Stream Processing: Streaming Data in Real Time, In Memory

Fern Halper
TDWI Director of Research for Advanced Analytics
@fhalper

August 12, 2015
Sponsors

intel®

SAP®
Speakers

Fern Halper
Research Director, Advanced Analytics, TDWI

Neil McGovern
Senior Director, Marketing, SAP

Jaan Leemet
Senior VP, Advanced Technology, Tangoe, Inc.
Agenda

• Stream processing overview
• Tango case study
• SAP stream processing introduction
• Round table discussion
• Audience Q&A
Data streams - Definition

• Data at rest vs. data in motion
• Data that arrives continuously as a sequence of instances
  – Sensor data, social media, traffic feeds, financial transactions
• Often, it needs to be processed immediately
Analyzing Data in Motion

Event processing

Stream mining

Complex event processing
Analyzing data streams

• Can involve summarizations of the stream
  – *Often uses a fixed length “window”*
  – *Can involve filtering*

• Historical sources can be involved
  – *Build models*

• Advanced analysis in real true real time
  – *Evolving to include advanced statistics*
Use cases for streams (1)

- Healthcare Monitoring
  - Patient Sensors
  - Analytics
  - Historical Data
Use cases for streams (2)

• Energy
  – *Environmental sensors*
  – *Preventive maintenance*
ETM Data Sources

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Now</th>
<th>3 years from now</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-time event streams</td>
<td>23%</td>
<td>36%</td>
</tr>
<tr>
<td>IoT</td>
<td>18%</td>
<td>39%</td>
</tr>
<tr>
<td>Machine data</td>
<td>27%</td>
<td>33%</td>
</tr>
</tbody>
</table>

(TDWI 2015, n=332)
Stream processing and analytics: status

ETM Processing and Analytics

- Using today:
  - Stream processing: 23%
  - Stream mining: 11%

- Plan to use in 3 years:
  - Stream processing: 37%
  - Stream mining: 36%

(TDWI 2015, n=332)
Transactional analysis of data streams with application to cost allocation

Jaan Leemet
SVP Adv Technology
August 2015
Who is Tangoe
Powering the Connected Enterprise

A leading global provider of Connection Lifecycle Management (CLM) software and related services

• Over $29.9Bn spend managed ($7.5Bn int'l)
• 1.2M invoices processed/month
• 72k fulfillment orders processed /month
• 3,100 carriers/1,800 bill formats
• 72k payments/month in 45+ currencies
• 92 invoice receipt centers
• 24 x 7 x 365 global support in 18 languages
• 37k mobile end-user contacts per month
• 165 call center agents in 8 global support centers

• Integrated processing centers
• Integrated wireless/fixed systems
• Integrated service delivery teams
• Integrated translation tools
• Regulatory support – 63 agencies
• 198 countries / territories served
• Application support for 12 languages
• Global SSAE16 certification
• Safe Harbor Certification
## Diversified Customer Base

Powering the Connected Enterprise

<table>
<thead>
<tr>
<th>Banking &amp; Financial</th>
<th>Technology</th>
<th>Systems Integrators &amp; Prof. Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank of America</td>
<td>comcast.</td>
<td>KPMG</td>
</tr>
<tr>
<td>Credit Suisse</td>
<td>DELL</td>
<td>xerox</td>
</tr>
<tr>
<td>Capital One</td>
<td>CITRIX</td>
<td>HP</td>
</tr>
<tr>
<td>RAYMOND JAMES</td>
<td>PHILIPS</td>
<td>Accenture</td>
</tr>
<tr>
<td>VISA</td>
<td>Aol.</td>
<td>BT</td>
</tr>
<tr>
<td>Deutsche Bank</td>
<td>Agilent Technologies</td>
<td></td>
</tr>
<tr>
<td>UBS</td>
<td>ORACLE</td>
<td>IBM</td>
</tr>
<tr>
<td>WELLS FARGO</td>
<td>MOTOROLA</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Healthcare</th>
<th>Other</th>
<th>Consumer &amp; Transportation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermountain</td>
<td>FirstEnergy</td>
<td>FedEx</td>
</tr>
<tr>
<td>Healthcare</td>
<td>tyco</td>
<td>BEST BUY</td>
</tr>
<tr>
<td>EXPRESS SCRIPTS</td>
<td>Constellation Energy</td>
<td>DELTA</td>
</tr>
<tr>
<td>CVS/pharmacy</td>
<td>ADP</td>
<td>AMR.</td>
</tr>
<tr>
<td>MCKESSON</td>
<td>SIEMENS</td>
<td>PEPSICO</td>
</tr>
<tr>
<td>Empowering Healthcare</td>
<td>Chemtura</td>
<td>Hertz.</td>
</tr>
<tr>
<td>Abbott</td>
<td>HALLIBURTON</td>
<td>Kraft Foods</td>
</tr>
<tr>
<td>Novartis</td>
<td></td>
<td>Johnson &amp; Johnson</td>
</tr>
<tr>
<td>DELTA DENTAL</td>
<td></td>
<td>GAP</td>
</tr>
<tr>
<td>novo nordisk</td>
<td></td>
<td>Bravo</td>
</tr>
</tbody>
</table>
The Application
Powering the Connected Enterprise

• A system with which we can analyze data traffic and extract transaction types for content analysis and allocation

• Not just per-application but rather at a transactional level where by transaction in a time window can be allocated separately.

• Data traffic can consist of a TCP/IP data stream, a flow of CDR records, and IoT feed of data, an input stream of invoice data from a carrier.

• Applications include: split billing for personal vs. business usage, cost allocation to departments, productivity analysis, security/threat detection, etc.
The Problem / Challenge
Powering the Connected Enterprise

• Lack of Native ‘on device’ methods to capture detailed usage
• Large volume of data flowing through the system
• Data interspersed and interwoven …
• Only interested in limited view of data ‘types’ not necessarily data content or all of the data
• Desire to act quickly when certain types of transactions are recognized
• Need for flexible and rapid methods to add and modify recognition algorithms and routines
• Ideally easy to integrate with existing subsystems used for Expense Management, Asset Management, Mobility Management
The Approach

Powering the Connected Enterprise

- Pull together access to data stream from both LAN and Mobile environments.
- Leverage feeds to access standard TCP/IP data streams
- Run continuous query type stream processor to extract patterns and usage
- Build up pattern recognition algorithms and heuristics
- Incorporate decryption and/or feeds from external syslog type resources, partner solutions
- Generate and Merge together multiple streams (with different schemas) if different input adapters (configurations) required
The Results and Observations

Powering the Connected Enterprise

- After evaluating numerous approaches ended up with SAP’s Event Stream Processor (ESP - now SAP HANA smart data streaming) as primary stream processing engine
- Complements our other SAP-based solutions such as HANA and BI tool sets
- Works with adaptor analogy: implement logic in SQL-like (CCL) language that can pull most of heavy lifting from outside the adaptor
- Benchmarking suggests this will allow us to do categorization in s/w where a lot of existing approaches use hardware assist.
- Like the flexibility of adaptor stream processing as it allows us to consume data sources for voice CDRs, data streams, invoice data ..
- Others may require audit capability demanding archival of the complete data (example a Hadoop lake). In our case we only consume and keep what we need trashing the rest, but have the option of capturing what/where we want
Tangoe (NASDAQ: TNGO) is a leading global provider of Connection Lifecycle Management software and services to a wide range of global enterprises and service providers.

The company's Connection Lifecycle Management technology, Matrix, is an on-demand suite of software and services designed to turn on, manage, secure, and support various connections in an enterprise's communications lifecycle, including mobile, fixed, machine, cloud, social, and IT.
Event stream processing uses continuous queries

Database Queries

Step 1: Store the data
Step 2: Query the data

Continuous Queries

Step 1: Define the continuous queries and the dataflow
Step 2: Wait for data to arrive. As it arrives, it flows through the continuous queries to produce immediate results
Streaming data sources are everywhere

- Sensors
- Click streams
- Social media
- Transactions
- Market prices
Extends the capabilities of the SAP HANA Platform with the addition of **real-time event stream processing**

1. **Capture** data arriving continuously from devices and applications

2. **Act** on new information as soon as it arrives: alerts, notifications and immediate response to changing conditions

3. **Stream** information to live operational dashboards

**Highly scalable** – process hundreds of thousands or even millions of events per second
Design time tools in SAP HANA Studio

Streaming plug-in for SAP HANA Studio includes a visual editor for defining continuous queries and directing stream flow, plus run/test tools.

- Event processing language: CCL (like SQL for streams)
- CCL Script (like event driven stored procedures for streams)
## Example adapters that are currently available

### Standard Adapters

(included with SDS/ESP licenses; both input and output unless otherwise noted)

- Message bus: JMS, IBM MQ, TIBCO
- Web Service: SOAP, REST
- Databases
- Files
- Sockets
- SAP RFC
- SAP Sybase Replication Server (in)
- Logfile (in)
- Microsoft Excel (out)
- Email (out)
- HTTP snapshot query (out)

### Optional Adapters

(licensed separately)

- NYSE Technologies
- FIX

### Parsing/Formatting

- JSON
- XML
- CSV
- FIX
- JMS Object Arrays
Roundtable Discussion
Speakers

Fern Halper
Research Director,
Advanced Analytics,
TDWI

Neil McGovern
Senior Director,
Marketing
SAP

Jaan Leemet
Senior VP,
Advanced Technology,
Tangoe, Inc.
Question 1

• What’s the value in combining multiple data sources? How do you do that?
Question 2

Is there such a thing as too much data in stream processing? How do you deal with that?
Question 3

Is it possible to have a real-time application without streaming analytics?
Question 4

Why and how does real-time streaming analytics impact business processes?
Audience Questions?
Contact Information, guest panelists

- If you have further questions or comments:
  
  Fern Halper, TDWI  
  fhalper@tdwi.org  
  Neil McGovern, SAP  
  neil.mcgovern@sap.com  
  Jaan Leemet, Tango  
  jaan.leemet@tango.com