Pervasive & KNIME
- data mining for the big data age

David Inbar
Snr Director, Business Development & Strategy
Pervasive Big Data & Analytics

Jim Falgout
Chief Technologist
Pervasive Big Data & Analytics
Agenda

• Quick intro –
  – Pervasive and DataRush
  – The big data challenge

• Fast, faster, fastest – the metrics

• Demos, use cases

• Integration with KNIME - flowable nodes, DataRush nodes, streaming API

• Q&A

• If there is time….. – scripting/extensions, parallelism within DataRush
Pervasive Software Overview

- 250 employees
- HQ in Austin, TX; offices - London, Frankfurt, Tokyo
- $50M annual sales (66% USA, 34% Europe/Asia)
- 25% of revenue re-invested in R&D

### Core Businesses

<table>
<thead>
<tr>
<th>Database</th>
<th>Integration</th>
<th>Business Xchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database for packaged business applications</td>
<td>Universal connectivity for cloud and on-premise applications</td>
<td>Cloud-based B2B Document Exchange Service</td>
</tr>
</tbody>
</table>

- PSQL Vx Server
- PSQL v11
- AuditMaster
- Backup Agent
-DataExchange
- Data Integrator
- Data Integrator Cloud
- Data Profiler
- Data Match Merge
- BXSupplierConnect
- BXOperatorConnect
- BXPortalConnect

### Innovation Businesses

<table>
<thead>
<tr>
<th>DataCloud</th>
<th>Big Data &amp; Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secure and reliable, Cloud Services platform (PaaS) on AWS</td>
<td>High-performance software for data-intensive processing and analytics</td>
</tr>
</tbody>
</table>

- Pervasive DataCloud
- Pervasive Galaxy Marketplace and Support Community
- RushAnalytics
- RushAccelerator for KNIME
- Pervasive DataRush
- Data Integrator – Hadoop Edition
Challenge of Scaling Analytics

1. Data is plentiful (but mostly under-used)
2. Processing times for tough/useful analytics are too long – both in the discovery phase and in production/deployment
3. Too expensive to scale with current state-of-the-art

Pervasive’s ‘Rush’ Solution:

Software that automatically leverages available hardware resources for extreme speed, across every environment.

Use software to unlock the power of your hardware.
## You Say It’s Fast, But How Fast?

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Task</th>
<th>TB/Hour*</th>
<th>Recs/Sec*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyber Security Threat Detection</td>
<td>Simple query to analyze log files</td>
<td>N/A</td>
<td>&gt;40 million</td>
</tr>
<tr>
<td>Network Optimization</td>
<td>Ingest, analyze, parse and persist Netflow data</td>
<td>N/A</td>
<td>3 million</td>
</tr>
<tr>
<td>Business Intelligence</td>
<td>Complex table scan &amp; aggregate query on TCP-H data</td>
<td>3</td>
<td>&gt;30 million</td>
</tr>
<tr>
<td>Analytics Data Prep</td>
<td>Bulk load weblog data into HBase</td>
<td>4.7</td>
<td>4 million</td>
</tr>
<tr>
<td>Benchmark</td>
<td>Analyze weblog data</td>
<td>4</td>
<td>&gt;10 million</td>
</tr>
</tbody>
</table>

*All on less than $100K of commodity hardware – a 5-node Hadoop cluster*
Performance Test

- DataRush versus PIG
  - Used TPC-H data
  - Generated 1TB data set in HDFS
  - Ran several "queries" coded in DataRush and PIG
  - Run times in seconds (smaller is better)

Cluster Configuration:
- 5 worker nodes
- 2 X Intel E5-2650 (8 core)
- 64GB RAM
- 24 X 1TB SATA 7200 rpm

TPC-H : 1 Terabyte Test : Run times
Demo
Architecture Overview

KNIME UI + parallel dataflow engine + application development framework for big data prep and analytics

Large volumes of data

High speed parallel access
- Flat files
- SQL databases
- NoSQL databases
- HDFS, HBase
- PMML
- SAS data files
- AWS S3, RedShift
- Google BigQuery

Quality data
Actionable analytics

Auto-scaling
- Vertically—take advantage of all cores
- Horizontally—take advantage of every node
When/Where Would I Use It?

- When speed is critical—for both model development and execution
- When R, SAS, SPSS, etc. are too slow, too expensive, or cannot scale
- For data mining with big and complex data
- For solutions that must run in multiple environments from laptops to big Hadoop clusters
- When you need solutions that use less hardware
DataRush/RushAnalytics Solutions

- Opera Solutions
  - Data science solutions provider
  - Embedding DataRush in engineered solutions
- Healthcare
  - Claims cleansing & processing
- Retail
  - Market basket analysis
  - Product category resolution (MDM)
- Telecom
  - CDR processing & analysis

"Pervasive DataRush’s efficiency and ability to automatically scale, whether on a single server or a Hadoop cluster, supports our vision for consistent, reusable, scalable Big Data analytics."

- Chief Operating Officer, Opera Solutions
Questions

david.inbar@pervasive.com
jim.falgout@pervasive.com