Bring Context To Your Machine Data With Hadoop, RDBMS & Splunk

Raanan Dagan and Rohit Pujari

September 25, 2017 | Washington, DC
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 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.

Splunk, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2017 Splunk Inc. All rights reserved.
Agenda

- Splunk Big Data Architecture
- Alternative Open Source Approach
- Real-World Customer Architecture
- End-to-end Demonstration
Big Data Technologies

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

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

Hadoop
- Semi-Structured
- Schema at Read
- MapReduce
- HDFS Storage
- Distributed File System

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

ETL
- Real-Time Indexing
- MapReduce
Splunk: Open And Extensible

Databases = Splunk DB Connect (Hive, Impala, Oracle)

Hadoop = Analytics for Hadoop, Hadoop Data Roll, Connect

Kafka = Splunk Kafka Add-On, Kafka with HEC

Spark = Spark SQL

NoSQL = MongoDB, Hbase, Cassandra apps
Splunk Enterprise Architecture

1. Modular
   - Syslog
   - TCP/UDP

2. Indexing Tier
   - Search Head Cluster
   - Splunk Analytics for Hadoop

3. NoSQL
   - Splunk Hadoop Data Roll

Forwarder
Windows/*NIX
HTTP

Wire Data
Syslog
HTTP
RDBMS

RDBMS
DBCConnect

DBCConnect

© 2017 SPLUNK INC.
Splunk And Hadoop - Products

- **Splunk Analytics for Hadoop:**
  - Analyze Hadoop Data using Hadoop MapReduce Processing

- **Splunk Hadoop Connect:**
  - Export data from Splunk to Hadoop

- **Hadoop Data Roll**
  - Archive Splunk indexers to Hadoop

- **Splunk Monitor Hadoop:**
  - Monitor Hadoop
Splunk & Hadoop Architecture

1. RDBMS, Syslog TCP/UDP, Wire Data, HTTP
   - 0101010
   - 0010101
   - 1010010

2. Search Head Cluster

Input Sources:
- Splunk
- SQoop
- Flume
- Kafka
- Manual scripts

Wire Data
- TCP/UDP

Output Destinations:
- RDBMS
- Syslog
- HTTP

Splunk Analytics for Hadoop
Splunk Big Data Technologies

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

**Splunk Analytics for Hadoop**
- Schema at Read
- Key-Value, Column, Document & Other Stores
- MapReduce
- HDFS Storage

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

- **RDBMS**
  - Oracle, MySQL, IBM DB2, Teradata

- **Cassandra, Hbase, MongoDB**

- **MapReduce**

- **Distributed File System**

- **Time-Series, Unstructured, Heterogeneous**
Splunk Scalability

Enterprise-class Availability and Scale

- 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 Splunk Indexers

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

Real-time Buffer

Index Pipeline

Raw data
Index Files

Splunk Index

Real-time Search Process
Splunk With Hadoop - 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

Security: Access Control, Pass Through Authentication, Kerberos

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

Media server logs (machine data)

Phone Number | IP Address | Track ID
---|---|---
Mar 01 19:18:50:000 aaa2 radiusd[12548]:[ID 959576 local1.info] INFO RADOP(13) acct start for 2172618992@splunktel.com 10.164.232.181 from 12.130.60.5 recorded OK.

Mar 01 19:18:50:150 10.2.1.34 GET /sync/addtolibrary/010112072010000056520000000000000053 - 80 - 10.164.232.181 "Mozilla/5.0 (iPhone; CPU iPhone OS 5_0_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 local1.info] INFO RADOP(13) acct stop for 2172618992@splunktel.com 10.164.232.181 from 12.130.60.5 recorded OK.

© 2017 SPLUNK INC.
Structured Data Contains Business Context

Phone number: 2172618992
IP address: 10.164.232.181
Track ID: 01011207201000005652000000000053

<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>

Customer, product databases

<table>
<thead>
<tr>
<th>Phone #</th>
<th>Subscriber ID</th>
<th>Subscriiber ID</th>
<th>First Name</th>
<th>Last Name</th>
<th>Age</th>
<th>State</th>
<th>Custome r 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
- Ingest, transform machine data in Splunk and export it to relational databases
- Integrate multiple DBs concurrently
- Simple set-up, non-evasive and secure
Open Source Alternatives
# Hadoop Complexity

## Ongoing Innovation in Apache

<table>
<thead>
<tr>
<th></th>
<th>HDP 2.6* 1H2017</th>
<th>HDP 2.5 Aug 2016</th>
<th>HDP 2.4 Mar 2016</th>
<th>HDP 2.3 Oct 2015</th>
<th>HDP 2.2 Dec 2014</th>
<th>HDP 2.1 April 2014</th>
<th>HDP 2.0 Oct 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.73 0.160 1.21* 2.1**</td>
<td>2.73 0.160 1.21* 2.1**</td>
<td>2.73 0.160 1.21* 2.1**</td>
<td>2.73 0.160 1.21* 2.1**</td>
<td>2.73 0.160 1.21* 2.1**</td>
<td>2.73 0.160 1.21* 2.1**</td>
<td>2.73 0.160 1.21* 2.1**</td>
</tr>
<tr>
<td></td>
<td>0.82 0.7 5.51 1.63* 2.1**</td>
<td>0.82 0.7 5.51 1.63* 2.1**</td>
<td>0.82 0.7 5.51 1.63* 2.1**</td>
<td>0.82 0.7 5.51 1.63* 2.1**</td>
<td>0.82 0.7 5.51 1.63* 2.1**</td>
<td>0.82 0.7 5.51 1.63* 2.1**</td>
<td>0.82 0.7 5.51 1.63* 2.1**</td>
</tr>
<tr>
<td></td>
<td>0.91 0.7 1.1 4.7 1.7 1.1 0.10 0.8</td>
<td>0.91 0.7 1.1 4.7 1.7 1.1 0.10 0.8</td>
<td>0.91 0.7 1.1 4.7 1.7 1.1 0.10 0.8</td>
<td>0.91 0.7 1.1 4.7 1.7 1.1 0.10 0.8</td>
<td>0.91 0.7 1.1 4.7 1.7 1.1 0.10 0.8</td>
<td>0.91 0.7 1.1 4.7 1.7 1.1 0.10 0.8</td>
<td>0.91 0.7 1.1 4.7 1.7 1.1 0.10 0.8</td>
</tr>
<tr>
<td>Pig</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Druid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tez</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zeppelin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slider</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phoenix</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accumulo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Falcon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atlas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sqoop</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kafka</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambari</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zookeeper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oozle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knox</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranger</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DATA MGMT** | **DATA ACCESS** | **GOVERNANCE & INTEGRATION** | **OPERATIONS** | **SECURITY**

**HORTONWORKS DATA PLATFORM**
Hadoop Use Cases

- Application Delivery
- IT Operations
- Security and Compliance

Splunk use cases

- Business Analytics
- IoT

Splunk or Hadoop use cases

- ETL for RDBMS

Hadoop use cases
Customer Architecture
Summary Architecture

Real Time Data - 25 Indexers

Historical data (VIX) - 60 Hortonworks nodes

Enrichment data (lookup) - MySQL DB

3 instances Splunk / Hadoop / DB

Connect Search Heads

2000 Forwarders

...
Splunk Deployment Architecture

- Web server
- 2,000 forwarders
- 3 search head
- 25 indexers
- ~2TB per day
- ~250 Users
- ~30 Concurrent Users
- Forwarder
Hadoop Architecture

~30 Flume Agents
~60 Data Nodes
~1.2 PB of storage
~2 Years data retention
Splunk + Hadoop = All The Data

- Real Time
- Analytics
- Alerts
- Apps

- Batch
- Compliment Splunk Analytics
- Historical searches
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>Outputs</strong></td>
<td>Lots of DB Connections (Small amount of data)</td>
<td>Search Head to DB (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</td>
<td>DB to Search Head</td>
</tr>
<tr>
<td><strong>Lookups</strong></td>
<td>Lots of DB Connections</td>
<td>DB to Search Head</td>
</tr>
<tr>
<td><strong>Selected this option</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary Architecture

- 3 instances Splunk / Hadoop / DB
- Connect Search Heads
- Real Time Data - 25 Indexers
- Historical data (VIX) - 60 Hortonworks nodes
- Enrichment data (lookup) - MySQL DB
- 2000 Forwarders
Customer Architecture Demo
Summary

- Splunk is open and extensible
- Splunk enables you to combine data from multiple sources for enriched insights
- Splunk can complement and fill the gaps in open source technologies
Thank You

Don't forget to rate this session in the .conf2017 mobile app