Real-time Insights and Analytics on Big Data

Archana Ganapathi
Research Engineer
Big Data Comes from Machines

Volume | Velocity | Variety | Variability

Machine-generated data is one of the fastest growing, most complex and most valuable segments of big data.

GPS, RFID, Hypervisor, Web Servers, Email, Messaging, Clickstreams, Mobile, Telephony, IVR, Databases, Sensors, Telematics, Storage, Servers, Security Devices, Desktops
Machine Data Contains a Goldmine of Insights

<table>
<thead>
<tr>
<th>Sources</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order Processing</td>
<td>Customer ID</td>
</tr>
<tr>
<td>Middleware Error</td>
<td>Order ID</td>
</tr>
<tr>
<td>Care IVR</td>
<td>Product ID</td>
</tr>
<tr>
<td>Twitter</td>
<td>Time Waiting On Hold</td>
</tr>
</tbody>
</table>

May 21 14:04:12.996 wl-01.acme.com Order 569281734 failed for customer 10098213
Exception follows: weblogic.jdbc.extensions.ConnectionDeadSQLException:
weblogic.common.resourcepool.ResourceDeadException Could not create pool.
The DBMS driver exception was: [BEA][Oracle JDBC Driver]Error establishing socket to host and port: ACMEDB-01:1521. Reason: Connection refused

ORDER,2012-05-21T14:04:12.484,10098213,569281734,67.17.10.12,43CD1A78322,SA-2100

05/21 16:33:11.238 [CONNEVENT] Ext 1207130 (0192033): Event 20111, CTI Num: ServID:Type 98#1, DNIS 5555685981, SerID 40489a07-7f6e-4251-801a-13ae51a6d092, trunk 7451.16
05/21 16:33:11.242 [SCREENPOEVENT] SerID 40489a07-7f6e-4251-801a-13ae51a6d092
CUSTID 10098213
05/21 16:37:49.732 [DISEVENT] SerID 40489a07-7f6e-4251-801a-13ae51a6d092

{actor:{displayName:"Go Boys!!",followersCount:1366,friendsCount:789,link:"http://dallascowboys.com/",location:{displayName:"Dallas, TX",objectType:"person",preferredUsername:"B0ysF@n80",statusesCount:6072},body:"Just bought this POS device from @ACME. Doesn't work! Called, gave up on waiting for them to answer! RT if you hate @ACME!"},objectType:"activity",postedTime:"2012-05-21T16:39:40.647-0600"}
Splunk as a Data Platform

Real-time Machine Data

Splunk storage

Custom dashboards
Report and analyze
Developer Platform

GPS, RFID, Hypervisor, Web Servers, Email, Messaging, Clickstreams, Mobile, Telephony, IVR, Databases

Sensors, Telematics, Storage, Servers, Security devices, Desktops, CDRs

Ad hoc search
Monitor and alert

Splunk as a Data Platform

Report and analyze
Developer Platform

Splunk storage

Custom dashboards

Ad hoc search
Monitor and alert

Splunk as a Data Platform
Deep Visibility and Insight for IT and Business

Over 5,600 organizations using Splunk across IT and business users
Under the hood...
Analyzing Heterogeneous Data

Universal Index
- No data normalization
- Automatically handles timestamps
- Parsers not required
- Index every term & pattern “blindly”
- No attempt to “understand” up front

Schema-on-the-fly
- Structure applied at search-time
- No brittle schema to work around
- Automatically find transactions, patterns and trends

Flexibility and Fast Time to Value
- Normalization as it’s needed
- Faster implementation
- Easy search language
- Multiple views into the same data
Real-time Analytics

Data Parsing

Source, event typing
Character set normalization
Line breaking
Timestamp identification
Regex transforms

Indexing

Real-time Search Process

Splunk Index
Splunk’s Search Language

command1 | command2 | command3

Filter Transform Enrich

Filter Transform Enrich

Filter Transform Enrich

search error | eval KB=bytes/1024.0 | stats avg(KB) by domain
Data Munging

Example: combine two different sources that have a different format for the score field.
Source X scores are based on a 0-100 scale, and source Y uses letter grades and needs to be changed to a 0-100 scale.

search source = X OR source = Y
    | eval score = if(source = ”X”, score, case(score = ”A”, 100, score = ”B”, 85, score = ”C”, 70, score = ”D”, 60, score = ”F”, 0))
Outlier Detection

Example: find scores more than 3 standard deviations more or less than the average.

search score = *
  | stats avg(score) as avg stdev(score) as stdev
  | where (score > avg + 3 * stdev) or (score < avg – 3*stdev)
Correlation

Example: find how correlated score and income are.

Search score = * income = *
  | stats avg(eval(score * income)) as avg_prod
  | avg(score) as avg_score
  | avg(income) as avg_income
  | eval cov = avg_prod − avg_score * avg_income
Clustering

Example: group people by score, income, age using kmeans.

search score = * income = * age = *
| kmeans k = 10 score, income, age
Deployment Architecture
Distributed Search
Splunk Big Data Integration

- **Ad hoc search**
- **Monitor and alert**
- **Report and analyze**
- **Custom dashboards**
- **Developer Platform**
  - API and SDKs to build Big Data apps

**Splunk Dev Platform**

**Splunk DB Connect**
- Real-time integration to relational DBs

**Hunk and Splunk Hadoop Connect**
- Reliable bi-directional integration to Hadoop
Turning big data into powerful insights
Analyze Insights as They Occur

Correlate CDRs with tariffs
Identify lowest cost routes

Real-time visibility
Heaviest users and abusers

Monitor cell towers
Detect major catastrophes
Understanding Customer Behavior

Content browsed, purchased and watched
All tracked by time and MAC address

Customer profile and MAC address/device assignments

Customer behavior analytics
Measuring User Experience on a Wide Scale

Web User
Clickstreams
12 million
monthly visits

Weblog
Traffic Data
750 million
queries per month

Maintain
high performance

Protect
content against malicious bots

Track
traffic sources for advertisers
Splunk Delivers Big Data in Days or Weeks

**Product-based Solution**
- Easy to download and deploy
- Pre-integrated, end-to-end functionality
- Enterprise-grade features

**Real-time Platform**
- Collects data from tens of thousands of sources
- Advanced real-time and historical analysis of data
- Interactive visualizations for IT and business users

**Performance at scale**
- Proven at multi-terabyte scale per day
- Upwards of PB under management
- Thousands of enterprise customers
Thank You

Splunk.com/bigdata