Finding a Needle in a Haystack at Speed

PLA1685

Leveraging Splunk 10
for Faster Searches

.conf25
splunk>

Forward-looking statements

This presentation may contain forward-looking statements regarding future events, plans or the expected financial performance of our company, including our expectations regarding our products, technology, strategy, customers, markets, acquisitions and investments. These statements reflect management's current expectations, estimates and assumptions based on the information currently available to us. These forward-looking statements are not guarantees of future performance and involve significant risks, uncertainties and other factors that may cause our actual results, performance or achievements to be materially different from results, performance or achievements expressed or implied by the forward-looking statements contained in this presentation.

For additional information about factors that could cause actual results to differ materially from those described in the forward-looking statements made in this presentation, please refer to our periodic reports and other filings with the SEC, including the risk factors identified in our most recent quarterly reports on Form 10-Q and annual reports on Form 10-K, copies of which may be obtained by visiting the Splunk Investor Relations website at www.investors.splunk.com or the SEC's website at www.sec.gov. The forward-looking statements made in this presentation are made as of the time and date of this presentation. If reviewed after the initial presentation, even if made available by us, on our website or otherwise, it may not contain current or accurate information. We disclaim any obligation to update or revise any forward-looking statement based on new information, future events or otherwise, except as required by applicable law.

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. We undertake no obligation either to develop the features or functionalities described, in beta or in preview (used interchangeably), or to include any such feature or functionality in a future release.

splunk> .conf25

Finding a Needle in a Haystack at Speed

Tomasz Sikora

Principal Performance Engineer | Splunk

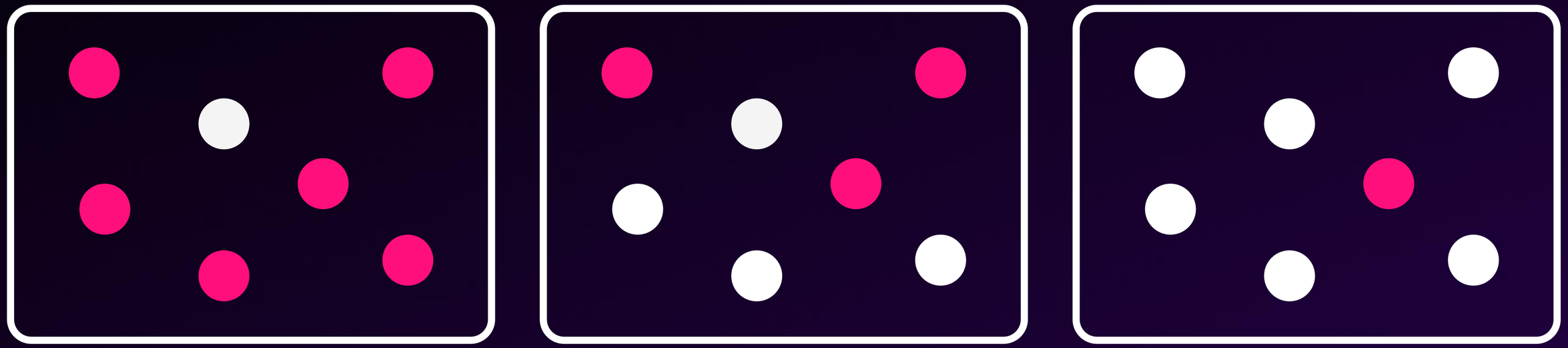
.conf25

splunk>

Areas I’m going to talk about today

- What is a needle in the haystack & why we care
- Bucket elimination techniques
- Search terms explained
- SmartStore – extra level of importance

What is “Needle-in-the-haystack” search



Low - selectivity

Mid - selectivity

High - selectivity

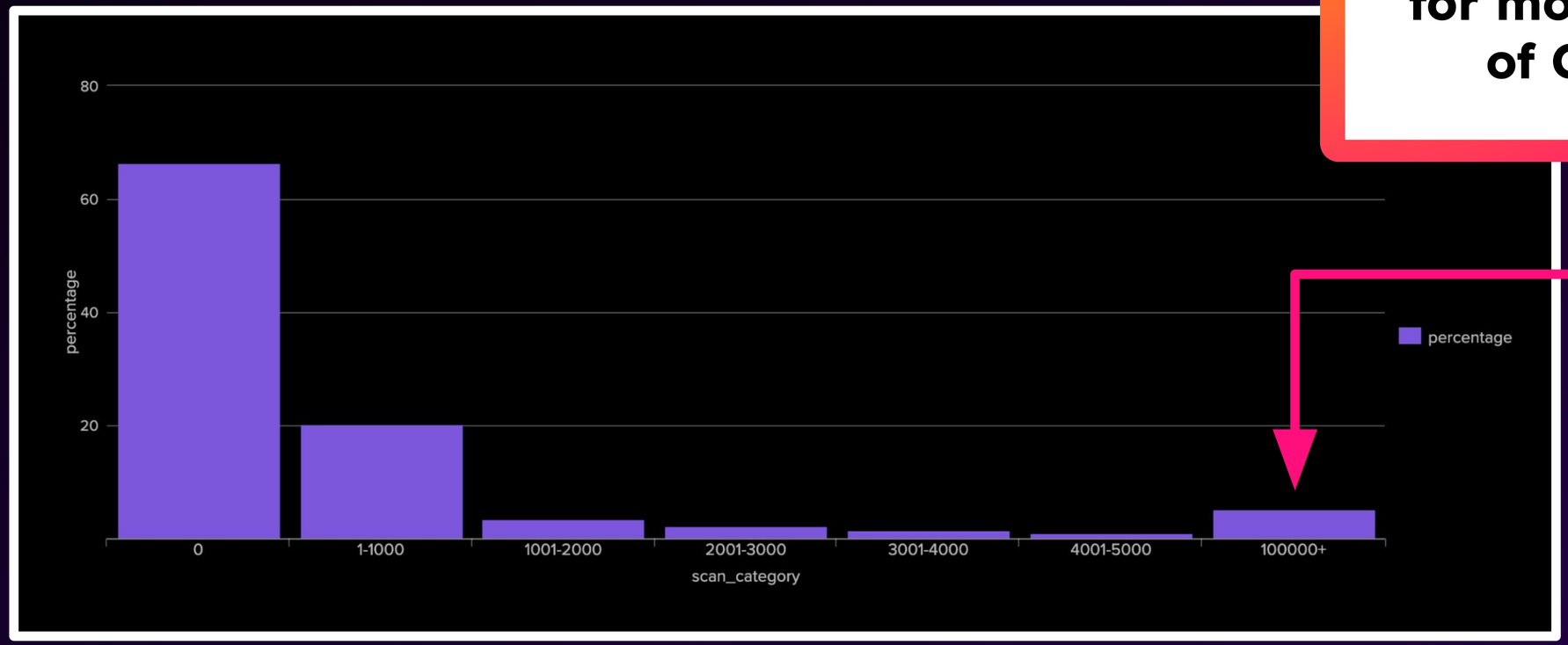
Needle-in-the-haystack!

Most searches scan few events only!

***However*, system performance is highly affected by more dense searches**

Splunk Enterprise Cloud Numbers

Zero count dominates and only ~6% of searches scan more than 100k events



Believe it or not!

These 6% of searches account for more than 35% of CPU time!

Splunk is good at elimination

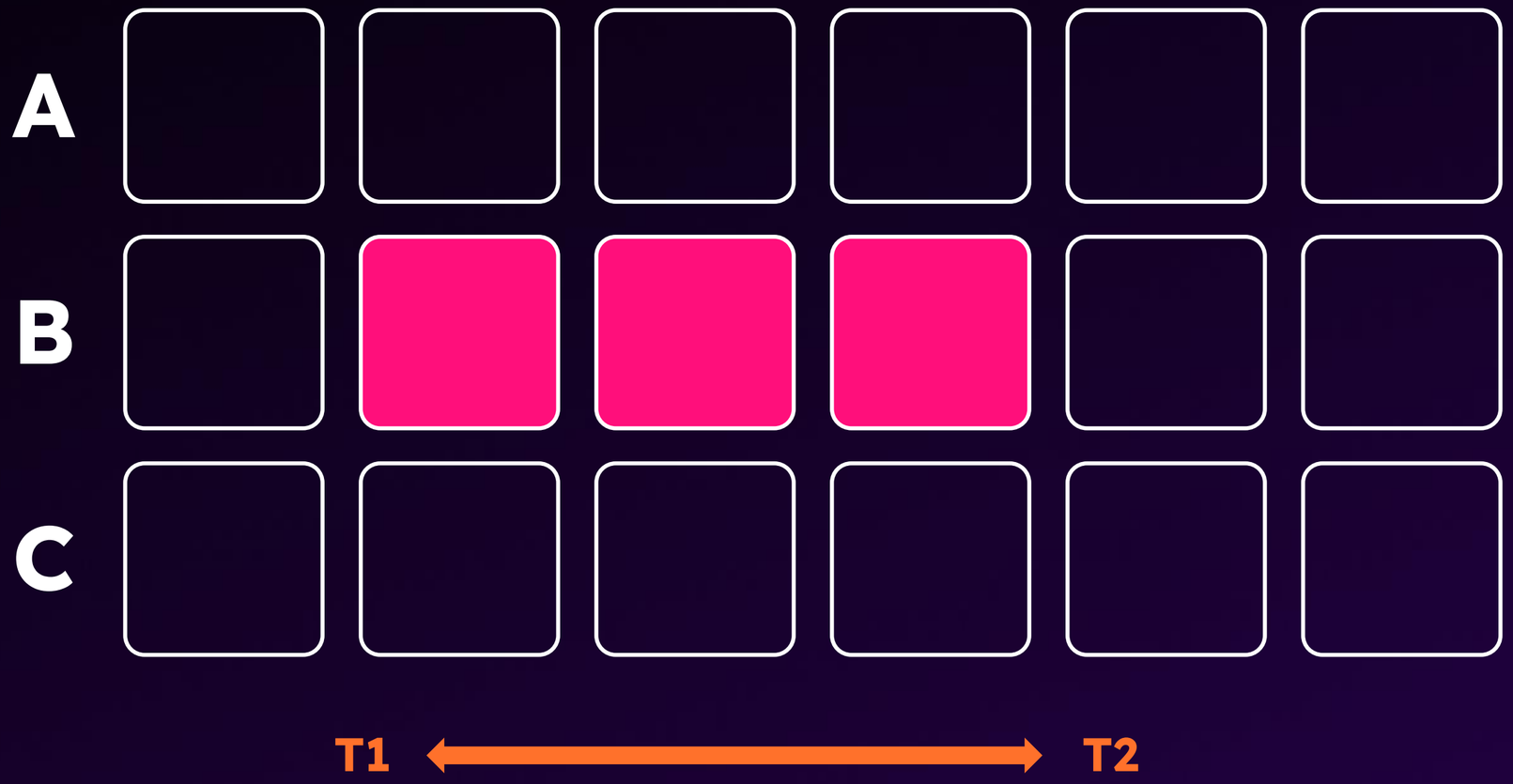
If you want:

- Faster searches
- Faster loading dashboards
- Reduced server load
- Lower TCO
- Support more use cases

Focus on elimination!

Level 1: Bucket elimination

Index and time range



DO

```
index=B earliest=T2
latest=T1
```

DON’t

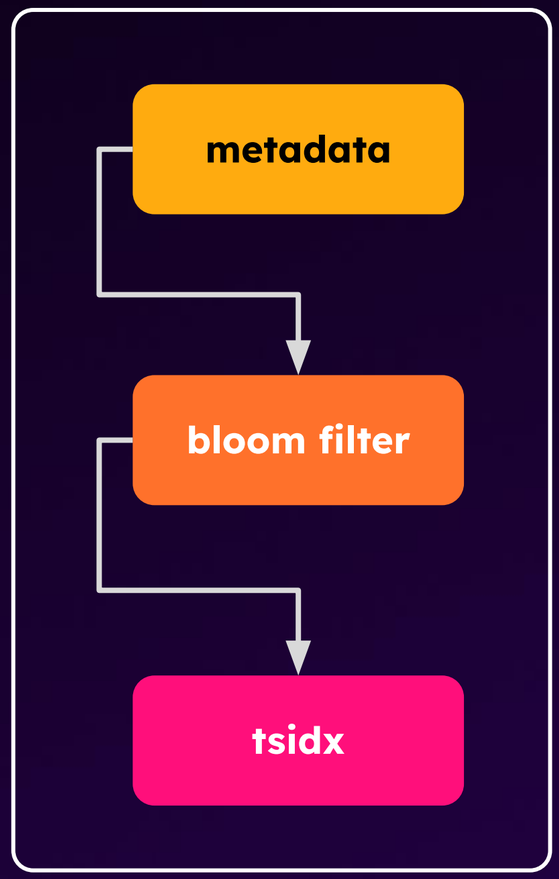
```
index=*
```

Level 2: Bucket elimination

metadata, bloom filter, tsidx

“The earlier the bucket is eliminated, the better.”

“Still better late than never…”



source, host, sourcetype

exact term match **(false positive)**

more sophisticated term matching (eg. prefix match)

Level 2: Bucket elimination

example queries & hints

Query	Possible elimination stage
index=B …	
index=B sourcetype=aws	**metadata**
index=B sourcetype=aws HTTP*	**metadata** **tsidx**
index=B sourcetype=aws (HTTP OR HTTPS)	**metadata** **bloom filter** **tsidx**

What is TERM in Splunk terms?

Tricky part :)

Consider a simple example

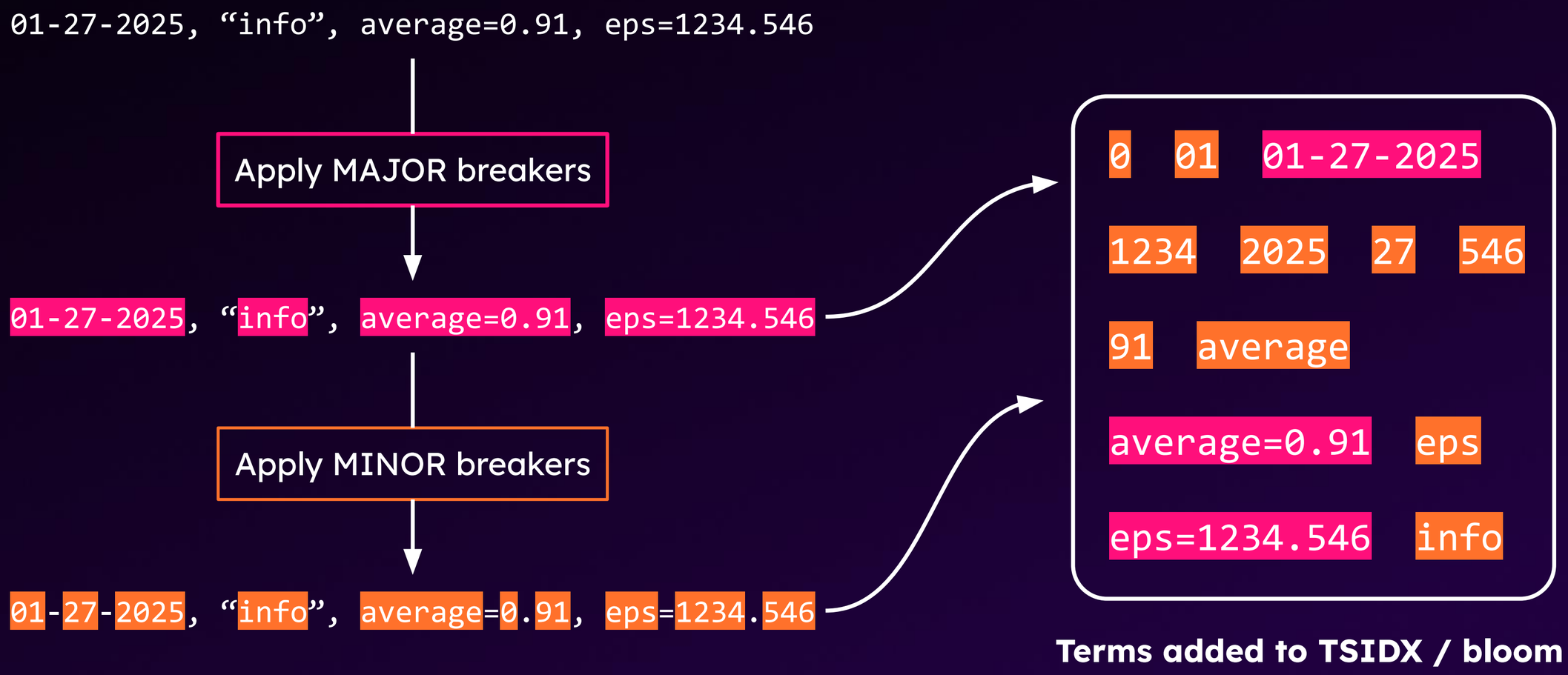
```
index=main average=0.9*
```

→ 20s

```
index=main TERM(average=0.9*)
```

→ 0.7s

MAJOR and MINOR breakers



Terms added to TSIDX / bloom

Power of TERM()

```
SPL: index=* average=0.9*

TERMS:  0  9*
```

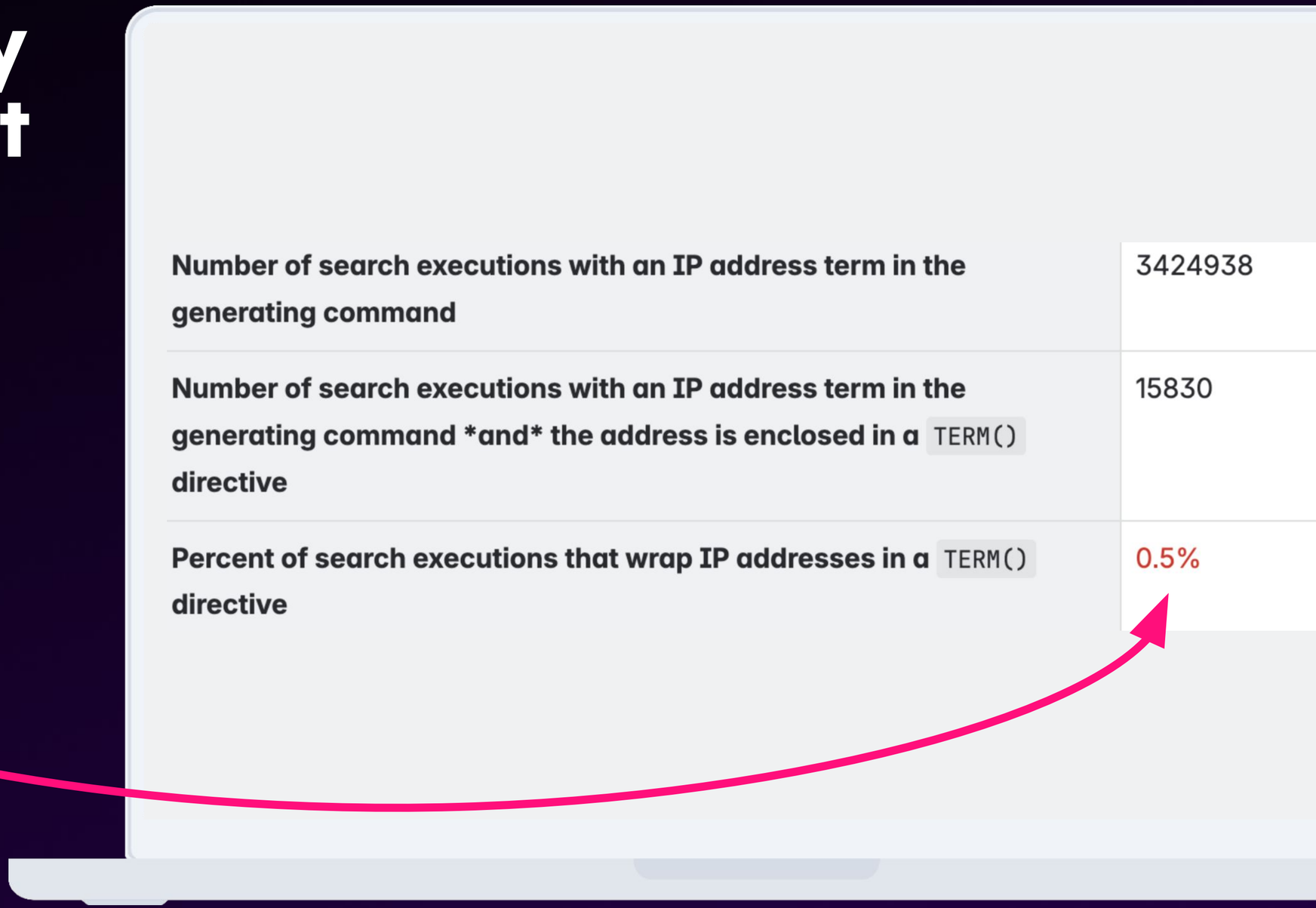
```
SPL: index=*
TERM(average=0.9*)

TERMS: average=0.9*
```

How effective will the elimination be?

This is pretty basic right?

Vast majority of users don't use it!

Less than 1% of searches are done right



Number of search executions with an IP address term in the generating command	3424938
Number of search executions with an IP address term in the generating command *and* the address is enclosed in a `TERM()` directive	15830
Percent of search executions that wrap IP addresses in a `TERM()` directive	0.5%

Splunk 10 to the rescue

Demo of search optimizer improvement

An extra level of importance for SmartStore

S2 Data Density

To enhance cache performance, we aim to maximize the percentage of **useful** data downloaded

AWS S3

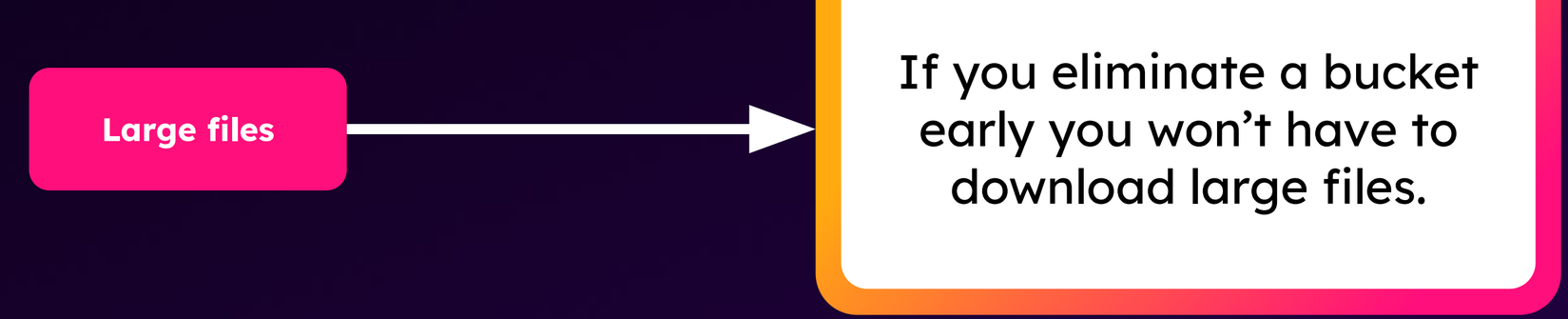
Local SSD

Data density: Eliminate early!

Craft your search to favor early elimination

Splunk downloads buckets in incremental way:

- bloom filter and metadata
- tsidx
- journal



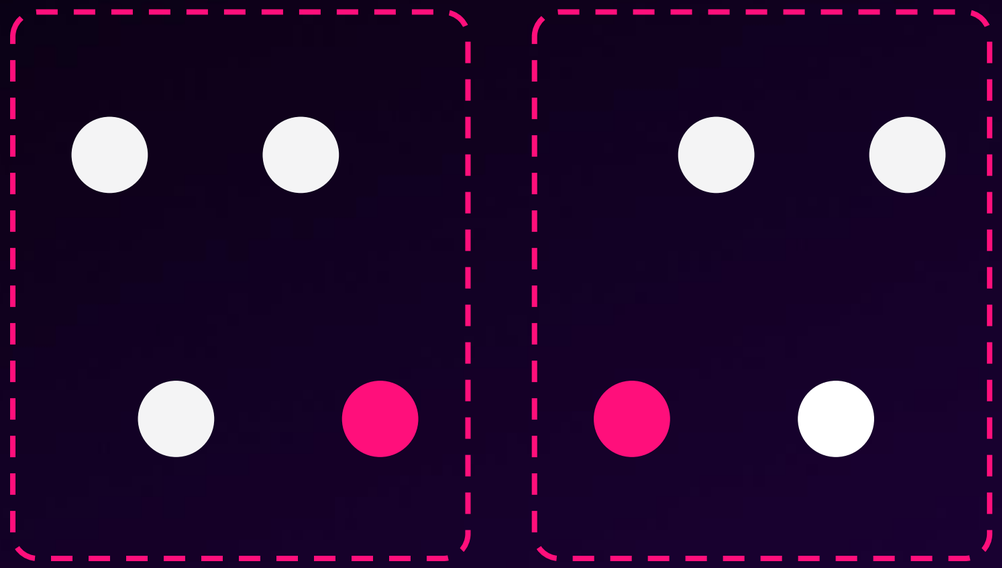
Data density: Small buckets

matching event

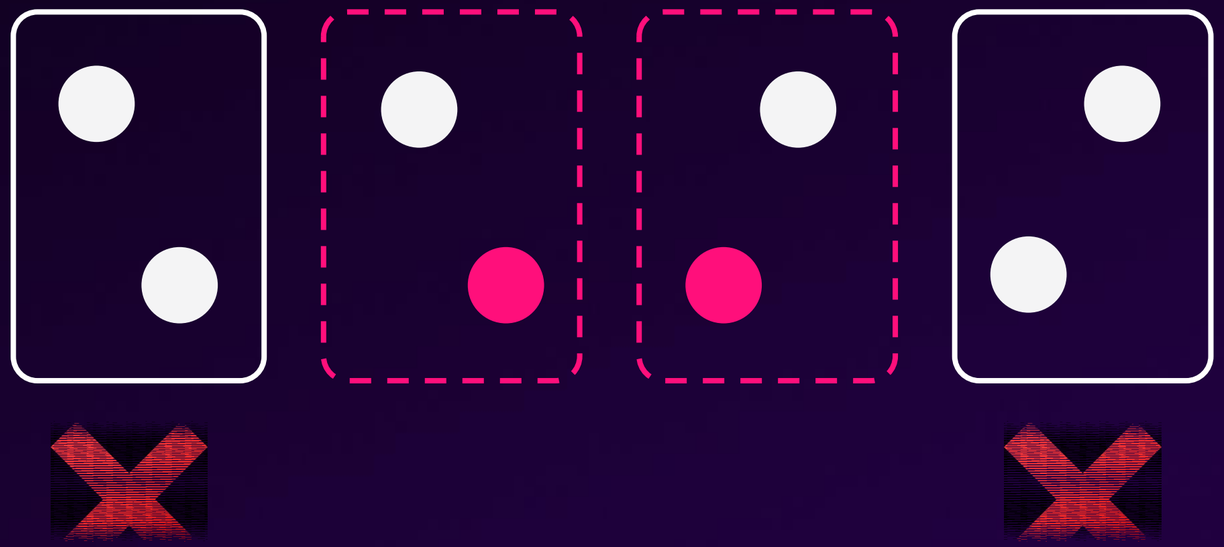
not matching event

bucket to download

Large buckets



Small buckets

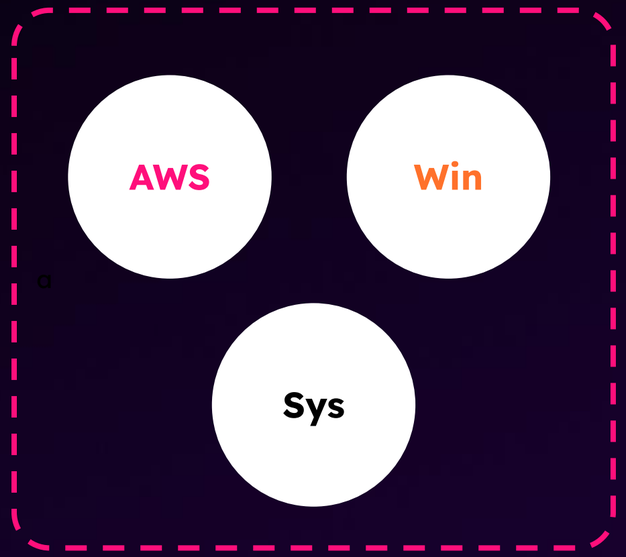


Data density: splitbyindexkeys

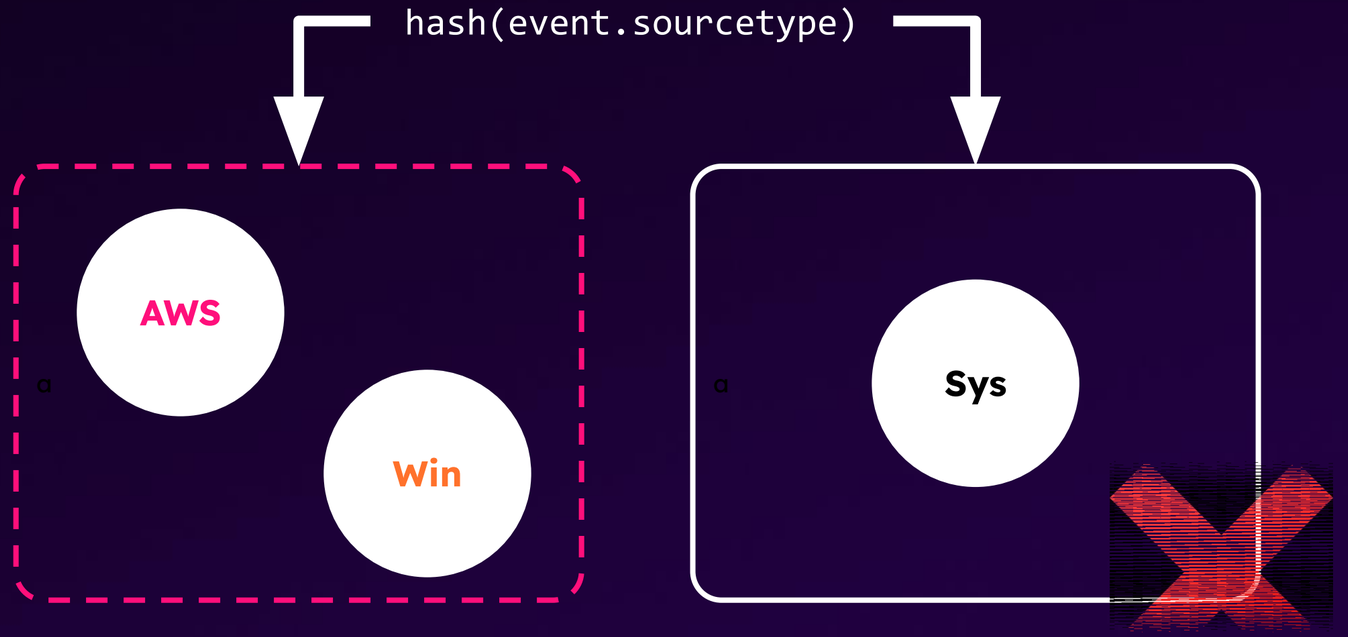
bucket to download

| search sourcetype=aws

Default



splitbyindexkeys (sourcetype)



Other S2 optimizations

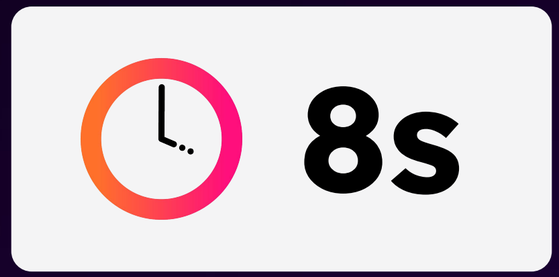
S3 multipart download	**increases download throughput for large files**
TSIDX compression	**Compressed data download is faster**
Bucket Predictor V2	**Optimize prefetch algorithm**

Selected SmartStore optimizations impact

index=testindex sourcetype="PerfmonMetrics:CPU"
| stats avg("metric_name:Processor.%_Idle_Time") by instance

Fast Mode

default



splitbyindexkeys

splitbyindexkeys &
bucket predictor v2

Thank you