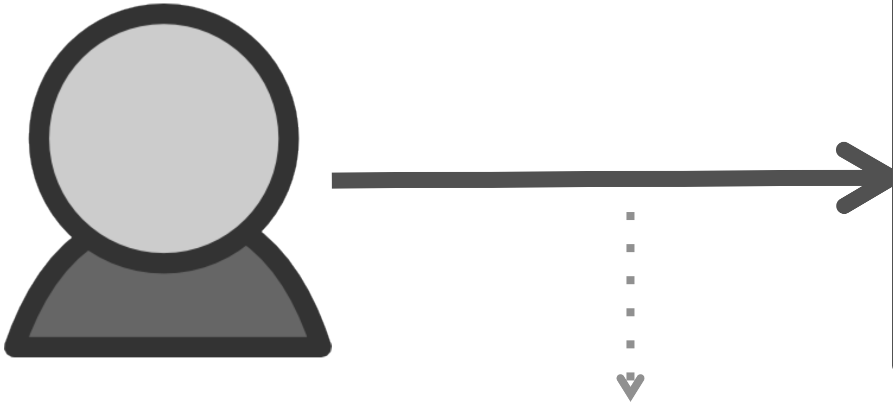
Analyzing log analysis: an empirical study of user log mining

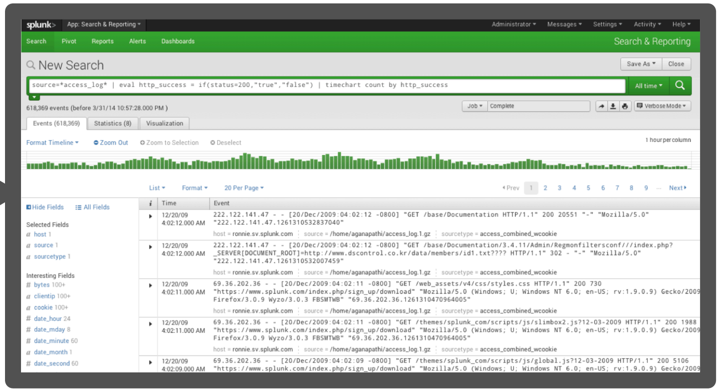
Sara Alspaugh
@salspaugh
AMPLab, EECS Department, UC Berkeley

in collaboration with:
Archana Ganapathi (Splunk),
Beidi Chen, Jessica Lin, Marti Hearst, Randy Katz (UC Berkeley)

USENIX LISA 2014

Data collected





```
source=*access_log* | eval http_success = if(status=200,"true","false") | timechart count by http_success
```

618,369 events (before 3/31/14 10:57:28.000 PM)

```
12/20/09 4:02:12.000 AM  222.122.141.47 - - [20/Dec/2009:04:02:12 -0800] "GET /base/Documentation HTTP/1.1" 200 20551 "-" "Mozilla/5.0"
                         "222.122.141.47.1261310532837040"
                         host = ronnie.sv.splunk.com | source = /home/aganapathi/access_log.1.gz | sourcetype = access_combined_wcookie
12/20/09 4:02:12.000 AM  222.122.141.47 - - [20/Dec/2009:04:02:12 -0800] "GET /base/Documentation/3.4.11/Admin/Regmonfiltersconf///index.php?
                         _SERVER[DOCUMENT_ROOT]=http://www.dscontrol.co.kr/data/members/id1.txt???? HTTP/1.1" 302 - "-" "Mozilla/5.0"
                         "222.122.141.47.1261310532007459"
                         host = ronnie.sv.splunk.com | source = /home/aganapathi/access_log.1.gz | sourcetype = access_combined_wcookie
12/20/09 4:02:11.000 AM  69.36.202.36 - - [20/Dec/2009:04:02:11 -0800] "GET /web_assets/v4/css/styles.css HTTP/1.1" 200 730
                         "https://www.splunk.com/index.php/sign_up/download" "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.9) Gecko/2009
                         Firefox/3.0.9 Wyzo/3.0.3 FBSMTWB" "69.36.202.36.1261310470964005"
                         host = ronnie.sv.splunk.com | source = /home/aganapathi/access_log.1.gz | sourcetype = access_combined_wcookie
12/20/09 4:02:11.000 AM  69.36.202.36 - - [20/Dec/2009:04:02:11 -0800] "GET /themes/splunk_com/scripts/js/slimbox2.js?12-03-2009 HTTP/1.1" 200 1988
                         "https://www.splunk.com/index.php/sign_up/download" "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.9) Gecko/2009
                         Firefox/3.0.9 Wyzo/3.0.3 FBSMTWB" "69.36.202.36.1261310470964005"
                         host = ronnie.sv.splunk.com | source = /home/aganapathi/access_log.1.gz | sourcetype = access_combined_wcookie
12/20/09 4:02:09.000 AM  69.36.202.36 - - [20/Dec/2009:04:02:09 -0800] "GET /themes/splunk_com/scripts/js/global.js?12-03-2009 HTTP/1.1" 200 5106
                         "https://www.splunk.com/index.php/sign_up/download" "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.9) Gecko/2009
```

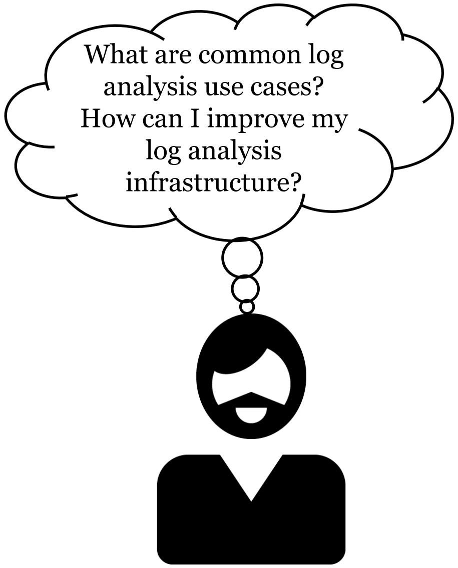
User

Log analysis tool

Queries

```
search index=os eventtype=linux-password-change-failed
search index=os eventtype=“Failed_SU” index=“os” sourcetype=“interfaces” host=* | multikv
fields name, inetAddr, RXbytes, TXbytes | streamstats current=f last(TXbytes) as lastTX,
last(RXbytes) as lastRX by Name  | eval time=_time | strcat Name “-” inetAddr “@” host
Interface_Host | eval RX_Thruput_KB = (lastRX-RXbytes)/1024 | eval TX_Thruput_KB =
(lastTX-TXbytes)/1024 | timechart eval(sum(TX_Thruput_KB)/dc(time)) by Interface_Host
search index=os sourcetype=openPorts | MULTIKV | STATS count BY Port | SORT count
search index=os source=ps | multikv | timechart avg(VSZ_KB) by USER useother=F limit=10
“” | strcat source “@” host changelist | timechart count by changelist
search sourcetype=syslog error OR failed OR severe NOT assignment starthoursago=1 |
fields +_raw
search index=os source=ps | multikv | timechart avg(RSZ_KB) by COMMAND
search index=os source=iostat | multikv | timechart avg(rReq_PS) avg(wReq_PS)
search index=os source=lsof | multikv | timechart count(USER) by USER
search index=os source=vmstat | multikv | timechart avg(memTotalMB) by host
```

Motivation

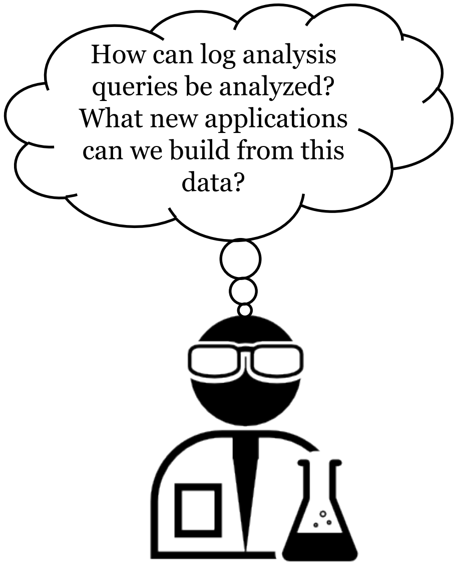


What are common log analysis use cases?
How can I improve my log analysis infrastructure?

log analysis practitioners

How can we be more relevant to our customers?
How can we improve customer experience?

data analysis and management companies

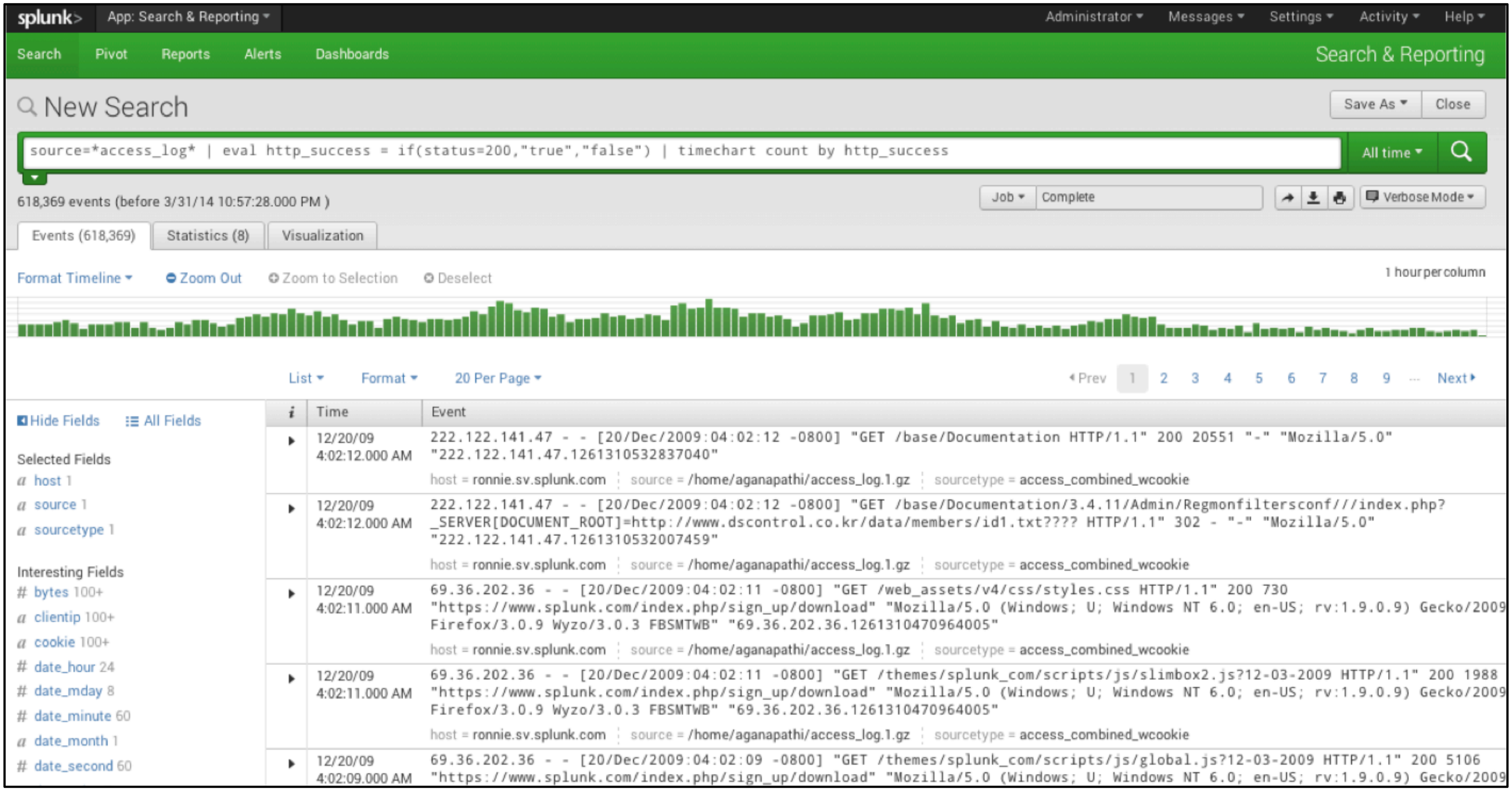


How can log analysis queries be analyzed?
What new applications can we build from this data?

computer science researchers

What do people use a highly popular commercial purpose-built log analysis tool to do?

Splunk screenshot



```
source=*access_log* | eval http_success = if(status=200,"true","false") | timechart count by http_success
```

618,369 events (before 3/31/14 10:57:28.000 PM)

Selected Fields

- a host 1
- a source 1
- a sourcetype 1

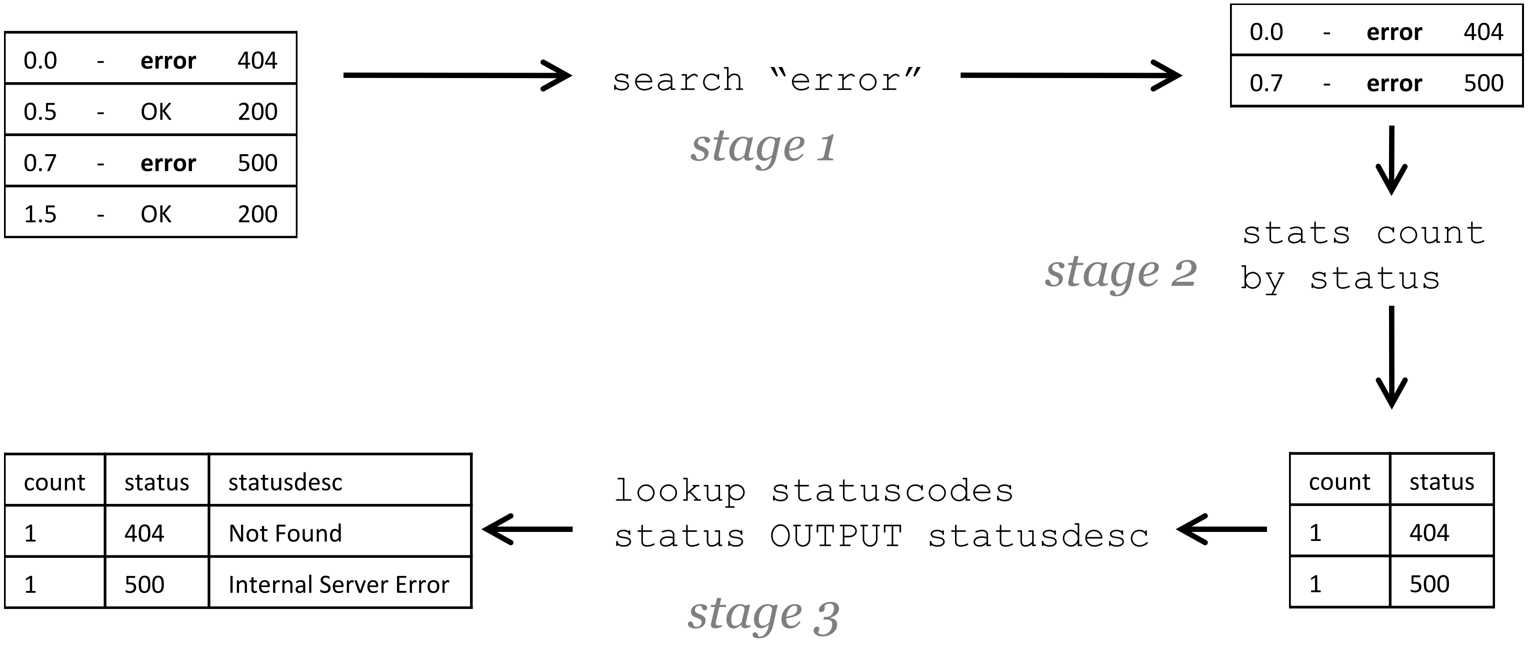
Interesting Fields

- # bytes 100+
- a clientip 100+
- a cookie 100+
- # date_hour 24
- # date_mday 8
- # date_minute 60
- a date_month 1
- # date_second 60

i	Time	Event
▸	12/20/09 4:02:12.000 AM	222.122.141.47 - - [20/Dec/2009:04:02:12 -0800] "GET /base/Documentation HTTP/1.1" 200 20551 "-" "Mozilla/5.0" "222.122.141.47.1261310532837040" host = ronnie.sv.splunk.com ┆ source = /home/aganapathi/access_log.1.gz ┆ sourcetype = access_combined_wcookie
▸	12/20/09 4:02:12.000 AM	222.122.141.47 - - [20/Dec/2009:04:02:12 -0800] "GET /base/Documentation/3.4.11/Admin/Regmonfiltersconf///index.php?_SERVER[DOCUMENT_ROOT]=http://www.dscontrol.co.kr/data/members/id1.txt???? HTTP/1.1" 302 - "-" "Mozilla/5.0" "222.122.141.47.1261310532007459" host = ronnie.sv.splunk.com ┆ source = /home/aganapathi/access_log.1.gz ┆ sourcetype = access_combined_wcookie
▸	12/20/09 4:02:11.000 AM	69.36.202.36 - - [20/Dec/2009:04:02:11 -0800] "GET /web_assets/v4/css/styles.css HTTP/1.1" 200 730 "https://www.splunk.com/index.php/sign_up/download" "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.9) Gecko/2009 Firefox/3.0.9 Wyzo/3.0.3 FBSMTWB" "69.36.202.36.1261310470964005" host = ronnie.sv.splunk.com ┆ source = /home/aganapathi/access_log.1.gz ┆ sourcetype = access_combined_wcookie
▸	12/20/09 4:02:11.000 AM	69.36.202.36 - - [20/Dec/2009:04:02:11 -0800] "GET /themes/splunk_com/scripts/js/slimbox2.js?12-03-2009 HTTP/1.1" 200 1988 "https://www.splunk.com/index.php/sign_up/download" "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.9) Gecko/2009 Firefox/3.0.9 Wyzo/3.0.3 FBSMTWB" "69.36.202.36.1261310470964005" host = ronnie.sv.splunk.com ┆ source = /home/aganapathi/access_log.1.gz ┆ sourcetype = access_combined_wcookie
▸	12/20/09 4:02:09.000 AM	69.36.202.36 - - [20/Dec/2009:04:02:09 -0800] "GET /themes/splunk_com/scripts/js/global.js?12-03-2009 HTTP/1.1" 200 5106 "https://www.splunk.com/index.php/sign_up/download" "Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.9) Gecko/2009

Example Splunk query

```
search “error”
| stats count by status
| lookup statuscodes status OUTPUT statusdesc
```



0.0	-	**error**	404
0.5	-	OK	200
0.7	-	**error**	500
1.5	-	OK	200

`search “error”`

stage 1

0.0	-	**error**	404
0.7	-	**error**	500

`stats count by status`

stage 2

count	status
1	404
1	500

`lookup statuscodes status OUTPUT statusdesc`

stage 3

count	status	statusdesc
1	404	Not Found
1	500	Internal Server Error

Practitioner viewpoint

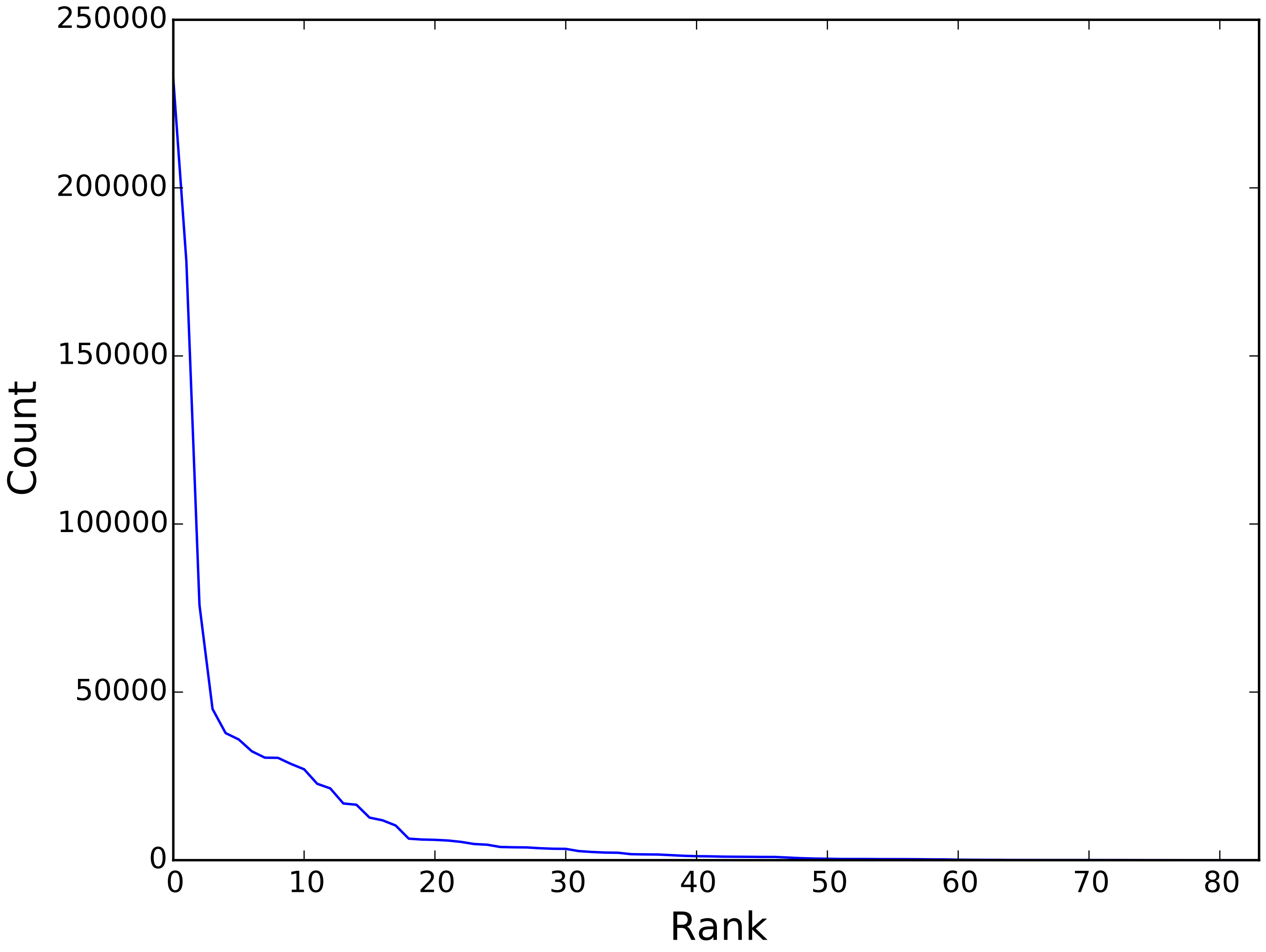
- What are the primitives of log analysis?
  - commands or transformations
- What are the main tasks of log analysis?
  - additional detail
- Why do users analyze logs?
  - context, roles, goals, use cases

Practitioner viewpoint

- What are the primitives of log analysis?
  - commands or transformations
- What are the main tasks of log analysis?
  - additional detail
- Why do users analyze logs?
  - context, roles, goals, use cases

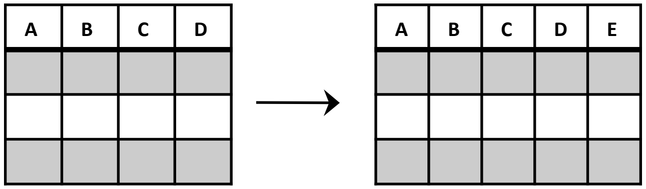
Splunk commands in order of frequency of appearance

command	**count**
search	232373
eval	178080
stats	75927
table	44967
fields	37803
rename	35919
where	32402
inputlookup	30490
sort	30442
lookup	28620
outputlookup	27042
dedup	22731
… snip …	
localop	27
reverse	15
abstract	10
map	7
anomalies	3
extract	2
outlier	2
datamodel	2
format	1
outputtext	1
dbinspect	1


250000
200000
150000
100000
50000
0
Count
0
10
20
30
40
50
60
70
80
Rank

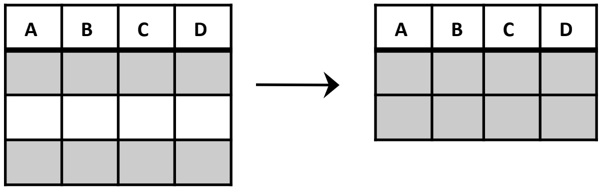
Approach: impose hierarchical organization into tasks, sub-tasks, lower-level activities

```
addinfo
appendcols
bin
bucket
eval
extract
iplocation
kv
outputtext
rangemap
rex
spath
strcat
xmlkv
```



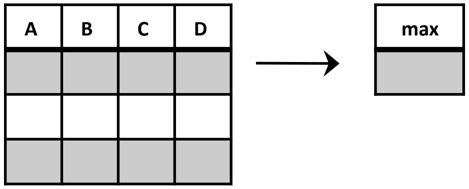
augment

```
dedup
head
regex
search
tail
uniq
where
```



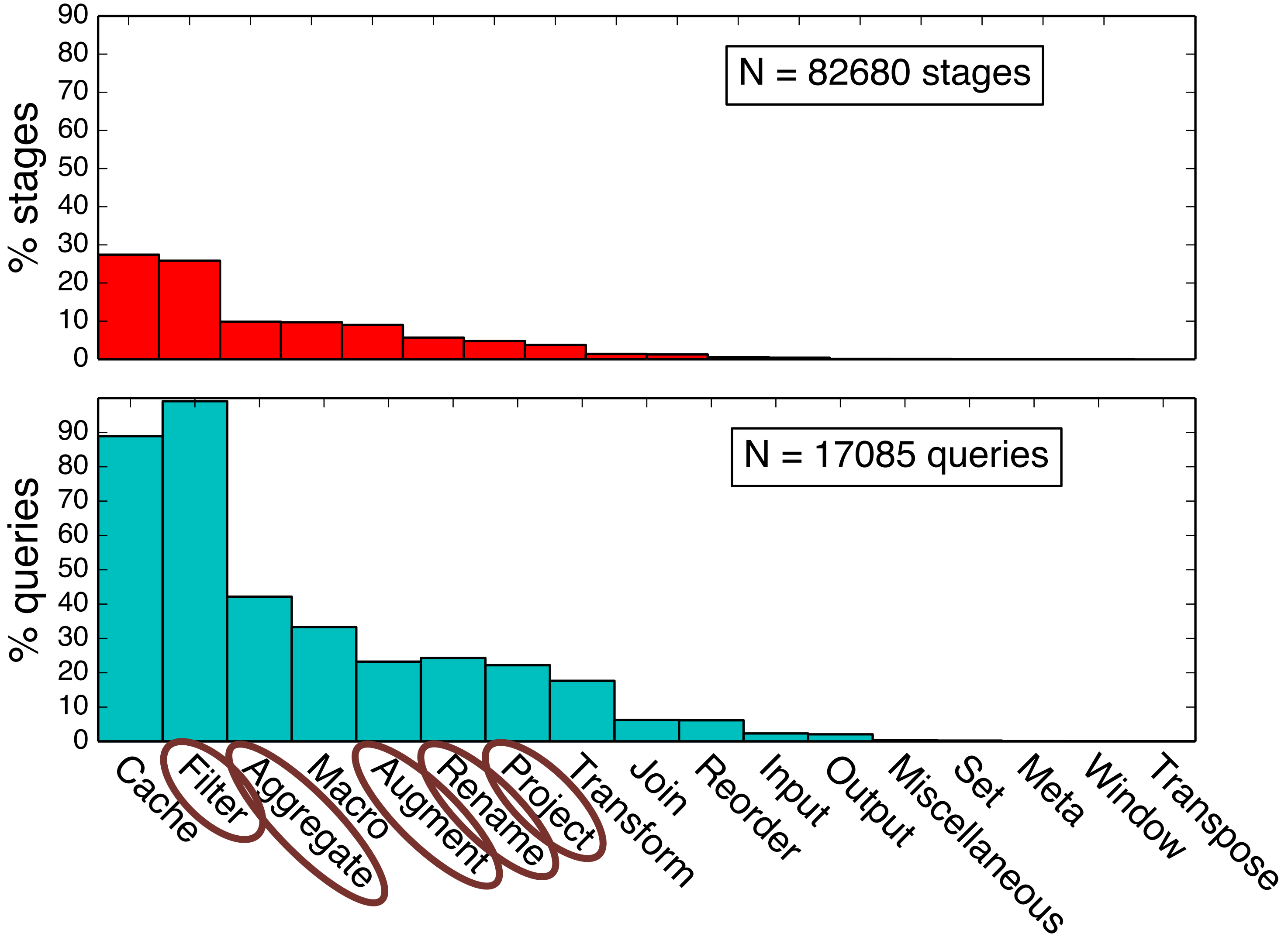
filter

```
addcoltotals
counttable
eventcount
geostats
stats
timechart
top
```



aggregate

Top log analysis transformations

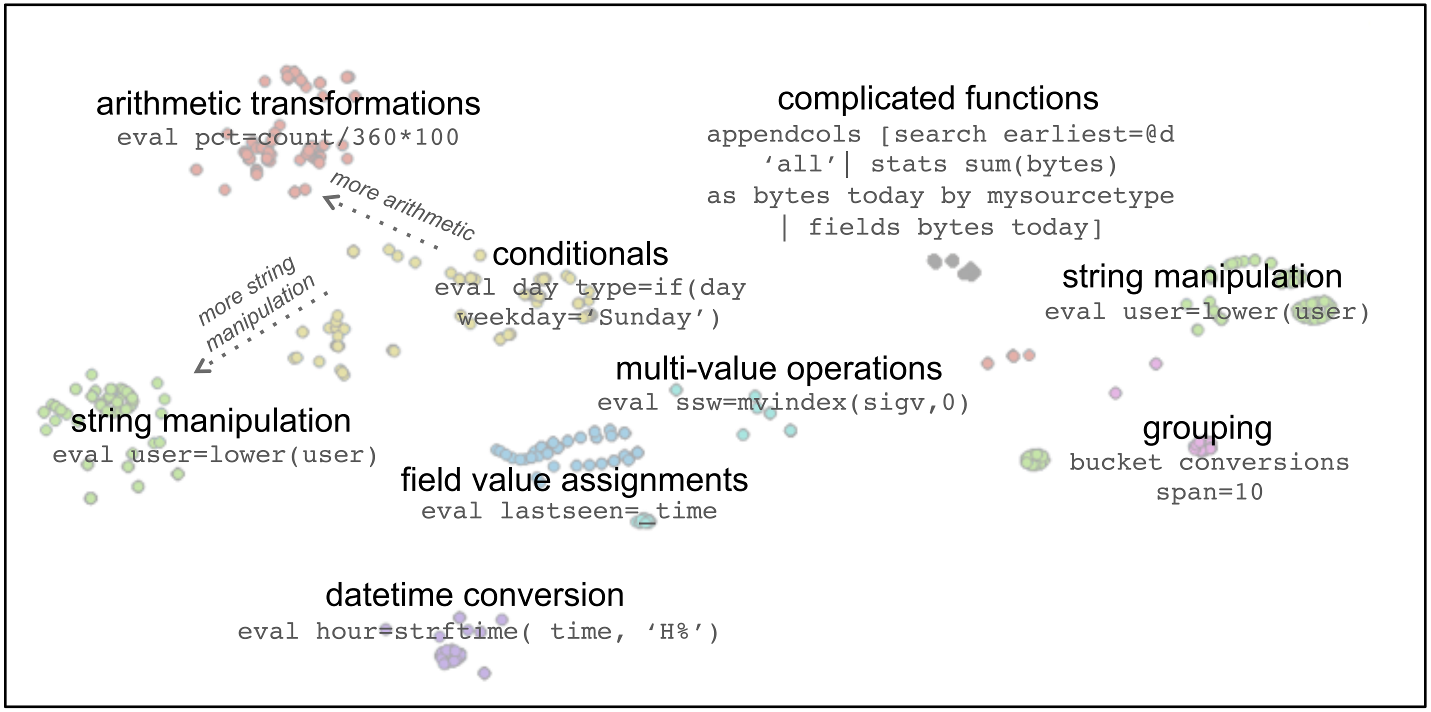


Practitioner viewpoint

- What are the primitives of log analysis?
  - commands or transformations
- What are the main tasks of log analysis?
  - additional detail
- Why do users analyze logs?
  - context, roles, goals, use cases

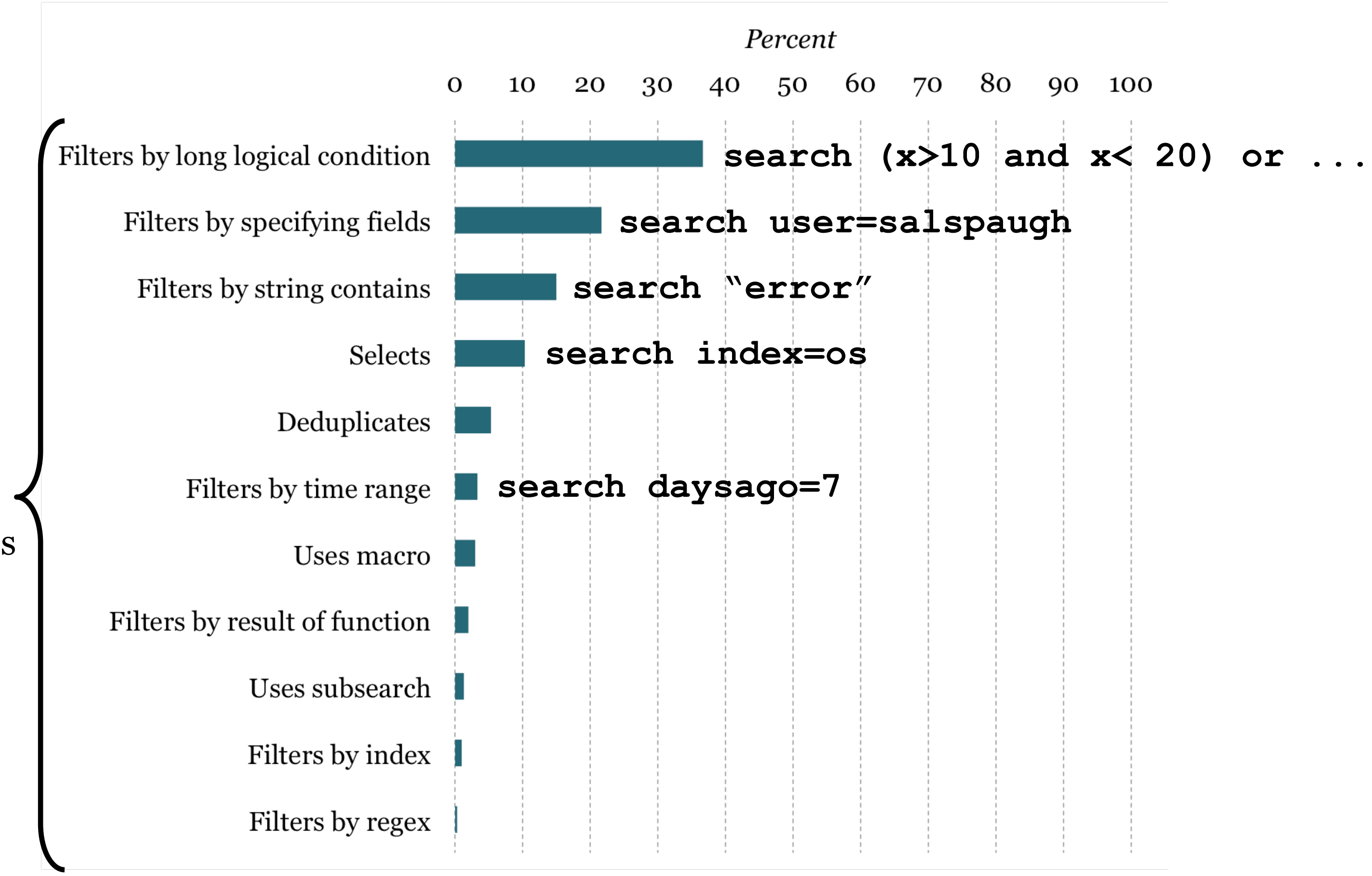
Transformation details

0. Split queries into stages.
1. Featurize each stage in given category (i.e., filter, augment, aggregate).
2. Perform PCA for dimensionality reduction.
3. Perform t-SNE to visualize.
4. Label clusters.



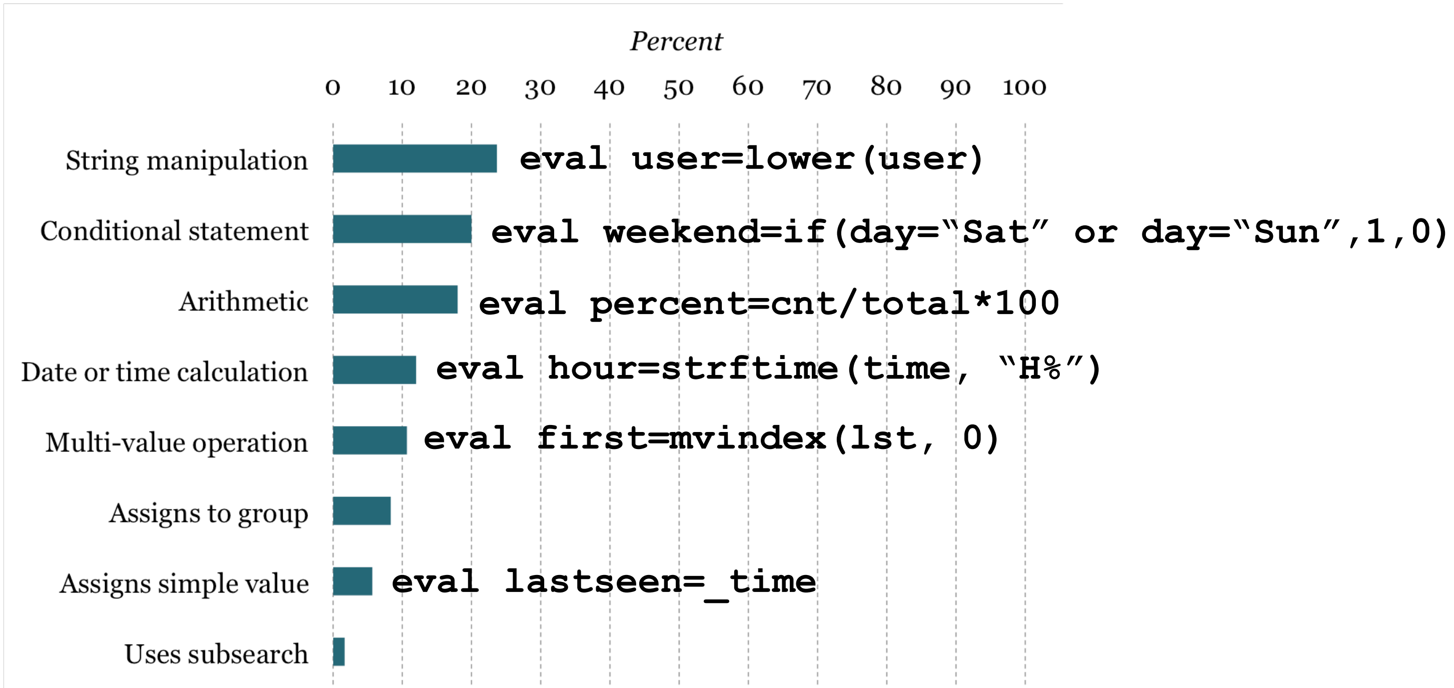
```
eval pct=count/360*100
appendcols [search earliest=@d
‘all’| stats sum(bytes)
as bytes today by mysourcetype
| fields bytes today]
eval day type=if(day
weekday=‘Sunday’)
eval user=lower(user)
eval ssw=mvindex(sigv,0)
eval user=lower(user)
bucket conversions
span=10
eval lastseen=_time
eval hour=strftime(_time, ‘H%’)
```

Types of
Filter
transformations



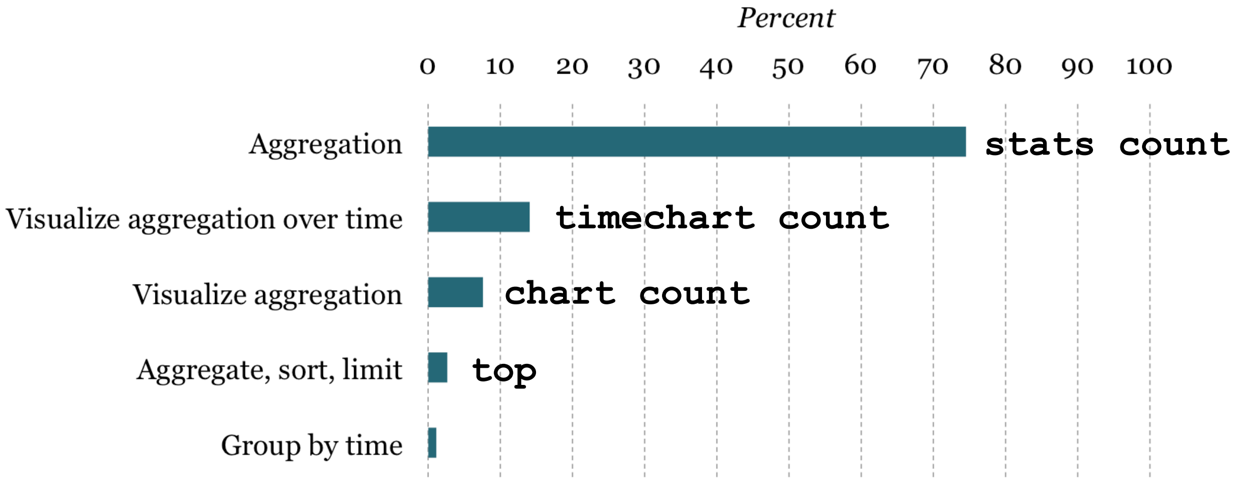
```
search (x>10 and x< 20) or ...
search user=salspaugh
search “error”
search index=os
search daysago=7
```

Types of
Augment
transformations



```
eval user=lower(user)
eval weekend=if(day="Sat" or day="Sun",1,0)
eval percent=cnt/total*100
eval hour=strftime(time, "H%")
eval first=mvindex(lst, 0)
eval lastseen=_time
```

Types of
Aggregate
transformations



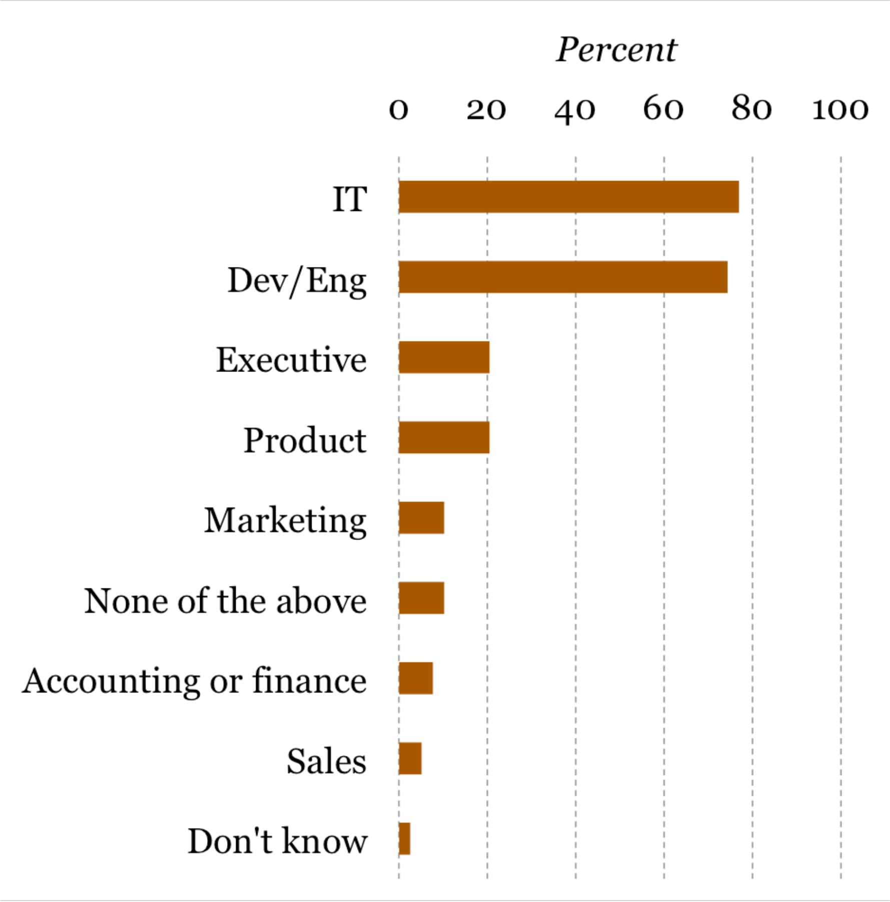
```
stats count
timechart count
chart count
top
```

Practitioner viewpoint

- What are the primitives of log analysis?
  - commands or transformations
- What are the main tasks of log analysis?
  - additional detail
- Why do users analyze logs?
  - context, roles, goals, use cases

Sales engineer survey

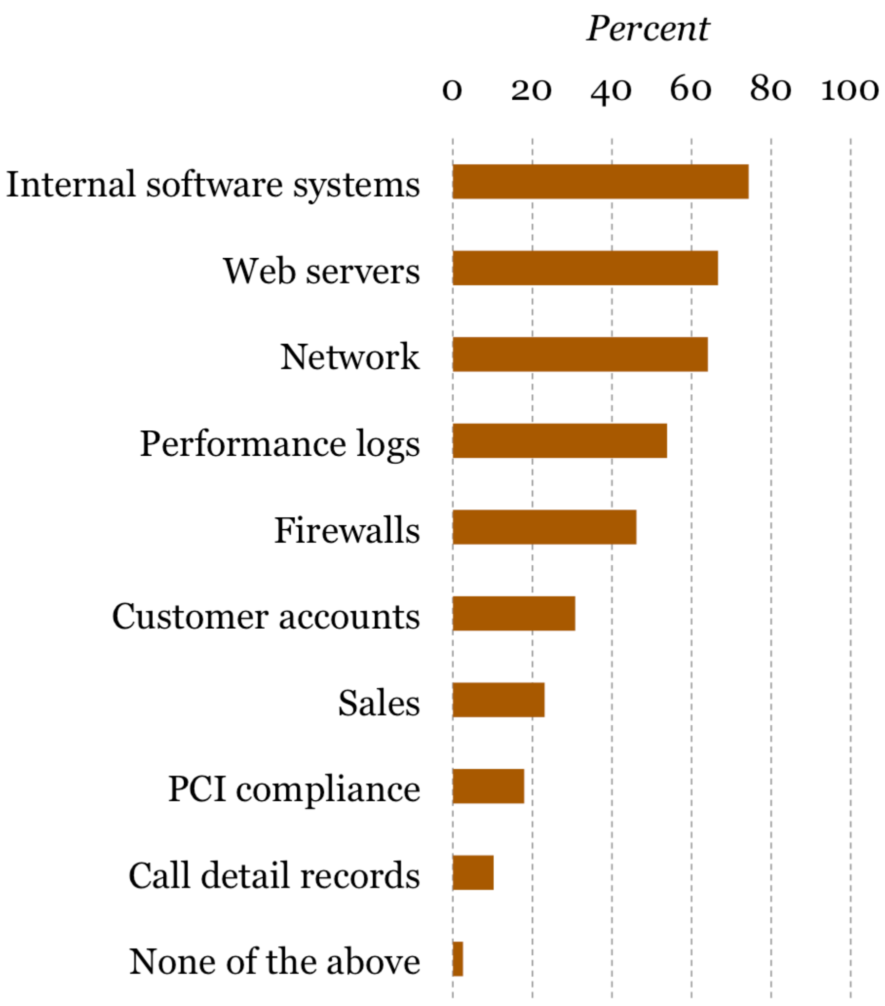
- What are the roles of the primary Splunk users within the organization?
- Write-in answers:
  - – manufacturing team
  - – data analysts
  - – compliance team
  - – security team
  - – email team



n=39

Sales engineer survey

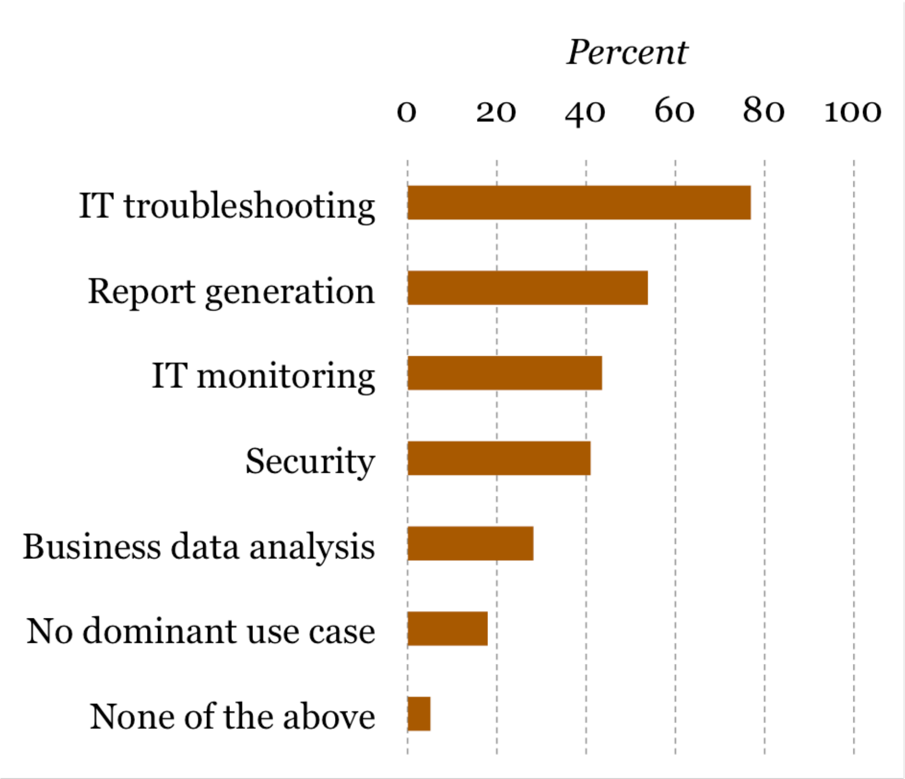
- Roughly what types of data sources did each of these customers have?
- Write-in answers:
  - email
  - sensors
  - mobile apps
  - middleware
  - custom applications



n=39

Sales engineer survey

- What problems does the organization typically try to address with Splunk?
- Write-in answers:
  - app management
  - customer satisfaction
  - workflow automation
  - email monitoring
  - manufacturing



n=39

In the paper

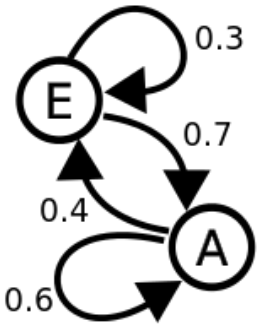
tasks

How many different types of tasks are there? How is the frequency of each of these tasks distributed? What are the most and least common tasks?

sub-tasks

Within the high-level transformation tasks, what sub-tasks are performed? What is the frequency of each of these subtasks?

How are sequences of tasks statistically distributed? What type of tasks usually come first? What comes last? What tasks typically follow a given other task? How many tasks are performed in an average query? What are common subsequences of tasks?



pipelines

What are the primary roles of Splunk users? What problems do they use Splunk for? What other software do they use along with Splunk? What is their level of Splunk expertise? How technical are Splunk users?

ecosystem

Try on your own Splunk data

How many users are there? How are user arrivals distributed? What is the user interarrival rate? What are the basic properties of user sessions? What is the average session length? How many queries are there per session?
How do tasks vary by user?

users

How do transformation frequencies vary with respect to data source? Are some tasks more common in certain contexts than others? How do pipelines vary with respect to data source? How similar is usage with respect to data source type?

data

Analysis challenges

- Messy, complex, with dirty provenance
- Discussion on improving logging: `www.eecs.berkeley.edu/~alspaugh/papers/alspaugh-idea2014-final.pdf`
- No query input/output, data, metadata, system details
- Very large command space with skewed distribution
- Representation mismatch between query language and analysis questions due to functionality overloading
- Caution: research-quality code!
  - query parser: `salspaugh.github.io/splparser`
  - query utilities (including command taxonomy): `salspaugh.github.io/queryutils`
  - paper code: `github.com/salspaugh/lupe`
- **`alspaugh@eecs.berkeley.edu`**

Analysis summary

- 87% of log analysis: data cleaning and troubleshooting

  [40%] **filter**: field-value match, string match, selection

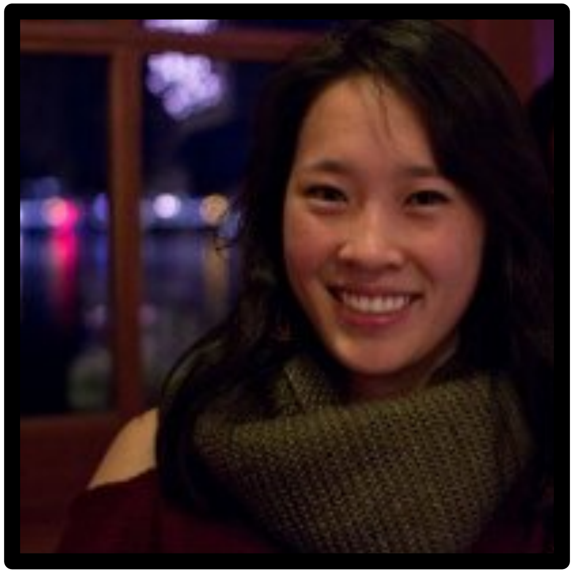
  [15%] **add a column**: that is function of other columns (string manipulation, categorization, simple arithmetic, datetime conversion, etc.)

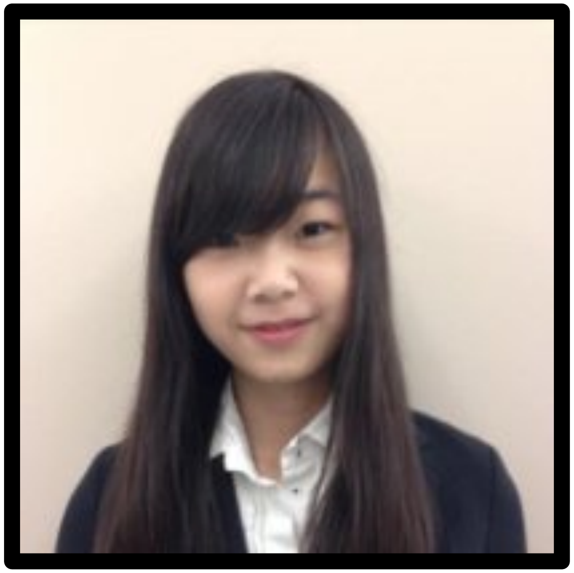
  [15%] **aggregate**: count, average, max, min

  [9%] **rename** columns

  [8%] **project** columns (filter by column)
- Use cases: troubleshooting, security, report generation, monitoring, business intelligence
- Log analysis is not just for IT: management, product, marketing, sales looking at logs

Berkeley CS: graduating May 2015!





`jessica.lin@berkeley.edu` `bettychen824@berkeley.edu`

LISA Lab Office Hours: 4:00-4:45 Thursday

Questions, comments, code requests:

alspaugh@eecs.berkeley.edu