**[Performance Best Practice: Flat File (Fast) Parse](http://microsoft-ssis.blogspot.com/2013/04/performance-best-practice-flat-file.html)**

Here are some tips to speed up the reading of flat files within the SSIS data flow. They are especially handy for importing large flat files (or when you merge join your small flat file to a large dataset). Start, where possible, by reading the files from a fast drive (a Solid State Disk / preferably not used by Windows or SQL Server) instead of some share.

* Optimize Buffer Size
* Use Source Query
* Adjust Row Size
* Adjust Column Size
* Use Where clause

**Minimum datatype**  
By default all columns are string 50 in a Flat File Connection Manager. To get as many rows in a [data flow buffer](http://microsoft-ssis.blogspot.com/2013/04/performance-best-practice-more-rows-per.html) it's important to use a minimum data type for each column. And if it's for example an integer column then parse it to an integer in the Flat File Connection Manager instead of parsing it in the data flow with a Data Conversion or Derived Column Transformation. Otherwise you end up with two columns in your buffer instead of one.  
  
A good start is to use the **Suggest Types** **button** in the Flat File Connection Manager editor. It will scan a couple of rows from your flat file and come up with a minimum data type.

|  |
| --- |
|  |
| Suggest Types |

**Unused columns**  
In the Flat File Connection Manager it's impossible to skip columns that you don't need, but to minimize pressure on performance you should not parse unneeded columns. Parsing/converting is expensive. **Just leave it string**. In the Flat File Source editor you can uncheck the unneeded columns.

|  |
| --- |
| [http://4.bp.blogspot.com/-YFmByXl-be8/UWa4XJSJIpI/AAAAAAAACHE/wD17OAbMHw0/s320/flatfile03.jpg](http://4.bp.blogspot.com/-YFmByXl-be8/UWa4XJSJIpI/AAAAAAAACHE/wD17OAbMHw0/s1600/flatfile03.jpg) |
| Leave it string, don't parse. |

|  |
| --- |
|  |
| Uncheck unneeded flat file columns |

**Fast Parse**  
If you have a so called 'trusted' source (for example a database export) then you don't have to worry for mistakes in the data types. To speed up the performance you can enable **fast parse** for all non-string columns. But for date columns you have to be sure that the [format is correct](http://msdn.microsoft.com/en-us/library/ms141005.aspx). Try using the ISO format YYYY-MM-DD.  
You can enable fast parse in the Advanced Editor of the Flat File Source. Right click it and choose "Show Advanced Editor...".

|  |
| --- |
| [http://3.bp.blogspot.com/-bPEW9TUhfqU/UWcNVl74feI/AAAAAAAACHc/Favh25_Ftd4/s320/flatfile05.jpg](http://3.bp.blogspot.com/-bPEW9TUhfqU/UWcNVl74feI/AAAAAAAACHc/Favh25_Ftd4/s1600/flatfile05.jpg) |
| Show Advanced Editor... |

Then go to the Input and Output Properties tab and then to the Output Columns. Select a non-string column and set the FastParse property to true. Repeat this for all non-string columns.

|  |
| --- |
| [http://1.bp.blogspot.com/-MIj_wFiE1K8/UWcN7N1pjoI/AAAAAAAACHk/oyuOd7yV6yg/s320/flatfile04.jpg](http://1.bp.blogspot.com/-MIj_wFiE1K8/UWcN7N1pjoI/AAAAAAAACHk/oyuOd7yV6yg/s1600/flatfile04.jpg) |
| Enable FastParse |

**Bulk Insert Task**  
If your destination is a SQL Server table and you don't need data transformations then you might want to consider/test the [Bulk Insert Task](http://msdn.microsoft.com/en-us/library/ms141239.aspx) as a alternative for the Data Flow Task.  
  
  
**You might also like:**

[Performance Best Practice: more rows per buffer](http://microsoft-ssis.blogspot.com/2013/04/performance-best-practice-more-rows-per.html)

[Performance Best Practice: Network Packet Size](http://microsoft-ssis.blogspot.com/2013/04/performance-best-practice-network.html)

[Performance Best Practices](http://microsoft-ssis.blogspot.com/2010/12/performance-best-practices.html)

[**Performance Best Practice: more rows per buffer**](http://microsoft-ssis.blogspot.com/2013/04/performance-best-practice-more-rows-per.html)

SSIS uses buffers to transport a set of rows through the data flow task. In general the less buffers you need to transport all rows from the source to the destination the faster it is.  
  
You can compare a buffer going through a data flow, with a fire bucket going through a medieval line of people trying to extinguish a fire.  
  
You can extinguish the fire faster by adding an extra line of people (parallel process) or my making the buckets larger (but not too large, otherwise they can't lift the buckets).  
  
For this example I use a table with [Google Analytics data](http://analyticssource.codeplex.com/) from my blog staged in a database table.

|  |
| --- |
| <http://2.bp.blogspot.com/-Nf6kJ1g42Lw/UWfZJnKZ32I/AAAAAAAACIE/ed_in4E_Rwk/s1600/buffers01.jpg> |
| Table with quite large columns |

**Changing Buffer Size**  
By default the buffer size is 10 Megabytes (1024 \* 1024 \* 10 = 10,485,760 bytes). You can adjust this size for each Data Flow Task (see Data Flow Task Property **DefaultBufferSize**).

|  |
| --- |
| [http://4.bp.blogspot.com/-gS_QtJ4Uv4Y/UWhRLAkRXqI/AAAAAAAACJ0/HeR7LTS4edk/s320/DefaultBufferSize.jpg](http://4.bp.blogspot.com/-gS_QtJ4Uv4Y/UWhRLAkRXqI/AAAAAAAACJ0/HeR7LTS4edk/s1600/DefaultBufferSize.jpg) |
| DefaultBufferSize 64KB <> 100MB |

I have a Package that pauses between buffers with a Script Component. This allows me to see how many records there are in one buffer. In my case 1 row of Google Analytics data is about 29000 bytes and if I run my Data Flow Task I can see that 360 rows of data fit in to my 10MB buffer.

|  |
| --- |
| [http://1.bp.blogspot.com/-W-bB92KpZkY/UWfqFoFaPNI/AAAAAAAACIU/6OIyi820SUs/s320/buffers02.jpg](http://1.bp.blogspot.com/-W-bB92KpZkY/UWfqFoFaPNI/AAAAAAAACIU/6OIyi820SUs/s1600/buffers02.jpg) |
| 360 rows = ±10MB |

By making my buffer twice the size you will see that the number of rows in my buffer doubles. You can change the buffer size up to 100MB, but you have to remember that there are multiple buffers going through your data flow at the same time. If you don't have enough memory to fit all those buffers, then SSIS will swap memory to disk making it very slow! So only use bigger buckets if you can lift them...

|  |
| --- |
| [http://2.bp.blogspot.com/-Bx6upmlip38/UWfsxx0KCII/AAAAAAAACIk/vHdxNczBcZc/s320/buffers03.jpg](http://2.bp.blogspot.com/-Bx6upmlip38/UWfsxx0KCII/AAAAAAAACIk/vHdxNczBcZc/s1600/buffers03.jpg) |
| 721 rows = ± 20MB |

**Adjust row size**  
Never use a table as a datasource! Selecting a table is akin to "SELECT \*..." which is universally recognized as bad practice(\*). Always use a query. This allows you to skip unwanted columns and rows.   
  
Unchecking "Available External Columns" at the Columns pane of the OLE DB Source Editor will prevent unwanted column in the buffer, but the data will have to go from SQL Server to SSIS. Removing unwanted column with a query will reduce the number of [network packets](http://microsoft-ssis.blogspot.com/2013/04/performance-best-practice-network.html) going from a SQL Server machine to your SSIS machine.  
﻿

|  |
| --- |
| [http://3.bp.blogspot.com/-piQfYGTfKlg/UWf18AJS4vI/AAAAAAAACI0/2A9UoZFuHbg/s640/buffers04.jpg](http://3.bp.blogspot.com/-piQfYGTfKlg/UWf18AJS4vI/AAAAAAAACI0/2A9UoZFuHbg/s1600/buffers04.jpg) |
| Use query instead of whole table |

Let's remove an unwanted column to see the effect. Removing the nvarchar(4000) column [ga:referralPath] will reduce the total recordsize from 29030 to 21028 bytes. Now 497 rows instead of 320 will fit in my 10MB buffer.

|  |
| --- |
| [http://4.bp.blogspot.com/-NU2lTyyFsbw/UWf-EHYWNVI/AAAAAAAACJE/zhi7X_sixW0/s320/buffers05.jpg](http://4.bp.blogspot.com/-NU2lTyyFsbw/UWf-EHYWNVI/AAAAAAAACJE/zhi7X_sixW0/s1600/buffers05.jpg) |
| Removing unwanted columns => more rows in buffer |

**Adjust Column Size**  
Often column are just unnecessary too big in the source. If you can't change the source then you could change it in the source query with a [CAST or CONVERT](http://microsoft-ssis.blogspot.com/2011/01/cast-and-convert-to-avoid-data.html). For example the nvarchar(4000) column for page titles is way to big. My titles are not that long and the characters should fit in varchar. Let's CAST it to varchar(255) and see how many extra rows will fit in the buffer.

|  |
| --- |
| [http://4.bp.blogspot.com/-N_3zyVDP9LE/UWhC-Iszj4I/AAAAAAAACJc/ikI-6Qs4ggM/s320/buffers06.jpg](http://4.bp.blogspot.com/-N_3zyVDP9LE/UWhC-Iszj4I/AAAAAAAACJc/ikI-6Qs4ggM/s1600/buffers06.jpg) |
| Resizing datatypes => more rows in buffer |

Now we have 787 rows instead of 320 rows in the 10mb buffer. That should speed up this package!  
  
**Note:**if you change datatypes in an existing source query, then you have to uncheck and check the changed columns in the Columns pane of the OLE DB Source editor. This will refresh the metadata in the data flow. Otherwise you will get an error like: *Validation error. DFT - Visits: DFT - Visits: Column "ga:pageTitle" cannot convert between unicode and non-unicode string data types.*

|  |
| --- |
| [http://4.bp.blogspot.com/-SmCN212dwdQ/UWhBrkFL_JI/AAAAAAAACJU/PuFdK8UqSfw/s400/ColumnsUncheck.gif](http://4.bp.blogspot.com/-SmCN212dwdQ/UWhBrkFL_JI/AAAAAAAACJU/PuFdK8UqSfw/s1600/ColumnsUncheck.gif) |
| Uncheck and check changed columns |

**Add WHERE clause**  
Don't use a Conditional Split Transformation to remove records after an database table source. Instead use a WHERE clause in your source query. The Conditional Split is a so called Magic Transformation. It will only visually hide records instead of removing them from the buffer. And the records are unnecessary transported from SQL Server to SSIS.

|  |
| --- |
| [http://1.bp.blogspot.com/-L_FtCeJe2Mo/UWnODOoPz4I/AAAAAAAACKc/AeEicDdHff4/s320/buffers07.jpg](http://1.bp.blogspot.com/-L_FtCeJe2Mo/UWnODOoPz4I/AAAAAAAACKc/AeEicDdHff4/s1600/buffers07.jpg) |
| Add WHERE clause instead of Conditional Split |

**RunInOptimizedMode**   
This is a Data Flow Task property that, if set to true, will 'remove' unused columns and unused transformations to improve performance. Default value is true, but this property isn't the holy grail and you shouldn't rely on it while developing your packages.

|  |
| --- |
| [http://2.bp.blogspot.com/-BwTT7wVn_C8/UWm-wuDZbaI/AAAAAAAACKM/hPkE6e6pL3o/s320/RunInOptimizedMode.jpg](http://2.bp.blogspot.com/-BwTT7wVn_C8/UWm-wuDZbaI/AAAAAAAACKM/hPkE6e6pL3o/s1600/RunInOptimizedMode.jpg) |
| RunInOptimizedMode |

**Enable logging on Data Flow Task**  
There are some events that you could log to get information about the buffer:  
User:BufferSizeTuning: This will tell you how many rows there were in the buffer.

* Rows in buffer type 0 would cause a buffer size greater than the configured maximum. There will be only 787 rows in buffers of this type.
* Rows in buffer type 0 would cause a buffer size greater than the configured maximum. There will be only 1538 rows in buffers of this type.

User:PipelineInitialization: This will tell you something about which settings are used.

* The default buffer size is 10485760 bytes.
* Buffers will have 10000 rows by default.
* The data flow will not remove unused components because its RunInOptimizedMode property is set to false.

|  |
| --- |
| [http://2.bp.blogspot.com/-cgpT7JKpKH0/UWm-MlTNyJI/AAAAAAAACKE/7CZa4LpUXjU/s320/BufferLogging.jpg](http://2.bp.blogspot.com/-cgpT7JKpKH0/UWm-MlTNyJI/AAAAAAAACKE/7CZa4LpUXjU/s1600/BufferLogging.jpg) |
| BufferSizeTuning and PipelineInitialization |

**Performance Counters**  
SSIS comes with a set of performance counters that you can use to monitor the performance of the Data Flow Task. More about that in a later blog. For now see [msdn](http://msdn.microsoft.com/en-us/library/ms137622.aspx" \t "_blank).