IBM Business Analytics Proven Practices: Entity Analytics Performance Guide for DB2

Product(s): IBM SPSS Modeler Premium; Area of Interest: Performance

Steve Schormann
Senior Performance Engineer
IBM
11 June 2014

DB2 tuning guidance around SPSS Modeler Premium Entity Analytics.

View more content in this series

Introduction

Purpose of Document

This document describes advanced settings for IBM SPSS Modeler Premium Entity Analytics (EA) which will assist in tuning Entity Analytics based on specific situations.

Applicability

This document is applicable to the IBM SPSS Modeler Premium Entity Analytics 16.0.0 release and its use of IBM DB2.

Exclusions and Exceptions

The information contained in this document may evolve over time. It is important to note that not all data or hardware is equal and your results may not be directly comparable to the results shown or described in this document.

Assumptions

Entity Analytics is a premium component of IBM SPSS Modeler. Users of this document should have a base understanding of SPSS Modeler, Entity Analytics, and the data they will be accessing.
Overview

Entity Analytics in IBM SPSS Modeler 16 utilizes an IBM DB2 database as its repository. While the default settings that are shipped with the product work well in most situations, as users start to push the limits of Entity Analytics it will be needed to change the settings to get optimal performance. This document will describe the settings in the environment that can have an effect on performance related tasks.

EA feature components and configuration settings

Entity Analytics installs a number of components that are useful to understand in order to get maximum performance from your system.

SPSS Modeler nodes

EA Export node

The EA Export node allows the user to map data fields within a dataset computed by an SPSS Modeler stream to the features in a selected EA Entity Type (for example PERSON) and export that dataset to a specified EA data source within the EA repository. EA performs a compute intensive resolution process as each record is loaded into the DB2 repository.

Tuning parameters in the EA system can improve the performance of this node in terms of the rate at which records can be loaded into the repository.

EA Source node

The EA Source Node provides a way to obtain from the repository a summary of which records exported to each EA data source were resolved to which Entity ID. The summary can be filtered by data source.

Executing streams which read data from this node should not have any performance issues that require tuning.

EA Streaming node

The EA Streaming node provides a mechanism for searching the EA repository for potential matches with records from an incoming stream. The user maps fields from the input records to features in the EA repository and selects the kind of matching to perform. The output will include each input record and has the option of including further records describing information known on each entity in the repository which is a potential match for the input record.

Tuning EA parameters can reduce the amount of time taken to process each record when this node is executed.

Entity Analytics processing location

Entity Analytics can be installed in a local installation mode on a client-only machine and can also be installed in distributed mode to a server. In both situations EA is installed on the client. When
making the recommended configuration changes always make the changes where EA is being processed and the database resides.

**IBM DB2**  
The EA feature contains an IBM DB2 database backend which uses a default configuration providing complete transaction recoverability and adequate performance for smaller datasets. The DB2 database offers the following benefits,

- Good configuration for getting started with EA.
- Guarantees that your work is recoverable.
- Configured for minimal system resource usage.

**Configuration settings**  
Configuration files are provided with system defaults. However, when dealing with larger datasets, configuration changes are required to achieve better performance. Some of these configuration changes are,

- Increase the number of records processed in a single batch.
- Increase EA concurrency with the concurrency setting in the g2.ini file.
- Reduce input/output (I/O) operations with configuration changes.
  - Page cleaners
  - Transaction logging
  - Optimizer hints
- Use a fast dedicated disk for database.

**EA workload characteristics**  
**Effects of EA on the system**  
Entity Analytics processes export records and inserts or updates features and elements into the backing database. Each export record can result in many inserts and updates to the database consuming memory and causing I/O.

I/O latency is the single highest contributing factor to overall performance for datasets larger than available memory. EA is self-optimizing and over time will reduce the strain on the backend database but eventually the volume of data can overwhelm the I/O capacity and performance will suffer.

**Effects of data on EA**  
The data itself can also affect performance of the EA system. A highly related dataset can result in lower performance due to more candidates being used in matching and scoring. Sorted datasets can also reduce performance when using multiple threads due to locking on the same entity. In addition, the number of features in the dataset will affect overall performance.

**EA feature performance best practices**  
**Dedicated database disk**  
If possible, place the DB2 database on a dedicated disk. I/O operations to the system disk and pagefile usage can cause bursts of activity to the disk which will affect performance. When
installing SPSS Modeler choose a dedicated disk other than the operating system disk, the DB2 databases will also be created on that disk.

**Best practice**
Install SPSS Modeler on a fast dedicated disk.

If SPSS Modeler is already installed, see the [Solid State Drives](#) best practice for moving the database location.

**Solid State Drives**
Solid state drives (SSDs) are becoming more commonplace in every type of system. They provide much higher I/O rates compared to standard hard drives.

When working with EA databases, much of the resulting database size is dependent on how much actual data is being exported. A guideline of 10,000 bytes per export record can be used as an approximation of database size per record. This would result in a database size of approximately 100 GB for 10,000,000 exported records and require large amounts of I/O capability to sustain performance over the whole export.

While SSDs are typically utilized as the primary system disk to improve boot times and overall system performance, they should also be used to improve I/O times for processes that would otherwise be waiting for disk activity to complete.

To move a DB2 database to other disks you will perform a database backup followed by a restore to the new drive. With SPSS Modeler and all Entity Analytic jobs stopped, issue the following commands for a Windows platform.

**Best practice**
Use an SSD when large record sets are exported.

```plaintext
• md c:\eabackup
• db2 backup database <repositoryName> to c:\eabackup
• db2 drop database <repositoryName>
• db2 restore database <repositoryName> from c:\eabackup on <SSD drive letter>
```

**Increase batch size**
When exporting larger datasets to EA you should increase the number of records in a batch. The cumulative overhead of starting the EA export can be reduced by increasing the batch size. The default batch size is 1000 and is adequate for small datasets. There should never be a need to reduce the batch size below 1000.

Please note that progress report updates are displayed by SPSS Modeler after each batch is completed and interrupting stream execution will wait for the current batch to complete loading.

**Best practice**
Set export_batch_size to 1/10th of the total records you expect to export.

`export_batch_size, datasetSize/10`
Add the `export_batch_size` parameter to the end of the `ea.cfg` file located under `<ModelerServerInstallLocation>\ext\bin\pasw.entityanalytics`. When updating the `ea.cfg` file, Modeler Server or Modeler Client will need to be restarted depending on where the EA processing is executed.

Example syntax: `export_batch_size, 1000`

**Increase EA concurrency**

Increasing concurrency increases the number of worker threads in the EA Export node to process records from the batch. If your system has more than 1 CPU you can increase the concurrency up to a maximum of 4.

Do not increase concurrency to more than the number of logical CPUs in your system. The EA feature can consume all CPU cycles available as long as the I/O latency remains low. If you intend on doing other CPU intensive tasks on the system you may want to decrease the threads to the number of CPUs - 1.

Add or change the concurrency parameter under the `[PIPELINE]` section in the `g2.ini`. The location of the `g2.ini` file depends on the operating system being used.

Windows Vista, Windows 7, Windows Server 2008:

`C:\ProgramData\IBM\SPSS\Modeler\16\EA`

UNIX:

`<modeler-install-directory>/ext/bin/pasw.entityanalytics/EA`

Example syntax: `CONCURRENCY=1`

**Best practice**

Set concurrency to a value of (maximum of the number of logical CPUs on your system – 1) to allow for other concurrent work.

```
[PIPELINE]
CONCURRENCY=#logicalCPUs-1
```

When updating the `g2.ini` file, Entity Analytics will need to be stopped and restarted depending on where the EA processing is executed.

Please note that progress report updates are displayed by SPSS Modeler after each batch is completed and interrupting stream execution will wait for the current batch to complete loading.

**Use alternate page cleaners**

DB2 has two different page cleaning algorithms to flush changed pages to disk. Flushing pages can incur some performance overhead if the amount of pages to flush is very large. This can happen with insert intensive workloads such as EA. Using DB2's alternate page cleaners causes changed pages to be written in a more consistent manner rather than flushing large amounts of pages at one time.
Best practice
Use alternate page cleaners.

db2set DB2_USE_ALTERNATE_PAGE_CLEANING=YES

To modify the page cleaning algorithm, use the **db2set** command setting the `DB2_USE_ALTERNATE_PAGE_CLEANING` parameter to **YES**. See the DB2 Command references link in the **Resources** section of this article for instructions on using the `db2set` command.

Use DB2 relaxed logging

Relational databases such as DB2 maintain database recoverability in the case of a failure by using a transaction log. All transactions are recorded in the log and the logs are flushed to disk before the next transaction can be started. At high transaction rates the flush time can be substantial and resulting in a performance impact. You can choose to relax the flush of the transaction log with a DB2 configuration change.

Setting the `DB2_SKIP_LOG_WAIT` environment variable to **YES** can greatly speed up performance. The downside is that during processing, if the system fails, it is possible that you may have lost some transactions. In the event this occurs, restart the export from the beginning.

Best practice
Direct DB2 to use relaxed logging on the transactional log.

db2set DB2_SKIP_LOG_WAIT=YES

To modify the transaction log setting, use the **db2set** command setting the `DB2_SKIP_LOG_WAIT` parameter to **YES**. See the DB2 Command references link in the **Resources** section of this article for instructions on using the `db2set` command.

DB2 optimizer hint

Entity Analytics's use of DB2 includes statements of the form `select c1, c2, c3 from t1 where c1 in (?, ?, ?, ?, ?)`, a form which is called an **inlist**. DB2 will normally use a nested loop join operation when accessing data which is optimal for EA, but if the number of parameters in the inlist is larger than 32, DB2's optimizer will use a merge join which is not optimal for EA.

Best practice
Force DB2 optimizer to always use nested loop join for inlists.

db2set DB2_INLIST_TO_NLJN=YES

To force the DB2 optimizer to use nested loop join for inlists, use the **db2set** command setting the `DB2_INLIST_TO_NLJN` parameter to **YES**. See the DB2 Command references link in the **Resources** section of this article for instructions on using the `db2set` command.
Resources

- DB2 Command references
- DB2 10.5 Backup overview
About the author

Steve Schormann

Steve Schormann has been with IBM in various positions for 32 years. Most recently in the Identity Insight development group as senior performance engineer. Prior to this Steve was in the DB2 development group in Toronto, much of that time working with the performance group in IBM DB2 to achieve some world record performance benchmarks. Steve then was involved in database compatibility efforts within DB2, developing tools and assisting ISVs and customers in their migrations to DB2.

© Copyright IBM Corporation 2014
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)