Trading Systems
- Financial Risk Analyses, Trading, Fraud Detection
- Weather monitoring
- e-Commerce personalization and recommendations
- User behavior analysis
- Ad targeting
- Trending topics
- Recommendations
- Supply Chain Management
- Business Activity Monitoring
Warehouse Analytics vs Streaming Analytics

Traditional Analysis

Which systems failed last month?

Streaming Data Analysis

Which systems are about to fail?
Once upon a time...

- Analysis
- Planning
- Strategy

Human Input

Transactions

Internal Data

MicroStrategy
Now we want it all....and as fast as it changes

- Analysis
- Planning
- Strategy
- Tactical
- Operational
- Automation

- Sensors
- Logs
- Blogs
- Social Media
- Audio
- EMail
- Documents
- Web Pages
- Press Reports
- Machine Logs
- Generated Data
- Video
- Transactions
- Internal Data
- Human Input
- Gov’t Data
- Public Data
- RSS Feeds
State of Streaming Technology

• Streaming Analytics is a fluid and dynamic area of development
  • Hortonworks first embedded Apache Storm and then added Spark
  • Cloudera is embedding Apache Spark and also accommodates Storm
  • MapR first embedded Apache Spark and then added Storm
  • Apache includes projects for Storm, Spark, Kafka, Samsa and Flink

• Products trace to complex event processing, graph and workflow technologies and Continuous Query Language research
  • In-memory window management for operators over streams
  • Apache Storm, Apache Spark Streaming, InfoSphere Streams
  • ESPER, SQLStream, StreamBase, TruViso, SAP Event Stream Processor

• A number of products trace to Complex Event processing in support of DoD, High Frequency Trading and telecommunications

• There are numerous successful custom implementations with common staging areas for the results
  • NoSQL and messaging products such as Cassandra, MongoDB, Kafka, RabbitMQ, Redis
Real-time Architecture
High Level View of Real-Time Processing

- Need Robust Stream collection and processing Platform
  - Guaranteed data processing
  - Fault tolerant and scalable
  - No message broker

- Require Event Correlation Engine
  - Threshold based
  - Filter Based
  - Sequence Based
  - Time Based

- Situation Refinement
  - Detected & Predicted Situations

- Patterns from Historical Data
  - Large scale Storage
  - Parallel processing
  - Reliable
  - Data replication

- Need Big Data Analytics System
Apache Storm

- Spout is a data source providing a continuous stream of data
- Bolt is a computation or processing unit that creates output
- Tuple is a record of data representing the event that is passed between spouts and bolts
- Topology is the computation graph that defines the Storm application. This is based on elements of graph theory
InfoSphere Streams
Lambda Architecture

Variations on mixing real-time and historical data
Real-time Integration with MicroStrategy
Streaming Objectives

• Visualize and alert from external streaming applications
  • Ingest feeds from external engines such as STORM, Spark Streaming, Streams, Amazon Kinesis, Manufacturing systems
  • Streaming updates of in-memory cubes
  • Real time widgets for web and mobile
  • Extend streaming analytic results with MicroStrategy analytics and platform features
  • Trigger events in real-time to trigger distribution services in MicroStrategy
  • Subscribe to queue sources

• Deliver connectors to sources with real-time features
  • For example Parstream, Oracle in-memory, Spark, Cassandra, JDBC, URL

• Some objectives require or are enhanced with custom SDK work
Real-time applications cannot batch-ETL data from an operational source to a data warehouse
- It takes too long

Streaming technologies like Apache Storm, Apache Spark or InfoSphere Streams process data as it flows through the system
- Instead of storing to a database
Typical Components for Real-time Implementations

- **Streaming Infrastructure**
  - **Real-time Data Streams**
  - **Stream Application**
  - **Staging**
  - **Visualization And Alerts**

- **Streaming Components**
  - Storm
  - Spark Streaming
  - Hive
  - HBase
  - Cassandra
  - Kafka
  - Redis
  - Flume
  - Elastic Search
  - SOLR
  - LogStash

- **Pull Connectors**
  - ODBC
  - JDBC
  - Web Services
Demo
1. Auto Refresh in Web and Mobile
2. Web Service/JDBC/ODBC accessed by Free-form report
3. Staged results data
4. Stream processing application
5. Read events from a queue or directly from an API
6. Read logs and other sources with logstash or Flume
Enhanced Implementations
Push Customization Using Visualization SDK

Streaming Infrastructure → Stage/API → Node.js Push

Admin App * → Add/Delete Start/Stop Monitor

Node.js Push * → Node.js/Socket.io Filtering Server

Static Cubes for Design

Historical Cubes for RT Blending, Analysis, and Alerts

Design Mode*

Run-time Mode*

MicroStrategy Analytics and Visualization
Summary

- New data sources abound at high volume, high speed and varied structure by virtue of the internet and an instrumented world.
- Applications which leverage these new sources often require acting on opportunities and threats ASAP.
- Many of these real-time applications require analysis of historical data to capture rules and models to apply in real-time.
- This historical/real-time process forms a cycle.
- The architecture to meet these requirements combines traditional warehouse analysis with Big Data batch and real-time streaming analytics.
- MicroStrategy gives you the ability to integrate streaming technologies for visualization and alerts.
References

- http://storm.apache.org/
- http://spark.apache.org/streaming/
- http://xinhstechblog.blogspot.com/2014/01/scaling-realtime-analytics-on-big-data.html
- https://www.mapr.com/developercentral/lambda-architecture
- https://education.emc.com/content/_common/docs/ks_articles/2013K_S_Bhattacharya_Mitra-Analytics_on_Big_FAST_Data.pdf
Reading List

**Learning Real-time Processing with Spark Streaming**
*By: Sumit Gupta*
*Publisher: Packt Publishing*
*Pub. Date: September 28, 2015*

**Real-time Analytics with Storm and Cassandra**
*By: Shilpi Saxena*
*Publisher: Packt Publishing*
*Pub. Date: March 27, 2015*
*Print ISBN-13: 978-1-78439-549-0*

**Storm Applied: Strategies for real-time event processing**
*By: Sean T. Allen, Matthew Jankowski, and Peter Pathirana*
*Publisher: Manning Publications*
*Pub. Date: March 31, 2015*
*Print ISBN-13: 978-1-61729-189-0*
*Print ISBN-10: 1-61729-189-7*

**Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data**
*By: Byron Ellis*
*Publisher: John Wiley & Sons*
*Pub. Date: July 21, 2014*
*ISBN-13: 978-1-118-83802-0*
*Web ISBN-10: 1-118-83791-6*
Reading List

Big Data: Principles and best practices of scalable realtime data systems
By: Nathan Marz with James Warren
Publisher: Manning Publications
Pub. Date: April 30, 2015