Test 1 : read 2.4GB file with 25M customer records

with CSV input, lazy conversion, 500k NIO buffer

Iteration	Duration	Speed (r/s)	Disk throughput : between 9 and 20 MB/s
1	165,7	150.875	CPU Utilisation: between 50 and 60%
2	146,0	171.233	
3	147,6	169.377	
4	146,2	170.999	
5	148,7	168.124	
Average	***150,8***	***166.121***	***1,0 GB/minute***

Test 2 : read 2.4GB file with 25M customer records

with CSV input, lazy conversion, 50M NIO buffer, 2 step copies

Iteration	Duration	Speed (r/s)	Disk throughput : between 22 and 32 MB/s
1	93,1	268.528	CPU Utilisation: between 70 and 100%
2	94,1	265.675	
3	95,4	262.055	
4	98,3	254.323	
5	90,2	277.162	
Average	***94,2***	***265.549***	***1,5 GB/minute***

Remark: Adding more step copies doesn't help : the hard disk can't read faster

Test 3 : read 2.4GB file, write back to different disk

with CSV input, lazy conversion, fast data dump, 2 readers, 1 writer

Iteration	Duration	Speed (r/s)	Disk throughput (iostat -k 5) is :	
1	196,0	127.551	Read:	Between 12 and 18MB/s
2	198,5	125.945	Write:	Around 14MB/s sustained
3	194,8	128.337		
4	194,4	128.601	CPU Utilisation: between 90 and 115%	
5	197,4	126.646		
Average	***196,2***	***127.416***	***0,7 GB/minute***	

Test 4 : read 2.4GB file, write back to different disk

with CSV input, lazy conversion, fast data dump, 2 readers, 2 writers to 2 files

Iteration	Duration	Speed (r/s)	Disk throughput (iostat -k 5) is :	
1	154,7	161.603	Read:	Between 12 and 18MB/s
2	147,9	169.033	Write:	Around 18MB/s sustained
3	147,5	169.492		
4	148,6	168.237	CPU Utilisation: between 90 and 130%	
5	147,9	169.033		
Average	***149,3***	***167.480***	***1,0 GB/minute***	

Test 1 : read 2.4GB file with 25M customer records

With a single delimited reader

Iteration	Duration	Speed (r/s)
1	114,5	218.341
2	109,0	229.415
3	117,5	212.780
4	111,6	223.994
5	108,3	230.947
Average	**112,2**	**223.095**

Disk throughput : between 9 and 20 MB/s
CPU Utilisation: between 75 and 85%

1,3 GB/minute

Test 2 : read 2.4GB file with 25M customer records

with multiple delimited readers

This was not possible with TOS or I couldn't find the options.

Test 3 : read 2.4GB file, write back to different disk

with multiple delimited readers and one writer

This was not possible with TOS or I couldn't find the options.

Test 4 : read 2.4GB file, write back to different disk

with CSV input, lazy conversion, fast data dump, 2 readers, 2 writers to 2 files

This was not possible with TOS or I couldn't find the options.

Test 5 : read 2.4GB file, write back to different disk

With single delimited read and writer

Iteration	Duration	Speed (r/s)
1	321,9	77.670
2	331,2	75.483
3	331,6	75.384
4	330,9	75.552
5	331,3	75.460
Average	***329,4***	***75.910***

Note: disk throughput (iostat -k 5) is :
Read: Between 6 and 10MB/s
Write: Around 8MB/s sustained

CPU Utilisation: between 95 and 99%

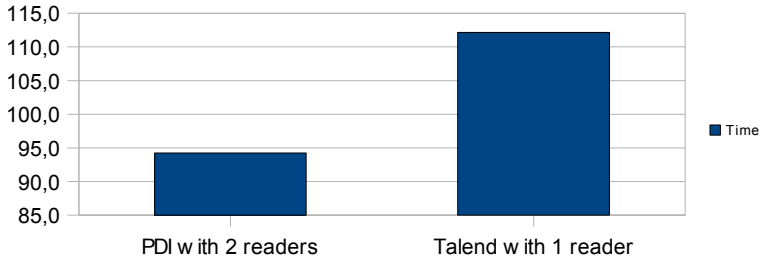
0,4 GB/minute

Comparing numbers: difficult because of different architectures.

Let's compare the best of the results:

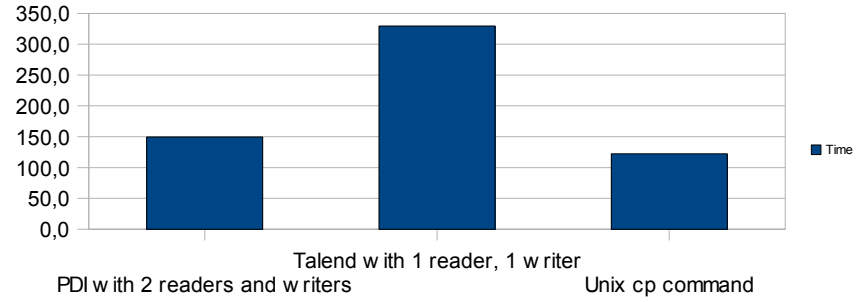
Description	**Time**
PDI with 2 readers	94,2
Talend with 1 reader	112,2

Difference : PDI is ***19% faster***



Reading and writing back	**Time**
PDI with 2 readers and writers	149,3
Talend with 1 reader, 1 writer	329,4
Unix cp command	122,2

Difference : PDI is ***121% faster***



Other interesting notes:

Talend uses a maximum of between 80-99% CPU (not multi-threaded)
PDI uses a maximum of 130% CPU (1.3 CPU used) in test 4
Talend is faster in the single threaded reading of a file.
The file size (2.4GB) is large enough to NOT fit into the cache.

Specs

CPU	Intel(R) Core(TM)2 CPU T7600 @ 2.33GHz
Disk	90GB 7200 rpm laptop disk
Memory	3.3GB, 666Mhz
OS	Kubuntu 8.10 : Intrepid Ibex
Linux kernel	2.6.27-8
Filesystem used	ext3
External USB disk	USB 2.0, 100GB, ext3 formatted, used to write target file to
Source file:	http://mattcasters.s3.amazonaws.com/customers-25M.txt
Size file	2.614.561.970 bytes
Nr of rows in file	25.000.001 with one header row