



CREATIVITY THROUGH EFFICIENCY

An automated predictive datamining tool

Server/Infrastructure Selection for TIMi v1.15

Creation Date: September 2012

Last Edited: March 2025

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 S3 storage, 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 $3 \times 2\text{GHz} > 4\text{GHz}$). 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.

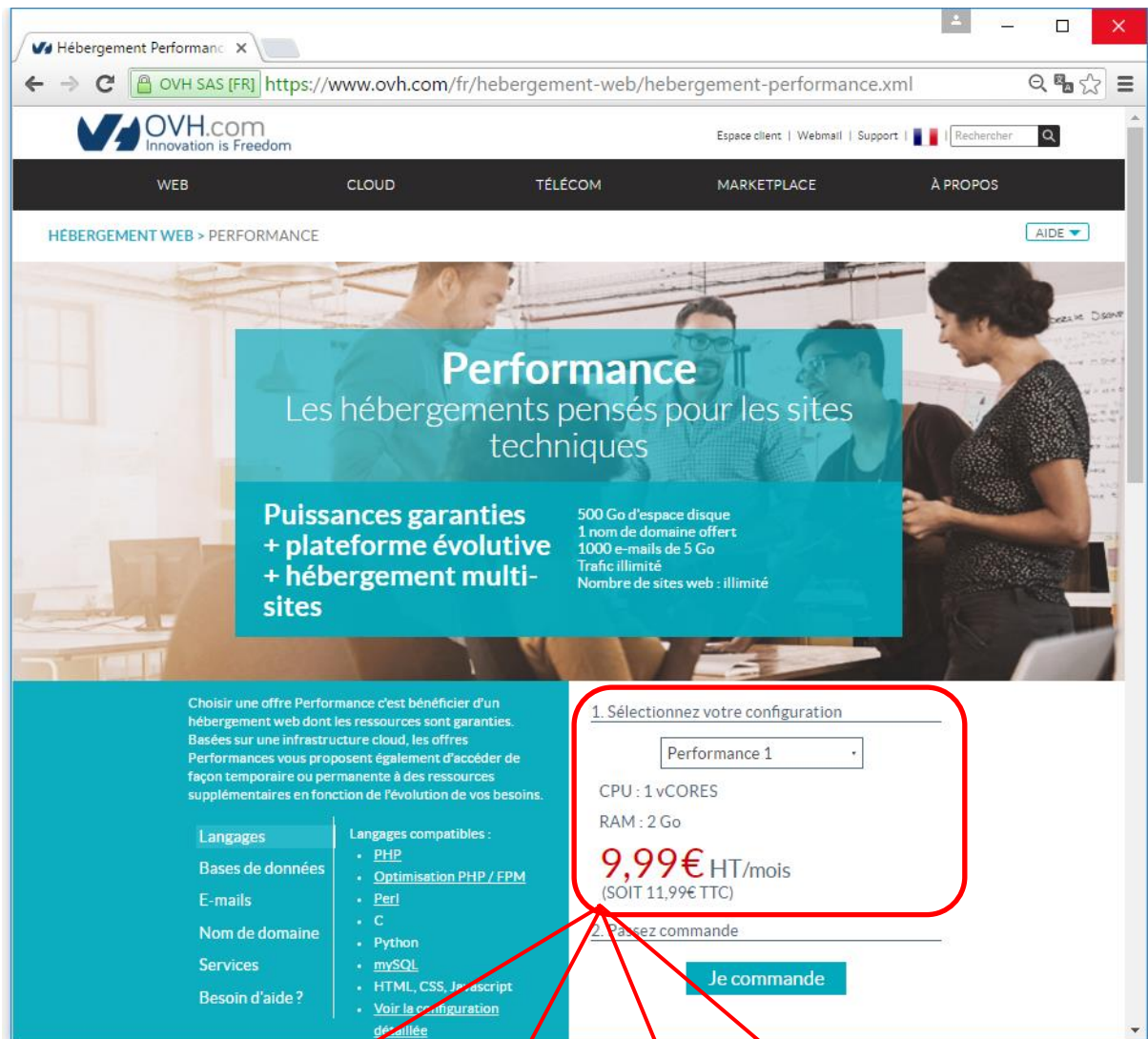
Most of the time, one of the best solution is simply to give to each of the analysts good "Core I7" laptops. In this way, each analyst has the usage of 100% of its own powerful Core 17 CPU: for more information about this subject/infrastructure, see the section 2.1 here below.

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:



Performance
Les hébergements pensés pour les sites techniques

Puissances garanties + plateforme évolutive + hébergement multi-sites

500 Go d'espace disque
1 nom de domaine offert
1000 e-mails de 5 Go
Trafic illimité
Nombre de sites web : illimité

Choisir une offre Performance c'est bénéficier d'un hébergement web dont les ressources sont garanties. Basées sur une infrastructure cloud, les offres Performances vous proposent également d'accéder de façon temporaire ou permanente à des ressources supplémentaires en fonction de l'évolution de vos besoins.

Langages compatibles :

- PHP
- Optimisation PHP / FPM
- Perl
- C
- Python
- MySQL
- HTML, CSS, Javascript
- Voir la configuration détaillée

1. Sélectionnez votre configuration

Performance 1

CPU : 1 vCORES
RAM : 2 Go
9,99€ HT/mois
(SOIT 11,99€ TTC)

2. Passez commande

Je commande

1. Sélectionnez votre configuration

Performance 1

CPU : 1 vCORES
RAM : 2 Go
9,99€ HT/mois
(SOIT 11,99€ TTC)

1. Sélectionnez votre configuration

Performance 2

CPU : 2 vCORES
RAM : 4 Go
18,99€ HT/mois
(SOIT 22,79€ TTC)

1. Sélectionnez votre configuration

Performance 3

CPU : 3 vCORES
RAM : 6 Go
26,99€ HT/mois
(SOIT 32,39€ TTC)

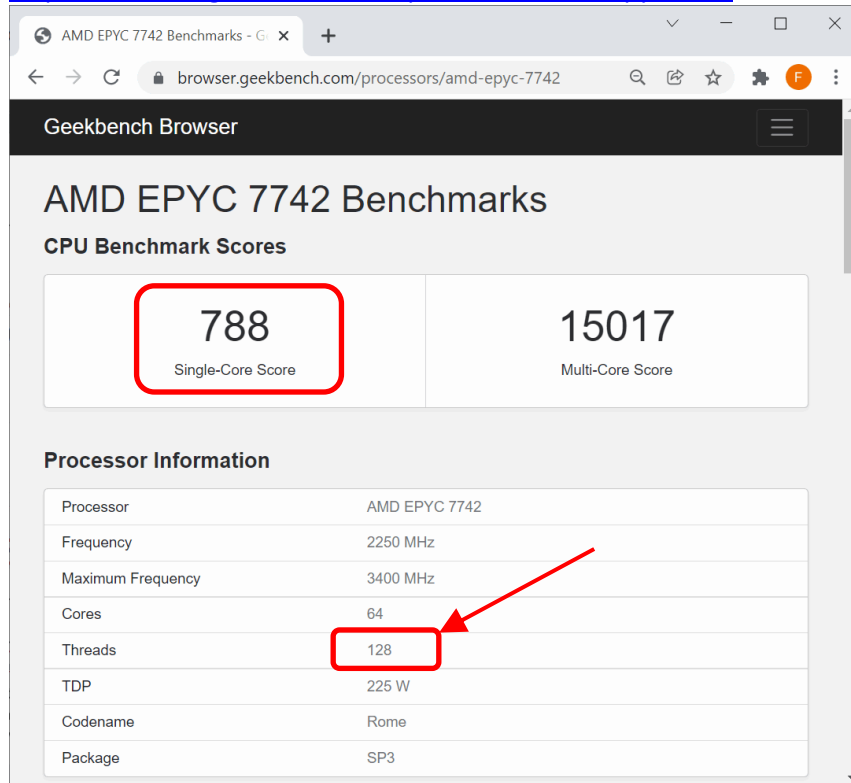
1. Sélectionnez votre configuration

Performance 4

CPU : 4 vCORES
RAM : 8 Go
33,99€ HT/mois
(SOIT 40,79€ TTC)

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 2025/4/17 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 i9-13900KS	3.2	24	3130
2	Intel Core i9-14900KF	3.2	24	3088
3	Intel Core i9-14900K	3.2	24	3070
4	Intel Core i9-13900K	3.0	24	2989
5	AMD Ryzen 7 7700X	4.5	8	2984
6	Intel Core i7-14700KF	3.4	20	2983
7	AMD Ryzen 9 7950X	4.5	16	2981
8	AMD Ryzen 9 7900X	4.7	12	2954
9	Intel Core i7-14700K	3.4	20	2937
10	AMD Ryzen 9 7950X3D	4.2	16	2929
11	Intel Core i9-13900KF	3.0	24	2921
12	AMD Ryzen 9 PRO 7945	3.7	12	2901
13	Intel Core i7-13700KF	3.4	16	2897
14	AMD Ryzen 5 7600X	4.7	6	2891
15	AMD Ryzen 7 7700	3.8	8	2883
16	Intel Core i9-14900F	2.0	24	2872
17	AMD Ryzen 9 7900	3.7	12	2867
18	AMD Ryzen 9 7900X3D	4.4	12	2863
19	Intel Core i7-13700K	3.4	16	2856
20	Intel Core i9-14900	2.0	24	2825
21	Intel Core i5-14600K	3.5	14	2817
22	Intel Core i9-13900F	2.0	24	2802
23	Intel Core i5-14600KF	3.5	14	2781
24	Intel Core i5-13600KF	3.5	14	2771
25	AMD Ryzen 9 7945HX3D	2.3	16	2758
26	AMD Ryzen 9 7945HX	2.5	16	2756
27	AMD Ryzen 5 7500F	3.7	6	2754

The best value CPU on the 2025/4/17 for Anatella/TIMi

28	AMD Ryzen 5 7600	3.8	6	2748
29	Intel Core i9-13900	2.0	24	2734
30	AMD Ryzen 7 7800X3D	4.2	8	2725
31	Intel Core i7-14700	2.1	20	2719
32	Intel Core i7-14700F	2.1	20	2716
33	Intel Core i9-12900KS	3.4	16	2705
34	Intel Core i9-14900HX	2.2	24	2694
35	Intel Core i5-13600K	3.5	14	2689
36	Intel Core i7-13700F	2.1	16	2687
37	AMD Ryzen 7 PRO 7745	3.8	8	2672
38	AMD Ryzen 5 7645HX	4.0	6	2657
39	Intel Core i7-13700	2.1	16	2653
40	AMD Ryzen 7 8700G	4.2	8	2650
41	AMD Ryzen 9 7845HX	3.0	12	2645
42	Intel Core i9-12900K	3.2	16	2640
43	Intel Core i5-14600	2.7	14	2634
44	AMD Ryzen 7 7745HX	3.6	8	2631
45	AMD Ryzen 5 PRO 7645	3.8	6	2631
46	Intel Core i9-12900KF	3.2	16	2608
47	Intel Core i7-12700K	3.6	12	2565
48	Intel Core i5-14600T	1.8	14	2553
49	Intel Core i5-14500	2.6	14	2537
50	Intel Core i5-12600KF	3.7	10	2534
51	Intel Core i7-12700KF	3.6	12	2533
52	Intel Core i9-12900F	2.4	16	2531
53	Intel Core i5-13600	2.7	14	2518
54	Intel Core i7-14700HX	2.1	20	2518
55	Intel Core i9-12900	2.4	16	2513
56	Intel Core i7-13700T	1.4	16	2491

57	Intel Core i5-12600K	3.7	10	2483
58	AMD Ryzen 9 7940HS	4.0	8	2453
59	Intel Core i7-14650HX	2.2	16	2448
60	Intel Core i5-13500	2.5	14	2426
61	Intel Core i3-14100	3.5	4	2417
62	Intel Core i7-13700E	1.9	16	2416
63	Intel Core i7-12700	2.1	8	2412
64	Intel Core i7-12700F	2.1	12	2403
65	Intel Core i3-14100F	3.5	4	2399
66	Intel Core i9-11900K	3.5	8	2388
67	Intel Core i5-13600T	1.8	14	2388
68	AMD Ryzen 9 8945HS	4.0	8	2385
69	Intel Core i5-12600	3.3	6	2384
70	Intel Core i5-14500T	1.7	14	2380
71	Intel Core i5-14400	2.5	10	2372
72	AMD Ryzen 7 7840HS	3.8	8	2364
73	AMD Ryzen 5 8500G	3.5	6	2354
74	AMD Ryzen 5 8600G	4.3	6	2349
75	Intel Core i5-14400F	2.5	10	2346
76	AMD Ryzen 7 8845HS	3.8	8	2343
77	Intel Core i5-12500	3.0	6	2327
78	Intel Core i5-14500HX	2.6	14	2327
79	AMD Ryzen 9 PRO 7940HS	4.0	8	2326
80	Intel Core i9-11900KF	3.5	8	2324
81	AMD EPYC 9374F	3.8	32	2321
82	Intel Core i5-13400F	2.5	10	2306
83	Intel Core i5-13400	2.5	10	2297
84	AMD Ryzen 5 7640HS	4.3	6	2296
85	Intel Core i7-11700KF	3.6	8	2286

86	Intel Core i5-13500T	1.6	14	2279
87	Intel Core i3-13100F	3.4	4	2270
88	Intel Core i7-12700T	1.4	12	2269
89	Intel Core i3-13100	3.4	4	2263
90	AMD Ryzen 5 PRO 7640HS	4.3	6	2257
91	AMD Ryzen 7 7840U	3.3	8	2256
92	Intel Core i9-12900T	1.4	16	2255
93	Intel Core i7-11700K	3.6	8	2250
94	Intel Xeon E-2388G	3.2	8	2245
95	Intel Core Ultra 9 185H	2.3	16	2245
96	AMD Ryzen Z1	3.2	6	2243
97	Intel Core i3-12300	3.5	4	2240
98	Intel Pentium II/III	?	1	2231
99	Intel Core i3-12100F	3.3	4	2230
100	AMD Ryzen 5 7540U	3.2	6	2228
101	Intel Core i5-11600KF	3.9	6	2225
102	Intel Core i5-12400F	2.5	6	2220
103	AMD Ryzen 7 PRO 7840HS	3.8	8	2218
104	Intel Core i9-11900F	2.5	8	2217
105	AMD EPYC 9174F	4.1	16	2214
106	AMD Ryzen 9 5950X	3.4	16	2212
107	Intel Core i5-11600K	3.9	6	2209
108	Intel Core i5-12400	2.5	6	2206
109	Intel Xeon W-1350P	4.0	6	2206
110	AMD Ryzen 9 5900X	3.7	12	2204
111	Intel Core i3-12100E	3.2	4	2200
112	AMD Ryzen 7 8840U	3.3	8	2198
113	AMD Ryzen 7 5800X	3.8	8	2194
114	Intel Core i5-13400T	1.3	10	2189

115	Intel Core i9-11900	2.5	8	2188
116	Intel U300E	1.1	5	2185
117	AMD Ryzen 9 PRO 5945	3.0	12	2183
118	AMD Ryzen Z1 Extreme	3.3	8	2174
119	AMD Ryzen 5 8640U	3.5	6	2170
120	Intel Core i5-12500T	2.0	6	2164
121	Intel Core 7 150U	1.8	10	2163
122	Intel Core i3-12100	3.3	4	2157
123	AMD Ryzen 7 7840U	3.3	8	2149
124	AMD Ryzen 7 PRO 7840U	3.3	8	2147
125	AMD Ryzen 7 5700X	3.4	8	2146
126	AMD Ryzen 5 8640HS	3.5	6	2140
127	AMD Ryzen 7 PRO 5845	3.4	8	2137
128	Intel Xeon W-1350	3.3	6	2135
129	Intel Core i7-11700F	2.5	8	2134
130	AMD EPYC 9254	2.9	24	2132
131	AMD Ryzen 5 7640U	3.5	6	2130
132	AMD Ryzen Threadripper PRO 5955WX	4.0	16	2129
133	AMD Ryzen 9 5900	3.0	12	2126
134	Intel Core i7-11700	2.5	8	2125
135	Intel Core i3-12300T	2.3	4	2121
136	AMD Ryzen 7 5800X3D	3.4	8	2117
137	AMD Ryzen 5 5600X	3.7	6	2114
138	Intel Xeon W-1370	2.9	8	2114
139	Intel Xeon W-11865MRE	2.6	8	2113
140	Intel Xeon w9-3495X	1.9	56	2110
141	AMD Ryzen 5 PRO 7540U	3.2	6	2110
142	Intel Xeon w5-2465X	3.1	16	2107
143	Intel Core i5-11600	2.8	6	2104

144	Intel Xeon w5-2455X	3.2	12	2102
145	Intel Xeon w5-3435X	3.1	16	2102
146	Intel Xeon w7-2495X	2.5	24	2097
147	Intel Xeon E-2324G	3.1	4	2095
148	Intel Core i7-12700E	2.1	12	2094
149	AMD Ryzen 5 7540U	3.2	6	2094
150	AMD Ryzen 5 5600X3D	3.3	6	2085
151	Intel Core i9-11900T	1.5	8	2083
152	AMD Ryzen 7 5800	3.4	8	2081
153	Intel Xeon w7-2475X	2.6	20	2077
154	AMD Ryzen 7 7840U	3.3	8	2076
155	AMD EPYC 9684X	2.6	96	2068
156	AMD Ryzen Threadripper PRO 5945WX	4.1	12	2061
157	Intel Xeon W-11955M	2.6	8	2059
158	AMD Ryzen 7 8840HS	3.3	8	2055
159	AMD Ryzen 5 5600	3.5	6	2053
160	AMD Ryzen Threadripper PRO 5975WX	3.6	32	2048
161	Intel Core i5-12600T	2.1	6	2048
162	Intel Xeon E-2336	2.9	6	2048
163	Intel Core i5-12400T	1.8	6	2044
164	Intel Xeon E-2356G	3.2	6	2044
165	AMD Ryzen 5 5600GT	3.6	6	2039
166	AMD Ryzen Threadripper PRO 5965WX	3.8	24	2024
167	AMD Ryzen 7 5700G	3.8	8	2014
168	AMD Ryzen 7 5700GE	3.2	8	2010
169	Intel Core i5-11500	2.7	6	2007
170	Intel Core i3-13100E	3.3	4	2001
171	Intel Xeon W-11855M	3.2	6	1999
172	Intel Core i7-11390H	3.4	4	1998

173	AMD Ryzen 9 PRO 6950H	3.3	8	1998
174	AMD Ryzen 7 5700	3.7	8	1997
175	Intel Xeon Gold 6442Y	2.6	24	1996
176	Intel Xeon w7-3445	2.6	20	1994
177	AMD Ryzen 9 6900HX	3.3	8	1993
178	Intel Core i7-11850HE	2.6	8	1990
179	Intel Core Ultra 5 134U	0.7	12	1987
180	AMD EPYC 9454P	2.8	48	1986
181	Intel Core i5-11400F	2.6	6	1982
182	Intel Xeon Platinum 8462Y+	2.8	32	1982
183	Intel Pentium Gold G7400	3.7	2	1976
184	AMD Ryzen 7 PRO 5750G	3.8	8	1975
185	AMD EPYC 9654P	2.4	96	1974
186	Intel Core i5-11400	2.6	6	1961
187	Intel Xeon w7-3465X	2.5	28	1959
188	Intel Core i3-12100T	2.2	4	1949
189	Intel Core i7-1265U	1.8	10	1944
190	AMD Ryzen 7 PRO 5750GE	3.2	8	1943
191	Intel Core i7-1260U	1.1	10	1942
192	AMD Ryzen 7 7735HS	3.2	8	1940
193	Intel Xeon E-2378	2.6	8	1940
194	AMD Ryzen 9 5980HX	3.3	8	1939
195	AMD Ryzen Threadripper PRO 5995WX	2.7	64	1938
196	AMD Ryzen 5 5600G	3.9	6	1936
197	Intel Core i7-11850H	2.5	8	1933
198	AMD Ryzen 5 5600GE	3.4	6	1930
199	Intel Core i7-11600H	2.9	6	1925
200	AMD Ryzen 7 5700X3D	3.0	8	1925
201	AMD Ryzen 5 PRO 5650GE	3.4	6	1914

202	AMD Ryzen 5 PRO 5650G	3.9	6	1913
203	AMD EPYC 9554P	3.1	64	1911
204	Intel Xeon w9-3475X	2.2	36	1907
205	AMD Ryzen 9 PRO 6950HS	3.3	8	1903
206	Intel Core i7-11800H	2.3	8	1902
207	AMD EPYC 9554	3.1	64	1902
208	Intel Core i5-12500TE	1.9	6	1900
209	Intel Xeon w3-2435	3.1	8	1876
210	Intel Xeon w5-3425	3.2	12	1875
211	AMD Ryzen 5 5500	3.6	6	1874
212	AMD Ryzen 7 6800H	3.2	8	1871
213	AMD Ryzen 7 PRO 6850HS	3.2	8	1864
214	AMD Ryzen 7 PRO 6850H	3.2	8	1857
215	AMD Ryzen 9 6900HS	3.3	8	1850
216	AMD Ryzen 9 5900HX	3.3	8	1845
217	Intel Core i5-1235U	1.3	10	1841
218	AMD EPYC 9654	2.4	96	1837
219	Intel Core i5-11320H	3.2	4	1834
220	Intel Core i3-1215U	1.2	6	1833
221	Intel Xeon w7-3455	2.5	24	1826
222	Intel Xeon w5-2445	3.1	10	1825
223	AMD Ryzen 5 6600H	3.3	6	1824
224	AMD EPYC 9124	3.0	16	1823
225	Intel Core i5-11400H	2.7	6	1820
226	Intel Core i7-1195G7	2.9	4	1820
227	Intel Xeon W-3375	2.5	38	1818
228	AMD Ryzen 7 6800HS	3.2	8	1817
229	Intel Core i7-11700T	1.4	8	1817
230	AMD Ryzen 3 PRO 5350G	4.0	4	1816

231	Intel Core i5-11500H	2.9	6	1812
232	Intel Core i9-9900KS	4.0	8	1810
233	Intel Xeon w3-2425	3.0	6	1808
234	Intel Core i7-11375H	3.3	4	1797
235	Intel Core i5-11600T	1.7	6	1791
236	Intel Xeon Silver 4416+	2.0	20	1785
237	Intel Core i9-10900K	3.7	10	1780
238	Intel Core i7-11370H	3.3	4	1777
239	AMD Ryzen 7 7735U	2.7	8	1777
240	AMD Ryzen 3 PRO 5350GE	3.6	4	1776
241	Intel Core i5-12500E	2.9	6	1776
242	Intel Core i5-7640X	4.0	4	1772
243	AMD Ryzen 5 7535HS	3.3	6	1772
244	AMD Ryzen 7 3800XT	3.9	8	1770
245	Intel Core i9-10900KF	3.7	10	1769
246	AMD Ryzen 3 5300G	4.0	4	1766
247	AMD Ryzen 5 7535U	2.9	6	1766
248	Intel Xeon W-1290	3.2	10	1764
249	Intel Core i9-10850K	3.6	10	1759
250	AMD Ryzen 7 5800H	3.2	8	1758
251	Intel Core i7-1185G7E	2.8	4	1756
252	AMD Ryzen 7 PRO 6850U	2.7	8	1754
253	AMD Ryzen 7 5800U	1.9	8	1744
254	Intel Core i9-10910	3.6	10	1742
255	Intel Core i3-13100T	2.5	4	1742
256	AMD Ryzen 9 3900XT	3.8	12	1737
257	Intel Core i7-1165G7	2.8	4	1736
258	Intel Core i7-1185G7	3.0	4	1736
259	AMD Ryzen 5 PRO 6650U	2.9	6	1736

260	AMD Ryzen 5 PRO 7530U	2.0	6	1734
261	Intel Core i5-1155G7	2.5	4	1731
262	AMD EPYC 9354	3.2	32	1731
263	Intel Core i5-11260H	2.6	6	1730
264	AMD Ryzen 5 PRO 6650H	3.3	6	1730
265	AMD Ryzen 7 7736U	2.7	8	1729
266	AMD Ryzen 9 3950X	3.5	16	1727
267	Intel Core i7-10700KF	3.8	8	1725
268	Intel Core i7-10700K	3.8	8	1722
269	AMD Ryzen 7 5825U	2.0	8	1722
270	Intel Core i9-12900E	2.3	16	1722
271	Intel Core i9-10900	2.8	10	1721
272	AMD Ryzen 5 3600XT	3.8	6	1720
273	Intel Xeon W-1270	3.4	8	1717
274	Intel Core i9-9900KF	3.6	8	1714
275	Intel Core i9-10900F	2.8	10	1712
276	AMD Ryzen 7 PRO 5850U	1.9	8	1712
277	AMD Ryzen 7 6800U	2.7	8	1709
278	AMD Ryzen 9 3900X	3.8	12	1708
279	Intel Xeon Gold 6426Y	2.5	16	1706
280	AMD Ryzen 3 3300X	3.8	4	1705
281	Intel Core i9-9900K	3.6	8	1703
282	AMD Ryzen 7 3800X	3.9	8	1703
283	Intel Xeon W-1290P	3.7	10	1703
284	Intel Core i5-9600K	3.7	6	1697
285	Intel Core i3-10325	3.9	4	1697
286	Intel Core i7-9700K	3.6	8	1695
287	Intel Core i7-8086K	4.0	6	1694
288	Intel Core i5-10600KF	4.1	6	1694

289	Intel Core i7-9700KF	3.6	8	1693
290	Intel Core i5-10600K	4.1	6	1692
291	Intel Celeron G6900	3.4	2	1686
292	AMD Ryzen Threadripper PRO 3955WX	3.9	16	1685
293	AMD Ryzen Threadripper 3960X	3.8	24	1682
294	Intel Core i9-10920X	3.5	12	1680
295	AMD Ryzen Threadripper PRO 3945WX	4.0	12	1680
296	AMD Ryzen 9 PRO 3900	3.1	12	1680
297	Intel Core i5-11300H	3.1	4	1679
298	AMD Ryzen 9 3900	3.1	12	1678
299	Intel Core i5-1145G7	2.6	4	1677
300	Intel Pentium Gold G7400T	3.1	2	1674
301	Intel Core i5-11500T	1.5	6	1673
302	AMD EPYC 7443P	2.8	24	1672
303	AMD Ryzen 7 3700X	3.6	8	1671
304	AMD Ryzen 5 5560U	2.3	6	1670
305	AMD Ryzen 5 3600X	3.8	6	1668
306	AMD Ryzen Threadripper 3970X	3.7	32	1661
307	Intel Core i7-1250U	1.1	10	1653
308	AMD Ryzen Threadripper PRO 3975WX	3.5	32	1648
309	Intel Xeon W-1250P	4.1	6	1647
310	Intel Core i9-13900E	1.8	24	1646
311	AMD EPYC 9534	2.4	64	1646
312	Intel Xeon w3-2423	2.1	6	1644
313	AMD Ryzen 7 PRO 3700	3.6	8	1641
314	AMD Ryzen 5 6600U	2.9	6	1640
315	Intel Core i5-9600KF	3.7	6	1638
316	AMD Ryzen 5 7530U	2.0	6	1638
317	AMD EPYC 75F3	3.0	32	1637

318	AMD Ryzen 7 7730U	2.0	8	1636
319	Intel Core i7-10700F	2.9	8	1635
320	AMD Ryzen 7 PRO 4750G	3.6	8	1635
321	AMD Ryzen 5 PRO 5675U	2.3	6	1635
322	Intel Xeon E-2288G	3.7	8	1634
323	Intel Xeon E-2278G	3.4	8	1633
324	AMD Ryzen 5 PRO 3600	3.6	6	1633
325	AMD Ryzen 5 5600U	2.3	6	1630
326	Intel Core i7-10700	2.9	8	1626
327	Intel Core i7-7740X	4.3	4	1626
328	Intel Xeon E-2276G	3.8	6	1623
329	Intel Xeon Platinum 8351N	2.4	36	1617
330	Intel Core i7-8700K	3.7	6	1616
331	AMD Ryzen 5 PRO 5650U	2.3	6	1616
332	AMD Ryzen 5 3600	3.6	6	1614
333	Intel Core i5-10600	3.3	6	1614
334	AMD Ryzen Threadripper 3990X	2.9	64	1613
335	AMD Ryzen 3 PRO 5475U	2.7	4	1612
336	Intel Core i5-1230U	1.0	10	1610
337	AMD Ryzen 7 PRO 4750GE	3.1	8	1610
338	Intel Xeon E-2274G	4.0	4	1610
339	Intel Xeon E-2246G	3.6	6	1610
340	Intel Core i9-9900	3.1	8	1608
341	AMD EPYC 9754	2.2	128	1603
342	Intel Xeon E-2334	3.4	4	1603
343	Intel Xeon E-2286G	4.0	6	1602
344	Intel Core i9-10980XE	3.0	18	1597
345	Intel Core i5-1135G7	2.4	4	1597
346	Intel Xeon W-2275	3.3	14	1593

347	AMD Ryzen 3 PRO 7330U	2.3	4	1593
348	Intel Core i7-7700K	4.2	4	1592
349	Intel Core i3-1115G4	3.0	2	1591
350	AMD Ryzen 5 5625U	2.3	6	1589
351	Intel Core i7-9700F	3.0	8	1585
352	Intel Core i9-10900X	3.7	10	1584
353	Intel Xeon E-2176G	3.7	6	1584
354	AMD Ryzen 5 5600H	3.3	6	1583
355	Intel Core i5-8600K	3.6	6	1581
356	Intel Core i5-11400T	1.3	6	1580
357	AMD Ryzen 5 3500X	3.6	6	1579
358	AMD Ryzen 5 4600G	3.7	6	1577
359	Intel Xeon E-2126G	3.3	6	1577
360	AMD Ryzen 7 PRO 5875U	2.0	8	1575
361	Intel Core i7-9700	3.0	8	1574
362	AMD Ryzen Threadripper PRO 3995WX	2.7	64	1574
363	Intel Xeon E-2236	3.4	6	1574
364	Intel Core i9-10900E	2.8	10	1572
365	Intel Core i3-10320	3.8	4	1571
366	Intel Xeon W-10885M	2.4	8	1568
367	Intel Core i7-1068NG7	2.3	4	1567
368	Intel Core i5-10505	3.2	6	1567
369	Intel Core i7-12700TE	1.4	12	1567
370	Intel Xeon W-2245	3.9	8	1565

371	AMD Ryzen 5 PRO 4650G	3.7	6	1563
372	AMD Ryzen 5 4500	3.6	6	1561
373	Intel Xeon E-2286M	2.4	8	1561
374	Intel Celeron G6900E	3.0	2	1561
375	Intel Core i9-10940X	3.3	14	1559
376	AMD EPYC 7313P	3.0	16	1558
377	AMD Ryzen 9 4900H	3.3	8	1557
378	Intel Xeon W-2255	3.7	10	1556
379	Intel Xeon E-2174G	3.8	4	1551
380	Intel Xeon W-2235	3.8	6	1549
381	Intel Xeon E-2224G	3.5	4	1548
382	Intel Xeon E-2186G	3.8	6	1544
383	AMD EPYC 9354P	3.2	32	1540
384	Intel Core i5-9600	3.1	6	1537
385	Intel Core i3-10300	3.7	4	1537
386	AMD EPYC 7773X	2.2	64	1536
387	Intel Core i3-10305	3.8	4	1536
388	Intel Core i7-8700	3.2	6	1535
389	Intel Xeon E-2146G	3.5	6	1534
390	AMD EPYC 7343	3.2	16	1533
391	AMD Ryzen 7 PRO 7730U	2.0	8	1531
392	Intel Core i5-10500	3.1	6	1528
393	Intel Xeon E-2244G	3.8	4	1527
394	Intel Core i5-1038NG7	2.0	4	1526

395	Intel Xeon Gold 6334	3.6	8	1526
396	Intel Core i5-7600K	3.8	4	1525
397	Intel Xeon E-2136	3.3	6	1525
398	AMD Ryzen 5 5500H	3.3	4	1525
399	Intel Core i7-10875H	2.3	8	1521
400	AMD Ryzen 3 PRO 5450U	2.6	4	1518
401	AMD Ryzen 5 3500	3.6	6	1517
402	Intel Xeon W-2295	3.0	18	1516
403	Intel Core i3-9350KF	4.0	4	1516
404	AMD Ryzen 3 4100	3.8	4	1513
405	AMD EPYC 7713	2.0	64	1511
406	AMD Ryzen 5 PRO 4650GE	3.3	6	1510
407	Intel Core i3-8350K	4.0	4	1509
408	AMD Ryzen 3 5425U	2.7	4	1509
409	Intel Xeon W-2225	4.1	4	1508
410	Intel Xeon Silver 4410Y	2.0	12	1508
411	Intel Xeon W-10855M	2.8	6	1507
412	Intel Core i5-9500F	3.0	6	1506
413	Intel Xeon E-2144G	3.6	4	1506
414	Intel Core i5-10500E	3.1	6	1503
415	Intel Core i7-6700K	4.0	4	1502
416	Intel Core i7-9700E	2.6	8	1501
417	Intel Core i9-9900T	2.1	8	1500

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 2200) 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 6 of the above table, we find a cheap CPU (the "Intel Core i7-14700KF" that costs around than \$350 on the 2025/4/17). 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 93 of the above table, we find a 4 years old CPU (the "Intel Core i7-11700K" that was released in Q1 2021). All the CPU's slower than this very, very old CPU (i.e. all the CPU's located at a ranking number above 93) 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/SSD BUT VERY BAD CPU's (i.e. they have low-grade XEON servers such as the "Intel Xeon Platinum 8351N") 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 April 2025)

1.5.1. You buy the hardware

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 2025/4/17):

<https://www.ldlc.com/fr-be/fiche/PB00677439.html>

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

- CPU: Intel Core i7-14700KF (3.4 GHz / 5.6 GHz)
- Mother board: MSI Z790 GAMING PLUS WIFI
- RAM: DDR5 32 Go (2 x 16 Go) 3200 MHz
- System disk: SSD PCI-E NVMe 3.0 1TB
- (optional) Mass Storage Disc: **to replace: see below**
- GPU: NVIDIA GeForce RTX 5060 Ti 16 Go
- Network Card : 2.5 Gigabit LAN
- 8 high-definition audio canals
- Power Supply: MSI MAG A650BE 650 watt
- CPU cooler: Fox Spirit LightFlow XT360 ARGB (watercooling)
- Box: Cooler Master MasterBox MB520 Mesh ARGB
- Microsoft Windows 11 Family 64 bits

The screenshot shows the LDLC website interface. At the top, there's a navigation bar with 'TOUS NOS PRODUITS', 'CONFIGURATEUR PC', 'VOS CONSOMMABLES', 'PROMOTIONS', 'NOUVEAUTÉS', and 'BESOIN D'AIDE'. The main content area displays the 'LDLC PC11 Plus Perfect' product page. It includes a product image, a star rating of 4.5 from 9 clients, and a 'Poser une question' button. The description highlights the Intel Core i7 processor, NVIDIA GeForce RTX 5060 Ti graphics, and Windows 11. The price is prominently displayed as 1949€95, with a 4x payment option for 497€84 x4. The page also features a 'Garantie 3 ans' badge and a 'Payer en 4x' button.

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).

- Add a mass-storage SSD drive: We recommend Samsung SSD drives (SSD drives from Samsung are a ***lot more reliable*** than any other SSD or HDD drives). With the motherboard included inside the “LDLC PC11 Plus Perfect” (the pc suggested here above), you can mount up to 4 of these “M.2. NVMe” drives (either 4TB each or 8TB each) and 4 “SSD SATA” (8TB each) for a whopping total of $4 \times 8 + 4 \times 8 = 64\text{TB}$ of storage.

<https://www.ldlc.com/fiche/PB00653109.html>

Samsung SSD 990 EVO Plus M.2 PCIe NVMe 4 To

SSD 4 To M.2 2280 NVMe 2.0 - PCIe 5.0 x2/PCIe 4.0 x4

★★★★★ 26 avis clients | 3 questions / réponses

Le disque SSD 990 EVO Plus 4 To de Samsung apporte un boost de performances à votre machine pour des temps de chargement réduits, des transferts rapides de fichiers volumineux. Ces performances sont rendues possibles grâce à des vitesses de lecture jusqu'à 7250 Mo/s.

Capacité : 4096 Go

2048 Go	1024 Go	4096 Go
139€95	91€95	279€95

279€95
Eco-part. : 0€05

Payer en 3x
95€98 x3
dont 7€99 de frais

Quantité : 1

AJOUTER AU PANIER

ACHETER CET ARTICLE

Être informé d'une baisse de prix

DISPO WEB : EN STOCK | DISPO BOUTIQUE : Dispo dans 8 boutiques

Livraison possible demain avant 13h
Livraison possible en soirée
Livré aujourd'hui en région lyonnaise
Vérifier mon code postal
Livraison possible en Chrono Relais
Reprise de votre ancien produit
En savoir plus

Read-Speed: 7.25 GByte/sec ; Write-speed: 6.3 GByte/sec

Samsung SSD 9100 PRO M.2 PCIe NVMe 8 To

SSD 8 To M.2 2280 NVMe 2.0 - PCIe 5.0 x4

★★★★★ Un avis client | Poser une question

Le disque Samsung SSD 9100 PRO M.2 PCIe NVMe 2 To offre des performances de nouvelle génération avec des taux de transferts élevés et latences ultra-réduites. Il est également compatible PlayStation 5 afin d'en amplifier les performances ou d'en augmenter la capacité de stockage.

Capacité : 8192 Go

1024 Go	2048 Go	4096 Go	8192 Go
169€95	259€95	469€95	1129€95

Dissipateur thermique : Non
Oui Non

1129€95
Eco-part. : 0€07

En 4x | En 10x
288€75 x4
dont 24€97 de frais

Être prévenu de la disponibilité

DISPO WEB : RUPTURE | DISPO BOUTIQUE : Choisir ma boutique

GARANTIE 5 ANS

Read-Speed: 14.8 GByte/sec ; Write-speed: 13.4 GByte/sec

1.5.2. You rent the hardware inside a Cloud data center

You'll find more details on this subject in this document:

http://download.timi.eu/docs/Azure_vm_review_2025-01_eng.pdf

We suggest you to rent a server on <http://hetzner.com>. Why?

- Hetzner is a 100% European cloud provider, so that you won't have any difficulties related to GDPR compliance.
- Hetzner has the servers with the best, more recent CPU's.
- The SSD storages inside all the Hetzner servers are directly connected to the server's motherboard through a high-speed "NVMe M.2" socket. This guarantee a read/write speed of around 2000 MByte/sec.

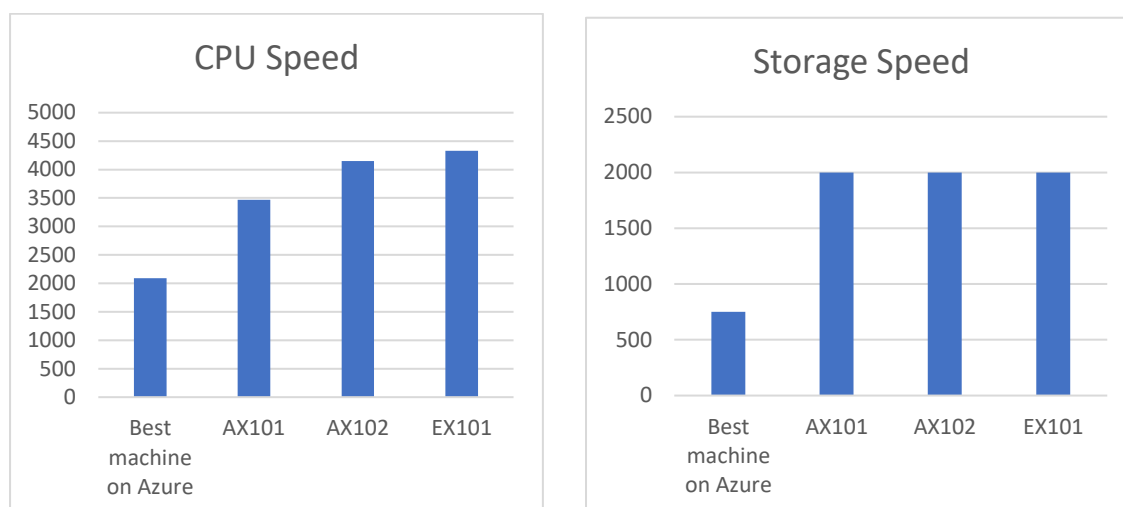
In opposition, in all other cloud providers (Azure, AWS, Google GCP, OVH, etc.), the SSD storage is deported: i.e. it's connected through an ethernet cable to the server instead of being directly connected to the motherboard using a "NVMe M.2" socket. This means that the read/write speed on the SSD will actually depends on the ethernet network load and will be limited by the ethernet cable speed. This is well visible on Azure where you can directly select the ethernet cable speed up to a maximum speed of 750 MB/sec. Warning: The price of the SSD storage increases dramatically with the allocated speed for the ethernet cable. It's common to pay **10K€ per month** for your "SSD Mass storage" to get the maximum speed of only 750 GB/sec.

On Hetzner, we currently recommend the EX101 server.

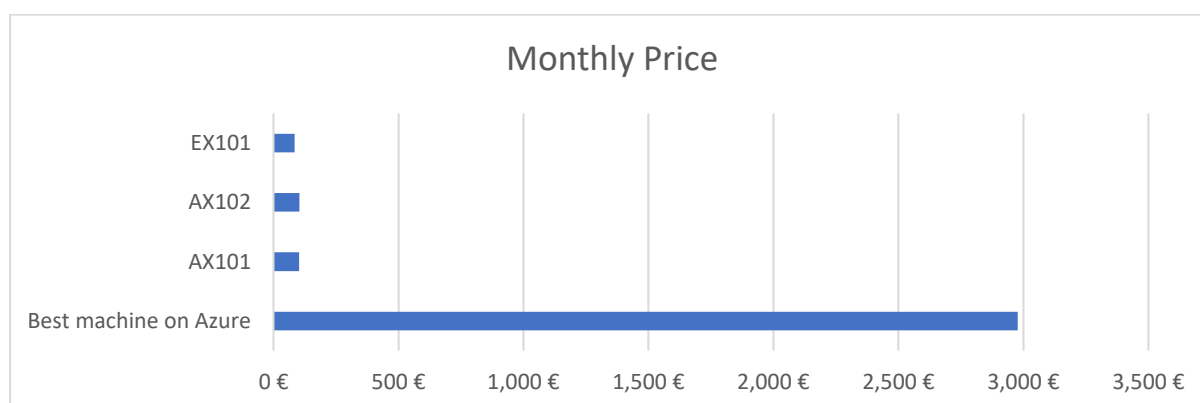
After a careful investigation of the CPU's available inside Azure (for more details, please refer to the PDF accessible through the URL given here above), we finally arrived to this comparison table between Hetzner and Azure:

	Best machine on Azure	AX101	AX102	EX101
CPU name	AMD EPYC 7551	AMD Ryzen 9 5950X	AMD Ryzen 9 7950X3D	Intel Core i9-13900
CPU Speed	2092	3469	4150	4330
RAM	200 GB	128 GB	128 GB	64 GB
Storage capacity	4 TB	8 TB	4 TB	4 TB
Storage Speed	750 MB/sec	2000 MB/sec	2000 MB/sec	2000 MB/sec
Monthly Price	2977 €	102 €	104 €	84 €

Graphically:



Monthly Price:



We recommend to use the EX101 server on Hetzner because it is:

- 35.4 times cheaper than the best available Azure server
- 2.07 times faster in terms of CPU than the best available Azure server
- 2.66 times faster in terms of Storage-Access-Speed (to read & write data on SSD) than the best available Azure server
- The Hetzner server has enough RAM (64GB) to do everything and anything with TIMi.
- The Hetzner server is GDPR compliant (i.e. it's operated by a 100% European company)

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. **Phase 1: The “Exploration phase”**

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, 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. **Phase 2: The “Production phase”**

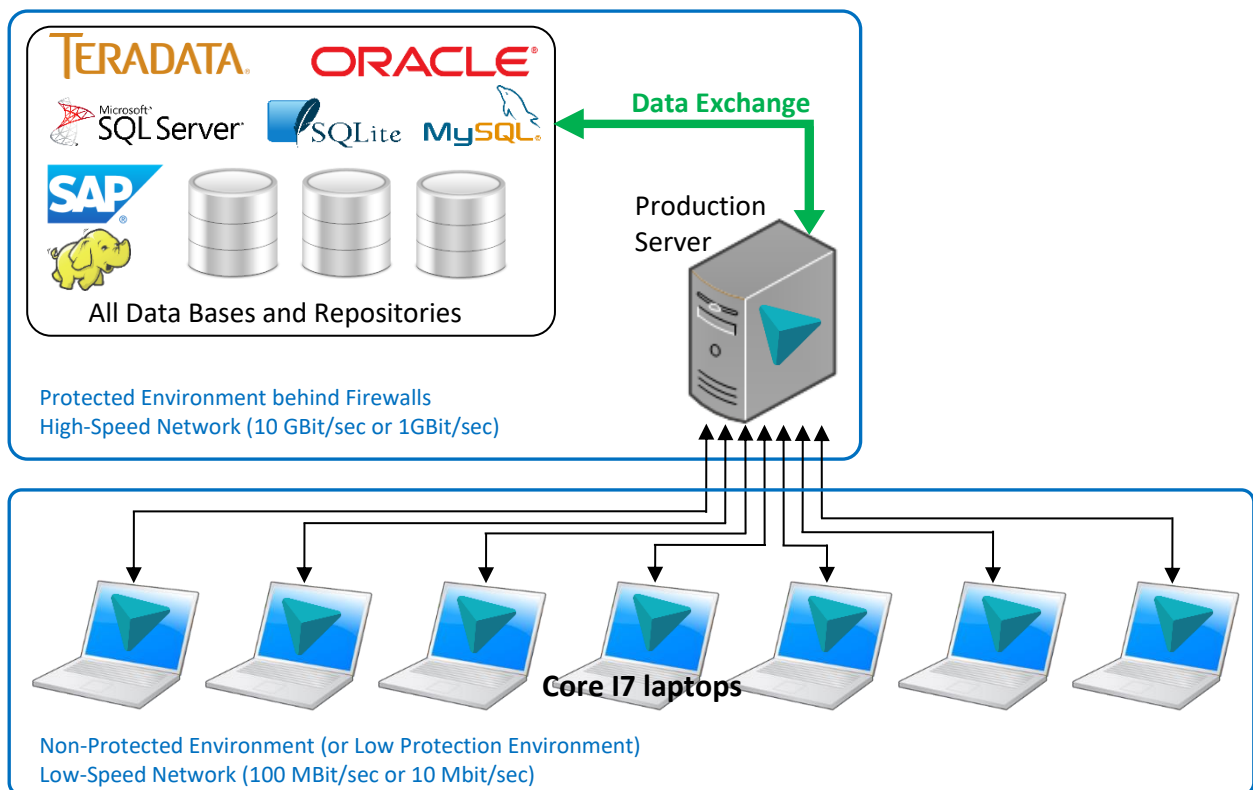
The “Production phase” comes after the “Exploration phase” and usually lasts for years. There are usually no really “heavy” computations during the “Production phase”.

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 expensive 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:

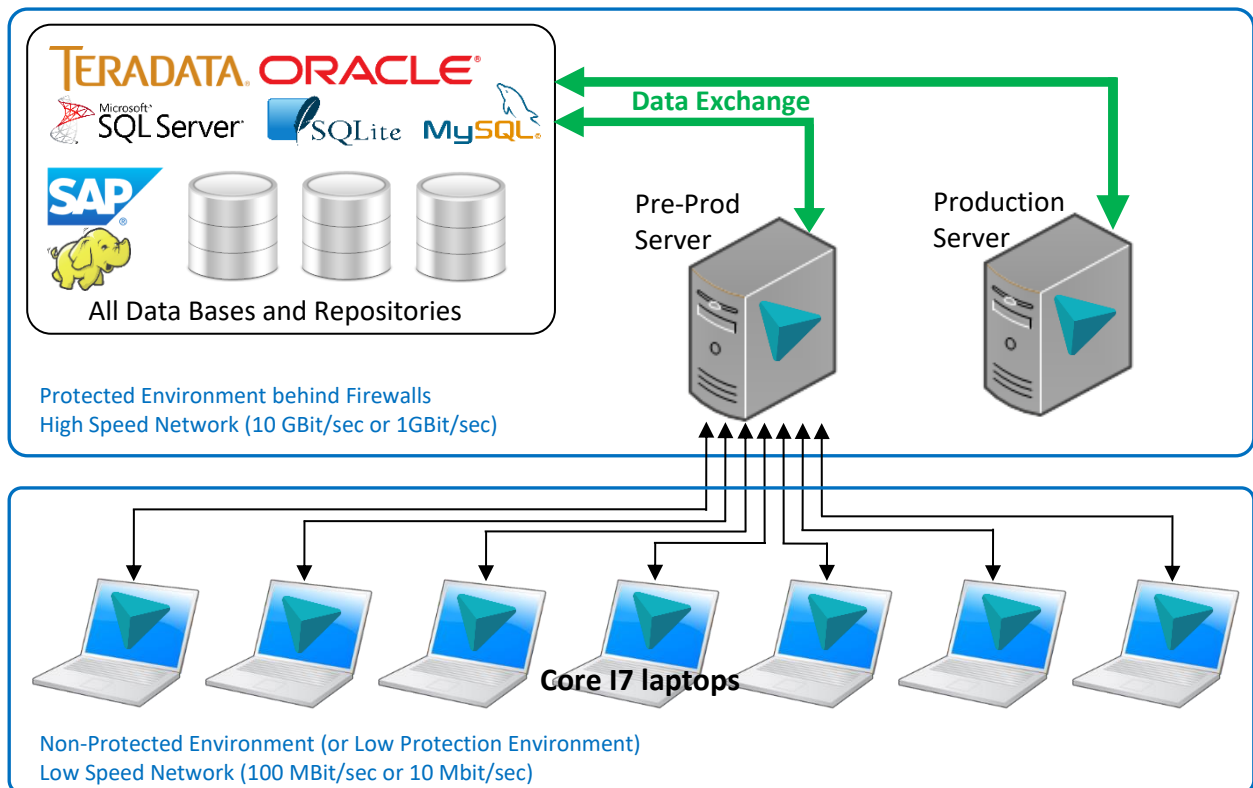
- Pro:
 - **Cheap:** No hardware investment required outside the purchase of a “Production Server”
 - **Scalable:** If you have more analysts, simply add more laptops.
 - **Secure:** Why? Because:
 - The Analysts/Data Scientists cannot delete any critical data from your operational systems because all 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

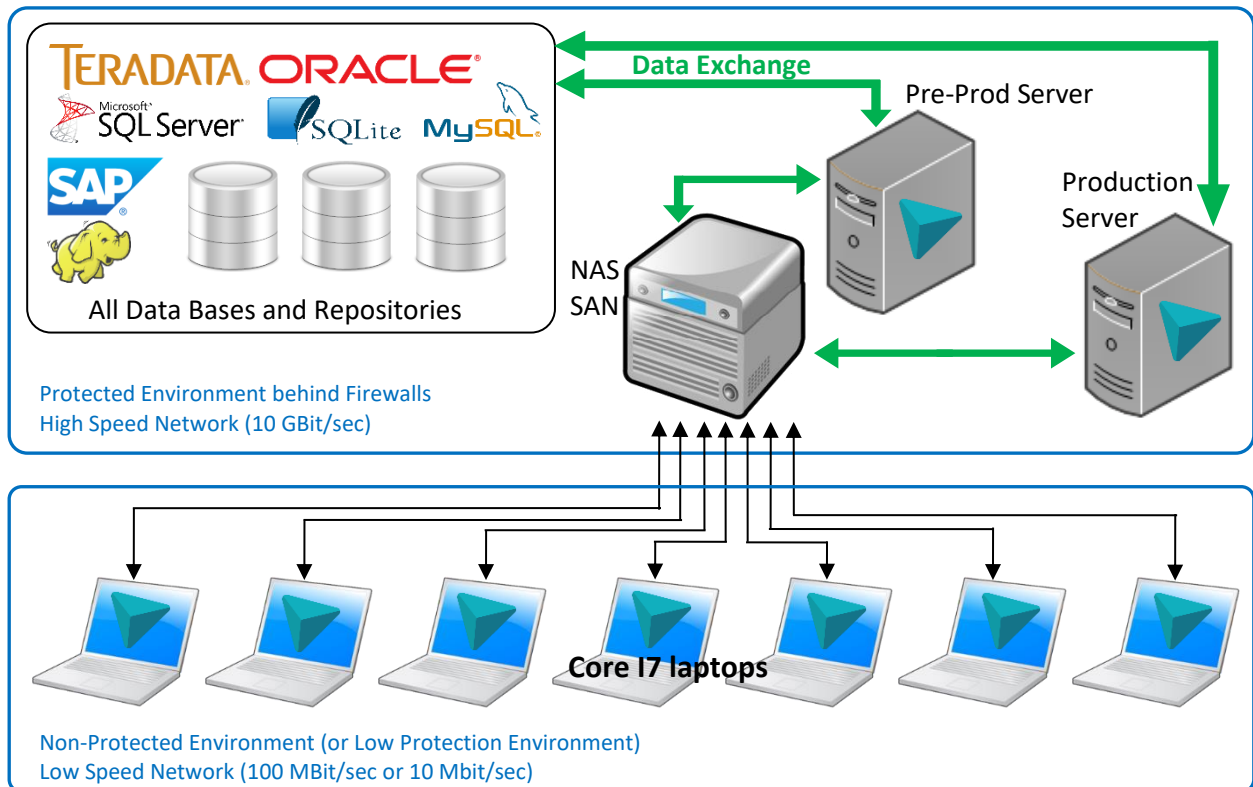
To increase reliability, use a “pre-production” server:



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 $7 \times 100 \text{ MB/sec} = 700 \text{ MB/sec}$ 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 or Python ecosystems (or even worse: in the Hadoop ecosystem), 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) by rather on "in-memory" computations.
- 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.



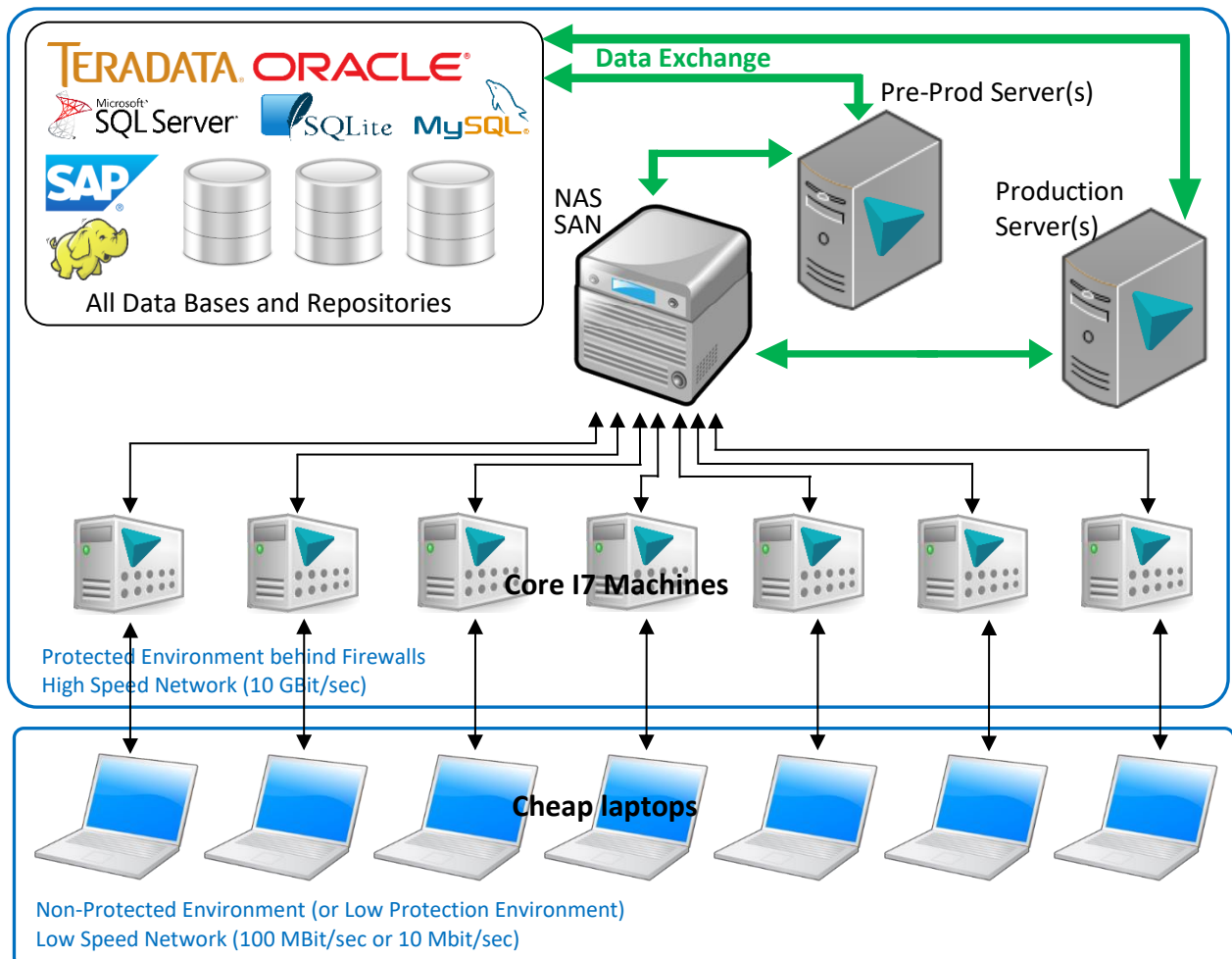
Anatella can also run little “boxes” coded in R or Python. What’s happening for such little box? It's true that the “normal” R and Python engines do not work "in streaming" (such as Anatella) but we managed to "transform & upgrade" these engines so that they can also work (most of the time) in streaming 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.

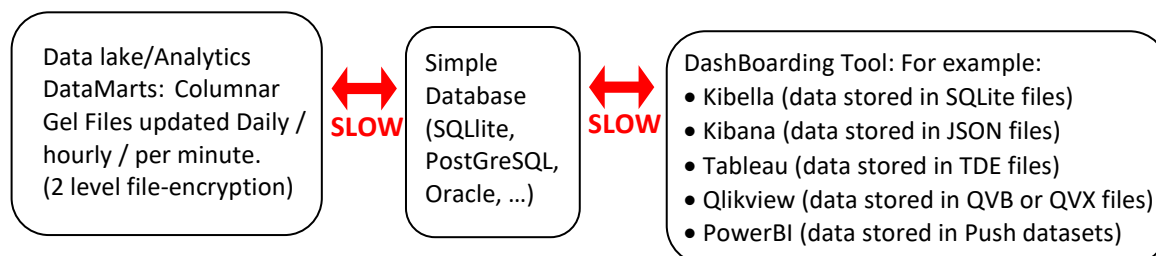
3. Integration with third party tools

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:

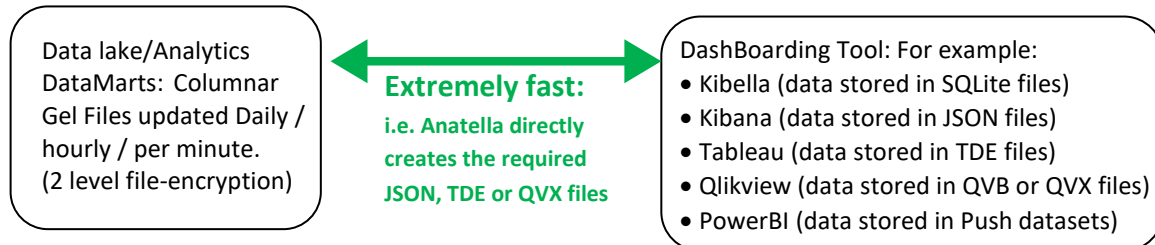
- It works with all Dashboarding tools because (nearly) all the Dashboarding tools can get their data from “common” SQL-based data sources (one exception is Kibana that requires to get its data from JSON files).

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 Qlickview documentation both agree that, getting data from a file (i.e. from a TDE file for Tableau or from a QVX file for Qlickview) 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 Qlickview, 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.



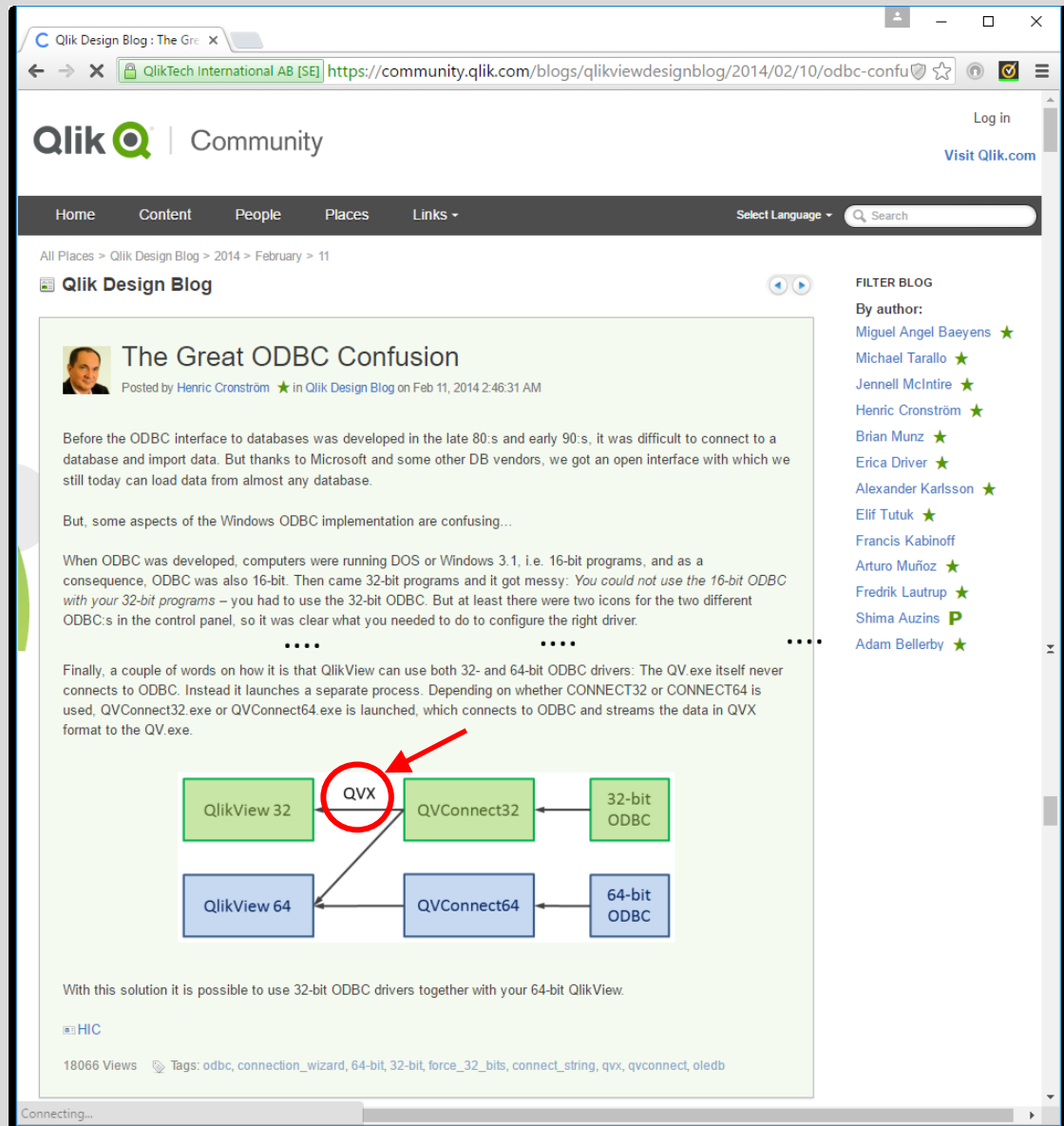
If you intend to use still another Dashboarding tool (i.e. not Tableau, Qlickview or Kibana), you can still get a decent speed by using a SQLite database as the “database-in-the-middle”. Please refer to the section 4.8.2.4. and 4.8.17.6 of the “AnatellaQuickGuide.pdf” for more information about SQLite 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:



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.

```

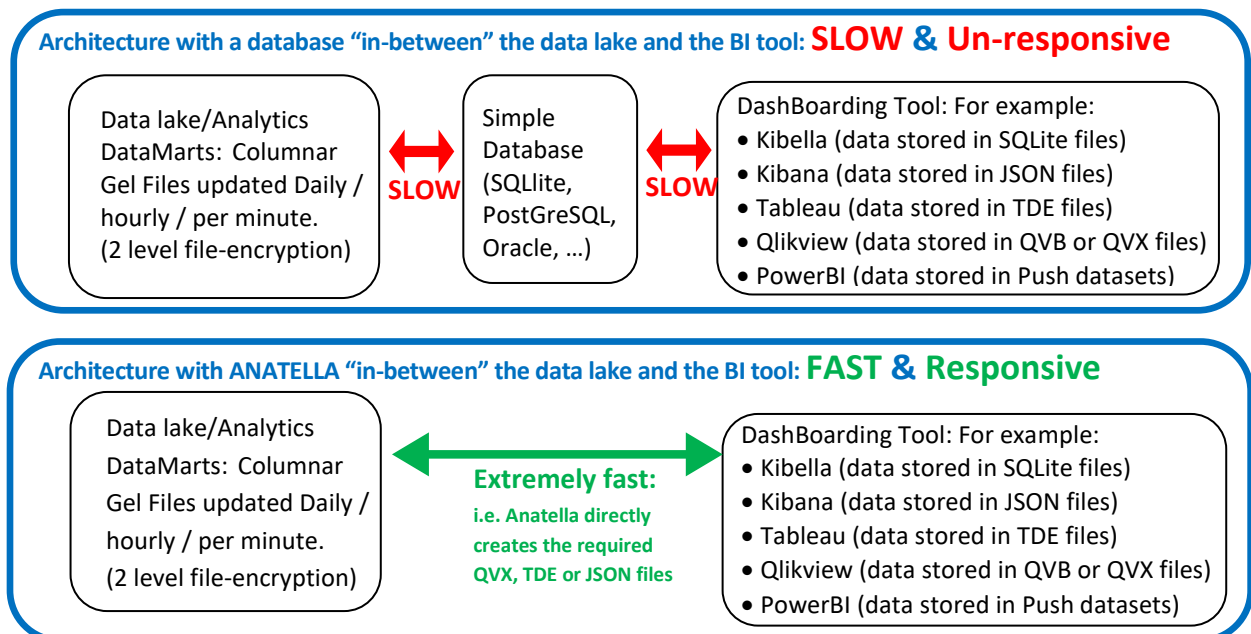
  graph LR
    QV32[QlikView 32] --> QVX((QVX))
    QV32 --> QVConnect32[QVConnect32]
    QVConnect32 --> ODBC32[32-bit ODBC]
    QV64[QlikView 64] --> QVConnect64[QVConnect64]
    QVConnect64 --> ODBC64[64-bit ODBC]
  
```

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

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

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. This situation is extremely uncommon: i.e. 95% of companies won’t need more than one TIMi server.



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.

