.conf2015

Hunk Performance and Troubleshooting best practice

Raanan Dagan

Praveen Burgu

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.

Who are you?

- Raanan Dagan - Sr. SE, Big Data specialist
- Praveen Burgu – Sr. Software Engineer

Agenda

- Performance
  - 10 ways to optimize Hunk search performance: MR Jobs, Timestamp Extraction, Caching
- Troubleshoot
  - Inspect search job issues: MR Jobs, Performance, Timestamp

.conf2015

Hunk Performance

splunk>

Hunk Performance Main Points

1. Run MR Jobs
2. HDFS Storage
3. VIX with Timestamp / indexes.conf
4. File Format
5. Compression types / File size
6. Event breaking / Props.conf
7. Report Acceleration
8. Hardware
9. Search Head Clustering
10. Other Flags (Threads, Splits)

#1: Make Sure you use MR Jobs

Not MR Jobs – Just Splunk

- Index=xyz

Not MR Jobs – Just Splunk

- Index=xyz | stats count and using Verbose Mode

Yes, this will run MR Jobs

- Index=xyz | stats count and using Smart Mode

Allows you to use the Power of Hadoop MR Jobs parallel processing

2: HDFS Storage

This is BAD

- /data/root/dir/...

This is GOOD

- /data/root/dir/**2014/10/01**/....
- /data/root/dir/**2014/10/02**/....

This is BETTER

- /data/root/dir/2014/10/01/**app=apache**/...
- /data/root/dir/2014/10/01/**app=mysql**/...

Allows you to bring subset of data from HDFS based on time extraction

3: VIX with Timestamp / Indexes.conf

HDFS = /user/splunk/data/20141123/14/SFServer/myfile.gz

```
[hadoop]
vix.provider = MyHadoopProvider
vix.input.1.path = /user/splunk/data/*/*/${server}/...
vix.input.1.accept = \.gz$
vix.input.1.et.regex = .*?/data/(\d+)/(\d+)/.*?.gz
vix.input.1.et.format = yyyyMMddHH
vix.input.1.et.offset = 0
vix.input.1.lt.regex = .*?/data/(\d+)/(\d+)/.*?.gz
vix.input.1.lt.format = yyyyMMddHH
vix.input.1.lt.offset = 3600
```

Time extraction will enable you to use the Time Picker in the Hunk UI to bring Subset of the data

#4: File Format

- Don't add multiple sources into one file
- Use a self-describing format for the data whenever possible; e.g. json, avro, csv, Parquet, ORC, RC, etc.
- If using a log file, look at this list for Splunk Known Source Types (sourcetype=access_combined) http://docs.splunk.com/Documentation/Splunk/latest/Data/Listofpretrainedsourcetypes
- Look at the Splunk App Store for 600 other options to break the events / fields http://apps.splunk.com

Hunk will benefit if the file has some structure. Otherwise we will need to use REGEX to extract fields

#5: Compression type / File size

This is BAD (Large Non-splittable)

- 500MB GZ file

This is BAD (too many MR Jobs)

- 10,000 X 1kb files

This is GOOD (Large spilittable)

- 500MB LZO or Snappy file

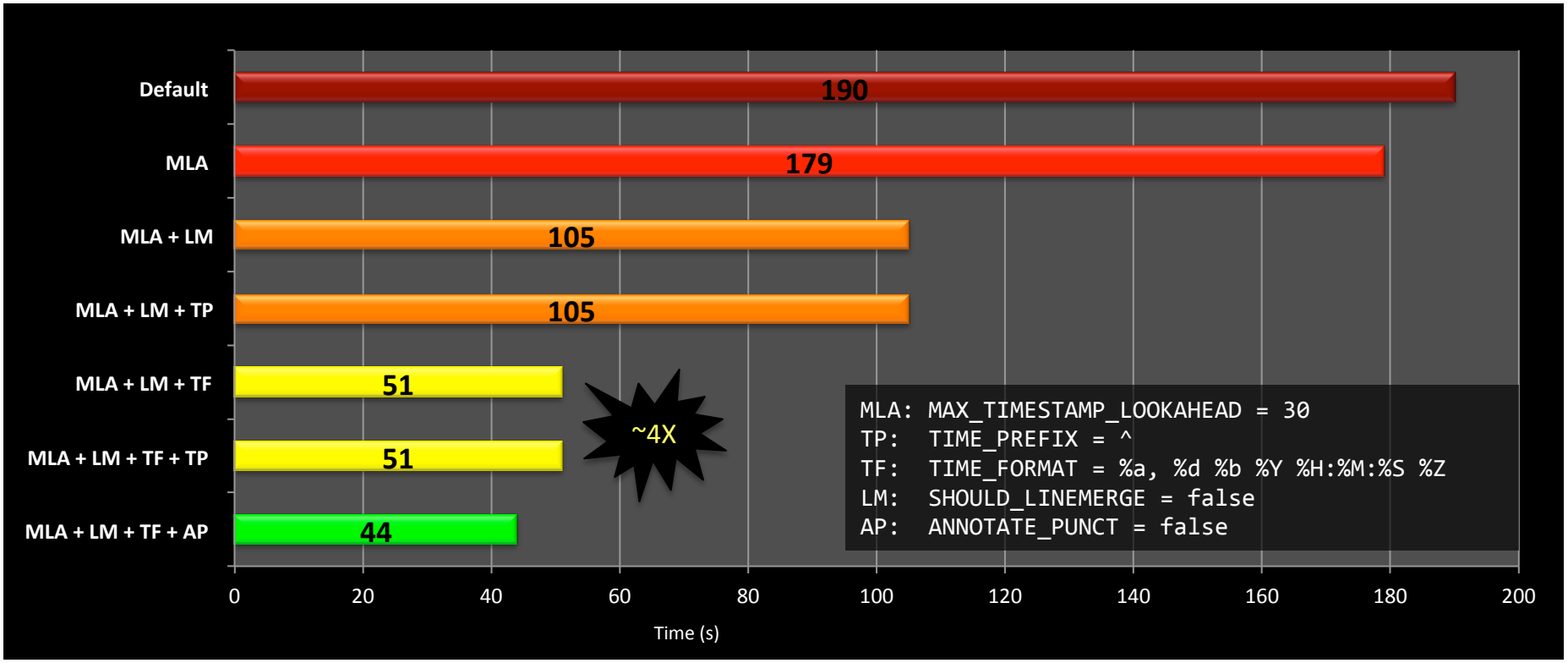
This is GOOD (Non-splittable, but 1 MR per file)

- 127MB or 63MB GZ file

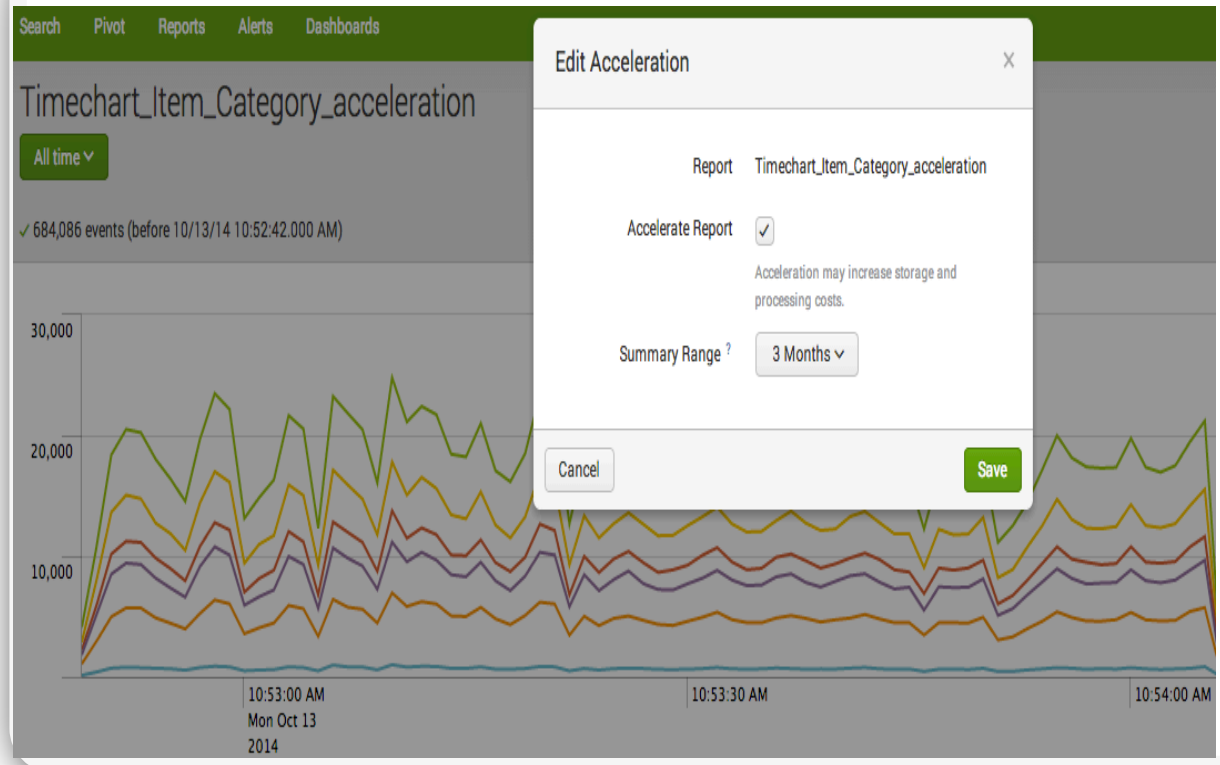
To avoid too many MR Jobs, or running out of memory make sure to use the correct compression or file size

#6: Index-time pipeline processing

http://docs.splunk.com/Documentation/Hunk/latest/Hunk/Performancebestpractices



#7: Report Acceleration



Report acceleration will improve performance – Bring data from Cache

NOTE: vix.env.HADOOP_HEAPSIZE = 1024 or above

Splunk and Hadoop - Caching options

#8: Hardware

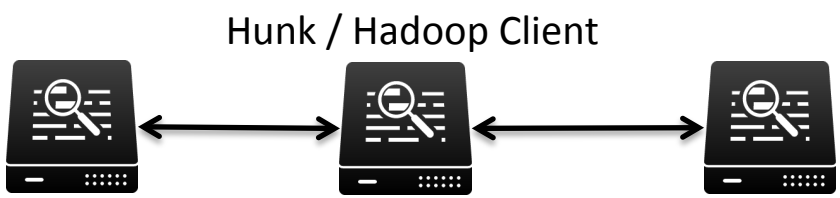
Dedicated search head

- Intel 64-bit chip architecture
- 4 CPUs, 4 cores per CPU, at least 2 Ghz per core
- 12 GB RAM
- 2 x 300 GB, 10,000 RPM SAS hard disks, configured in RAID 1
- Standard 1Gb Ethernet NIC, optional 2nd NIC for a management network
- Standard 64-bit Linux

Data Nodes = The SplunkD indexer is installed, by default, on each data node '/tmp/splunk' directory. You just need to make sure you have about 40MB, or more, space in that directory

A good Hardware with multiple cores can be very beneficial to interact with hundreds of end users

#9: Search Head Clustering



Add Many Concurrent Users

1. No Single Point of Failures = Dynamic Captain
2. "One Configuration" across SH = Automatic Config replication
3. Horizontal Scaling = Ability to add / remove SH nodes on running cluster

#10: Other Optimization Flags

Number of Jobs:

- **vix.splunk.search.mr.threads** - # of threads to use when reading map results from HDFS
- **vix.splunk.search.mr.maxsplits** - maximum number of splits in an MR job (Default to 10000)

Number of copies to each data node:

- **vix.splunk.setup.bundle.setup.timelimit** - time limit in ms for setting up bundle on TT
- **vix.splunk.setup.bundle.replication** - set custom replication factor for bundle on hdfs
- **vix.splunk.setup.package.replication** - set custom replication factor for splunk package on hdfs

VIX overrides:

- **vix.input.[N].recordreader** - list of recordreaders to use when processing this input, these RR are tried before those at the provider level. For example, ImageRecordReader – PCapRecordReader – ZipRecordReader – EncryptionRecordReader
- **vix.input.[N].splitter** – For example, ParquetSplitGenerator
- **vix.input.[N].required.fields** – For example, In smart mode always extract Timestamp field

.conf2015

Hunk Troubleshooting

splunk>

Troubleshooting Main Points

1. Hunk UI shows errors
2. Search.log to debug Hunk / Hadoop client issues
3. Hadoop logs to debug Hadoop Server issues
4. Job -> Inspect Job to debug many performance issues

Troubleshooting – Enable Debugging

Each log line in the file that involves Hunk ERP operations is annotated with **`ERP.<provider>`**… and contains links for spawned MR job(s). You may need to follow these links to troubleshoot MR issues.

To enable more detailed logging and monitoring flow modes, edit the following parameters in the provider setting:

By default, Hunk makes the best effort to prune unnecessary columns/fields to improve search performance.
For debugging, you can turn this off and have ERP return all columns to Hunk to do the filtering and final processing at the search head.

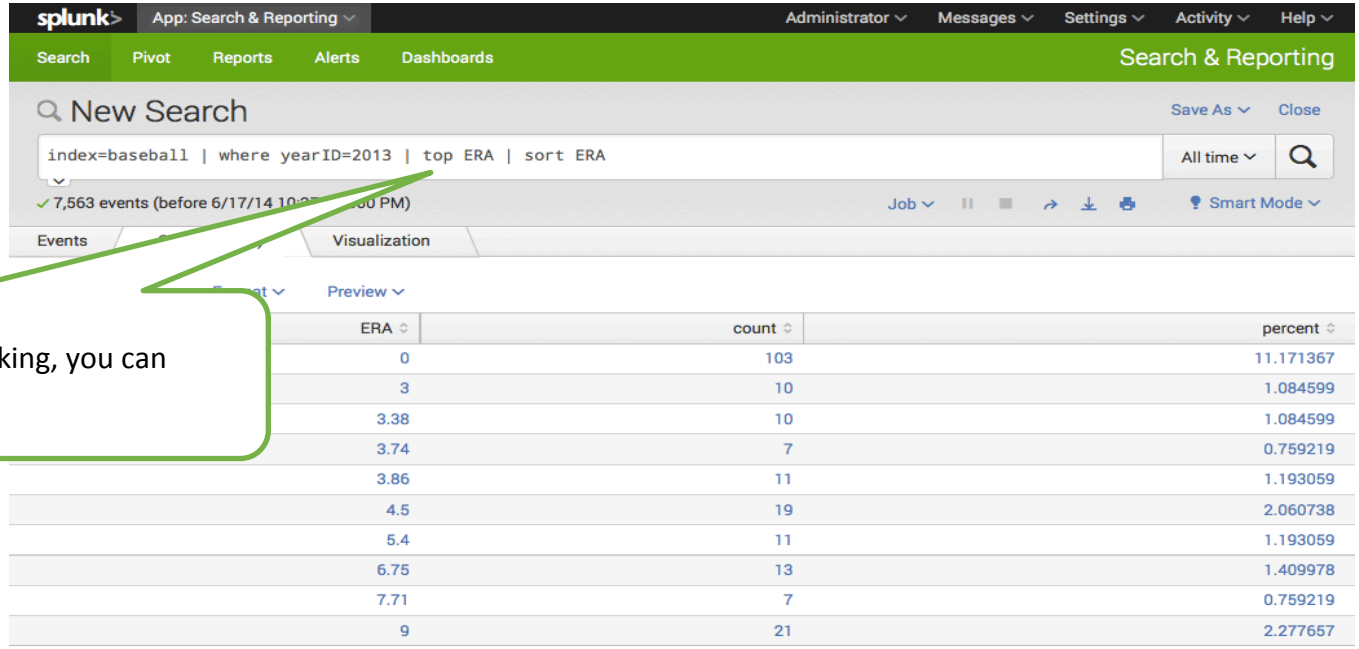
By setting to **1**, search.log will have DEBUG level logging events.

By default, Hunk searches run in mixed mode.
To disable, set the value to **0**.

vix.splunk.search.column.filter	1
vix.splunk.search.debug	1
vix.splunk.search.mixedmode	1

Example # 1, No MapReduce Job in Hadoop

Troubleshooting – No Map Reduce Job



To check if a MapReduce job is working, you can append a reporting search job.

Find search.log

New Search

Save As Close

index="pony" | stats count by customer.age

All time

684,084 of 684,084 events matched

Job Smart Mode

Events Statistics (53) Visualization

reduceSearch	nt by customer.age
remoteSearch	dex="pony" | addinfo type=count label=prereport_events |
reportSear	
request	atest_time s count by , 'preview
resultCoun	
resultIsStre	
resultPrevie	
runDuration	982.6
runtime	{'auto_pause': '0', 'auto_cancel': '30'}
scanCount	684084
search	search index="pony" | stats count by customer.age
searchProviders	['ERP:hadoop220', 'ip-10-167-137-149']
sid	1403047234.550
statusBuckets	0
	600
Additional info	search.log

1

In this example, a search returns some results but it seems like it is stuck after the initial streaming results. Just the fact that it has returned some result indicates that Hunk can access data in HDFS.

2

If you encounter an error while running a basic search, you can find a complete search job detail in the job inspector.

3

If you encounter issues while building your reports, **`search.log`** is the place to look. You can access the file via the job inspector.

In Search.log – Pinpoint the error

Hunk log lines are denoted with **ERP.** followed by a provider name. In this example, a job was submitted and Hunk is contacting ResourceManager (YARN).

```
10-166-41-118.ap-northeast-1.compute.internal_1403047234.550_0
06-17-2014 23:20:41.352 INFO  ERP.hadoop220 -  AsyncMRJob - AsyncMRJob job.name=SPLK_ip-10-166-41-118.ap-
northeast-1.compute.internal_1403047234.550_0 running ...
06-17-2014 23:20:41.352 INFO  ERP.hadoop220 -  AsyncMRJob - Submitting job.name=SPLK_ip-10-166-41-118.ap-
northeast-1.compute.internal_1403047234.550_0 ...
                INFO  SearchOperator:stdin - Initializing from configuration
                      ERP.hadoop220 -  RMProxy - Connecting to ResourceManager at

                O  PipelineComponent - registering timer callback name=triggerCollection
                f7ba085cf80
                NFO  LineBreakingProcessor - Initializing
                NFO  regexExtractionProcessor - Initializing
                NFO  PipelineComponent - Launching the pipelines.
                NFO  SearchOperator:stdin - setting up new preview state and writer ...
                NFO  ConfPathMapper - /home/splunker/hunk/etc/system/local: Skipping on-disk
                nges in memory only for write of: /nobody/system/props/source::/user/splunker
06-17-2014 23:20:41.662 INFO  ConfObjectManagerDB - /home/splunker/hunk/etc/system/metadata/local.meta:
Skipping flush, keeping changes in memory only
06-17-2014 23:20:41.663 INFO  pipeline - Registering metrics callback for: Pipeline:vix
06-17-2014 23:20:41.737 INFO  SearchParser - PARSING: typer | tags
06-17-2014 23:20:41.738 INFO  FastTyper - found nodes count: comparisons=6, unique_comparisons=5, terms=4,
unique_terms=4, phrases=12, unique_phrases=12, total leaves=22
06-17-2014 23:20:41.809 INFO  SearchOperator:stdin - setting _need_timestamp_fields=0, required time field
name=
06-17-2014 23:20:41.809 INFO  SearchOperator:stdin - required fields list =
Message,_time,customer.age,host,index,prestats_reserved_*,psrsvd_*,source,sourcetype
06-17-2014 23:20:41.823 INFO  UserManager - Setting user context: splunk-system-user
06-17-2014 23:20:41.823 INFO  UserManager - Done setting user context: NULL -> splunk-system-user
06-17-2014 23:20:41.823 INFO  UserManager - Unwound user context: splunk-system-user -> NULL
06-17-2014 23:20:41.823 INFO  SearchOperator:stdin - started writer thread, conf=source::/user/splunker
/data/Hunkdata.json.gz|host::ip-10-166-41-118|preprocess-gzip| ...
```

In Search.log – Pinpoint the error

```
06-17-2014 23:20:41.829 WARN  SearchOperator:kv - source is an indexed field, ignoring TOKENIZER
06-17-2014 23:20:41.829 WARN  SearchOperator:kv - sourcetype is an indexed field, ignoring TOKENIZER
06-17-2014 23:20:42.302 INFO  ERP.hadoop220 -  Client$Connection - Retrying connect to server:
localhost/127.0.0.1:8039. Already tried 0 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1 SECONDS)
06-17-2014 23:20:42.824 INFO  DispatchThread - Generating results preview took 1 ms
06-17-2014 23:20:43.458 INFO  ERP.hadoop220 -  Client$Connection - Retrying connect to server:
localhost/127.0.0.1:8039. Already tried 1 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1 SECONDS)
06-17-2014 23:20:44.090 INFO  DispatchThread - Generating results preview took 1 ms
06-17-2014 23:20:44.304 INFO  ERP.hadoop220 -  Client$Connection - Retrying connect to server:
localhost/127.0.0.1:8039. Already tried 2 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1 SECONDS)
06-17-2014 23:20:45.306 INFO  ERP.hadoop220 -  Client$Connection - Retrying connect to server:
localhost/127.0.0.1:8039. Already tried 3 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1 SECONDS)
06-17-2014 23:20:45.672 INFO  DispatchThread - Generating results preview took 1 ms
06-17-2014 23:20:46.306 INFO  ERP.hadoop220 -  Client$Connection - Retrying connect to server:
localhost/127.0.0.1:8039. Already tried 4 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1 SECONDS)
06-17-2014 23:20:47.583 INFO  DispatchThread - Generating results preview took 1 ms
06-17-2014 23:20:47.587 INFO  ERP.hadoop220 -  Client$Connection - Retrying connect to server:
localhost/127.0.0.1:8039. Already tried 5 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1 SECONDS)
06-17-2014 23:20:48.328 INFO  ERP.hadoop220 -  Client$Connection - Retrying connect to server:
localhost/127.0.0.1:8039. Already tried 6 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1 SECONDS)
06-17-2014 23:20:48.857 INFO  DispatchThread - Generating results preview took 1 ms
06-17-2014 23:20:49.383 INFO  ERP.hadoop220 -  Client$Connection - Retrying connect to server:
localhost/127.0.0.1:8039. Already tried 7 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1 SECONDS)
06-17-2014 23:20:50.122 INFO  DispatchThread - Generating results preview took 1 ms
```

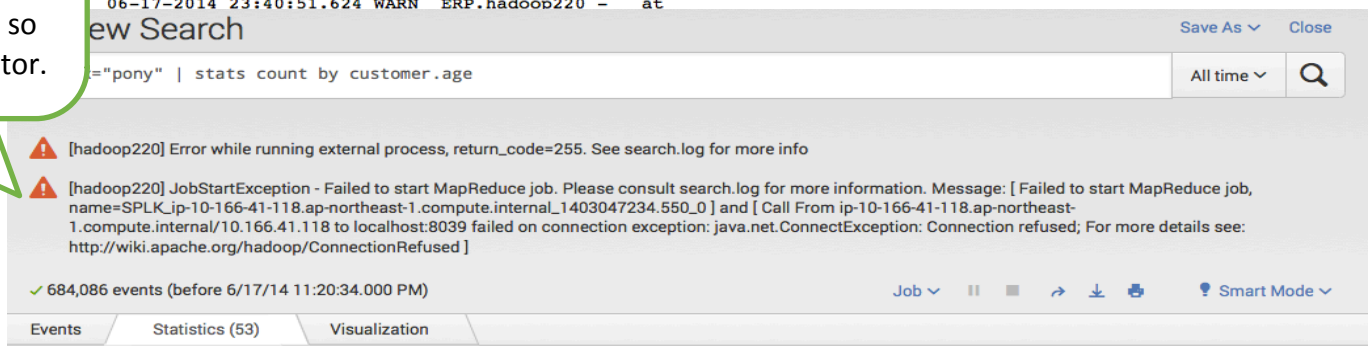
However, it looks like Hunk cannot connect to the ResourceManager.

Error will be display in UI and search.log

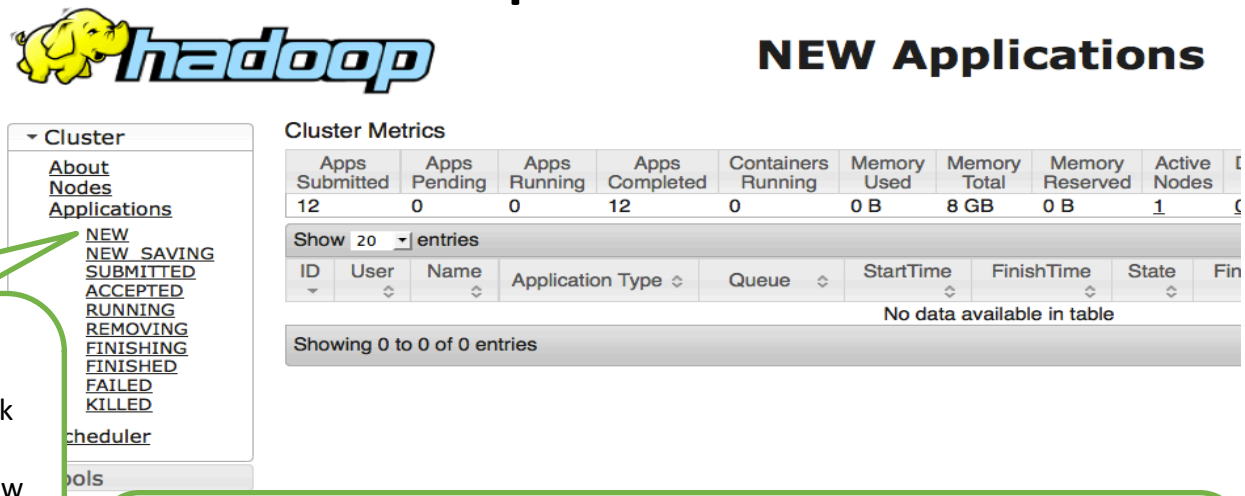
Eventually repeated attempts failed and the ERP throws an exception.

And the error message is shown on the partial results page indicating that the MapReduce job was unable to start. You suspect that maybe the ResourceManger node is down and so you contact the Hadoop administrator.

```
localhost/127.0.0.1:8039. Already tried 9 time(s); retry policy is
RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1 SECONDS)
06-17-2014 23:40:51.620 ERROR ERP.hadoop220 -  UserGroupInformation - PriviledgedActionException as:splunker
(auth:SIMPLE) cause:java.net.ConnectException: Call From ip-10-166-41-118.ap-northeast-
1.compute.internal/10.166.41.118 to localhost:8039 failed on connection exception: java.net.ConnectException:
Connection refused; For more details see:  http://wiki.apache.org/hadoop/ConnectionRefused
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -  ClusterInfoLogger - Exception thrown while logging cluster info
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -  java.net.ConnectException: Call From ip-10-166-41-118.ap-
northeast-1.compute.internal/10.166.41.118 to localhost:8039 failed on connection exception:
java.net.ConnectException: Connection refused; For more details see:  http://wiki.apache.org/hadoop
/ConnectionRefused
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at
sun.reflect.GeneratedConstructorAccessor5.newInstance(Unknown Source)
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at
java.lang.reflect.Constructor.newInstance(Constructor.java:526)
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at
org.apache.hadoop.net.NetUtils.wrapWithMessage(NetUtils.java:783)
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at
org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:730)
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at org.apache.hadoop.ipc.Client.call(Client.java:1351)
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at org.apache.hadoop.ipc.Client.call(Client.java:1300)
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at
org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:206)
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at com.sun.proxy.$Proxy14.getClusterMetrics(Unknown Source)
06-17-2014 23:40:51.624 WARN  ERP.hadoop220 -   at
```



Troubleshoot Hadoop Server issues



A Hadoop administrator checks the ResourceManager and finds that the node is running and no job from Hunk has been queued.
With that information, you can narrow down the issue to a network connection or a Hunk configuration error.

In this example, the culprit was misconfigured address to the ResourceManager.
After fixing the address, the job was able to complete successfully.
For more examples of error message, check:
http://docs.splunk.com/Documentation/Hunk/latest/Hunk/TroubleshootHunk

Example # 2, Real World - Bad Performance

No MapReduce Job = Not a Good start

	3.23	dispatch.fetch	21	-	-
	0.00	dispatch.localSearch	1	-	-
	0.00	dispatch.preview	3	-	-
	0.01	dispatch.readEventsInResults	1	-	-
	0.00	dispatch.stream.local	1	-	-
	0.85	dispatch.timeline	21	-	-
	0.02	dispatch.writeStatus	11	-	-
	1.60	erp.hortonworks.stream.bytes	7	239,522,304	17,790,151
	0.16	erp.hortonworks.vix.hadoop.dirs.listed	1	-	-
	0.16	erp.hortonworks.vix.hadoop.files.listed	-	-	-
	0.16	erp.hortonworks.vix.hadoop.splits.generation.time	1	-	-
	0.01	startup.configuration	1	-	-
	0.02	startup.handoff	1	-	-

Steam.bytes = Splunk generate results

Yes, MapReduce Job = Better

52.90	erp.hortonworks.MR	29	8	8
32.04	erp.hortonworks.MR.SPLK_ubuntu_1436386697.388_0	17	5	5
20.86	erp.hortonworks.MR.SPLK_ubuntu_1436386697.388_1	12	3	3
0.00	erp.hortonworks.MR.failed.tasks	2	-	-
0.00	erp.hortonworks.MR.failed.tasks.SPLK_ubuntu_1436386697.388_0	1	-	-
0.00	erp.hortonworks.MR.failed.tasks.SPLK_ubuntu_1436386697.388_1	1	-	-
52.56	erp.hortonworks.report.wait	9	-	-
22.26	erp.hortonworks.report.delay	1	-	-
0.08	erp.hortonworks.setup	1	-	-
0.02	erp.hortonworks.report.bytes	8	7,300	21,898
0.00	erp.hortonworks.setup.bundles	1	-	-
0.01	erp.hortonworks.setup.splunk	1	-	-
0.21	erp.hortonworks.vix.hadoop.dirs.listed	5	-	-
0.21	erp.hortonworks.vix.hadoop.files.listed	8	-	-
1.19	erp.hortonworks.vix.hadoop.splits.generation.time	2	-	-
0.22	startup.configuration	9	-	-
		9	-	-

report.bytes = Hadoop generate results
MR.SPLK = Leverage Hadoop

Examine HDFS Storage

52.90	erp.hortonworks.MR	29	8	8
32.04	erp.hortonworks.MR.SPLK_ubuntu_1436386697.388_0	17	5	5
20.86	erp.hortonworks.MR.SPLK_ubuntu_1436386697.388_1	12	3	3
0.00	erp.hortonworks.MR.failed.tasks	2	-	-
0.00	erp.hortonworks.MR.failed.tasks.SPLK_ubuntu_1436386697.388_0	1	-	-
0.00	erp.hortonworks.MR.failed.tasks.SPLK_ubuntu_1436386697.388_1	1	-	-
52.56	erp.hortonworks.report.wait	9	-	-
22.26	erp.hortonworks.report.delay	1	-	-
0.08	erp.hortonworks.setup	1	-	-
0.02	erp.hortonworks.report.bytes	8	7,300	21,898
0.00	erp.hortonworks.setup.bundles	1	-	-
0.01	erp.hortonworks.setup.splunk	1	-	-
0.21	erp.hortonworks.vix.hadoop.dirs.listed	5	-	-
0.21	erp.hortonworks.vix.hadoop.files.listed	8	-	-
1.19	erp.hortonworks.vix.hadoop.splits.generation.time	2	-	-
0.22	startup.configuration	9	-	-
		9	-	-

Hadoop.dirs / files .listed = How many directories Splunk need to scan

VIX with Timestamp on the files = Not great

1,140.31	erp.hunky_erp-1482.MR	10	72	72
678.77	erp.hunky_erp-1482.MR.SPLK_apatil-centos65x64-001_1435613899.334_0	2	40	40
461.54	erp.hunky_erp-1482.MR.SPLK_apatil-centos65x64-001_1435613899.334_1	8	32	32
0.00	erp.hunky_erp-1482.MR.failed.tasks	1	-	-
0.00	erp.hunky_erp-1482.MR.failed.tasks.SPLK_apatil-centos65x64-001_1435613899.334_1	1	-	-
657.74	erp.hunky_erp-1482.report.wait	7	-	-
579.74	erp.hunky_erp-1482.report.delay	1	-	-
6.55	erp.hunky_erp-1482.setup	1	-	-
0.09	erp.hunky_erp-1482.report.bytes	72	84,551	240,343
0.50	erp.hunky_erp-1482.setup.bundles	1	-	-
2.01	erp.hunky_erp-1482.setup.splunk	1	-	-
376.59	erp.hunky_erp-1482.vix.hunky_erp-1482.dirs.listed	365	-	-
0.00	erp.hunky_erp-1482.vix.hunky_erp-1482.files.filter.time	8,688	-	-
376.59	erp.hunky_erp-1482.vix.hunky_erp-1482.files.listed	8,760	-	-
570.39	erp.hunky_erp-1482.vix.hunky_erp-1482.splits.generation.time	223	-	-
1.73	startup.configuration	73	-	-
0.06	startup.handoff	73	-	-

Scan 8,760 files – filter out 8,688 = Only 72 files used for search
Recommendation is to build Timestamp on Directories

No-Splittable Very Large File = Bad

32.01	erp.hortonworks.MR	17	1	1
32.01	erp.hortonworks.MR.SPLK_ubuntu_1436227204.315_0	17	1	1
0.00	erp.hortonworks.MR.failed.tasks	1	-	-
0.00	erp.hortonworks.MR.failed.tasks.SPLK_ubuntu_1436227204.315_0	1	-	-
32.75	erp.hortonworks.report.delay	1	-	-
32.03	erp.hortonworks.report.wait	2	-	-
0.09	erp.hortonworks.setup	1	-	-
0.04	erp.hortonworks.report.bytes	1	39,306	135,058
0.00	erp.hortonworks.setup.bundles	1	-	-
0.01	erp.hortonworks.setup.splunk	1	-	-
0.15	erp.hortonworks.vix.orders.dirs.listed	-	-	-
0.15	erp.hortonworks.vix.orders.files.listed	1	-	-
0.23	erp.hortonworks.vix.orders.splits.generation.time	2	-	-
0.10	startup.configuration	2	-	-
0.02	startup.handoff	2	-	-

1 MR Job for very large file is not ideal

Yes-Splittable Very Large File = Good

140.96	erp.hortonworks.MR	76	20	20
39.99	erp.hortonworks.MR.SPLK_ubuntu_1436226266.312_0	21	5	5
34.95	erp.hortonworks.MR.SPLK_ubuntu_1436226266.312_2	19	5	5
33.02	erp.hortonworks.MR.SPLK_ubuntu_1436226266.312_1	18	5	5
33.00	erp.hortonworks.MR.SPLK_ubuntu_1436226266.312_3	18	5	5
0.00	erp.hortonworks.MR.failed.tasks	4	-	-
0.00	erp.hortonworks.MR.failed.tasks.SPLK_ubuntu_1436226266.312_0	1	-	-
0.00	erp.hortonworks.MR.failed.tasks.SPLK_ubuntu_1436226266.312_1	1	-	-
0.00	erp.hortonworks.MR.failed.tasks.SPLK_ubuntu_1436226266.312_2	1	-	-
0.00	erp.hortonworks.MR.failed.tasks.SPLK_ubuntu_1436226266.312_3	1	-	-
140.13	erp.hortonworks.report.wait	21	-	-
18.74	erp.hortonworks.report.delay	1	-	-
0.25	erp.hortonworks.report.bytes	20	639,038	2,435,196
0.08	erp.hortonworks.setup	1	-	-
0.00	erp.hortonworks.setup.bundles	1	-	-
		1	-	-

Multiple Jobs means we leverage Hadoop parallel system

Report Acceleration = Great

0.00	dispatch.localSearch	1	-	-
0.00	dispatch.preview	1	-	-
0.00	dispatch.preview.stats.execute_output	1	-	-
0.00	dispatch.preview.write_results_to_disk	1	-	-
0.00	dispatch.stream.local	1	-	-
0.01	dispatch.writeStatus	7	-	-
0.02	erp.hortonworks.cache.bytes	8	7,249	21,800
1.66	erp.hortonworks.cache.delay	1	-	-
1.00	erp.hortonworks.report.wait	2	-	-
0.14	erp.hortonworks.vix.hadoop.dirs.listed	5	-	-
0.14	erp.hortonworks.vix.hadoop.files.listed	8	-	-
0.34	erp.hortonworks.vix.hadoop.splits.generation.time	2	-	-
0.01	startup.configuration	1	-	-
0.02	startup.handoff	1	-	-

cache.bytes = HDFS results (No need for MR)

Summary

Summary - Performance

1. Run MR Jobs
2. HDFS Storage
3. VIX with Timestamp / indexes.conf
4. File Format
5. Compression types / File size
6. Event breaking / Props.conf
7. Report Acceleration
8. Hardware
9. Search Head Clustering
10. Many Other Flags (Threads, Splits)

Summary - Troubleshooting

1. Hunk UI shows errors
2. Search.log to debug Hunk / Hadoop client issues
3. Hadoop logs to debug Hadoop Server issues
4. Job -> Inspect Job to debug many performance issues

.conf2015

THANK YOU

splunk>

Common Issues We See

Issue	Clue for Issue	Potential Solution
Performance	Job takes a long time	Most likely customer is not running MR Jobs Change to index = xyz | stats count by xyz + smart mode
Memory	No Error! Job is just hanging ..	Lower vix.mapred.job.map.memory.mb = 1024 OR Increase the memory on the Hadoop side
Heartbeat	In the search.log you will see “operation took longer than the heartbeat interval”	vix.splunk.heartbeat = 0
Timestamp / Fields Extraction in Smart Mode	Events are not showing correctly	vix.input.[N].required.fields = Timestamp Or Props.conf
Hive Jars missing or Hive issues	In search.log you will see Exception in thread "main" java.lang.NoSuchFieldError	Add Jars to vix.env.HUNK_THIRDPARTY_JARS Or Look in answers for Hive
Data nodes /tmp directory will not install SplunkD	In Hadoop logs (not in Splunk logs) you will see permission or issues writing to / tmp/splunk	Change vix.splunk.home.hdfs Or Fix permission / size