TIMi

CREATIVITY THROUGH EFFICIENCY

An automated predictive datamining tool

Server/Infrastructure Selection for TIMi v1.13

Creation Date: September 2012
Last Edited: January 2022

Introduction

This document is a guide on how to select a good hardware infrastructure for TIMi.

The optimal hardware infrastructure for TIMi is composed of several PC's (laptops & servers) interconnected together. This documents contains 4 sections:

- Section 1 gives advices on which PC to buy to run TIMi efficiently.
- Section 2 explains the different roles of each PC in the global infrastructure.
- Section 3 explains how this infrastructure integrates with other tools such as Hadoop, Ordinary BI tools (Kibella, Tableau, Qlick, etc.) and Jenkins (Jenkins is a scheduler that allows to run TIMi-based batch job automatically each day, week, month).
- Section 4 summarizes this document in a global info-graphic.

1. How to select a good machine to run TIMi?

1.1. Minimum Requirements

Minimum system Requirements for TIMi installation on a server or workstation are:

OS: Windows 2000, XP, Vista, Win7, Windows8, Win10

Strongly advised: No virtual machine: By default, TIMi won't run in a VMWare virtual Machine or equivalent software. If you need to run TIMi in a VMWare virtual Machine, **please request a Network License** (the Network License is free when you buy a TIMi license).

RAM: Minimum 2GB per simultaneous user. Large data-transformations require more RAM.

Video Card: To use StarDust, you need a 3D-hardware-accelerated-graphical card that supports OpenGL2.0 (any computer produced after 2007 will do the trick). The specific OpenGL drivers from you video card manufacturer must be installed (do not use "generic" OpenGL drivers).

For increased processing speed, we strongly suggest a server with a good SSD drive (Samsung) and a multi-core CPU (TIM is 100% mutli-threaded). A server with an "Intel Core I7 at 3.0 GHz (or above)" CPU is always a good option. To process large volumes of data with Anatella (to do "Big Data" analytics), you also need a 64-bit OS and plenty of RAM (16GB or 32GB).

1.1.1. Minimum requirement: Anatella.

Although Anatella (our ETL) is able of achieving 99% of the required data transformations on "simple commodity PC's" (typically we are using the simple laptops from the dataminers), it is nevertheless advisable to use a server with a large quantity of RAM (16 GB or 32GB) for transformations involving tables containing tens of billions of rows.

1.1.2. Minimum requirement: Modeler.

TIMi Modeler is the only "Machine Learning tool" that is 100% multi-threaded. In other words: TIMi Modeler performs its computations using all the available CPUs on the server. This means that the computations made by a user who is "alone" on a quad-core server will be four times faster (approximately) than when there are 4 users working simultaneously on the server. In contrast, other datamining softwares perform all their computation using a single CPU. This means that when a user is "alone" on a quad-core server, the computation time (of the software competitors) is the same as when there are 4 users on the server (because 3 of the 4 CPUs remain unused).

This unique feature of TIMi Modeler affects the QoS ("Quality of Service") provided by TIMi Modeler: a limited number of users on each server greatly improves the computation speed and therefore the QoS. Therefore, a larger number of TIMi servers (and CPU's) provide a higher QoS. The purchase of additional "TIMI" servers is nearly always justified (to improve the QoS).

On a given hardware, the computation time of TIMi Modeler mainly depends on the size of the analyzed datasets. When handling large datasets, computation time increases. When working on very large datasets, to obtain a satisfactory QoS (i.e. a computation time that is reasonable), it's necessary to provide more CPU resources to the TIMi users.

Here is a table that summarizes the situation:

Dataset Size	Minimum Required RAM per concurrent user for TIMi modeler	Minimum Required CPU Resources For TIMi Modeler
Large	2 GB	from 0.4 to 8 CPU's per concurrent user (ideally 1 CPU/ concurrent user)
		On a quadcore server: 1 to 10 concurrent users (for a good QoS: 4 simultaneous users)
Small (less than 1 MB)	100 MB	from 1 to 110 simultaneous users on a quadcore server

The recommendations given the table above reflects the fact that Modeler's users are usually analyzing datasets at the Gigabyte size (common volumes are: ¼, ½, 1, 2, 5, 10, 20 GB), which is a common situation for data mining analysis, yet quite exceptional for "old" statistical packages... To ensure a greater modeling accuracy (and therefore a higher ROI), TIMi practically never does any sampling and always work on the "full" data set (thanks to its unique compression algorithm, TIMi can store in internal RAM datasets of several dozen gigabytes). This approach is more costly in CPU and RAM but consistently deliver superior models and thus a higher ROI.

1.2. Introduction to the Optimal hardware selection for TIMi

The objective of this document is to help you select the best server for an efficient working environment with ***The Intelligent Mining Machine (TIMi).*** The TIMi software solution contains 4 tools: Anatella, TIMi Modeler, Stardust and Kibella.

Fortunately, these four tools share the same needs in terms of hardware. More precisely, the main bottleneck (that slows down all computations) when computing some results with Anatella, TIMi, Stardust or Kibella is nearly always the CPU.

The main limiting factor in terms of processing speed for Anatella is usually not the hard drive (as it's the case with other ETL's) but rather the processor (i.e. the CPU) that performs the computations. Since "simple modern laptops" are now usually provided with rather slow hard drive but with good processors (i.e. they are equipped with Intel Core i7 and Intel Core i5 3GHz processors). These "simple laptops" are good candidates to run Anatella.

Thanks to its unique data compression technology, Anatella reduces to the minimum the bandwidth used on your corporate computer network. So there is no reason to let idle the good processors on your dataminer's computers.

Some of our clients have done several benchmarks that demonstrates that Anatella installed on one simple laptop equipped with a CoreI7 CPU and 32GB RAM is approximately three times faster than a "1 million euro" Oracle Exadata solution. Thus, the total available "computing power" that you get when installing Anatella on each&every dataminer laptop is ***worth several millions euros when purchased from another vendor***.

To Summarize: The possibility to use Anatella directly on the dataminers laptops allows you to provide to your dataminer team a very comfortable working environment (i.e. it allows you to achieve a very high QoS = "Quality of Service") and a tremendous computing power.

In an attempt to better extract all the power available inside the CPU (since the CPU is the main bottleneck), we created tools that are heavily multithreaded (i.e. under some specific conditions the tools might use several CPU cores simultaneously). Despite using inside our tool the most advanced multithreading techniques (e.g. lock-free code), we still advise our customer to buy CPU that have the best speed on ONE core (i.e. single-core execution) because of the large overhead that multithreading and parallel computations always exhibit.

Let's give an example: 99% of the time, it's more efficient (i.e. faster) to run all the computations on one core on a (fast) "Intel Core I7 at 4GHz" (4GHz is the speed of the CPU) than to run the same computation using 3 cores/CPU on a (slower) "Intel Core I7 at 2GHz" (despite the fact that 3x2GHz > 4GHz). This happens because of the multithreading overhead. This overhead is analysed and explained in more details in this youtube video.

To summarize, when buying some hardware to run TIMi & Anatella, you should buy a machine with a fast CPU (more precisely, you should pay attention to the "single-core execution" speed of the CPU): The objective of this section is to help you on this process.

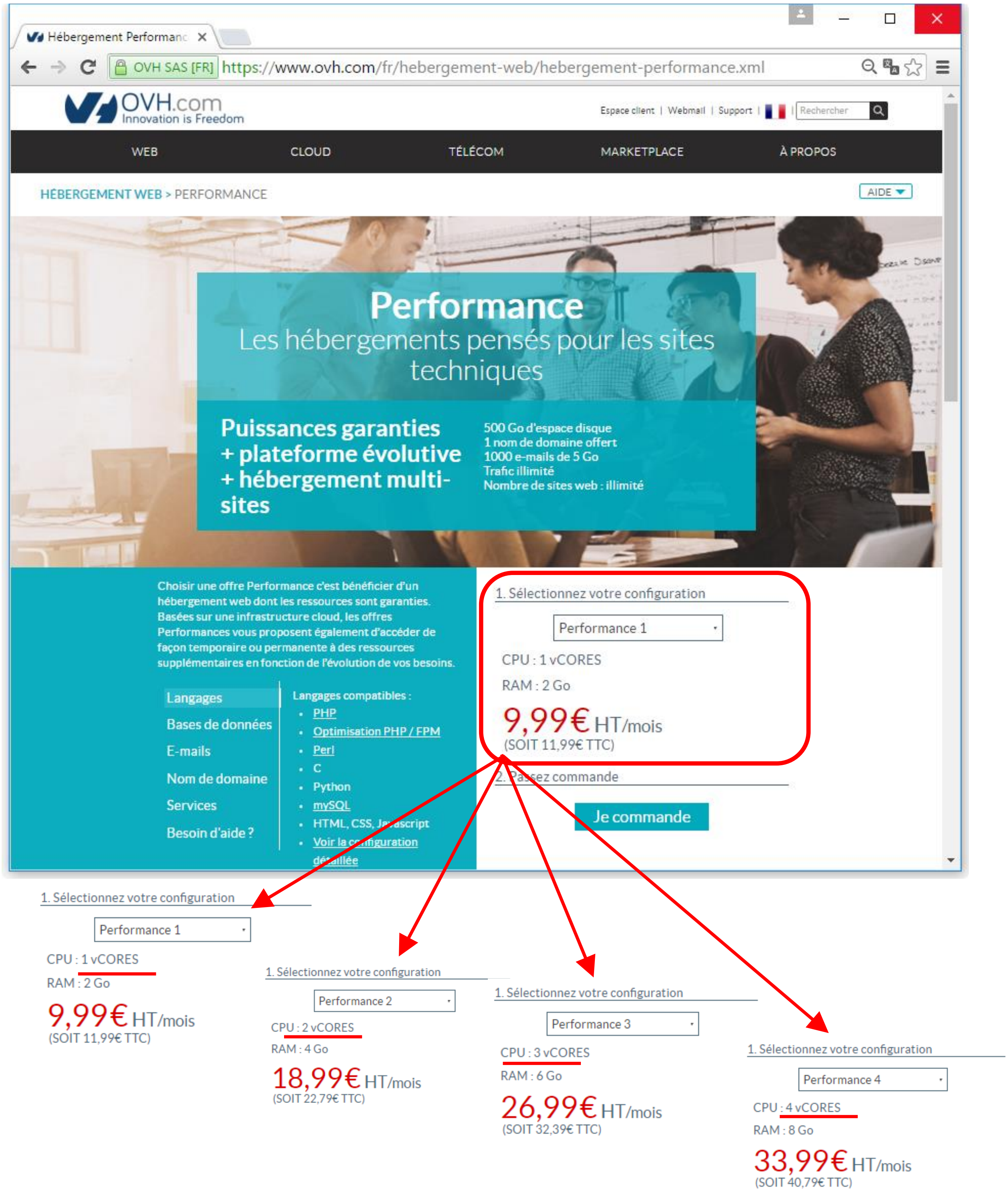
Most of the time, one of the best solution is simply to give to each of the analysts good "Core I7" laptops. In this way, each analyst has the usage of 100% of its own powerful Core 17 CPU: for more information about this subject/infrastructure, see the section 2.1 here below.

To help you select the right CPU for your server, we advise you to use the "geek benchmark" website that list the ("single-core") performances of all current CPU's.

1.3. About common “Big” Servers in Data Centers

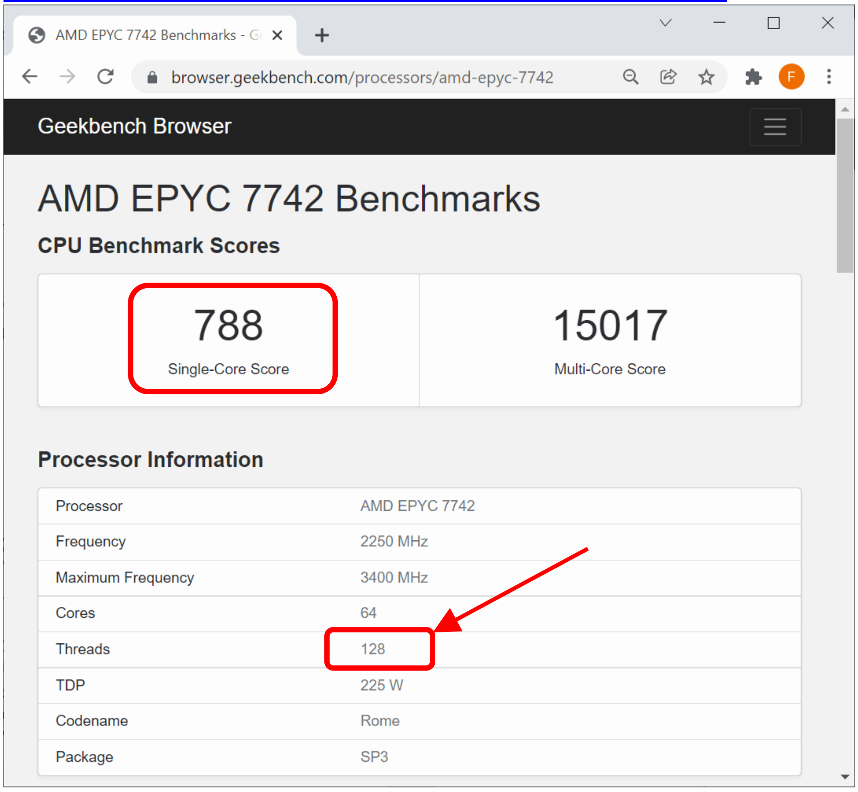
Unfortunately, currently, most “big servers” in large data centers are optimized to be “web servers”. Such type of “big servers” possess very efficient hard drives but always some VERY BAD cpu. Thus, if you decided to install TIMi inside a data center, it most certainly means that you selected a VERY BAD machine for TIMi.

Basically, it's always the same thing: The owners of the "data centers" want to tell their customers that they will have their own "**vcore**/CPU" dedicated for themselves only. Here is a good example:



To be able to provide to all their customers with this really large number of cores, the data center's owners are buying (or building) servers based on processors such as this one:

https://browser.geekbench.com/processors/amd-epyc-7742



Please note in the above screenshot the very low "single-core" Score: 788 (less than 1000: pitiful!).

This means that the speed of each of the (128) vcores/threads inside this server is really slow (but there are many of them!). To summarize, the data center shows you an advertisement that guarantees to you your own, private vcore but nobody said that this vcore will be fast! You must also realize that, very often, the situation is even worse than the above screenshot: The CPU's inside common servers in data centers are, most of the time, far far worse.

What's happening if you install Anatella/TIMi on a server using the "AMD EPYC 7742" cpu described here above? Everything will be very, very slow compared to a small, inexpensive "stupid" Core I7 laptop. The only way to get this (extremely expensive) server running a little bit faster than the "stupid" Core I7 laptop would be to use many cores (i.e. more than 8) at the same time (since this server has 160 vcores/threads compared to a simple "Core I7" laptop that still has already 8 vcores/threads). For example, you could try to run many graphs at the same time (at least more than 8). Unfortunately, in reality, this situation almost never happens: 99% of the time, the order in which the graphs are executed is just Sequential: one graph than another graph and so on… (i.e. most of the time, you run one graph at-a-time).

1.4. About XEON processors

The "professional, high-grade" CPU's for data centers are, most of time, XEON processors (and not Core I7 or Core I9 processors). What to think about XEON processors in general? Let's look at the GeekBench website to have an answer (the table below has been extracted on the 2022/1/17 from the page https://browser.primatelabs.com/processor-benchmarks) :

The best CPU on the 2022/1/17 for Anatella/TIMi

#	CPU Name	Frequency [GHz]	Core count	Score
1	Intel Core i9-12900K	3.2	16	2000
2	Intel Core i9-12900KF	3.2	16	1966
3	Intel Core i7-12700KF	3.6	12	1913
4	Intel Core i7-12700K	3.6	12	1905
5	Intel Core i5-12600K	3.7	10	1858
6	Intel Core i9-11900K	3.5	8	1840
7	Intel Core i9-11900KF	3.5	8	1751
8	Intel Core i9-11900F	2.5	8	1704
9	Intel Core i7-11700K	3.6	8	1695
10	Intel Core i7-11700KF	3.6	8	1695
11	AMD Ryzen 9 5950X	3.4	16	1687
12	Intel Core i9-11900	2.5	8	1678
13	Intel Core i5-11600K	3.9	6	1673
14	AMD Ryzen 9 5900X	3.7	12	1672
15	AMD Ryzen 7 5800X	3.8	8	1672
16	Intel Core i5-11600KF	3.9	6	1652
17	Intel Core i7-11700	2.5	8	1620
18	Intel Core i9-11900T	1.5	8	1619
19	AMD Ryzen 5 5600X	3.7	6	1615
20	Intel Core i5-11600	2.8	6	1604
21	Intel Core i7-11700F	2.5	8	1579
22	Intel Core i5-11500	2.7	6	1551
23	AMD Ryzen 7 5700G	3.8	8	1539
24	Intel Core i5-11400	2.6	6	1489
25	AMD Ryzen 5 5600G	3.9	6	1480
26	Intel Core i7-11700T	1.4	8	1477
27	Intel Core i5-11400F	2.6	6	1468
28	Intel Core i7-11800H	2.3	8	1466
29	Intel Core i7-1195G7	2.9	4	1451
30	Intel Core i7-11375H	3.3	4	1430
31	AMD Ryzen 9 5900HX	3.3	8	1417
32	Intel Core i7-1185G7	3.0	4	1406
33	Intel Core i7-11370H	3.3	4	1403
34	Intel Core i9-10900K	3.7	10	1393
35	Intel Core i7-1185G7E	2.8	4	1382
36	Intel Core i9-10900KF	3.7	10	1379
37	Intel Core i7-1165G7	2.8	4	1379
38	Intel Core i9-9900KS	4.0	8	1367
39	Intel Core i9-10850K	3.6	10	1362
40	AMD Ryzen 7 3800XT	3.9	8	1341
41	Intel Core i7-10700K	3.8	8	1337
42	AMD Ryzen 7 5800H	3.2	8	1331
43	Intel Core i9-10910	3.6	10	1324
44	Intel Core i9-10900F	2.8	10	1323
45	Intel Core i7-10700KF	3.8	8	1321
46	AMD Ryzen 9 3900XT	3.8	12	1319
47	Intel Core i9-9900KF	3.6	8	1318
48	Intel Core i5-1145G7	2.6	4	1318
49	Intel Core i7-8086K	4.0	6	1317
50	Intel Core i9-10900	2.8	10	1315
51	Intel Core i5-11400H	2.7	6	1312
52	Intel Core i9-9900K	3.6	8	1310
53	Intel Core i5-10600K	4.1	6	1308
54	Intel Core i5-10600KF	4.1	6	1307
55	AMD Ryzen 5 3600XT	3.8	6	1302
56	AMD Ryzen 9 3950X	3.5	16	1296
57	Intel Core i5-11300H	3.1	4	1291
58	AMD Ryzen 7 3800X	3.9	8	1289
59	AMD Ryzen 7 5800U	1.9	8	1288
60	Intel Core i7-9700KF	3.6	8	1284
61	Intel Core i7-9700K	3.6	8	1281
62	AMD Ryzen 9 3900X	3.8	12	1277
63	Intel Core i3-9350KF	4.0	4	1276
64	AMD Ryzen Threadripper 3960X	3.8	24	1270
65	Intel Core i7-7740X	4.3	4	1269
66	AMD Ryzen 7 PRO 5850U	1.9	8	1269
67	AMD Ryzen 3 3300X	3.8	4	1265
68	Intel Core i7-10700	2.9	8	1259
69	AMD Ryzen Threadripper 3970X	3.7	32	1257
70	Intel Core i7-10700F	2.9	8	1255
71	AMD Ryzen 7 3700X	3.6	8	1254
72	Intel Core i9-9900	3.1	8	1252
73	Intel Core i5-1135G7	2.4	4	1249
74	Intel Core i7-8700K	3.7	6	1246
75	AMD Ryzen 5 3600X	3.8	6	1246
76	Intel Core i5-9600KF	3.7	6	1246
77	AMD Ryzen 5 5600H	3.3	6	1245
78	Intel Core i5-1145G7E	1.5	4	1243
79	Intel Core i5-11260H	2.6	6	1243
80	Intel Core i5-9600K	3.7	6	1235
81	Intel Core i5-10600	3.3	6	1234
82	Intel Core i3-1115G4E	2.2	2	1233
83	Intel Core i7-1068NG7	2.3	4	1228
84	Intel Core i3-10320	3.8	4	1217
85	AMD Ryzen 9 3900	3.1	12	1216

The best value CPU on the 2022/1/17 for Anatella/TIMi

86	Intel Core i7-7700K	4.2	4	1213
87	AMD Ryzen Threadripper 3990X	2.9	64	1213
88	AMD Ryzen 5 5600U	2.3	6	1213
89	Intel Core i5-8600K	3.6	6	1212
90	Intel Core i3-7350K	4.2	2	1205
91	AMD Ryzen 5 3600	3.6	6	1205
92	Intel Core i7-9700F	3.0	8	1204
93	Intel Core i9-10885H	2.4	8	1203
94	AMD Ryzen 7 PRO 4750G	3.6	8	1201
95	Intel Core i5-9600	3.1	6	1200
96	**Intel Xeon E-2288G**	**3.7**	**8**	**1200**
97	Intel Core i7-9700	3.0	8	1197
98	Intel Core i9-10920X	3.5	12	1193
99	Intel Core i7-10875H	2.3	8	1193
100	Intel Core i9-10980XE	3.0	18	1184
101	Intel Core i5-7640X	4.0	4	1179
102	Intel Core i3-8350K	4.0	4	1177
103	AMD Ryzen 5 3500X	3.6	6	1173
104	Intel Core i5-10500	3.1	6	1173
105	Intel Core i3-1115G4	3.0	2	1170
106	**Intel Xeon W-2295**	**3.0**	**18**	**1170**
107	Intel Core i3-10300	3.7	4	1170
108	AMD Ryzen 5 PRO 4650G	3.7	6	1169
109	Intel Core i7-8700B	3.2	6	1168
110	Intel Core i7-9800X	3.8	8	1162
111	Intel Core i7-10870H	2.2	8	1161
112	Intel Core i9-9940X	3.3	14	1159
113	Intel Core i9-10940X	3.3	14	1159
114	Intel Core i5-7600K	3.8	4	1158
115	Intel Core i9-10900X	3.7	10	1157
116	Intel Core i7-8700	3.2	6	1155
117	Intel Core i7-10850H	2.7	6	1153
118	Intel Core i3-1125G4	2.0	4	1141
119	Intel Core i7-6700K	4.0	4	1136
120	AMD Ryzen 5 3500	3.6	6	1134
121	Intel Core i7-10750H	2.6	6	1134
122	AMD Ryzen 9 4900H	3.3	8	1134
123	Intel Xeon E-2276M	2.8	6	1131
124	AMD Ryzen 3 3100	3.6	4	1130
125	Intel Core i5-11500T	1.5	6	1128
126	Intel Core i9-9900X	3.5	10	1126
127	Intel Core i9-9980XE	3.0	18	1126
128	Intel Core i7-8569U	2.8	4	1125
129	Intel Core i5-1038NG7	2.0	4	1125
130	Intel Core i9-7940X	3.1	14	1122
131	Intel Core i9-7960X	2.8	16	1116
132	Intel Xeon W-3245	3.2	16	1115
133	Intel Xeon W-3265M	2.7	24	1115
134	Intel Core i5-8600	3.1	6	1114
135	Intel Core i9-9980HK	2.4	8	1114
136	Intel Core i5-9500	3.0	6	1113
137	Intel Core i9-7980XE	2.6	18	1111
138	Intel Core i9-9960X	3.1	16	1109
139	Intel Core i9-7920X	2.9	12	1108
140	Intel Core i5-10400	2.9	6	1108
141	Intel Core i9-7900X	3.3	10	1107
142	Intel Core i5-10400F	2.9	6	1107
143	Intel Core i9-9920X	3.5	12	1106
144	Intel Core i7-7820X	3.6	8	1103
145	**Intel Xeon W-2145**	**3.7**	**8**	**1103**
146	**Intel Xeon W-3235**	**3.3**	**12**	**1102**
147	Intel Core i7-7800X	3.5	6	1099
148	Intel Core i5-6600K	3.5	4	1095
149	Intel Core i3-10100	3.6	4	1095
150	Intel Core i3-10100F	3.6	4	1092
151	AMD Ryzen 9 4900HS	3.0	8	1081
152	Intel Core i7-9850H	2.6	6	1079
153	Intel Xeon W-3175X	3.1	28	1079
154	AMD Ryzen 7 4800H	2.9	8	1079
155	**Intel Xeon W-2150B**	**3.0**	**10**	**1076**
156	**Intel Xeon W-3275M**	**2.5**	**28**	**1074**
157	**Intel Xeon W-2191B**	**2.3**	**18**	**1073**
158	Intel Core i7-8559U	2.7	4	1071
159	**Intel Xeon W-2135**	**3.7**	**6**	**1071**
160	AMD Ryzen Threadripper 2920X	3.5	12	1070
161	**Intel Xeon E-2136**	**3.3**	**6**	**1070**
162	Intel Core i9-8950HK	2.9	6	1067
163	Intel Core i9-9880H	2.3	8	1065
164	**Intel Xeon W-2155**	**3.3**	**10**	**1064**
165	AMD Ryzen Threadripper 2950X	3.5	16	1063
166	Intel Core i5-10300H	2.5	4	1063
167	Intel Core i5-8500B	3.0	6	1062
168	Intel Core i5-8500	3.0	6	1061
169	Intel Core i3-9100	3.6	4	1060
170	Intel Core i7-4790K	4.0	4	1059
171	Intel Core i5-9400F	2.9	6	1058
172	Intel Core i7-9700T	2.0	8	1054

173	AMD Ryzen 7 4800HS	2.9	8	1051
174	**Intel Xeon E3-1270 v6**	**3.8**	**4**	**1050**
175	**Intel Xeon E5-1660 v3**	**3.0**	**8**	**1047**
176	Intel Core i5-9400	2.9	6	1047
177	Intel Core i7-9750H	2.6	6	1045
178	Intel Core i7-7700	3.6	4	1044
179	**Intel Xeon E3-1245 v6**	**3.7**	**4**	**1044**
180	**Intel Xeon W-2170B**	**2.5**	**14**	**1043**
181	Intel Core i7-8809G	3.1	4	1042
182	**Intel Xeon W-2140B**	**3.2**	**8**	**1041**
183	**Intel Xeon E3-1240 v6**	**3.7**	**4**	**1041**
184	**Intel Xeon E3-1275 v6**	**3.8**	**4**	**1040**
185	AMD Ryzen 7 2700X	3.7	8	1039
186	Intel Core i3-9100F	3.6	4	1039
187	Intel Core i5-8400	2.8	6	1036
188	**Intel Xeon W-3223**	**3.5**	**8**	**1036**
189	AMD Ryzen 7 4800U	1.8	8	1036
190	Intel Core i7-8700T	2.4	6	1035
191	AMD Ryzen 5 2600X	3.6	6	1034
192	Intel Core i5-7600	3.5	4	1030
193	AMD Ryzen 7 5700U	1.8	8	1028
194	**Intel Xeon E-2176M**	**2.7**	**6**	**1026**
195	Intel Core i3-6320	3.9	2	1023
196	AMD Ryzen 7 PRO 4750U	1.7	8	1023
197	Intel Core i3-7300	4.0	2	1022
198	Intel Core i7-1060NG7	1.2	4	1022
199	Intel Xeon W-2125	4.0	4	1019
200	Intel Core i7-5775C	3.3	4	1017
201	AMD Ryzen Threadripper 2990WX	3.0	32	1014

202	**Intel Xeon E3-1535M v6**	**3.1**	**4**	**1009**
203	Intel Core i7-8850H	2.6	6	1007
204	Intel Core i5-5675C	3.1	4	1006
205	Intel Core i5-8279U	2.4	4	1002
206	**Intel Xeon E3-1280 v6**	**3.9**	**4**	**1000**
207	AMD Ryzen Threadripper 2970WX	3.0	24	999
208	**Intel Xeon E3-1285L v4**	**3.4**	**4**	**998**
209	AMD Ryzen Threadripper 1900X	3.8	8	996
210	Intel Core i3-1005G1	1.2	2	992
211	Intel Core i7-6850K	3.6	6	989
212	AMD Ryzen 3 2300X	3.5	4	989
213	AMD Ryzen 5 4600H	3.0	6	989
214	Intel Core i5-4690K	3.5	4	987
215	Intel Core i7-7820HK	2.9	4	986
216	**Intel Xeon Gold 6134**	**3.2**	**8**	**983**
217	Intel Core i7-6900K	3.2	8	982
218	AMD Ryzen Threadripper 1920X	3.5	12	982
219	AMD EPYC 7542	2.9	32	981
220	Intel Xeon E3-1585L v5	3.0	4	980
221	**Intel Xeon Gold 6128**	**3.4**	**6**	**979**
222	**Intel Xeon E-2146G**	**3.5**	**6**	**978**
223	**Intel Xeon Gold 6254**	**3.1**	**18**	**977**
224	Intel Core i7-8750H	2.2	6	976
225	**Intel Xeon E3-1275 v3**	**3.5**	**4**	**975**
226	**Intel Xeon E3-1280 v3**	**3.6**	**4**	**975**
227	AMD Ryzen 5 4600HS	3.0	6	975
228	**Intel Xeon E3-1240 v5**	**3.5**	**4**	**974**
229	Intel Core i7-4770K	3.5	4	970
230	Intel Core i7-6700	3.4	4	970

231	**Intel Xeon E3-1275 v5**	**3.6**	**4**	**969**
232	AMD Ryzen 7 4700U	2.0	8	969
233	AMD Ryzen Threadripper 1950X	3.4	16	968
234	Intel Core i3-8300	3.7	4	967
235	Intel Core i7-6950X	3.0	10	966
236	AMD Ryzen 5 2600	3.4	6	966
237	**Intel Xeon E3-1271 v3**	**3.6**	**4**	**965**
238	Intel Core i5-9300HF	2.4	4	965
239	Intel Core i5-1030NG7	1.1	4	964
240	Intel Core i7-7920HQ	3.1	4	963
241	AMD Ryzen 5 2500X	3.6	4	963
242	AMD Ryzen 5 4600U	2.1	6	962
243	Intel Core i7-8705G	3.1	4	961
244	Intel Core i3-8100B	3.6	4	960
245	Intel Core i5-4670K	3.4	4	959
246	Intel Core i5-7500	3.4	4	959
247	Intel Core i7-7820HQ	2.9	4	959
248	Intel Core i5-8400H	2.5	4	959
249	Intel Core i5-6600	3.3	4	957
250	**Intel Xeon W-2133**	**3.6**	**6**	**957**
251	Intel Core i7-4910MQ	2.9	4	956
252	Intel Core i5-8257U	1.4	4	953
253	AMD Ryzen 5 PRO 4650U	2.1	6	952
254	Intel Core i7-7567U	3.5	2	950
255	Intel Core i5-9300H	2.4	4	949
256	Intel Core i5-10210U	1.6	4	948
257	Intel Core i5-8259U	2.3	4	947
258	AMD Ryzen 5 1600X	3.6	6	945
259	Intel Core i3-10100T	3.0	4	945

260	Intel Xeon E3-1230 v6	3.5	4	944
261	Intel Xeon E3-1225 v6	3.3	4	944
262	Intel Core i7-7700T	2.9	4	943
263	Intel Core i5-9500T	2.2	6	942
264	AMD Ryzen 7 1800X	3.6	8	940
265	Intel Core i3-6300	3.8	2	940
266	Intel Xeon W-2123	3.6	4	940
267	Intel Core i3-8100	3.6	4	939
268	Intel Core i3-7100	3.9	2	938
269	Intel Core i5-7600T	2.8	4	938
270	AMD Ryzen 5 5500U	2.1	6	938
271	Intel Core i7-4960HQ	2.6	4	937
272	Intel Core i5-8300H	2.3	4	936
273	Intel Xeon E3-1276 v3	3.6	4	935
274	Intel Xeon E3-1230 v5	3.4	4	934
275	Intel Core i7-4790	3.6	4	933
276	Intel Core i7-10510U	1.8	4	933
277	Intel Core i7-5930K	3.5	6	931
278	Intel Xeon E3-1270 v5	3.6	4	931
279	AMD Ryzen 7 2700	3.2	8	931
280	Intel Xeon E3-1281 v3	3.7	4	931
281	Intel Core i7-8557U	1.7	4	931
282	Intel Xeon E3-1246 v3	3.5	4	930
283	Intel Xeon E3-1245 v5	3.5	4	930
284	Intel Pentium Gold G5400	3.7	2	930
285	Intel Core i7-4940MX	3.1	4	929
286	Intel Core i7-8650U	1.9	4	927
287	AMD Ryzen 5 3400G	3.7	4	925
288	Intel Core i7-5960X	3.0	8	923
289	Intel Xeon Processor E5-1680 v3	3.2	8	923
290	Intel Xeon E5-1650 v4	3.6	6	921
291	Intel Xeon E3-1241 v3	3.5	4	919
292	Intel Core i7-6800K	3.4	6	919
293	Intel Xeon E3-1280 v5	3.7	4	919
294	Intel Core i3-9100T	3.1	4	919
295	Intel Core i7-4790S	3.2	4	918
296	Intel Core i7-4771	3.5	4	917
297	AMD EPYC 7282	2.8	16	917
298	Intel Xeon E3-1225 v5	3.3	4	916
299	Intel Core i7-4770R	3.2	4	912
300	Intel Xeon E5-2696 v3	2.3	18	912
301	AMD Ryzen 5 PRO 1500	3.5	4	911
302	Intel Core i7-4770	3.4	4	910
303	Intel Xeon E3-1270 v3	3.5	4	910
304	Intel Core i5-4690	3.5	4	909
305	Intel Xeon E3-1275L v3	2.7	4	907
306	Intel Pentium Gold G5500	3.8	2	906
307	AMD Ryzen 3 PRO 1300	3.5	4	906
308	Intel Xeon E3-1240 v3	3.4	4	904
309	Intel Core i5-4690S	3.2	4	904
310	Intel Xeon E3-1245 v3	3.4	4	903
311	Intel Xeon Platinum 8124M	3.0	18	903
312	Intel Core i7-5820K	3.3	6	901
313	Intel Core i7-6820HK	2.7	4	901
314	Intel Xeon E5-1630 v4	3.7	4	901
315	AMD Ryzen 3 5300U	2.6	4	900
316	AMD Ryzen 5 1600	3.2	6	898
317	Intel Core i5-6600T	2.7	4	898
318	AMD Ryzen 7 PRO 1700X	3.4	8	898
319	Intel Xeon E5-1650 v3	3.5	6	897
320	AMD Ryzen 5 1500X	3.5	4	896
321	Intel Core i5-6500	3.2	4	895
322	Intel Core i3-6100	3.7	2	894
323	Intel Core i7-6700TE	2.4	4	894
324	Intel Xeon E5-1660 v4	3.2	8	894
325	Intel Core i5-4670	3.4	4	893
326	Intel Core i3-8109U	3.0	2	892
327	Intel Xeon E3-1231 v3	3.4	4	891
328	AMD Ryzen 7 1700X	3.4	8	891
329	Intel Pentium G5600	3.9	2	891
330	AMD Ryzen 3 1300X	3.5	4	890
331	Intel Xeon E3-1220 v6	3.0	4	890
332	Intel Core i3-10110U	2.1	2	889
333	Intel Xeon E5-1680 v2	3.0	8	885
334	Intel Core i7-6700T	2.8	4	885
335	Intel Core i7-4790T	2.7	4	884
336	Intel Core i5-8500T	2.1	6	884
337	Intel Core i5-7400	3.0	4	880
338	Intel Pentium G4620	3.7	2	880
339	Intel Xeon Gold 5218R	2.1	20	880
340	Intel Core i7-4980HQ	2.8	4	879
341	Intel Core i7-5775R	3.3	4	879
342	Intel Xeon E5-1620 v4	3.5	4	879
343	Intel Core i7-6920HQ	2.9	4	878
344	Intel Core i7-7660U	2.5	2	878
345	AMD Ryzen 3 4300U	2.7	4	878
346	Intel Xeon Gold 6126	2.6	12	877

347	Intel Core i3-1000NG4	1.1	2	876
348	Intel Core i5-7360U	2.3	2	875
349	**Intel Xeon E3-1285L v3**	**3.1**	**4**	**875**
350	Intel Core i7-4930MX	3.0	4	874
351	**Intel Xeon E3-1545M v5**	**2.9**	**4**	**874**
352	**Intel Xeon E3-1230 v3**	**3.3**	**4**	**873**
353	**Intel Core i5-7440HQ**	**2.8**	**4**	**873**
354	**Intel Xeon E3-1226 v3**	**3.3**	**4**	**872**
355	**Intel Xeon E5-1630 v3**	**3.7**	**4**	**872**
356	AMD Ryzen 3 3200G	3.6	4	872
357	Intel Core i3-4360	3.7	2	871
358	Intel Core i3-6098P	3.6	2	870
359	Intel Core i5-4590	3.3	4	868
360	Intel Core i7-4770S	3.1	4	866
361	**Intel Xeon Gold 6154**	**3.0**	**18**	**863**
362	Intel Core i7-4870HQ	2.5	4	862
363	**Intel Xeon E3-1575M v5**	**3.0**	**4**	**862**
364	Intel Core i3-4350	3.6	2	861
365	Intel Core i5-4670S	3.1	4	860
366	Intel Core i7-8550U	1.8	4	860
367	Intel Core i5-4590S	3.0	4	859
368	**Intel Xeon E3-1220 v5**	**3.0**	**4**	**857**
369	**Intel Xeon E5-2637 v3**	**3.5**	**4**	**854**
370	Intel Core i5-2550K	3.4	4	853
371	**Intel Xeon E3-1225 v3**	**3.2**	**4**	**853**
372	**Intel Xeon E3-1265L v3**	**2.5**	**4**	**853**
373	Intel Core i7-7700HQ	2.8	4	853
374	**Intel Xeon E3-1505M v5**	**2.8**	**4**	**852**
375	Intel Core i7-4900MQ	2.8	4	851
376	Intel Core i7-6567U	3.3	2	851
377	Intel Core i7-3770K	3.5	4	850
378	**Intel Xeon E3-1535M v5**	**2.9**	**4**	**850**
379	AMD EPYC 7302P	3.0	16	850
380	AMD Ryzen 7 PRO 1700	3.0	8	849
381	Intel Core i5-5675R	3.1	4	849
382	Intel Core i3-4370	3.8	2	848
383	Intel Core i5-7300HQ	2.5	4	848
384	Intel Core i7-7560U	2.4	2	848
385	**Intel Xeon E3-1290 v2**	**3.7**	**4**	**847**
386	**Intel Xeon E5-2697 v4**	**2.3**	**18**	**846**
387	Intel Pentium G4600	3.6	2	844
388	Intel Core i3-7100T	3.4	2	844
389	AMD Ryzen 5 2400G	3.6	4	843
390	Intel Core i5-4570	3.2	4	841
391	Intel Pentium G4500	3.5	2	840
392	AMD Ryzen 7 1700	3.0	8	839
393	Intel Pentium G4520	3.6	2	838
394	Intel Core i5-7500T	2.7	4	837
395	Intel Core i7-4960X	3.6	6	836
396	**Intel Xeon E5-2697 v3**	**2.6**	**14**	**836**
397	AMD Ryzen 3 2200G	3.5	4	836
398	Intel Core i5-3570K	3.4	4	835
399	Intel Core i5-8400T	1.7	6	835
400	Intel Core i7-5700HQ	2.7	4	834
401	Intel Core i5-7267U	3.1	2	834
402	**Intel Xeon Silver 4214R**	**2.4**	**12**	**834**
403	Intel Core i5-5575R	2.8	4	831
404	Intel Core i7-4770T	2.5	4	830
405	Intel Core i5-6400	2.7	4	830
406	**Intel Xeon E5-2697A v4**	**2.6**	**16**	**830**
407	Intel Core i7-6770HQ	2.6	4	828
408	Intel Core i3-4330	3.5	2	827
409	**Intel Xeon Gold 6130**	**2.1**	**16**	**827**
410	Intel Core i7-6820HQ	2.7	4	826
411	**Intel Xeon E5-1620 v3**	**3.5**	**4**	**826**
412	Intel Core i7-4810MQ	2.8	4	824
413	**Intel Xeon E5-2687W v4**	**3.0**	**12**	**823**
414	Intel Core i4-4690T	2.5	4	823
415	**Intel Xeon E5-2618L v4**	**2.2**	**10**	**823**
416	Intel Core i5-6440HQ	2.6	4	822
417	Intel Core i5-7260U	2.2	2	822
418	Intel Core i5-4570S	2.9	4	821
419	**Intel Xeon E3-1220 v3**	**3.1**	**4**	**820**
420	Intel Pentium G4560	3.5	2	819
421	AMD Ryzen 3 1200	3.1	4	819
422	Intel Core i5-8305G	2.8	4	819
423	Intel Core i3-8100T	3.1	4	817
424	**Intel Xeon E3-1270 V2**	**3.5**	**4**	**815**
425	**Intel Xeon E5-2690 v3**	**2.6**	**12**	**815**
426	**Intel Xeon E5-2696 v4**	**2.2**	**22**	**814**
427	Intel Pentium G3258	3.2	2	813
428	AMD Ryzen 7 3750H	2.3	4	811
429	AMD Ryzen 3 PRO 1200	3.1	4	811
430	Intel Core i7-2600K	3.4	4	810
431	AMD Athlon 3000G	3.5	2	810
432	**Intel Xeon E5-2690 v4**	**2.6**	**14**	**809**
433	Intel Core i3-8121U	2.2	2	806

434	**Intel Xeon E5-1607 v4**	**3.1**	**4**	**806**
435	Intel Core i5-6402P	2.8	4	805
436	Intel Core i7-8565U	1.8	4	804
437	Intel Core i7-4860HQ	2.4	4	803
438	**Intel Xeon E5-2699 v4**	**2.2**	**22**	**803**
439	Intel Core i7-3960X	3.3	6	802
440	Intel Core i7-4800MQ	2.7	4	802
441	Intel Core i7-4850HQ	2.3	4	802
442	**Intel Xeon E5-1650 v2**	**3.5**	**6**	**802**
443	Intel Core i5-4460	3.2	4	802
444	**Intel Xeon E5-2667 v4**	**3.2**	**8**	**801**
445	Intel Core i7-3770	3.4	4	799
446	Intel Core i7-3940XM	3.0	4	799
447	Intel Core i7-4820K	3.7	4	799
448	Intel Core i7-4930K	3.4	6	798
449	Intel Xeon E5-1660 v2	3.7	6	797
450	Intel Core i7-6700HQ	2.6	4	797
451	Intel Core i3-4170	3.7	2	797
452	AMD EPYC 7601	2.2	32	797
453	Intel Core i7-2700K	3.5	4	796
454	Intel Core i7-4720HQ	2.6	4	794
455	Intel Core i5-6287U	3.1	2	794
456	Intel Core i5-2500K	3.3	4	793
457	**Intel Xeon E5-2640 v3**	**2.6**	**8**	**793**
458	Intel Core i3-4160	3.6	2	792
459	AMD Ryzen 5 1400	3.2	4	792
460	Intel Core i3-6100T	3.2	2	789

461	AMD EPYC 7742	2.2	64	788
462	Intel Pentium G4400	3.3	2	787
463	**Intel Xeon E5-2678 v3**	**2.5**	**12**	**787**
464	Intel Core i7-3970X	3.5	6	786
465	Intel Xeon E3-1280 V2	3.6	4	786
466	Intel Core i7-4770HQ	2.2	4	782
467	Intel Core i5-4570T	2.9	2	781
468	Intel Xeon E3-1240 v2	3.4	4	780
469	Intel Core i5-6500T	2.5	4	780
470	**Intel Xeon Silver 4214**	**2.2**	**12**	**779**
471	Intel Core i3-4360T	3.2	2	778
472	Intel Core i7-5557U	3.1	2	777
473	Intel Core i3-8145U	2.1	2	776
474	**Intel Xeon E5-1607 v3**	**3.1**	**4**	**775**
475	Intel Core i5-3570	3.4	4	773
476	**Intel Xeon E3-1245 V2**	**3.4**	**4**	**773**
477	Intel Core i5-4440	3.1	4	773
478	**Intel Xeon E3-1505M v6**	**3.0**	**4**	**773**
479	**Intel Xeon E5-2698 v4**	**2.2**	**20**	**773**
480	Intel Core i5-8250U	1.6	4	771
481	Intel Core i7-3930K	3.2	6	770
482	Intel Core i5-6500TE	2.3	4	769
483	AMD EPYC 7702P	2.0	64	769
484	Intel Core i7-3770S	3.1	4	766
485	**Intel Xeon E5-1660**	**3.3**	**6**	**766**
486	AMD Ryzen 5 3550H	2.1	4	766
487	**Intel Xeon E5-2698 v3**	**2.3**	**16**	**766**

488	**Intel Xeon E5-1680 v4**	**3.4**	**8**	**766**
489	Intel Core i3-4340	3.6	2	765
490	Intel Core i3-4150	3.5	2	765
491	Intel Core i7-4712MQ	2.3	4	765
492	Intel Core i7-8665U	1.9	4	765
493	**Intel Xeon E5-2680 v4**	**2.4**	**14**	**764**
494	Intel Core i7-4710MQ	2.5	4	763
495	Intel Core i5-7400T	2.4	4	763
496	Intel Core i7-6650U	2.2	2	762
497	**Intel Xeon E3-1230V2**	**3.3**	**4**	**761**
498	Intel Core i5-4460S	2.9	4	761
499	Intel Core i7-4710HQ	2.5	4	760
457	**Intel Xeon E5-2640 v3**	**2.6**	**8**	**793**
458	Intel Core i3-4160	3.6	2	792
459	AMD Ryzen 5 1400	3.2	4	792
460	Intel Core i3-6100T	3.2	2	789
461	AMD EPYC 7742	2.2	64	788
462	Intel Pentium G4400	3.3	2	787
463	**Intel Xeon E5-2678 v3**	**2.5**	**12**	**787**
464	Intel Core i7-3970X	3.5	6	786
465	**Intel Xeon E3-1280 V2**	**3.6**	**4**	**786**
466	Intel Core i7-4770HQ	2.2	4	782
467	Intel Core i5-4570T	2.9	2	781
468	**Intel Xeon E3-1240 v2**	**3.4**	**4**	**780**
469	Intel Core i5-6500T	2.5	4	780

As you can see in the table here above, there exists only very few correct XEON processors (the XEON CPU's are written with a **red** font). Nearly all the best processors (with a score above 1100) are "Core I7" or "Core I9". So, my advice regarding XEON processors in general is: **Take extreme caution or just avoid them**.

At the position 3 of the above table, we find a cheap CPU (the "Intel Core i7-12700KF" that costs around than $350 on the 2022/1/17). All XEON processors are always much more expensive than that and, furthermore, all of them are much slower than this cheap "Core I7" CPU. One more reason to just avoid Xeon processors.

At the position 119 of the above table, we find a 7 years old CPU (the "Intel Core i7-6700K" that was released in Q3 2015). All the CPU's slower than this very, very old CPU (i.e. all the CPU's located at a ranking number above 119) should be avoided at all costs.

You must realize that the biggest buyers of large "professional" servers are the (web) data centers. This means that, overtime, the offer (from the PC manufacturers) has adapted to the demand (from the web data centers) and nowadays, most large "professional" servers have quite good hard-drives BUT VERY BAD CPU's (i.e. they have low-grade XEON servers such as the E5-2660 v3) because this is what's most commonly required in standard web-data-centers.

1.5. The Best CPU and Motherboard for a TIMi Server (as of May 2019)

1.5.1. The cost is not an issue

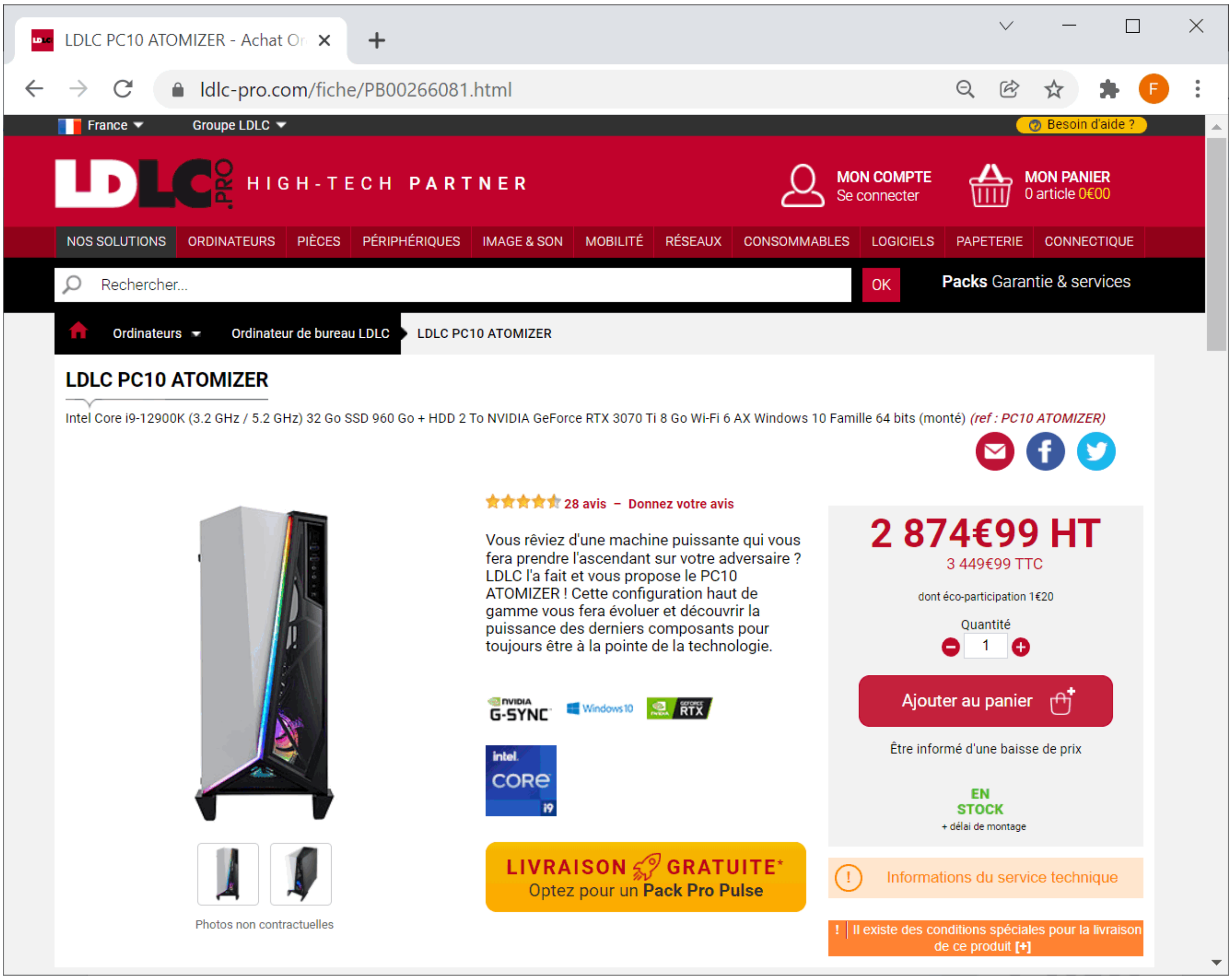
As you can see in the above table, the best CPU on the 2022/1/17 for Anatella/TIMi is the Intel Core i9-12900K at 3200MHz. When selecting a server, you must also pay attention to the motherboard: A good CPU installed on a bad motherboard also gives a poor result (i.e. low speed).

If your IT department allows it, you can buy an assembled, tested & configured server (with a good motherboard) for a good price here (as of 2022/1/17):

https://www.ldlc-pro.com/fiche/PB00266081.html

The components of the " LDLC PC10 ATOMIZER" desktop computer are:

- CPU: Intel Core i9-12900K (3.2 GHz / 5.2 GHz)
- Mother board: MSI MAG Z690 TOMAHAWK WIFI DDR4
- RAM: DDR4 32 Go (2 x 16 Go) 3200 MHz
- System disk: SSD PCI-E NVMe 3.0 4x 960 Go
- Mass Storage Disc: 2 TB HD *(**to replace: see below**)*
- GPU: NVIDIA GeForce RTX 3070 Ti 8 Go
- Network Card : Gigabit LAN
- 8 high-definition audio canals
- Power Supply: Corsair HX850 80PLUS Platinum
- CPU cooler: MSI MAG CORELIQUID C280
- Box: Corsair Carbide SPEC-OMEGA RGB Blanc/Noir
- Microsoft Windows 10 Family 64 bits



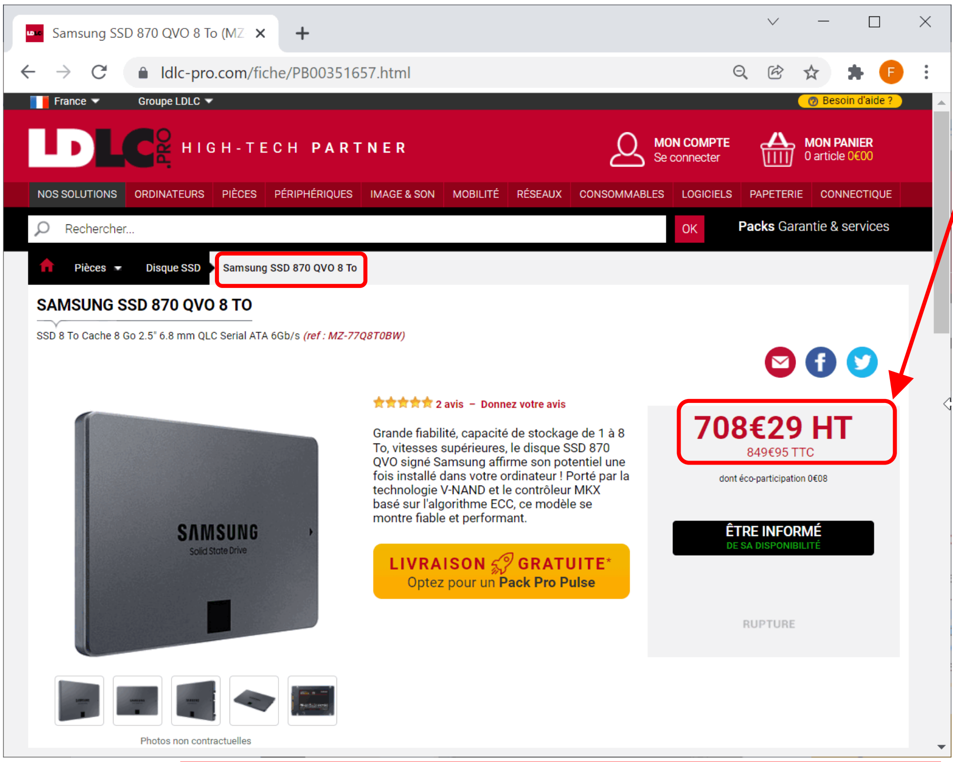
The above PC makes a solid Anatella/TIMi server. It's one of the best server you can buy for Anatella/TIMi (maybe the best one).

If you want the best performances, you should also:

- Upgrade your windows home edition to a professional edition (this is a one-click purchase inside the Windows Store. It costs $99).

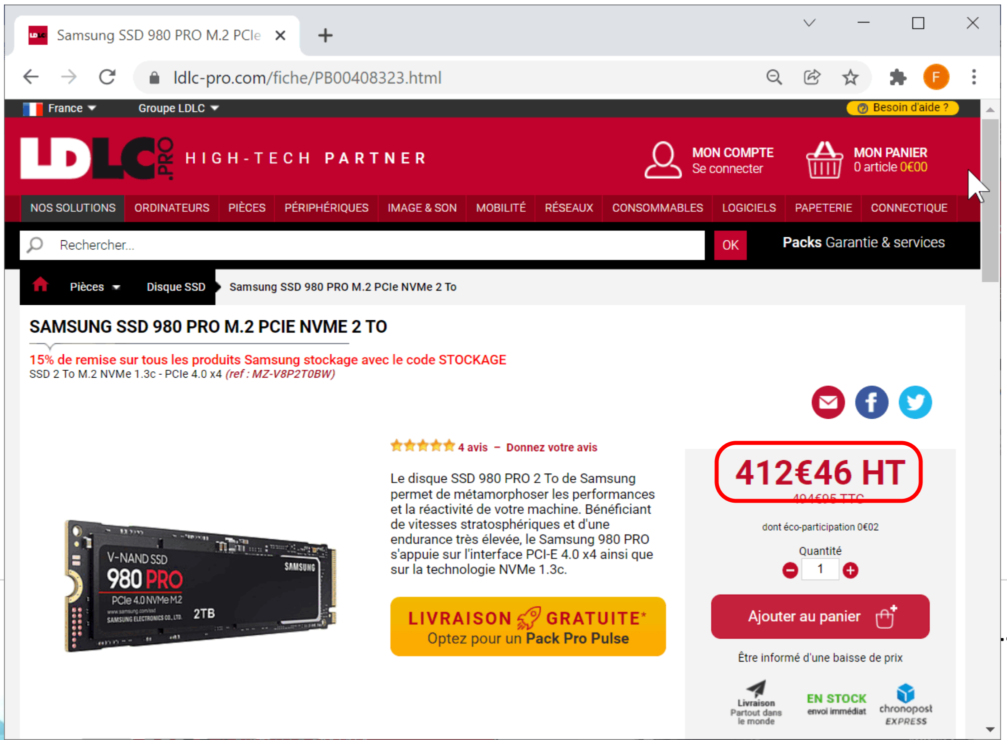
- Replace the HD drive with a Samsung SSD drive (SSD drives from Samsung are a ***lot more reliable*** than classical, old magnetic HD drives, so it makes sense to use a Samsung SSD drive to securely store all your data). I would suggest you to directly buy the 8TB capacity (for 708€): With the motherboard included inside the "LDLC PC10 Atomizer" (suggested here above), you can mount up to 6 of these drives for a whopping total of 6 x 8TB = 48TB of storage.

  https://www.ldlc-pro.com/fiche/PB00351657.html



Read-Speed: 560 MByte/sec ; Write-speed: 530 MByte/sec

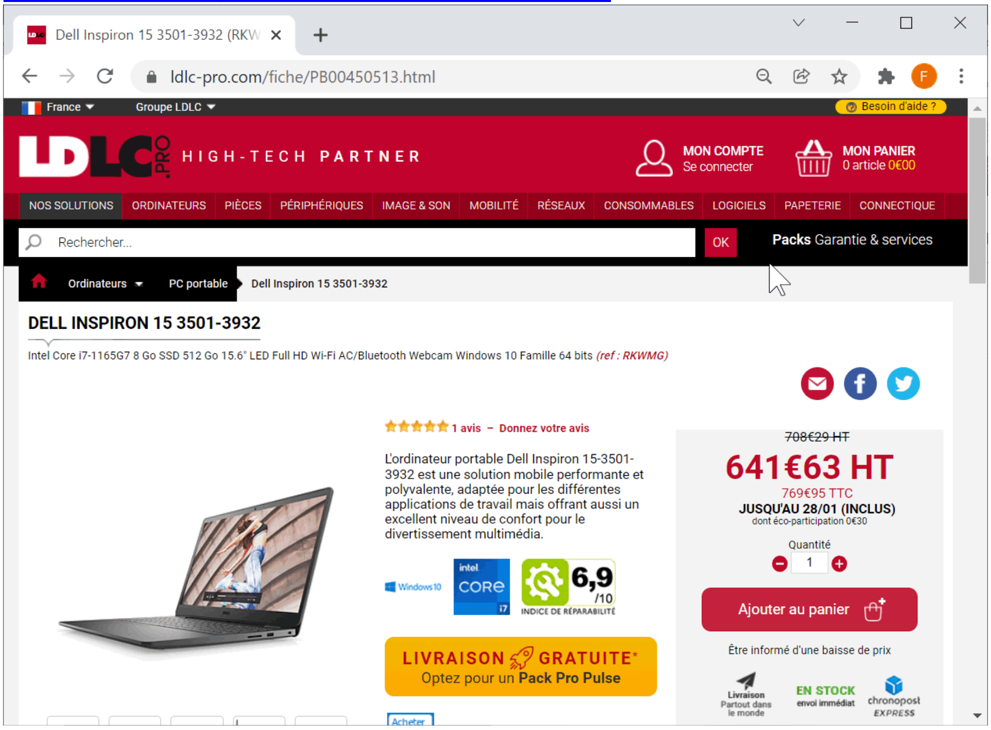
- **Optional:** Replace the single internal 960 GB SSD drive with two "*Samsung SSD 980 PRO M.2 PCIe NVMe 2 TB SSD drives*". The read-speed of each of these SSD drives is 3.5GB/sec. With the motherboard included inside the "LDLC PC10 Atomizer" (suggested here above), you can mount your two SSD drives in "RAID 0" (also named "stripping"), to get a whopping read-speed of 3.5 + 3.5 = 7 GB/sec! You can buy one 2TB "*Samsung SSD 980 PRO M.2 PCIe NVMe*" for 412€ here: https://www.ldlc-pro.com/fiche/PB00408323.html



1.5.2. I want the smallest cost!

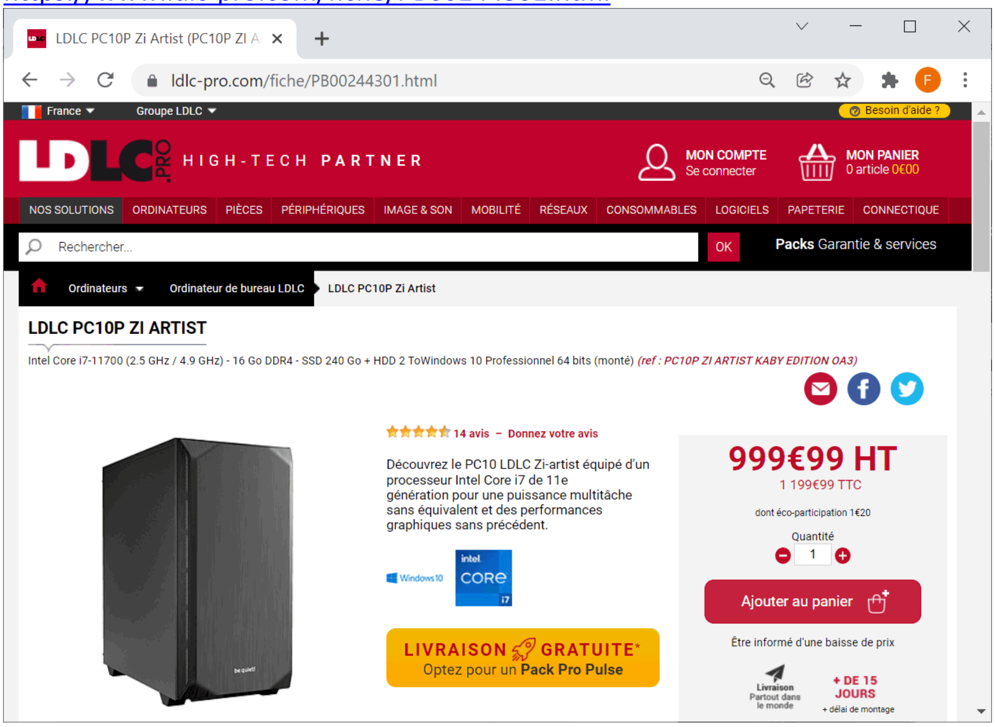
If you want a small laptop, select this one:

https://www.ldlc-pro.com/fiche/PB00450513.html



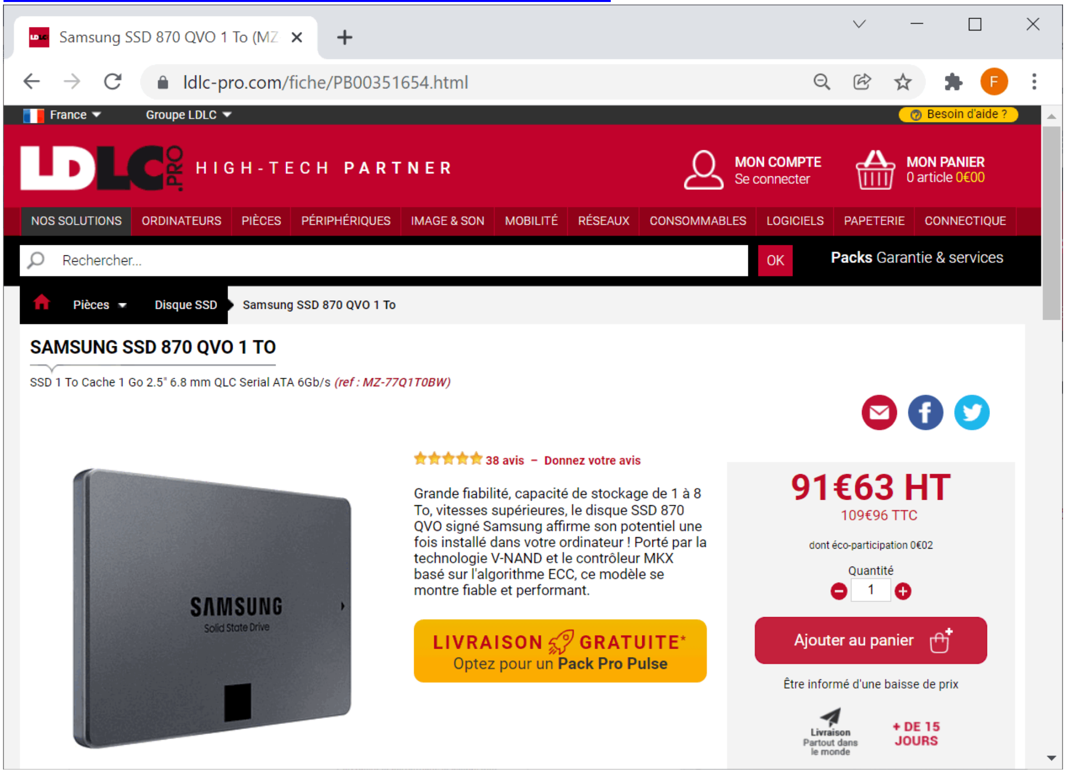
If you want a tower, select this one:

https://www.ldlc-pro.com/fiche/PB00244301.html

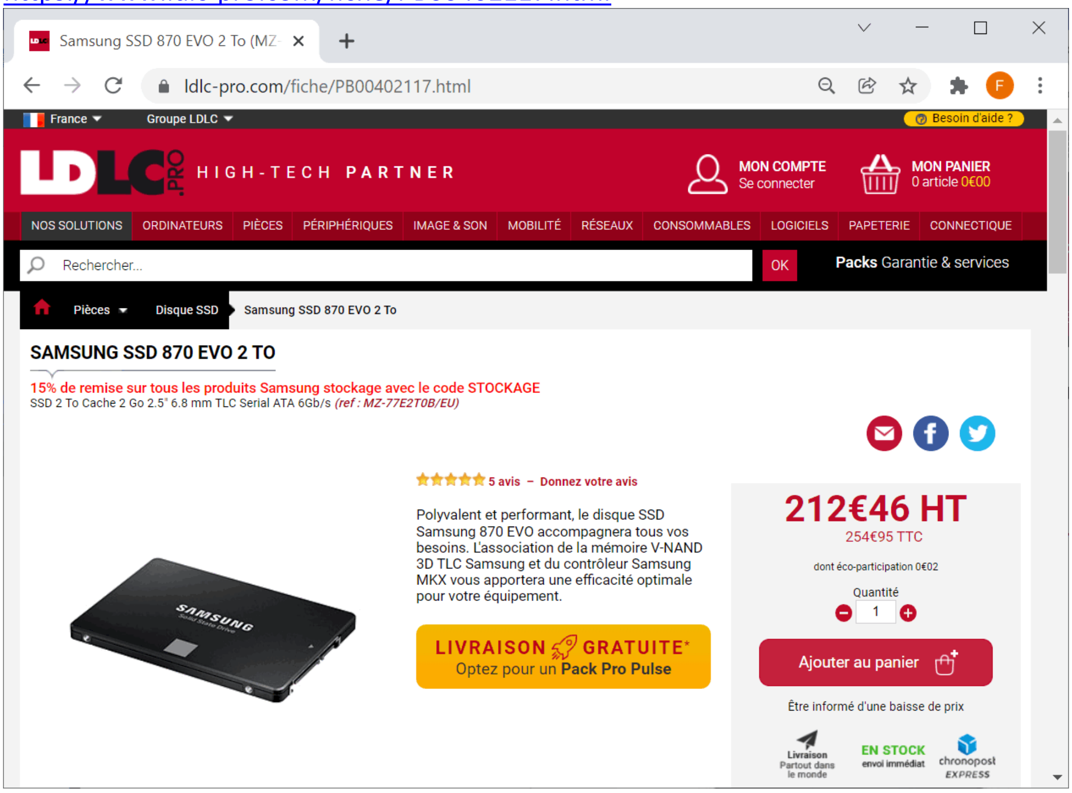


If you buy the tower, replace the HDD with an SSD. Select one of these SSD (1TB or 2TB):

https://www.ldlc-pro.com/fiche/PB00351654.html



https://www.ldlc-pro.com/fiche/PB00402117.html



2. Common infrastructures to run TIMi/Anatella

First a little bit of terminology: Any “Advanced Analytics” architecture/infrastructure must take into account that an advanced analytic project has always two phases:

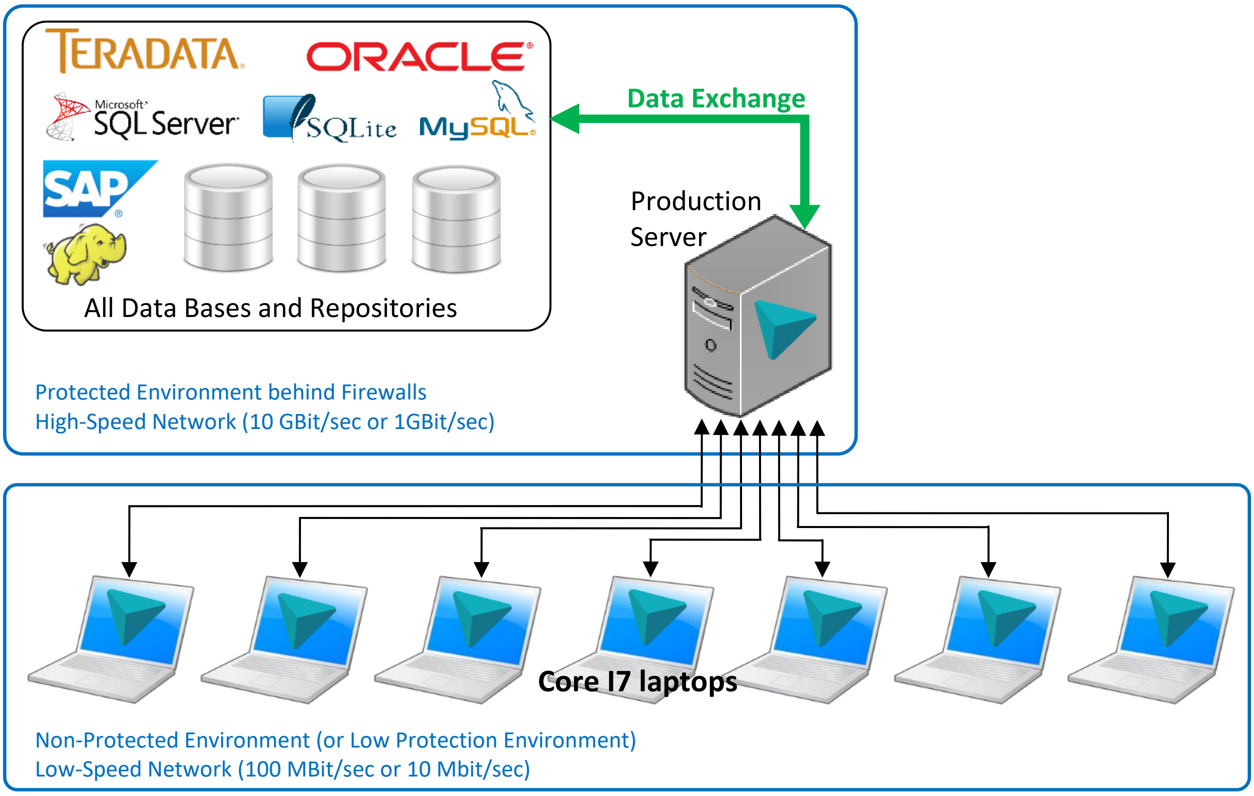
1. **<u>Phase 1: The “Exploration phase”</u>**
   What are the characteristics of the “Exploration Phase”?
   - The Analysts/Data Scientists are developing a new KPI, a new predictive Model or, in general, creating new results through the analysis of data.
   - The Analysts/Data Scientists typically run very heavy data transformations, very heavy computations, searching for the “golden egg”. On a “standard” infrastructure where all the computations are “centralized” on a central database or a central hadoop cluster, these heavy data transformations might disrupt the work of other analyst or, even worse, jeopardize the global stability of the whole IT infrastructure of the company (This is why, in most companies, the Data Scientists are not the “friends” of the IT people).
   - It doesn’t matter so much if one “heavy” computation fails (e.g. because of a bad parameterization).
   - The duration of the “Exploration phase” is, typically, from a few hours to a few weeks.

2. **<u>Phase 2: The “Production phase”</u>**
   The “Production phase” comes after the “Exploration phase” and usually lasts for years. There are usually no really “heavy” computations during the “Production phase”.
   The main concerns of the “Production phase” is the stability: All processes must run smoothly, without ever failing.

We’ll now review 4 different infrastructures, from the cheapest one to the most expansive one. There are no “best” infrastructure: It all depends on your budget and your particular business needs.

2.1. A first (cheap) infrastructure

Let's start with the most common, cheapest infrastructure:



What are the general principles behind such an infrastructure?

- The datasets on which the Analyst are working are centralized on the "Production server" (they are stored inside .gel_anatella files or .cgel_anatella files). Since these file formats are (heavily) compressed, the storage space is practically ***never*** an issue.

- The "Production Server" runs automatically, on a regular basis (typically: every day or week), the different Anatella-data-transformation-graphs and the different timi-models to create all the KPI's, dashboards or scores required for the normal operation of the company.

  The computations on the "Production Server" are always based on "fresh data" originating from the various data centers, data warehouse and, or any other data sources.

- During the "exploration phase", all the Analysts & all the Data Scientists are working on their own (Core I7) laptops. The above picture illustrates the situation when there are 7 Data Scientists inside the company (since there are 7 laptops inside the picture). Since each laptop has roughly three times the computing power of an Oracle Exadata machine, each analyst has, basically, enough computing power to compute any data transformations. The only limiting factor is the access & the storage of the datasets on which the Analyst are working.

- During the "exploration phase", an Analyst/Data Scientist needs some data to work with (obviously). This data is typically originally stored on the "Production Server" (although, it can change: see some alternative solutions below).

Typically (for performance reasons), the Analyst/Data Scientist will make a copy of the required data on its laptop and work on the copy to find new results. If the "exploration phase" lasts for a long period of time, the Analyst might need to refresh its local copy of the datasets, but 99% of the time, the Analyst can work with slightly outdated data to produce the required analysis results.

In general, as an analyst, you should avoid using data stored on a distant machine or accessing data through a slow network interface (That's ok if the network interface that is connected to your laptop is a 10GBit/sec network interface but I guess that it won't be the case).

For efficiency reasons, 99% of the time, it's better to copy locally **one time** your data on an encrypted partition on your local PC and then work with the local copy.

To encrypt a partition, you can use the free "bitlocker" application included inside MS-Windows or the famous & free TrueCrypt application.

Once the analysis is complete (i.e. once the new KPI looks good, once the new predictive models are ok, once the new reports are ok, etc.), the Analysts "moves" all his graphs & models to the "Production Server". In technical terms, we'll say that the "exploration phase" is finished and the "production phase" starts. Once the graphs & the models are on the "Production Server", they will be applied on the "fresh" data, to always get "fresh" results (i.e. the "production phase" is always on "fresh" data).

Moving a data-transformation process from one computer to another (i.e. from the Analyst's laptop to the "production server") is usually an error-prone procedure (e.g. it's usually a real nightmare with SAS). Contrary to all the other solutions, Anatella possesses some unique functionalities (e.g. relative path to the datasets, self-contained .anatella files, etc.) that allow you to migrate effortlessly all your work (graphs & models) from one machine to another.

Here are the Pro & Con of the above infrastructure:

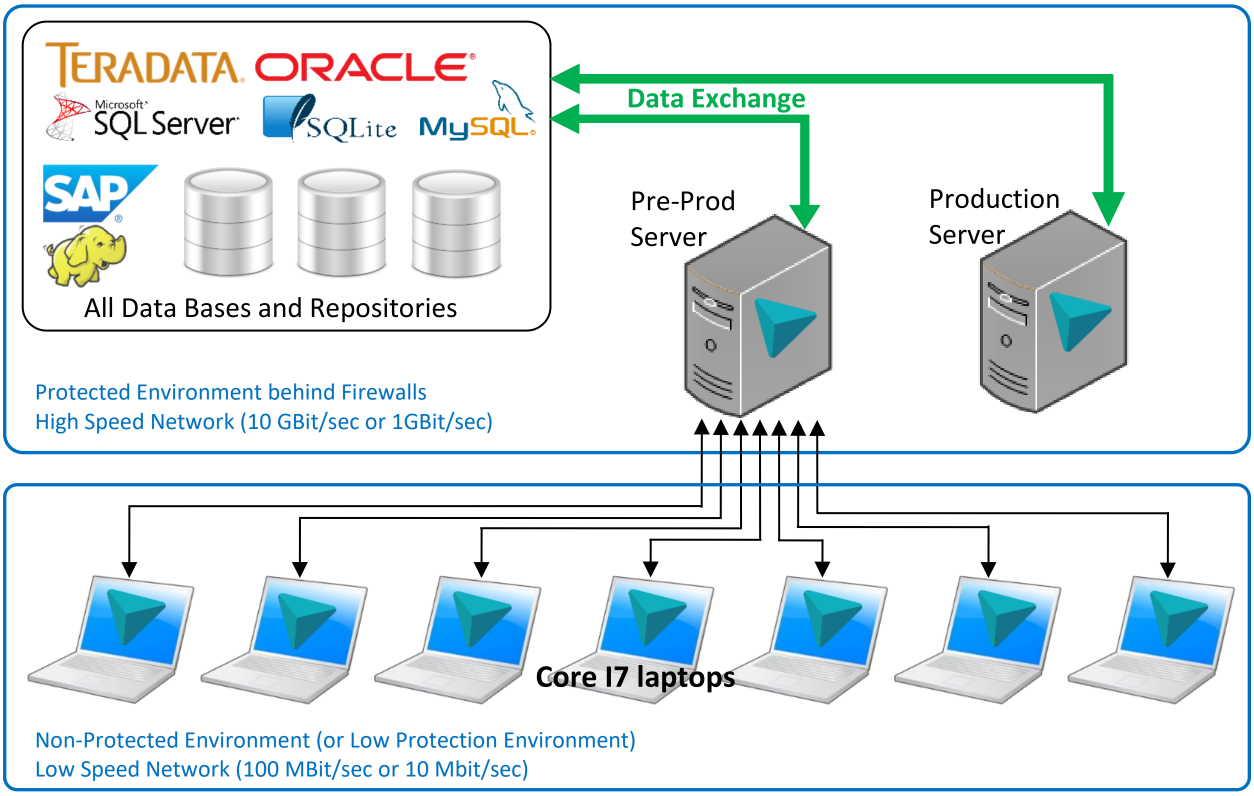
- Pro:
  - **Cheap:** No hardware investment required outside the purchase of a "Production Server"
  - **Scalable:** If you have more analysts, simply add more laptops.
  - **Secure:** Why? Because:
    - The Analysts/Data Scientists cannot delete any critical data from your operational systems because all they are allowed to do is to copy, from time-to-time, some .gel_anatella files or some .cgel_anatella files on their local hard drive. They can't even damage the .gel_anatella files (or .cgel_anatella files) that other analysts might require because the .gel_anatella files (or .cgel_anatella files) are read-only files.
    - During the "exploration phase", the Analysts/Data Scientists typically run very heavy data transformations, searching for the "golden egg" in your data. On a standard, centralized infrastructure, these *heavy* data transformations might disrupt the work of other analysts or, even worse, jeopardize the global stability of the whole IT infrastructure of the company.
      This disastrous situation will never happen with the proposed solution here above: Indeed, each of the Data Scientists is using its own CPU without consuming any resource from the production servers (or from other Analysts).

- Con:
    - Once an analysis is complete (i.e. once the "exploration phase" is complete) and a new process is developed, it's directly "published" to the "Production Server" without any testing-period first. This means that a bad, untested, process could consume so much processing power that it could "bring to its knees" the production server (which is a bad thing). We should add a pre-production server to avoid such situation, to improve reliability.
    - All the datasets are stored on the "production server" (i.e. the .gel_anatella and the .cgel_anatella files). If many Analysts/Data Scientists decide to simultaneously copy some large datasets on their laptop, the "production server" might slow-down briefly (due to the many simultaneous "copy operations").
    - During the "exploration phase", some datasets are copied on the laptops from the Analysts/Data Scientists for a brief period of time (i.e. for the time required for them to produce new results: i.e. to produce new graphs and new models). This might be a concern for very sensitive data (especially for banks, insurance, etc.). To alleviate this problem, the data is usually stored on the laptops on an encrypted partition (the partition is encrypted with bitlocker or truecrypt) but this might not be a solution that is "secure enough".

In the next sections, we'll review each of these "con" arguments and give different solutions to each of them.

2.2. A second infrastructure

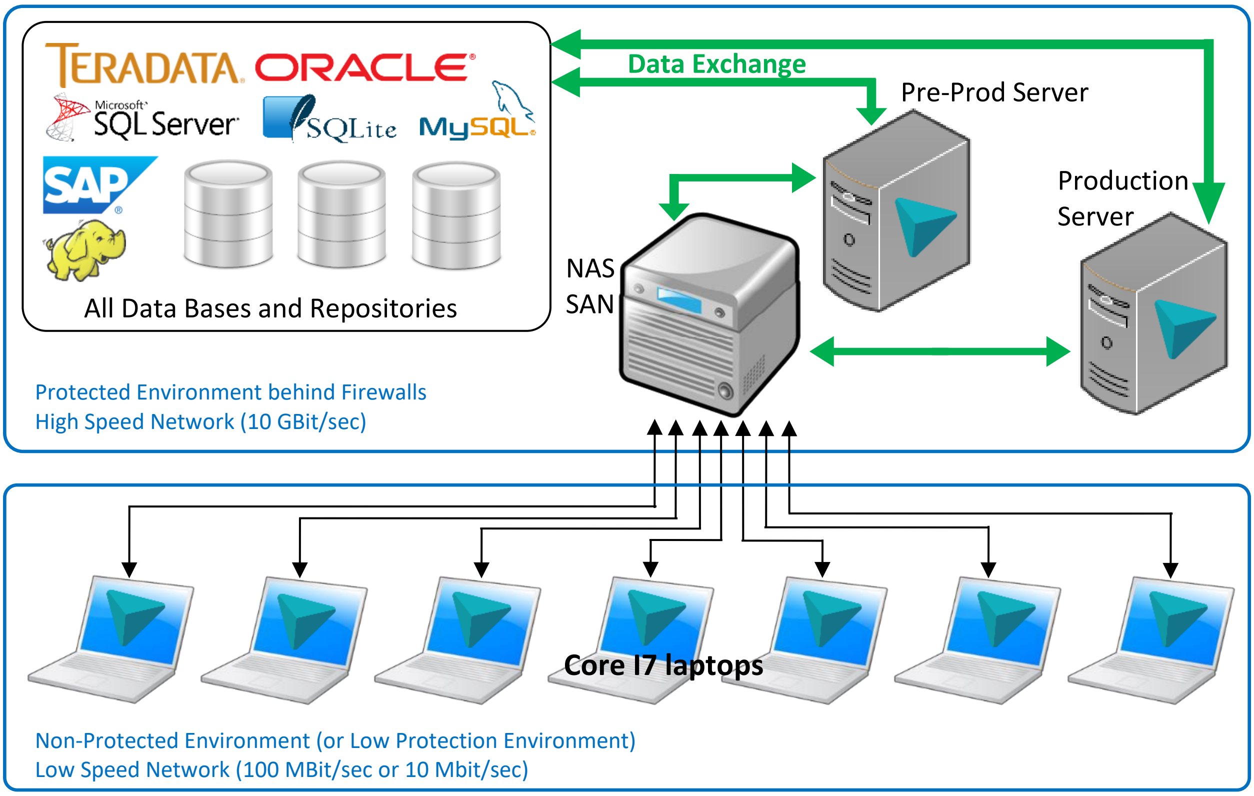
To increase reliability, use a “pre-production” server:



Once an analysis is complete (i.e. once the “exploration phase” is complete) and a new process is developed, it’s “published” to the “Pre-Production Server” for a few weeks. If the “Pre-Production” server remains stable & responsive, you can thereafter “publish” your new graphs&models from the “Pre-Production” environment to the final “Production Server” environment.

2.3. A Third infrastructure

To increase throughput when many laptops are accessing the datasets (stored in the .gel_anatella file and the .cgel_anatella files), use a NAS (Network Attached Storage):



Please note that, since all your datasets are now on a NAS, you need some fast network interface between the "Production Server", the "Pre-Prod server" and the "NAS". Ideally, you should use a 10 GBit/sec network infrastructure. If you are using a more standard (i.e. less expensive) infrastructure based on standard 1GBit/sec network cards, you might occasionally experience some slower disk access.

Please refer to the following table to know more about this subject:

Recommended solutions (depending on your budget)

Physical Location of the data	Maximum I/O Speed for processes running inside the "Production Server"
One new-generation NVMe SSD drive inside the "Production Server" (this is a bargain price!)	3500 MByte/sec (but NVMe SSD drives are relatively small: less than 4GB capacity)
RAID6-Drive inside the "Production Server" (or a SAN inside the "Production Server")	2000 MByte/sec (e.g. using 4 standard SSD drives)
NAS within a 10GBit/sec network	1000 MByte/sec on a "BIG" NAS 500 MByte/sec on a "small" NAS
One standard SSD drive inside the "Production Server"	550 MByte/sec (up to 48 GB of storage)
NAS within a 1GBit/sec network	100 MByte/sec This might increase up to 800 MByte/sec if you use "port trunking" but this is uncommon. For more information, see here: https://en.wikipedia.org/wiki/Link_aggregation
One Magnetic drive inside the "Production Server"	80 MByte/sec
HDFS drive (we strongly advise to avoid using HDFS)	From 5 to 50 Mbyte/sec

Most of the time, a server equipped with a standard SSD drive, already delivers optimal I/O performances (see section 1.5 for some advices about SSD drives).

2.3.1. A small Note about I/O speed

The objective of this section is to explain why the I/O's are not usually a bottleneck when manipulating data with Anatella (i.e. the CPU is usually the bottleneck and not the I/O's).

Anatella works in "streaming" (in opposition to R or python that works, by default, in "batch"). This means that, inside Anatella, there exists a data flow that is going "through" all the operations inside the "Anatella data transformation graph" (abbreviated to "graph"). For example, if you want to join two tables, you must (of course) read the data from both tables (you can read each table at 80 MB/sec "compressed data" or 800 MB/sec "uncompressed data") and, at the same time, compute the join, line-by-line (in streaming). However, the calculation of the join in itself is very expensive: It can only be done at 100 MB/sec on average (on a standard telecom table). So, there's a "bottleneck" at 100 MB/sec: i.e. it's useless to extract/read "data lines" out of the hard drive at a speed of 800 MB/sec if, right after the lecture, the data flow can only be processed at a maximum speed of 100 MB/sec.

> In the above example, the execution time is 100% governed by the speed of the join (and not at all by the speed of the hard drive or the speed of the I/O accesses).
>
> It's a little in opposition to codes in R/Python, where the execution times of the different components add up: With Anatella, the execution time of a graph is (usually) proportional to the slowest element of the graph.
>
> This is why it is possible to tell Anatella to allocate a larger number of CPU's to a particular operation (i.e. to a particular "box"), to avoid/reduce this "bottleneck" effect (for more information about this subject, see section 5.3.2 of the "AnatellaQuickGuide.pdf"). For example, Anatella could be told to use 7 CPU's to calculate the join (instead of using one CPU by default), to get a throughput of 7x100MB/sec = 700 MB/sec at the end (and, therefore, removing the "Bottleneck" of the join).

In practice, one quickly realizes that the hard disk (or the I/O speed) is practically NEVER the "bottleneck element" that decides of the overall execution time of an Anatella-data-transformation-graph. That's why we are now putting most of our development efforts in improving the speed of all other components (join, sort, filter, scoring) inside Anatella. For example, I think no one can beat the "sort routines" included in Anatella.

Currently, 99% of developers that are working in the R, Python, Hadoop ecosystems, etc. did not yet arrived to the same conclusion as us, and thus they are still (and quite stupidly) trying to get better I/O's. These developers did not manage to get the same conclusions as us because:

- They have a different architecture (that is not based on "data streaming", as in Anatella).
- They are using the Java language (that has such terrible I/O performances that it's always blocking everything).
- They have different "workloads": Anatella is built for analytics and predictive analytics tasks in mind. Such type of workloads typically requires complex, CPU-intensive computations (to create refined KPI's or to do "feature engineering") that dominates the computation time: i.e. these CPU-intensive operations usually represent 95% of the computation time (i.e. inside Anatella, the CPU is usually the bottleneck). So, if we can read the data faster, we will (maybe!) just gain a few percent out of the 5% of time that Anatella devotes to reading the data.

On the other hand, it's true that, for a very simple "Anatella-data-transformation-graphs" (e.g. for example, a graph that contains only one aggregate to compute), it's worth reading the data faster.

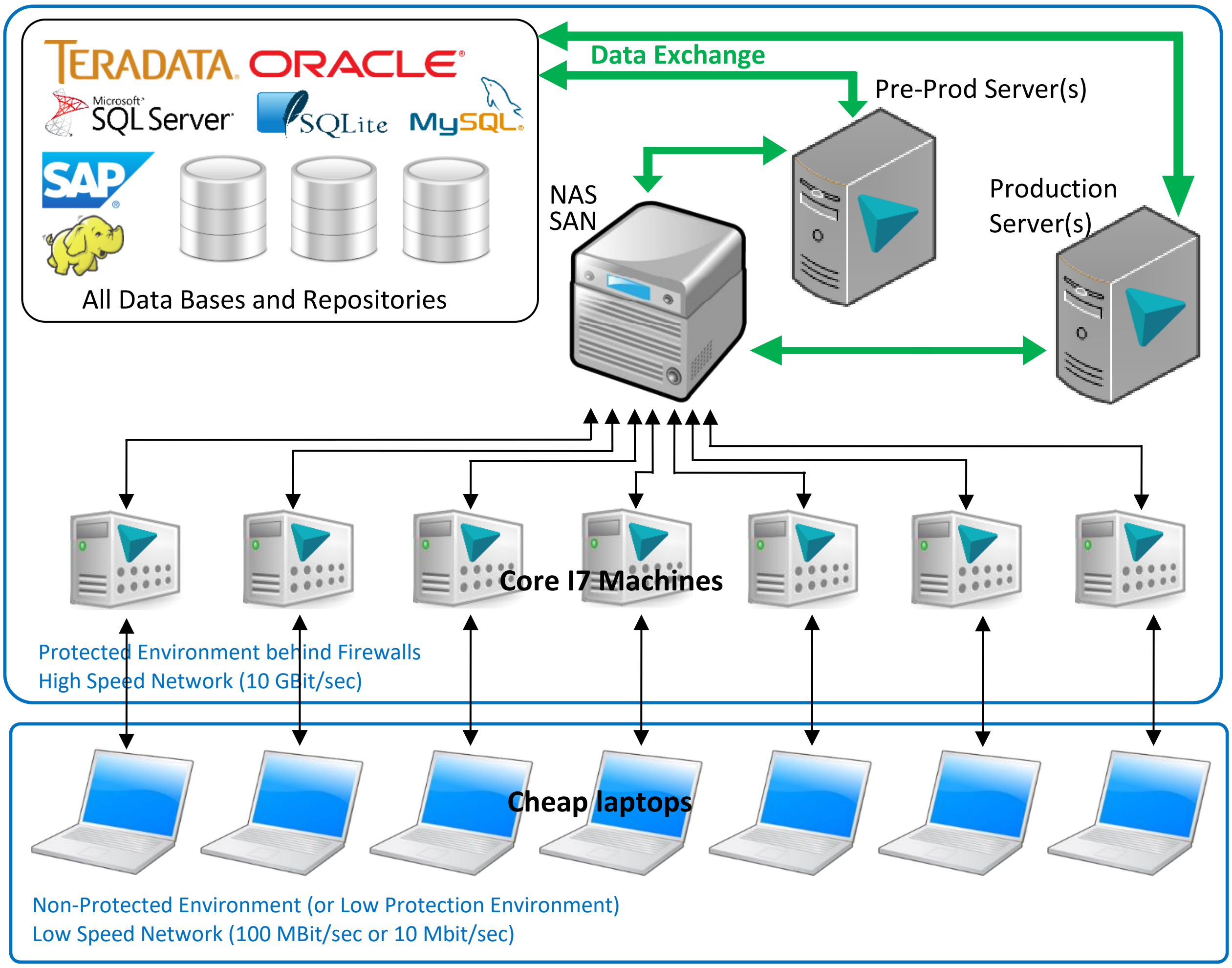
Anatella can also run little “boxes” coded in R or Python. What’s happening for such little box? It's true that the “normal” R and Python engines do not work "in streaming" (such as Anatella) but we managed to "transform & upgrade" these engines so that they can also work (most of the time) in streaming like the rest of the “boxes” inside Anatella. So, no worries! 😉

An amusing side-effect of this “upgrade” is the possibility to easily run in "parallel" (i.e. on several CPU’s) R and Python codes (i.e. there is an almost automatic parallelization of the R/Python code).

A good example is the “R_ApplyModel” box in Anatella that runs 100% in streaming mode and on multiple CPU’s (inside a N-Way multithreaded section: See the section 4.8.3.2. of the “AnatellaQuickGuide.pdf” for more information about this subject).

2.4. A fourth infrastructure

If you have very sensitive data, it might be better that your data always stays only inside your "*Protected Environment behind Firewalls*" and you'll have something like:



Each Analyst/Data Scientist is accessing its own "Core I7" machine using the standard Remote Desktop Protocol (The analysts only use their laptop as a simple terminal, so the laptops can be "cheap"). This architecture is slightly more expensive because you need to buy two machines (a good "Core I7" machine and a "cheap" laptop) for each new Data Scientist (instead of only one previously). The big advantage is, obviously, that your confidential datasets won't leave your "*Protected Environment behind Firewalls*" and you have a very secure solution.

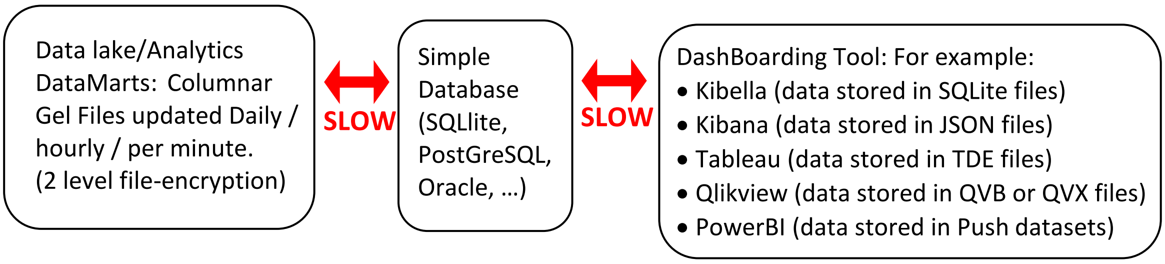
3. Integration with third party tools

3.1. Integration with "simple" BI tools

All the proposed infrastructures allow an easy&efficient integration with "simple" BI/Reporting/Dashboarding tools.

Typically, these BI tools are used to display some reports or dashboards inside a browser. There are many different techniques to give to the BI tools the datasets required to compute the reports and the dashboards.

One first solution is the following:



The advantages of the above solutions are:

- It works with all Dashboarding tools because (nearly) all the Dashboarding tools can get their data from "common" SQL-based data sources (one exception is Kibana that requires to get its data from JSON files).

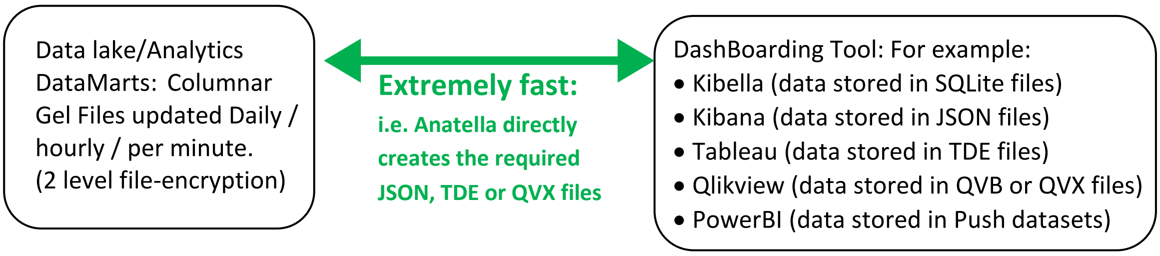
The dis-advantages of the above solutions are:

- The display-speed is "sluggish". More precisely: Qlickview, Tableau and Kibana are much more efficient when the data to display is stored inside their own proprietary format (i.e. when it's stored inside TDE files for Tableau or inside QVB/QVX files for Qlickview, or inside JSON files for Kibana). The time required to update a webpage containing an interactive dashboard is a lot longer when the dataset is stored in a remote database. At the end, the end-user experience when using a Dashboarding that runs many extraction out of a remote database is the following: it "feels sluggish": i.e. the user is enduring (intolerable) delays in all webpage-refresh (this is why Qlickview promotes so much "in-memory" analytics). In particular, you should really avoid "slowly reacting" databases (such as Hive or PostGreSQL).
- Some advanced functionalities are only available when the data is properly saved in the proprietary format of the tool (see the qlickview documentation: e.g. in Qlickview, some aggregates are only available when the data source is a QVB/QVX file ; The same exists in Tableau: i.e. Some aggregates are only available when the data source is a TDE file).
- It's inefficient in terms of hard-drive-space-consumption because there exists several copies of the datasets required for display. The first copy is stored into the Dashboarding tool in itself (i.e. it's stored inside TDE files for Tableau, or inside QVB/QVX files for Qlickview, or inside JSON files for Kibana). …and the second copy of the data is stored inside the "Simple Database" (located in the central position of the above chart).

- It's inefficient in terms of refresh-speed, when refreshing/updating the data source:
  - i.e. copying data from the data lake and into the database "in-the middle" (these are "insert"-type of operation) is ***extremely* slow** ("insert" operations in databases are slow). You can (partially) avoid this very slow "copy operation" by using a special database type: a SQLite database. One unique particularity of the SQLite databases is the extremely high speed of the "bulk-insert-operation" (i.e. writing data inside a SQLite database is nearly as fast as writing the same data inside a simple flat text file). Any other database will lead to a very slow running time.
  - i.e. copying data from the database "in-the middle" into the Dashboarding tool (typically, using an ODBC connection). We are actually talking about executing "*database-extraction procedures*". Such kind of procedures are **always slow**. The Tableau and the Qlikview documentation both agree that, getting data from a file (i.e. from a TDE file for Tableau or from a QVX file for Qlikview) is the most efficient way

of getting data-access: The Tableau & Qlickview documentation states that, getting data through a TDE file (for Tableau) or a QVX file (for Qlickview) is usually around 100 times faster, compared to a database-extraction.

To solve all the above problems, we propose the following:

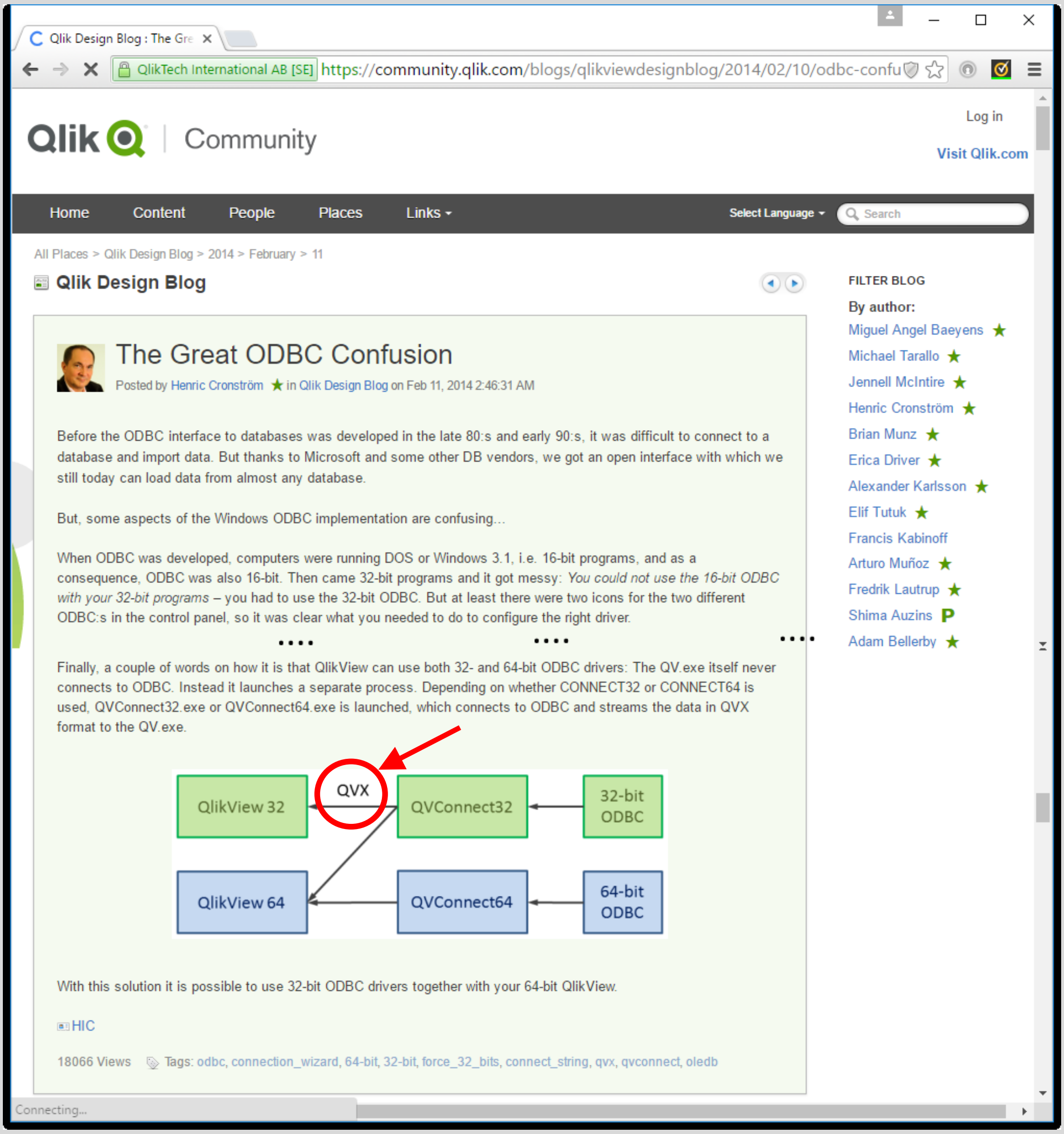


Anatella creates the required dataset files directly inside the proprietary format of the Dashboarding tools. i.e. Anatella creates QVX files for Qlikview, Anatella creates TDE files for Tableau, Anatella creates JSON files for Kibana. The "bottleneck-in-the-middle" (i.e. the database) has disappeared: it has been replaced by a very fast Anatella procedure that generates (at a high speed, since it's Anatella that is running !) all the required files.

If you intend to use still another Dashboarding tool (i.e. not Tableau, Qlikview or Kibana), you can still get a decent speed by using a SQLite database as the "database-in-the-middle". Please refer to the section 4.8.2.4. and 4.8.17.6 of the "AnatellaQuickGuide.pdf" for more information about SQLite databases (e.g. why they are so great for interacting with such kind of Reporting/Dashboarding tools)

If you're using Qlikview as a BI tools, there's a great article on the subject of QVX files inside Qlickview here:
https://community.qlik.com/blogs/qlikviewdesignblog/2014/02/10/odbc-confusion

Here is an excerpt out of this webpage:


Qlik Community

Home | Content | People | Places | Links

All Places > Qlik Design Blog > 2014 > February > 11

Qlik Design Blog

The Great ODBC Confusion

Posted by Henric Cronström in Qlik Design Blog on Feb 11, 2014 2:46:31 AM

Before the ODBC interface to databases was developed in the late 80:s and early 90:s, it was difficult to connect to a database and import data. But thanks to Microsoft and some other DB vendors, we got an open interface with which we still today can load data from almost any database.

But, some aspects of the Windows ODBC implementation are confusing...

When ODBC was developed, computers were running DOS or Windows 3.1, i.e. 16-bit programs, and as a consequence, ODBC was also 16-bit. Then came 32-bit programs and it got messy: *You could not use the 16-bit ODBC with your 32-bit programs* – you had to use the 32-bit ODBC. But at least there were two icons for the two different ODBC:s in the control panel, so it was clear what you needed to do to configure the right driver.

....

Finally, a couple of words on how it is that QlikView can use both 32- and 64-bit ODBC drivers: The QV.exe itself never connects to ODBC. Instead it launches a separate process. Depending on whether CONNECT32 or CONNECT64 is used, QVConnect32.exe or QVConnect64.exe is launched, which connects to ODBC and streams the data in QVX format to the QV.exe.

With this solution it is possible to use 32-bit ODBC drivers together with your 64-bit QlikView.

HIC

18066 Views Tags: odbc, connection_wizard, 64-bit, 32-bit, force_32_bits, connect_string, qvx, qvconnect, oledb

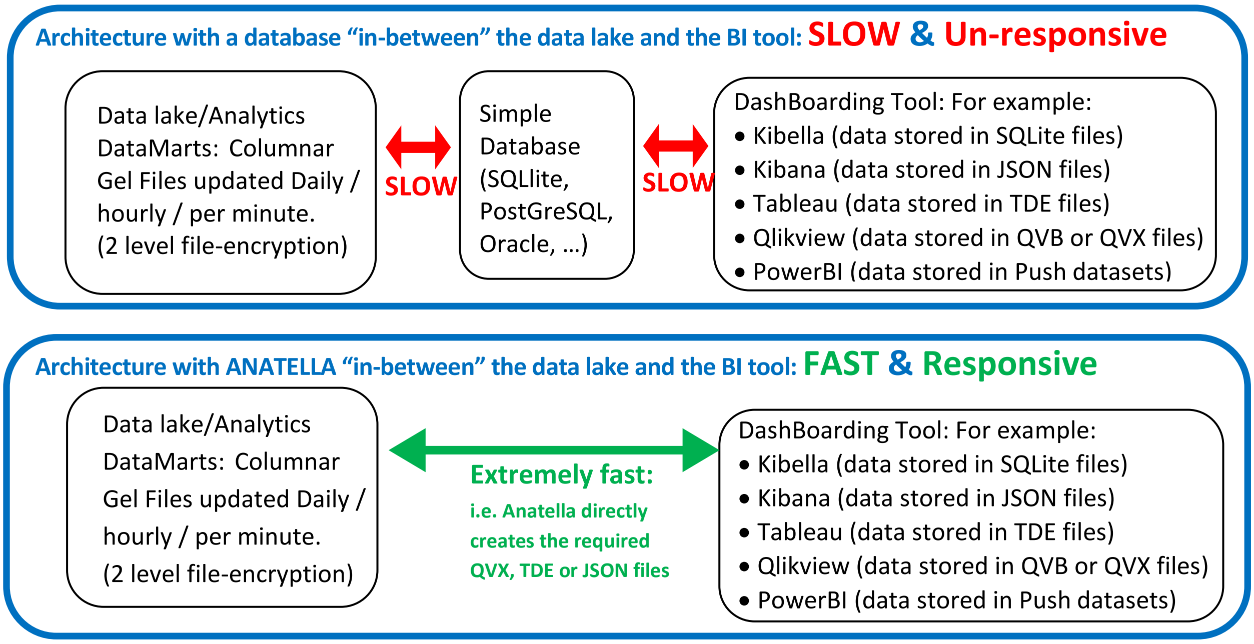
FILTER BLOG

By author:

Miguel Angel Baeyens
Michael Tarallo
Jennell McIntire
Henric Cronström
Brian Munz
Erica Driver
Alexander Karlsson
Elif Tutuk
Francis Kabinoff
Arturo Muñoz
Fredrik Lautrup
Shima Auzins
Adam Bellerby

In particular, you can read in the above article: "*QlikView itself never connects to ODBC* (i.e. the QlikView process never connects directly to the SQL database), *instead it launches a separate process ... that connects to ODBC and streams the data in QVX format to the QlikView process*". We now understand better the great interest of creating directly, and at very high speed, the famous QVX files rather than using ODBC to make a slow and unreliable data extraction out of your database.

This chart summarizes the interaction between a "data lake" and a "BI tool" inside two different infrastructures: i.e. inside an architecture based on a SQL database "in-the-middle" and inside the proposed optimal architecture based on Anatella:



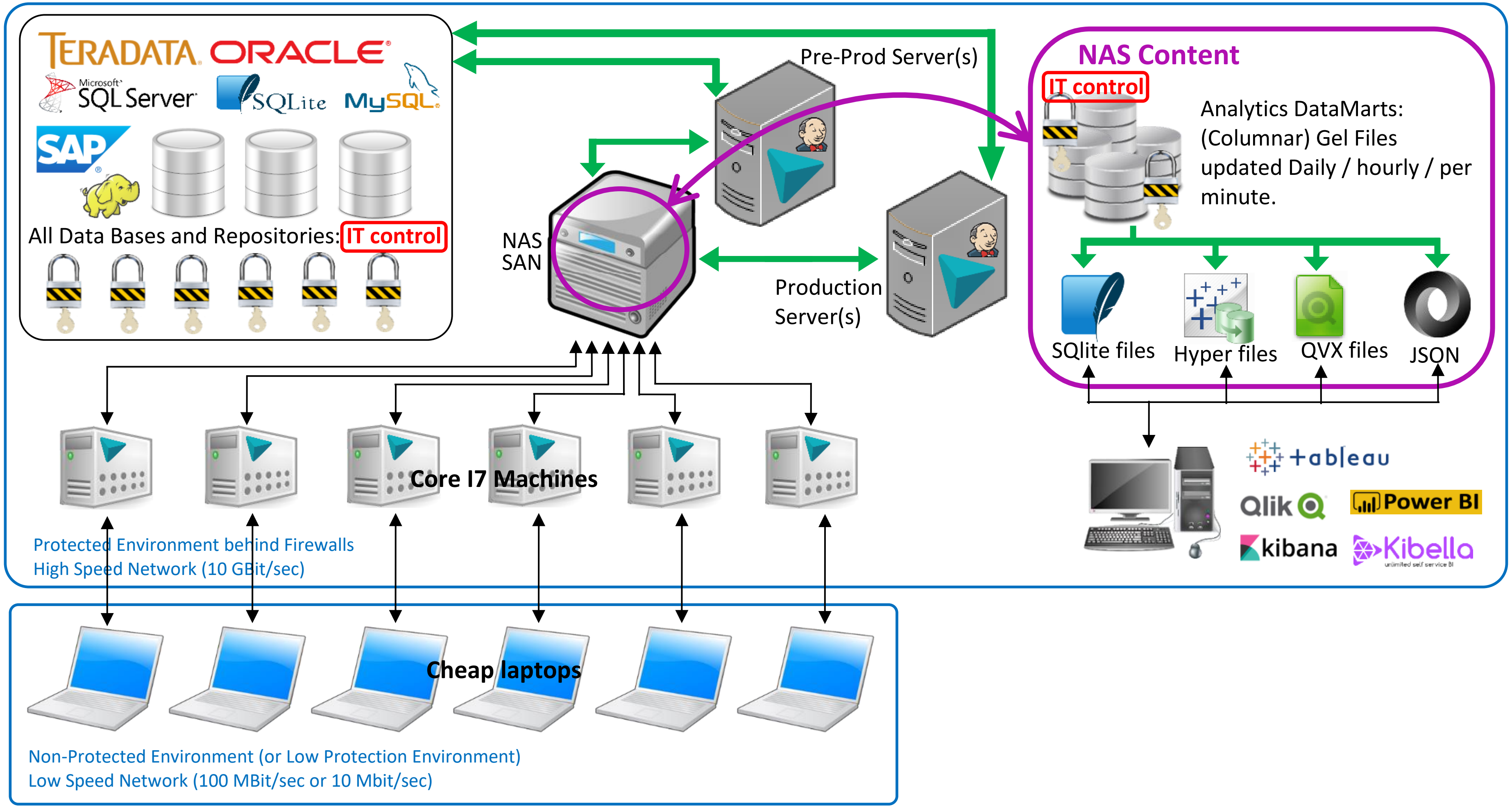
3.2. Scheduler: Jenkins

At one point, you might have so many jobs (so many data-transformations and scoring) running on your "Production Server" every night that you might need to add a second (or even a third) "Production Server" to still be able to compute everything during the short time-span of the night. To manage all the jobs running on the several different "Production Servers", one easy solution is to use "Jenkins": See the section 4.8.7.2. of the "AnatellaQuickGuide.pdf" to have more information about the integration between "Jenkins" and Anatella. Here is an extract of this section:

> *Jenkins can transparently manage a fleet of many computers (i.e. it manages many "nodes" in technical terms). When Jenkins needs to run a job, Jenkins can easily connect to an "idle" node and run the required job there (in technical term, this is called "distributed computation"). This gives to the final user/company a tremendous computing power: There are actually no limits to the delivered computing power: if you need more computing, simply add some more "nodes".*

4. Summary on the optimal infrastructure

You will find on the next page a chart that summarizes the proposed architecture.


TERADATA
ORACLE
Microsoft SQL Server
SQLite
MySQL
SAP
All Data Bases and Repositories:
IT control
NAS
SAN
Pre-Prod Server(s)
Production
Server(s)
NAS Content
IT control
Analytics DataMarts:
(Columnar) Gel Files
updated Daily / hourly / per
minute.
SQlite files
Hyper files
QVX files
JSON
Core I7 Machines
tableau
Qlik
Power BI
kibana
Kibella
Protected Environment behind Firewalls
High Speed Network (10 GBit/sec)
Cheap laptops
Non-Protected Environment (or Low Protection Environment)
Low Speed Network (100 MBit/sec or 10 Mbit/sec)