© 2019 SPLUNK INC.

How to effectively run high cardinality and federated searches using Data Fabric Search

Nikhil Roy
Principal Software Engineer | Splunk

Asha Andrade
Principal Software Engineer | Splunk

splunk> .conf19

© 2019 SPLUNK INC.

Nikhil Roy
Principal Software Engineer | Splunk

Asha Andrade
Principal Software Engineer | Splunk

splunk> .conf19

© 2019 SPLUNK INC.

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or plans 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 may differ materially. 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, it may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements made herein.

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 functionalities described or to include any such feature or functionality in a future release.

Splunk, Splunk>, Turn Data Into Doing, 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. © 2019 Splunk Inc. All rights reserved.


WE'VE DECIDED TO TAKE BIG DATA TO THE NEXT LEVEL...
HUMONGOUS DATA
© D.Fletcher for CloudTweaks.com

© 2019 SPLUNK INC.

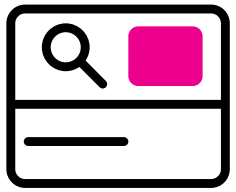
Data Fabric Search

Next Generation Search Platform

splunk> .conf19

© 2019 SPLUNK INC.

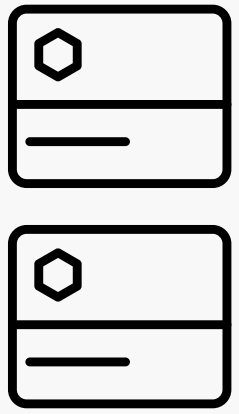
Splunk Infrastructure



Search Head

Request

Data Flow

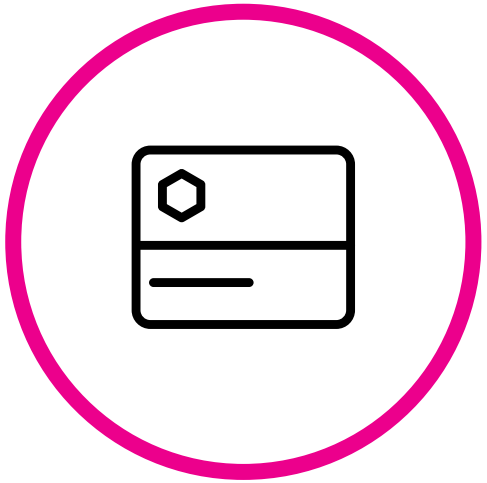


Search Peers

© 2019 SPLUNK INC.

Search Phases

Understanding how the search is orchestrated



#1 Search Peers (Indexers)

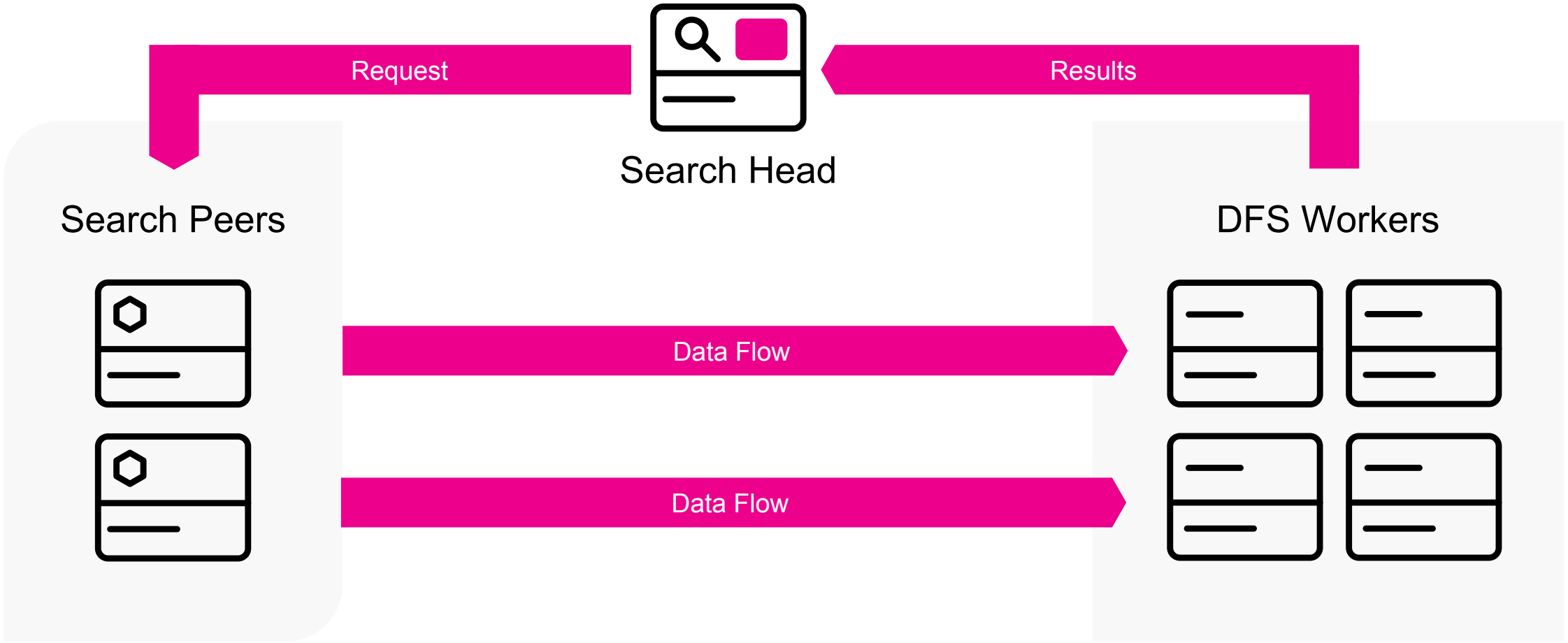
- Map Phase

#2 Splunk Search Head

- Reduce Phase
- Single point of reduction

© 2019 SPLUNK INC.

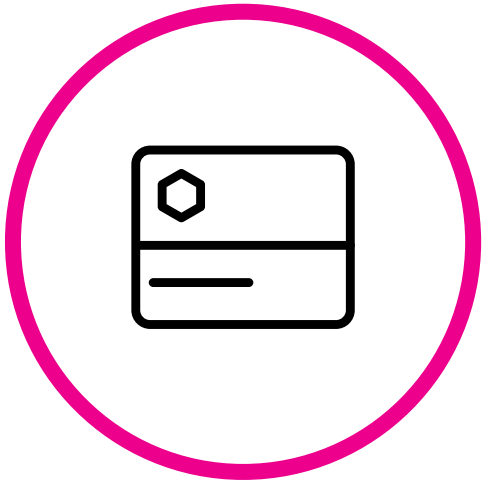
Data Fabric Infrastructure



© 2019 SPLUNK INC.

Search Phases with Data Fabric

Understanding how the search is orchestrated



#1 Search Peers (Indexers)

- Map Phase

#2 DFS Workers

- Distributed Reduce
- Number of workers can be scaled out

#3 Splunk Search Head

- Now responsible mostly for collection of results.

© 2019 SPLUNK INC.

© 2019 SPLUNK INC.

High Cardinality Searches

(fields with unique values)

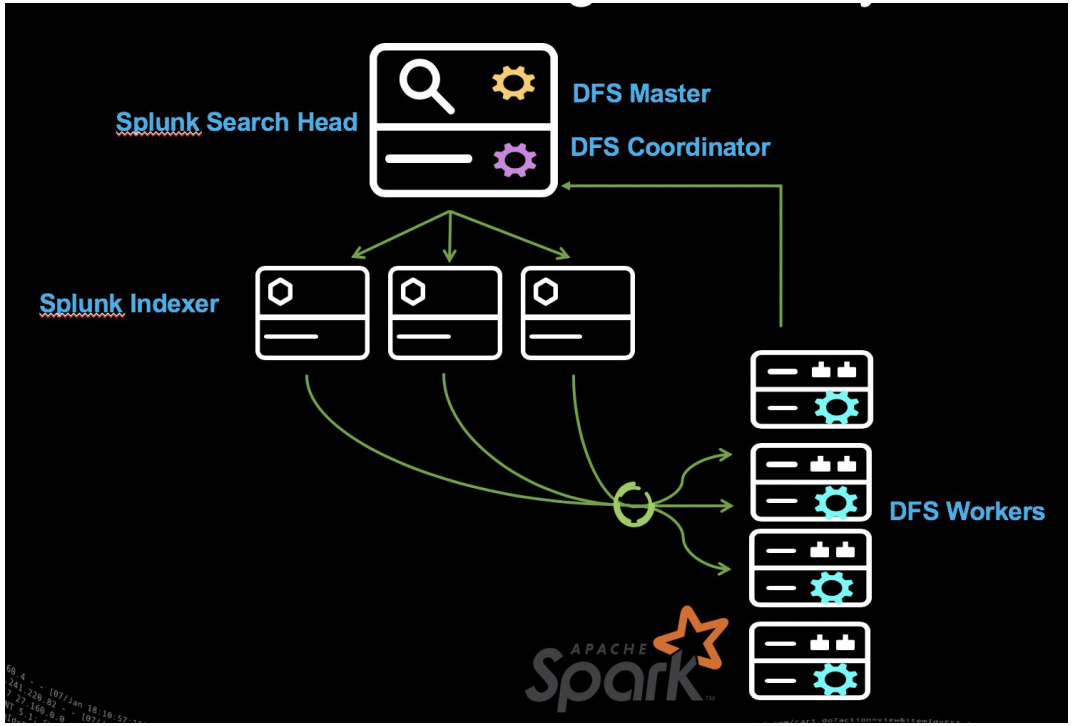
Searches which are aggregating voluminous data

- Volume > 100 Million events

Set of distinct field values which are in the order of millions

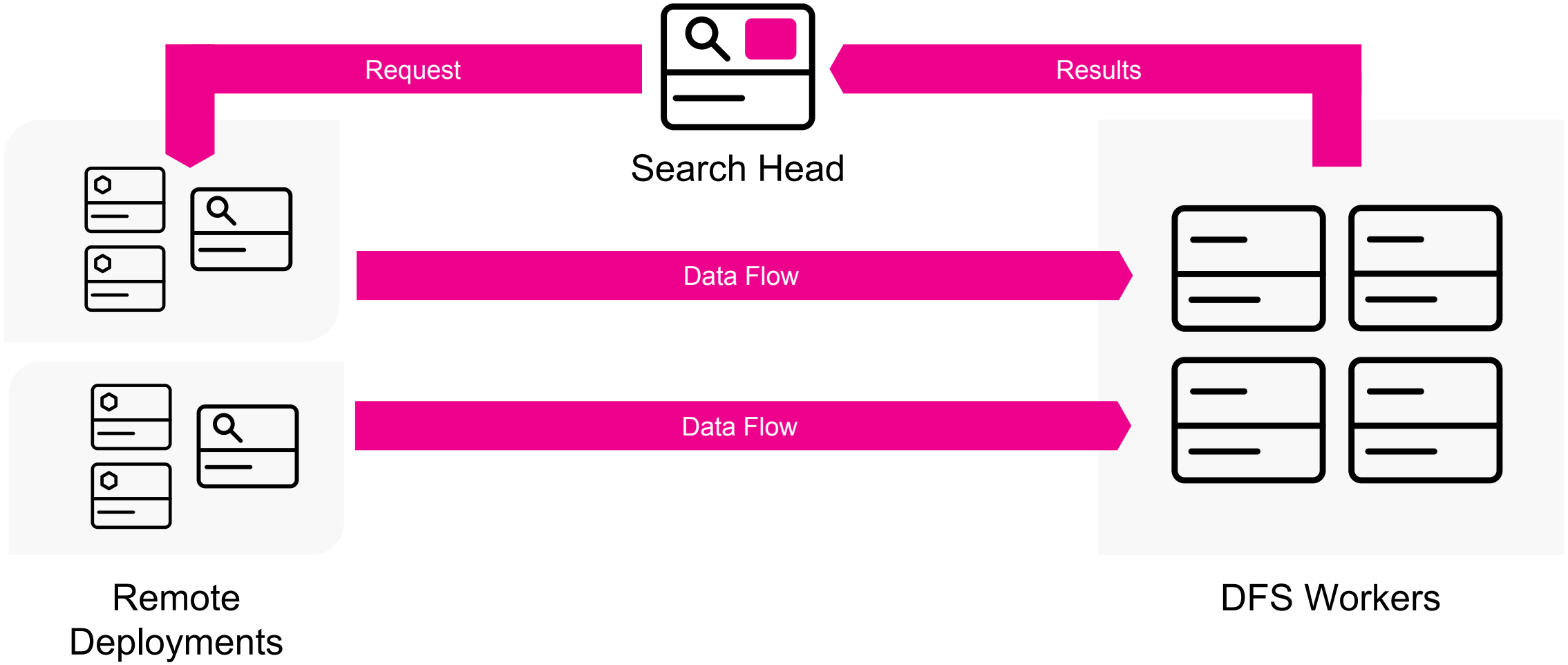
- Cardinality > 10 Million

Running these searches orders of magnitude faster



© 2019 SPLUNK INC.

Federate Using Data Fabric Search



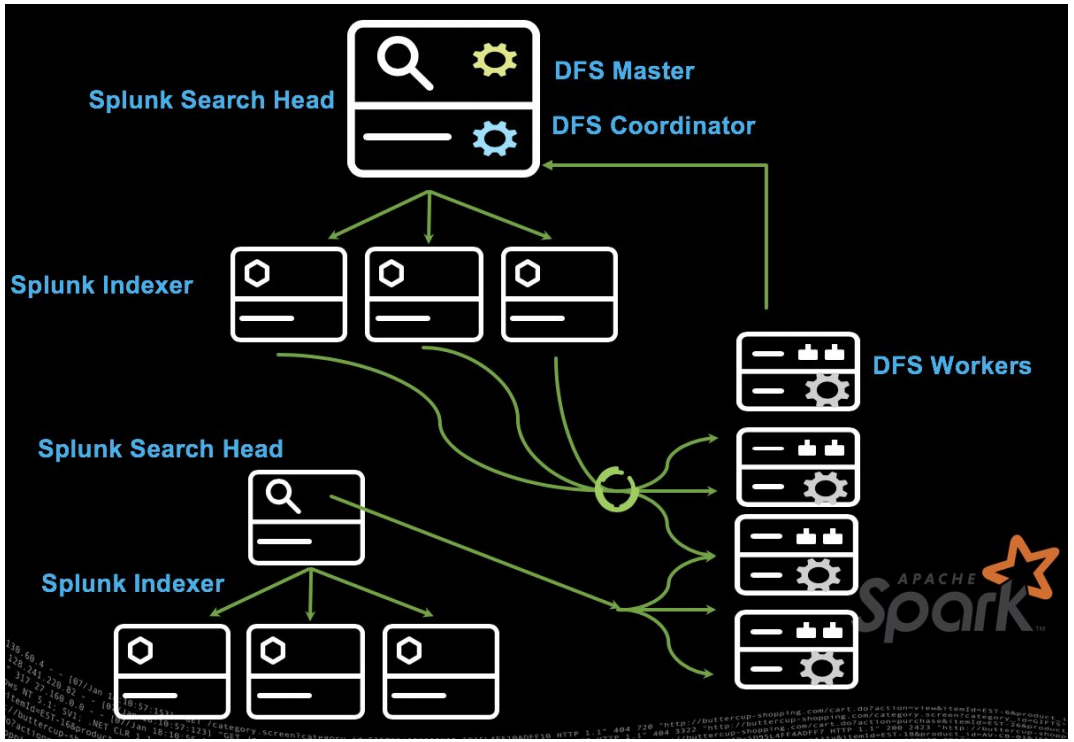
© 2019 SPLUNK INC.

Federated Searches

Searches which allow you to search across multiple Splunk deployments

Allows users to reuse knowledge objects on remote deployments

Allows users to have consistent data-access restrictions



© 2019 SPLUNK INC.

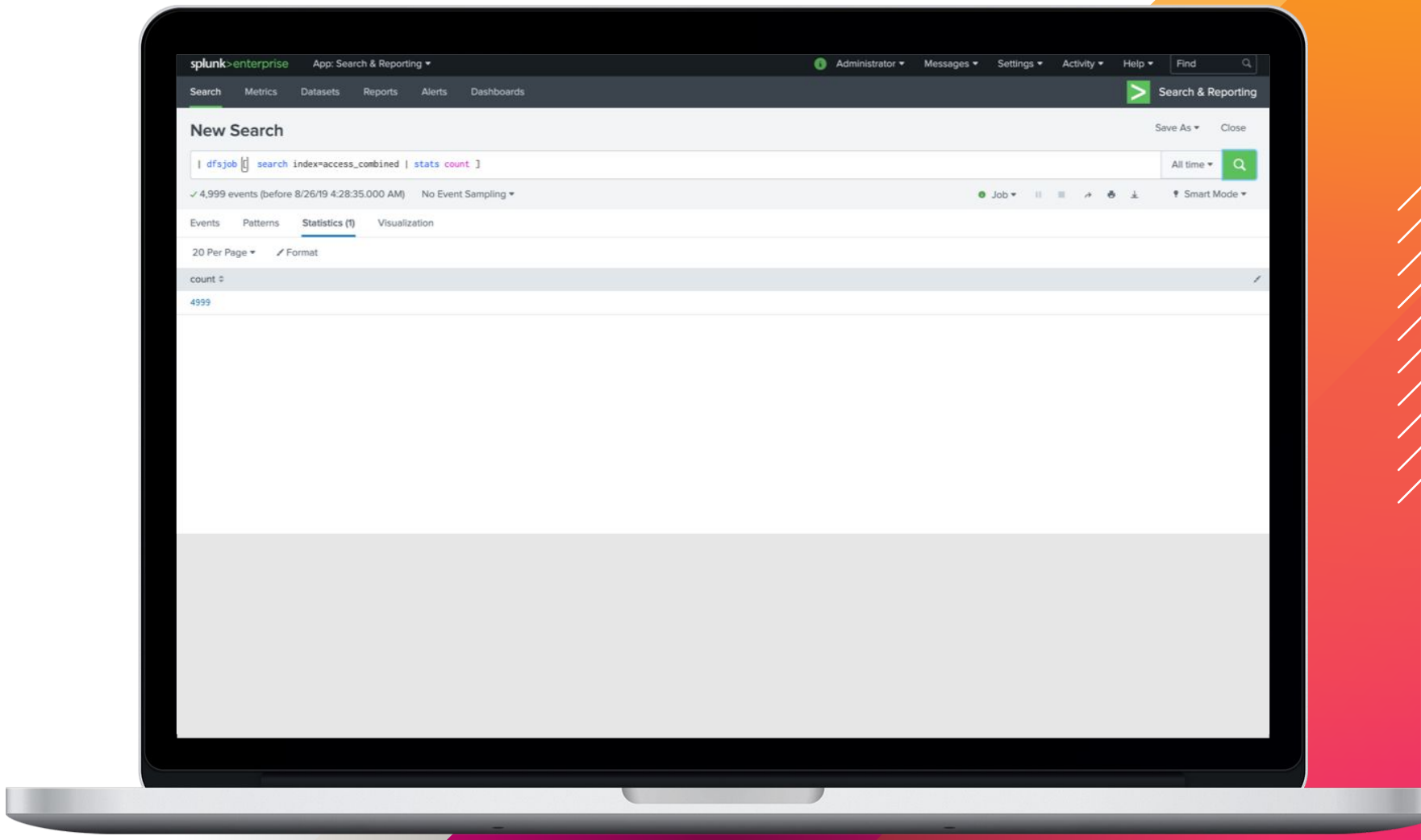
Splunking Your Data Assets

How to Search ?

splunk> .conf19

Data fabric search command

Allowing one to search using data fabric search infrastructure



Using dfsjob command

- | dfsjob [<insert-spl>]

© 2019 SPLUNK INC.

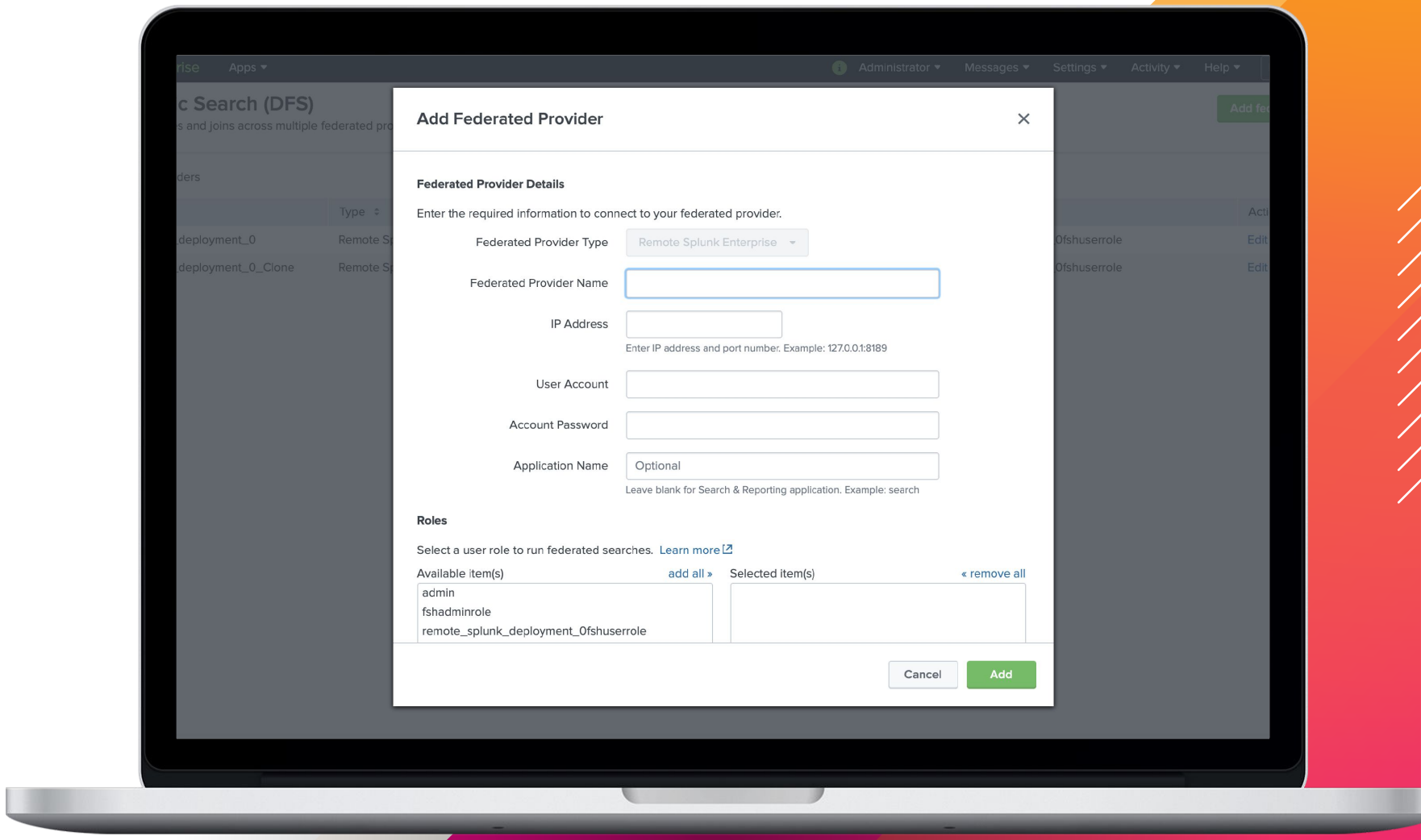
Federated Searches

Creating your provider

Federated providers refer to other Splunk deployments.

Configuration files

- federated.conf



© 2019 SPLUNK INC.

Federated Searches

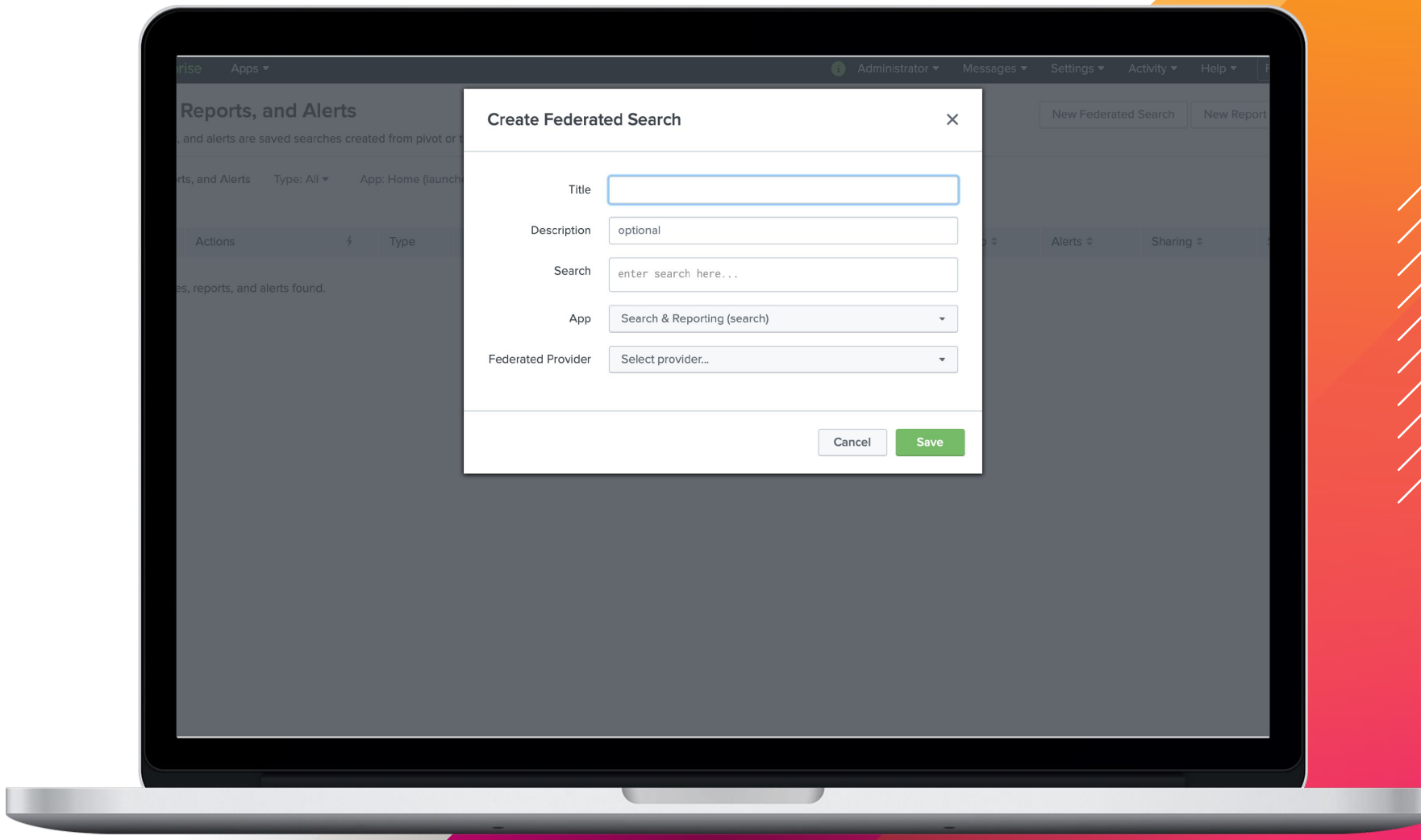
Creating your remote deployment search

Creating your federated searches and linking them to the appropriate remote deployment

Configuration files

- savedsearches.conf (searches)

Ensure you have access to data on remote deployment



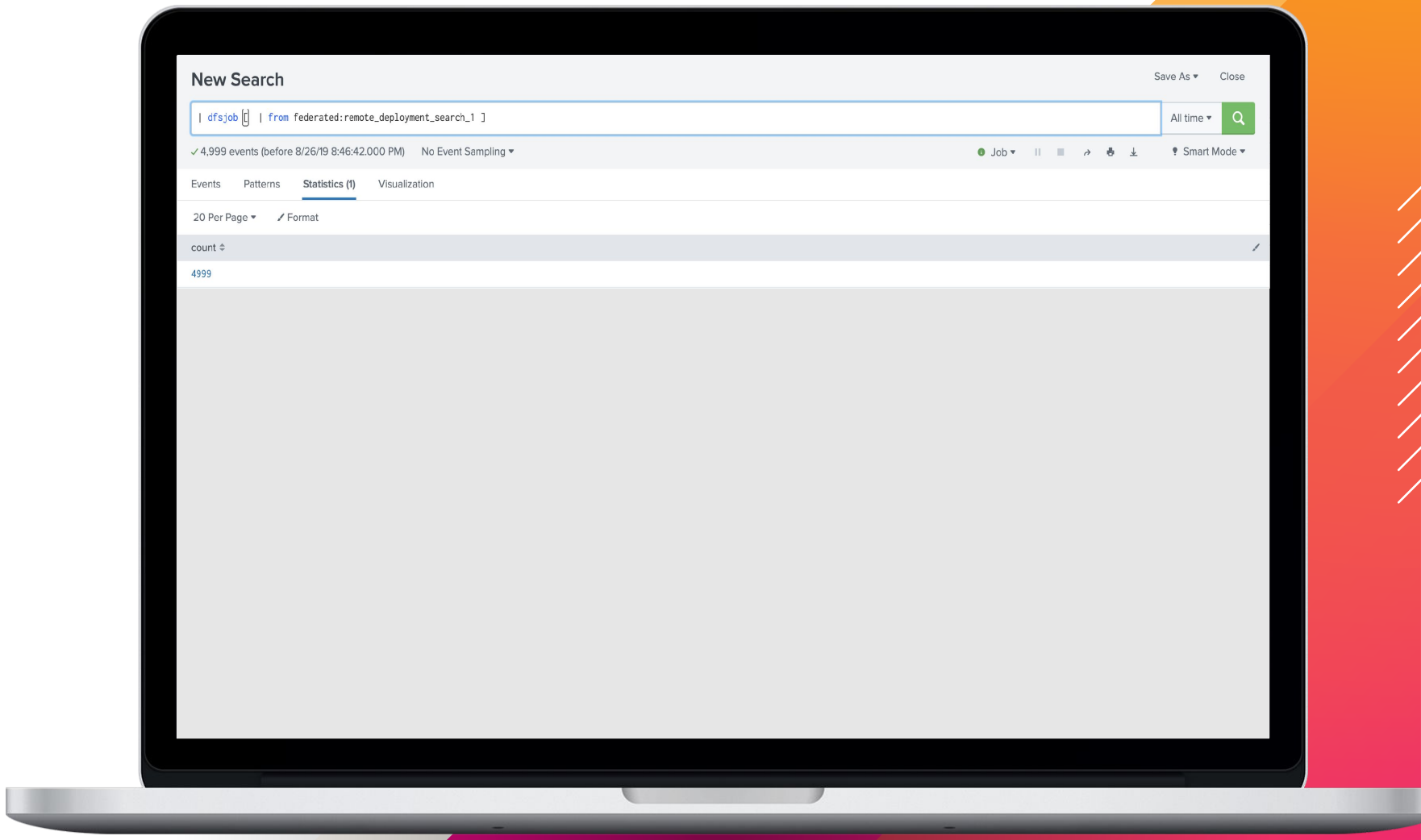
© 2019 SPLUNK INC.

Federate Searches

How to federate

Using federated saved seaches

- | dfsjob [| from federated:<insert_remote_federated_search_name>]



© 2019 SPLUNK INC.

© 2019 SPLUNK INC.

Splunk Data Fabric Search

What do we Currently Offer ?

splunk> .conf19

© 2019 SPLUNK INC.

DFS Supportability

Remote Deployment Variants

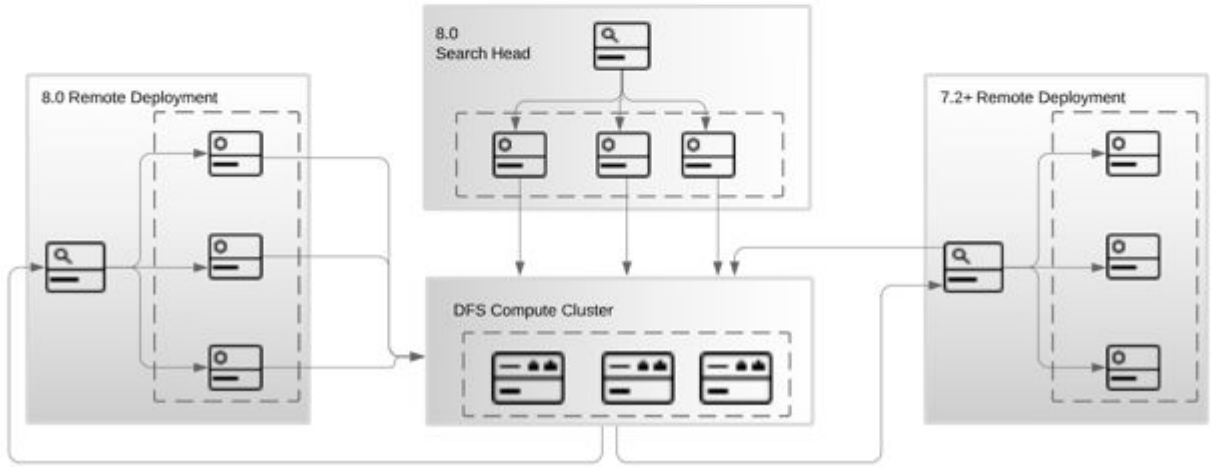
8.0 Remote Deployments

- Search at scale across multiple deployments
- Indexers distribute data directly to the DFS workers

7.2+ Remote Deployments

- Perform complex correlations across data from Remote Deployments
- Search Head will distribute data to the DFS workers

Note: Switch to 7.2+ with executionMode=sh in federated.conf



© 2019 SPLUNK INC.

SPL Support

DFS supports only **Transforming** Searches

- A search that uses transforming commands like stats to transform event data returned by a search into statistical tables that can be used as the basis for charts and other kinds of data visualizations.

An SPL in DFS is composed of three phases

- Map phase – Indexers
- Reduce Phase - Scalable DFS Compute Cluster NEW
  - Supported commands - dedup, eval, fields, head, join, rename, reverse, sort, stats, tail, union, where, from
- Final Reduce Phase – Search Head

© 2019 SPLUNK INC.

SPL Support

| dfsjob [| <map-phase> | <reduce-phase>] | <sh-phase>

| <map-phase> `comment("High Volume (exceeds 100M), High Cardinality (exceeds 10M)")`
| <(stats | join | union)>
| (dedup | eval | fields | head | join | rename | reverse | sort | stats | tail | union | where)*
| <sh-phase>

Note: <term>. is required
(<term>)* is optional and repeated 0 or more times

© 2019 SPLUNK INC.

High Cardinality: Stats

```
| dfsjob [ | search sourcetype="websense::cg::kv"
      | stats avg(bytes_in) as avg_bytes_in by url `comment("url with high distinct value")`
      | sort – avg_bytes_in
      | head 1000]
```

Map Phase

```
      | search sourcetype="websense::cg::kv"
      | addinfo type=count label=prereport_events
      | fields keepcolorder=t "bytes_in" "prestats_reserved_*" "psrsvd_*" "url"
      | prestats mean(bytes_in) by url | …
```

DFS Phase NEW

```
  … | stats allnum=false delim=" " partitions=1 avg(bytes) AS avg_bytes BY clientip
      | sort avg_bytes
      | head limit=1000
```

© 2019 SPLUNK INC.

Job Inspector

Information about various phases

dfs


```
{ [-]
   column_order: [ [-]
     count
   ]
   reduce_search: | rdin remote_ip-10-224-23-8.us-west-2.compute.internal_prd.ph1_1566845340.13 pushmode=true | stats allnum=false delim=" " partitions=1 count
   remote_search: [ [-]
     search index=access_combined | addinfo type=count label=prereport_events | fields keepcolorder=true "prestats_reserved_*", "psrsvd_*" | prestats allnum=false delim="
" count | rdout partition_method="vpc" workers="10.224.118.27:17501,10.224.163.171:17500,10.224.118.27:17500,10.224.159.220:17500"
   ]
   remote_sid: [ [-]
     prd.ph1_1566845340.13
   ]
   search: | dfsjob [| search index=access_combined | stats count]
}
```

© 2019 SPLUNK INC.

Stats in DFS

search	scanCount	doneProgress	resultCount	runDuration
search index=airlinedata | stats count by FlightNum Origin Dest | stats count	9941683	1.00000	1	5.307
| dfsjob [search index=airlinedata | stats count by FlightNum Origin Dest | stats count]	9941683	1.00000	1	9.617
search index=airlinedata | stats count by FlightNum Origin Dest ArrDelay ArrTime | stats count	9941683	1.00000	1	19.438
| dfsjob [search index=airlinedata | stats count by FlightNum Origin Dest ArrDelay ArrTime | stats count]	9941683	1.00000	1	14.996
search index=airlinedata | stats count by FlightNum Origin Dest ArrDelay ArrTime AirTime CarrierDelay Distance | stats count	9941683	1.00000	1	28.687
| dfsjob [search index=airlinedata | stats count by FlightNum Origin Dest ArrDelay ArrTime AirTime CarrierDelay Distance | stats count]	9941683	1.00000	1	17.864

© 2019 SPLUNK INC.

High Cardinality: Join

```
| dfsjob [ | search index=dogfood2007
           | stats count(clientip) as clientipcnt by clients, source
           | join usetime=f left=LHS right=RHS where LHS.clientipcnt = RHS.clientipcnt
               [ | search index=dogfood2007
                 | stats count(clientip) as clientipcnt by clientip, sourcetype
               ] `comment("No subsearch limits!")`
         ]
Map Phase(s) – Executed in parallel
    | search  index=dogfood2007 | addinfo … | fields … | prestats count(clientip) AS clientipcnt BY clients, source | …
    | search  index=dogfood2007 | addinfo … | fields … | prestats count(clientip) AS clientipcnt BY clients sourcetype |
  …
DFS Phase
    … | stats allnum=false delim=" " partitions=1 count(clientip) AS clientipcnt BY clients, source
      | join usetime=false left=LHS right=RHS where LHS.clientipcnt=RHS.clientipcnt
        [ … | stats allnum=false delim=" " partitions=1 count(clientip) AS clientipcnt BY clients, sourcetype]
```

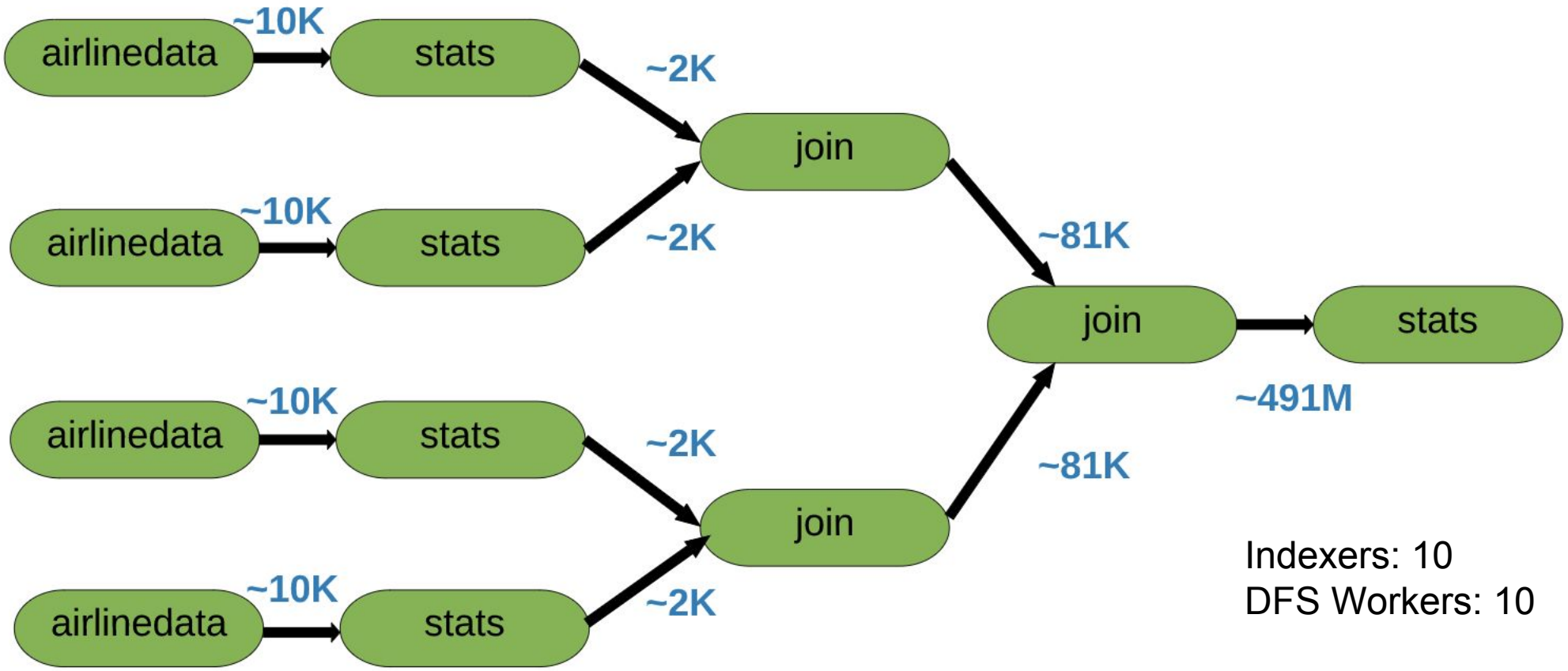
NEW

NEW

splunk> .conf19

© 2019 SPLUNK INC.

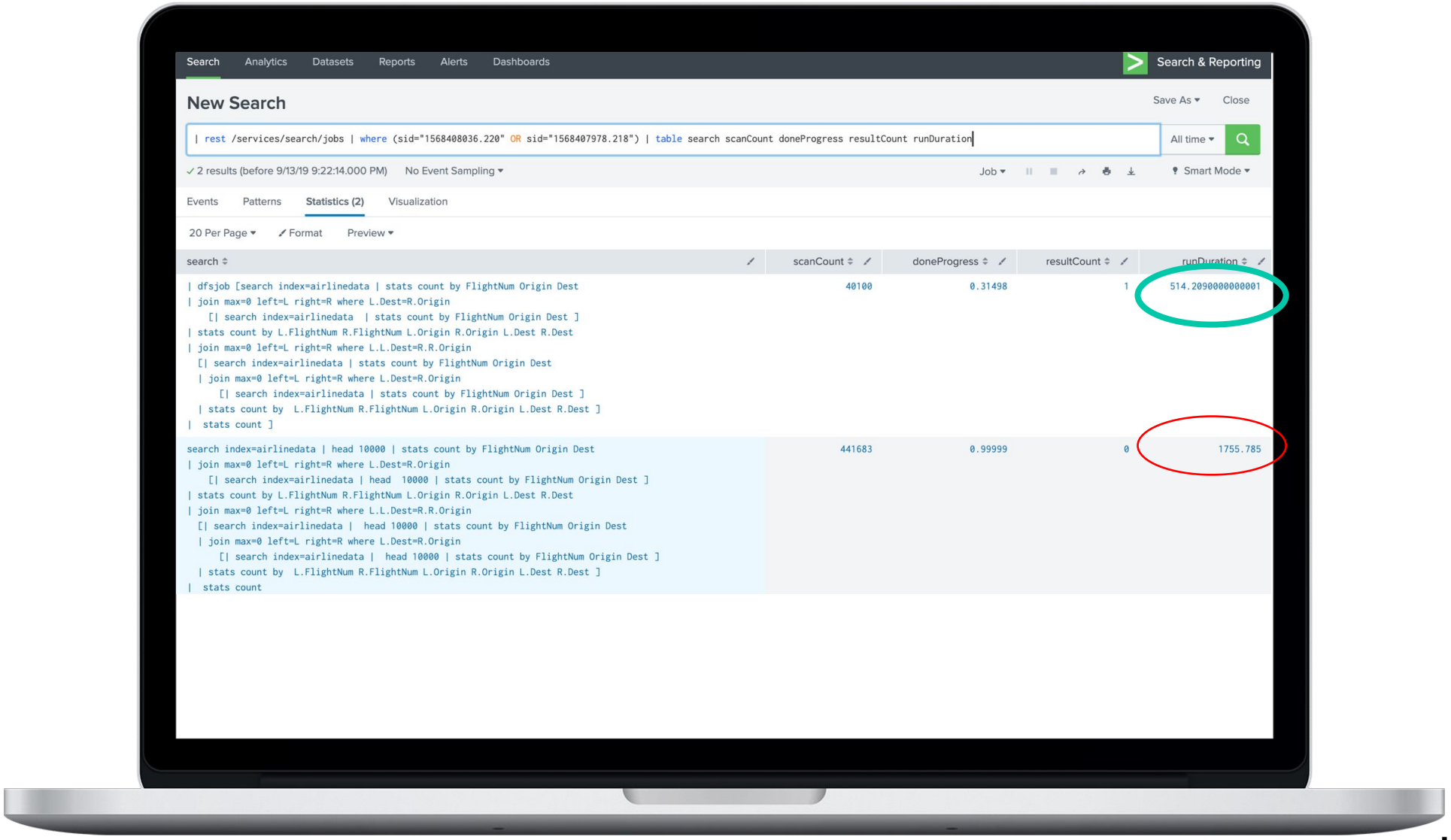
Join in DFS



Indexers: 10
DFS Workers: 10

© 2019 SPLUNK INC.

Join in DFS



© 2019 SPLUNK INC.

© 2019 SPLUNK INC.

Federated Search: Union, Join, Stats

```
| dfsjob [ | union [ | from federated:search_on_deployment_A ]
                     [ | from federated:search_on_deployment_B ]
                     [ | from federated:search_on_deployment_C ]
                     `comment("| search index=dogfood | stats count(clientip) as clientipcnt by clients, source")`
          | stats count by clients
          | join usetime=f left=L right=R where L.clients = R.clients
                [ | union [ | from federated:search_on_deployment_AA ]
                            [ | from federated:search_on_deployment_BB ]
                            [ | from federated:search_on_deployment_CC ]
                            `comment("| search index=dogfood | stats count(clientip) as clientipcnt by clients, sourcetype")`
                  | stats count by clients ]
          | sort -L.clients
          | head 100 ]
```

Join of a reduced union of three Splunk deployments, followed by a sort and head

© 2019 SPLUNK INC.

Debugging: Job Inspector

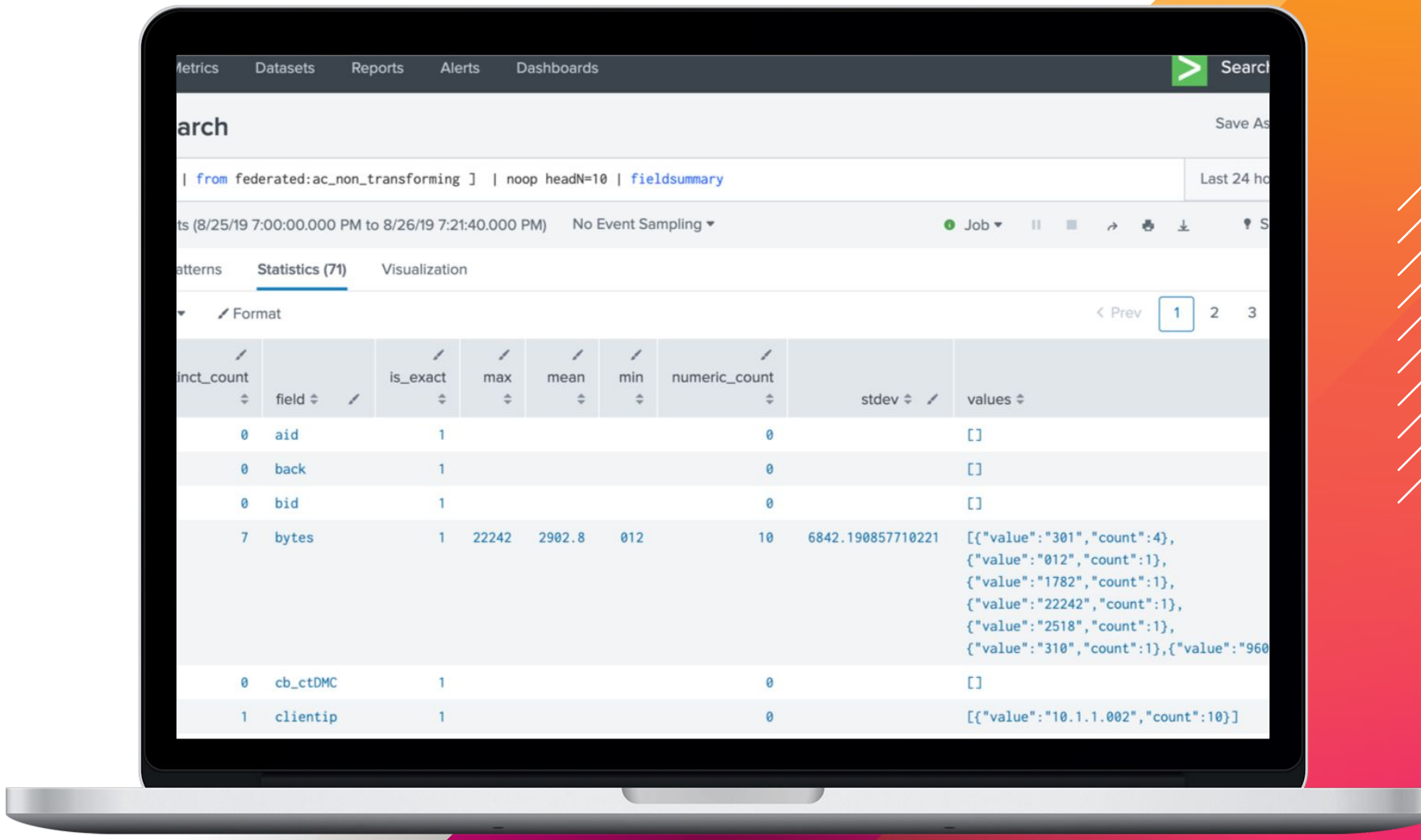
Information about various phases in a federated search

dfs	{ [-] column_order: [[-] clientip count] reduce_search: | rdin remote_ip-10-224-23-8.us-west-2.compute.internal_fshScaled_prd.ph1_2_1566846368.40 pushmode=true | stats allnum=false delim=" " partitions=1 count BY clientip search: | dfsjob [| from federated:ac] }
dfsFshMetrics	[[-] Remote Deployment Info: deploymentName=remote_splunk_deployment_0 datasetName=ac remoteSearch=search index=access_combined | stats count by clientip splunk.serviceAccount=remotefshuser remoteSid=fshScaled_2_1566846368.40 remoteEvent=4999 remoteScanCount=4999 duration=1.456000 FSH Total Counts: totalEventCount=4999 totalScanCount=4999]

Debugging: Federated Remote Deployment

DFS only supports Transforming searches by default.

However, you can use the headN option with the noop command in streaming and non-reporting federated searches to retrieve a smaller number of events from a search



© 2019 SPLUNK INC.

© 2019 SPLUNK INC.

© 2019 SPLUNK INC.

Key Takeaways

1. Use DFS when search head becomes a bottleneck for processing.
2. Use DFS when data processing volumes exceed 100 MM (and/or) 10 MM cardinality.
3. Use DFS when sub-search limits have been reached
4. Use DFS to federate searches across multiple deployments to perform joins and other correlations.

© 2019 SPLUNK INC.

DFS Relevant Conf Sessions and Links

- FN2124 - Data Fabric Search (DFS) - Under the Hood
- FN1727 - Tailoring Your Data Fabric to Custom Fit Your (SOCs/NOCs) Data Needs: Data Fabric Search Best Practices, Configuration, and Troubleshooting
- FN2276 – DFS Use Cases – Real World Applications
- FN2030 - Data Fabric Search : Opening doors to unprecedented levels of scale and performance
- DFS Docs: https://docs.splunk.com/Documentation/DFS/1.0.0/DFS/Overview

© 2019 SPLUNK INC.

.conf19

splunk>

Thank You !

Go to the .conf19 mobile app to

RATE THIS SESSION