Splunk for Ad Hoc Exploration of Twitter (and more)

Stephen Sorkin
VP Engineering, Splunk
Who am I

- Berkeley PhD dropout.
- Left to work at HP Labs.
- At Splunk since 2005.
- VP Engineering since 2010.
- Run the core product team.
Agenda

• Inspiration for Splunk
• Architecture:
  – Collection
  – Indexing
  – Search
  – Real-time Search
• Demo
What Does Machine Data Look Like?

ORDER, 2012-05-21T14:04:12.484, 10098213, 569281734, 67.17.10.12, 43CD1A7B8322, SA-2100


05/21 16:33:11.238 [CONNEVENT] Ext 1207130 (0192033): Event 20111, CTI Num: ServID: Type 0:19:9, App 0, ANI T7998#1, DNIS 5555685981, SerID 40489a07-7f6e-4251-801a-13ae51a6d092, Trunk T451.16
05/21 16:33:11.242 [SCREENPOPEVENT] SerID 40489a07-7f6e-4251-801a-13ae51a6d092 CUSTID 10098213
05/21 16:37:49.732 [DISCEVENT] SerID 40489a07-7f6e-4251-801a-13ae51a6d092

{actor:{displayName:"Go Boys!!",followersCount:1366,friendsCount:789,link: "http://dallascowboys.com/",location:{displayName:"Dallas, TX",objectType:"place"},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"}
**Machine Data Contains Critical Insights**

<table>
<thead>
<tr>
<th>Sources</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order Processing</td>
<td>believed data from various sources is critical for insights.</td>
</tr>
<tr>
<td>Middleware Error</td>
<td>Exception: 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.</td>
</tr>
<tr>
<td>Care IVR</td>
<td>Time Waiting On Hold: 05/21 16:33:11.238 [CONNEVENT] Ext 1207130 (0192633): Event 20111, CTI Num: ServID: Type D99#1, DNIS 5555685981, SerID 40489a07-7f6e-4251-801a-13ae516d092, Trunk 1451.16</td>
</tr>
<tr>
<td>Twitter</td>
<td>Customer’s Tweet: {actor: {displayName: &quot;Go Boys!!&quot;, followersCount: 1366, friendsCount: 789, link: &quot;<a href="http://dallascowboys.com%22%7D">http://dallascowboys.com&quot;}</a>, location: {displayLocation: &quot;Dallas, TX&quot;, objectType: &quot;person&quot;}, preferredUsername: &quot;B0ysF@n80&quot;, statusesCount: 6072}, body: &quot;Just bought this POS device from @ACME. Doesn’t work! Called, gave up on waiting for them to answer! RT if you hate @ACME!!&quot;, objectType: &quot;activity&quot;, postedTime: &quot;2012-05-21T16:39:40.647-0600&quot;}</td>
</tr>
</tbody>
</table>

- **Customer ID**: 10098213
- **Order ID**: 569281734
- **Product ID**: SA-2100
Splunk Enterprise with Hadoop

GPS, RFID, Hypervisor, Web Servers, Email, Messaging, Clickstreams, Mobile, Telephony, IVR, Databases, Telematics, Storage, Servers, Security devices, Desktops, CDRs, Applications

Ad hoc search
Add knowledge
Monitor and alert
Custom dashboards
Report and analyze

Splunk storage
Other Data Stores

Splunk Enterprise with Hadoop

GPS, RFID, Hypervisor, Web Servers, Email, Messaging, Clickstreams, Mobile, Telephony, IVR, Databases, Telematics, Storage, Servers, Security devices, Desktops, CDRs, Applications
Getting Data into Splunk
Agent and Agent-less Approach for Flexibility

- syslog
- TCP/UDP

Mounted File Systems
/\hostname\mount

Unix, Linux and Windows hosts

syslog compatible hosts and network devices

WMI
Event Logs Performance

Windows hosts

Active Director

Custom apps and scripted API connections

Virtual host

Scripted Inputs
- shell scripts
- custom parsers
- batch loading

Windows Inputs
- Event Logs
- performance counters
- registry monitoring
- Active Directory monitoring

Local File Monitoring
- log files
- config files
- dumps and trace files

Agent-less Data Input

Splunk Forwarder
Bucket Lifecycle

Events

- Hot Bucket is Full
- Too Many Warms
- Out of Space or Bucket is Old
- Explicit User Action

Storage Paths:

- $Home Path
- $Cold Path
- $Thawed Path
- $Frozen Path or Deleted

- Cheaper Storage
- Thawed Path
- Frozen Path

Hot Bucket

Warm Bucket

Cold Bucket

Frozen Bucket
Scales to TBs/day and Thousands of Users

- Automatic load balancing linearly scales indexing
- Distributed search and MapReduce linearly scales search and reporting

Offload search load to Splunk Search Heads

Auto load-balanced forwarding to as many Splunk Indexers as you need to index terabytes/day

Send data from 1000s of servers using any combination of Splunk Forwarders, syslog, WMI, message queues, or other remote protocols
Search Model

- Splunk Database as a table:
  - Columns = fields, rows
  - No fixed schema
  - Unlimited number of rows, can be very sparse
  - Special fields: `_raw`, `_time`, `host`, `source`, `sourcetype`

- search: series of commands with arguments
  - implicit search command usually first
  - Input/output of every command is a table
Search Model Example

sourcetype=syslog ERROR | top user | fields - percent

1. Fetch events from disk that match
2. Summarize into table of top 10 users
3. Remove column showing percentage
Search Command

- **Expand Search:** lookups, tags, savedsearch, eventtypes, etc.
- **LISPY Expression** (per index)
- **DB**

Flow diagram:
- **Lookup**
- **Calculated fields (5.0+)**
- **Field aliasing**
- **Field extractions**
- **sourcetype renaming**

- **Filter**
- **Apply eventtypes**
- **Apply tags**
Inside Universal Indexing

Automatic event boundary identification

Automatic timestamp normalization

...enable accurate searching and trending by time across all data:
Inside Search-time Knowledge Extraction

... enable statistics and precise search on specific fields:

```
misc activity NOT reply | stats avg(IpLen) by dest_ip
```

Automatically discovered fields

And user-defined fields
Inside Search-time Knowledge Extraction

Searches saved as event types

Plus tagging of event types, hosts and other fields

... enable normalized reporting, knowledge sharing and granular access control.
Integrate External Data
Extend analysis with lookups to external data sources

- LDAP, AD
- CMDB
- CRM/ERP
- Watch Lists

Correlate IP addresses with locations, accounts with regions

Sample data:

```
<TRANSACTION date="2010-01-04 15:57:18,307" activityCode="1010" sequenceNumber="100233683" callerID="MAR10043LA" transactionStatus="COMPLETE" result="SUCCESS" host="10.34.51.95"

accountNumber="COT9198616006"
domain=Atlanta
market=Raleigh-Durham

<TRANSACTION date="2010-01-04 15:57:17,194" activityCode="1010" sequenceNumber="106263965" callerID="MAR10619LA" transactionStatus="COMPLETE" result="SUCCESS" host="10.52.60.102"

accountNumber="COT4158884945"
domain=Bothell
market=San Francisco
```
Distributed Searching

1. POST to /services/search/jobs on search head

2. Search head spawns search in a separate process

3. Send ‘remote’ version of search to each search peers via /services/streams/search

4. Each search peer spawns another search process to run remote search

5. Read data from indexes 5b. For realtime, connect back to splunkd
Real-time Alerting

source="/var/log/secure.log"  "BAD SU"
Demo

- http://socialsplunk.com/
- http://socialsplunk.com:8081/map
- https://splunk4good-rtv.s3.amazonaws.com/rtv.png
The 2012 Election

source="twitter_httpstream" romney OR obama
| eval text=lower(body) | fields text | rex field=text max_match=1000 "(?<token>[@a-zA-Z]{5,})"
| eval token=mvfilter(NOT match(token, "@.*"))
| `clean_tweets`
| eval candidate=if(searchmatch("*obama* AND *romney*"), "obama:romney", if(searchmatch("*romney*"), "romney", if(searchmatch("*obama*"), "obama", null)))
| where NOT isnull(candidate)
| makemv delim=":" candidate
| top token by candidate limit=50