Best Practices for Flow Performance in Prep Builder

Rapinder Jawanda
Her/she
Senior Product Manager
Tableau

Christine Xu
Her/she
Senior Software Engineer
Tableau
Agenda

Architecture overview

Factors that impact performance for flows
  • Data, DB, sample, complexity of Flow…

Command line interface

Q&A
Interactive Mode vs. Run Mode

Interactive mode

Orders (central)
- Fix dates
  - Rename states

Orders (west)

Returns (all)

Run mode

Orders and returns
- All orders
  - Orders and returns

Clean 2
- Create 'Annual...
Architecture Overview

1. First time: Create caches
2. Additional time: Use caches
What Impacts Performance for Flows?

- Amount of data
- Databases you connect to
- Sample size
- Flow complexity
- Output file
- Machine specification
<table>
<thead>
<tr>
<th>Type</th>
<th>Field Name</th>
<th>Original Field Name</th>
<th>Changes</th>
<th>Sample Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>Row ID</td>
<td>Row ID</td>
<td></td>
<td>183, 184, 185</td>
</tr>
<tr>
<td>✔</td>
<td>Product ID</td>
<td>Product ID</td>
<td></td>
<td>TEC-PH-10003273, TEC-PH-10004896, TEC-AC-10002345</td>
</tr>
<tr>
<td>✔</td>
<td>File Paths</td>
<td>File Paths</td>
<td></td>
<td>orders_south_2015.csv</td>
</tr>
<tr>
<td>✔</td>
<td>Discount</td>
<td>Discount</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>✔</td>
<td>Postal Code</td>
<td>Postal Code</td>
<td></td>
<td>71, 203</td>
</tr>
<tr>
<td>✔</td>
<td>Category</td>
<td>Category</td>
<td></td>
<td>Technology</td>
</tr>
<tr>
<td>✔</td>
<td>Order ID</td>
<td>Order ID</td>
<td></td>
<td>CA-2015-158274</td>
</tr>
<tr>
<td>✔</td>
<td>Product Name</td>
<td>Product Name</td>
<td></td>
<td>AT&amp;T TR1909W, Nokia Lumia 521 (T-Mobile), HP Standard 104 ...</td>
</tr>
<tr>
<td>✔</td>
<td>Ship Date</td>
<td>Ship Date</td>
<td></td>
<td>11/24/2015, 12:00:00 AM</td>
</tr>
<tr>
<td>✔</td>
<td>Quantity</td>
<td>Quantity</td>
<td></td>
<td>4, 5, 2</td>
</tr>
<tr>
<td>✔</td>
<td>City</td>
<td>City</td>
<td></td>
<td>Monroe</td>
</tr>
<tr>
<td>✔</td>
<td>Sub-Category</td>
<td>Sub-Category</td>
<td></td>
<td>Phones, Accessories</td>
</tr>
<tr>
<td>✔</td>
<td>Segment</td>
<td>Segment</td>
<td></td>
<td>Home Office</td>
</tr>
</tbody>
</table>
## Amount of Data

Exclude unnecessary rows

<table>
<thead>
<tr>
<th>State</th>
<th>Order ID</th>
<th>Row ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illinois</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indiana</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iowa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kansas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Michigan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minnesota</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missouri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nebraska</td>
<td></td>
<td></td>
</tr>
<tr>
<td>North Dakota</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oklahoma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Dakota</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Add Filter**

Calculation:

\[ \text{[State] = "Nebraska"} \]
Demo: Remove Data
Database tips

Inputting data takes a while. Help!

- Copy remote files locally
- Wired connection
- Index the DB
- Custom SQL is a “blackbox” to Prep
- SQL server database isolation “Read Uncommitted”
Sampling

- Default
- 1 Row
- Random

For large data sets, you can improve performance by working with a subset of your data. Use these settings to select the data to include in the flow.

Select the amount of data to include in the flow:
- Default sample amount
- Use all data
- Fixed number of rows: \( \leq 1,000,000 \)

Sampling method:
- Quick select
- Random sample (more thorough but may impact performance)
Reduce Complexity of the Flow

Join: Supernode

Orders and returns

Orders and returns

Clean 2
Reduce Complexity of the Flow

Join: Cross DB joins

SQL Server DB table

Teradata DB table

Final joined result set in a hyper table

MS SQL table

Teradata table

= Results
Reduce Complexity of the Flow

Join: Cross DB joins

DB table

Excel file

Final joined result set in a hyper table
Reduce Complexity of the Flow

Join: Faster data types

String 35 sec. vs. Int 6 sec.
Reduce Complexity of the Flow

Join: Moving filters, aggregates before the join

• Joins are expensive operations
• Fewer rows = faster processing
Demo: Filter Before and After Join
Reduce Complexity of the Flow

Script node

- Python and R script steps
Reduce Complexity of the Flow

Aggregates

- Consider how result will be structured
- How do we aggregate multiple values down to a single value

- Sum
- Average or median
- Count or count distinct
- Minimum or maximum

Roll Up Sales
Demo: Show Aggregate
Type of Flow Output File

CSV vs. TDE vs. hyper:

Hyper is better equipped to handle large extracts

CSV 2 hrs 47 mins.

TDE 19 mins.

Hyper 11 mins.
Machine Specs

Follow recommended machine specs and deployment topologies

• Anti-virus/indexer scan
• Low memory->swapping in and out of memory which results in spooling
• Low disk->instability
Command Line Interface

• Directly run Prep Flows without having to open Builder
• Saves time for runs
• Schedule runs through command line
Refresh Output from Command Line

What you need

- Administrator privileges
- Path of Tableau Prep Builder installation
- Path of Flow file location
- Credentials .json file


- c, --connections <arg>   Path to a file with all connection information
- d, --debug              This option is for debugging
- dsv, --disableSslValidation  Disable SSL validation
- h, --help               Print usage message
- t, --tflFile <arg>      The Tableau Prep Flow file
Questions?
Thank You
Please complete the session survey in the mobile app

View ‘My Evaluations’ in the menu or find your session under ‘Schedule’