2020-02-26 Special Data Migration Meeting Notes

Date
26 Feb 2020

Attendees
- Stephen Pampell
- William Welling
- Zeno Tajoli
- Wayne Schneider
- Tod Olson
- Anne L. Highsmith
- Ian Walls
- Jason Root
- Lisa Sjögren
- Mark Arnold
- Marc Johnson
- Theodor Tolstoy
- Mark Veksler
- Charles Ledvina

Goals
- List of items needed to facilitate data migration to be taken back to Tech Council.

Discussion items

<table>
<thead>
<tr>
<th>Time</th>
<th>Item</th>
<th>Who</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>15m</td>
<td>Discuss data migration experiences</td>
<td>University of Chicago - Index Data</td>
<td>Charles Ledvina - Synchronous endpoints work well for instances, holdings, and items. Problems appear to be in SRS. After about 2M records, the process grinds to a halt. Wayne Schneider - Strategy was to load inventory and SRS separately and link afterward. Dropping indices dramatically improved performance. Theodor Tolstoy - Complexity of SRS affects performance. Mark Veksler - What is acceptable performance for data load and subsequent data upgrades? Ian Walls - My benchmark is 70 records/second. Tod Olson - Batch endpoints for CRUD operations need to be fleshed out. Concern about large data operations (bulk updates), efficient reads (paging vs. streaming) <a href="https://docs.google.com/document/d/10TiO1vn6D8U3f4GsgLCbr2_-405kJFu5YydcVnu-Toh/edit#heading=h.9domz8d2r0ni">https://docs.google.com/document/d/10TiO1vn6D8U3f4GsgLCbr2_-405kJFu5YydcVnu-Toh/edit#heading=h.9domz8d2r0ni</a></td>
</tr>
<tr>
<td>15m</td>
<td>Discuss data migration experiences</td>
<td>Texas A&amp;M - William Welling</td>
<td>William Welling - First attempt using Camunda to page records out of Oracle database and posting to SRM endpoint. Ended up with Duplicates in SRS. Took 10 days, never reached 4m records. Second attempt by going direct to DB results in errors in UI. Wayne Schneider - Data migration has historically been considered out of scope for data import. William Welling - Is the community going to have a common practice for moving data into the system in a migration?</td>
</tr>
<tr>
<td>10m</td>
<td>Discuss data migration experiences</td>
<td>Others</td>
<td>Zeno Tajoli - Shared script used to do data migration. 2.6m took 2h 45m. Using batch API, not using SRS, only inventory. Theodor Tolstoy - 1.3M bibs and 1.7M items, happy with the batch API's. Source records loading direct to DB. Re-implementing business logic in migration tooling not desirable. Tod Olson agrees. Anne L. Highsmith - At times we receive errors that aren't very specific in regards to which records are failing. Theodor Tolstoy has similar experiences.</td>
</tr>
<tr>
<td>10m</td>
<td>Status of SRS module rewrite</td>
<td>???</td>
<td>Theodor Tolstoy - Understands this has been pushed to Q2. MODSOURCE-97 - Improve performance and refactoring of SRS with standard DB tables CLOSED is locus of discussion for SRS redesign.</td>
</tr>
</tbody>
</table>
Action items

- Develop Data Migration community best practice
- Task force to work on SRS problems and workarounds for data migration
  - Task force should be able to directly interface with developers working on SRS
  - We would prefer the efficiency of using the existing data import endpoints to do migrations. If the endpoints are not robust enough, they should be re-worked to handle the amount of data inherent in large migrations.