Real World Big Data Architecture - Splunk, Hadoop, RDBMS

Raanan Dagan, Big Data Specialist, Splunk
Disclaimer

During the course of this presentation, we may make forward looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not, be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.
About Your Presenter

• Raanan Dagan - Sr. SE, Big Data specialist
• In Splunk for 3.5 Years
• Before Splunk worked for Cloudera and Oracle
Agenda

• Splunk Big Data Architecture
• Alternative Open Source Approach
• Real-World Customer Architecture
• Customer Roadmap Architecture
• End-to-end Demonstration (Splunk, Hadoop, RDBMS)
Big Data Technologies

Relational Database
- Structured
  - Schema at Write
  - SQL
  - ETL
- RDBMS
  - Oracle, MySQL, IBM DB2, Teradata

NoSQL
- Semi-Structured
  - Key-Value, Column, Document & Other Stores
  - Schema at Read
- Cassandra, Accumulo, MongoDB

Hadoop
- Semi-Structured
  - Schema at Read
  - MapReduce
  - HDFS Storage

Splunk
- Schema at Read
- Search
- Real-Time Indexing
- Time-Series, Unstructured, Heterogenous

Distributed File System
- MapReduce
Are you getting the most out of your Big Data investments?
Big Data in Government

How can you accelerate your time to value?
Splunk Big Data Technologies

**DB Connect**
- Schema at Write
- SQL
- ETL

**Hunk**
- Schema at Read
- Key-Value, Column, Document & Other Stores
- MapReduce
- HDFS Storage

**Splunk**
- Schema at Read
- Search
- Real-Time Indexing

**Storage Technologies**
- RDBMS: Oracle, MySQL, IBM DB2, Teradata
- Cassandra, Accumulo, MongoDB
- Distributed File System
- Time-Series, Unstructured, Heterogenous
Splunk Scalability

Enterprise-class Availability and Scale

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

Send data from thousands of servers using any combination of Splunk forwarders

Offload search load to Splunk Search Heads

Auto load-balanced forwarding to Splunk Indexers
Splunk Real-Time Analytics

### Data

- **Monitor Input**
- **TCP/UDP Input**
- **Scripted Input**

### Parsing Pipeline
- Source, event typing
- Character set normalization
- Line breaking
- Timestamp identification
- Regex transforms

### Index Queue

### Indexing Pipeline

### Real-time Buffer

### Real-time Search Process

### Splunk Index
Hunk - Analytics Platform for Hadoop

Full-featured, Integrated Product

Insights for Everyone

Works with What You Have Today

Explore  Analyze  Visualize  Dashboard  Share

Hunk

Hadoop Client Libraries

Hadoop Clusters

NoSQL, EMR, S3 Buckets
Hunk Unique Features

Virtual Index
• Enables seamless use of the Splunk technology stack on data wherever it rests
• Natively handles MapReduce

Schema-on-the-fly
• Structure applied at search time
• No brittle schema
• Automatically find patterns and trends

Flexibility and Fast Time to Value
• Interactive search
• Preview results while MapReduce jobs run
• Drag-and-drop analytics

Security: Access Control, Pass Through Authentication, Kerberos
Hunk Provides Self-Service Analytics for Hadoop

Hunk = Indexing + Data Preview + Security + Great UI
What About Structured Data?

Customer profile  Product attributes  Employee details  Pricing and Rate plans  Asset info
Use cases for structured data in Splunk

- Index machine data from databases, such as logs or sales records
- Enrich machine data with high-level data, such as customer records
- Update structured databases with Splunk info, such as risk scores
- Interactively browse structured and unstructured data from Splunk reports
Machine Data – Delivers Real-time Insights
Structured Data – Contains Business Context

Media server logs (machine data)

Customer, product databases

Phone number

IP address

Track ID

Mar 01 19:18:50 Mar 01 19:18:50 Mar 01 19:18:50
aaa2 radiusd[12548]:[ID 959576 locall.info] INFO RADOP(13) acct start for
2172618992@splunktel.com 10.164.232.181 from 12.130.60.5 recorded OK.
2013-03-01 19:18:50:150 10.2.1.34 GET /sync/addtolibrary 01011207201000005652000000000053 - 80 -
10.164.232.181 “Mozilla/5.0 (iPhone; CPU iPhone OS 5.1 like Mac OS x) AppleWebKit/534.46 (KHTML,
like Gecko) Version/5.1 Mobile/9A405 Safari/7534.48.3” 503 0 0 825 1680
Mar 01 19:18:50:163 aaa2 radiusd[12548]:[ID 959576 locall.info] INFO RADOP(13) acct stop for
2172618992@splunktel.com 10.164.232.181 from 12.130.60.5 recorded OK.

<table>
<thead>
<tr>
<th>Track ID</th>
<th>Artist</th>
<th>Title</th>
<th>Format ID</th>
<th>Run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>01011207201000005652000000000053</td>
<td>Maroon 5</td>
<td>Moves like Jagger</td>
<td>MP3</td>
<td>4:30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phone #</th>
<th>Subscriber ID</th>
<th>Subscriber ID</th>
<th>First Name</th>
<th>Last Name</th>
<th>Age</th>
<th>State</th>
<th>Customer Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2172618992</td>
<td>53546</td>
<td>53546</td>
<td>Jim</td>
<td>Morrison</td>
<td>25</td>
<td>CA</td>
<td>93</td>
</tr>
</tbody>
</table>
Splunk DB Connect

Reliable, scalable, real-time integration between Splunk and traditional relational databases

- Enrich search results with additional business context
- Easily import data into Splunk for deeper analysis
- Integrate multiple DBs concurrently
- Simple set-up, non-evasive and secure
Customer Open Source Alternative
## Hadoop Advantage / Disadvantage

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheap Storage</td>
<td>Requires Coding for most Analytics</td>
</tr>
<tr>
<td>Batch Distributed Processing</td>
<td>No Visualization Tools</td>
</tr>
<tr>
<td></td>
<td>No OOTB Apps / Solutions</td>
</tr>
</tbody>
</table>
Real-World Customer Architecture
Summary Architecture

Real Time Data - 25 Indexers

Historical data (VIX) 60 Hortonworks nodes

Enrichment data (lookup) - MySQL DB

2000 Forwarders

3 instances
Splunk / Hunk / DB Connect
Search Heads
Splunk Deployment Architecture

2,000 forwarders

Web server

~2TB per day

25 indexers

~250 Users

~30 Concurrent Users

3 search head

Web server

forwarder
Hadoop Architecture

~30 Flume Agents
~60 Data Nodes
~1.2 PB of storage
~2 years data retention
Splunk + Hunk = All the Data

- Real Time
- Analytics
- Alerts
- Apps

- Batch
- Compliment Splunk Analytics
- Historical searches

Web server

App server

indexer

indexer

data node

data node

data node

Hive
DB Connect Architecture

• Install DB Connect on a Search Head
• Use DB Connect for Lookup
• Several Lookups coming from two different MySQL Databases
• Lookup Enrich log data with business insight
# DB - Architecture Performance Impact

<table>
<thead>
<tr>
<th>Command</th>
<th>Connection</th>
<th>Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indexing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inputs - dbmon-tail</td>
<td>Medium number of connections (Small amount of data - only delta)</td>
<td>DB to Index (connection pooling)</td>
</tr>
<tr>
<td><strong>Recommended</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inputs – dbmon-dump</td>
<td>Small amount of connections (Lots of data per connection)</td>
<td>DB to Index (connection pooling)</td>
</tr>
<tr>
<td><strong>Not Indexing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Search – DBXQuery</td>
<td>Lots of DB Connections (Small amount of data)</td>
<td>DB to Search Head</td>
</tr>
<tr>
<td>Lookups <strong>Selected this option</strong></td>
<td>Lots of DB Connections</td>
<td>DB to Search Head</td>
</tr>
</tbody>
</table>
Summary Architecture

Real Time Data - 25 Indexers

Historical data (VIX) 60 Hortonworks nodes

Enrichment data (lookup) - MySQL DB

3 instances Splunk / Hunk / DB Connect
Search Heads

2000 Forwarders
Archive Splunk Data to HDFS or AWS S3

Drive Down TCO by Archiving Historical Data to Commodity Hardware
Unified Search

Intelligently Search Across Real-Time and Historical Data using the Same Splunk Interface

Real-Time Data

Historical Data in Hadoop
Roadmap Architecture

- 3 instances
  - Splunk / Hunk / DB Connect
  - Search Heads

Real Time Data - 25 Indexers

Historical data (VIX)
- 60 Hortonworks nodes

Enrichment data (lookup) - MySQL DB

2000 Forwarders

Unified Search

Archive

Splunk / Hunk / DB Connect

Search Heads
Customer Chosen Architecture Demo
Support Our Military Kids

Take our Survey!

Splunk will Donate $10 Dollars to our Military Kids
Plus a bonus if we hit 350 number of completed surveys onsite.
Text Splunk to 878787.
Thank You