Mariana Ávalos
Cristina Vázquez
Susana Jaramillo
Juan Carlos Medina
Marcelo Álvarez



# Análisis Cualitativo de riesgos

El nivel de apetito por el riesgo de este proyecto se clasifica como Pareto Risk, ya que estamos dispuestos a aceptar riesgos con una alta razón recompensa-riesgo, donde de ser necesario, se tomarán riesgos al contemplar su recompensa. Este nivel de riesgo hace que **la matriz de probabilidad e impacto** tenga un número equitativo de celdas en cada color verde, amarillo, naranja y rojo. Si clasificamos los riesgos identificados en el Risk Registry (16), podemos visualizar su severidad (multiplicación de probabilidad x impacto) en el siguiente diagrama, que muestra los riesgos con su ID (identificación de riesgos describe cada ID).

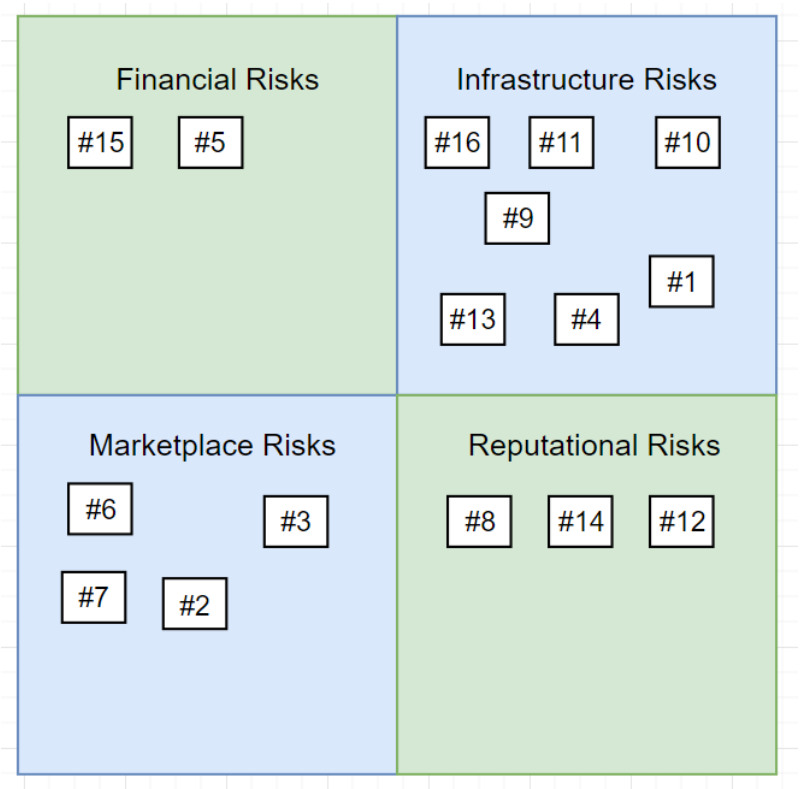
| Probability | Riesgo | | | |
|---|---|---|---|---|
| Probable (4) | Moderate (4) | Major (8) | #11 Severe (12) | #16 Severe (16) |
| Possible (3) | #2 #3 Minor (3) | Moderate (6) | Major (9) #6 | #10 Severe (12) |
| Unlikely (2) | #1 Minor (2) | #15 Moderate (4) #8 | Moderate (6) | Major (8) |
| Rare (1) | #4 Minor (1) | #9 Minor (2) | #13 #14 Minor(3) #5 #12 | #7 Moderate (4) |
| Impact | Low (1) | Medium (2) | High (3) | Very high (4) |

De esta manera, se identifican 3 riesgos con alta razón de riesgo-recompensa: riesgos con ID #11, #16 y #10. Si recordamos a qué riesgos corresponden dichos IDs sabemos que son:

- Riesgo #11: Almacenamiento excedido en el clúster de datos
- Riesgo #16: Ataque cibernético debido a vulnerabilidad en clúster de datos
- Riesgo #10: Sistema resultante se ejecuta demasiado lento por falta de optimización de algoritmos

Por lo tanto, **dichos riesgos serán el objeto de análisis cuantitativo** en las siguientes secciones, ya que según el apetito por el riesgo, al tratarse de riesgos con alto nivel de recompensa estamos comprometidos a atajarlos debidamente y a enfocarnos en ellos por encima de los restantes al ser los de alto impacto y probabilidad. Asimismo, los **tres riesgos más importantes** o severos del proyecto se clasifican como riesgos de **infraestructura** siguiendo el modelo FIRM del Risk Management Institute. Se utiliza este modelo para tener una identificación y preparación al riesgo lo más completa posible. Se busca que lo estratégico, táctico y operativo sea involucrado en cada uno de los tres riesgos que se analizan a continuación de forma cuantitativa. Al ser riesgos de infraestructura, y según el PMI, de tecnología interna, su análisis es meramente cuantitativo, a excepción de la clasificación utilizando la matriz de probabilidad e impacto y FIRM.

Siguiendo con el análisis cualitativo de los riesgos, se presenta la clasificación de los riesgos (con su ID, donde la sección de identificación de riesgos describe cada ID) por su naturaleza según el modelo FIRM del IRM, para asegurar una identificación completa de riesgos.



Mariana Ávalos
Cristina Vázquez
Susana Jaramillo
Juan Carlos Medina
Marcelo Álvarez

# Análisis Riesgo #11 - Simulación Montecarlo

El riesgo #11 de **almacenamiento excedido** en el clúster de datos se daría debido a una **estimación errónea** acerca de la cantidad de memoria en la nube que un usuario promedio tendría en el sistema. Por ello, es preciso realizar una simulación de Montecarlo para modelar la probabilidad de que los usuarios se excedan en la cantidad de memoria destinada para ellos. Para ello, lo primero que hay que estimar es cuál es la cantidad promedio que usaría una persona en un servicio cloud, lo cual se obtiene después de simular, con una distribución de probabilidad, la cantidad que N usuarios utilizarían del servicio.

La cantidad de almacenamiento digital requerido por una persona no parece seguir una distribución normal, ya que hay muchos factores que pueden influenciar las necesidades de almacenamiento y las preferencias de éste. Por ejemplo, pueden existir usuarios que utilizan almacenamiento en la nube de forma básica, mientras que otros usuarios dependen más de grandes archivos. Además, el uso de cloud está influenciado por muchos factores externos como avance tecnológico, cambios en los formatos de almacenamiento, y procesamiento.

Por lo tanto, la distribución de probabilidad del almacenamiento digital estará muy probablemente sesgada y será no-normal, con una cola larga a la derecha. Esto por la suposición de que la gran mayoría de la población sólo necesitará una porción de almacenamiento pequeña, mientras que el subconjunto más pequeño de individuos con necesidades especializadas o intensamente enfocadas en archivos pesados podrá requerir de cantidades más grandes de almacenamiento.

Esta distribución se conoce como power law distribution, donde los casos extremos, que en este caso son los usuarios con grandes necesidades de almacenamiento, tienen un impacto considerable en la distribución. Esta distribución coincide con otro fenómeno del ámbito, que es el tráfico de redes de internet. Un estudio publicado en Journal of Information Science en 2011, titulado A power law distribution for file sizes in a large network of personal computers (Lehman et al., 2011), analiza la distribución de los tamaños de los archivos a través de una red local inalámbrica con 250 000 archivos, y encuentra que el almacenamiento requerido sigue una distribución power law. De forma similar, un artículo publicado en Proceedings of the International Conference on Management of Emergent Digital EcoSystems también en 2011 titulado Modeling User Demand for Cloud Storage: An Empirical Study (Jiang et al., 2011), examina la distribución de almacenamiento en un grupo de usuarios y encontró que también seguía una power law distribution.

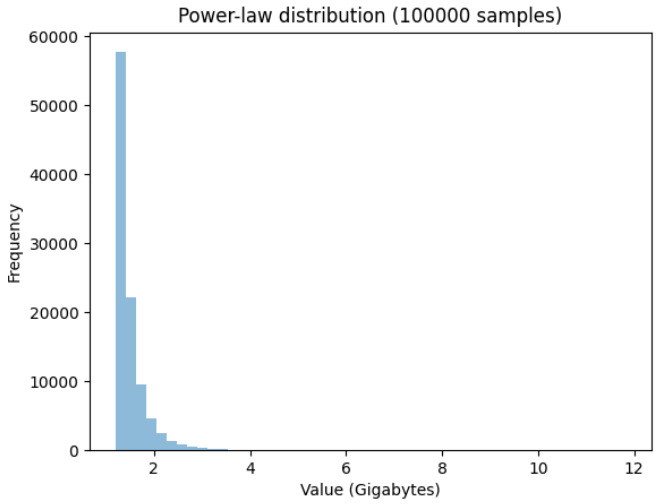
La distribución power law sigue la siguiente función:

$$f(x) = (\alpha - 1) * (x / xmin)^{(-\alpha)},$$

donde x es la variable aleatoria que representa el almacenamiento requerido de un usuario de un servicio en la nube, α es el parámetro de escala y xmin es el valor mínimo (Sarasola, 2023).

Mariana Ávalos
Cristina Vázquez
Susana Jaramillo
Juan Carlos Medina
Marcelo Álvarez

En el estudio de Jiang, se estima que estos parámetros son: α = 1.19 GB y xmin 6.16 GB (Jiang et al., 2011). Cabe resaltar que este tamaño en GB no es del tamaño de los archivos, sino de los datos en dichos archivos, medidos por bytes. Entonces, si utilizamos la herramienta de Python para crear 100,000 muestras de valores que siguen esta distribución, se obtiene el siguiente histograma de frecuencia:




Adicionalmente, se obtienen las siguientes medidas de tendencia central:

- Promedio:  1.475 GB por usuario
- Desviación estándar:  0.3648 GB de la media
- IQR: 0.298 GB de la media

A pesar de que la distribución no es normal, la desviación estándar nos habla de la variación de los datos con respecto de la media aritmética. En distribuciones no-normales, la desviación estándar resulta una medida demasiado sensible a outliers, que en caso de powerlaw, son muy relevantes. Por lo mismo, se puede calcular el IQR o rango intercuartil para saber la variación de los datos más concretamente. Para resolver la cuestión de qué probabilidad existe de que un usuario sobrepase el almacenamiento que el proyecto estima, primero tenemos que recordar el supuesto de que en el proyecto el usuario tendrá como máximo 1.5 GB de espacio aproximadamente. Por lo tanto, si en la simulación calculamos el número de datos que sobrepasan el valor de 1.5 y dividimos esta cuenta entre 100,000:

P(x > 1.5 GB) = casos donde x > 1.5 / 100,000 = 3,206 / 100,000 = 0.03206

Así, sabemos ahora que **existe una probabilidad del 3.21% de que un usuario sobrepase el espacio estimado en el servicio de la nube.** Si tomamos en cuenta que estos datos son de 2011, y tomamos el crecimiento anual promedio de almacenamiento en cloud computing de 4.5% (New York Times, 2021), obtenemos que el promedio es 2.55 GB y que existe una probabilidad del 3.21% de que se exceda. El código que se usó para este apartado se encuentra en el repositorio de la documentación del proyecto: https://github.com/the-other-mariana/pm/blob/master/PM2/risks/week6/montecarlo.py

Mariana Ávalos
Cristina Vázquez
Susana Jaramillo
Juan Carlos Medina
Marcelo Álvarez

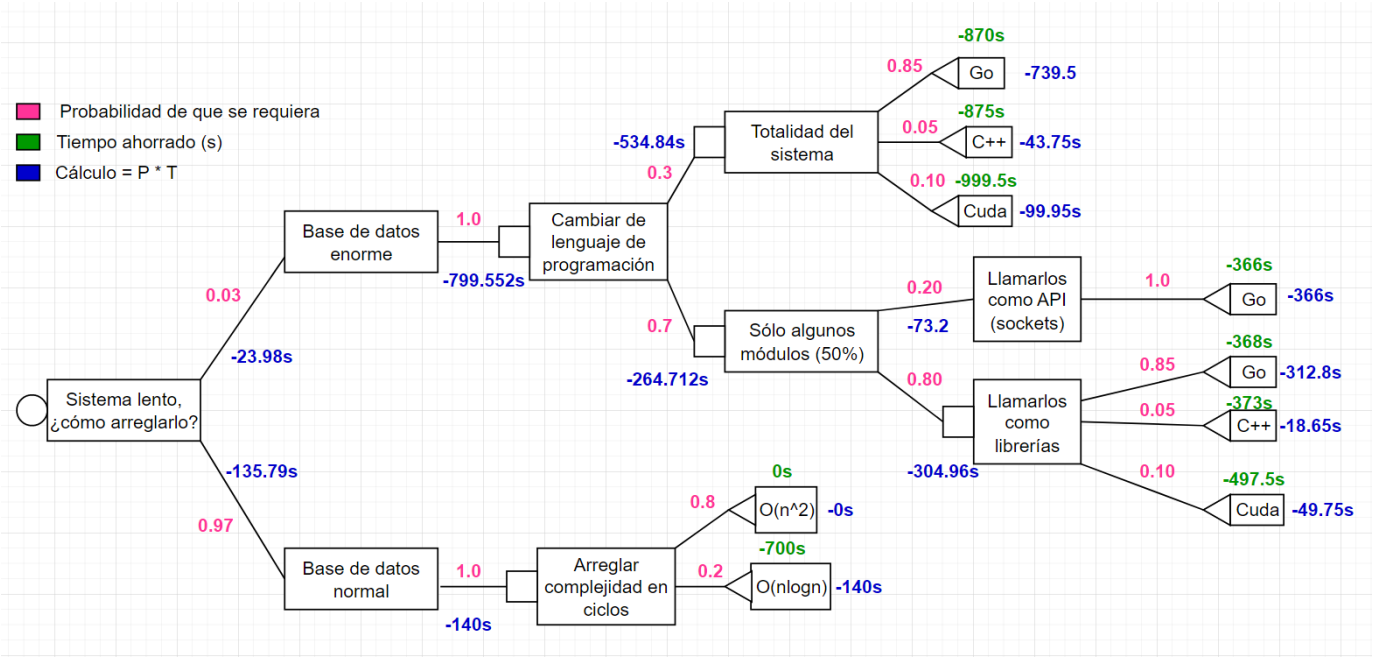
Mariana Ávalos
Cristina Vázquez
Susana Jaramillo
Juan Carlos Medina
Marcelo Álvarez

# Análisis Riesgo #10 - Árbol de decisiones

El riesgo #10 se refiere a aquel caso donde **el sistema resultante** de algún sprint termine siendo **demasiado lento** para el usuario. Este riesgo depende del feedback y de la definición nominal de “demasiado” lento. Si esto sucede, se debe analizar cuál es el camino a tomar, ya que existe una gran variedad de caminos a seguir cuando un sistema se percibe como “lento”.

Se realiza el siguiente análisis de árbol de decisión: suponiendo, para simpleza de cálculos, que el sistema realiza 1 operación en 1 segundo (s), y que un usuario requiere, en un momento dado t, 1000 operaciones, entonces el siguiente análisis muestra como “impacto” los segundos que se ahorraría en la ejecución del sistema si se compara con un sistema actual “lento” que tarda 1000s por 1000 operaciones. Es decir, cada rama muestra cada decisión y los segundos que se reducirían del sistema que tarda 1000s. Para el cálculo de tiempos en opciones paralelas en CPU (Go / C++), se asumen sistemas de 8 cores o núcleos; para opciones paralelas con GPU (CUDA), se asumen sistemas de 2000 núcleos. El “EMV” en lugar de ser de valor monetario esperado, será de valor de tiempo a reducir esperado por cada opción. Es una especie de árbol de decisión con base en los segundos ahorrados respecto al sistema “actual” de 1000s.

En el diagrama de abajo, se denota la probabilidad de que una opción se requiera en la implementación del sistema, el tiempo ahorrado con dicha opción en segundos y el cálculo del “EMV” de tiempo ahorrado probable.




Así, aunque resulte atractivo utilizar GPUs o paralelismo hoy en día, la opción que reduce con mayor probabilidad la ejecución del sistema en situaciones “normales” sería la

alternativa de **Arreglar complejidad en ciclos**, y buscar la opción de métodos con **complejidad O(nlogn)**. Sin embargo, como se vio en el análisis de Montecarlo, existe un 3.21% de probabilidad de que la base de datos requiere almacenamiento fuera de lo estimado. Si se presenta este caso, la opción que mejoraría los tiempos de ejecución sería cambiar el lenguaje de la **totalidad** del sistema, ya que cambiar algunas partes a otros lenguajes requiere de comunicación por internet (sockets) o tiempo de espera en ejecuciones de librerías externas.

Sin embargo, se sugiere un análisis previo de tiempos para desarrollar dichas alternativas, ya que cambiar el lenguaje del sistema podría alargar la fecha de entrega, mientras que cambiar sólo unas partes resulta en la reducción del 50% de la reducción que ofrece el cambio de lenguaje en su totalidad.

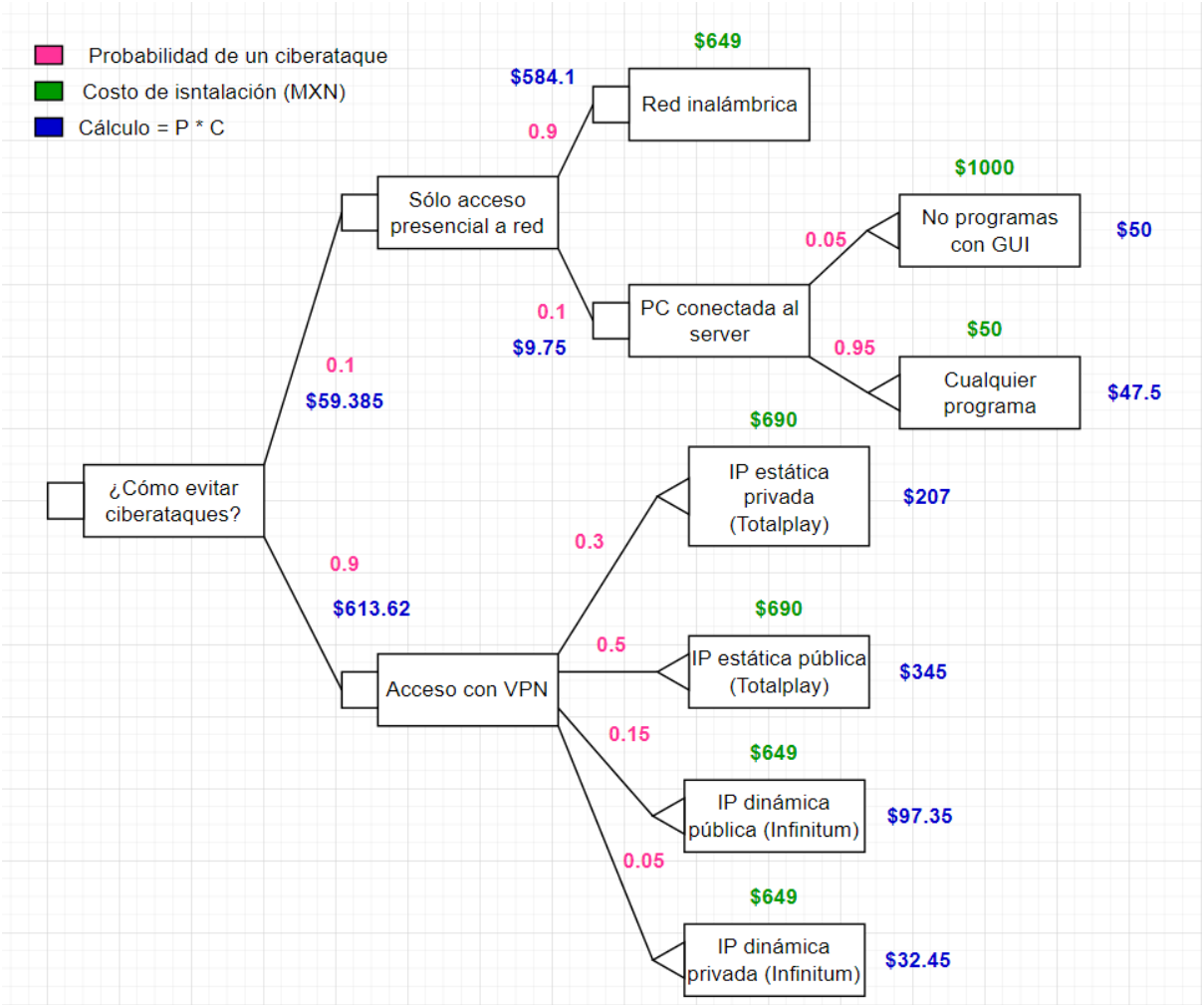
Este análisis también sirve de guía no sólo en contingencia o reducción de impacto del riesgo #10, sino también para saber cuál alternativa tomar para la reducción de tiempo de ejecución **antes de desarrollar el sistema**, es decir, mitigación de probabilidad de que el riesgo #10 se presente. Con esto, sabemos de antemano que debemos buscar orientar las pruebas de calidad para mantener una complejidad del sistema no mayor a O(nlogn) para asegurar que el sistema no resulte “lento”.

Mariana Ávalos
Cristina Vázquez
Susana Jaramillo
Juan Carlos Medina
Marcelo Álvarez

# Análisis de riesgo #16 - Árbol de decisiones

El riesgo #16 se refiere a la **posibilidad de sufrir un ataque cibernético**, por lo que se realizará un árbol de decisiones para identificar el camino que se debe seguir para prevenir (reducción de probabilidad) un ataque cibernético por vulnerabilidades del sistema.

Se establece en cada rama la probabilidad o vulnerabilidad a un ciberataque dada dicha opción, y en cada nodo adicionalmente se establece el costo de dicha opción. De esta manera, al tener una opción costosa pero que reduce la probabilidad de un ataque a un porcentaje pequeño, entonces se "reduce" su costo al multiplicarse. Así, la opción a seguir será la de menor costo ponderado.




Podemos concluir que la opción que presenta menor "costo" ponderado es la de permitir acceso a la Base de datos solamente a través de acceso presencial, y únicamente a través de una PC que está conectada permanentemente al servidor, y que restringe cualquier instalación de programa que utilice una GUI (Graphical User Interface), ya que con esto obliga a realizar cualquier actividad a través de la terminal únicamente. Esta alternativa reduce el costo y la probabilidad de un ciberataque. Por lo tanto, resulta ser el camino más conveniente para evitar un ciberataque por vulnerabilidades de red, es decir, reducir la probabilidad del riesgo.

Mariana Ávalos
Cristina Vázquez
Susana Jaramillo
Juan Carlos Medina
Marcelo Álvarez

# Referencias

(Lehman et al., 2011) Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2011). A power law distribution for file sizes in a large network of personal computers. Journal of Information Science, 37(3), 263-270.

(Jiang et al., 2011) Jiang, W., Li, X., & Wu, J. (2011, Septiembre). Modeling user demand for cloud storage: An empirical study. In Proceedings of the International Conference on Management of Emergent Digital EcoSystems (pp. 93-98). ACM.

(Totalplay, 2023) Totalplay. Promociones. Consultado 15/03/2023 desde: https://www.totalplay.com.mx/landings/promociones

(Infinitum, 2023) Infinitum. Internet Empresarial. Consultado 15/03/2023 desde: https://telmex.com/web/guest/mkt/hogar?pqi=PQI64&gclid=Cj0KCQjw2cWgBhDYARIsALggUhptbuAypn_V-iUkc1cdRYF8p080snYhmh2DOZGfxcUWM7zNIy2tWNIaAidFEALw_wcB&gclsrc=aw.ds

(Sarasola, 2023) Sarasola, J. Power Law Distributions: Statistics for business. Consultado 15/03/2023 desde: https://gizapedia.org/static/786bb254b9f75741600fc9ca31fcde81/english_beamer_powerlaw.pdf

(New York Times, 2021) New York Times. (Enero, 2021). Microsoft’s Profits Rise by 33%, Driven by Cloud Computing and Xbox. Consultado 21/03/2023 desde: https://www.nytimes.com/live/2021/01/26/business/us-economy-coronavirus