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?

Sources

Order Processing

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

Middleware Error

```
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 connection. The
DBMS driver exception was: [BEA][Oracle JDBC Driver]Error establishing socket to host and port:
ACMEDB-01:1521. Reason: Connection refused
```

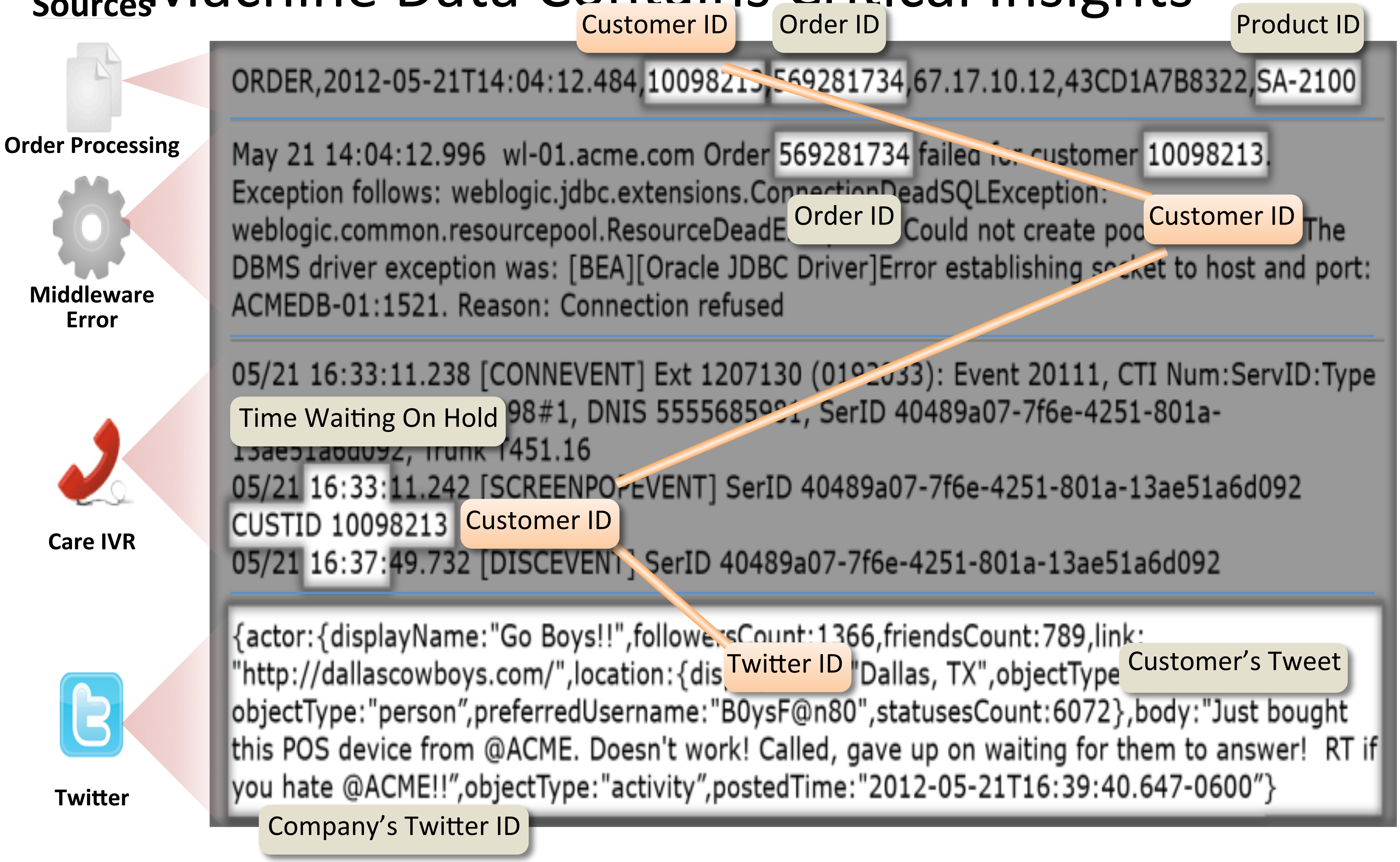
Care IVR

```
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
```

Twitter

```
{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



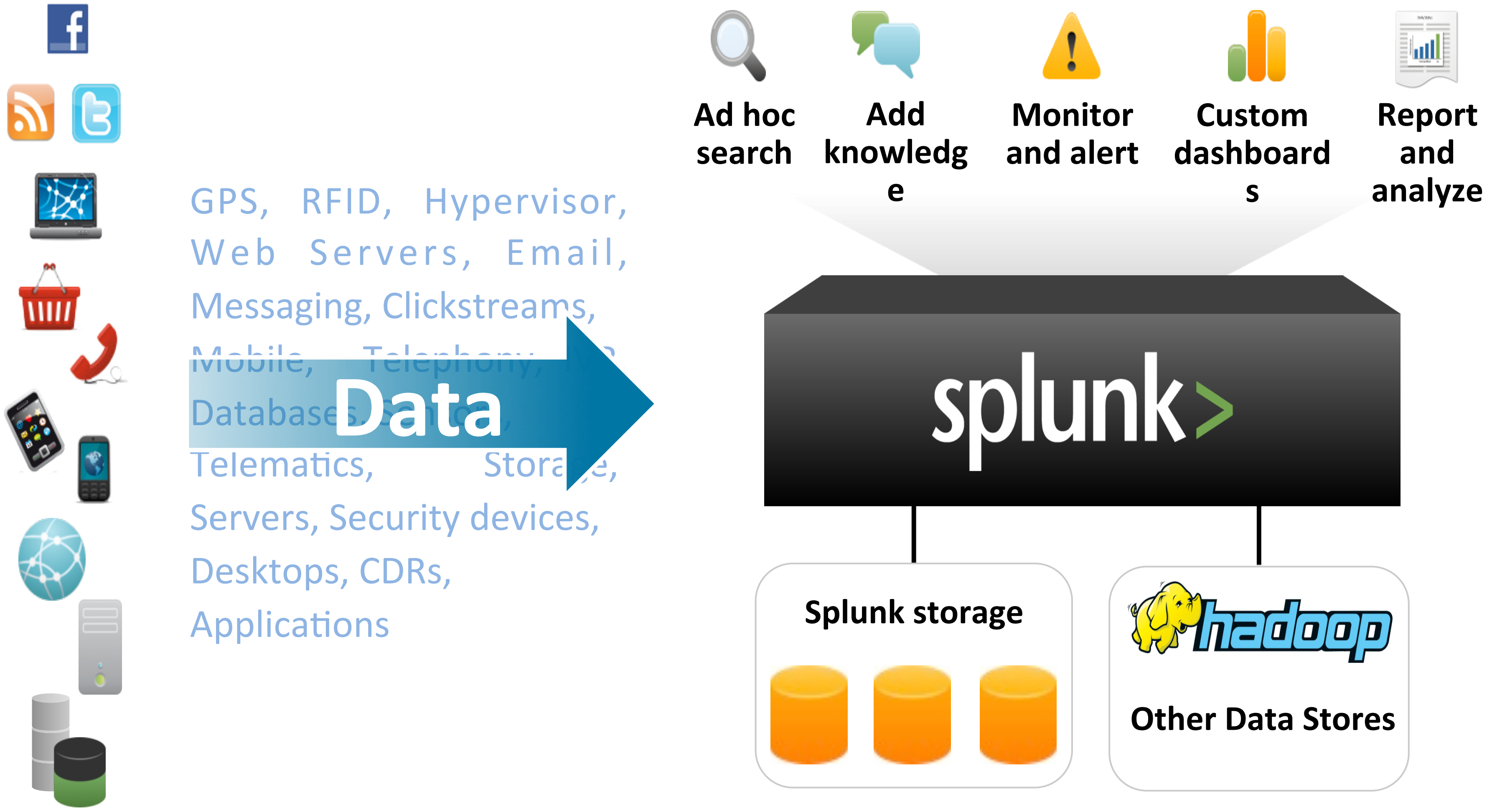
```
ORDER,2012-05-21T14:04:12.484,10098213,569281734,67.17.10.12,43CD1A7B8322,SA-2100
```

```
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.ResourceDeadE[illegible] Could not create poo[illegible] The
DBMS driver exception was: [BEA][Oracle JDBC Driver]Error establishing socket to host and port:
ACMEDB-01:1521. Reason: Connection refused
```

```
05/21 16:33:11.238 [CONNEVENT] Ext 1207130 (0192033): Event 20111, CTI Num:ServID:Type
[illegible]98#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:{dis[illegible]"Dallas, TX",objectType[illegible]
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 Enterprise with Hadoop



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

Data

Ad hoc search

Add knowledge

Monitor and alert

Custom dashboards

Report and analyze

splunk>

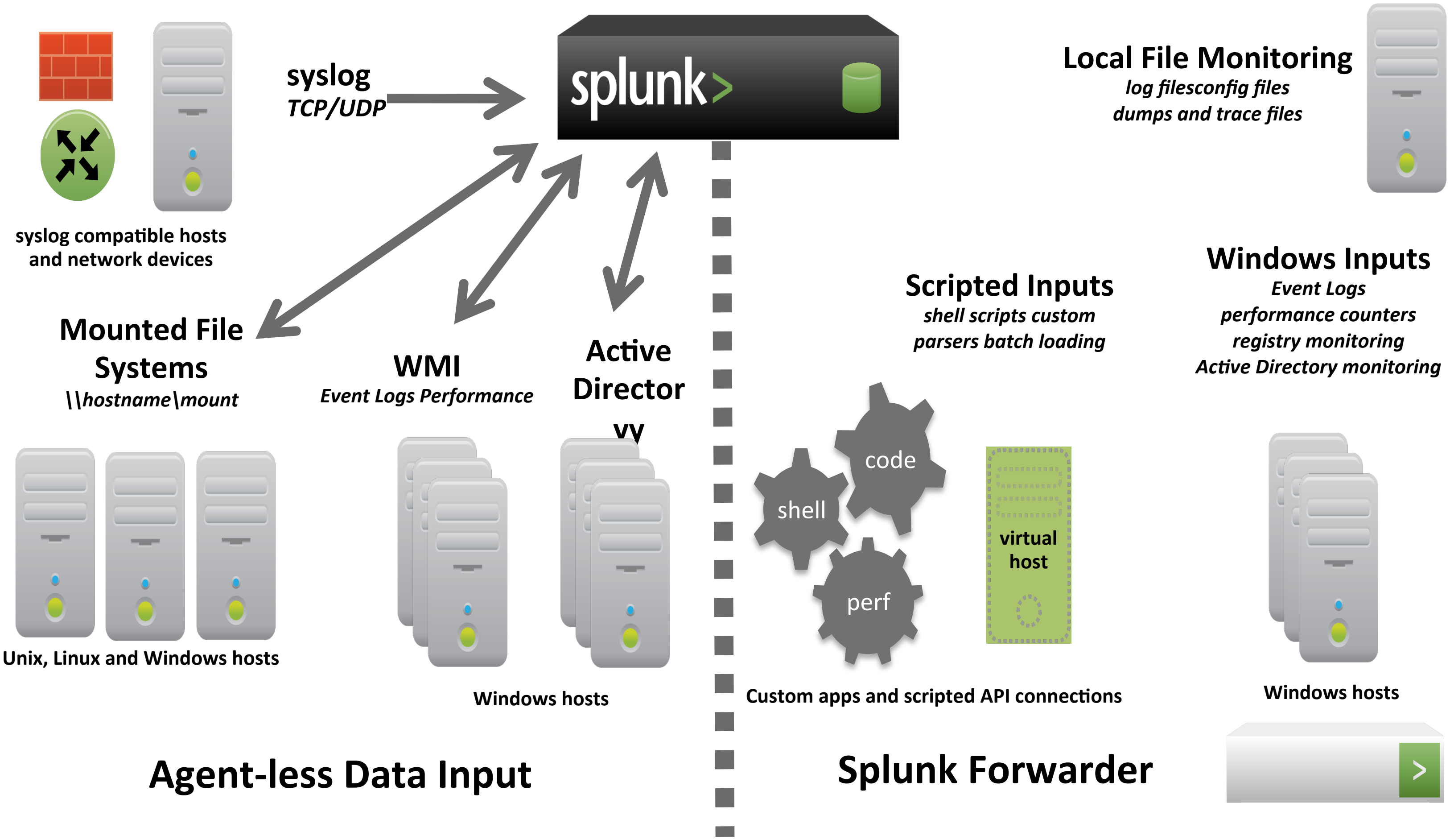
Splunk storage

hadoop

Other Data Stores

Getting Data into Splunk

Agent and Agent-less Approach for Flexibility



syslog
TCP/UDP

syslog compatible hosts
and network devices

Mounted File Systems
\\hostname\mount

WMI
Event Logs Performance

Active Director
vy

Unix, Linux and Windows hosts

Windows hosts

Agent-less Data Input

Local File Monitoring
log filesconfig files
dumps and trace files

Scripted Inputs
shell scripts custom
parsers batch loading

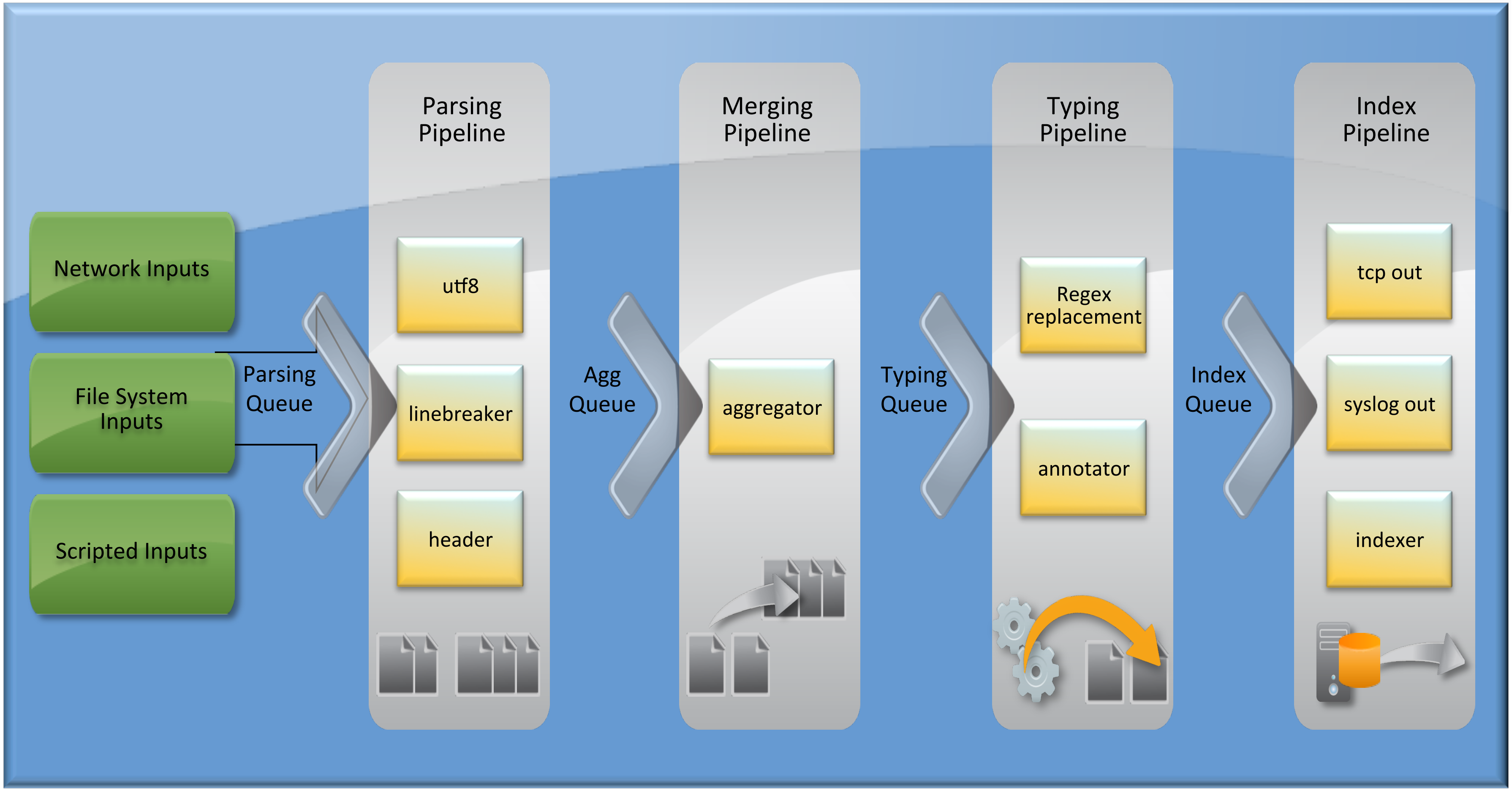
Windows Inputs
Event Logs
performance counters
registry monitoring
Active Directory monitoring

Custom apps and scripted API connections

Windows hosts

Splunk Forwarder

Pipelines/Processors



Network Inputs

File System Inputs

Scripted Inputs

Parsing Queue

Parsing Pipeline

utf8

linebreaker

header

Agg Queue

Merging Pipeline

aggregator

Typing Queue

Typing Pipeline

Regex replacement

annotator

Index Queue

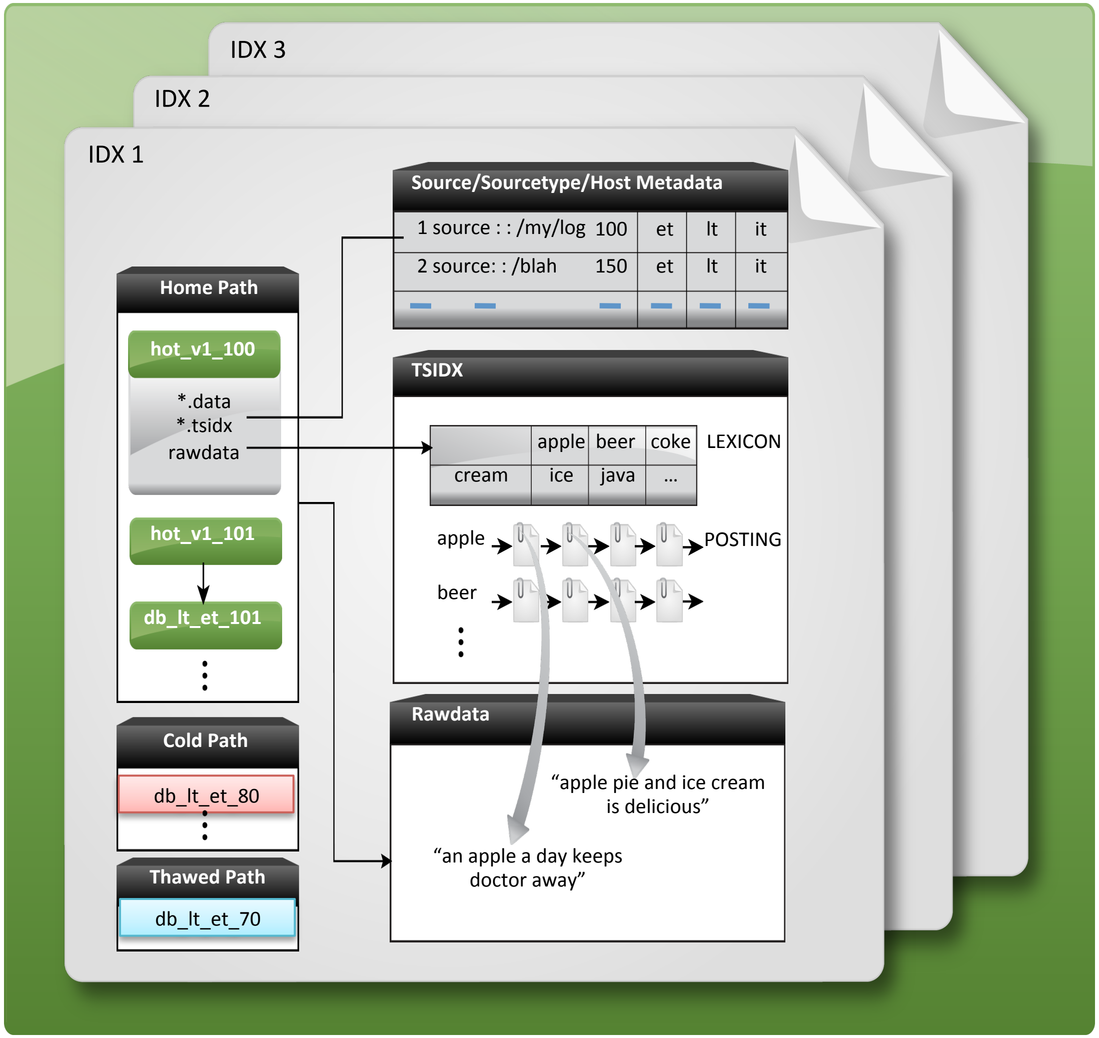
Index Pipeline

tcp out

syslog out

indexer

Index Processor

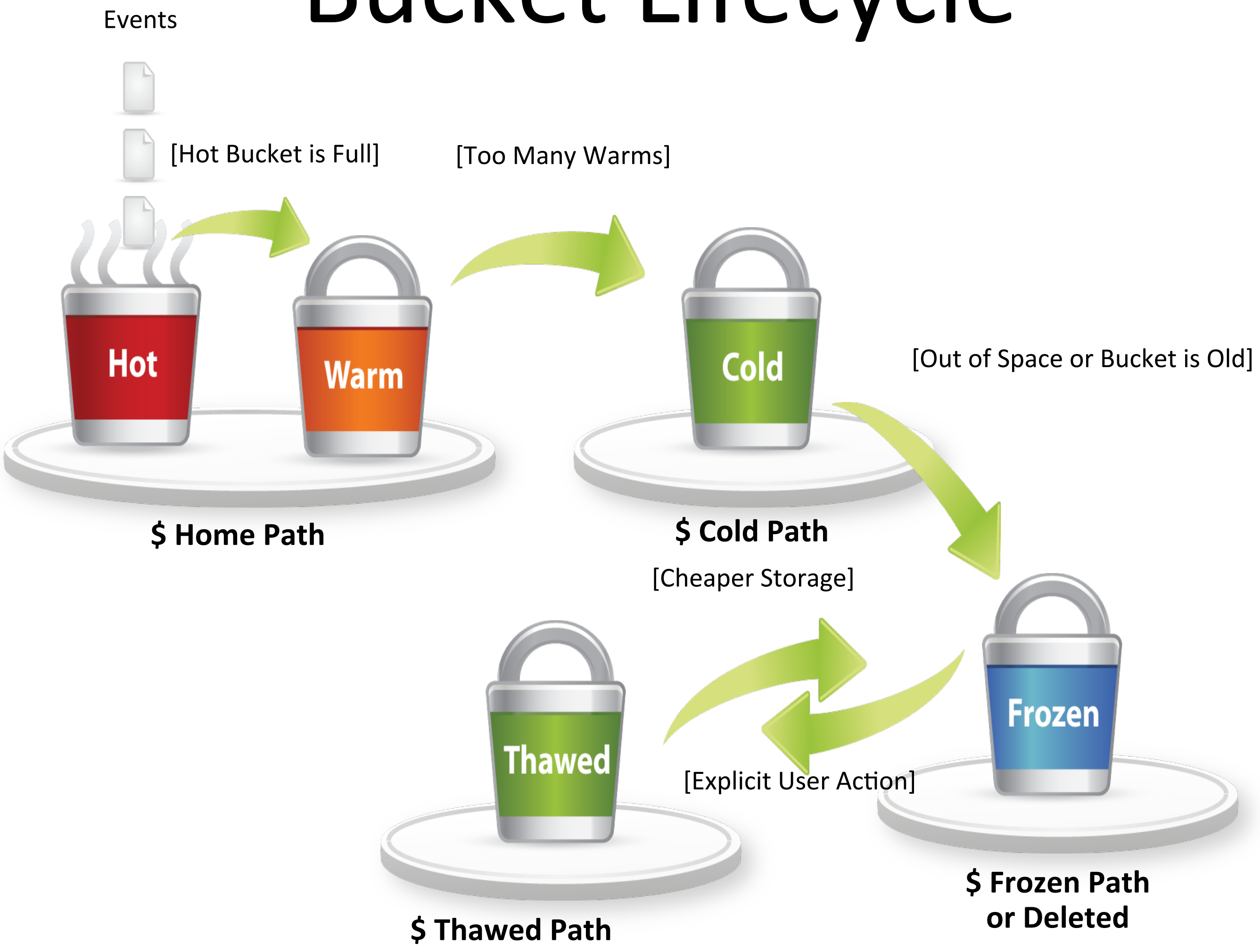


1 source : : /my/log	100	et	lt	it
2 source: : /blah	150	et	lt	it

	apple	beer	coke
cream	ice	java	...

apple
beer
POSTING
Cold Path
db_lt_et_80
Thawed Path
db_lt_et_70
Rawdata
"apple pie and ice cream is delicious"
"an apple a day keeps doctor away"

Bucket Lifecycle



Events

[Hot Bucket is Full]

[Too Many Warms]

[Out of Space or Bucket is Old]

$ Home Path

$ Cold Path

[Cheaper Storage]

[Explicit User Action]

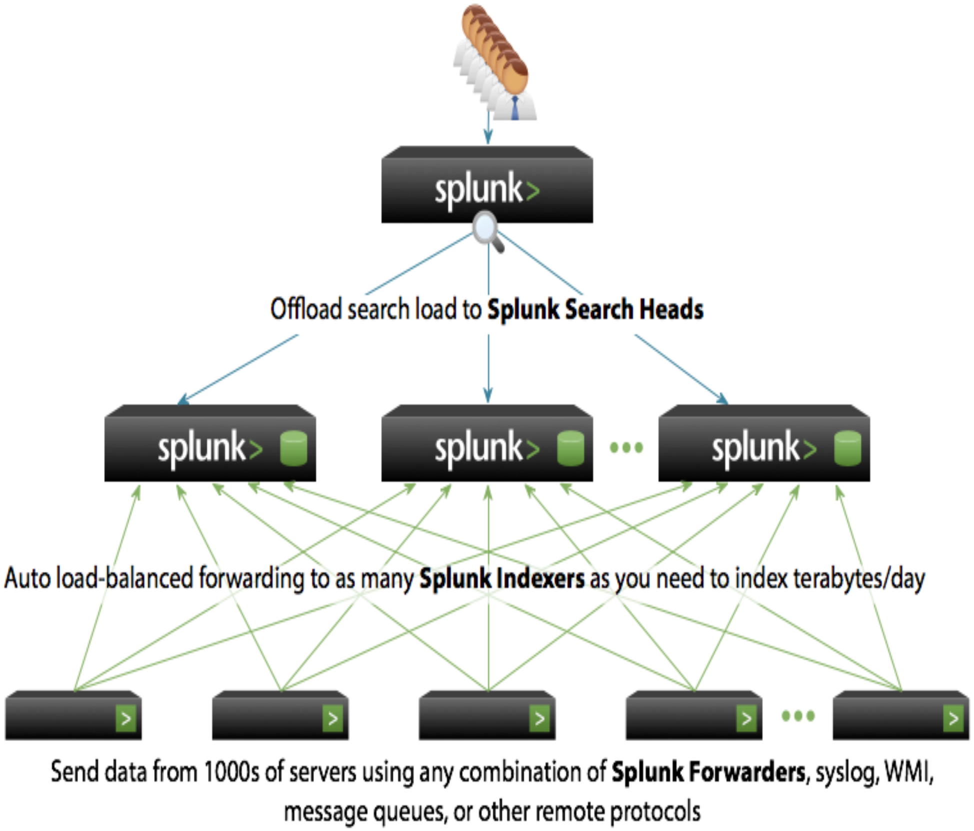
$ Frozen Path or Deleted

$ Thawed Path

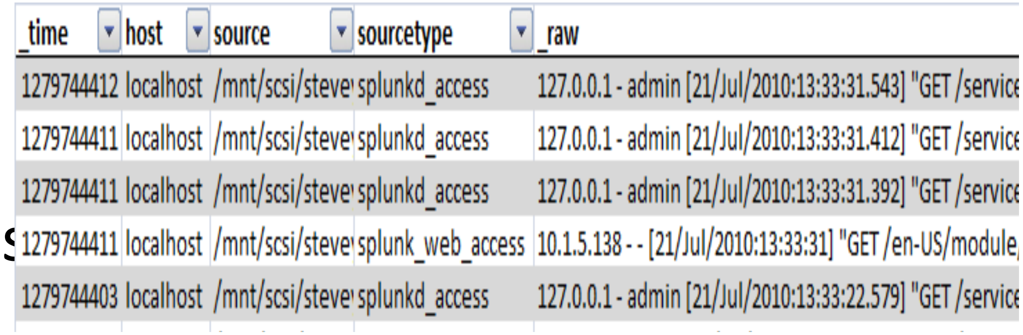
Scales to TBs/day and Thousands of Users

- Automatic load balancing linearly scales indexing

- Distributed search and MapReduce linearly scales search and reporting


splunk>
Offload search load to **Splunk Search Heads**
splunk>
splunk>
•••
splunk>
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

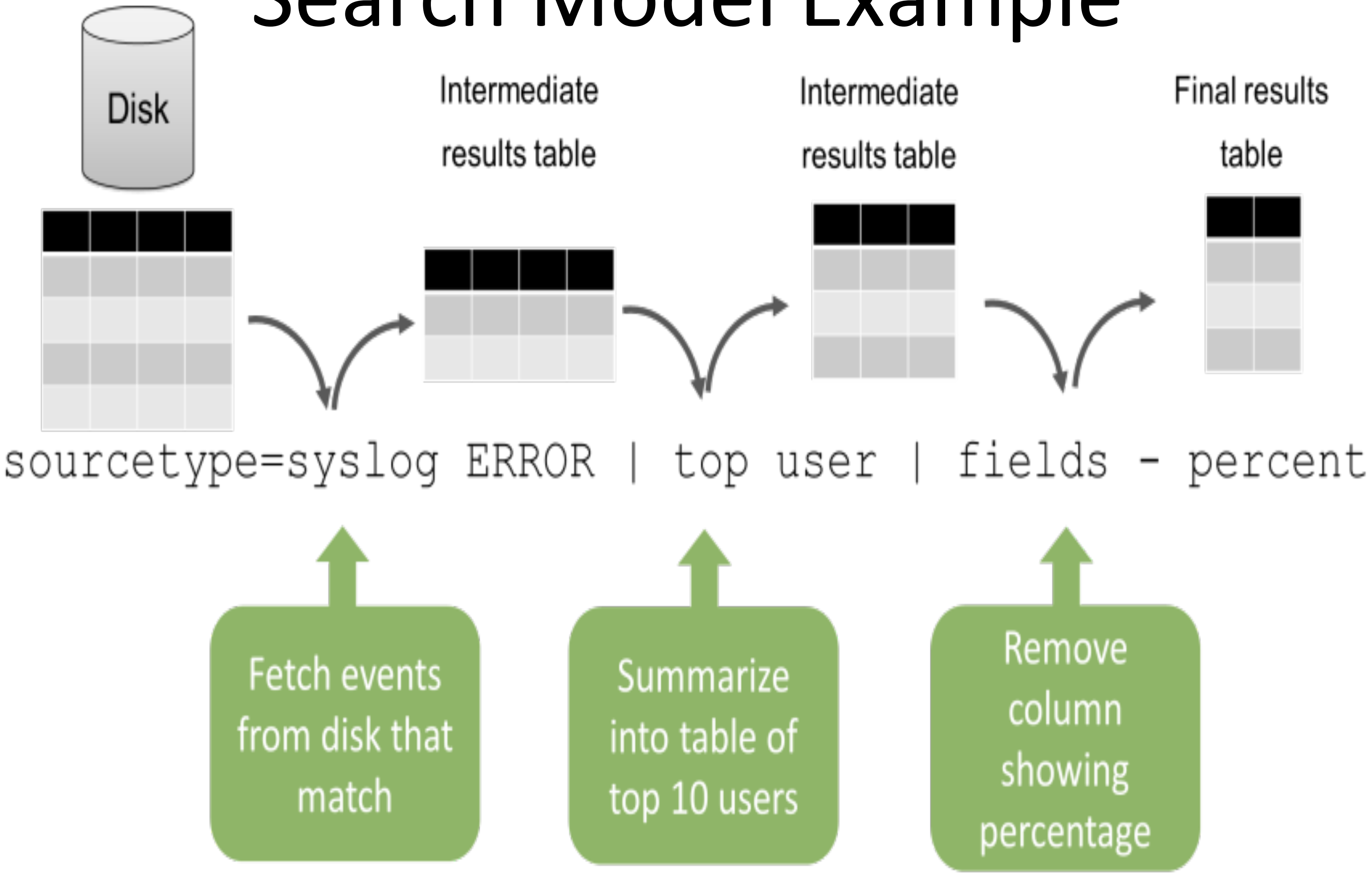
Search Model



_time	host	source	sourcetype	_raw
1279744412	localhost	/mnt/scsi/steve	splunkd_access	127.0.0.1 - admin [21/Jul/2010:13:33:31.543] "GET /service
1279744411	localhost	/mnt/scsi/steve	splunkd_access	127.0.0.1 - admin [21/Jul/2010:13:33:31.412] "GET /service
1279744411	localhost	/mnt/scsi/steve	splunkd_access	127.0.0.1 - admin [21/Jul/2010:13:33:31.392] "GET /service
1279744411	localhost	/mnt/scsi/steve	splunk_web_access	10.1.5.138 - - [21/Jul/2010:13:33:31] "GET /en-US/module
1279744403	localhost	/mnt/scsi/steve	splunkd_access	127.0.0.1 - admin [21/Jul/2010:13:33:22.579] "GET /service

- Splunk Database as a
  - 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



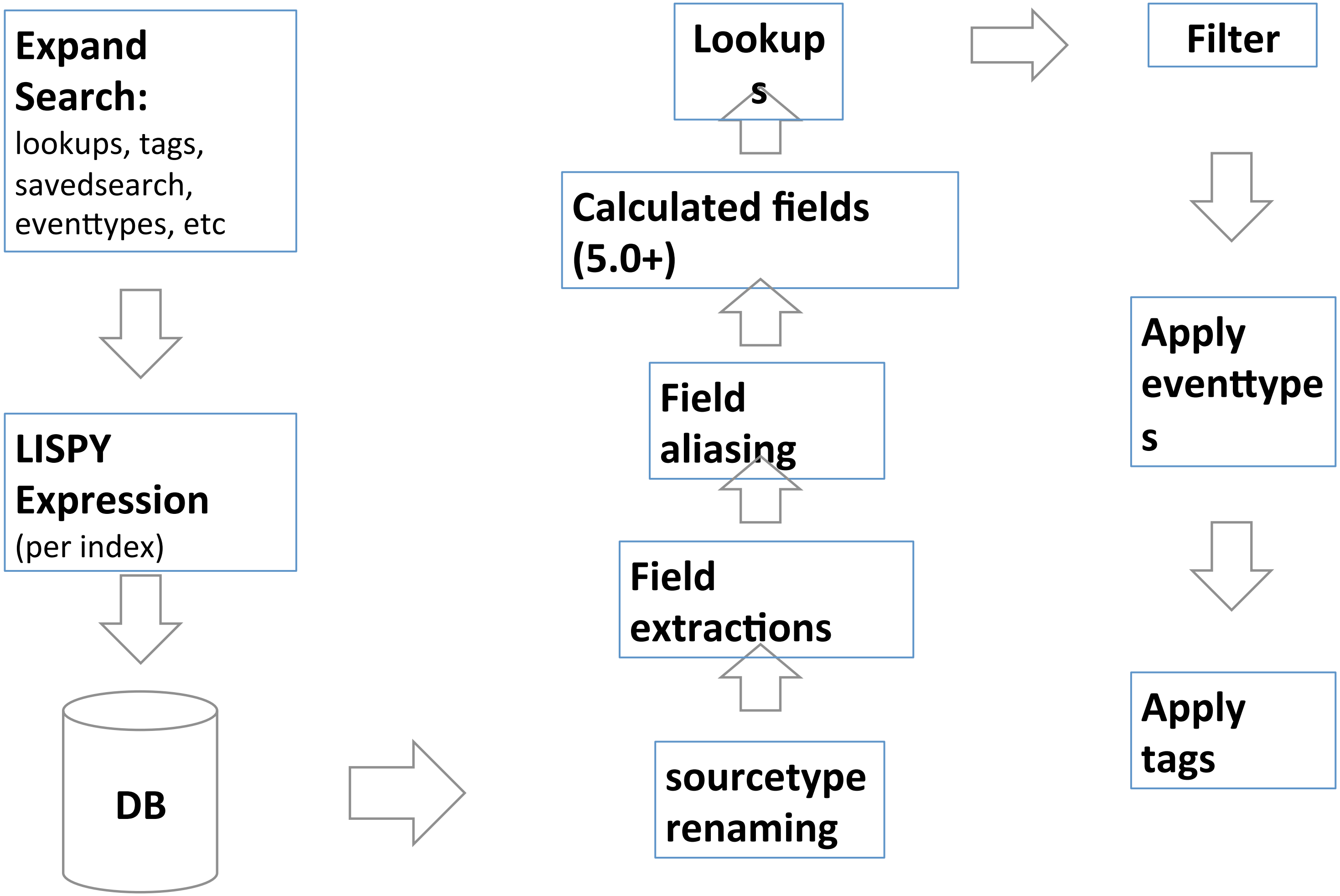
```
sourcetype=syslog ERROR | top user | fields - percent
```

Fetch events from disk that match

Summarize into table of top 10 users

Remove column showing percentage

Search Command



Expand Search:
lookups, tags, savedsearch, eventtypes, etc

LISPY Expression
(per index)

DB

sourcetype renaming

Field extractions

Field aliasing

Calculated fields (5.0+)

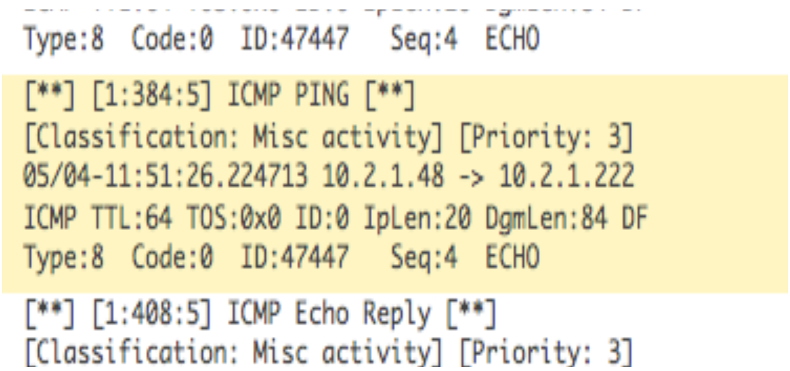
Lookup
s

Filter

Apply eventtype
s

Apply tags

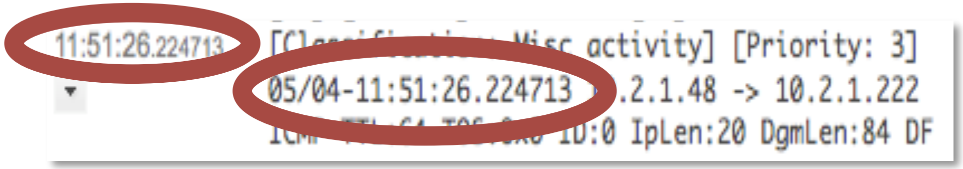
Inside Universal Indexing



```
Type:8  Code:0  ID:47447   Seq:4  ECHO
[**] [1:384:5] ICMP PING [**]
[Classification: Misc activity] [Priority: 3]
05/04-11:51:26.224713 10.2.1.48 -> 10.2.1.222
ICMP TTL:64 TOS:0x0 ID:0 IpLen:20 DgmLen:84 DF
Type:8  Code:0  ID:47447   Seq:4  ECHO
[**] [1:408:5] ICMP Echo Reply [**]
[Classification: Misc activity] [Priority: 3]
```

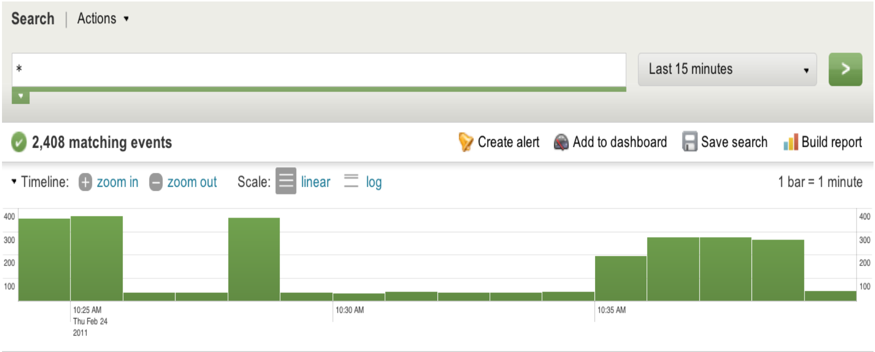
Automatic event boundary identification

Automatic timestamp normalization



```
11:51:26.224713   [Cl           Misc activity] [Priority: 3]
                  05/04-11:51:26.224713 .2.1.48 -> 10.2.1.222
                  ICM                  ID:0 IpLen:20 DgmLen:84 DF
```

...enable accurate searching and trending by time across all data:



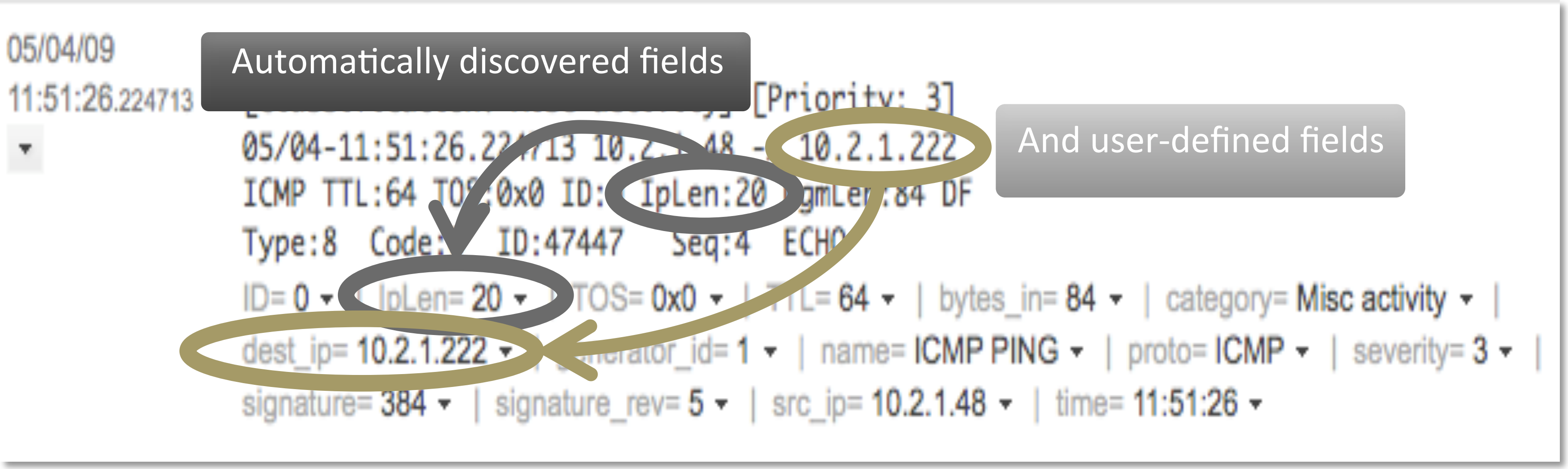
Search: *

Last 15 minutes

2,408 matching events

1 bar = 1 minute

Inside Search-time Knowledge Extraction

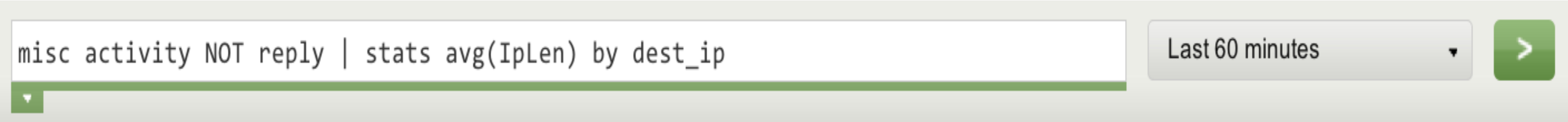


```
05/04/09
11:51:26.224713
▾
                 [Priority: 3]
05/04-11:51:26.224713 10.2.1.48 -> 10.2.1.222
ICMP TTL:64 TOS:0x0 ID:0 IpLen:20 DgmLen:84 DF
Type:8  Code:0  ID:47447   Seq:4  ECHO
ID= 0 ▾ | IpLen= 20 ▾ | TOS= 0x0 ▾ | TTL= 64 ▾ | bytes_in= 84 ▾ | category= Misc activity ▾ |
dest_ip= 10.2.1.222 ▾ | generator_id= 1 ▾ | name= ICMP PING ▾ | proto= ICMP ▾ | severity= 3 ▾ |
signature= 384 ▾ | signature_rev= 5 ▾ | src_ip= 10.2.1.48 ▾ | time= 11:51:26 ▾
```

Automatically discovered fields

And user-defined fields

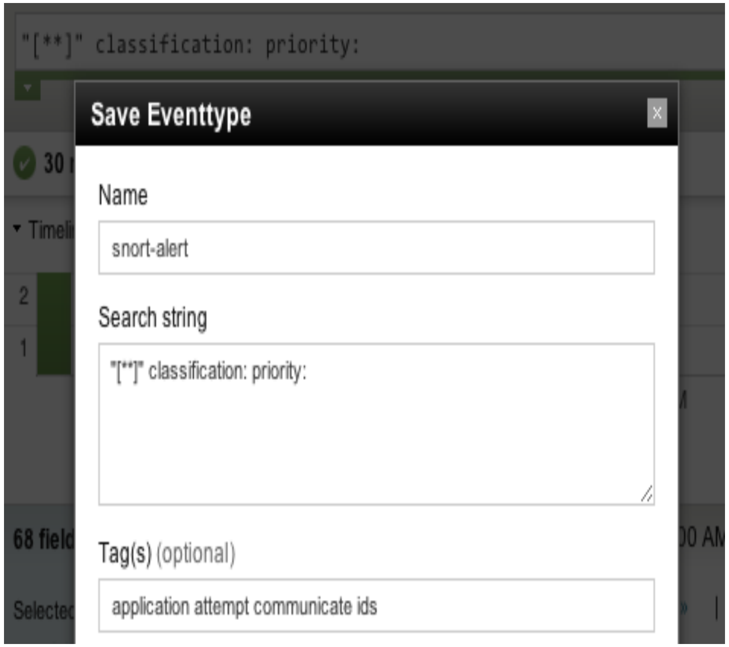
… enable statistics and precise search on specific fields:



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

Inside Search-time Knowledge Extraction

Searches saved as event types



```
"[**]" classification: priority:
```

Save Eventtype

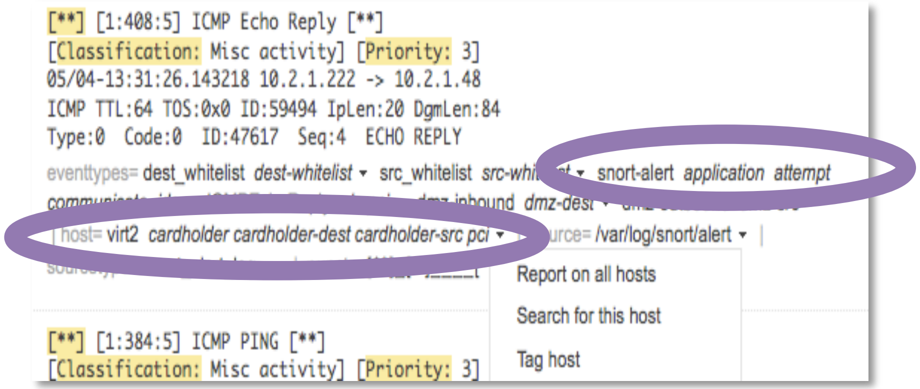
Name: snort-alert

Search string: "[**]" classification: priority:

Tag(s) (optional): application attempt communicate ids

… enable normalized reporting, knowledge sharing and granular access control.

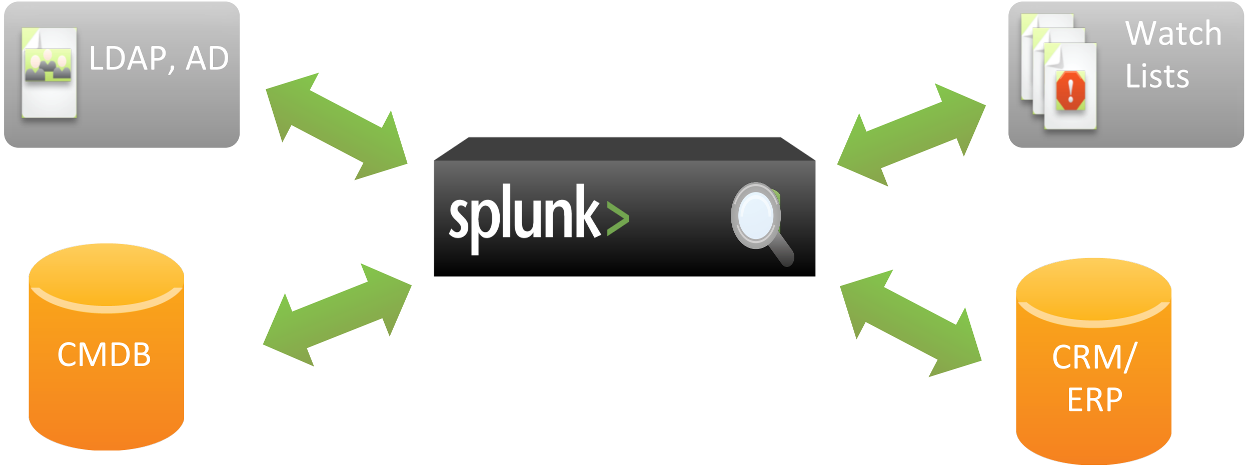
Plus tagging of event types, hosts and other fields



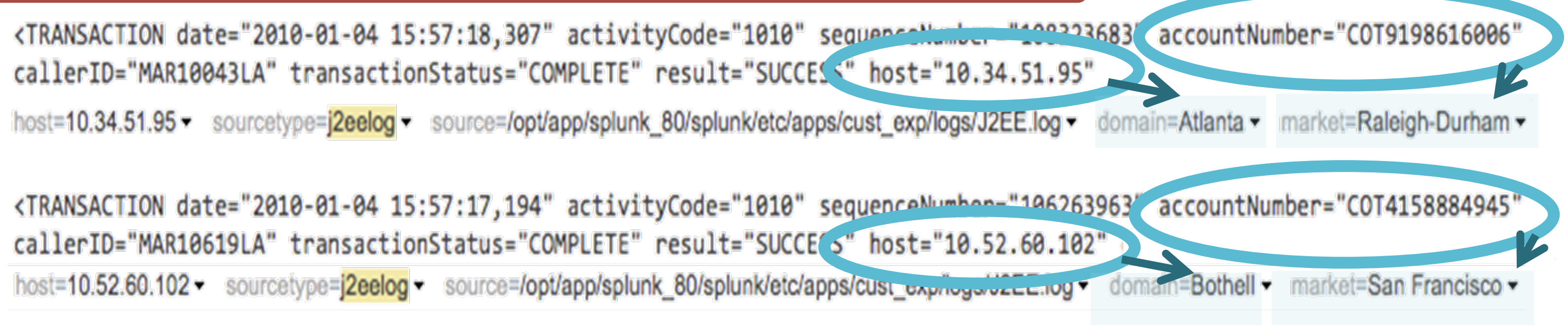
```
[**] [1:408:5] ICMP Echo Reply [**]
[Classification: Misc activity] [Priority: 3]
05/04-13:31:26.143218 10.2.1.222 -> 10.2.1.48
ICMP TTL:64 TOS:0x0 ID:59494 IpLen:20 DgmLen:84
Type:0  Code:0  ID:47617  Seq:4  ECHO REPLY
eventtypes= dest_whitelist dest-whitelist ▾ src_whitelist src-whi ▾ snort-alert application attempt
communicate ids dmz inbound dmz-dest ▾
host= virt2 cardholder cardholder-dest cardholder-src pci ▾ source= /var/log/snort/alert ▾

[**] [1:384:5] ICMP PING [**]
[Classification: Misc activity] [Priority: 3]
```

Integrate External Data

Extend analysis with lookups to external data sources



Correlate IP addresses with locations, accounts with regions

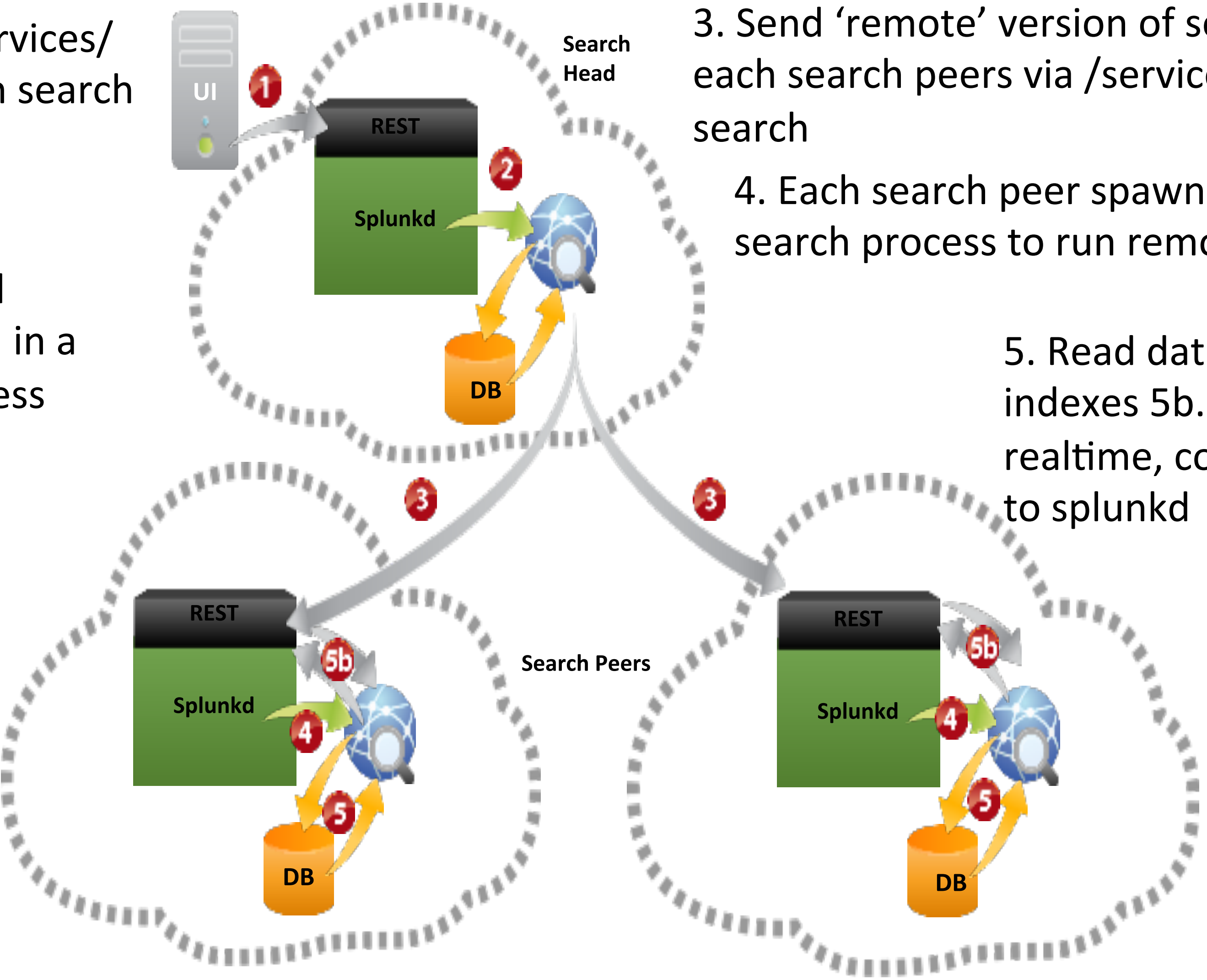


```
<TRANSACTION date="2010-01-04 15:57:18,307" activityCode="1010" sequenceNumber="[illegible]3683" accountNumber="COT9198616006"
callerID="MAR10043LA" transactionStatus="COMPLETE" result="SUCCESS" host="10.34.51.95"
host=10.34.51.95 ▾  sourcetype=j2eelog ▾  source=/opt/app/splunk_80/splunk/etc/apps/cust_exp/logs/J2EE.log ▾  domain=Atlanta ▾  market=Raleigh-Durham ▾

<TRANSACTION date="2010-01-04 15:57:17,194" activityCode="1010" sequenceNumber="106263963" accountNumber="COT4158884945"
callerID="MAR10619LA" transactionStatus="COMPLETE" result="SUCCESS" host="10.52.60.102"
host=10.52.60.102 ▾  sourcetype=j2eelog ▾  source=/opt/app/splunk_80/splunk/etc/apps/cust_exp/logs/J2EE.log ▾  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

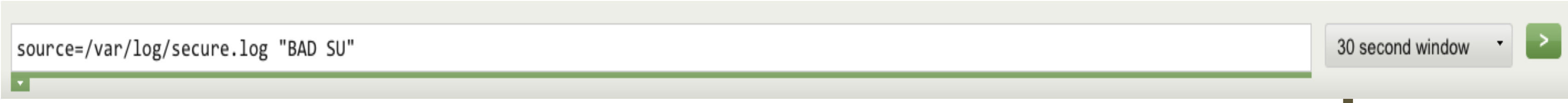


3. Send ‘remote’ version of search to each search peers via /services/streams/search

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

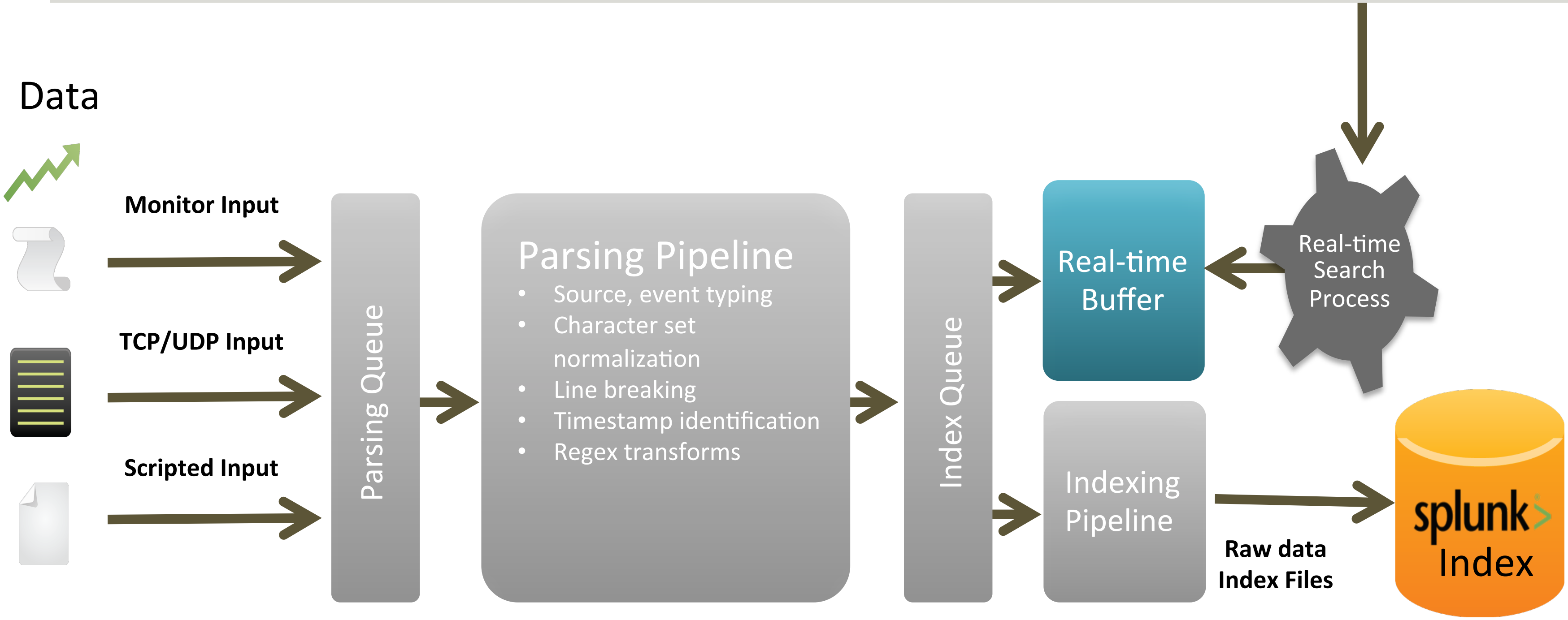
5. Read data from indexes 5b. For realtime, connect back to splunkd

Real-time Search



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

30 second window



Data

Monitor Input

TCP/UDP Input

Scripted Input

Parsing Queue

Parsing Pipeline

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

Index Queue

Real-time Buffer

Indexing Pipeline

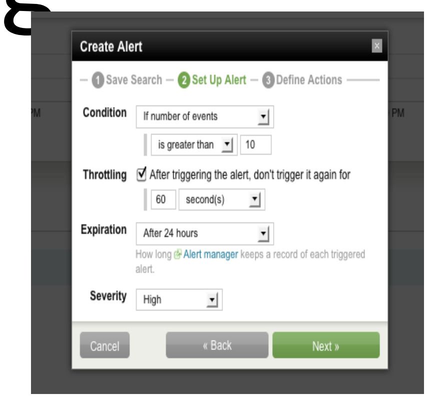
Real-time Search Process

Raw data Index Files

Index

Real-time Alerting

source=“/var/log/secure.log” “BAD SU”



Create Alert

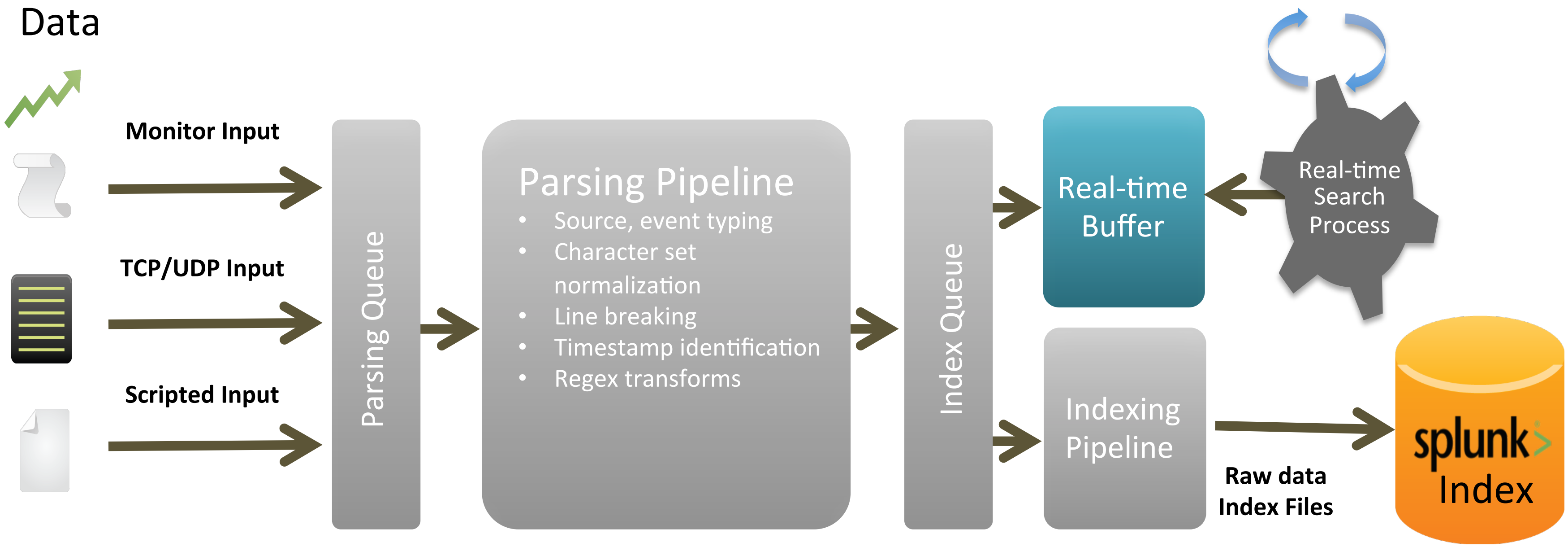
Condition: If number of events; is greater than; 10

Throttling: ☑ After triggering the alert, don't trigger it again for 60 second(s)

Expiration: After 24 hours

How long Alert manager keeps a record of each triggered alert.

Severity: High



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
```