Find the most affordable Cloud computing choice to get a Hadoop cluster

APACHE HADOOP IN AMAZON EMR

Link: https://aws.amazon.com/es/emr/details/hadoop/

Amazon EMR makes easy to create and manage fully configured, elastic clusters of Amazon EC2 instances running Hadoop and other applications in the Hadoop ecosystem.

Amazon EMR can be used to easily install and configure tools such as Hive, Pig, Hue, Ganglia, Oozie, and HBase on the user's cluster.

Amazon S3, at the same time, is defined to be highly scalable, low cost, and designed for durability. By storing the data in Amazon S3, the user can decouple his compute layer from his storage layer, allowing him to size his Amazon EMR cluster for the amount of CPU and memory required for his workloads instead of having extra nodes.

1. Advantages of Hadoop on Amazon EMR

- Increased speed and agility

The user can initialize a new Hadoop cluster dynamically and quickly, or add servers to the existing Amazon EMR cluster

- Reduced administrative complexity

Amazon EMR addresses the user's Hadoop infrastructure requirements so he can focus on his core business.

- Integration with other cloud services

Hadoop environment can be easily integrated with other services of Amazon.

- Pay for clusters only when the customer needs them

Amazon EMR allows to start workload clusters easily, save the results, and shut down the Hadoop resources when they are no longer needed, to avoid unnecessary infrastructure costs.

- Improved availability and disaster recover

A potential problem in a determined region can be easily circumvented by launching a cluster in another zone in minutes.

- Flexible capacity

Clusters can be created with the required capacity within minutes and use Auto Scaling to dynamically scale out and scale in nodes.

2. Amazon EMR with the MapR Distribution for Hadoop Pricing

Link: https://aws.amazon.com/emr/mapr/pricing/

Once we have seen the features that Amazon provides in relation with Hadoop Cloud Computing, the different prices (depending on the Amazon Instances) are shown below.

We can select nine regions, in which these prices differ (The collected results belong to Ireland, the nearest available region to Spain with the most available options of models on sale).

There are five Amazon EC2 Instance Types, with different features, perfomances, capacities...:

General Purpose, Computed Optimized, Memory Optimized,Accelerated Computing and Storage Optimized.

Let's inquire into them:

- General Purpose:

This instance has at the same time three different modalities:

T2:

T2 instances are "Burstable Perfomance Instances" that provide a baseline level of CPU performance with the ability to burst above the baseline.

For most general-purpose workloads, T2 Unlimited instances will provide ample performance without any additional charges

Use cases:

Websites and web applications, development environments, build servers, code repositories, micro services, test and staging environments, and line of business applications.

This sub-type summes a total of 7 models, with different capacity of memory and maximum number of vCPU.

M5:

They are the latest generation of General Purpose Instances. This family provides a balance of compute, memory, and network resources, and it is a good choice for many applications.

Use Cases:

Small and mid-size databases, data processing tasks that require additional memory, caching fleets, and for running backend servers for SAP, Microsoft SharePoint, cluster computing, and other enterprise applications.

This sub-type summes a total of 6 models.

M4:

The utilization of this sub-type of instance is recommended in similar cases as the M5 sub-type

This sub-type summes a total of 6 models.

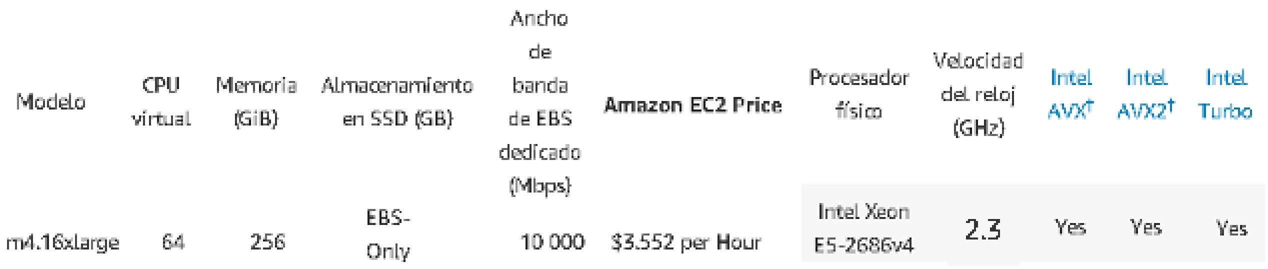
The prices reflected in the following list just collect those ones with Amazon EC2 (there are two groups of cloud cluster providing: Amazon EMR (with higher perfomance but more expensive and Amazon EC2)).

Not all the instances described in the web page are on sale, but just a percentage of them.

I have listed only the sub-types that meet the requierements described in the header of the task for each instance. For example, for "General Purposes", just the model M4 meets the requirements.

Note: The memory is provided in GiB (gibibyte), not in GB, so it is necessary to transform the quantites to see if the requirements of memory are met.

- M4:



Modelo	CPU virtual	Memoria (GiB)	Almacenamiento en SSD (GB)	Ancho de banda de EBS dedicado (Mbps)	**Amazon EC2 Price**	Procesador físico	Velocidad del reloj (GHz)	Intel AVX†	Intel AVX2†	Intel Turbo
m4.16xlarge	64	256	EBS-Only	10 000	$3.552 per Hour	Intel Xeon E5-2686v4	2.3	Yes	Yes	Yes

- Computed Optimized

C5:

Optimized for compute-intensive workloads and deliver very cost-effective high performance at a low price.

Use Cases

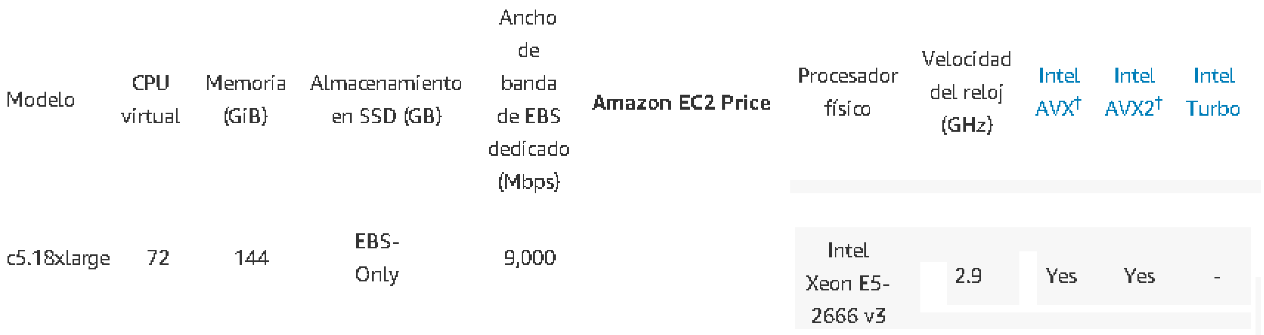
High performance web servers, scientific modelling, batch processing, distributed analytics, high-performance computing (HPC), machine/deep learning inference, ad serving, highly scalable multiplayer gaming, and video encoding.

C4:

Use Cases

High performance front-end fleets, web-servers, batch processing, distributed analytics, high performance science and engineering applications, ad serving, MMO gaming, and video-encoding.

From this instance, we cannot select any model since those ones on sale do not meet the requirements and the just the model c5.18xlarge does. Nevertheless I will display the features (without the price, since it is not given) of it.



Modelo	CPU virtual	Memoria (GiB)	Almacenamiento en SSD (GB)	Ancho de banda de EBS dedicado (Mbps)	Amazon EC2 Price	Procesador físico	Velocidad del reloj (GHz)	Intel AVX†	Intel AVX2†	Intel Turbo
c5.18xlarge	72	144	EBS-Only	9,000		Intel Xeon E5-2666 v3	2.9	Yes	Yes	-

- Accelerated computing

P3:

P3 instances are the latest generation of general purpose GPU instances.

Use Cases:

Machine/Deep learning, high performance computing, computational fluid dynamics, computational finance, seismic analysis, speech recognition, autonomous vehicles, drug discovery.

P2:

P2 instances are intended for general-purpose GPU compute applications.

Use Cases:

Machine learning, high performance databases, computational fluid dynamics, computational finance, seismic analysis, molecular modeling, genomics, rendering, and other server-side GPU compute workloads.

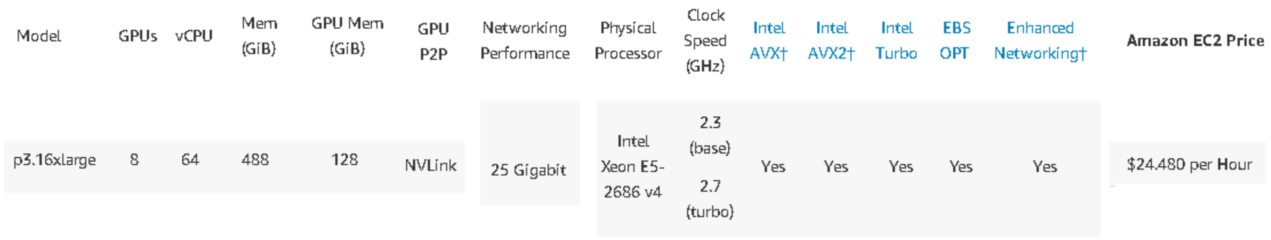
F1:

F1 instances offer customizable hardware acceleration with field programmable gate arrays (FPGAs).

Use Cases:

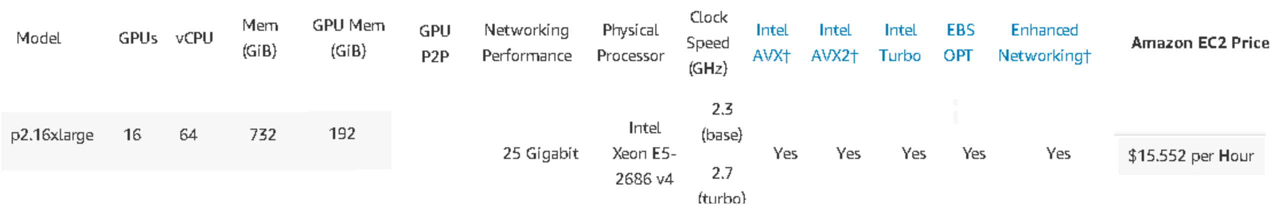
Genomics research, financial analytics, real-time video processing, big data search and analysis, and security.

– P3:



Model	GPUs	vCPU	Mem (GiB)	GPU Mem (GiB)	GPU P2P	Networking Performance	Physical Processor	Clock Speed (GHz)	Intel AVX†	Intel AVX2†	Intel Turbo	EBS OPT	Enhanced Networking†	Amazon EC2 Price
p3.16xlarge	8	64	488	128	NVLink	25 Gigabit	Intel Xeon E5-2686 v4	2.3 (base) 2.7 (turbo)	Yes	Yes	Yes	Yes	Yes	$24.480 per Hour

– P2:



Model	GPUs	vCPU	Mem (GiB)	GPU Mem (GiB)	GPU P2P	Networking Performance	Physical Processor	Clock Speed (GHz)	Intel AVX†	Intel AVX2†	Intel Turbo	EBS OPT	Enhanced Networking†	Amazon EC2 Price
p2.16xlarge	16	64	732	192		25 Gigabit	Intel Xeon E5-2686 v4	2.3 (base) 2.7 (turbo)	Yes	Yes	Yes	Yes	Yes	$15.552 per Hour

- Memory optimized

X1e:

Use for memory intensive enterprise applications, such as high-performance databases or in-memory databases.

X1:

They are optimized for large-scale, enterprise-class and in-memory application.

Use Cases:
In-memory databases (e.g. SAP HANA), big data processing engines (e.g. Apache Spark or Presto), high performance computing (HPC). Certified by SAP to run Business Warehouse on HANA

R4:

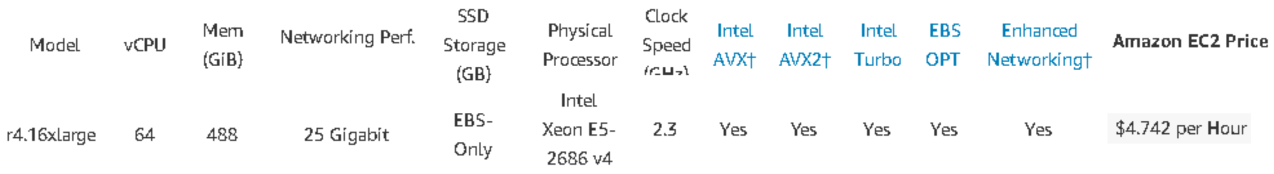
R4 instances are optimized for memory-intensive applications.

Use Cases:

High performance databases, data mining & analysis, in-memory databases, distributed web scale in-memory caches, applications performing real-time processing of unstructured big data, etc.

Just R4 instances are on sale, and, in this group, the unique model that matches the requirements is r4.16xlarge, whose features and price are exposed below:

– R4:



Model	vCPU	Mem (GiB)	Networking Perf.	SSD Storage (GB)	Physical Processor	Clock Speed (GHz)	Intel AVX†	Intel AVX2†	Intel Turbo	EBS OPT	Enhanced Networking†	Amazon EC2 Price
r4.16xlarge	64	488	25 Gigabit	EBS-Only	Intel Xeon E5-2686 v4	2.3	Yes	Yes	Yes	Yes	Yes	$4.742 per Hour

- Storage optimized

H1:

H1 instances feature up to 16 TB of HDD-based local storage.

Use Cases:
MapReduce-based workloads, distributed file systems such as HDFS and MapR-FS, network file systems...

I3:

This instance family provides Non-Volatile Memory Express (NVMe) SSD-backed Instance storage optimized for low latency and very high random I/O performance.

Use Cases:

NoSQL databases (e.g. Cassandra, MongoDB, Redis), in-memory databases (e.g. Aerospike), scale-out transactional databases,

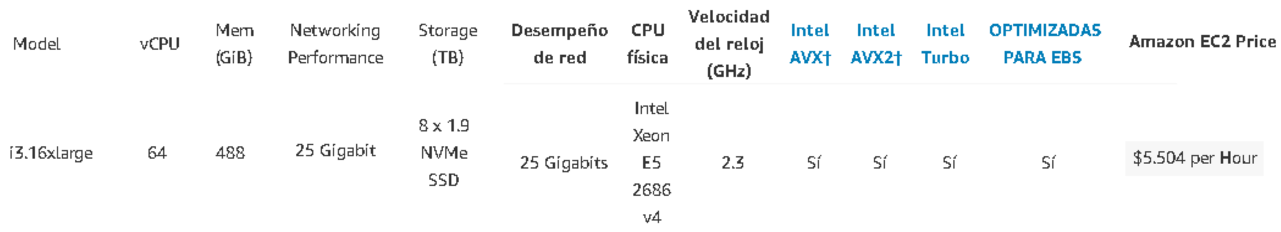
D2:

D2 instances feature up to 48 TB of HDD-based local storage.

Use Cases:
Massively Parallel Processing (MPP) data warehousing, MapReduce and Hadoop distributed computing...

Just I3 and D2 instances are on sale, and, in these group, the unique model that matches the requirements is i3.16xlarge, whose features and price are exposed below:

– I3:



Model	vCPU	Mem (GiB)	Networking Performance	Storage (TB)	Desempeño de red	CPU física	Velocidad del reloj (GHz)	Intel AVX†	Intel AVX2†	Intel Turbo	OPTIMIZADAS PARA EBS	Amazon EC2 Price
i3.16xlarge	64	488	25 Gigabit	8 x 1.9 NVMe SSD	25 Gigabits	Intel Xeon E5 2686 v4	2.3	Sí	Sí	Sí	Sí	$5.504 per Hour

The next table summarizes all the models exposed, showing the prices to make comparisons among them (according to the Amazon calculator, for an hour per month and 50 cores)

Model	**Price($/hour)**	**vCPU**	**Memory(GiB)**	**Clock Speed (GHz)**
m4.16xlarge	**191,50**	64	256	2.3
p3.16xlarge	**1.361,25**	64	488	2.7 (turbo)
p2.16xlarge	**791,50**	64	732	2.7 (turbo)
r4.16xlarge	**251,00**	64	448	2.3
i3.16xlarge	**289,00**	64	448	2.3

To these previous prices it is necessary to add the prices of hiring Amazon EBS (Amazon Elastic Block Store) which provides persistent block storage volumes for use with Amazon EC2 instances in the AWS Cloud.

By choosing the modality “Amazon EBS General Purpose SSD (gp2) volumes” (the cheaper one for SSD storage), the price for the model m4.16xlarge would be increased in 0.16$ per hour per month, giving a total of **191,66 $** per hour.

APACHE HADOOP IN MICROSOFT AZURE

Link: https://azure.microsoft.com/en-us/services/hdinsight/

The product from Azure that provides the cloud service is HDInsight, which makes it easy, fast, and cost-effective to process massive amounts of data.
It allows to use open-source frameworks such as Hadoop, Spark, Hive, LLAP, Kafka, Storm, R & more.

In this case, similarly to Amazon, we can choose the region where the product will be delivered.

There are two main options: Memory Optimized nodes and General Purpose nodes.

However, unfortunately, none of those two options provides an enough number of cores (the highest one is 16) for our application, therefore Microsoft Azure is an option that must be rejected.

HAPACHE HADOOP EN GOOGLE CLOUD DATAPROC

Link: https://cloud.google.com/dataproc/pricing?hl=es-419

1. Advantages of Hadoop on Google Cloud Dataproc

Google Cloud Dataproc provides a service of Apache Hadoop, Apache Spark, Apache Pigy and Apache Hive to process with no effort large data sets at low cost. The user can create clusters of any size and deactivate them for economical reasons.

As remarked features of this products, we have:

- Fast and reliable data processing:

The size of the clusters can be fast changed, from three nodes to hundreds of them. Additionally, each action of cluster takes less than 90 seconds on average.

- Affordable prices:

Prices structure at low cost very easy to understand, based on the real usage (measured by minute).

-Open source ecosystem:

The nearest location to Spain is London, so the prices are shown in relation with that city:

2. Modalities offered in Cloud Dataproc

In the web page we can retrieve the next models:

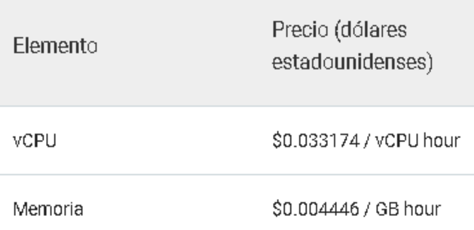
– Standard machines: type n1-standard-X

– High memory capacity machines: these models have 6,5 GB of RAM per virtual core. These instances are ideal for tasks which require higher memory capacity with respect to the virtual CPU.
They are type n1-highmem-X

– High CPU capacity machines: these models have 0.9 GB of RAM per virtual core. These instances are ideal for tasks which require higher virtual CPU capacity than capacity memory.
They are type n1-highcpu-X

– Customized machines: it is another interesting option, where the user can create a customized machine with a determined number of vCPU and memory.

The amount charged in Cloud Dataproc for the type of customized machine depends on the total number of vCPUs in each node.



Elemento	Precio (dólares estadounidenses)
vCPU	$0.033174 / vCPU hour
Memoria	$0.004446 / GB hour

3. Prices

The models shown below and their respective prices correspond only to those ones which meet the conditions.

The prices shown below have been calculated with the calculator hosted in the web page (London):

Model	vCPU	Memory(GB)	Splitted prices ($/hour)	Total prices ($/hour)
n1-standard-64	64	240	1 master core – 3,92 49 worker cores- 191,92 Cloud Dataproc – 0,5	**196,34**
n1-standard-96	96	360	1 master core – 5,88 49 worker cores- 287,88 Cloud Dataproc – 0,5	**294,26**
n1-highmem-64	64	416	1 master core – 4,87 49 worker cores- 238,81 Cloud Dataproc – 0,5	**244,18**
n1-highmem-96	96	624	1 master core – 7,31 49 worker cores- 358.21 Cloud Dataproc – 0,5	**366,02**

Apart from the models, it is necessary to add the price of the storage. For 10 TB, this price is 2.088,9 $ for SSD Storage and 491,52 $ for HDD Storage.
But as opposed to Amazon, the price does not compute for hours, so it can be an interesting option for times greater than 3072 hours (for HDD Storage), that is, a little more than 4 months of work.

But for lower work-times, Amazon product is more affordable.

Most affordable option provided by Google is n1-standard-64, with a price of 196,34 $/hour. (omitting the price of the SSD or HDD storage).

APACHE HADOOP IN CLOUD PLATFORM

LinK: https://cloudplatform.sap.com/dmp/capabilities/us/product/SAP-Cloud-Platform-Big-Data-Services/b62d6db5-ffe8-4587-b5fb-8e6fd97fbc55,

A full-service Big Data cloud, based on Hadoop and Spark, that meets rigorous demands for reliability, scalability, and security.

The features we can retrieve in the web page are:

- Secure and Ready Solutions
- Operations Excellence Included: A Big Data operations team comes with every subscription.
- Automatic Elasticity :Management of data without worrying about capacity,
- Automatically Updated for You
- Foster Digital Innovation

This provider have different modalities depending of the addressee, and the most basic model we have is Big Data Service Starter Edition, with a storage capacity of 20 TB.
The price of this service is 6.075 $, but the fee is fixed per month, not per hour like Amazon or Google services.

Once more, the selection of the model depends on the time that the application will require, since it is more profitable to buy this last service (Cloud Platform) if the number of hours is greater than 20 hours per month (more or less) than services provided by Amazon or Google.

Nevertheless, in the last section I will expose my own conclusions that I have been able to draw up, describing in my opinion which are the advantages and disadvantages of each cloud computing provider and when it is most affordable depending on the time that we need for our application.

APACHE HADOOP IN ORACLE CLOUD

Link: https://cloud.oracle.com/en_US/big-data-cloud

Oracle Cloud provides several big data services and deployment models, among them : Oracle Big Data Cloud, which is built as a big data cloud platform for enterprises to run big data workloads. As described in the web page, Oracle Big Data Cloud is offered as a managed, secure, performant, elastic and integrated platform cloud service on Oracle public cloud.

The features this service can provide are described in the page, and they are:

- Auto-Healing: Failing cluster components are automatically fixed without any human intervention.

- Encryption Support to protect all your data.

- Smart Data Movement: Intelligent connectors to database and cloud storage move only the needed data

- Platform for All Data: Integrated with other Cloud services such as Event Hub (Streaming Data) aor Database (Relational Data).

Products offered are:

– Oracle Big Data Cloud Service - Compute Edition - Compute Capacity: $0.242 per OCPU and hour.

– Oracle Big Data Cloud Service - Compute Edition - High Performance Storage Capacity: $0.1412 Gigabyte Storage Capacity Per Month

If we want at least 50 cores and a minimun storage capacity of 10TB, the price amounts to **156,69 $.**

OTHER CLOUD COMPUTING PROVIDERS

In the following web pages, extensive lists of possible provider are exposed.

1) → https://www.kdnuggets.com/2015/04/hadoop-as-service-18-cloud-options.html
2) → https://www.technavio.com/blog/top-16-companies-in-the-hadoop-as-a-service-hdaas-market

However, excluding the 4 providers described in this task, I have found for the rest of providers have too much cumbersome and complexity to find the configuration looked for.
Other providers are based on these 4 described above.

COMPARISON AMONG STUDIED MODELS

Model	vCPU	Memory(GB)	Total prices ($/hour)
AMAZON EC2			
m4.16xlarge	64	256	**191,50**
p3.16xlarge	64	488	**1.361,25**
p2.16xlarge	64	732	**791,50**
r4.16xlarge	64	448	**251,00**
i3.16xlarge	64	448	**289,00**
GOOGLE CLOUD DATAPROC			
n1-standard-64	64	240	**196,34**
n1-standard-96	96	360	**294,26**
n1-highmem-64	64	416	**244,18**
n1-highmem-96	96	624	**366,02**
ORACLE CLOUD			
Compute Edition	50	100	**156,69**

Looking at the summarazing table, I can highlight the following notes:

- The best model offered is p3.16xlarge, whose perfomance and features are superiors to the rest of models, independently of the provider, even for our purpose (memory almost five times greater than the one we need, enough vCores and even this model includes gCPU and clock speed higher than the average, 2.7 Ghz).

  Despite the previous information, this model is at the same time the most expensive, just for works and applications that require a great perfomance (this is not the case).

- Modalities of pricing are different depending on the provider, as described in previous sections.
  Although Google and Cloud Platform offers models which meets enough capacity for our application, the payment is montly in respect to the purchase of the storage , resulting in an higher price that by the payment of a model from Amazon (where the vCPU and storage is bought per hour, not monthly).

  Therefore, for long execution and working times, Google and Cloud Platform may be an interesting option, but in our case I think is most affordable to select the models from Amazon or Oracle Cloud.

- Having discarded the providers Google and Cloud Platform, the last decision must be made between Amazon and Oracle.
  The cheaper option is offered by Oracle, but the features and parameters are most limited.

Therefore I could conclude that the best option for our application is the **model m4.16xlarge from Amazon AWS**, as it provides enough capacity and perfomance to develop and run the application without any problem.

It is important to remark that the chosen option (m4.16xlarge) offers as type of storage “EBS General Purpose SSD Volumes “, that is, it provides an SSD device, which is normally most expensive that HDD.