Outline for "How to reduce the Splunk cost?"

1. Discuss about cost of licensing model
    - Volume based licensing
    - Sumologic, Splunk, Elasticsearch etc

2. Daily volume of data results in the cost
3. Agents:
    a. Reducing the volume of log data sent to Splunk to process or store
    b. How agents can be used to process the raw events before ingestion into the Splunk
    c. Agent's architecture
4. Reducing the event size
    a. How to reduce the event size?
    b. Storing only key details from log events.
    c. How to use the key details in adhoc queries or reports
    d. Examples of how to do a raw event comparison with key detail storage

5. How to reduce the amount of data stored?

    a. Explaining Patterns for log events. Each type of action is an event, same actions – similar patterns, creating database tables for similar patterns to analyze.
    b. Formatting data before uploading into Splunk
    c. Converting json data into csv, that occupies less space and also to extract the information

---

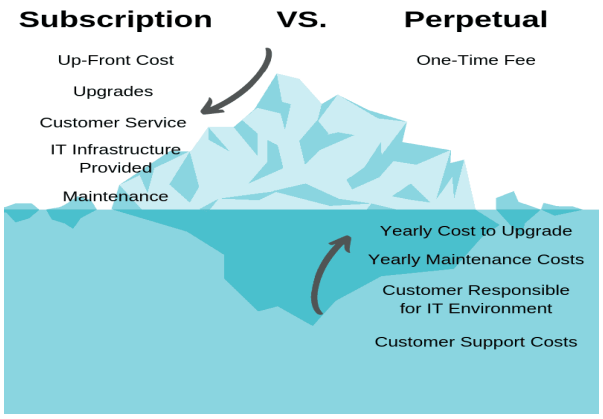
APPROACH FOR SPLUNK COST REDUCTION

1 . Discuss about cost of licensing model

There are three different types of licensing model:

1. Annual

   It is subscription of a software or software as a service (SaaS),
   We should pay the less Upfront Cost and have to pay annually.



2. Perpetual

   It is buying license upfront by paying large amount for single time and owning it.

3. Volume-based licensing model.
   I. It is like paying as you go, you will never own the software even if you use it forever. Low upfront costs and will get charged only the amount of volume you have used.
   II. Vendor is responsible for software and hosting infrastructure and maintenance.

2. Daily volume of data results in the cost

The following are the different types of Service providers and their pricing is listed by Daily Volume wise

Splunk:

Perpetual and Term Licensing are two options for licensing Splunk Enterprise:

Perpetual license:

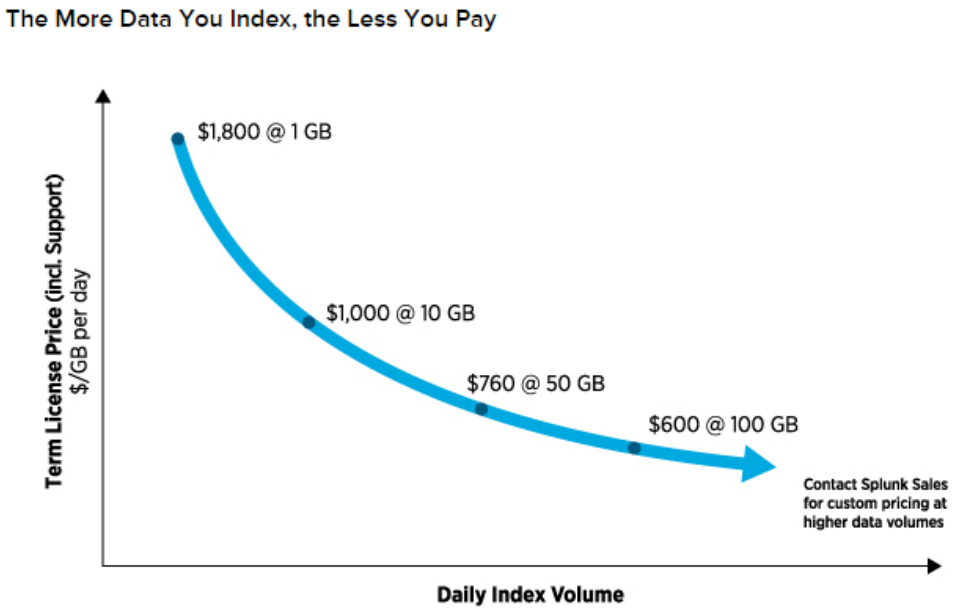
this includes the full functionality of Splunk Enterprise and starts as

Term license:

this provides the option of paying a yearly fee instead of the one-time perpetual license fee. Term licenses start at $1,800 per year*, which includes annual support fees

Index Volume	Perpetual License (per GB)	Annual Term License (per GB)	Volume Purchase Discount
1 GB/Day	$4,500	$1,800	0%
10 GB/Day	$2,500	$1,000	44%
50 GB/Day	$1,900	$760	58%
100 GB/Day	$1,500	$600	67%
>100 GB/Day	Contact sales for custom pricing with additional volume discounts		

The data above reflects pricing for Americas customers only. Pricing for Splunk products varies in EMEA



Sumo Logic:

Sumo Logic Provides Four types of services:

1. Essentials

   Primarily focused on monitoring and troubleshooting use cases.

   The price is starting at $2.25 Per GB of logs

2. Enterprise Operations

Designed for best-in-class monitoring & troubleshooting, backed by full enterprise level 24x7 support.

The price is starting at $4.95 per GB of Logs

3. Enterprise Security

Designed for teams with SaaS SIEM, threat analytics & reporting, and audit compliance use cases, backed by full enterprise level 24x7 support.

The price is starting at $4.95 per GB of Logs

4. Enterprise Suit

All in one platform – most economical plan for organizations with multiple use cases. Hyper-scale on-demand or continuous log analytics, monitoring & troubleshooting, SIEM and much more.

$ 0.11/ GB log search on-demand

$2.475/ GB log search unlimited

$5.50/ GB log search, alerts, dashboard

Elastic Search:

Elastic search is not giving service as volume based, it will charge monthly.

Standard	Gold	Platinum	Enterprise
$16/month	**$19/month**	**$22/month**	We should contact the Sales.

3.Reducing the volume of log data sent to Splunk to process or store

- **Why do we need to reduce the volume of log data before sending to splunk?**

    a. In the industry, Log analysis is mainly based on two key factors:

        1. Volume of data
        2. Retention of data

    The expansion of any organization is directly proportional to the expansion of the log data, which in turn is directly proportional to the expense.
    In many cases a huge amount of the data does not require long-term retention. Today, this plain fact is dismissed while estimating the costs by the log analysis solutions and offer an either fully or not at all operative model that compels companies to pay for retention schemes that do not consider this difference and so are gratuitously costly and ineffectual.

    b. As Splunk is licensed by daily indexed data volume , you need to pay for the total amount of data you send to Splunk per day.So,it is in every customer's interest to keep the data volume generated **as low as possible.**

    c. It is important to estimate the cost based on data generated from the Splunk add-on and data stored .

    Benefits of reducing log data:

    1. Minimizes cost effectively
    2. Reduces log noise

- **How can we reduce the volume of log data?**
    1. Filtering logs that are not needed:
        a. Firstly, this can be done by making a choice of having a default configuration which can deliver the complete and high resolution detail or just optimized for data volume.
        b. Once we know what is amount of generated data ,having an idea , of how much and what are the details that can be reduced without actually affecting the reliability of data is important.
        c. Move the log statement to a lower level to such as debug, so that we don't actually lose the log, when we need it.
        d. Drop logs containing specific key words
    2. Sampling logs
    3. Reduce log size

    For example :

    <u>REDUCE LOG (Decrease the Assigned Capacity of the Recovery Log)</u>

    1. Use this command to decrease the space that can be used by the recovery log.

2. REDUCE LOG command can be used while users are accessing the server.
3. QUERY LOG command can be used to know how much you can reduce the assigned capacity of the recovery log.
4. This command can generate a background process that can be canceled with the CANCEL PROCESS command.
5. If a REDUCE LOG background process is canceled, the assigned capacity of the recovery log may be partially reduced.
6. To display information on background processes, use the QUERY PROCESS command.

- **Requirements to achieve this:**

b.How agents can be used to process the raw events before ingestion into the Splunk

- From the agent's perspective, the data volume per host depends on the various dependencies like the types of applications, the system configurations, types of browsers, background processes running etc.
- Agents could be having dashboards displaying the data volume details is very much essential.

4.How to Reduce the Event size?

1. Log Events are generated based on the actions done by the user or system on machine
2. These actions may or may not be repetitive, but most of them are repetitive
3. For each action done on a machine, log events are stored regarding that specific action
4. Since most of the actions are repetitive, events regarding to those actions will have same structure but field values may differ.
5. When we need to store the logs in a database, we store the entire event, this will increase the size of logs (basically writing the same event multiple times with different field values)
6. Instead of storing the entire event, storing only those field values which are varying in the events will drastically reduce the size of an event.

How the flow looks like?

1. The flow starts with the collecting all the log events from the virtual machine
2. Create Patterns for the events which are having similar structure and name the fields which are varying across the events
3. Store the field values in the database
4. When we want to recreate the events, we will reverse the process what we had done, fetch the pattern that which event belongs and fill the fields with their respective values from the database.

Explaining the steps with few log events:

```
Aug 5 10:08:58 logminer1 systemd[1]: Started Session 2396 of user farooq.
Jul 23 15:10:49 logminer1 systemd[1]: Stopping User Manager for UID 1013...
Aug 5 08:03:14 logminer1 sshd[31788]: error: maximum authentication attempts exceeded for
invalid user test2 from 40.65.126.238 port 38170 ssh2 [preauth]
Jul 27 08:55:54 logminer1 systemd[1]: Started Session 2143 of user malleswari
Jul 23 16:15:17 logminer1 systemd[1]: Started Session 2039 of user pratap.
Jul 23 16:15:57 logminer1 systemd[1]: Stopping User Manager for UID 1013...
Jul 23 03:55:33 logminer1 systemd[1]: Created slice User Slice of pratap.
```

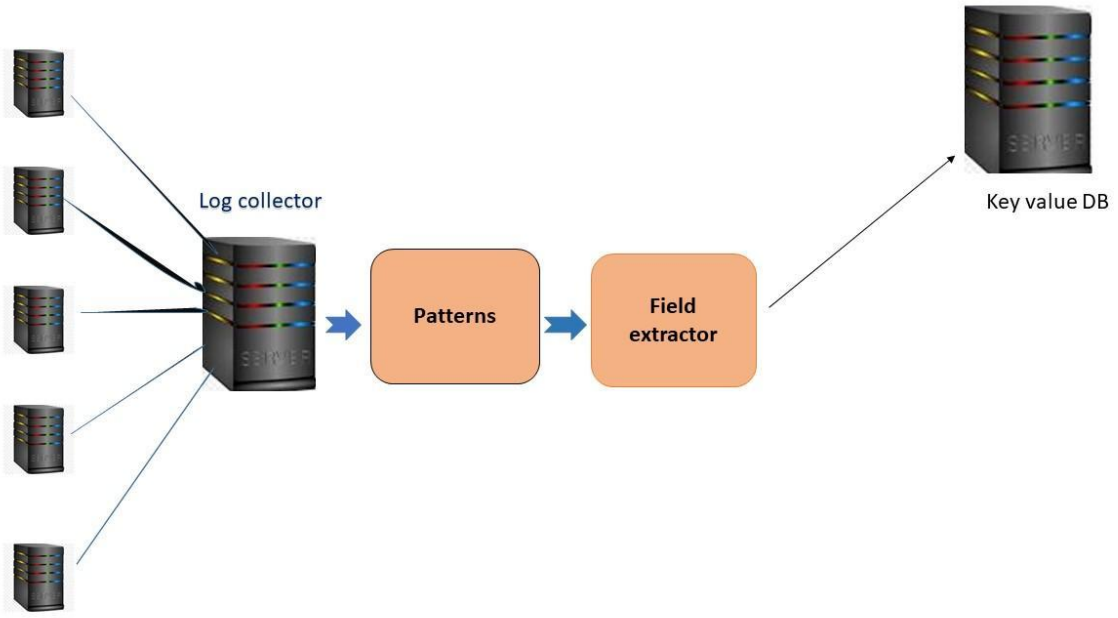
These are the sample syslog events that are collected from a machine. If we analyze them carefully events 1, 5, 6 from top have similar structure but with different field values. So let us group and form them as a pattern.

```
Aug 5 10:08:58 logminer1 systemd[1]: Started Session 2396 of user farooq.
Jul 27 08:55:54 logminer1 systemd[1]: Started Session 2143 of user malleswari
Jul 23 16:15:17 logminer1 systemd[1]: Started Session 2039 of user pratap.
```

The above red marked values are the varying values from these events.Now name and extract the fields from the above pattern.

```
EVENTIME logminer1 systemd[1]: Started Session PID of user USER
```

So from the above pattern we can extract EVENTIME, PID, USER and these values are stored in database (a key value storage) rather than events. We will store this pattern and use to recreate the events. A single pattern covers multiple events.



---

4. Reducing the event size

b) How to use the key details in adhoc queries or reports

- A non-standard inquiry. An **ad hoc query** is created to obtain information as the need arises. Contrast with a **query** that is predefined and routinely processed. See **query** and **ad hoc**

- These queries can be created base don the key data we have stored.To create ad-hoc queries, you first specify the source view for the query to determine the type of **records** to include. Then you can specify output fields and filter criteria for the query. You can use categories to group your queries. When you view ad-hoc queries, you can use filters such as Type or Category to limit how many queries are shown.

c) Examples of how to do a raw event comparison with key detail storage

Windows account login attempts generate events containing copious metadata. See below for an authentication event in Snare Syslog format.

However, only important ingress authentication details are needed. Procedure can decrease the data size. leaving only the important event details required by the system

The difference in size between the two is 2,360 characters (2,917 vs only 557 characters), which is a reduction of 80.9%.

Example:

Full event sample of a Windows Failed Authentication Event in Syslog


```
01 Aug 2019 17:46:45.291{
  "timestamp": "2019-08-01T21:46:43.000Z",
  "hostname": "NXLOG-AGENT",
  "event_code": "4625",
  "description": "An account failed to log on.",
  "subject_user_sid": "S-1-0-0",
  "subject_user_name": "-",
  "subject_domain_name": "-",
  "subject_logon_id": "0x0",
  "logon_type": "Network",
  "target_user_sid": "S-1-0-0",
  "target_user_name": "ADMINISTRATOR",
  "target_domain_name": "",
  "failure_reason": "Unknown user name or bad password.",
  "status": "username or password incorrect",
  "sub_status": "user name is correct but the password is wrong",
  "process_id": "0x0",
  "process_name": "-",
  "workstation_name": "-",
  "ip_address": "212.92.116.56",
  "ip_port": "0",
  "logon_process_name": "NtLmSsp",
  "authentication_package_name": "NTLM",
  "transmitted_services": "-",
  "lm_package_name": "-",
  "key_length": "0",
  "source_data": "<11>Aug  1 17:46:43 NXLOG-AGENT
MSWinEventLog\t3\tSecurity\t77\tThu Aug 01 17:46:43
2019\t4625\tMicrosoft-Windows-Security-Auditing\tN/A\tN/A\tFailure
Audit\tNXLOG-AGENT\tLogon\t\tAn account failed to log on.    Subject:
Security ID:  S-1-0-0   Account Name:  -   Account Domain:  -   Logon ID:
0x0    Logon Type:   3    Account For Which Logon Failed:   Security ID:
S-1-0-0   Account Name:  ADMINISTRATOR   Account Domain:      Failure
Information:   Failure Reason:  Unknown user name or bad password.
Status:   0xC000006D   Sub Status:  0xC000006A    Process Information:
Caller Process ID: 0x0   Caller Process Name: -    Network Information:
Workstation Name: -   Source Network Address: 212.92.116.56   Source Port:
0    Detailed Authentication Information:   Logon Process:  NtLmSsp
Authentication Package: NTLM   Transited Services: -   Package Name (NTLM
only): -   Key Length:  0    This event is generated when a logon request
fails. It is generated on the computer where access was attempted.    The
Subject fields indicate the account on the local system which requested
```

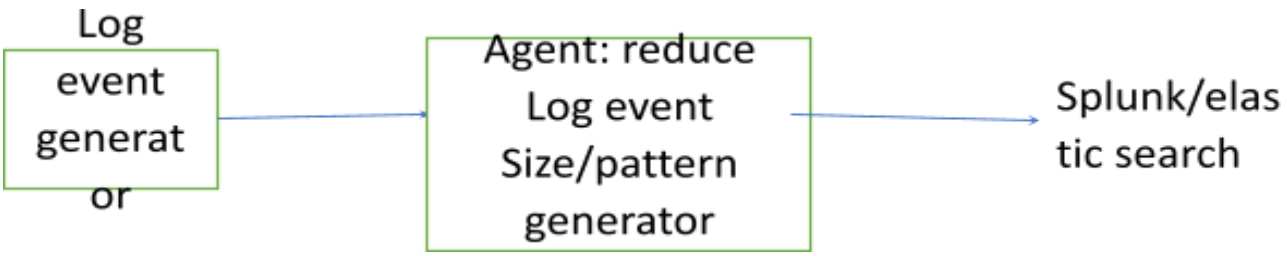
the logon. This is most commonly a service such as the Server service, or a local process such as Winlogon.exe or Services.exe. The Logon Type field indicates the kind of logon that was requested. The most common types are 2 (interactive) and 3 (network). The Process Information fields indicate which account and process on the system requested the logon. The Network Information fields indicate where a remote logon request originated. Workstation name is not always available and may be left blank in some cases. The authentication information fields provide detailed information about this specific logon request. - Transited services indicate which intermediate services have participated in this logon request. - Package name indicates which sub-protocol was used among the NTLM protocols. - Key length indicates the length of the generated session key. This will be 0 if no session key was requested.\t223050"
}

Windows Failed Authentication Event after log enrichment

{"timestamp":"2019-08-01T20:33:41.000Z","user":"NXLOG-AGENT","account":"NXLOG-AGENT","result":"FAILED_OTHER","source_ip":"212.92.117.25","service":"CUSTOM UNIVERSAL EVENT","geoip_organization":"NFOrce Entertainment B.V.","geoip_country_code":"NL","geoip_country_name":"Netherlands","geoip_city":"","geoip_region":"","authentication_target":"-","source_json":{"time":"2019-08-01T20:33:41Z","account":"NXLOG-AGENT","version":"v1","authentication_target":"-","source_ip":"212.92.117.25","event_type":"INGRESS_AUTHENTICATION","authentication_result":"FAILURE"}}

d) Agents:

a. Agent's architecture



---

5. How to reduce the amount of data stored?

Explaining Patterns for log events. Each type of action is an event, same actions – similar patterns, creating database tables for similar patterns to analyze.

Similar events will be considered as patterns
If similarity is not found then that will be considered as separate patterns
Below are the different types of log events

- Application logs

- Database logs
- Network data
- Configuration files
- Performance data
- Time-based data

More data captured = more visibility
In Database store key once and values for same patterns

Formatting data before uploading into Splunk

Timestamp key=value key=value key=value key=value key=value

- May 26 18:14:15 myhostname DBIP=10.5.10.2 Service=Oracle ClientIP=75.149.38.65 SrcPort=80 DestPort=8080 UID=10534 **Sql_Text=Select * from Table1 where uname="dummy".**

keys and values are faster readable to Splunk
Keys are same. So, we will retain keys and pass only values.
Splunk will index and search as fast as we can write it out.
The time stamp can be in ISO 8601 form - i.e. YYYY-MM-DD HH:MM:SS.mmm TZ DST.

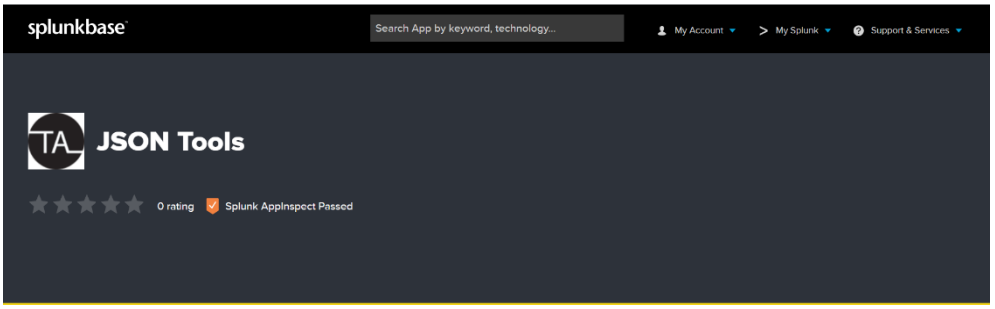
- Example: 2011-10-24 14:04:02 +0200 DST

If you do not want or need the time zone then that can be omitted.

How to reduce the amount of data stored?

Converting json data into csv, that occupies less space and also to extract the information.

- First, we need to convert the raw data into JSON, then export a "CSV with one field containing the JSON.
- Splunk can export events in JSON via the web interface and when queried via the REST api can return JSON output.
- For example…



- **Hardware requirements:** None
- **Software requirements:** Splunk Enterprise 6.3+
- **Installation:** Simply install this app on your search head/s and restart Splunk.
- **Configuration:** No configuration is required.
- With “JSON Tools” app you could convert _raw (and any other fields not from _raw) to JSON, then export a "csv" with one field containing the JSON.
- It can also parse JSON at index/search-time, but it can't *create* JSON at search-time.
- This app provides a 'mkjson' command that can create a JSON field from a given list or all fields in an event.
- The **mkjson** command will include all except hidden fields by default (those that start with an underscore) in the JSON, unless includehidden=true or a list of fields is provided.

```
… | mkjson [outputfield=<fieldname>]
[includehidden=<true|false>] [<fields>]
```

- To simply convert events to JSON:

```
... | mkjson
```

- Before the **mkjson** command if you don't want to include any other hidden fields(_time is technically a hidden field)

```
... | eval time=_time or ... | convert ctime(_time) AS time
```

- When exporting to CSV.. fields with multiple values can be easily preserved. For example, we can convert a single field to JSON:

```
... | mkjson outputfield=src src | outputlookup mylookup
```

- After converting _raw (and any other fields not from _raw) to JSON, then export a "csv" with one field containing the JSON.

```
... | mkjson outputfield=json | table json | outputcsv mycsv
```

==========

- Using converter also we can covert the JSON to CSV. Below link is mentioned.

  https://konklone.io/json/

==========