Running Enterprise Security at Capacity Accurately Thanks to Data Model Acceleration

Gabriel Vasseur, PhD | Senior Cyber Security Analyst
September 2017 | Washington, DC
Forward-Looking Statements

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 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.

Splunk, Splunk>, Listen to Your Data, 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. © 2017 Splunk Inc. All rights reserved.

By attending this talk, you agree to owe a beer to the speaker.
Who is This Guy?

- French
- Lives in England
- Works for **THALES** UK
- PhD in theoretical physics
- 10 years in the IT security industry
- Currently
  - On paper: Senior Cyber Security Analyst
  - In reality: Resident Data Scientist / Splunk Guru
- Spoke at .conf2016:
  
What is He Talking About?

1. Our **story** and why we're here giving this talk
2. The best introduction to **Data Models** ever!
3. Dive into the world of **DM Acceleration**
4. The **Big Picture**: accuracy, load and Data Model acceleration
5. How to **optimize** DM Acceleration
What is This Talk **Not** About?

- tstats tutorial
- Enterprise Security (ES)
- ITSI
- Spelunking
- ...
Company Meets Splunk (2013)

- Assesses data inputs (~50 GB daily) and use cases (Security, ~10 users)
- Looks at minimum requirements for Splunk 6.0.1 ES 3.0 (12 cores & 12GB RAM!)
- Buys over-spec'd hardware (16 cores & 32GB RAM)

Honeymoon period: the relationship feels fresh and exciting!
Trouble (2013 -> 2016)

- Many upgrades later (Splunk 6.0 ES 3.0 to Splunk 6.4 ES 4.4), minimum requirements are 16 cores & 32GB RAM, hardware is no longer over spec'd
- More data has been on-boarded (~150GB daily)
- More use cases / searches / alerts have been developed
- Best practices not necessarily followed...

- Splunk gives and gives, but reaches its limits:
  - Skipped scheduled searches
  - Laggy & crashy data model acceleration
  - Risk of false negatives...
Money Doesn't Buy Happiness
(But it helps)

- Throw hardware at the problem and scale up:
  - Indexer cluster (add more indexers)
  - Search head cluster (a bit complex with ES)
  - More RAM
  - SSDs?

- But...
  - Can't afford it and/or will take too long, need a fix NOW
  - Most importantly, will it be enough? Do we really understand what affects accuracy?
Bring in the Counselors

- Health check with Splunk PS
  - Kick-start ideas for tune-up
- Long-term Splunk PS engagement
  - Implemented many improvements
  - Made data model acceleration really key to managing load

This talk is the summary of what we learned

Disclaimer
- This is what works for us, your mileage may vary!
What Is A Data Model?
“A Data Model is a Hierarchically Structured Search-time Mapping of Semantic Knowledge About One or More Datasets.”
Life *Without* Data Models

- **index=endpoints sourcetype=calvin:AV**
  - 25-sep-2017 4pm OPS threat definition updated to revision 1986
  - 25-sep-2017 5pm DETECTION path=hobbes.exe verdict=BAD threat=water.balloon

<table>
<thead>
<tr>
<th>_time</th>
<th>log_type</th>
<th>path</th>
<th>verdict</th>
<th>threat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-25 4pm</td>
<td>OPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep-25 5pm</td>
<td>DETECTION</td>
<td>hobbes.exe</td>
<td>BAD</td>
<td>water.balloon</td>
</tr>
</tbody>
</table>

- **index=servers sourcetype=hobbes:malwarebuster**
  - 2017/09/25T6pm loglevel=alert state=infected virusname=baseball-cheat path=calvin.exe
  - 2017/09/25T7pm loglevel=alert state=infected boot sector is compromised

<table>
<thead>
<tr>
<th>_time</th>
<th>loglevel</th>
<th>state</th>
<th>virusname</th>
<th>path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-25 6pm</td>
<td>alert</td>
<td>infected</td>
<td>baseball-cheat</td>
<td>calvin.exe</td>
</tr>
<tr>
<td>Sep-25 7pm</td>
<td>alert</td>
<td>infected</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Malware Data Model

As part of the Common Information Model (CIM) all malware events should:

▶ Be tagged with "malware" and "attack"
▶ Have a "signature" field
▶ Have an "action" field
▶ "action" must be either "allowed" or "blocked"
▶ ... etc ...

Now, let's write some props and transforms for each vendor to make their data CIM-compliant...
Achieving CIM Compliance

- index=endpoints sourcetype=calvin:AV

We package this configuration in an "Technical Add-on" (TA) called TA-calvinAV

- Or we get it ready made from Splunkbase

- We do something similar for the Hobbes AV sourcetype

---

eventtype: "AV-calvin-detection"
definition: sourcetype=calvin:AV log_type=DETECTION
tags: malware, attack
eval'ed field "action":
case( verdict=="BAD", "blocked",
verdict=="GOOD", "allowed")

Defined only for calvin:AV sourcetype

Aliased to "signature"
Life With CIM Compliant Data

- `index=endpoints sourcetype=calvin:AV`

  25-sep-2017 4pm OPS threat definition updated to revision 1986
  25-sep-2017 5pm DETECTION path=hobbes.exe verdict=BAD threat=water.balloon

<table>
<thead>
<tr>
<th>_time</th>
<th>log_type</th>
<th>path</th>
<th>verdict</th>
<th>threat</th>
<th>action</th>
<th>signature</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-25 4pm</td>
<td>OPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>malware</td>
</tr>
<tr>
<td>Sep-25 5pm</td>
<td>DETECTION</td>
<td>hobbes.exe</td>
<td>BAD</td>
<td>water.balloon</td>
<td>blocked</td>
<td>water.balloon</td>
<td>malware, attack</td>
</tr>
</tbody>
</table>

- `index=servers sourcetype=hobbes:malwarebuster`

  2017/09/25T6pm loglevel=alert state=infected virusname=baseball-cheat path=calvin.exe
  2017/09/25T7pm loglevel=alert state=infected boot sector is compromised

<table>
<thead>
<tr>
<th>_time</th>
<th>loglevel</th>
<th>state</th>
<th>virusname</th>
<th>path</th>
<th>action</th>
<th>signature</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-25 6pm</td>
<td>alert</td>
<td>infected</td>
<td>baseball-cheat</td>
<td>calvin.exe</td>
<td>blocked</td>
<td>baseball-cheat</td>
<td>malware, attack</td>
</tr>
<tr>
<td>Sep-25 7pm</td>
<td>alert</td>
<td>infected</td>
<td></td>
<td>calvin.exe</td>
<td>blocked</td>
<td></td>
<td>malware, attack</td>
</tr>
</tbody>
</table>
Life With CIM Compliant Data

- tag=malware tag=attack

25-sep-2017 5pm DETECTION path=hobbes.exe verdict=BAD threat=water.balloon
2017/09/25T6pm loglevel=alert state=infected virusname=baseball-cheat object=calvin.exe
2017/09/25T7pm loglevel=alert state=infected boot sector is compromised

<table>
<thead>
<tr>
<th>_time</th>
<th>log_type</th>
<th>loglevel</th>
<th>path</th>
<th>state</th>
<th>verdict</th>
<th>virusname</th>
<th>threat</th>
<th>action</th>
<th>signature</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-25</td>
<td>DETECTION</td>
<td></td>
<td>hobbes.exe</td>
<td>BAD</td>
<td></td>
<td>water.balloon</td>
<td></td>
<td>blocked</td>
<td>water.balloon</td>
<td>malware, attack</td>
</tr>
<tr>
<td>5pm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep-25</td>
<td>alert</td>
<td></td>
<td>calvin.exe</td>
<td>infected</td>
<td></td>
<td>baseball-cheat</td>
<td></td>
<td>blocked</td>
<td>baseball-cheat</td>
<td>malware, attack</td>
</tr>
<tr>
<td>6pm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep-25</td>
<td>alert</td>
<td></td>
<td></td>
<td>infected</td>
<td></td>
<td></td>
<td></td>
<td>blocked</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7pm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- CIM = abstraction layer = you can do vendor-agnostic searches
Life With CIM Compliant Data

- Let's take it to the next level...

<table>
<thead>
<tr>
<th>signature</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseball-cheat</td>
<td>1</td>
</tr>
<tr>
<td>water.balloon</td>
<td>1</td>
</tr>
</tbody>
</table>
The Malware Data Model

### Calculated fields
- eval or lookup
- Not for CIM-compliance!
- `signature = if(isnull(signature), "unknown", signature)`
The `datamodel` Command

```
| datamodel Malware Malware_Attacks search
```

25-sep-2017 5pm DETECTION path=hobbes.exe verdict=BAD threat=water.balloon
2017/09/25T6pm loglevel=alert state=infected virusname=baseball-cheat object=calvin.exe
2017/09/25T7pm loglevel=alert state=infected boot sector is compromised

<table>
<thead>
<tr>
<th>_time</th>
<th>...</th>
<th>...</th>
<th>path</th>
<th>state</th>
<th>verdict</th>
<th>virusname</th>
<th>threat</th>
<th>Malware_Attacks.action</th>
<th>Malware_Attacks.signature</th>
<th>Malware_Attacks.tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-25 5pm</td>
<td>...</td>
<td>...</td>
<td>hobbes.exe</td>
<td>BAD</td>
<td></td>
<td>water.balloon</td>
<td>blocked</td>
<td>water.balloon</td>
<td>malware, attack</td>
<td></td>
</tr>
<tr>
<td>Sep-25 6pm</td>
<td>...</td>
<td>...</td>
<td>calvin.exe</td>
<td>infected</td>
<td></td>
<td>baseball-cheat</td>
<td>blocked</td>
<td>baseball-cheat</td>
<td>malware, attack</td>
<td></td>
</tr>
<tr>
<td>Sep-25 7pm</td>
<td>...</td>
<td>...</td>
<td>Infected</td>
<td></td>
<td></td>
<td></td>
<td>blocked</td>
<td>unknown</td>
<td>malware, attack</td>
<td></td>
</tr>
</tbody>
</table>
The **datamodel** Command

- | **datamodel** Malware Malware_Attacks search | search Malware_Attacks.action=blocked | **stats** count by Malware_Attacks.signature

<table>
<thead>
<tr>
<th>Malware_Attacks.signature</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseball-cheat</td>
<td>1</td>
</tr>
<tr>
<td>water.balloon</td>
<td>1</td>
</tr>
<tr>
<td>unknown</td>
<td>1</td>
</tr>
</tbody>
</table>

- Constraint and vendor stuff is abstracted away
- Still access to raw event and all fields
- Freshly brewed: uses latest search-time configuration and DM definition
- Doesn't benefit from DM acceleration :-(
- **Note**: **datamodel** command to be replaced with **from** command
The `tstats` Command

```
| tstats count from datamodel=Malware where Malware_Attacks.action=blocked by Malware_Attacks.signature
```

<table>
<thead>
<tr>
<th>Malware_Attacks.signature</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseball-cheat</td>
<td>1</td>
</tr>
<tr>
<td>water.balloon</td>
<td>1</td>
</tr>
<tr>
<td>unknown</td>
<td>1</td>
</tr>
</tbody>
</table>

Does the same as the `datamodel` command BUT:

- No access to `_raw` or any non-DM fields
- Stats oriented, no way to just get a table of all the events one by one
- **Will benefit from DM acceleration!!!**
- Can also be used outside DM with indexed fields
The Pivot Command & Datasets UI

- Don't ask me!
- It's powered by tstats...
- ...so go learn tstats!
Accelerating Data Models
# Data Model Acceleration

**Data Models**

Data models enable users to easily create reports and visualizations.

<table>
<thead>
<tr>
<th>Title</th>
<th>Type</th>
<th>Owner</th>
<th>Sharing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td>data model</td>
<td>nobody</td>
<td>Global</td>
</tr>
<tr>
<td>Incident Management</td>
<td>data model</td>
<td>nobody</td>
<td>Global</td>
</tr>
<tr>
<td>Interprocess Messaging</td>
<td>data model</td>
<td>nobody</td>
<td>Global</td>
</tr>
<tr>
<td>Intrusion Detection</td>
<td>data model</td>
<td>nobody</td>
<td>Global</td>
</tr>
<tr>
<td>Inventory</td>
<td>data model</td>
<td>nobody</td>
<td>Global</td>
</tr>
<tr>
<td>JVM</td>
<td>data model</td>
<td>nobody</td>
<td>Global</td>
</tr>
<tr>
<td>Malware</td>
<td>data model</td>
<td>nobody</td>
<td>Global</td>
</tr>
<tr>
<td>Malware Olivier Tmp</td>
<td>data model</td>
<td>olivier</td>
<td>Private</td>
</tr>
<tr>
<td>Microsoft Exchange</td>
<td>data model</td>
<td>admin</td>
<td>Global</td>
</tr>
<tr>
<td>Network Resolution (DNS)</td>
<td>data model</td>
<td>nobody</td>
<td>Global</td>
</tr>
</tbody>
</table>

**Edit Acceleration**

- **Data Model**: Malware
- **Accelerate**: 
- **Summary Range**: 3 Months

Acceleration may increase storage and processing costs.

[Image of Splunk interface with data model options and Edit Acceleration dialog box]
How DM Acceleration Works

▶ Remember this search?
▶ | datamodel Malware Malware_Attacks search

25-sep-2017 5pm DETECTION path=hobbes.exe verdict=BAD threat=water.balloon
2017/09/25T6pm loglevel=alert state=infected virusname=baseball-cheat object=calvin.exe
2017/09/25T7pm loglevel=alert state=infected boot sector is compromised

<table>
<thead>
<tr>
<th>_time</th>
<th>...</th>
<th>...</th>
<th>path</th>
<th>state</th>
<th>verdict</th>
<th>virusname</th>
<th>threat</th>
<th>Malware_Attacks.action</th>
<th>Malware_Attacks.signature</th>
<th>Malware_Attacks.tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-25</td>
<td></td>
<td></td>
<td>hobbes.exe</td>
<td>BAD</td>
<td></td>
<td>water. balloon</td>
<td>blocked</td>
<td>water. balloon</td>
<td>malware, attack</td>
<td></td>
</tr>
<tr>
<td>5pm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep-25</td>
<td></td>
<td></td>
<td>calvin.exe</td>
<td>infected</td>
<td></td>
<td>baseball-cheat</td>
<td>blocked</td>
<td>baseball-cheat</td>
<td>malware, attack</td>
<td></td>
</tr>
<tr>
<td>6pm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep-25</td>
<td></td>
<td></td>
<td>Infected</td>
<td></td>
<td></td>
<td></td>
<td>blocked</td>
<td>unknown</td>
<td>malware, attack</td>
<td></td>
</tr>
<tr>
<td>7pm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How DM Acceleration Works

- Something similar now runs every 5 minutes
- Any non-DM fields are filtered out

<table>
<thead>
<tr>
<th>_time</th>
<th>Malware_Attacks.action</th>
<th>Malware_Attacks.signature</th>
<th>Malware_Attacks.tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-25 5pm</td>
<td>blocked</td>
<td>water.balloon</td>
<td>malware attack</td>
</tr>
<tr>
<td>Sep-25 6pm</td>
<td>blocked</td>
<td>baseball-cheat</td>
<td>malware attack</td>
</tr>
<tr>
<td>Sep-25 7pm</td>
<td>blocked</td>
<td>unknown</td>
<td>malware attack</td>
</tr>
</tbody>
</table>

- The results are shoved into the Data Model summary
- Earliest/latest cover whole retention period
- Only freshly indexed data is considered (late-arriving data included!)
- DM summary = extra .tsidx files in buckets on the indexer
- No _raw included
Think of the DM summary as a traditional DB table:

- DM definition = schema (list of columns)
- Constraint is run as a search, each result = one row

Except:

- Implementation is really not like a DB table...!
- No way to just list rows/events one by one

With DM acceleration, tstats command is automagically much faster!

Best of both worlds:

- Late binding schema (not much set in stone at index-time, flexible search-time configuration)
- Speed of rigid schema with accelerated DM
tstats summariesonly=t

- tstats automatically use DM summaries if they are available
- The last few minutes won't be, most likely
- Doesn't sound like much, but impact is big
- Use `summariesonly=t`
- Careful:
  - Make sure the retention period covers your earliest/latest
  - If you have a time picker in a dashboard, make use you tell your users!

- Make sure the DM summary is fully populated (backfilled + no lag)
ES's Data Model Audit Dashboard
ES's Data Model Audit Dashboard

Focus optimization on worst offenders

Anything here is BAAAD!
Monitoring DM Acceleration Lag

<table>
<thead>
<tr>
<th>New Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>tstats summariesonly=true prestats=t count from datamodel=Network_Traffic by _time span=10s</td>
</tr>
<tr>
<td>eval datamodel=&quot;Network_Traffic&quot;</td>
</tr>
<tr>
<td>tstats append=t summariesonly=true prestats=t count from datamodel=Authentication by _time span=10s</td>
</tr>
<tr>
<td>eval datamodel=Iff(isnull(datamodel),&quot;Authentication&quot;,datamodel)</td>
</tr>
<tr>
<td>timechart span=10s count by datamodel</td>
</tr>
</tbody>
</table>

- Events: 11,329,637 events (01/08/2017 07:32:00.000 to 01/08/2017 08:32:28.000)
Data Model Acceleration

In `datamodels.conf`:

```plaintext
[Web_TEMP]
acceleration = 1
acceleration.earliest_time = -1w
```
Understanding DM Backfilling

Dashboard Source included!
Understanding DM Backfilling

Web_TEMP data model config

-1w  Retention (earliest)  -1w  Backfill target  6.4%  Backfill complete  3  max concurrent  3,600  max time  7012.0 MB  data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon Aug 7 03:40:00 2017</td>
<td>delegated_remote</td>
<td>09:15:38</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 03:35:00 2017</td>
<td>delegated_remote</td>
<td>09:18:38</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 03:30:00 2017</td>
<td>delegated_remote</td>
<td>09:23:38</td>
<td>running</td>
</tr>
</tbody>
</table>
Understanding DM Backfilling

Web_TEMP data model config

<table>
<thead>
<tr>
<th>Retention (earliest)</th>
<th>Backfill target</th>
<th>Backfill complete</th>
<th>max concurrent</th>
<th>max time</th>
<th>data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1w</td>
<td>-1w</td>
<td>16.7%</td>
<td>3</td>
<td>3,600</td>
<td>16013.3 MB</td>
</tr>
</tbody>
</table>

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon Aug 7 06:40:00</td>
<td>delegated_remote</td>
<td>00:27:47</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 06:35:00</td>
<td>delegated_remote</td>
<td>00:32:47</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 06:30:00</td>
<td>delegated_remote</td>
<td>00:37:47</td>
<td>running</td>
</tr>
</tbody>
</table>
Understanding DM Backfilling

Web_TEMP data model config

- Retention (earliest): -1w
- Backfill target: -1w
- Backfill complete: 45.6%
- Max concurrent: 3
- Max time: 3,600
- Data size: 27392.4 MB

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>Scheduled</th>
<th>Statuses</th>
<th>Duration</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon Aug 7 08:30:00</td>
<td>delegated, remote</td>
<td>01:00:42:55</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 08:30:00</td>
<td>delegated, remote</td>
<td>01:00:43</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 08:30:00</td>
<td>delegated, remote</td>
<td>01:00:43</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 08:30:00</td>
<td>delegated, remote</td>
<td>01:00:43</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 08:30:00</td>
<td>delegated, remote</td>
<td>01:00:43</td>
<td>running</td>
</tr>
</tbody>
</table>

© 2017 SPLUNK INC.
Understanding DM Backfilling

Web_TEM config
-1w Retention (earliest)
-1w Backfill target
70.8% Backfill complete
3 max concurrent
3,600 max time
29555.2 MB data size

Web_TEM data model acceleration state

Web_TEM event counts - Monitor lag and backfill

Web_TEM recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon Aug 7 09:40:00 2017</td>
<td>delegated_remote</td>
<td>00:00:24</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 09:45:00 2017</td>
<td>delegated_remote</td>
<td>00:00:24</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 09:40:00 2017</td>
<td>delegated_remote</td>
<td>01:00:35:273</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 09:35:00 2017</td>
<td>delegated_remote</td>
<td>01:00:30:271</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 09:30:00 2017</td>
<td>delegated_remote</td>
<td>01:00:42:558</td>
<td>done</td>
</tr>
</tbody>
</table>
Understanding DM Backfilling

### Web_TEMP data model config

<table>
<thead>
<tr>
<th>Retention (earliest)</th>
<th>Backfill target</th>
<th>Backfill complete</th>
<th>max concurrent</th>
<th>max time</th>
<th>data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1w</td>
<td>-1w</td>
<td>71.6%</td>
<td>3</td>
<td>3,600</td>
<td>31556.3 MB</td>
</tr>
</tbody>
</table>

### Web_TEMP data model acceleration state

### Web_TEMP event counts - Monitor lag and backfill

![Event count chart]

### Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon Aug 7 09:45:00 2017</td>
<td>delegated_remote</td>
<td>00:06:53</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 09:40:00 2017</td>
<td>delegated_remote</td>
<td>00:11:53</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 09:35:00 2017</td>
<td>delegated_remote</td>
<td>00:16:52</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 08:40:00 2017</td>
<td>success, delegated_remote, delegated_remote_completion</td>
<td>01:00:35:273</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 08:35:00 2017</td>
<td>delegated_remote, delegated_remote_completion</td>
<td>01:00:30:271</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 08:30:00 2017</td>
<td>delegated_remote, delegated_remote_completion, success</td>
<td>01:00:42:568</td>
<td>done</td>
</tr>
</tbody>
</table>
Understanding DM Backfilling
# Understanding DM Backfilling

## Web_TEMP data model config

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Retention (earliest)</td>
<td>-1w</td>
<td>Backfill target</td>
<td>-1w</td>
<td>97.9%</td>
<td>Backfill complete</td>
</tr>
<tr>
<td>max concurrent</td>
<td>3</td>
<td>max time</td>
<td>3,600</td>
<td>64577.9 MB</td>
<td>data size</td>
</tr>
</tbody>
</table>

## Web_TEMP data model acceleration state

### Web_TEMP event counts - Monitor lag and backfill

![Graph showing event counts over time]

### Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon Aug 7 11:45 2017</td>
<td>delegated, remote</td>
<td>00:62:21</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 19:00 2017</td>
<td>delegated, remote</td>
<td>00:57:21</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 19:45 2017</td>
<td>success, delegated, remote, completion</td>
<td>01:00:32,752</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 19:40 2017</td>
<td>delegated, remote, success</td>
<td>01:00:38,849</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 09:45 2017</td>
<td>success, delegated, remote, completion</td>
<td>01:00:28,254</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 09:40 2017</td>
<td>success, delegated, remote</td>
<td>01:00:36,169</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 09:35 2017</td>
<td>delegated, remote, completion, success</td>
<td>01:00:35,933</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 08:40 2017</td>
<td>delegated, remote, completion, success</td>
<td>01:00:35,273</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 08:35 2017</td>
<td>delegated, remote, completion, success</td>
<td>01:00:30,271</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 08:30 2017</td>
<td>delegated, remote, completion, success</td>
<td>01:00:42,568</td>
<td>done</td>
</tr>
</tbody>
</table>
Understanding DM Backfilling

Web_TEMP data model config

-1w  Retention (earliest)  -1w  Backfill target  100.0%  Backfill complete  3  max concurrent  3,600  max time  74,858.9 MB  data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon Aug 7 12:50</td>
<td>delegated, remote</td>
<td>00:03:06</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 11:55</td>
<td>delegated, remote</td>
<td>00:08:06</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 11:50</td>
<td>success, delegated, remote_completion</td>
<td>01:00:36:743</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 11:45</td>
<td>success, delegated, remote</td>
<td>01:00:45:203</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 10:50</td>
<td>delegated, remote_completion, success, delegated, remote</td>
<td>01:00:25:025</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 10:45</td>
<td>delegated, remote_completion, success, delegated, remote</td>
<td>01:00:32:752</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 10:40</td>
<td>success, delegated, remote_completion, success, delegated, remote</td>
<td>01:00:38:849</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 09:45</td>
<td>delegated, remote_completion, success, delegated, remote</td>
<td>01:00:28:254</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 09:40</td>
<td>delegated, remote_completion, success, delegated, remote</td>
<td>01:00:36:169</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 09:35</td>
<td>success, delegated, remote_completion, success, delegated, remote</td>
<td>01:00:35:933</td>
<td>done</td>
</tr>
</tbody>
</table>
Understanding DM Backfilling

Web_TEMP data model config

-1w  -1w  99.7%  3  3,600  80793.5 MB

Retention (earliest)  Backfill target  Backfill complete  max concurrent  max time  data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon Aug 7 12:00</td>
<td>delegated_remote</td>
<td></td>
<td></td>
</tr>
<tr>
<td>00 2017</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mon Aug 7 12:30</td>
<td>delegated_remote</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 2017</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mon Aug 7 13:00</td>
<td>success</td>
<td>01:00:34:854</td>
<td>done</td>
</tr>
<tr>
<td>00 2017</td>
<td>delegated_remote_completion</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>delegat...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Understanding DM Backfilling

Web_TEMP data model config

-1w Retention (earliest)
-1w Backfill target
99.6% Backfill complete
3 max concurrent
3,600 max time
81794.1 MB data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

<table>
<thead>
<tr>
<th>Web_TEMP recent acceleration jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>scheduled</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Mon Aug 7 13:00:00 2017</td>
</tr>
<tr>
<td>Mon Aug 7 12:55:00 2017</td>
</tr>
<tr>
<td>Mon Aug 7 12:50:00 2017</td>
</tr>
<tr>
<td>Mon Aug 7 11:35:00 2017</td>
</tr>
<tr>
<td>Mon Aug 7 11:50:00 2017</td>
</tr>
<tr>
<td>Mon Aug 7 11:45:00 2017</td>
</tr>
<tr>
<td>Mon Aug 7 10:50:00 2017</td>
</tr>
<tr>
<td>Mon Aug 7 10:45:00 2017</td>
</tr>
<tr>
<td>Mon Aug 7 10:40:00 2017</td>
</tr>
<tr>
<td>Mon Aug 7 09:45:00 2017</td>
</tr>
</tbody>
</table>
Understanding DM Backfilling

Web_TEMP data model config

-1w Retention (earliest) -1w Rackfill target 100.0% Rackfill complete 3 max concurrent 3,600 max time 85477.0 MB data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon Aug 7 14 05 00 2017</td>
<td>delegated_remote completion delegated_remote success</td>
<td>01:06:27:755</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 14 00 00 2017</td>
<td>delegated_remote completion delegated_remote success</td>
<td>00:15:03</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 13 55 00 2017</td>
<td>delegated_remote completion delegated_remote success</td>
<td>00:30:03</td>
<td>running</td>
</tr>
<tr>
<td>Mon Aug 7 13 00 00 2017</td>
<td>delegated_remote completion delegated_remote success</td>
<td>01:00:26:166</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 12 50 00 2017</td>
<td>delegated_remote completion delegated_remote completion</td>
<td>01:00:45:038</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 11 55 00 2017</td>
<td>success delegated_remote completion delegated_remote</td>
<td>01:00:34:854</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 11 50 00 2017</td>
<td>success delegated_remote completion delegated_remote</td>
<td>01:00:30:743</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 11 45 00 2017</td>
<td>success delegated_remote completion delegated_remote</td>
<td>01:00:45:263</td>
<td>done</td>
</tr>
<tr>
<td>Mon Aug 7 10 50 00 2017</td>
<td>delegated_remote completion delegated_remote success</td>
<td>01:00:45:263</td>
<td>done</td>
</tr>
</tbody>
</table>
In `datamodels.conf`:

```
[Web_TEMP]
acceleration = 1
acceleration.earliest_time = -1w
acceleration.backfill_time = -1d
```
Limiting Data Model Backfilling

Web_TEMP data model config

-1w Retention (earliest)
-1d Backfill target
0.0% Backfill complete
3 max concurrent
3,600 max time
0.0 MB data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 06:15:06 2017</td>
<td>delegated_remote</td>
<td>09:03:42</td>
<td>running</td>
</tr>
</tbody>
</table>
Limiting Data Model Backfilling

Web_TEMP data model config

-1w  -1d  3.9%  3  3,600  1246.4 MB

Retention (earliest)  Backfill target  Backfill complete  max concurrent  max time  data size

Web_TEMP data model acceleration state

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 08:25:00 2017</td>
<td>delegated,remote</td>
<td>00:00:37</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 08:20:00 2017</td>
<td>delegated,remote</td>
<td>00:05:37</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 08:15:00 2017</td>
<td>delegated,remote</td>
<td>00:10:37</td>
<td>running</td>
</tr>
</tbody>
</table>
Limiting Data Model Backfilling

Web_TEMP data model config

-1w  -1d  67.2%  3  3,600  21919.7 MB

Retention (earliest)  Backfill target  Backfill complete  max concurrent  max time  data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 08:25:09 2017</td>
<td>delegated_remote</td>
<td>00:54:26</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 08:20:09 2017</td>
<td>delegated_remote</td>
<td>00:59:29</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 08:15:09 2017</td>
<td>delegated_remote</td>
<td>01:00:42.725</td>
<td>done</td>
</tr>
</tbody>
</table>

Web_TEMP recent acceleration jobs
Limiting Data Model Backfilling

Web_TEMP data model config

- 1w
- 1d
79.3%
3
3,600
2,2514.0 MB

Retention (earliest)
Backfill target
Backfill complete
max concurrent
max time
data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 08:23:00</td>
<td>delegated, remote</td>
<td>09:00:42</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 08:20:00</td>
<td>delegated, remote</td>
<td>08:05:42</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 08:23:00</td>
<td>delegated, remote</td>
<td>01:00:42</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 08:29:00</td>
<td>delegated, remote, completion</td>
<td>01:30.26.010</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 08:15:00</td>
<td>delegated, remote, completion</td>
<td>01:00.42.725</td>
<td>done</td>
</tr>
</tbody>
</table>
Limiting Data Model Backfilling

**Web_TEMP data model config**

- Retention (earliest): -1w
- Backfill target: -1d
- Backfill complete: 95.8%
- max concurrent: 3
- max time: 3,600
- data size: 31559.2 MB

**Web_TEMP data model acceleration state**

**Web_TEMP recent acceleration jobs**

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 15:00:26 2017</td>
<td>delegated_remote</td>
<td>00:03:17</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 09:30:00 2017</td>
<td>delegated_remote</td>
<td>00:48:17</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 09:33:00 2017</td>
<td>delegated_remote</td>
<td>00:53:17</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 09:25:00 2017</td>
<td>success, delegated_remote</td>
<td>00:31:42:994</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 08:25:03 2017</td>
<td>delegated_remote, completion</td>
<td>01:00:32:225</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 08:20:00 2017</td>
<td>delegated_remote, completion</td>
<td>01:00:36:010</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 08:15:00 2017</td>
<td>delegated_remote, completion</td>
<td>01:00:42:725</td>
<td>done</td>
</tr>
</tbody>
</table>
Limiting Data Model Backfilling

Web_TEMP data model config

-1w  Retention (earliest)
-1d  Backfill target
100.0% Backfill complete
3  max concurrent
3,600 max time
33148.6 MB data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>min_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 10:40:00</td>
<td>delegated_remote</td>
<td>00:03:56</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 10:35:00</td>
<td>success delegated_remote</td>
<td>00:03:46</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 10:30:00</td>
<td>delegated_remote</td>
<td>00:13:56</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 10:15:00</td>
<td>success delegated_remote</td>
<td>00:21:17</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 09:30:00</td>
<td>success delegated_remote</td>
<td>00:55:18</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 09:25:00</td>
<td>delegated_remote</td>
<td>01:00:34</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 09:20:00</td>
<td>success delegated_remote</td>
<td>00:51:42</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 08:29:00</td>
<td>delegated_remote</td>
<td>01:00:32</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 08:20:00</td>
<td>success delegated_remote</td>
<td>01:00:26</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 08:15:00</td>
<td>delegated_remote</td>
<td>01:00:42</td>
<td>done</td>
</tr>
</tbody>
</table>
Backfilling Without Lag

In `datamodels.conf`:

```
[Web_TEMP]
acceleration = 1
acceleration.earliest_time = -1w
acceleration.backfill_time = -1d
acceleration.max_time = 900
```
### Backfilling Without Lag

<table>
<thead>
<tr>
<th>Web_TEMP data model config</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1w</td>
</tr>
<tr>
<td>Retention (earliest)</td>
</tr>
</tbody>
</table>

### Web_TEMP data model acceleration state

No results found.

### Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 12:55:00 2017</td>
<td>delegated_remote</td>
<td>00:01:42</td>
<td>running</td>
</tr>
</tbody>
</table>

No results found.
Backfilling Without Lag

Web_TEMP data model config

-1w  -1d  5.7%  3  900  1896.7 MB

Retention (earliest)  Badfill target  Badfill complete  max concurrent  max time  data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 13:05:09 2017</td>
<td>delegated, remote</td>
<td>00:09:55</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:00:00 2017</td>
<td>delegated, remote</td>
<td>00:05:55</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 12:55:00 2017</td>
<td>delegated, remote</td>
<td>00:10:55</td>
<td>running</td>
</tr>
</tbody>
</table>
Backfilling Without Lag

Web_TEMP data model config

-1w Retention (earliest)  -1d Backfill target  15.3% Backfill complete  3 max concurrent  900 max time  5081.7 MB data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 12:05:00</td>
<td>delegated_remote</td>
<td>00:09:49</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:00:00</td>
<td>delegated_remote</td>
<td>00:14:49</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 12:55:00</td>
<td>delegated_remote_completion success delegated_remote</td>
<td>00:15:33:22</td>
<td>done</td>
</tr>
</tbody>
</table>
Backfilling Without Lag

Web_TEMP data model config

-1w  -1d  53.8%  3  900  9509.0 MB

- Retention (earliest)
- Backfill target
- Backfill complete
- max concurrent
- max time
- data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 13:25:06 2017</td>
<td>delegated, remote</td>
<td>00:04:19</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:20:06 2017</td>
<td>delegated, remote</td>
<td>00:09:19</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:15:06 2017</td>
<td>delegated, remote</td>
<td>00:14:19</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:06:06 2017</td>
<td>success, delegated, remote, completion</td>
<td>00:15:24:075</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 13:00:06 2017</td>
<td>success, delegated, remote</td>
<td>00:15:29:185</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 12:55:06 2017</td>
<td>success, delegated, remote, completion</td>
<td>00:15:33:225</td>
<td>done</td>
</tr>
</tbody>
</table>
Backfilling Without Lag

Web_TEMP data model config

-1w  -1d  60.1%  3  900  12669.1 MB

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue Aug 8 13:25:00 2017</td>
<td>delegated, remote</td>
<td>00:00:56</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:25:00 2017</td>
<td>delegated, remote</td>
<td>00:10:55</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:20:00 2017</td>
<td>delegated, remote</td>
<td>00:15:56</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:15:00 2017</td>
<td>delegated, remote, completion, success</td>
<td>00:15:22:081</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 13:05:00 2017</td>
<td>success, delegated, remote, completion, delegated, remote</td>
<td>00:15:24:075</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 13:00:00 2017</td>
<td>success, delegated, remote, delegated, remote</td>
<td>00:15:20:185</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 12:55:00 2017</td>
<td>success, delegated, remote, completion, delegated, remote</td>
<td>00:18:33:225</td>
<td>done</td>
</tr>
</tbody>
</table>
### Backfilling Without Lag

**Web_TEMP data model config**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retention (earliest)</td>
<td>-1w</td>
</tr>
<tr>
<td>Backfill target</td>
<td>-1d</td>
</tr>
<tr>
<td>Backfill complete</td>
<td>90.4%</td>
</tr>
<tr>
<td>max concurrent</td>
<td>3</td>
</tr>
<tr>
<td>max time</td>
<td>900</td>
</tr>
<tr>
<td>data size</td>
<td>14777.6 MB</td>
</tr>
</tbody>
</table>

**Web_TEMP data model acceleration state**

**Web_TEMP event counts - Monitor lag and backfill**

![Graph showing event counts](image)

**Web_TEMP recent acceleration jobs**

<table>
<thead>
<tr>
<th>Date/Time</th>
<th>Status</th>
<th>Run Time</th>
<th>Done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 13:40:00</td>
<td>delegated_remote</td>
<td>09:04:37</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:25:00</td>
<td>delegated_remote</td>
<td>09:10:47</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 13:20:00</td>
<td>delegated_remote</td>
<td>09:15:35</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 12:15:00</td>
<td>delegated_remote</td>
<td>09:15:28</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 13:05:00</td>
<td>success</td>
<td>09:15:24</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 13:00:00</td>
<td>success</td>
<td>09:15:29</td>
<td>done</td>
</tr>
<tr>
<td>Tue Aug 8 12:55:00</td>
<td>success</td>
<td>09:15:33</td>
<td>done</td>
</tr>
</tbody>
</table>
Backfilling Without Lag

Web_TEMP data model config

-1w  -1d  95.2%  3  900  19032.5 MB

-1w  -1d  95.2%  3  900  19032.5 MB

Retention (earliest)  Backfill target  Backfill complete  max concurrent  max time  data size

Web_TEMP data model acceleration state

Web_TEMP event counts - Monitor lag and backfill

Web_TEMP recent acceleration jobs

<table>
<thead>
<tr>
<th>scheduled</th>
<th>statuses</th>
<th>run_time</th>
<th>done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue Aug 8 14:05:06 2017</td>
<td>delegated_remote</td>
<td>09:34</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 14:04:06 2017</td>
<td>delegated.remote</td>
<td>09:34</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:03:06 2017</td>
<td>delegated.remote</td>
<td>09:34</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:24:06 2017</td>
<td>delegated.remote</td>
<td>09:34</td>
<td>running</td>
</tr>
<tr>
<td>Tue Aug 8 13:15:06 2017</td>
<td>delegated.remote</td>
<td>09:34</td>
<td>running</td>
</tr>
</tbody>
</table>
The Big Picture
Accuracy, Load & Data Models
What Do People Really Want?

- Get alerts
- Drill-down to dashboards
- Drill-down to raw events
- Check out dashboards/reports
- Needle-in-haystack search
Accuracy, Load & Data Models

DM summaries

- Correlation searches
- Notable events
- Drill down to dashboard
- Drill down to raw
- DM acceleration
- Needle-in-haystack

Sources:
- RAW DATA

Scheduler

2017 SPLUNK INC.
Accuracy, Load & Data Models

Healthy sources
- Accurate time (NTP)
- Accurate TZ
- No lag
- No rubbish

DM summaries

Correlation searches
- Notable events
- Drill down to dashboard
- Drill down to raw
- DM acceleration
- Needle-in-haystack

Scheduler

Sources

RAW DATA

Accuracy, Load & Data Models

Healthy sources
- Accurate time (NTP)
- Accurate TZ
- No lag
- No rubbish

DM summaries

Correlation searches
- Notable events
- Drill down to dashboard
- Drill down to raw
- DM acceleration
- Needle-in-haystack

Scheduler

Sources

RAW DATA
Accuracy, Load & Data Models

DM summaries

Correlation searches

notable events

drill down to dashboard

Scheduler

Healthy index-time

- Accurate timestamp extraction
- Accurate event breaking
- Accurate routing (index & sourcetype)

DM acceleration

needle-in-haystack

 RAW DATA

Healthy index-time

▶ Accurate timestamp extraction
▶ Accurate event breaking
▶ Accurate routing (index & sourcetype)
Healthy search-time
- CIM-compliant field extractions
- CIM-compliant eventtypes & tags

Healthy acceleration
- Timely acceleration searches (no lag)

DM acceleration

RAW DATA
Accuracy, Load & Data Models

DM summaries

Correlation searches

notable events

drill down to dashboard

RAW DATA

Minimize load:
- tstats summaries only = t
  (15s VS 1s, matters if ran every 5 min.)
- Targeted raw data drilldowns
Optimizing DM Acceleration
Update a DM Without Rebuilding

- datamodels.conf

```
[Web_TEMP]
acceleration = 1
acceleration.earliest_time = -1w
acceleration.backfill_time = -1d
acceleration.max_time = 900
```

- limits.conf

```
[tstats]
allow_old_summaries = true
```

- Splunk Web doesn't let you change an accelerated DM :-(
- Turning off the acceleration and back on again may trigger a rebuild anyway
Update a DM Without Rebuilding

1. Ensure acceleration is 100% up to date

2. Clone DM via web UI

3. Implement changes via web UI to benefit from validation

4. Manually cp and edit modelName & displayName

5. Manually cp clone to overwrite original

6. Delete clone via web UI

Malware clone2.json

Malware clone2.json

Malware clone.json

Malware clone.json

Malware.json
Optimize Constraint – Optimize Tags/Eventtypes

- Look at eventtypes and tags to narrow them down
  - Remove eventtypes that you know will never happen in your data
  - Tweak eventtypes to make them faster
  - Painstaking! Only if you know what you're doing!

`tag=malware tag=attack`
Optimize Constraint – Specify The Index

- Remember when I said the constraint for the Malware DM is this?
  
tag=malware tag=attack

- I lied! It's actually:
  
  `cim_Malware_indexes` tag=malware tag=attack

- The CIM setup page sets that macro

- For other non splunk_SA_CIM DMs, you must manually edit the constraint in the same spirit
Optimize Constraint – Specify The Index

Splunk Enterprise Security

Splunk Enterprise Security extends the security analysis functionality of the Splunk platform, allowing you to centralize your security posture and easily investigate your data. Discover, image, and investigate potential security incidents, coordinate response and remediation, and maintain security posture across security domains, and correlate your data with threat intelligence.

Security Posture
See real-time status of the organization's security posture over the last 24 hours

Incident Review
Work directly with notable events

App Configuration
Configure the application

Documentation
View the User manual, Use Cases, and the Installation and Upgrade manual

Community
Explore Splunk Answers for relevant questions and answers

Product Tour
Go through a product tour to understand Splunk Enterprise Security at a high level
Optimize Constraint – Specify The Index
Don't put all your data in the same index
(nor in hundreds!)
Let's start the story with version A of the Malware data model:

/opt/splunk/etc/apps/Splunk_SA_CIM/default/data/models/Malware.json

constraint: tag=malware tag=attack
fields: ...
....
The Default VS Local Problem

▶ One day we add a field to the data model, creating version A*:

```
/opt/splunk/etc/apps/Splunk_SA_CIM/default/data/models/Malware.json
```

```
constraint: tag=malware tag=attack
fields: ...
.....
```

```
A
```

```
/opt/splunk/etc/apps/Splunk_SA_CIM/local/data/models/Malware.json
```

```
constraint: tag=malware tag=attack
fields: ... + custom field
.....
```

```
A*
```
Later we upgrade Splunk_SA_CIM from version A to version B:

/opt/splunk/etc/apps/Splunk_SA_CIM/default/data/models/Malware.json

constraint: `cim_Malware_indexes` tag=malware tag=attack
fields: ...

```
+ custom field
```

B

A*

constraint: tag=malware tag=attack
fields: ...

```
+ custom field
```

B
The advice you hear: "clone before you modify"

- I don't see the point as it doesn't solve the problem

My advice: be paranoid with upgrades!

The Default VS Local Solution

- Out-of-the-box Splunk A
- Production Splunk A*
- Out-of-the-box Splunk B

custom changes

upgrade changes

conflicts?
Other Optimization

▶ Disable DM acceleration for DM you don't use (duh!)
  • Warning: they can come back to life automatically:
    https://docs.splunk.com/Documentation/ES/4.7.2/Install/Datamodels#Data_model_acceleration_enforcement

▶ Don't enable DM acceleration on all your search heads (or move to SHC!)

▶ Reduce cardinality in DM:
  • E.g. session ID in network traffic

▶ Tweak retention:
  • Performance impact is small, space impact is potentially big
Monitoring Skipped Searches

Dashboard Source included!
Reducing Skipped Searches

► Consider increasing quota of scheduled OR auto-summary searches (last resort!)
  • limits.conf:
    
    [scheduler]
    max_searches_perc = 50
    max_auto_summary_searches = 50

► **Careful:** this might affect non-summary searches negatively!

► **Careful:** don't allow a powerful Search Head tier to overwhelm a struggling Indexer tier
Enable data summary replication on your indexer cluster

- Not default, but recommended by splunk
- Don't do it for performance, do it for accuracy!
- In server.conf on Cluster Master:
  ```
  [clustering]
  summary_replication=true
  ```

Consider increasing max_concurrent for a lagging DM:

- In datamodels.conf:
  ```
  acceleration.max_concurrent = 3
  ```
- Only do so if you have spare resources! (last-ish resort)

**Last Word: Protect Your Accuracy**
That's all for Data Model Acceleration!

There is a lot more you should do to optimize everything else in splunk (see further reading)
1. **Slides** and **recording** available on [http://conf.splunk.com/sessions/2017-sessions.html](http://conf.splunk.com/sessions/2017-sessions.html) in a few weeks.

2. Slides available **now** if you email me at gabriel.vasseur@uk.thalesgroup.com
   Includes **source code & further reading**!

3. **Rate** this session in the app :-)

4. **Poke** me on the .conf app if you want to hang out!

**The End...**

**Thanks!**

I hope you liked it!
Thank You

Don't forget to rate this session in the .conf2017 mobile app
Thank You

Don't forget to rate this session in the .conf2017 mobile app
Further Reading
On Data Models

▶ Acceleration docs
http://docs.splunk.com/Documentation/Splunk/latest/Knowledge/Accelerateddatamodels

▶ ES & Data Models docs
https://docs.splunk.com/Documentation/ES/4.7.2/Install/Datamodels

▶ conf 2016 talk @ http://conf.splunk.com/sessions/2016-sessions.html
"The power of data normalization: a look at CIM under the hood"
Mark Bonsack & Vladimir Skoryk
On Searches

conf 2016 talks @ http://conf.splunk.com/sessions/2016-sessions.html

▶ "Behind the magnifying glass: how search works"
Jeff Champagne

▶ "How to scale: from _raw to tstats"
David Veuve

▶ "Fields, indexed tokens and you"
Martin Muller
On Optimizing Everything Else

conf 2016 talks @ http://conf.splunk.com/sessions/2016-sessions.html

▶ Jiffy-lube quick tune-up for your splunk environment
Jeff Champagne & Sean Delaney

▶ Architectural anti-pattern: it seemed like a good idea at the time
David Paper & Duane Waddle

▶ Worst practices and how to fix them
Jeff Champagne
Source Code
Search Source

- Shared for inspiration only – no guarantee!
- No support is included :-)
- But I might help if you ask kindly and I'm not crazy busy
- Works for me and my splunk setup, but your mileage may vary
TODO: Get the custom_decorations.css from the "Splunk 6.x Dashboard Examples" app and place it in etc/apps/YOURAPPNAMEHERE/appserver/static/

TODO: Replace any occurrence of "=Web" below with whichever DM you are using in your dashboard.

NOTE: ${value_never_set} is not supposed to be set :-)

```html
<form
  <row>
    <panel depends="${value_never_set}">
      <input type="dropdown" token="DM_earliest_token" searchWhenChanged="false">
        <query>
          datamodel("Splunk_Audit", "Datamodel_Acceleration") | drop_dm_object_name("Datamodel_Acceleration")
          | eval retention_days=retention/(24*60*60) | eval DM_earliest=if(retention==0,0,now()-retention)
          | table datamodel retention_days DM_earliest | search datamodel=Web
          | earliest=DM_earliest
        </query>
        <fieldForLabel>DM_earliest</fieldForLabel>
        <fieldForValue>DM_earliest</fieldForValue>
      </input>
    </panel>
    <panel>
      <single>
        <title>Oldest record in DM</title>
        <search>
          | tstats summariesonly min(_time) as oldest from datamodel=Web | eval oldest=strftime(oldest, "%c")
        </search>
      </single>
    </panel>
    <panel>
      <search>
        | stats count as DM_earliest | eval DM_earliest="$DM_earliest_token"
        | eval timepicker="$time_token.earliest" | eval timepicker_earliest=if(match(timepicker, "^\d+$"), timepicker, relative_time(now(), timepicker))
        | eval is_in_range=if(timepicker_earliest>=DM_earliest,1,0)
        | eval is_in_range_human=if(is_in_range==1,"YES","NO")
        | rangemap field=is_in_range low=1 severe=0
        | progress
        | set token=dm_range_value=result.is_in_range_human
        | add token=dm_range ranged_value=dm_range_value
        | earliness=DM_earliest
      </search>
    </panel>
    <panel>
      <html>
        <div class="panel-head dashboard-element-header">
          <h3 class="dashboard-element-title">Search within DM range?</h3>
        </div>
        <div class="custom-result-value icon-only $dm_range$">
        </div>
      </html>
    </panel>
  </row>
</form>
```
Rebuild Monitor Source
Skipped Searches Status Source
“Don't believe everything you read on the internet.”

Charles Babbage