UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Estudios Superiores Acatlán

Pronosticar el crecimiento de una base de datos por medio de un modelo matemático para un servidor Oracle Exadata 11g R2 con la metodología Box-Jenkins.

TESIS

QUE PARA OBTENER EL TÍTUO DE:

Licenciado en Matemáticas Aplicadas y Computación

PRESENTA:

Vladimir Giles Carmona

ASESOR:

Lic. Christian Carlos Delgado Elizondo

México, D.F. Noviembre de 2015

Resumen

En lo que va de la revolución en las tecnologías de la información que se inicia aproximadamente en la década de los 60 del siglo pasado cuando los transistores y los componentes de almacenamiento comenzaron una competencia por ser más eficientes en su procesamiento de datos y en el almacenamiento de los mismos, comenzó a ser cada vez más difícil tener ambientes similares para simular y someter a ciertos controles de posibles escenarios para probar con variantes en cargas de datos, procesamiento o modificación de parámetros y poder experimentar con diferentes umbrales de espacios en discos de almacenamiento y tiempos a los que es necesario someter a pruebas simuladas con la finalidad de tener pronósticos para mantener un control en el futuro.

Las empresas en nuestro país se ven limitadas en recursos para poder tener un ambiente de producción y un ambiente de pruebas en paralelo debido a los elevados costos en hardware y licencias, sean para uso en aplicaciones usadas en áreas de administración del negocio o para otras que tenga demasiada dependencia con el hardware y que además es muy difícil su instalación en alguna máquina virtual como las que existen actualmente en el mercado: virtual box, vmware, etc. Que son muy útiles pero que no pueden ser usadas como es el caso de Exadata.

En el caso de los servidores Oracle Exadata, su costo es exageradamente elevado y es por ello que surge la necesidad de hacer un modelo que nos permita pronosticar el crecimiento y presentar la cantidad más óptima para agregar discos a una base de datos que en exceso serían componentes subutilizados con los cuales se podría ahorrar recursos que se debieran aprovechar en otras necesidades más importantes o por el contrario si se sobrepasara el umbral o límite de discos disponibles se corre el riesgo de gastar una cantidad de dinero inesperado debido a un desbordamiento derivado de una mala estimación del espacio.

En este trabajo se usó una serie de tiempo que comprende del año 1984 al 2013 tomando una observación mensual que se estimó en unidades de megabytes por mes, estos acontecimientos consecutivos se registraron en la base de datos con un trigger o lanzador en automático y se fueron almacenando en el área de auditoría de sistemas de Telcel que estuvieron en funcionamiento hasta el año 2007. Otra parte se sacó desde el año del 2007 al 2013 de manera manual y con ayuda de las tablas que almacenan información histórica.

El trabajo se dividió en cuatro capítulos, el primero presenta el planteamiento de investigación en la que se muestra la problemática, la hipótesis, los objetivos y la metodología que es el factor indispensable para poder desarrollar la investigación debido a que en ella se pudo hacer la

estimación de la gráfica de series de tiempo y posteriormente el pronóstico para llegar a los resultados de la investigación.

En el capítulo dos se hace una reseña histórica de las bases de datos principalmente de la empresa que más ha creado tecnología asociada al modelo relacional, se presentan las características del sistema manejador de bases de datos relacional, posteriormente se describe específicamente la redundancia en el almacenamiento y el concepto de ASM que es una manera de generar espejeo para el almacenamiento en función de los discos y la instancia que los administra, finalmente se describe de manera general el almacenamiento en Exadata que es la tecnología a la que se van a migrar los datos y a la cual no se tiene acceso.

En el capítulo tres se presenta la teoría general de distribuciones en los sistemas de colas, los conceptos de procesos y optimización así como los modelos AR, MA y los modelos autorregresivos integrados y de medias móviles ARIMA que son los que usa la metodología de Box-Jenkins, finalmente se da una explicación y breve introducción del software GRETL.

En el capítulo cuatro, desarrollo de la investigación, se encuentra la aproximación de la serie y el pronóstico con el software anteriormente comentado, también se presentan los resultados de los datos y las conclusiones que serán de mucha ayuda para mostrarlas al cliente para que se tome la recomendación de la compra de discos mencionados al final. Se usó la norma APA para las citas y la bibliografía.

Para las citas de referencias en inglés y español se usó letra cursiva. Las gráficas están numeradas en función de cada capítulo consecutivo.

Se presenta un glosario con la finalidad de definir los conceptos del trabajo a detalle y una bibliografía que está en inglés y español respectivamente.

Agradecimientos

A la Facultad de Estudios Superiores Acatlán por ser una institución que me permitió durante todos estos años adquirir conocimientos y obtener una formación universitaria tan privilegiada como la vida misma que nos da la cualidad de comprender y existir.

También a todo el personal que labora en el programa de Matemáticas Aplicadas y Computación por tener la virtud de servir y formar a los alumnos y tesistas, en especial a la Mtra. Jeanett López García, jefa del programa, a la Mtra. Georgina Eslava García Jefatura de Sección de Informática, por haberme guiado en el planteamiento de investigación, al Lic. Christian Carlos Delgado Elizondo jefe de la sección de probabilidad y estadística e investigación de operaciones por toda la atención brindada en cada detalle y por compartirme su experiencia en el trabajo de investigación, al Lic. Francisco Javier López Rodríguez secretario técnico del programa, por su apoyo en todos los trámites administrativos. A todos ellos gracias por haber formado parte de este trabajo.

A los sinodales que me apoyaron en todo momento Dra. Ma del Carmen González Videgaray, Mtra. Sara Camacho Cancino, Mtra. Marisol Velázquez Salazar, Mtro Gabriel Delgado Juárez a todos ellos mis mas sinceros agradecimientos por sus comentarios y experiencia brindada durante el periodo de revisión y el examen profesional.

Al Mtro. Alejandro Rodríguez Trejo, al Mtro. Oswaldo Palma Coca, al Lic. Aldo Leonel Prado Núñez y al Ing. Oscar Rentería Sebastián por compartirme bibliografía y experiencias en sus trabajos de investigación para grado de licenciatura asi como ayuda técnica, a todos los maestros que me fueron asignados durante la licenciatura de los cuales seria imposible mencionar a cada uno de ellos, a todos gracias por su entrega y labor cotidiana en la tarea de formar profesionistas. También agradecer a todos mis compañeros del proyecto en Telcel por sus facilidad otorgada y sus experiencias para el trabajo, por su amistad y su aprendizaje gracias a todos.

Doy también gracias a mis padres por darme la vida y por mostrarme el camino del trabajo y la educación, a mis hermanos y mi familia en general por estar al pendiente durante todo este tiempo, a muchos de mis amigos de la licenciatura y de otras licenciaturas de la FES Acatlán así como a todos los que forman parte de mis amistades mas apreciadas. A todos ellos dedico este trabajo de formación profesional y de experiencia de vida, por estar en todos esos momentos que no regresarán pero que fueron parte fundamental de la vida profesional, académica y de momentos entrañables.

Capítulos

Índice de figuras y tablas

Capítulo I Planteamiento de investigación

Introducción

En este capítulo se hace una descripción del origen de la investigación así como de la hipótesis y los objetivos planteados para la realización del proyecto, se explica también el origen de la elección del tema y finalmente la metodología empleada.

Capítulo I Planteamiento de investigación

Introducción

1.1 Problemática

El siguiente trabajo tiene su origen en la puesta en práctica de la instalación de bases de datos Oracle 11gR2. Desde el requerimiento por parte del cliente pasando por la investigación técnica para instalar el RDBMS relational database management system (sistema manejador de base de datos relacional) y tomando en cuenta una estimación aproximada del crecimiento con la finalidad de no desperdiciar espacio; pero tampoco solicitar al área de storage (área encargada de discos) espacio insuficiente ya que en ambos casos se generan pérdidas económicas y en tiempos, tanto para el proveedor como para el cliente.

La fase de instalación del RDBMS corresponde al seguimiento planificado y documentación técnica de la empresa de software Oracle y podríamos tomarla como algo que ya está probado dentro de las diferentes arquitecturas.

El siguiente paso es instalar la base de datos Oracle (almacén de datos) en el cual se solicita hacer un dimensionamiento de los espacios que se deberán de utilizar.

Por servidor de base de datos Oracle entendemos un conjunto de procesos de segundo plano y estructuras de memoria que administran a un conjunto de archivos donde se almacenan los datos que con el tiempo van creciendo principalmente los archivos llamados datafiles o archivos de datos los demás se mencionarán más adelante aunque son de menor crecimiento como pueden ser los de password los de archivamiento de cambios y los de control entre otros. (Alapati, 2009 p. 189)

Lo anterior nos llevó a encontrar una manera de estimar el crecimiento con métodos estadísticos para poder saber a tres años como será el crecimiento, lo anterior debido a que el cliente así lo pide habitualmente. La tarea siguiente fue revisar métodos de series de tiempos y programas de computación que facilitan el cálculo. Una tarea también fue considerar que varios paquetes de estimaciones estadísticas no fueran muy costosos o que fueran accesibles para poder usarlos en

futuras estimaciones no solo en este trabajo sino también para los alumnos de la licenciatura en Matemáticas Aplicadas y Computación, ya que es el día a día de sus actividades y que además tuvieran algunas mejoras. Así que finalmente después de revisar varios programas se decidió usar el paquete estadístico de distribución libre GNU conocido como GRETL.

También se encontró un paquete de simulación que no es de distribución libre pero que para uso académico se puede obtener la versión básica denominado ARENA. Este paquete puede hacer simulaciones básicas y estimar el tiempo de llegada en líneas de espera, se optó no usarlo debido a que el problema era de estimación de series de tiempo.

La finalidad de hacer este trabajo fue que ya se tiene conocimiento de las bases de datos Oracle y queda por investigar más a detalle la parte de los procesos estocásticos y los procesos ARIMA específicamente el método de Box-Jenkins.

El trabajo tiene la finalidad a futuro de implementar una metodología de estimación estadística en los proyectos de implementación de bases de datos en instituciones públicas, privadas y academias de educación superior.

Se hace uso también de las herramientas matemáticas y computacionales para optimizar tiempo y dinero en este tipo de proyectos solicitados por el cliente, en este caso es Telcel, empresa que se dedica a prestar servicios de telecomunicaciones a nivel nacional e internacional principalmente voz y datos. Aunque su mercado se está diversificando a servicios de banca y créditos para el uso de servicios en telecomunicaciones para el usuario final dentro otras áreas del comercio electrónico móvil y el Internet en dispositivos móviles principalmente con sistema operativo Android distribución basada en Linux.

A continuación se citan varias formas de aplicar las series de tiempo y la metodología estadística que nos ocupa en este trabajo.

"*The methods to be presented in this book are designed for the purpose of analyzing series of statistical observations taken at regular intervals in time. The methods have a wide range of applications. We can cite astronomy [539], meteorology [444], seismology [491], oceanography [232], [251], communications engineering and signal processing [425], the control of continuous process plants [479], neurology and electroencephalography [151], [540], and economics [233]; and this list is by no means complete.*" (Green, 1999 p. 29)

El campo de investigación de las series de tiempo comprende una enorme área de aplicación aunque desafortunadamente algunas veces es conocido solo en estudios económicos y no es muy explotado en otros campos de conocimiento no menos importantes como es el caso del crecimiento de bases de datos en un departamento de informática.

En el siguiente trabajo se investigará la tendencia del pronóstico de crecimiento de almacenamiento de datos en un servidor poco accesible para hacer pruebas de laboratorio, lo importante del resultado es conocer el valor aproximado total al final de la estimación.

Cabe mencionar que este trabajo se busca alcanzar un pronóstico con las herramientas anteriormente mencionadas y presentar el resultado para conocer la tasa de crecimiento y específicamente el número de discos que se propondrá sugerir compre el cliente para instalar la base de datos, con la finalidad de no hacer uso de un servidor Exadata que es demasiado costoso.

1.2 Hipótesis

Conocer la tasa de crecimiento de una base de datos Oracle, permitirá establecer el tipo y número de dispositivos de almacenamiento físico que se requerirán en el futuro.

1.3 Objetivo general

Pronosticar el crecimiento de una base de datos en un servidor Oracle Exadata 11g R2 mediante un modelo de simulación con la metodología Box-Jenkins, para obtener un dimensionamiento adecuado de los espacios al asignarlos cuando se instala la base de datos.

1.4 Objetivos específicos

Con lo antes mencionado se puede describir lo siguiente:

- Estimar el espacio más óptimo para no generar pérdidas económicas cuando se implementa por primera vez tomando el volumen de almacenamiento vigente.
- Hacer una descripción a lo largo del trabajo de investigación, de las herramientas matemáticas y computacionales.
- Establecer un modelo general para poder hacer estimaciones en servidores a los que no se tiene acceso a un ambiente de pruebas debido a su alto costo como es el caso de Oracle Exadata 11g Release 2.

1.5 Metodología

Para definir la metodología se hace una descripción general de lo que a lo largo del trabajo se revisó, en primera instancia existen las ciencias formales que son en este caso el estudio de la lógica, los números y las matemáticas para llegar a la estadística matemática que es la que se usará en este trabajo; pero haciendo uso de las realizaciones que son algo muy parecido a la muestra de una población la diferencia es que las realizaciones son un conjunto de observaciones históricas en el tiempo que es imposible replicar. Por otro lado se encuentran las ciencias fácticas que requieren de la experimentación y que en gran medida también se usa en este trabajo ya que se experimenta con una serie de tiempo. Hasta aquí se describen dos tipos de ciencias y su manera de buscar el resultado, finalmente el camino para llegar a una hipótesis planteada es el método.

"…la ciencia se concibe como un constante progreso, progreso que consiste en llegar a leyes cada vez más universales, de tal manera que la representación del mundo sea cada vez más perfecta, aunque nunca llegue a ser del todo completa." (Xirau, 2011 p. 363)

"*El calificativo <<científico>> sugiere que un conocimiento es objetivo, verdadero, riguroso, bien comprobado. En cambio, lo que no es <<científico>>suele considerarse como subjetivo, como algo que depende de circunstancias cambiantes o que es poco fiable en general. Parece que todo conocimiento que se presenta con pretensiones de objetividad debería ser científico.*"(Artigas, 1999 págs. 14-15)

La metodología es el estudio o elección de un método adecuado aplicable a determinado objeto de estudio. Metodología es un concepto demasiado amplio siendo preferible usar el término método para fines generales ya que existen varias metodologías.

Para el caso de la metodología Box-Jenkins desarrollada en 1970 por George E. P. Box y Gwilym M. Jenkins de la universidad de Wisconsin-Madison y formalizada en 1976 también se le conoce como modelos Box-Jenkins, esto debido a que se parte del hecho de que la serie temporal que se trata de pronosticar es generada por un proceso estocástico cuya naturaleza está caracterizada mediante un modelo.

El objeto de estudio para esta metodología es la serie de tiempo, conocida como una serie cronológica o histórica, también definida como una sucesión de observaciones de una variable

tomando valores secuencialmente a lo largo del tiempo. También conocido en estadística como unidad experimental.

La metodología Box-Jenkins se basa en dos principios filosóficos que sustentan su funcionalidad para efectos prácticos y de investigación: parsimonia y aproximaciones sucesivas.

La parsimonia: fue propuesta en el siglo XIV por el lógico inglés Guillermo de Ockham, su idea postula que: la explicación más sencilla es, probablemente, más correcta que la más difícil y compleja, en esta idea se encuentra implícito el concepto de que la naturaleza prefiere lo simple antes que lo complejo. Posteriormente otros filósofos agregaron el concepto de pasar una navaja por la descripción del problema como una manera de rasurar lo que no es necesario, fue por esa razón que se le conoce actualmente como navaja de Ockham o principio de economía. (Xirau, 2011 págs. 186-187)

Mejoramiento iterativo: este consiste en comenzar con un modelo o diagrama en el que se van realizando modificaciones para mejorar su calidad. El objetivo de un algoritmo iterativo consiste en explorar el estado de soluciones para llegar a la resolución más óptima, resulta muy práctico este método en problemas difíciles de carácter práctico. Esta metodología es muy utilizada en los actuales sistemas de información para desarrollo de algoritmos e ingeniería de software. Tentativamente los métodos iterativos aparecen con una carta del matemático Gauss a un estudiante, en ella se proponían ecuaciones de 4x4 mediante la repetición de la solución del componente en donde el residuo era mayor. Posteriormente se estableció la teoría de los métodos estacionarios.

Young en los años 50 del siglo XX, en esa década aparece el método de gradiente conjugado y posteriormente los métodos de Cornelius Lanczos, Magnus Hestenes y Eduard Stiefel aunque por su naturaleza y el contexto de esa época se mal entendieron. No fue sino hasta la década de los 70 cuando se comprendieron mejor estos métodos los cuales sirven de mucho en la solución de ecuaciones de derivadas parciales y también en la metodología Box-Jenkins. En resumen esta metodología busca separar las observaciones predecibles de las no predecibles, para trabajar con los primeros y los segundos reducirlos a una mínima parte conocida como error o ruido blanco que se incluye dentro del modelo.

El procedimiento específico visto como procesos de Box-Jenkins se puede separar en cuatro etapas:

Identificación: Utilizando los datos se intentará plantear un modelo ARIMA (p,d,q) que sea la mejor propuesta para ser investigada, se define también si se incluye una constante, aquí se puede sugerir más de un modelo.

Estimación: se realiza inferencia sobre los parámetros condicionada a que el modelo investigado sea apropiado

Validación: generación de contrastes de diagnostico para comprobar si el modelo se ajusta a los datos, de no ser así, revelar las discrepancias del modelo para poder mejorarlo.

Predicción: obtención de pronósticos en términos probabilísticos de valores futuros de la variable, en esta etapa se intentará también evaluar la capacidad predictiva del modelo.

Las herramientas que se usan para la metodología Box-Jenkins, también llamados filtros son: la función de autocorrelacion FAC, función de autocorrelación parcial FACP el periodograma así como el periodograma integrado, esto hace referencia a herramientas de investigación y no a las de tipo computacional. (Videgaray, 2011 p. 23)

Capítulo II Contexto de bases de datos Oracle

A lo largo de este capítulo se expone la historia de las bases de datos relacionales y las características que se van agregando durante el tiempo, se hace también una descripción técnica del almacenamiento específicamente en el servidor Exadata y los procesos de crecimiento de espacio en disco duro.

Capítulo II Contexto de bases de datos Oracle

2.1 Historia y arquitectura Oracle

Las bases de datos se iniciaron históricamente con modelos de almacenamiento; como el jerárquico, el de red y el de listas invertidas, estos tres antecediendo a lo que hoy se conoce como Bases de Datos Relacionales que no solo tienen un éxito comercial por su funcionalidad sino que están sólidamente fundamentadas en modelos matemáticos con amplias posibilidades de usarse para fines didácticos pedagógicos y científicos.

El modelo relacional aparece entre 1969 y 1970, aunque comercialmente las bases de datos relacionales aparecen entre 1979 y 1980. Actualmente este modelo es el más predominante entre los diferentes desarrolladores de software de bases de datos como IBM, Microsoft y Oracle entre otros. (Date, 2003 p. 25)

Una aproximación ambigua a la definición de bases de datos o sistema relacional es: que los datos están organizados en tablas y es así como lo percibe el usuario. Y también una definición aproximada se puede entender cuando nos vamos a la historia de los manejadores de bases de datos que se origina en los sistemas de procesamiento de archivos que tenían limitantes y gracias a ellas se dio la necesidad de desarrollar el modelo de bases de datos relacionales.

A continuación se presentan las limitantes que tienen los sistemas de almacenamiento basados en archivos esto para tener un comparativo con la arquitectura relacional en cuanto a sus mejoras.

Limitaciones de los sistemas de procesamiento de archivos, estos puntos se describen a partir de la experiencia personal en el campo de trabajo:

- Redundancia e inconsistencia
- Dificultad en el acceso a la información
- Problemas de seguridad
- Problemas de Integridad
- Inadecuados métodos para recuperar datos

Los problemas antes mencionados son los que originan la necesidad de crear un sistema manejador de bases de datos más eficiente. (Osorio, 2008 p. 11)

Surgidos todos en ambientes productivos y académicos para mejorar la productividad y eficiencia en el manejo de los datos. A ellos se les ha agregado tecnología que les ayuda también a hacer más eficiente el funcionamiento. Aunque en esencia el modelo relacional optimiza la manera de seleccionar los datos.

El concepto de Bases de Datos Relacionales fue iniciado comercialmente por el Doctor Edgar F. Codd en una publicación de una Investigación de la compañía IBM denominada System R4 Relational, Publicado en 1970. Posteriormente una empresa llamada Software Development Laboratories Relational desarrollo un producto de Bases de Datos relacionales en 1977 llamado Oracle V.2, la empresa después cambio su nombre a Relational Software Incorporated (RSI) y finalmente la llamaron simplemente Oracle corporation. (Wreenwald, 2009 p. 22)

Y la razón por la que se hicieron populares en el ámbito comercial los modelos relacionales comenzó con la presentación de un modelo de Red propuesto por Charles Bachman de General Electric con una organización de registros de datos ligados entre sí. Este modelo origino lo que se conoció como CODASYL Database Task Group. Mientras tanto La División de Aviación Espacial de América del Norte e IBM desarrollaron un segundo modelo aproximado al modelo jerárquico de 1965.

El modelo Relacional se conceptualiza en el ligado a tablas bidimensionales que se organiza o contienen columnas y registros. (Wreenwald, 2009 págs. 22 - 24)

El servidor de Bases de Datos Oracle está diseñado para soportar bases de datos relacionales y su arquitectura se basa principalmente en los archivos físicos de datos y la instancia que se conforma por estructuras de memoria y procesos de segundo plano.

La evolución que han tenido las bases de datos Oracle es la siguiente de acuerdo al libro Oracle Essentials figura 2.1.1:

Año	*Innovaciones*
1977	Software Development Laboratories es fundado por Larry Ellison, Bob Miner y Ed Oates
1979	Oracle V 2: La Primera versión comercial de base de datos relacional que usa SQL
1983	Oracle V 3: Código base simple para uso de oracle en múltiples plataformas
1984	Oracle V 4: Con herramientas portables y lectura de consistencia
1986	Oracle V 5 Aparece la arquitectura: Cliente/Servidor En una base de datos relacional Oracle
1987	Herramientas CASE y 4GL
1988	Oracle Financial Applications: Creado en una base de datos relacional Oracle
1989	Oracle 6: Bloqueo de registros y respaldos en caliente
1991	Oracle Parallel Server: En Plataformas de paralelismo masivo
1993	Oracle 7: Optimizador Basado en Costos
1994	Oracle V 7.1 : Opciones de Paralelismo incluyendo carga de query´s y creación de índices
1996	Base de Datos universal Con SQL extendido vía Cartuchos, Cliente Ligero. Servidor de Aplicaciones
1997	Oracle 8: Objeto Relacional y características Very Large Database (VLDB)
1999	Oracle 8i : Java Virtual Machine (JVM) en el Servidor de Bases de Datos
2000	Oracle Database 9i : Herramientas Oracle Integradas en la capa media
2001	Oracle Database 9i: Real Application Clusters, OLAP, y data mining en el Servidor de Bases de Datos
2003	Oracle Database 10g y Oracle Application Server 10g: “grid” computing Oracle Database 10gAutomatiza el uso de claves
2005	Oracle Adquiere PeopleSoft y anuncia la adquisición de Siebel, Incrementando la oferta de aplicaciones de inteligencia de negocios en aplicaciones ERP y CRM
2007	Oracle Database 11g: Extensiones de capacidades de auto administración y manejo de cambios en la base de datos end-to-end. Adquisición de Hyperion, se agrega Independencia en OLAP y Manejo de desempeño en aplicaciones financieras

Figura 2.1.1información tomada de (Wreenwald, 2009 p. 26)

Oracle Exadata es un sistema integrado creado principalmente para incrementar el desempeño tanto en bases de datos OLTP como en DatawareHousing. Es un concepto de servidor que integra almacenamiento, red y software que es masivamente escalable seguro y redundante en cuanto a su almacén de datos. Cabe aclarar que Oracle Exadata nació como un sistema de almacenamiento propiamente. (Osorio, 2010 p. 11)

Por esta razón es que es también conocido en el argot como SAGE (Storage Appliance for Grid Enviroments), uno de los paradigmas principales en esta nueva tecnología es su tipo de almacenamiento que son los discos de estado sólido.

2.2 Características importantes de Oracle RDBMS

El RDBMS permite mantener una o más bases de datos activas con el diseño del manejador Oracle, las características permiten entender de manera general el funcionamiento y parte de la arquitectura.

Para comenzar, tomaremos en cuenta que en la arquitectura Oracle el proceso de consulta no es igual que el proceso de escritura de datos, esto debido a que como se explicó anteriormente para el caso de la lectura de registros estos son tomados de dos partes: Si es por primera vez que se va a realizar una consulta un proceso en particular busca los datos en la estructura especifica llamada database buffer cache al no existir en memoria tendrán que leerse directamente del disco y almacenarse en el DB Buffer Cache para estar disponible en consultas posteriores, este trabajo lo hace el proceso servidor el cual se encarga de escribir de la memoria principal a la memoria secundaria. En caso de que ya exista la consulta en la estructura de memoria entonces se presentarán otras condiciones en las que esta puede ser de rápida respuesta o por el contrario de tardanza significativa.

Lo anterior técnicamente es conocido como:

Parseo Duro (Hard parsing):
Se presenta cuando es ingresada una consulta SQL por primera vez y no se encuentra en el share pool área (estructura de memoria compartida) del sga, este evento es muy pesado debido a que se utilizan más recursos debido a que se ejecutan todas las etapas de parseo de una consulta.

Parseo Suave (Soft parsing):
Una instrucción sql se presenta y se encuentra una coincidencia en el shared pool área esto facilita el ahorro de algunos recursos ya que no se hace completamente el parseo, no obstante, no es del todo optimo el parseo suave ya que se utilizan también recursos para la sintaxis y la seguridad. (Alapati, 2009 p. 191)

Se puede decir que una consulta al ser almacenada por primera vez, la siguiente ocasión en que sea usada se ahorrará una cantidad de recursos, como el uso de memoria principal que es en donde se almacenan temporalmente todas las consultas, específicamente el SQL Área.

Es importante identificar otras características de ventaja en el tiempo de ejecución de una consulta con la cláusula select. Y que para los casos anteriores de parseo duro y suave se generan un conjunto de combinaciones que es importante identificar para con ello optimizar el tiempo de ejecución. Otra propiedad que habrá que considerar es la red, a continuación se menciona parte de esta característica.

Red:

Es importante saber que el ancho de banda (Volumen) y la velocidad (Latencia) de la red influyen ya que esto permite una mejor transmisión de bloques de datos en tiempo de ejecución. En la red es importante también considerar la transmisión interna que puede estar determinada incluso por los dispositivos de transmisión interna de las tarjetas y también por tecnologías como el Interconnect usado para arquitectura cluster Oracle rac o también para la infiniband de muy alta velocidad usada para la tecnología de servidor integrado Oracle Exadata.

Memoria:

La memoria está directamente relacionada con los bloques o segmentos de datos almacenados en las diferentes estructuras de memoria que se pueden redimensionar según las necesidades o demanda de uso y tomando en cuenta la capacidad de memoria que se tiene disponible. Actualmente los servidores tienen memoria suficiente para trabajar con alta demanda, lo más importante es la configuración de parámetros. El más importante es el SGA_TARGET y en últimas versiones el MEMORY_TARGET que permite incrementar proporcionalmente los tamaños de las áreas contenidas dentro del SGA (System Global Area).

Discos:

Aquí se encuentra un punto crucial para el modelo de simulación que se presenta ya que en la parte de la escritura es de suma importancia profundizar en este conjunto de dispositivos. Para el caso de la ejecución de sentencias en el parseo duro es donde los discos deben tener una combinación óptima, esto quiere decir que al configurar la redundancia y el orden de los arreglos de discos se deben de tomar en cuenta todas las recomendaciones y mejores prácticas mencionadas en los manuales de referencia del fabricante. Es importante también hacer una

afinación a la instancia que administra los arreglos de discos conocida como Oracle ASM (automatic storage management).

Un parámetro que tiene la característica de incrementar el número de agentes que atienden al cliente o petición (servidores) estos servidores podrían verse como procesos en espera de que atiendan a cada solicitud que va llegando y en muchas ocasiones están asociados a un procesador en particular. En 11g R2 ya bien probado y dimensionado es el parallel_degree_policy que muy someramente significa política de grado del paralelismo de las consultas. (Alapati, 2009 págs. 93-94)

Este tiene otros parámetros asociados como son:

MGMT_Pn

PARALLEL_TARGET_PERCENTAGE

PARALLEL_QUEUE_TIMEOUT

PARALLEL_DEGREE_LIMIT_P1

Otros parámetros importantes que están relacionados con el almacenamiento son:

DB_BLOCK_SIZE es un parámetro que establece el tamaño de los bloques de las bases de datos Oracle almacenados en los archivos de datos y caché en el SGA. El rango de valores depende del sistema operativo, pero es típicamente 8192 para bases transaccionales OLTP y valores más altos para los demás sistemas de almacenamiento de datos.

Para el caso de nuestro modelo de simulación existe un parámetro específico llamado DB_WRITER_PROCESS el cual determina el número de escritores vistos como procesos en hilo que hacen la labor de almacenar en el disco duro una vez que se dan las siguientes condiciones:

Existen unos archivos llamados redo log files que en donde se guardan los cambios hechos en las bases de datos y usarlos por una posible falla en el hardware, software o medios de comunicación en un escenario futuro no esperado, los redo log files por ser archivos pequeños se van borrando ya que están configurados para reescribirse para poderse guardar de manera histórica se necesita poner en modo archive log (modo archivamiento en automático). Aunque es importante considerar que ésta configuración hace poco más lenta la escritura debido a que hay duplicidad de escritura, en redo logs y archive logs.

La manera en que funcionan los redo log files es cuando el proceso de grabado LGWR escribe los registros hechos de los cambios del buffer de la memoria a los redo log groups hasta alcanzar el límite permitido en cuanto al tamaño originalmente definido al crearlos o también se da un cambio de grupo en los redo logs cuando se hace una operación de switch de los mencionados archivos. Es así que el proceso LGWR continua escribiendo en el siguiente grupo esto se ejecuta de forma circular, de tal manera que los miembros de los grupos son sobrescritos por los más recientes cambios hechos en las bases de datos.

El tamaño de los redo log files puede también afectar al tiempo de los procesos encargados de escribir datos DBWR y el proceso de archivamiento ARCH. Por lo regular un redo log file grande ofrece un mejor desempeño y contrariamente un tamaño pequeño asignado a los redo log files incrementa el la actividad de checkpoint reduciendo por consecuencia el desempeño. Pero cada base de datos es distinta y no se puede tener una regla general para definir un tamaño óptimo, es necesario hacer un diagnostico especifico de la base en cuestión.

El tamaño de los redo log files no afecta al proceso LGWR, este más bien puede afectar al comportamiento del DBWR y del checkpoint.

El checkpoint es eventualmente afectado por varios eventos: el tamaño de los redo log files el valor del parámetros FAST_START_MTTR_TARGET. Si el valor de éste parámetro está configurado para limitar el tiempo de recuperación de la instancia, Oracle automáticamente intenta ejecutar un checkpoint tan frecuente como sea necesario, dada esta condición el tamaño del redo log file debe ser lo suficientemente grande para evitar eventos de checkpoint innecesarios. El tamaño óptimo puede obtenerse por un query en la columna OPTIMAL_LOGFILE_SIZE de la vista V$INSTANCE_RECOVERY y también se puede obtener información en las alertas y recomendaciones del Oracle enterprice manager.

No existe una recomendación precisa del tamaño de los redo log files pero se recomienda hacerlos de entre 100M y algunos gigabytes, un tiempo aproximado para que se cambien los redo log files y se complete el ciclo de intercambio de datos es de aproximadamente 20 minutos. (Chan, 2011 p. 67)

2.3 Redundancia en discos y ASM

Algo muy importante a considerar es el tipo de almacenamiento que se configura para bases de datos OLTP (online transaction processing) o DSS (decision support system) ya que el desempeño de escritura en los bloques de datos depende del tipo de configuración para bases de datos altamente transaccionales, es decir en la arquitectura OLTP las sentencias INSERT y UPDATE son muy frecuentes. Por otro lado una base de datos de tipo Datawarehouse o DSS en la que el volumen de datos es más alto pero no se hacen muchos cambios en los registros, en estas lo que más se utiliza es la sentencia SELECT mas a menudo para realizar consultas de selección de datos.

Una vez que se analiza lo anterior ya sea para implementar bases de datos por primera vez o para hacer una investigación de la configuración para hacer un modelo de simulación como es en este caso. Se deben tomar en cuenta características como el tamaño de las particiones y el modo en que se organizara la configuración.

Haciendo una breve descripción de almacenamiento mencionaremos que tanto las particiones con formato fat32, ext4, ntfs etc. Como también las particiones crudas (cook partition) raw devices llevan tiempo usándose en servidores, en su mayoría los últimos mencionados ya que el acceso es más rápido debido a su arquitectura; Para tener una administración y configuración adecuada cada controladora de discos de un fabricante en particular desarrollaba su propio software.

La empresa Oracle sacó al mercado a partir de la versión Oracle 10g la característica conocida como ASM para administrar los discos, esta nueva forma de trabajar desde los arreglos de discos tiene componentes muy estándar de software que trabaja con una instancia (memoria y procesos) similar a la de bases de datos de almacenamiento.

A diferencia de ocuparse de la administración de los datos se encarga del control de acceso a los discos, esto anterior, con la finalidad de hacer general el uso de un solo software controlador de dispositivos de almacenamiento anteriormente proporcionado por cada una de las controladoras de fabricantes.

De esta forma cuando el unix sysadmin (el encargado de administrar los dispositivos de almacenamiento en sistema operativo unix) termina de configura los arreglos de discos le presenta estos al Oracle dba y este último se encarga de verlos al instalar el ASM (automatic storage management) y posteriormente configura la instancia para alcanzar un nivel óptimo de desempeño

al acceder los bloques a los discos, el ASM es configurable de manera muy parecida a una instancia convencional a la de bases de datos. Tiene parámetros genéricos y otros adicionales específicos de ASM; pero menor en número de parámetros al de las bases de datos convencionales.

En la Figura 2.3.1 se muestra una pre instalación de almacenamiento en discos con ASM con tipos de redundancia normal y externa, se debe mencionar que existen 3 tipos de redundancia en la tecnología ASM que son: redundancia externa que usa todos los discos físicos para almacenar meta datos, redundancia normal que usa la mitad de los discos para datos y la otra mitad es el espejo que esta por seguridad, redundancia alta que usa una tercera parte de los discos físicos las otras dos terceras partes son para espejo del tipo de redundancia en cuestión usada.

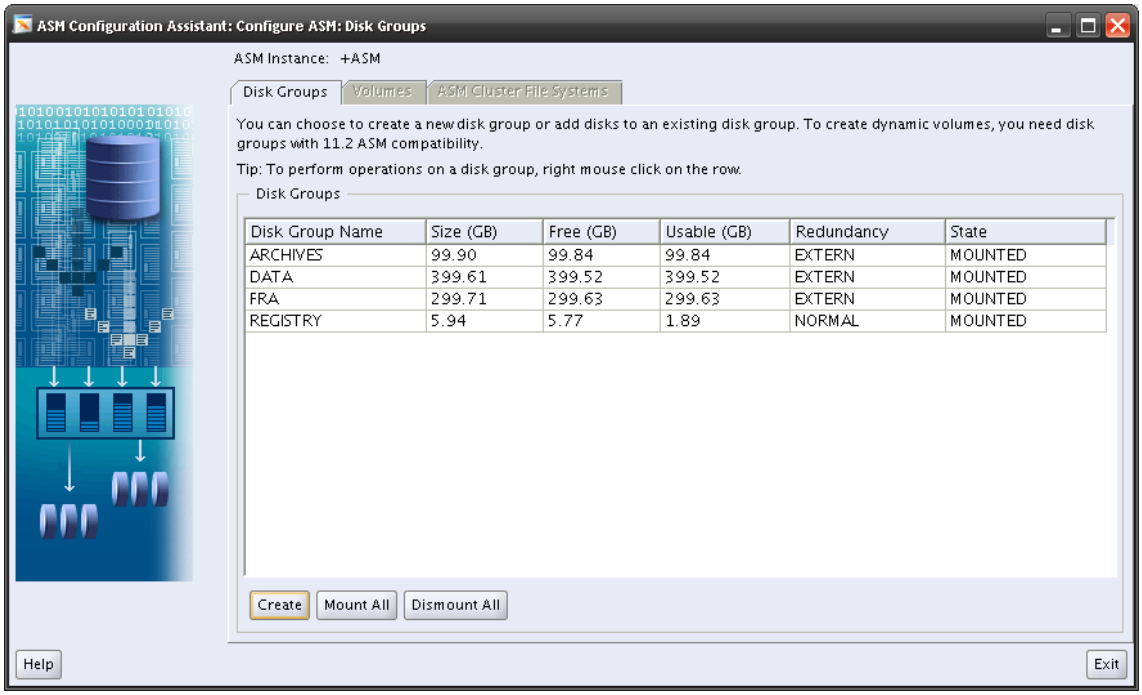

Figura 2.3.1 imagen tomada de una instalación real

La redundancia es el nivel de duplicidad que puede configurarse para un conjunto de discos, esto permite tener un margen de error significativo cuando se presenta un fallo en alguno de los discos. Esta configuración se hace desde el Hardware, es decir, se puede hacer un arreglo con alta redundancia o baja redundancia. Así también desde el Software ASM se puede configurar un determinado tipo de redundancia esta característica es muy estable ya en la versión 11g release 2.

Todos los discos son candidatos a fallos por un posible defecto de fabricación o por el tiempo de uso que puede suceder en cientos o miles de horas y todos los discos tienen ese factor de proporción conocido como MTBF (mean time between failures), este promedio de vida de un disco para una base de datos de alta disponibilidad es muy importante. También para este estudio lo es ya que no hay una misma velocidad en la escritura de discos en un arreglo de baja redundancia que en uno de configuración alta, esto hace que la característica de escritura pueda tomarse como un factor de proporción diferente en uno u otro tipo de redundancia para el modelo de simulación en el arreglo de discos. (Alapati, 2009 p. 134)

El tipo de redundancia que se escoge al hacer esa configuración depende mucho de la necesidad de la base de datos sea esta OLTP uso general o DSS datawarehouse y para efecto de este modelo es importante tomar en cuenta la proporción de escritura a los discos según el arreglo o redundancia que se haya seleccionado.

Como se mencionó la redundancia se hace desde los discos (hardware) o se puede hacer de manera muy estable en la versión 11g release 2 desde el ASM. Se deduce que si se configura una alta redundancia en ambas partes se tiene un mínimo error a fallos pero también queda menos espacio disponible para almacenamiento de datos productivos y lo más importante, la velocidad de escritura es más baja que si no existiera redundancia y los discos estuvieran configurados directamente (redundancia externa).

Antes de usar un disco, la primer tarea es dividirlo, el particionamiento permite almacenar datos del sistema y la aplicación, correspondientes en secciones separadas y estructuradas del disco respectivamente, así también el particionamiento permite un manejo más fácil de todos en su conjunto.

Incluso después de haber particionado los discos no se tiene una manera conveniente de acceso a los datos o de almacenarlos, es por ello que se debe usar un sistema de archivos o filesystem en ingles, esto nos permite tener propietarios individuales de archivos y directorios, seguimiento y monitoreo en tiempo de creación y modificación, control de acceso a los datos y contabilidad de distribución de uso y espacio.

El Filesystem se puede configurar para un solo disco o para varios discos, en este caso el usuario verá presentes los discos como uno solo, esto significa que se ven como un volumen lógico. En el sistema operativo UNIX existen varias maneras de combinar múltiples discos físicos en un único volumen lógico.

Una manera de crear un dispositivo lógico en múltiples sistemas UNIX es usar una herramienta conocida como LVM (Logical Volume Manager), con esta utilería se pueden crear diez discos de 72GB cada uno dando como resultado un volumen lógico de 720GB. En general, cada disco físico es visto como parte de un todo y un único sistema de archivos (filesystem), está conformado por cada uno de los discos físicos. Los más modernos y sofisticados sistemas de almacenamiento usan el LVM debido a que es muy flexible, un ejemplo es que se puede agregar espacio y modificar particiones en un sistema productivo en línea. Finalmente una vez que se crean volúmenes lógicos se pueden designar volúmenes de disco como puntos de montaje haciendo incluso referencia a una carpeta y finalmente se pueden crear archivos individuales en esos puntos de montaje.

Se llama arreglos redundantes o RAID (Redundant Array of Inexpensive Disks) a la forma de configurar una gran cantidad de volúmenes lógicos en uno solo, y sirve para ganar un alto desempeño y una alta disponibilidad en cuanto a la perdida de datos por posible falla en alguno de los discos. Esto también permite ahorrar de una manera considerable y reducir costos de los discos pequeños por separado en tiempo y costo la sustitución de uno dañado en particular. Los datos son separados en pedazos de 32K y 64K generalmente para mantener un estándar además de ser cantidades proporcionales al bloque de sistema operativo. (Alapati, 2009 p. 135)

Y estos pedazos son escritos en cada disco la exacta distribución de los datos es determinada por el RAID. Cuando los datos son leídos en reversa el proceso es invertido dando la impresión de que se está leyendo un único disco de gran tamaño.

Los dispositivos RAID proveen redundancia, si un disco en un dispositivo RAID falla, se puede hacer una reconstrucción inmediata en el disco que presentó el problema.

Dos factores en el desempeño del disco son importantes: La velocidad de transferencia y el número de operaciones I/O por segundo. La velocidad de transferencia se refiere al factor de transferencia con la cual los datos se mueven por el sistema controlador de discos. En cuanto a las operaciones I/O (lectura y escritura a discos)más de un sistema de disco puede ser manejado en un periodo de tiempo determinado.

La mayoría de los sistemas RAID permiten el reemplazo de discos, a este evento se le conoce también como (Hot Swapping).

El llenado de discos se va ocupando secuencialmente, es decir, en una configuración de bloques de 8KB para una cadena de 24KB de información los primeros 8KB se almacenan en el primer disco, los segundos 8KB se almacenan en el segundo y los otros 8KB en el tercero respectivamente.

A continuación se describen los tipos de RAID: (Alapati, 2009 págs. 135-140)

RAID 0

En realidad el raid 0 no tiene una configuración de redundancia como los demás ya que todo el disco está disponible para el almacenamiento de los datos sin que se duplique ningún bloque. Esto por obvias razones permite que la escritura sea la más veloz de todas las configuraciones.

RAID 1

En esta configuración los datos son duplicados y espejeados es decir para cada bloque de datos de 8KB se hace una copia exacta en otro de los discos exactamente del mismo tamaño y con la misma información. Por lo anterior se deduce que la escritura es más lenta que en el caso pierde entre un 10 y 20 % de desempeño ya que la redundancia es más alta. De manera contraria a la escritura, la lectura en este tipo de raid es más óptimo y en menos tiempo ya que se realiza la lectura en paralelo. Se usa raid 1 cuando el valor de los datos es más importante que el desempeño.

RAID 2

Utiliza un sistema de corrección de errores, garantiza un alto rendimiento no obstante es un sistema demasiado costoso ya que el sistema es muy voluminoso y ocupa demasiado espacio además la experiencia en bases de datos productivas han demostrado que es ineficiente.

RAID 3

La diferencia en este tipo de raid la hace un disco de paridad adicional el cual mantiene la información necesaria para la corrección de errores por posible falla. Las paridades traen consigo algoritmos que permiten la reconstrucción de datos en el disco que sufrió daños el raid 3 es más lento que el raid 0 pero con todo ello es mejor que el raid 2 ya que es más rápido y para crear el tipo de arreglo solo se necesita un disco adicional.

RAID 4

En raid 4 los trozos de datos son más grandes que en raid 3 lo que permite que el sistema de

lectura I/O procese más rápido por ser en paralelo las respuestas no obstante que la escritura a los discos presenta lentitud debido a que se debe escribir adicionalmente al disco de paridad antes de escribir al arreglo de los demás disco en este caso el disco de paridad se convierte en un cuello de botella.

RAID 5

En el raid 5 se presenta lentitud en la escritura pero no es comparable como en el arreglo de raid 4 debido a que su sistema permite manejar la concurrencia con más eficiencia varios proveedores de hardware y software han mejorado sus algoritmos y usando memoria no volátil para iniciar el proceso de escritura en cada parseo. El raid 5 virtualmente proporciona los beneficios de velocidad en escritura y alta disponibilidad por posibles fallos. Mejorados en comparación a los anteriores.

RAID 0+1

El raid 0+1 es un espejo de divisiones la diferencia de este y un raid 1+0 es la localización de cada nivel de raid dentro de la localización final. Se hace uso de todos los discos para la recuperación de uno de ellos y no se toleran dos fallas consecutivas. Debido a que existen dudas y falta de confianza para este tipo de arreglo en bases de datos de alta disponibilidad algunas empresas han optado por usar tipos de arreglos como son RAID 0+1+5 (Espejo sobre paridad única) y RAID 0+1+6 (Espejo sobre paridad dual).

Existen también niveles de RAID Anidados por ejemplo:

- RAID 0+1: Un espejo de divisiones
- RAID 1+0: Una división de espejos
- RAID 30: Una división de niveles RAID con paridad dedicada
- RAID 100: Una división de una división de espejos
- RAID 10+1: Un Espejo de espejos
- Otros tipos de RAID son:
- RAID S o RAID de Paridad
- Matrix RAID
- Linux MD RAID 10
- IBM ServerRAID 1E
- RAID Z de Sun Microsystem

Existen ventajas sobre las configuraciones por RAID de Software sobre las de Hardware una de

ellas es la de la migración ya que en ocasiones ni siquiera los discos de un mismo fabricante son compatibles entre una versión y otra.

Existen otro tipo de almacenamientos que solo se mencionarán como son la SAN (Storage Área Networks) basado en almacenamiento compartido para arquitecturas cluster, NAS (Networked Attached Storage) la diferencia es que ésta es más difícil de escalar que la SAN. (Alapati,2009 p. 140)

InfiniBand Es una nueva tecnología que trata de eliminar el protocolo TCP/IP tiene el suyo llamado Sockets Direct Protocol (SDP), reduce los cuellos de botella, no requiere Bus I/O, una liga puede operar a 2.5 GB por segundo.

La infiniBand se usa para tecnologías como Oracle Exadata y además se usa una tecnología que sirve para configurar los volúmenes y la redundancia antes de instalar la instancia de bases de datos o configurar y hacer modificaciones en tiempo productivo.(Alapati, 2009 p. 136)

En el caso de la instancia Oracle ASM que se entiende como un conjunto de procesos de segundo plano y estructuras de memoria que administran los grupos de discos o volúmenes de discos, se puede ver también a esta instancia como un espacio, arreglo de discos o sistema de archivos que administran la información que está ahí. Su función también es la de dar seguridad por posibles fallos en alguno de los discos de algún grupo de ellos en particular.

Estos arreglos tiene también una redundancia que se divide en:

Externa (external): todos los discos se presentan como si fuera un RAID 0 a nivel físico, lo que significa que no usa el espejeo

Normal (normal): se requieren al menos dos grupos de discos para posibles fallos, usa espejeo.

Alta (high): se requieren al menos tres grupos de discos para posibles fallos, usa espejeo (Abraham 2012, p. 42).

Cabe mencionar que para efecto de la configuración del voting disk, los ocr y olr para bases de datos en cluster y stand alone se usa una redundancia normal, es decir, con 3 discos dentro del asm para poder recuperar en caso de pérdida la configuración del arreglo de discos en la instancia +ASM, la redundancia externa no es recomendable ya que se tiene solo una copia de la configuración y en caso de pérdida no se tendría una copia, también la configuración alta se puede

configurar pero sin demasiados discos en este caso 5, con la configuración normal es más que suficiente.

El nivel de redundancia controla cuantos fallos a disco son tolerantes sin desmontar el grupo de discos o sin la perdida de datos, como se mencionó anteriormente el nivel o tipo de grupo de discos determina el nivel de espejeo con el cual Oracle crea archivos en el grupo de discos.

En este caso ASM es más específico y flexible que un RAID de hardware tradicional, en cuanto a la redundancia normal ya que en ASM se puede especificar el nivel de redundancia para cada archivo.

Cuando se crea un grupo de discos es posible definir también el tamaño de la unidad de almacenamiento AU (alocation unit) con el parámetro AU_SIZE que puede medir 1,2,4,8,16,32 o 64 MB esto también depende del nivel de compatibilidad del grupo de discos, lo que se almacena son extents o extensiones cada uno se puede almacenar en un disco diferente. Un LUN Logical unit number: es un disco presentado a un sistema computacional por un arreglo de almacenamiento, Oracle recomienda usar la utilidad de RAID de hardware para crear los LUN's.

Un volumen lógico es mapeado hacia un LUN esto resulta más fácil de configurar, además Oracle no recomienda usar manejadores de volumen lógico para espejeo ya que ASM provee espejeo. (Abraham,2012, p. 62)

2.4 Almacenamiento y el crecimiento en Exadata

La forma en que está organizado el almacenamiento en Oracle se clasifica de dos formas: físico y lógico, el primero se refiere a los archivos de datos que están alojados en el sistema de archivos o en el volumen con un nombre en particular, los más conocidos son los datafiles (archivos de datos) control files (archivos de control), redo log files (archivos de los cambios hechos en la base de datos), archive log files (archivos históricos de los redo logs) server parameter file (archivo de la configuración de parámetros del servidor password file (archivo de contraseñas) principalmente.

Todos ellos se almacenan en el espacio asignado para los datos externos visto desde el sistema operativo. A lo anterior también se debe agregar el bloque de datos del sistema operativo con sus características específicas de cada sistema y el tamaño en particular que le fue asignado.

Posteriormente esta la organización lógica que se conforma por el tablespace o unidad de espacio de los distintos objetos existentes en Oracle estos pueden ser: tablas, procedimientos almacenados, secuencias, índices etc. El tablespace es la unidad lógica que está directamente organizada en la base de datos. Los objetos se llaman segmentos de datos cuando ocupan un espacio en el tablespace y por consecuencia en la base de datos. Un tablespace puede estar formado de uno o más datafiles; pero un datafile no puede apuntar a más de un tablespace.

En versiones anteriores se administraban directamente los denominados extents o extensiones de los segmentos y son incrementos al segmento originalmente creado. Después se clasifican también los segmentos que pueden ser de tipo índice, de tipo datos, de tipo cluster, de tipo undo que es donde se encuentran los segmentos de roll back que anteriormente también eran administrados de forma manual (versiones Oracle 8).

Los bloques Oracle también ocupan un lugar importante en la clasificación lógica ya que están administrados de forma interna en el servidor y están formados de 1 o más bloques de sistema operativo, esto es que, los bloques Oracle son múltiplos de los bloques de sistema operativo.

Con lo anterior se puede entender como está organizado el bloque lógico desde un concepto de almacenamiento organizado internamente y que es la unidad mínima de almacenamiento en la base de datos. Esta es la que nos interesa simular para el caso del modelo de escritura en el almacén además la intenció del tipo que se va a simular es un modo OLTP.

Para hablar más de almacenamiento en la bases de datos Oracle y específicamente de Oracle Exadata, cabe mencionar que éste se inició originalmente como una arquitectura de almacenamiento en discos denominado o conocido como SAGE (Storage Appliance for Grid Environments) que significa una aplicación para manejo de almacenamiento y orientada a la administración modular de componentes Oracle. También fue un código para denominar el proyecto aun en una fase beta. Posteriormente se transformó en una aplicación de almacenamiento integral que incluía a hardware y software. El principal motivo para el desarrollo de Oracle Exadata fue dar solución a los cuellos de botella en bases de datos muy grandes.

En la figura 2.4.1 se muestra la arquitectura general de Oracle Exadata que toma casi toda la arquitectura de un servidor de bases de datos Oracle convencional.

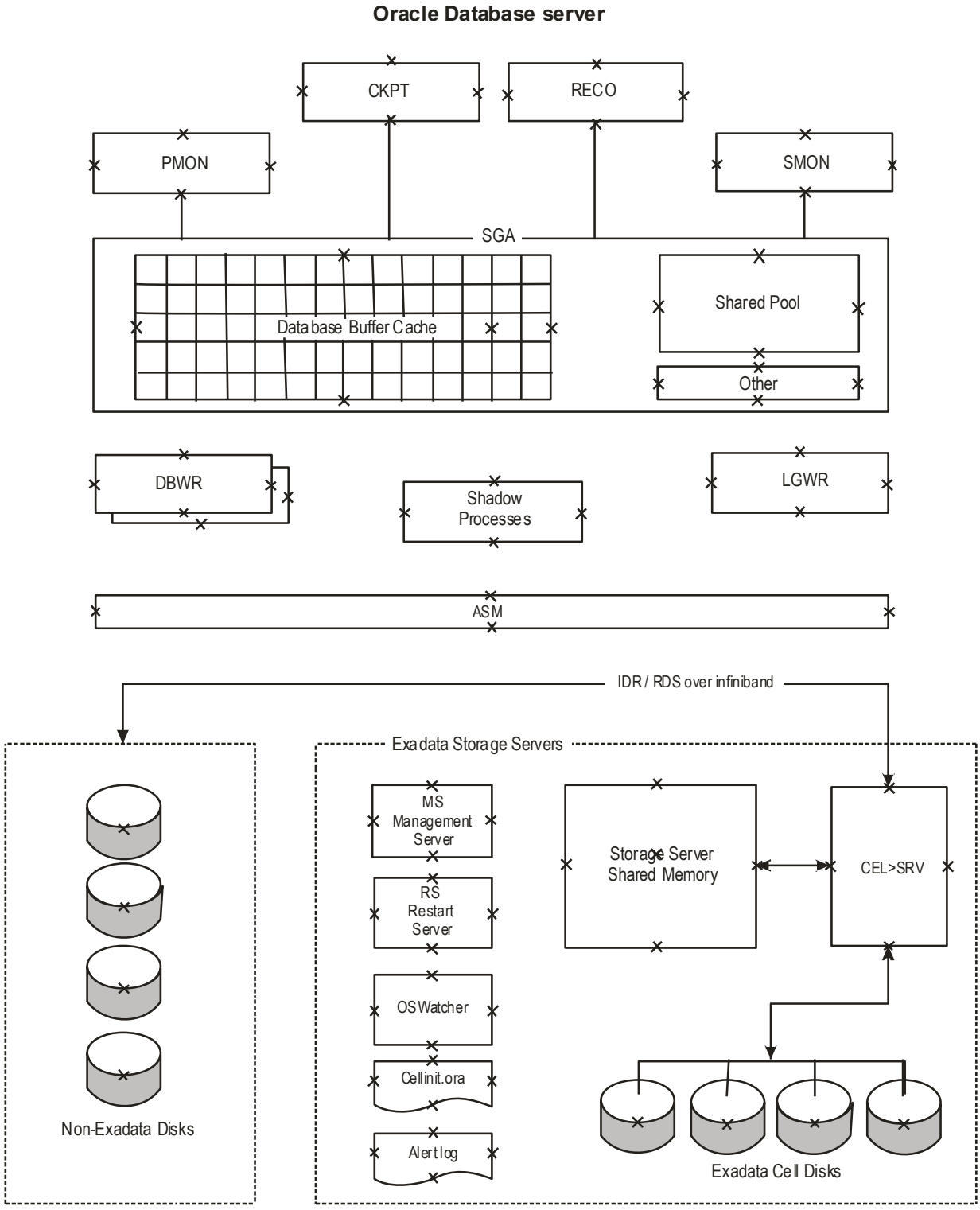

Figura 2.4.1 (Osborne, 2010 p. 18)

Otra característica muy importante es la infiniband que es la parte de comunicación entre dos o más nodos para acelerar el proceso de comunicación, este proceso no será tratado a profundidad en este trabajo.

Algunas características del software que son usadas en un servidor Exadata son:

- Usa Oracle 11g Release 2 que está diseñado para optimizar las capacidades en Oracle Exadata.
- Exadata Smart Scan: mejora el desempeño de query's cuando hay sobrecarga de procesamiento y uso de minería de datos para un servidor de almacenamiento inteligente escalable.
- Exadata Smart Flash Cache
- Exadata Hybrid Columnal Compression: Reduce el tamaño de tablas de Bases de Datos DSS
- Infiniband Network: Permite conectar una multitud de servidores con un alto grado de transmisión entre ellos, con un ancho de banda aproximado de 40 Gigabits

En la figura 2.4.2 se muestra la arquitectura de comunicación física denominada infiniband que es la que pasa a sustituir al interconect en un servidor de bases de datos Oracle RAC convencional.

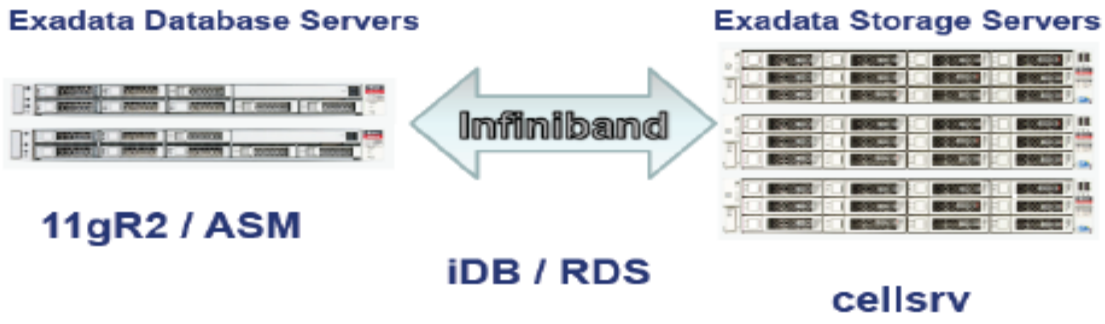

Figura 2.4.2 (Osborne 2010 p. 3)

En el libro Oracle Exadata Recipies se hace un cálculo estimado de la base de datos en cuestión y para efectos de un ejemplo de crecimiento, en el primer query se calcula el tamaño total como se hace en este trabajo pero a diferencia de que se usa una metodología diferente y más simple.

```
SQL> select 'Data Files' type, sum(bytes)/1024/1024/1024 szgb,count(*) cnt
from    v$datafile group by substr(name,1,instr(name,'/',1,2)-1)
union
select 'Temp Files', sum(bytes)/1024/1024/1024 szgb,count(*) cnt
from v$tempfile group by substr(name,1,instr(name,'/',1,2)-1)
union
select 'Redo Member',sum(l.bytes)/1024/1024/1024 szgb,count(*) cnt
from v$logfile lf, v$log l
where l.group#=lf.group# group by substr(member,1,instr(member,'/',1,2)-1)
/

File Type      Size (GB)
-----------    ---------
Number of files
---------------
36 8 4

Data Files
Redo Membe
Temp Files
2950.68
32.00
72.00
```

Para la parte de pronóstico se hace una segunda consulta que contabiliza el crecimiento de 3 días para poder hacer una sumatoria y pronosticar el crecimiento; pero siempre y cuando sea constante, esto obviamente limita un pronóstico con más factores que pudieran modificar el resultado. Por tal motivo es más objetivo el método de esta tesis para pronosticar el crecimiento con la metodología Box-Jenkins.

```
SQL> select avg(gbpd) AvgGBpd, max(gbpd) MaxGBpd,
avg(gbpd) * &&num_days AvgReq,
max(gbpd) * &&num_days MaxReq
from (
        select  al.dt, (al.blks * kc.blksize)/1024/1024/1024 gbpd
        from
                (select trunc(completion_time) dt, sum(blocks) blks
                from v$archived_log
                group by trunc(completion_time)) al,
                (select max(lebsz) blksize
from x$kccle) kc Enter value for num_days: 3)

/

old   2:
new   2:
old   3:
new   3:
avg(gbpd) * &&num_days AvgReq,
avg(gbpd) * 3 AvgReq,
max(gbpd) * &&num_days MaxReq
max(gbpd) * 3 MaxReq
Avg GB per Day
------------------
           6.11
1 row selected.
Max GB per Day
--------------------
123.16

GB Req Avg
---------------
18.34

GB Req Max
----------------
369.48
```

Esta última consulta es una suma proporcional de la acumulación de espacio en 3 días de crecimiento. Solo se menciona como ejemplo para aclarar que el crecimiento se pronostica de manera muy superficial sin tomar en cuenta el costo beneficio y un procedimiento más sistemático y científico. (Clarke, 2013 págs. 122-123).

Capítulo III Modelos matemáticos

En este capítulo se exponen los eventos continuos y discretos así como las cadenas de markov asociadas, los procesos estocásticos, así como las series de tiempo con las que posteriormente se desarrollará la investigación.

Capítulo III Modelos matemáticos

3.1 Eventos continuos y discretos

En primera instancia se expone lo siguiente: existen dos formas de dividir el conocimiento científico, las ciencias formales en las que se encuentran la matemática, la lógica y el universo de los números, y en segundo lugar las ciencias fácticas en las que se encuentran todos los campos de investigación en los que intervienen objetos tangibles o en los que es necesario generar modelos a partir de formas concretas o existentes físicamente, puede hablarse aquí de ejemplos como las ciencias naturales, sociales físicas etc.

"*La estadística es una rama de las matemáticas que tiene aplicaciones en cada faceta de nuestra vida. Es un lenguaje nuevo y poco conocido para casi todas las personas, pero, al igual que cualquier idioma nuevo, la estadística puede parecer agobiante a primera vista. Queremos que el lector "entrene su cerebro" para entender este nuevo lenguaje paso a paso. Una vez aprendido y entendido el lenguaje de la estadística, veremos que es una poderosa herramienta para el análisis de datos en numerosos campos de aplicación diferentes.*" (Mendenhall, 2010 p. 3)

Por otro lado se divide en dos ramas que son la estadística descriptiva que se relaciona mas con la observación y la estimación a partir de una muestra como parte de una población (presentación de gráficas) y la estadística inferencial que se encarga mas de generar los modelos estadísticos y usar mas la aleatoriedad en función de las realizaciones que sustituyen a la muestra, es aquí en donde se estudia la serie de tiempo y sus procesos estocásticos respectivos para generar uno o varios análisis a partir de un modelo en este caso usando la metodología de Box-Jenkins.

"*El objetivo de la estadística inferencial es hacer inferencias (es decir, sacar conclusiones, hacer predicciones, tomar decisiones) acerca de las características de una población a partir de información contenida en una muestra.*" (Mundanal, 2010 p. 4)

A la metodología se le puede definir como la forma de manipular al método de forma crítica para llegar a un resultado esperado haciendo uso de técnicas que en su conjunto nos facilitan hacer una investigación más eficiente, en este caso se usan herramientas computacionales como es el caso del software de distribución libre GRETL.

En las Ciencias Matemáticas existen dos formas de acceder a un resultado a partir del quehacer científico: el método deductivo y el método inductivo.

La diferencia principal en cada uno de ellos es que el método deductivo aspira a demostrar, mediante lógica pura, la conclusión en su totalidad a partir de unas premisas, de manera que se garantice la veracidad de las conclusiones. El método inductivo crea leyes a partir de la observación de los hechos mediante la generalización del comportamiento observado, desarrolla una generalización que mediante la lógica demuestra las leyes o conjunto de conclusiones. En el segundo se intenta a partir de casos particulares o partes por separado crear un modelo de un sistema específico, sometido a pruebas con valores discretos que nos permite llegar a una conclusión que se acerque a la realidad.

El pensamiento del hombre es en comparación al de otros seres vivos característico por abstraer y crear modelos de su realidad de lo que le rodea, y todo esto lo usa para su beneficio en función de lo que se plantea para poder entender y estimar la manera en que se comporta un fenómeno en particular de la naturaleza o de su entorno. No por ello los seres vivos que no son humanos no tienen inteligencia para usarla en su entorno su inteligencia es innata a su conducta sin que independiente mente generen cuestionamientos e hipótesis.

El ser humano crea modelos que pueden representar a una realidad en particular y con ello entender su comportamiento con la finalidad de hacer una estimación y resolver un problema. Existen pues, modelos que representan sistemas.

Un sistema es un conjunto de objetos o ideas inter relacionadas y para delimitarlas se usan criterios. Tiene también propiedades como la Sinergia: La interrelación de las partes es mayor o menor que la simple suma de las partes y Entropía: Indica el grado de desorden del sistema. Se puede reducir la entropía ingresando información al sistema. Véase también pensamiento sistemático empírico y racional (Xirau, 2011 págs. 299-327)

Todos los modelos matemáticos son sistemas que interactúan entre sí, como ejemplo de ellos tenemos los siguientes ejemplos explicados como variables aleatorias:

Variables Aleatorias Discretas:

- El número de posibles valores de la variable es finito o Infinito pero al final contable
- Para cada posible valor x_i de la variable X se tiene que $p(x_i) = p(X = x_i)$ es la probabilidad de que X tome la variable de x_i.

- Se Cumplen las siguientes condiciones

  a) $p(X_i)$ Para todo x_i

  b) $\sum_{i=1}^{\infty} \mathrm{p(xi)} = 1$

- Distribución de probabilidad o función masa de probabilidad *(puf)* de X es el conjunto de pares $(x_i, p(x_i))$ con $i=1, 2 \ldots$

Variables aleatorias continúas:

- El espacio de valores de la variable *X(Rx)* es un intervalo o un conjunto de intervalos.
- La probabilidad de que el valor de X se encuentre en un intervalo $[\ a,b\ ]$ está dada por la expresión:

$$\mathrm{P(a \leq x \leq b)} = \int_a^b f(x)\,dx$$

f(x) se llama función de densidad de probabilidad (*pdf*) de la variable X

- pdf satisface las siguientes condiciones:

  a) $f(x) \geq 0$ Para todo x

  b) $\int_{Rx} f(x) = 1$

  c) $f(x)=0$ si $x \notin Rx$

En la figura 3.1.1 se muestra un esquema que muestra la división entre modelos continuos y modelos discretos

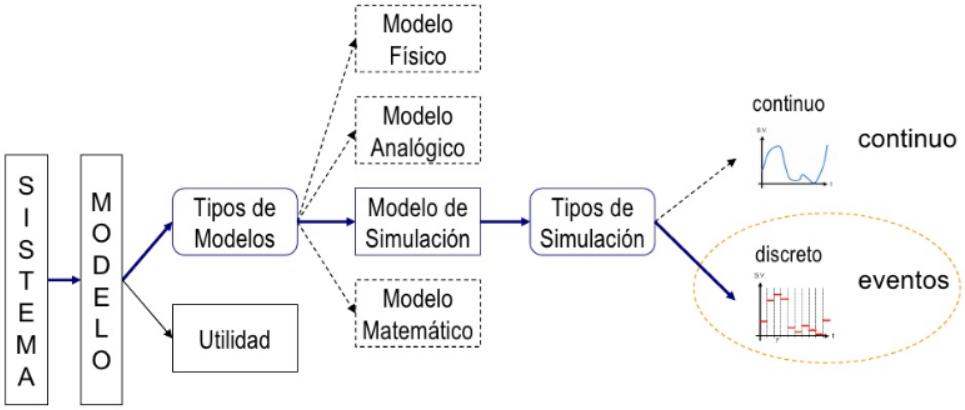

Figura 3.1.1

3.2 Funciones continuas y discretas en teoría de colas

Funciones de distribución acumulativa cumulative distribution function CDF se denota por $F(x)$, mide la probabilidad de que la variable X tenga un valor menor o igual que x; es decir:

$$F(x) = P(X \leq x)$$

- Si X es discreta $\quad F(x) = \sum_{xi \leq x} p(xi)$

- Si X es continua $\quad F(x) = \int_{\infty}^{x} f(t)dt$

Propiedades de $F(x)$:

a) F es una función no decreciente. Si $a<b$ entonces $F(a) \leq F(b)$

b) $\lim_{x \to \infty} f(x) = 1$

c) $\lim_{x \to \infty} f(x) = 0$

Valor esperado

E(x) se denomina media y se define de la siguiente forma:

Si X es continua $\quad E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$

Si X es Discreta $\quad E(X) = \sum_{\substack{X_i P(X_i) \\ todo\ i}}$

Media: la media es la medida de la tendencia central de la variable aleatoria

Varianza:

*V(X)*o σ^2 Se denomina varianza y se define como:

$$V(X)=E\left[(X-E(X))^2\right]$$

Equivalentemente

$$V(X)=E(X^2)-\left[E(X)^2\right]$$

La varianza de X mide la variación de los valores de x respecto a la media y la desviación estándar σ se define como la raíz cuadrada de la varianza de X.

La Moda se define como sigue:

Para variable discreta es el valor de la variable que aparece más frecuentemente

Para variable continua la moda es el valor máximo de*(pdf)*

La moda puede no ser única. Si el valor de la moda aparece en dos valores la distribución es bimodal.

3.3 Distribuciones discretas en los sistemas de colas

Bernoulli

Es una distribución de probabilidad con dos puntos de probabilidad discreta definida por:

$p(0)=q$

$p(1)=p$ siendo $p+q=1$ dado $p,q > 0$

Sea un experimento de manera constante en n ensayos y supóngase que cada uno puede tener éxito o fracaso (1 o 0) sea por ejemplo $x_j = 1$ éxito y $x_j = 0$ fracaso, se tiene entonces:

$$p(x_1, x_2 \ldots x_n) = p_1(x_1) \cdot p_2(x_2) \ldots p_n(x_n)$$

$$P_j(x_j) = P(x_j) = \begin{cases} p & x_j = 1 \quad j = 1,2 \ldots n \\ 1 - p = q & x_j = 0 \quad j = 1,2 \ldots n \\ 0 & otros \end{cases}$$

La distribución de Bernoulli sirve para representar la estacionariedad en sentido estricto para el caso de ejemplos binomiales, no es del todo fácil demostrar esta característica para efectos teóricos, así que se usan otros métodos de demostración.

La estacionariedad en sentido amplio o débil es un criterio para generar un modelo ARIMA estabilizando la media, la varianza y haciendo a que la covarianza no dependa del tiempo t sino del valor inmediato anterior k. La explicación de este concepto escapa al alcance de este trabajo así que se sugiere consultar otra bibliografía más avanzada en el tema.

Una sugerida es: (Videgaray, 2011 págs. 63, 67).

Binomial

$$p(x) = \begin{cases} \binom{n}{x} p^x q^{n-x} & x = 0,1,2 \ldots n \\ 0 & otros \end{cases}$$

La media y la varianza son: p

$E(x)=np$ $V(x)=npq$

Distribución Geométrica (Relacionada con la secuencia de ensayos de Bernoulli)

X indica el número de ensayos para obtener el primer éxito. La distribución de esta variable es:

$$p(x) = \begin{cases} q^{x-1}p & x = 1,2 \dots \\ 0 & otros \end{cases}$$

El evento $\{X=x\}$ ocurre cuando hay $x-1$ fallos seguidos de un éxito cada uno de los fallos tiene asignada una probabilidad de q=1-q y cada uno de los éxitos tiene probabilidad p, así:

$$P(FFF \dots FS) = q^{x-1} * p$$

Poisson

Se utiliza para modelar tiempos entre eventos aleatorios ocurridos en un intervalo de tiempo fijo, la función masa de probabilidad (PMF) está dada por:

$$p(x) = \begin{cases} \dfrac{e^{-\alpha}\alpha^{x}}{x!} & x = 1,2 \dots \\ 0 & otros \end{cases} \qquad con\ \alpha > 0$$

Una propiedad importante de la distribución de Poisson es $E(x) = V(x) = \alpha$ La función de distribución acumulativa es:

$$f(x) = \sum_{1=0}^{x} \frac{e^{-\alpha}\alpha^{i}}{i!}$$

3.4 Distribuciones continuas en los sistemas de colas

Distribución uniforme

Se usa cuando todos los valores en un rango finito se pueden considerar iguales, la variable aleatoria X esta uniformemente distribuida en el intervalo (a,b) si pdf está dada por:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otros \end{cases}$$

cdf está dada por:

$$f(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$$

Se tiene que

$$p(x_1 < x < x_2) = F(x_2) - F(x_1) = \frac{x_2 - x_1}{b-a}$$

La media y la varianza son:

$$E(X) = \frac{a+b}{2} \qquad V(X) = \frac{(b-a)^2}{12}$$

Distribución exponencial

Utilizada para modelar tiempo entre llegadas y también tiempos entre servicios, una variable aleatoria X se dice que tiene distribución exponencial con parámetro $\lambda > 0$ si su pdf está dada por:

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x \geq 0\\ \\ 0 & resto\end{cases} \quad \Rightarrow \quad F(x)=\begin{cases}1-e^{-\lambda x} & x \geq 0\\ \\ 0 & resto\end{cases}$$

La media y la varianza son:

$$E(x)=\frac{1}{\lambda} \qquad V(x)=\frac{1}{\lambda^2}$$

Distribución gamma

Se utiliza para representar el tiempo requerido para finalizar una tarea; una variable aleatoria X tiene una distribución gamma con parámetros β y θ si PDF es:

$$f(x)=\begin{cases}\dfrac{\beta\theta}{\Gamma(\beta)}(\beta\theta x)^{\beta-1}e^{-\beta\theta x} & x>0\\ \\ 0 & otros\end{cases}$$

Donde β es un parámetro de forma y θ es un parámetro de escala y además:

$$\Gamma(\beta)=\int_0^\infty x^{\beta-1}e^{-x}dx \qquad Para\ \beta\ entero \qquad \Gamma(\beta)=(\beta-1)!$$

La media y la varianza son:

$$E(x)=\frac{1}{\theta} \qquad V(x)=\frac{1}{\beta\theta^2}$$

Cuando es entero, la distribución gamma está relacionada con la exponencial para $\beta=1$ se obtiene una distribución exponencial

Distribución Erlang-k

La expresión PDF de gamma, para $\beta = k$, con K entero se denomina distribución Erlang de orden K, la media y la varianza son:

$$E(x) = \frac{1}{\theta} \qquad\qquad V(x) = \frac{1}{K\theta^2}$$

Con lo anterior se verifica qué:

$$F(x) = \begin{cases} 1 - \sum_{1=0}^{k-1} \frac{e^{-k\theta x}(K\theta x)^i}{i!} & x > 0 \\ 0 & x \le 0 \end{cases}$$

Distribución normal

Una variable aleatoria X con media $\mu(-\infty < \mu < \infty)$ y varianza σ^2 tiene una distribución normal si PDF es:

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad -\infty < x < \infty$$

Se utiliza la notación $N(\mu, \sigma)$

La CDF de la distribución normal es:

$$F(x) = p(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\Pi}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] dt$$

La función acumulativa CDF está tabulada y es:

$$\varphi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\Pi}} e^{-\frac{t^2}{2}} dt$$

Distribución weibull

Se utiliza en modelos de fiabilidad para representar tiempos de vida de dispositivos. Un sistema formado por muchas partes independientes y el sistema falla cuando uno de ellos también falla. Una variable aleatoria tiene una distribución weibull si *pdf* es:

$$f(x)=\begin{cases}\dfrac{\beta}{\alpha}\left(\dfrac{x-v}{\alpha}\right)^{\beta-1}\exp\left[-\left(\dfrac{x-v}{\alpha}\right)^{\beta}\right] & x \geq v \\ \\ 0 & otros\end{cases}$$

Los tres parámetros de una distribución weibull son $v(-\infty < v < \infty)$ que es el parámetro de localización; $\alpha(\alpha > 0)$ que es el parámetro de escala y $\beta(\beta > 0)$ que es el parámetro de forma. Para se tiene que la PDF es:

$$f(x)=\begin{cases}\dfrac{\beta}{\alpha}\left(\dfrac{x}{\alpha}\right)^{\beta-1}\exp\left[-\left(\dfrac{x}{\alpha}\right)^{\beta}\right] & x \geq 0 \\ \\ 0 & otros\end{cases}$$

Distribución triangular

Se utiliza cuando no se conoce la forma exacta de la distribución pero se estima el mínimo, el máximo y la moda de una variable.

Una variable aleatoria X tiene distribución triangular si *pdf* es:

$$f(x)=\begin{cases}\dfrac{2(x-a)}{(b-a)(c-a)} & a\le x\le b\\[2ex] \dfrac{2(c-x)}{(c-b)(c-a)} & b<x\le c\\[2ex] 0 & otros\end{cases}$$

La *cdf* de una distribución triangular es:

$$F(x)=\begin{cases}0 & x\le a\\[2ex] \dfrac{(x-a)^2}{(b-a)(c-a)} & a<x\le b\\[2ex] 1-\dfrac{(c-x)^2}{(c-b)(c-a)} & b<x<c\\[2ex] 1 & x>c\end{cases}$$

3.5 Procesos y optimización

Para comprender las características de la simulación una vez planteado el modelo es importante definir los tipos de procesos que existen y su significado, así también hacer una introducción a los procesos matemáticos formales, a continuación se mencionan los conceptos del siguiente autor.

Proceso: es el conjunto de las fases de un fenómeno en evolución, un ejemplo de ello podría ser el proceso de alguna enfermedad, desarrollo del cuerpo humano, fabricación de productos, etc.

Proceso transitorio: es aquel que una vez ocurrido un suceso jamás se volverá a repetir; por ejemplo el proceso de la vida humana, una vez pasado a la adolescencia jamás volveremos a estar en la niñez.

Proceso absorbente: una vez ocurrido un suceso jamás podremos salir de ese estado, por ejemplo la muerte.

Proceso recurrente: es aquel que tiene una probabilidad de volver a ocurrir, por ejemplo, visitar un determinado lugar que nos agrade como la Feria de Chapultepec, etc.

Proceso periódico: es aquel en que tiene cierta frecuencia, un ejemplo de ello son las estaciones del año.

Proceso estable o estacionario: es aquel que se encuentra dentro ciertos límites por ejemplo el ritmo de nuestro corazón.

Proceso estocástico: es una colección de variables aleatorias x_t , en donde t toma valores de un conjunto T dado. Con frecuencia T se toma como el conjunto de enteros no negativos y x_t representa una característica de interés medible en el tiempo. Por ejemplo $x_1, x_2, x_3, \ldots$ podrían representar el número de visitantes a la feria de Chapultepec cada semana. (Delgado, 2000 p. 88)

Como se puede observar los procesos estocásticos son los más importantes en la investigación ya que aquí se encuentran puntos específicos de un comportamiento diferente en el tiempo en este caso x_n representa cada uno de los pedazos de tiempo.

En los proyectos de investigación matemática se presentan dos tipos de estimaciones que se pueden medir y que ambas son importantes, la cuantitativa y la cualitativa la primera en el presente

estudio puede ser: número de bloques que llegan en un periodo de tiempo de alta demanda al servidor, en la parte cualitativa puede presentarse un ejemplo cuando se tiene que medir la proporción de crecimiento de los datos en una empresa en particular para el caso de su base de datos.

La parte cuantitativa es más puntual en cuanto al número de lo que se mide y la parte cualitativa hace una estimación de una proporción de los datos llegando a lo que se conoce como una estimación aproximada.

Un proceso estocástico se puede representar con t=1,2,3… y cada variable aleatoria x toma un valor en el rango M+1 que se caracterizan por los estados de un proceso en particular. Al conjunto de estados posibles que se pueden presentar Espacio de Estados y se simbolizan por S. Así al conjunto de valores que puede tomar el tiempo t se llama espacio paramétrico, representado por ζ.

Un proceso estocástico tiene la propiedad markoviana si:

$$P(X_{t+1} = j \mid X_0 = k_0, X_1 = k_1, \ldots, X_{t-1} = k_{t-1}, X_r = i) = P(X_{t-1} = j \mid X_t = i)$$

Para $t = 0,1,\ldots$ y toda sucesión $i, j, k_0, k_1, \ldots k_{t-1}$

Con lo anterior se puede explicar el hecho de no depender de los estados históricos para determinar la probabilidad actual de una variable, esto es que para x_t se calcula la probabilidad del evento siguiente en función del estado actual, las probabilidades condicionales $P(X_{t+1} = j \mid X_t = i)$ se llaman probabilidades de transición. Si para cada i y j se tiene:

$$P(X_{t+1} = j \mid X_t = i) = P(X_t = j \mid X_0 = i) \text{ Para todo } \quad t = 0,1,\ldots,$$

Por lo tanto se tiene que las probabilidades de transición son estacionarias y se representan por p_{ij} .

Se deduce que al tener probabilidades estacionarias de transición no cambiarán con el tiempo y esto asegura tener más certeza en la estimación de un conjunto de partes en particular.

Se tiene: i,j y $(n = 0,1,2,\ldots)$

$$P(X_{t+n} = j \mid X_t = i) = P(X_n = j \mid X_0 = i)$$

Para todo t=0,1,… se tiene que p_{ij}^{n} son probabilidades condicionales y se llaman probabilidades de transición a los n pasos, después de un numero de iteraciones (unidades de tiempo) en i pasa a j.

Las p_{ij}^{n} deberán cumplir lo siguiente:

a) $0 \leq p_{ij}^{n} \leq 1$ Para toda i, j y n=0,1,2,…

b) $\sum_{j=0}^{M} p_{ij}^{n} = 1$ Para toda i, j y n=0,1,2,…

Un proceso estocástico es una cadena de markov si cumple con las siguientes características:

1 Un número finito de estados
2 La propiedad markoviana
3 Probabilidades de transición estacionaria
4 Un conjunto de probabilidades iniciales $P(X_0 = i)$ para toda i

(Delgado, 2000 p. 89)

El caso más simple de un proceso estocástico en que los resultados dependen de otros, ocurre cuando el resultado en cada etapa sólo depende del resultado de la etapa anterior y no de cualquiera de los resultados previos. Tal proceso se denomina proceso de Markov o cadena de Markov (una cadena de eventos, cada evento ligado al precedente) Estas cadenas reciben su nombre del matemático ruso Andrei Andreevitch Markov (1856-1922).

Por lo tanto se resume que una cadena de Markov es una sucesión de ensayos similares u observaciones en la cual cada ensayo tiene el mismo número finito de resultados posibles y en donde la probabilidad de cada resultado para un ensayo dado depende solo del resultado de ensayo inmediatamente precedente y no de cualquier resultado previo.

La autocorrelación es una herramienta básica utilizada en casi todos los métodos avanzados como son: ARMA de Wold, ARIMA de Box-Jenkins, ARARMA de Parzen y filtrado de Kalman.

3.6 Modelos ARIMA y Box-Jenkins

Los modelos autoregresivos y de medias móviles se originan considerando las propuestas de dos matemáticos Norbert Wiene y A. Khintchine quienes realizaron estudios en las series armónicas principalmente, comparadas con los modelos usados en campos como la astronomía y posteriormente en el campo de la economía, antes de ellos y en esta misma área, se hizo uso de series armónicas por Lagrange quien investigo sobre irregularidades en estas.
Yule explico con un ejemplo la serie de tiempo autoregresiva, describiendo la existencia de algo que grabara el movimiento de un péndulo que se pone a oscilar y que posteriormente se perturba este movimiento por alguien que modifica las trayectorias o formas de este evento Slutsky (Rusia)
Propuso una descripción parecida a la de Yule; llegaron a conclusiones muy similares por diferentes caminos del lado analítico y también del lado inductivo.

La discusión de Norbert Wiene y A. Khintchinese basó principalmente en el margen de error o ruido blanco que caracteriza a una serie como autoregresiva. (Pullock, 1999 págs. 30 - 41)

En 1970, George E. P. Box y Gwilym M. Jenkins desarrollaron una metodología destinada a identificar, estimar y diagnosticar modelos dinámicos de series temporales en los que la variable tiempo juega un papel fundamental.

Una parte importante de esta metodología está pensada para liberar al Matemático de la tarea de especificación de los modelos dejando que los propios datos temporales de la variable a estudiar nos indiquen las características de la estructura probabilística subyacente. En parte, los procedimientos que se van a analizar se contraponen a la manera de identificar y especificar un modelo que se apoya en las teorías que anteceden al fenómeno analizado aunque, convenientemente utilizados, los conceptos y procedimientos que se examinan constituyen una herramienta útil para ampliar y complementar los conocimientos básicos.

Se deberá definir una estructura que permita, por sus características, cumplir el fin último de predicción: proceso estocástico estacionario. Se mencionan cuáles son las condiciones que debe de cumplir esta función para poder calcularla y así definir el proceso estocástico estacionario lineal y discreto. Posteriormente, se analizan los modelos más simples (que emplean menos retardos) conforme a una serie de funciones características (covarianza, autocorrelación total y autocorrelación parcial), describiendo sus condiciones y planteando estructuras teóricas que luego

puedan ser identificables con series temporales reales, esto es de manera general un método para hacer diagnósticos y pronósticos de series temporales. Las series de tiempo en los modelos (ARIMA) y la metodología Box–Jenkins son de suma importancia en la estimación de la tasa de crecimiento de algún tipo de modelo como es el caso del crecimiento de bases de datos y así también en el conocimiento de otras áreas de estudio:

"Una serie temporal es una secuencia ordenada de observaciones cada una de las cuales está asociada a un momento de tiempo. Ejemplos de series temporales las podemos encontrar en cualquier Campo de la ciencia. En Economía cuando buscamos datos para estudiar el comportamiento de una variable económica y su relación con otras a lo largo del tiempo, estos datos se presentan Frecuentemente en forma de series temporales. Así, podemos pensar en series como los precios diarios de las acciones, las exportaciones mensuales, el consumo mensual, los beneficios trimestrales, etc. En Meteorología, tenemos series temporales de temperatura, cantidad de lluvia caída en una región, velocidad del viento, etc. En Marketing son de gran interés las series de ventas Mensuales o semanales. En Demográfica se estudian las series de Población Total, tasas de natalidad, etc. En Medicina, los electrocardiogramas o electroencefalogramas. En Astronomía, la actividad solar, o en Sociología, datos como el número de crímenes, etc."(González, 2009 p. 88)

Un proceso estocástico es una sucesión de variables aleatorias Y_t ordenadas, pudiendo tomar t cualquier valor entre -∞ y ∞.

Para caracterizar un proceso estocástico, basta con especificar la media y la varianza para cada Y_t la covarianza para variables referidas a distintos valores de *t*:

$$E[Y_t] = \mu_t$$

$$\sigma_t^2 = Var\,(y_t) = E[y_t - \mu_t]^2$$

$$\gamma_t = Cov(Y_t - Y_s) = E[(y_t - \mu_t)(y_s - \mu_s)]$$

Existen dos tipos de procesos estocásticos que se usan en los procesos ARIMA en especial:

Ruido Blanco: Sucesión de variables aleatorias (proceso estocástico) con esperanza media 0 cero y varianza constante e independiente para distintos valores de t (covarianza nula)

Proceso estocástico estacionario: Un proceso estocástico se dice que es estacionario si las funciones de distribución conjuntas son invariantes con respecto a un desplazamiento en el tiempo (variación de ***t***). Es decir, considerando que$t, t+1, t+2, \ldots., t+k$ reflejan periodos sucesivos:

$$F(Y_t\ , Y_{t+1}, \ldots..Y_{t+1}) = F(Y_{t+m}, Y_{t+1+m}, \ldots..., Y_{t+k+m})$$

Esta definición anterior se conoce como: estacionariedad en sentido estricto o fuerte y puede cambiar sustancialmente usando la estacionariedad en sentido amplio o débil.
Un proceso es débilmente estacionario si:

Las esperanzas matemáticas de las variables aleatorias no dependen del tiempo, son constantes.

$$E[Y_t] = E[Y_{t+m}] \quad \forall\, m$$

Las varianzas tampoco dependen del tiempo y son finitas

$$Var[Y_t] = Var[Y_{t+m}] \neq \infty \quad \forall\, m$$

Las covarianzas entre dos variables aleatorias de proceso correspondientes a periodos distintos de tiempo (distintos valores para t) únicamente dependen del rango de tiempo transcurrido entre ellas:

$$Cov(Y_t, Y_s) = Cov(Y_{t+m}, Y_{s+m}) \forall\, m$$

De esto último se concluye que, si un fenómeno es estacionario, sus variables pueden estar relacionadas linealmente entre sí, pero de forma que la relación entre dos variables depende de la distancia temporal **k** transcurridas entre ellas.

Principalmente las funciones que se utilizan son la Función de auto correlación simple o total (FAS), la función de auto correlación parcial (FAP), estos en un gráfico llamado correlograma. Y el espectrograma.

Hay dos tipos de series en general: multivariantes y univariantes, para este trabajo se usan las series univariantes, o en otro sentido se dice que se usa solo una unidad experimental y no más de dos. La estacionariedad en sentido estricto garantiza la estacionariedad en sentido amplio o débil pero no a la inversa.

La autocorrelación, en adelante “AC”, sirve para identificar el patrón básico y determinar el modelo apropiado de la serie de tiempo. El coeficiente de correlación es la asociación entre dos variables y describe lo que acontece con una de ellas si se presenta un cambio en la otra. El grado de esta relación mencionada anteriormente fluctúa entre -1 y +1 un aumento en ambos sentidos se relaciona con una disminución y el valor de cero indica que no existe relación entre las dos variables. El grado de la relación de una autocorrelación se mide mediante el coeficiente de correlación.

Existen tres modelos principalmente que describen cualquier tipo de datos para la metodología Box-Jenkins:

- Autoregresivos (AR)
- De medias móviles (MA)
- De promedio móvil autorregresivo mixto (ARIMA) Usado para la metodología Box-Jenkins.

Una importante clase de modelos estocásticos para describir series de tiempos que ha recibido gran atención, son los llamados modelos estacionarios, los que asumen que el proceso permanece en equilibrio estadístico con una propiedad probabilística que no cambia sobre el tiempo, en particular variando un nivel promedio constante fijo y con una varianza constante.

3.7 Modelos autorregresivos y de medias móviles ARIMA

ARIMA significa: modelos autorregresivos integrados y de medias móviles.

Un modelo es autorregresivo si la variable endógena de un periodo t es explicada por las observaciones de ella misma correspondientes a periodos inmediatamente anteriores añadiéndose una variable error. En el caso de procesos estacionarios con distribución normal, la teoría estadística de los procesos estocásticos nos dice que, bajo determinadas condiciones previas toda Y_t puede expresarse como una combinación lineal de sus valores pasados (parte sistemática) más un término de error (innovación o ruido blanco).

Los modelos autorregresivos se abrevian con AR, el orden de un modelo de este tipo se representa como: $AR(1), AR(2), \ldots . AR(n)$ un modelo $AR(1)$ se expresa:

$$Y_t = \phi_0 + \phi_1 + Y_{t-1} + a_t$$

El error de este tipo se denomina Ruido Blanco, cuando cumple las tres hipótesis básicas tradicionales:

- Media nula
- Varianza constante
- Covarianza nula entre errores correspondientes a observaciones diferentes

La ecuación general de $AR(p)$ es:

$$Y_t = \phi_0 + \phi_1 + Y_{t-1} + \phi_2 Y_{t-2} + \ldots\ldots + \phi_p Y_{t-p} + a_t$$

La forma abreviada es:

$$\phi_1(L) Y_t = \phi_0 + a_t$$

Donde $\phi_1(L)$ es el operador polinomial de retardos.

Medias móviles: es aquel modelo que describe el valor de una determinada variable en un periodo de tiempo t en una función de un término independiente y una sucesión de errores correspondientes a periodos precedentes, ponderados convenientemente. Estos modelos se denotan como $MA(q)$ y la ecuación general es:

$$Y_t = \mu + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}$$

Abreviando:

$$Y_t = \theta_q(L) a_t + \mu$$

Un modelo de medias móviles $MA(q)$ puede obtenerse a partir de un modelo autorregresivo realizando sustituciones.

Para que a un proceso estocástico estacionario pueda ser usado se deben cumplir las siguientes condiciones:

- No debe ser anticipante (Hipótesis de recursividad temporal), esto es que los valores de una variable en un momento t no dependerán de lo que ésta tome en t+j, siendo j cualquier valor superior a cero.
- El proceso debe ser invertible, esto es: que la correlación entre una variable y su pasado va reduciéndose a medida que hay mas distancia en el tiempo, justo en donde se esta considerando la correlación (proceso ergódico), esto quiere decir que si se especifica una variable en función de ciertos coeficientes que determinen la correlación con valores pasados de ella misma, los valores deberán ser menores a uno porque de lo contrario el proceso de valores infinitos seria explosivo.

Existen otros modelos que se usan para simulación en contextos con información histórica como son: Modelos de vectores autorregresivos (VAR), Modelos de vectores de corrección de error (VEC), Modelos autorregresivos y condicionales heteroscedásticos (ARCH) que no se verán en este trabajo pero que junto con los modelos ARIMA se pueden considerar dentro de un mismo contexto de simulación y pronóstico de resultado de comportamientos de fenómenos matemáticos.

Algo muy importante en la ciencia estadística es que se debe hacer una explicación de lo que para el caso de los procesos estocásticos y/o estudio de las series de tiempo se conoce como una realización. Esta característica es el resultado de los datos que se presentan por única vez y que no es posible volver a repetir el proceso o fenómeno. Además la realización es posible denominarla como una muestra de una población como es el caso de la estadística inferencial y además usarla para hacer el análisis para el planteamiento del modelo en Box-Jenkins y finalmente pronosticar, cuando se cumplen determinadas características de estacionariedad en la serie que se describe.

La función de la media es constante en el tiempo esto se puede ver en la gráfica ya que se distribuyen los valores de la serie de manera horizontal, para llegar a esto es necesario aplicar logaritmos y hacer iteraciones de ser necesario.

Homoscedasticidad, esto es que la varianza es constante en el tiempo esto gráficamente se verifica si los valores o puntos en la grafica se mantienen dentro de un rango constante muy cerca de la media, esto se logra usando también logaritmos y aplicación de primeras diferencias entre otros métodos.

La función de autocorrelación no depende de un tiempo t especifico de todo el universo de la serie en cuestión sino más bien de un intervalo k de entre dos variables, la actual y una anterior que puede parecerse a la cadena de Markov o proceso markoviano descrito anteriormente, esto se presenta como:

$$E[Y_t] = \mu_t$$

$$\sigma_t^2 = Var\ (y_t)\ = E[y_t - \ \mu_t]^2$$

$$\gamma_t = Cov(Y_t - Y_s) = E[(y_t - \mu_t)(y_s - \mu_s)]$$

El principio de estacionariedad en sentido amplio o débil es: que la media sea constante, que la varianza sea constante y que la covarianza o correlación no dependa del tiempo t sino de un estado anterior k, la existencia de estacionariedad en sentido estricto o fuerte puede ser demostrada con la estacionariedad en sentido amplio o débil, no a la inversa. (Videgaray, 2011 p. 49).

3.8 Metodología Box-Jenkins

George E. P. Box y Gwilym M. Jenkins desarrollaron una metodología para modelar series temporales de tipo univariante.

Es importante además mencionar que la metodología Box-Jenkins tiene como principal objetivo hacer una separación entre los datos predecibles que se encuentran en el conjunto de los datos observados, de la parte que está compuesta por los datos totalmente aleatorios o no predecibles.

Así, podemos decir que para la parte predecible se usan criterios para llegar a ella denominados filtros, Autorregresivos (AR), Integrados (I) y de Medias móviles (MA).

Por otro lado la parte no predecible se reduce a una parte no significativa también conocido como Ruido blanco y en consecuencia no influye en el resultado final, ya que la proporción debería ser mínima.

Esta metodología se basa en dos principios o filosofías fundamentales:

- Principio de parsimonia: Elegir siempre el modelo más sencillo y representativo en cuanto a los datos. En resumen es que no se deben agregar al modelo elementos que no están plenamente justificados. William de Occam
- Principio de mejoramiento iterativo: se inicia con un modelo sencillo e ir mejorando con iteraciones sucesivas el modelo mismo. Iterando un diagrama de flujo.

(Videgaray, 2011 p. 23)

Las etapas de la metodología son las siguientes:

1 **Recolección de datos**: se recomienda disponer de 50 o más datos y en el caso de series mensuales, es habitual usar un rango de entre seis y diez años completos de información.

2 **Representación gráfica de la serie**: para emitir criterios en relación a la estacionariedad y estacionalidad de la serie es de gran utilidad disponer de un gráfico. Es común usar medias y desviaciones típicas por subperiodo para juzgar en relación a la estacionariedad de la serie.

3 **Transformación previa de la serie y eliminación de la tendencia**: la transformación logarítmica es necesaria para series no estacionarias en varianza y es muy frecuente en series con dispersión relativamente constante en el tiempo.

4 **Identificación del modelo**: consiste en determinar el tipo de modelo más adecuado para la serie, es decir el orden de los procesos autorregresivos y de medias móviles de las componentes regular y estacional. Técnicamente esta decisión se tomará en base a las funciones de las gráficas del correlograma de la función autocorrelación (FAC) y la función autocorrelación parcia (FACP)

5 **Estimación de los coeficientes del modelo**: Una vez seleccionado el modelo, se procede a la estimación de sus parámetros, dado que se trata de un procedimiento iterativo de cálculo, pueden sugerirse valores iniciales.

6 **Contraste de validez conjunta del modelo**: Utilizaremos diversos procedimientos para valorar el modelo o modelos inicialmente seleccionados sean estos: Contraste de significación de parámetros, Covarianzas entre estimadores, Coeficiente de correlación, Suma de cuadrados de errores, etc.

7 **Análisis detallado de los errores**: Las diferencias históricas entre los valores reales y los estimados por el modelo constituyen una fuente de especial interés para una valoración final del modelo. Deberá comprobarse un comportamiento no sistemático de los mismos, así como analizarse la posible existencia de errores especialmente significativos.

8 **Selección del modelo y predicción**: En función de las etapas anteriores se selecciona el modelo y se utilizará como forma inicial de predicción.

En el punto 4 para la FAC se va a estimar si es decreciente infinita esto con ayuda de la observación directa del correlograma y también con ayuda del intervalo de confianza, se deberá probar la siguiente hipótesis.

$$H_0: \rho_k = 0$$

y

$$H_a: \rho_k \neq 0$$

Hipótesis nula e Hipótesis alternativa respectivamente para H_0 la autocorrelacion correspondiente al intervalo k es estadísticamente insignificante esto quiere decir que no hay autocorrelacion entre el intervalo que va de Y_t y Y_{t-k} .

Para el caso contrario se rechaza H_0 ; se acepta H_a debido a que $\rho_k$es significativa.
Para el intervalo de confianza se tiene que $(1-\alpha)100\%$ con un valor para t-de-student $\alpha = 0.05$
Lo cual nos da como resultado el valor aproximado a 2, y de esta forma quedan de la siguiente forma las hipótesis:

Aceptar H_0 si

$$H_0 \; si \; |t_{rk}| < 2$$

En caso contrario rechazar H_0

Para el caso de la FACP se tiene algo semejante

H_0: $\rho_{kk} = 0$

y

H_a: $\rho_{kk} \neq 0$

(Videgaray, 2011 págs. 108,116)

Las dos funciones anteriores se verán soportadas fuertemente por los gráficos que genera el software GRETL.

En 6 se identifican los parámetros aquí se estima el orden del tipo de filtro que se usará, haciendo uso del método de máxima verosimilitud o el método de mínimos cuadrados, cabe aclarar que el software que usaremos en este trabajo será el de máxima verosimilitud debido a que es mejor que el de mínimos cuadrados. (Videgaray, 2011 págs. 119,120)

Esta regla se sigue para determinar el nivel de filtro ya sea: autorregresivo AR(p), integrado i/o de medias móviles MA(q), al observar las gráficas de FAC y FACP.

En la figura 3.8.1 se muestran los criterios generales para determinar el nivel de la función autorregresivo, integrados y de medias móviles, aunque existen más criterios, para este trabajo solo se usan estos tres métodos o filtros para Box-Jenkins.

TIPO	FAS	FAP
AR(p)	Muchos coeficientes no nulos	Primeros p no nulos, resto 0
MA(q)	Primeros p no nulos, resto 0	Muchos coeficientes no nulos
ARMA(p,q)	Muchos coeficientes no nulos	Muchos coeficientes no nulos

Figura 3.8.1

Existen otros modelos que se usan para simulación en contextos con información histórica como son: Modelos de vectores autorregresivos (VAR), Modelos de vectores de corrección de error (VEC), Modelos autorregresivos y condicionales heteroscedásticos (ARCH) que no se verán en este trabajo pero que junto con los modelos ARIMA se pueden considerar dentro de un mismo contexto de simulación y pronóstico de resultado de comportamientos de fenómenos matemáticos, es por esta razón que no se mencionan los criterios de ellos.

3.9 Software GRETL

El código fuente original del software GRETL parte del programa ESL *Econometrics Software Library* desarrollado por el profesor de la universidad de San Diego California Ramu Ramanathan, creado bajo la licencia de código abierto GNU General Public Licence.

Estos autores también han contribuido. El profesor Fiorentini, Calzolari y Panattoni; el código para generar los p-values para el test Dickey–Fuller, el test es trabajo de James MacKinnon.

Otros autores de diferentes nacionalidades son: Ignacio Díaz-Emparanza (España), Michel Robitaille and Florent Bresson (Francia), Cristian Rigamonti (Italia), Tadeusz Kufel and Pawel Kufel (Polonia), Markus Hahn and Sven Schreiber (Alemania), Hélio Guilherme and Henrique Andrade (Portugal), Susan Orbe (pais Basco), Talha Yalta (Turquia) and Alexander Gedranovich (Rusia).
GRETL está disponible para sistemas operativos como: Linux, MS Windows, Mac OS X,

"*Sophisticated Gretl offers a full range of least-squares based estimators, either for single equations and for systems, including vector autoregressions and vector error correction models. Several specific maximum likelihood estimators (e.g. probit, ARIMA, GARCH) are also provided natively; more advanced estimation methods can be implemented by the user via generic maximum likelihood or nonlinear GMM.*" (Cottrell, 2014 p. 1)

GRETL es una aplicación muy diversa que ofrece varias soluciones estadísticas pero lo más importante es que tiene los filtros que son necesarios para estimar y pronosticar con la metodología Box-Jenkins.

Originalmente este programa fue desarrollado en Linux pero actualmente se encuentra disponible para otras plataformas como son Windows y Mac OS y para otras plataformas basadas en UNIX BSD.

Como GRETL es un software de código abierto el usuario puede desarrollar los cambios al programa que le sean de utilidad.

Algunos de los países en los que está disponible el lenguaje son: Inglaterra, Francia, Italia, España, Polonia, Portugal, Alemania, País Vasco, Turquía, Rusia, Albania, o Grecia dependiendo del tipo de teclado de la computadora.

El código original está basado en el programa ESL Econometrics Software Library desarrollado por el Profesor Ramu Ramanathan de la universidad de San Diego California, gracias a él, su código está disponible bajo la licencia GNU General Public Licence.

Uno de los métodos que contiene GRETL es el de mínimos cuadrados ordinarios, Ordinary Least Squares (OLS). Por sus siglas en ingles. (Cottrell, 2014 p. 5)

Menú

Figura 3.9.1

Este es el menú de GRETL que de manera general contiene los campos que casi todos los programas tienen de Archivo, Herramientas, Datos, Versión etc. Con ligeras variantes.

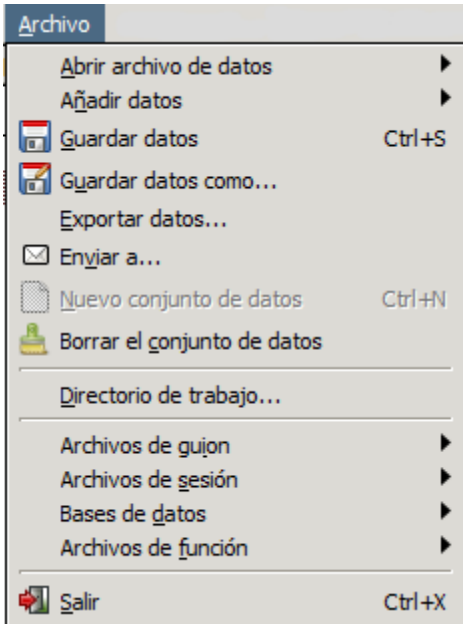

Figura 3.9.2

En la imagen 3.9.2 se despliega campos básicos para mantenimiento de archivos así como para exportar datos.

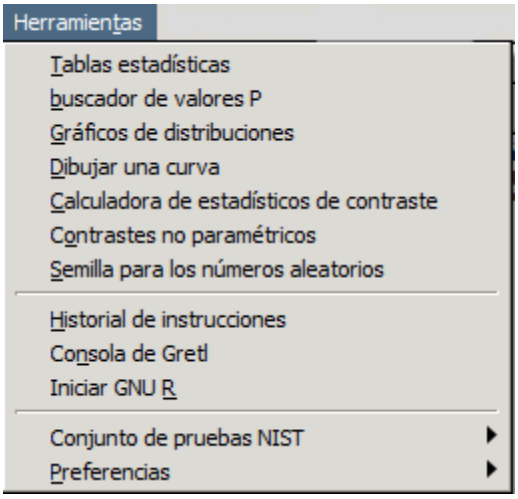

Figura 3.9.3

En la figura 3.9.3 se muestran los campos que pertenecen a herramientas, estos van desde las tablas estadísticas hasta la consola GRETL para ejecutar comandos en línea

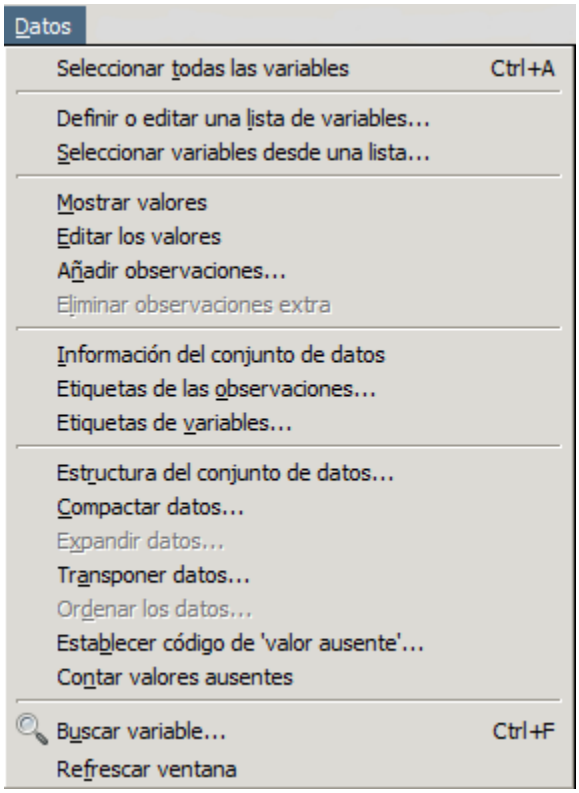

Figura 3.9.4

En figura 3.9.4 se muestran los campos que pertenecen a datos principalmente para el manejo de valores y uso de etiquetas.

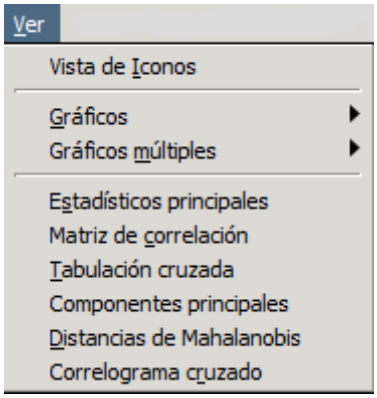

Figura 3.9.5

Figura 3.9.5 se presenta el campo ver para configuración de vista de salidas y matriz de correlación, entre otros.

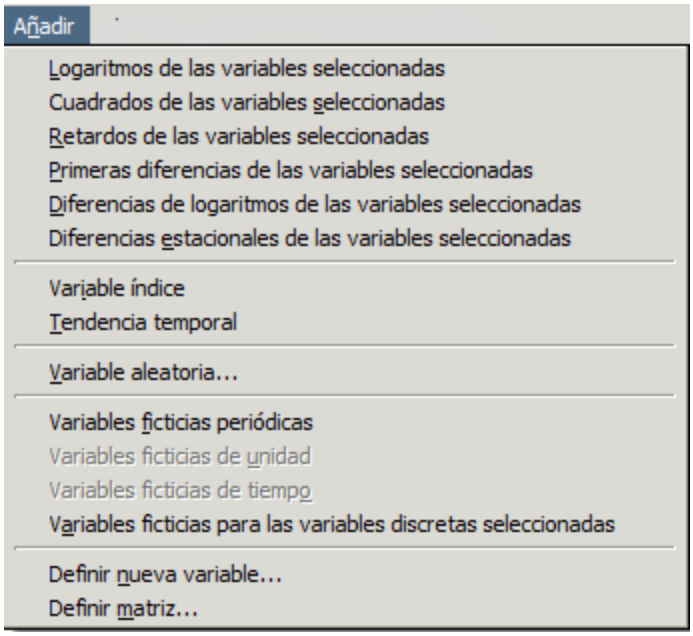

Figura 3.9.6

Figura 3.9.6 añadir es para el manejo de funciones, como transformación logarítmica de funciones, retardo de variables, variable índice etc.

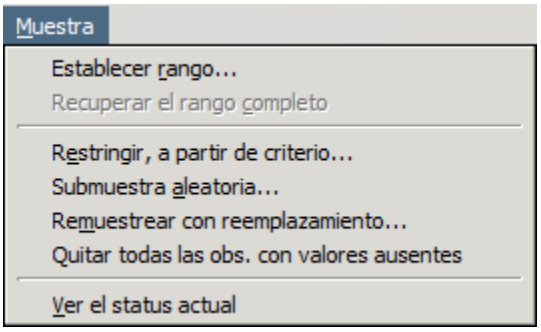

Figura 3.9.7

En 3.9.7 se tienen muestras que principalmente sirven para el manejo de rangos y criterios.

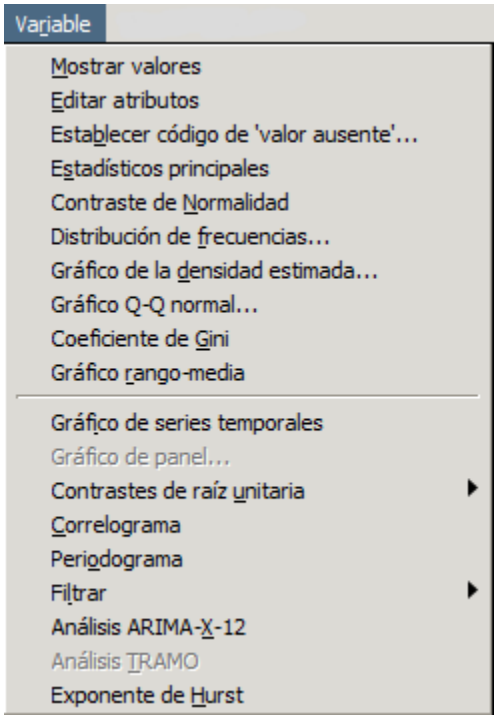

Figura 3.9.8

3.9.8 enfocado al manejo y uso de variables para las diferentes aplicaciones estadísticas.

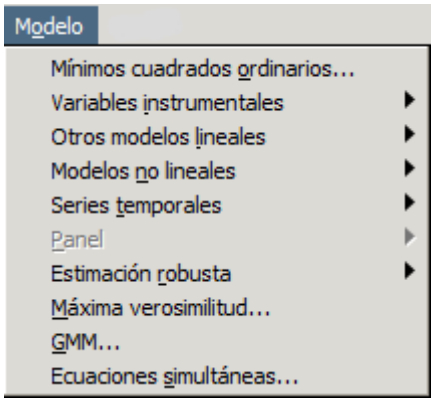

Figura 3.9.9

En 3.9.9 tenemos Modelo | Series Temporales | ARIMA que es en donde vamos a generar el modelo Box-Jenkins desde GRETL.

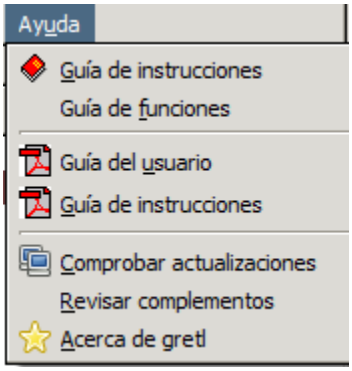

Figura 3.9.10

En 3.9.10 está la ayuda para el programa, que es de mucho apoyo ya que está en su mayoría en español.

La parte inferior tiene un menú más sencillo que tiene acciones como la de ejecutar la calculadora, el editor de texto, abrir la consola GRETL, vista de iconos de sesión, paquetes de funciones, la guía de usuario, guía de instrucciones gráficos X-Y, mínimos, cuadrados ordinarios MCO, base de datos GRETL y la opción de abrir un conjunto de datos. (Cottrell, 2014 págs. 7-11)

Figura 3.9.11

“The default estimation method for ARMA models is exact maximum likelihood estimation (under the assumption that the error term is normally distributed), using the Kalman filter in conjunction with the BFGS maximization algorithm.” (Cottrell, 2014 p. 212)

El método por default en GRETL es el de máxima verosimilitud a diferencia de otras aplicaciones que hacen el cálculo con los mínimos cuadrados.

“Gretl supports forecasting on the basis of ARMA models using the method set out by Box and Jenkins (1976).2 The Box and Jenkins algorithm produces a set of integrated AR coefficients which take into account any differencing of the dependent variable (seasonal and/or non-seasonal) in the ARIMA context, thus making it possible to generate a forecast for the level of the original variable. By contrast, if you first difference a series manually and then apply ARMA to the differenced series, forecasts will be for the differenced series, not the level...” (Cottrell, 2014 p. 214)

Capitulo IV Desarrollo de investigación

En este capítulo se lleva a cabo la investigación haciendo uso de un programa llamado GRETL con el que se sustituye a otros más conocidos pero que algunos requieren licencia para usarse. Así también se exponen los resultados, las conclusiones y los trabajos futuros.

Capítulo IV Desarrollo de investigación

4.1 Solución del modelo con el programa GRETL

De acuerdo a lo descrito en la fundamentación del tema, el problema de simular el crecimiento de bases de datos de un servidor Oracle es complejo debido a la dificultad de tener físicamente un servidor Exadata demasiado costoso.

Esto da como resultado la necesidad de simular el pronóstico de crecimiento, usando la metodología Box-Jenkins con ayuda del software de simulación de distribución libre GRETL.

Estos son los datos que se tienen para la simulación del crecimiento de datos en GRETL con la metodología Box-Jenkins. Los valores por mes van desde el año 1984 al 2013 tomando como unidad de medida el Mega Byte para almacenar cadenas de datos binarias compuestas por 0 y 1.

El Mega Byte es el resultado de usar múltiplos de la unidad más básica de medida de almacenamiento como es el Bite. Luego le sigue el K Byte y después el megabyte

Obs , MBytes		
1984:01 , 978	1994:01 , 54863	2004:01 , 96955
1984:02 , 1142	1994:02 , 55996	2004:02 , 97078
1984:03 , 1335	1994:03 , 56653	2004:03 , 98413
1984:04 , 1908	1994:04 , 57585	2004:04 , 99447
1984:05 , 2127	1994:05 , 57927	2004:05 , 99756
1984:06 , 2564	1994:06 , 58368	2004:06 , 100261
1984:07 , 3351	1994:07 , 58466	2004:07 , 100296
1984:08 , 4412	1994:08 , 58899	2004:08 , 100698
1984:09 , 4536	1994:09 , 59035	2004:09 , 101252
1984:10 , 4772	1994:10 , 59132	2004:10 , 101527
1984:11 , 5246	1994:11 , 59930	2004:11 , 102146

1984:12 , 5276	1994:12 , 60638	2004:12 , 102181
1985:01 , 5469	1995:01 , 60884	2005:01 , 102557
1985:02 , 5622	1995:02 , 61343	2005:02 , 103074
1985:03 , 6031	1995:03 , 61738	2005:03 , 103188
1985:04 , 6214	1995:04 , 61805	2005:04 , 103429
1985:05 , 6487	1995:05 , 62311	2005:05 , 104394
1985:06 , 6553	1995:06 , 62671	2005:06 , 105462
1985:07 , 6903	1995:07 , 62840	2005:07 , 106201
1985:08 , 7110	1995:08 , 63137	2005:08 , 106601
1985:09 , 7138	1995:09 , 63900	2005:09 , 106806
1985:10 , 8064	1995:10 , 64072	2005:10 , 107153
1985:11 , 8069	1995:11 , 64082	2005:11 , 107968
1985:12 , 8279	1995:12 , 64090	2005:12 , 110392
1986:01 , 8395	1996:01 , 64777	2006:01 , 111724
1986:02 , 8458	1996:02 , 64995	2006:02 , 112147
1986:03 , 8807	1996:03 , 65043	2006:03 , 112176
1986:04 , 8966	1996:04 , 65453	2006:04 , 112343
1986:05 , 9060	1996:05 , 65887	2006:05 , 112809
1986:06 , 9225	1996:06 , 66054	2006:06 , 114922
1986:07 , 9236	1996:07 , 66439	2006:07 , 115301
1986:08 , 9810	1996:08 , 66475	2006:08 , 116273
1986:09 , 9934	1996:09 , 66918	2006:09 , 117355
1986:10 , 10033	1996:10 , 66987	2006:10 , 117559
1986:11 , 11060	1996:11 , 67239	2006:11 , 117566
1986:12 , 11230	1996:12 , 68482	2006:12 , 117615
1987:01 , 11697	1997:01 , 69256	2007:01 , 117671
1987:02 , 12261	1997:02 , 70029	2007:02 , 117818
1987:03 , 12380	1997:03 , 70515	2007:03 , 118013
1987:04 , 13333	1997:04 , 71077	2007:04 , 118219
1987:05 , 14035	1997:05 , 71084	2007:05 , 118621
1987:06 , 14115	1997:06 , 71109	2007:06 , 119280
1987:07 , 14938	1997:07 , 71385	2007:07 , 119454
1987:08 , 15714	1997:08 , 71642	2007:08 , 121092
1987:09 , 16115	1997:09 , 71847	2007:09 , 121166

1987:10 , 17140	1997:10 , 71911	2007:10 , 121547
1987:11 , 17623	1997:11 , 71932	2007:11 , 121583
1987:12 , 19020	1997:12 , 72137	2007:12 , 122084
1988:01 , 19256	1998:01 , 72804	2008:01 , 122459
1988:02 , 19274	1998:02 , 72870	2008:02 , 122720
1988:03 , 19697	1998:03 , 72876	2008:03 , 123503
1988:04 , 19701	1998:04 , 73278	2008:04 , 123871
1988:05 , 19816	1998:05 , 73725	2008:05 , 124132
1988:06 , 19966	1998:06 , 73862	2008:06 , 124188
1988:07 , 20062	1998:07 , 73984	2008:07 , 124867
1988:08 , 20243	1998:08 , 74160	2008:08 , 125562
1988:09 , 20306	1998:09 , 74565	2008:09 , 125602
1988:10 , 20467	1998:10 , 74853	2008:10 , 125779
1988:11 , 20613	1998:11 , 74898	2008:11 , 126520
1988:12 , 20987	1998:12 , 75295	2008:12 , 126587
1989:01 , 21314	1999:01 , 75446	2009:01 , 126764
1989:02 , 21692	1999:02 , 75451	2009:02 , 127187
1989:03 , 22026	1999:03 , 75616	2009:03 , 127866
1989:04 , 22763	1999:04 , 75987	2009:04 , 127874
1989:05 , 23385	1999:05 , 76107	2009:05 , 128592
1989:06 , 25493	1999:06 , 77204	2009:06 , 129045
1989:07 , 28253	1999:07 , 77245	2009:07 , 129550
1989:08 , 28597	1999:08 , 77511	2009:08 , 130319
1989:09 , 28854	1999:09 , 77766	2009:09 , 130810
1989:10 , 29540	1999:10 , 78172	2009:10 , 131209
1989:11 , 30778	1999:11 , 78851	2009:11 , 131925
1989:12 , 31641	1999:12 , 79002	2009:12 , 132249
1990:01 , 31924	2000:01 , 79085	2010:01 , 132367
1990:02 , 32066	2000:02 , 79260	2010:02 , 132491
1990:03 , 32396	2000:03 , 79653	2010:03 , 132538
1990:04 , 32791	2000:04 , 79736	2010:04 , 132668
1990:05 , 33528	2000:05 , 79875	2010:05 , 132712
1990:06 , 33639	2000:06 , 80467	2010:06 , 132861
1990:07 , 33863	2000:07 , 80604	2010:07 , 133694

1990:08 , 34942	2000:08 , 81054	2010:08 , 133919
1990:09 , 35166	2000:09 , 81072	2010:09 , 134234
1990:10 , 35551	2000:10 , 82439	2010:10 , 134678
1990:11 , 35811	2000:11 , 83254	2010:11 , 134715
1990:12 , 35864	2000:12 , 85531	2010:12 , 134758
1991:01 , 36317	2001:01 , 86211	2011:01 , 134807
1991:02 , 37180	2001:02 , 86510	2011:02 , 135007
1991:03 , 38121	2001:03 , 86541	2011:03 , 135266
1991:04 , 38811	2001:04 , 87023	2011:04 , 135496
1991:05 , 38914	2001:05 , 87072	2011:05 , 136642
1991:06 , 39014	2001:06 , 87152	2011:06 , 137117
1991:07 , 39723	2001:07 , 87216	2011:07 , 137543
1991:08 , 39782	2001:08 , 87262	2011:08 , 138150
1991:09 , 39827	2001:09 , 87424	2011:09 , 138213
1991:10 , 40082	2001:10 , 88015	2011:10 , 138309
1991:11 , 40222	2001:11 , 88617	2011:11 , 138580
1991:12 , 40309	2001:12 , 88777	2011:12 , 138695
1992:01 , 40432	2002:01 , 89441	2012:01 , 138952
1992:02 , 40513	2002:02 , 89683	2012:02 , 139624
1992:03 , 40656	2002:03 , 90105	2012:03 , 140002
1992:04 , 40875	2002:04 , 90211	2012:04 , 140350
1992:05 , 41488	2002:05 , 90236	2012:05 , 140828
1992:06 , 42533	2002:06 , 91440	2012:06 , 141262
1992:07 , 42645	2002:07 , 91595	2012:07 , 141711
1992:08 , 42863	2002:08 , 91900	2012:08 , 141962
1992:09 , 45111	2002:09 , 92041	2012:09 , 142240
1992:10 , 45384	2002:10 , 92042	2012:10 , 142479
1992:11 , 45547	2002:11 , 92430	2012:11 , 142603
1992:12 , 46737	2002:12 , 92527	2012:12 , 142884
1993:01 , 46739	2003:01 , 92638	2013:01 , 142914
1993:02 , 47043	2003:02 , 93727	2013:02 , 143217
1993:03 , 47050	2003:03 , 93811	2013:03 , 144153
1993:04 , 48731	2003:04 , 94100	2013:04 , 144192
1993:05 , 48033	2003:05 , 94188	2013:05 , 144641

1993:06 , 48136	2003:06 , 94197	2013:06 , 144972
1993:07 , 48483	2003:07 , 95349	2013:07 , 145768
1993:08 , 49018	2003:08 , 95362	2013:08 , 145999
1993:09 , 49318	2003:09 , 95492	2013:09 , 146629
1993:10 , 49182	2003:10 , 95500	2013:10 , 146729
1993:11 , 49492	2003:11 , 95648	2013:11 , 146943
1993:12 , 49721	2003:12 , 95754	2013:12 , 147016

Figura 4.1.1

Con los datos de 4.1.1 Se abre el archivo de texto llamado Crecimiento Base de Datos Radio Móvil Telcel 1984_2013 que contiene los datos anteriores en GRETL:

En a figura 4.1.2 se muestra la serie Bytes en GRETL

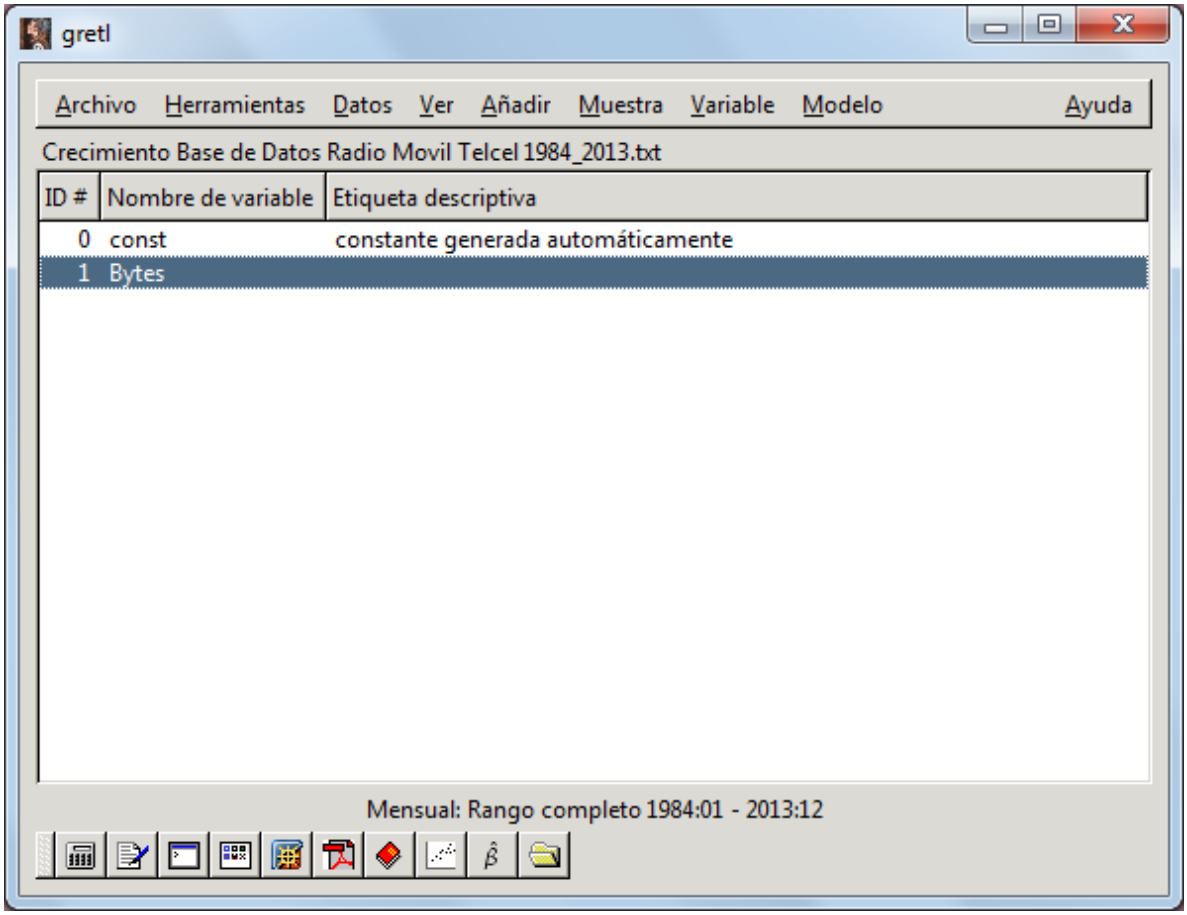

Figura 4.1.2

Se genera la grafica de series temporales y notamos que no es estacionaria Fig. 4.1.3

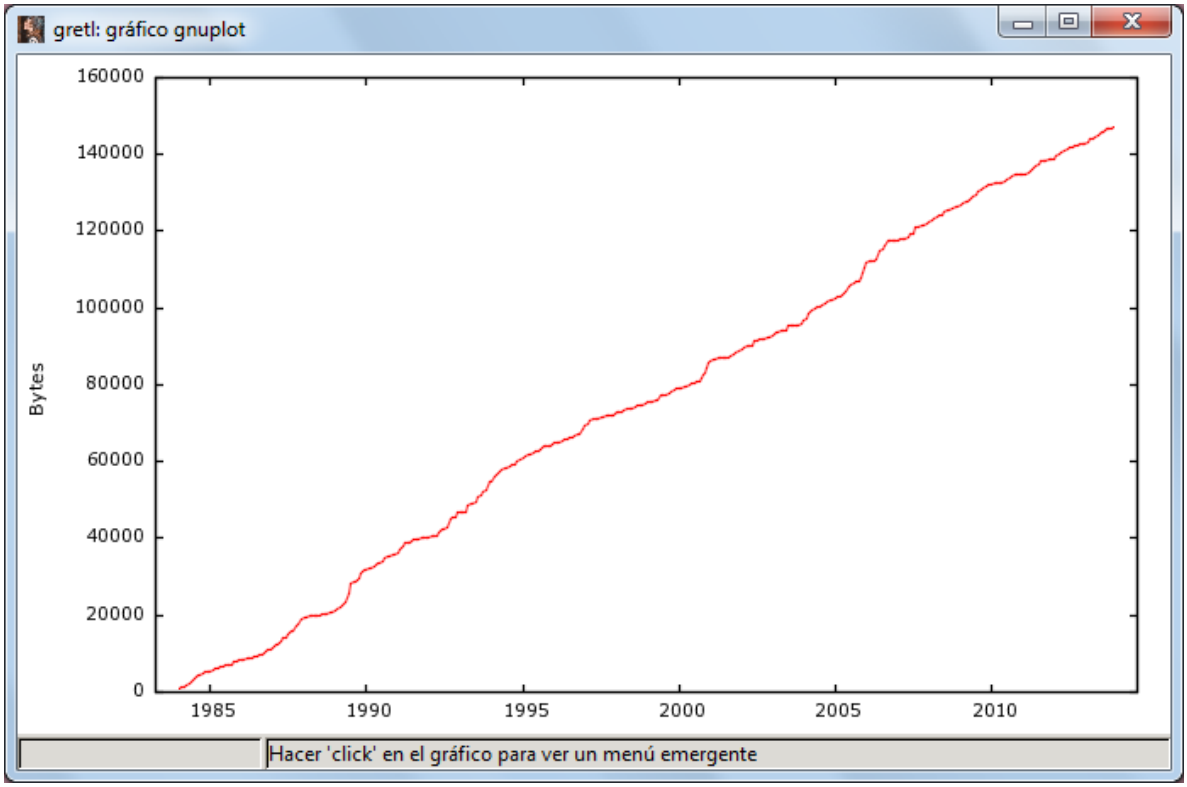

Figura 4.1.3

Se generan los estadísticos principales Fig. 4.1.4

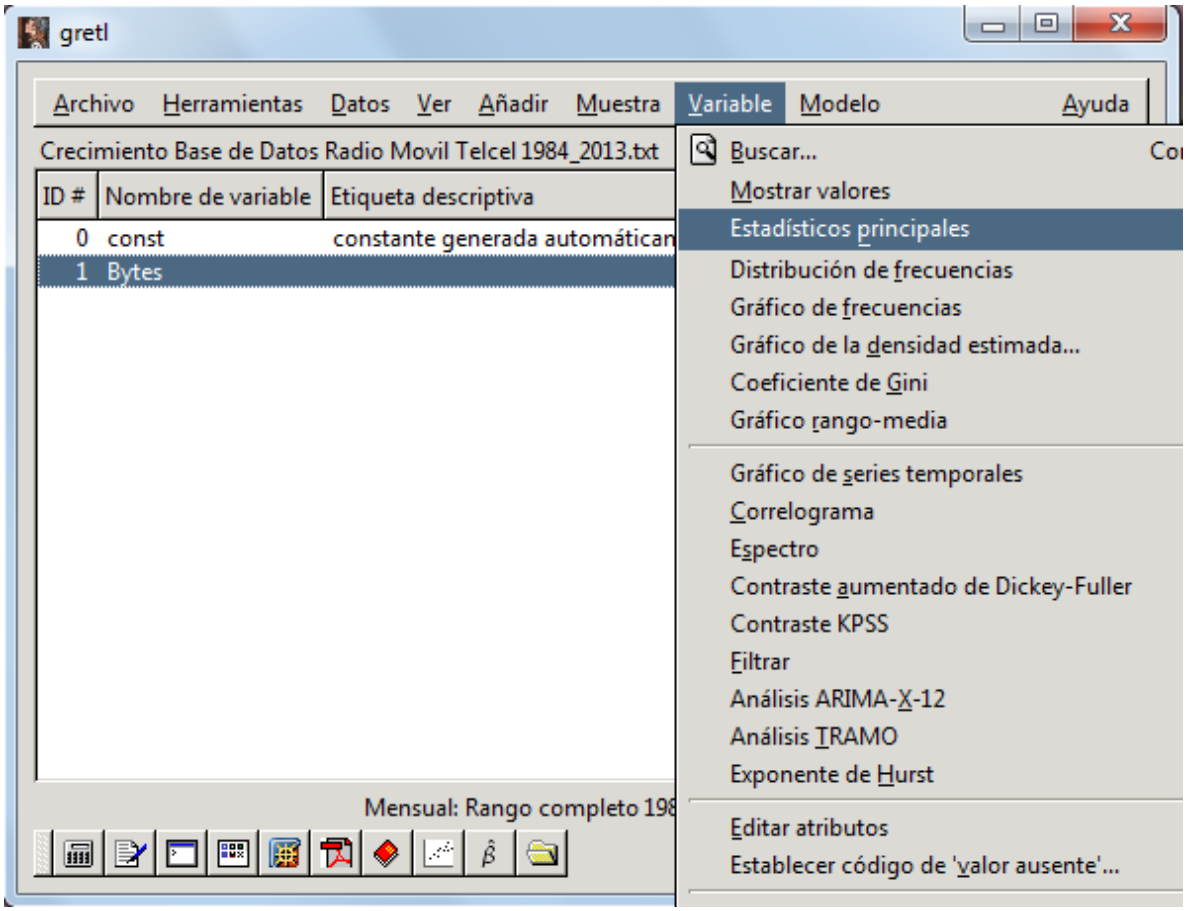

Figura 4.1.4

Aquí se presentan los estadísticos principales, media, varianza, desviación estándar, asimetría y curtosis fig. 4.1.5

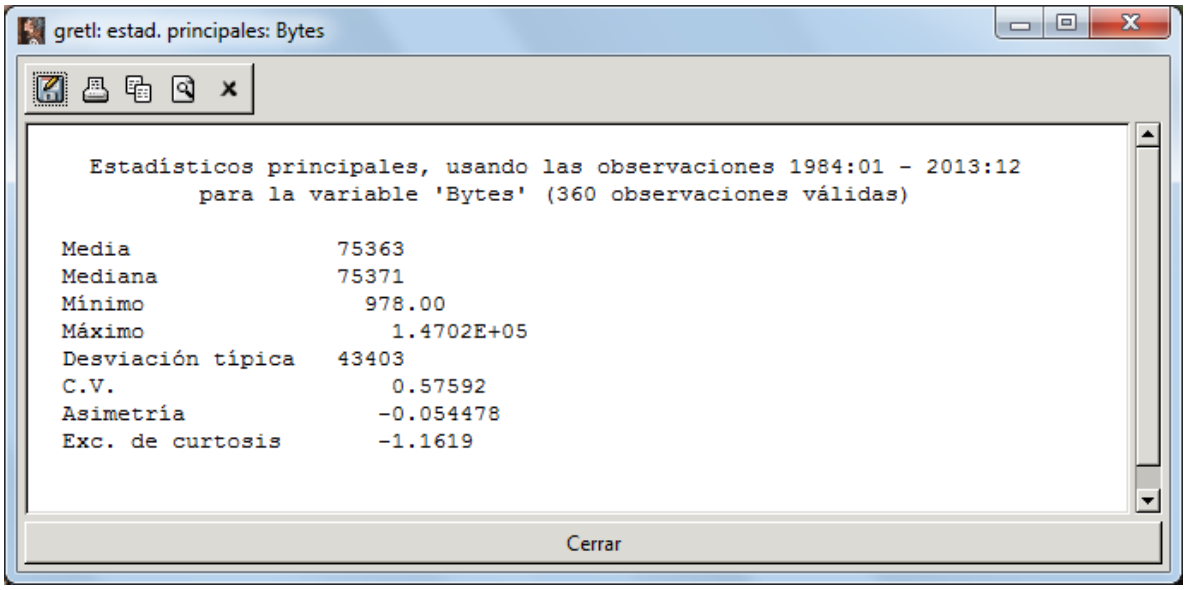

Figura 4.1.5

Se aprecia que no es una distribución normal debido a que tiene asimetría, generamos el grafico de frecuencia Normal para hacer una observación más precisa fig. 4.1.6

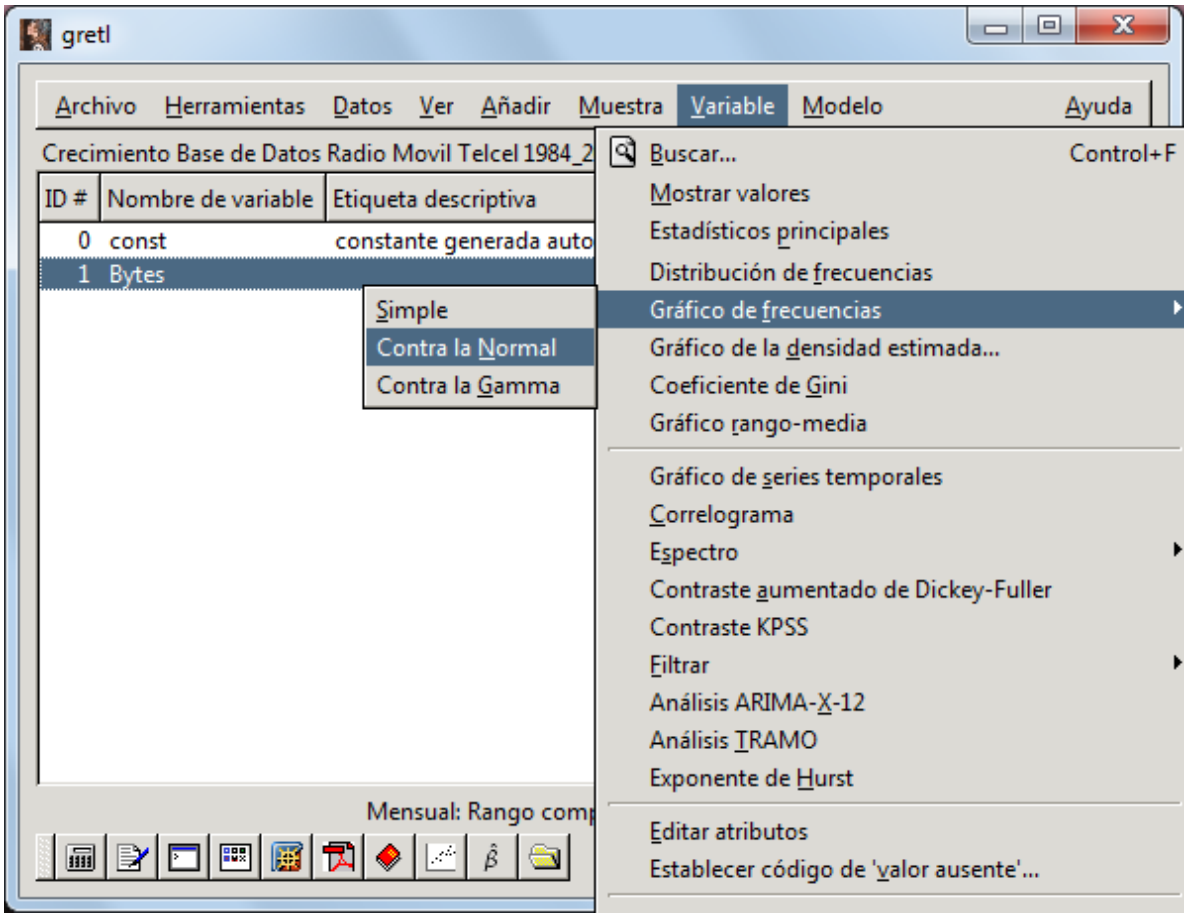

Figura 4.1.6

Se aprecia que el contraste de normalidad es muy simétrico y aceptable fig. 4.1.7

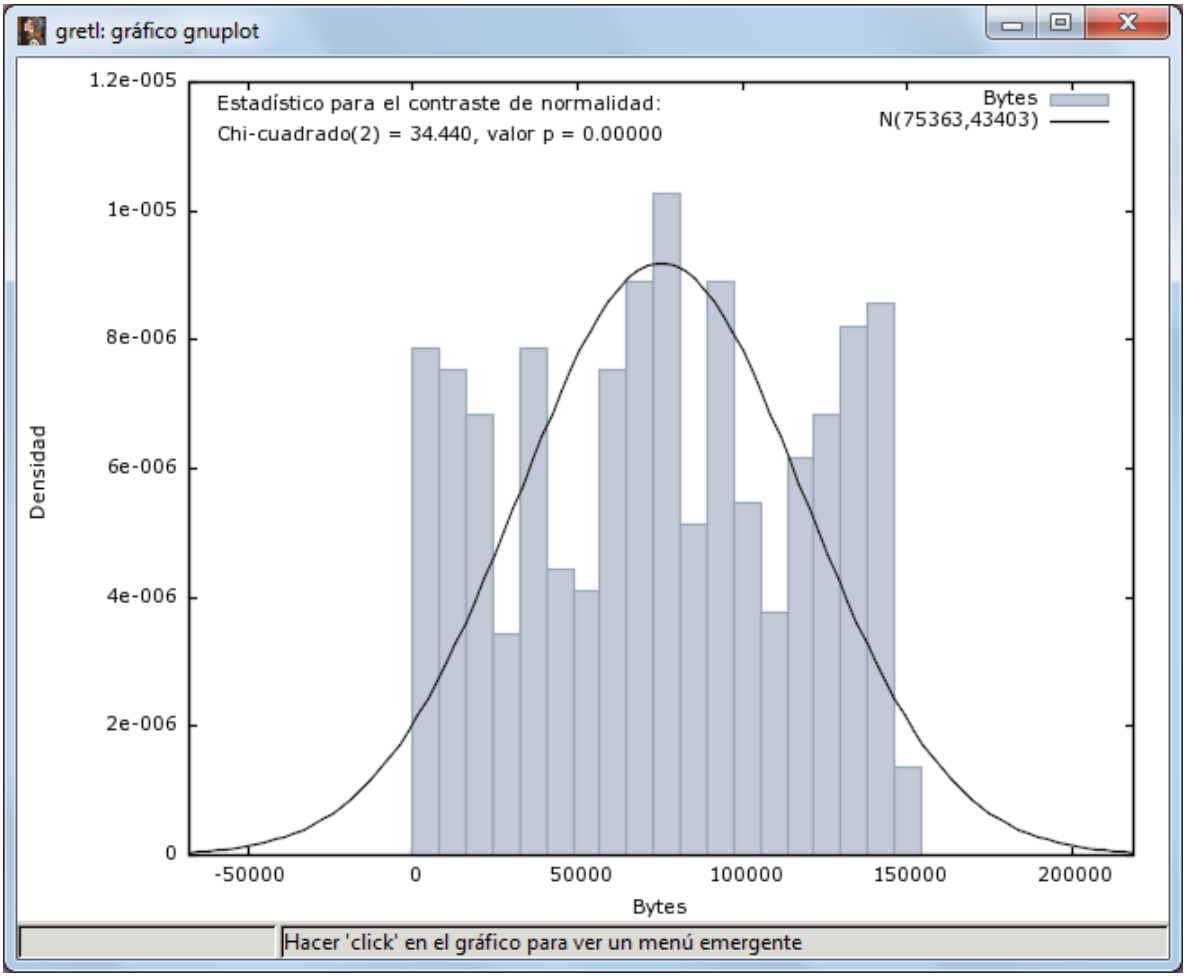

Figura 4.1.7

Se genera una serie de logaritmos fig. 4.1.8

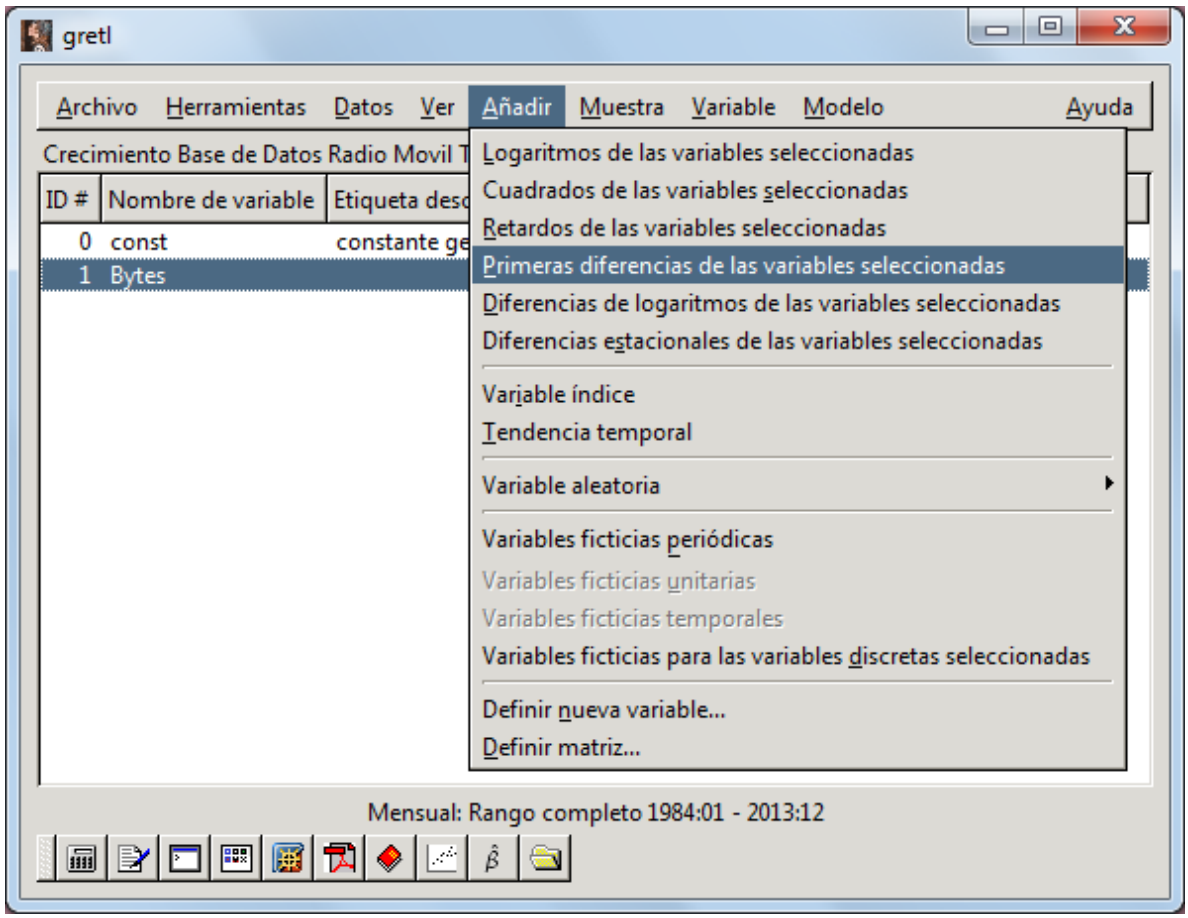

Figura 4.1.8

Se genera las primeras diferencias de la variable seleccionada y a partir de ella corremos el grafico de series temporales fig. 4.1.9

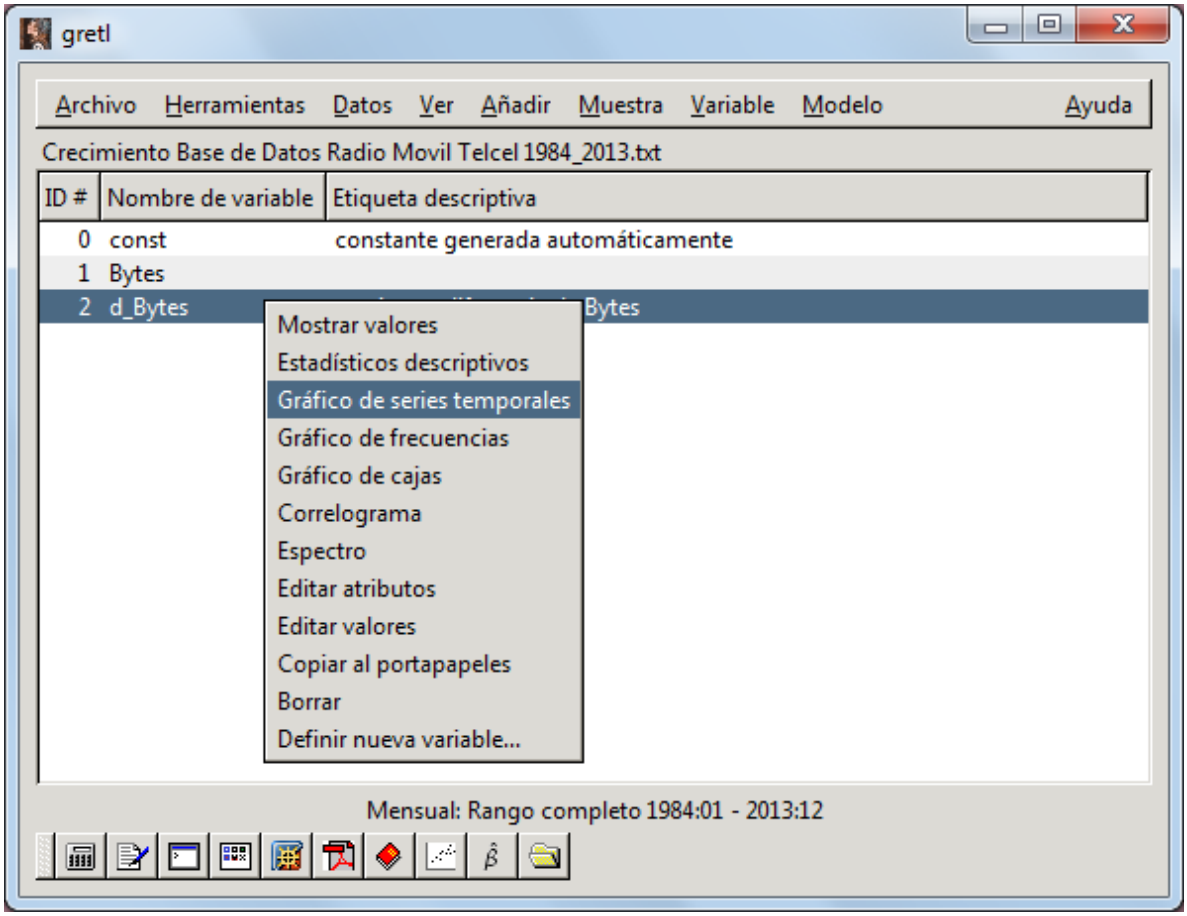

Figura 4.1.9

Se aprecia que la serie se ha hecho estacionaria fig. 4.1.10

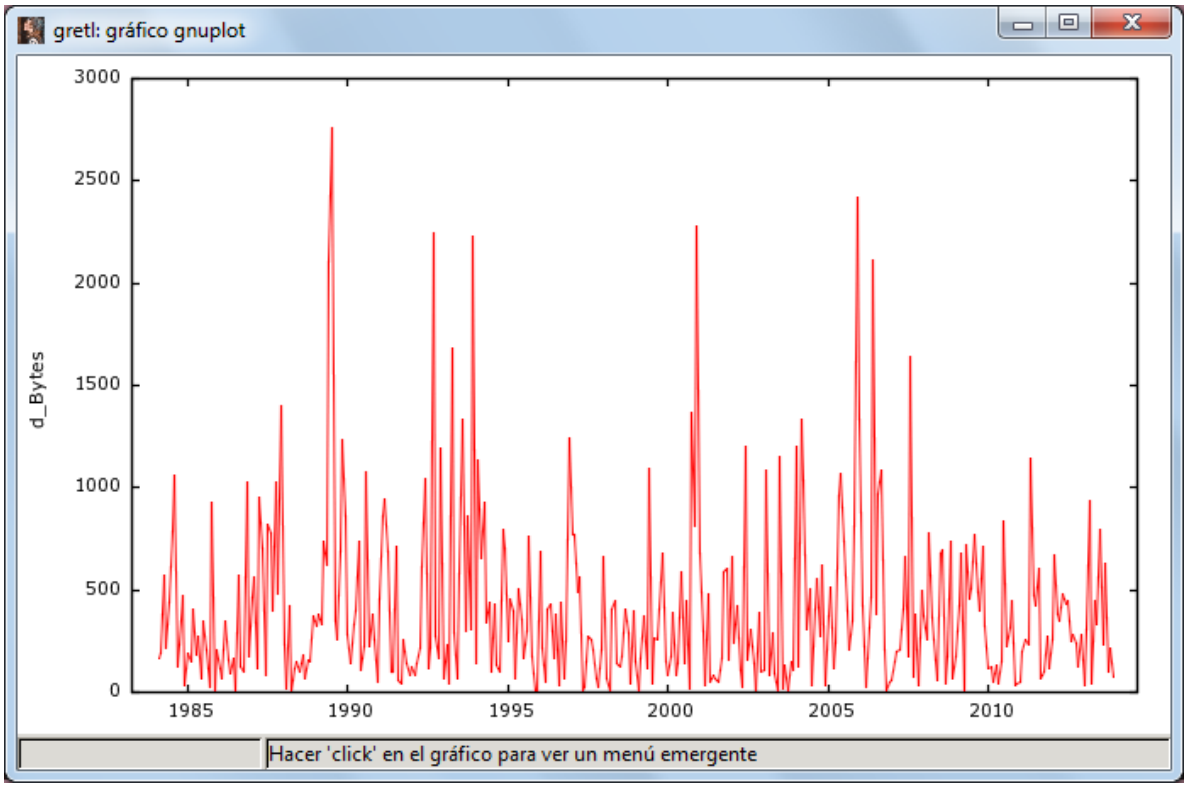

Figura 4.1.10

A partir de la serie original se genera un correlograma para ver la tendencia: se nota que en la FAC se observa una tendencia decreciente y la FACP tiene solo un componente significativo, así que una primera opción puede ser una AR(0) MA(1) o ARIMA(0,0,1) fig. 4.1.11

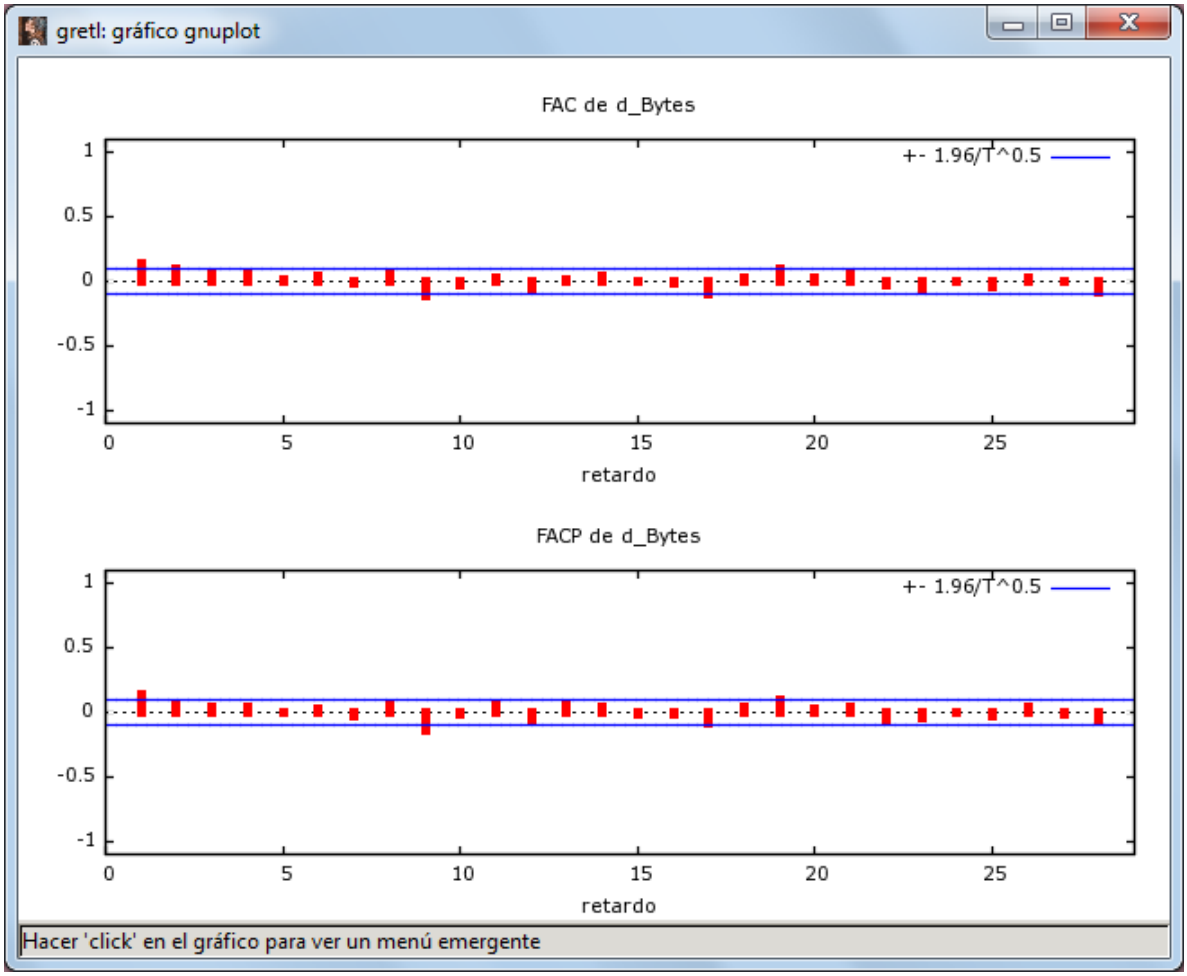

Figura 4.1.11

Se genera las diferencias de logaritmos de la serie seleccionada fig. 4.1.12

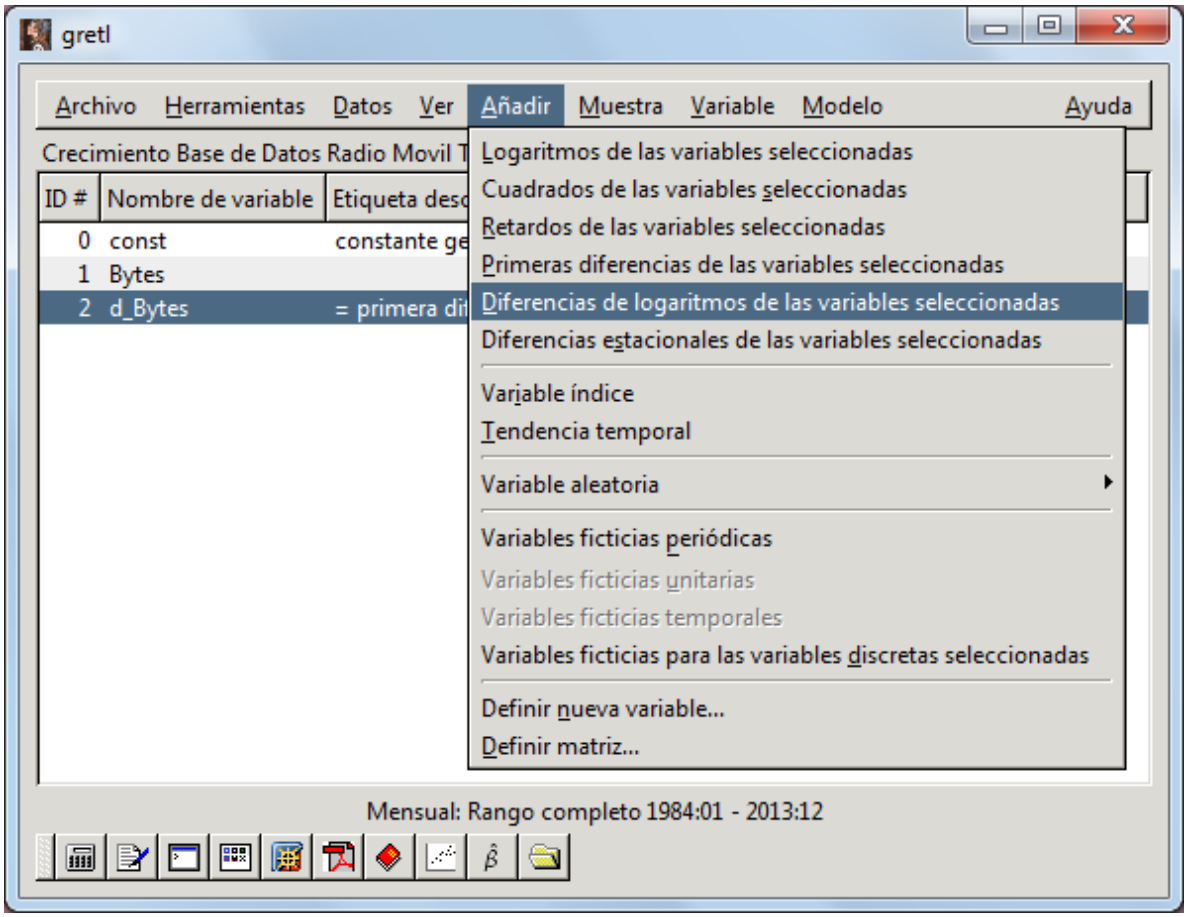

Figura 4.1.12

Se genera el gráfico de series temporales nueva mente 4.1.13

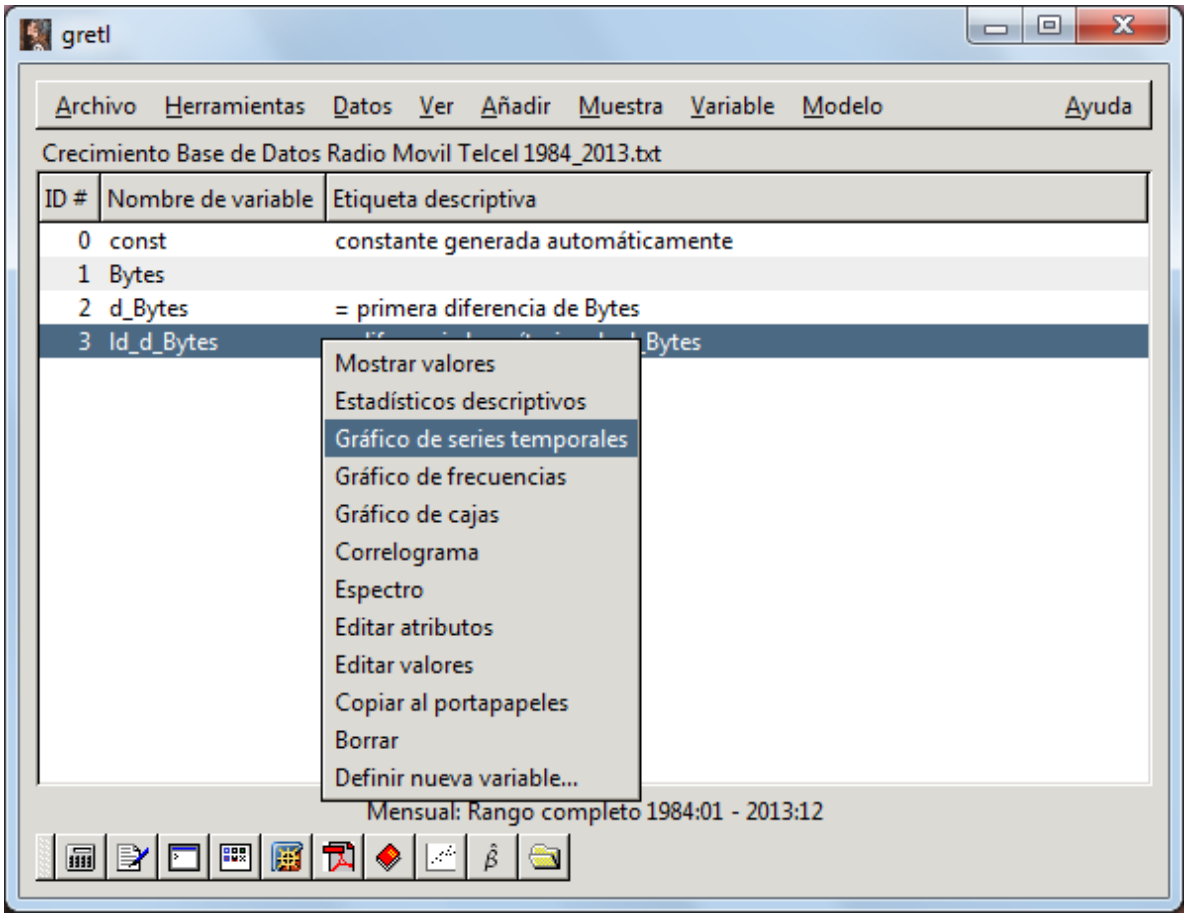

Figura 4.1.13

Grafico de series temporales estacionario fig. 4.1.14

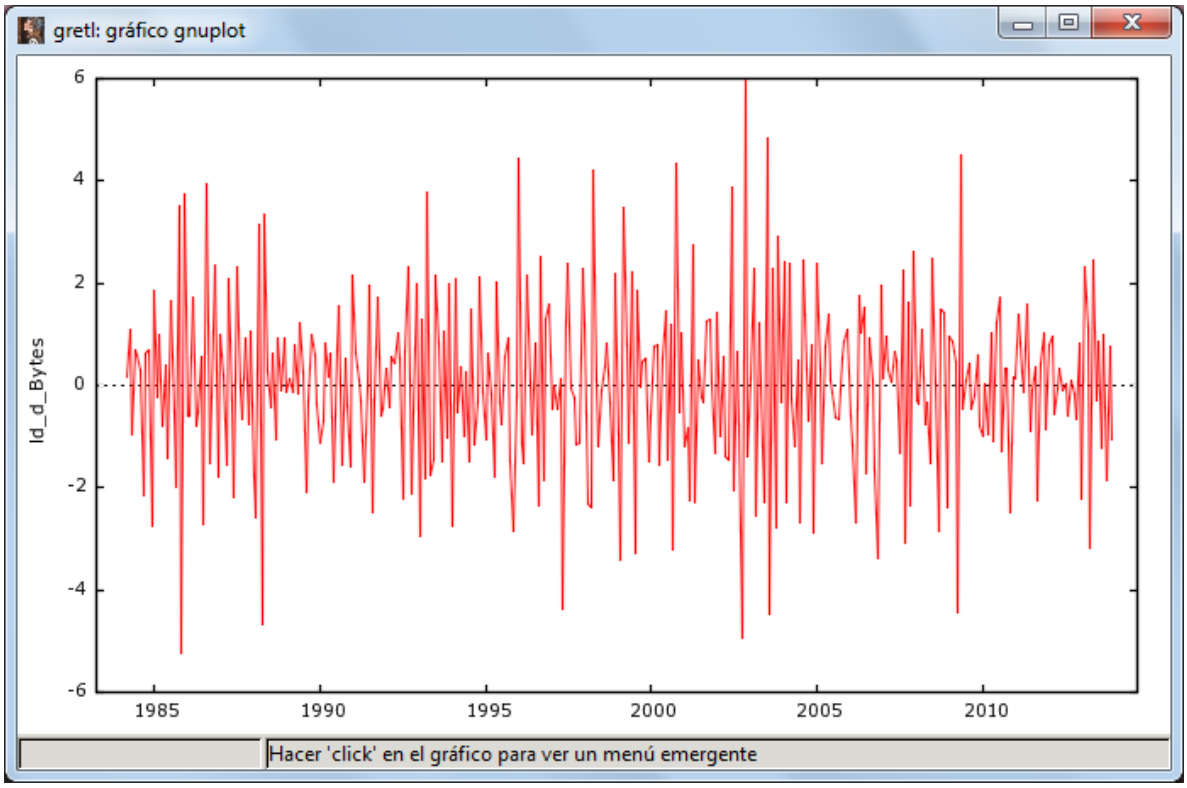

Figura 4.1.14

Correlograma con posibles ARIMA (2, 1,1), fig. 4.1.15

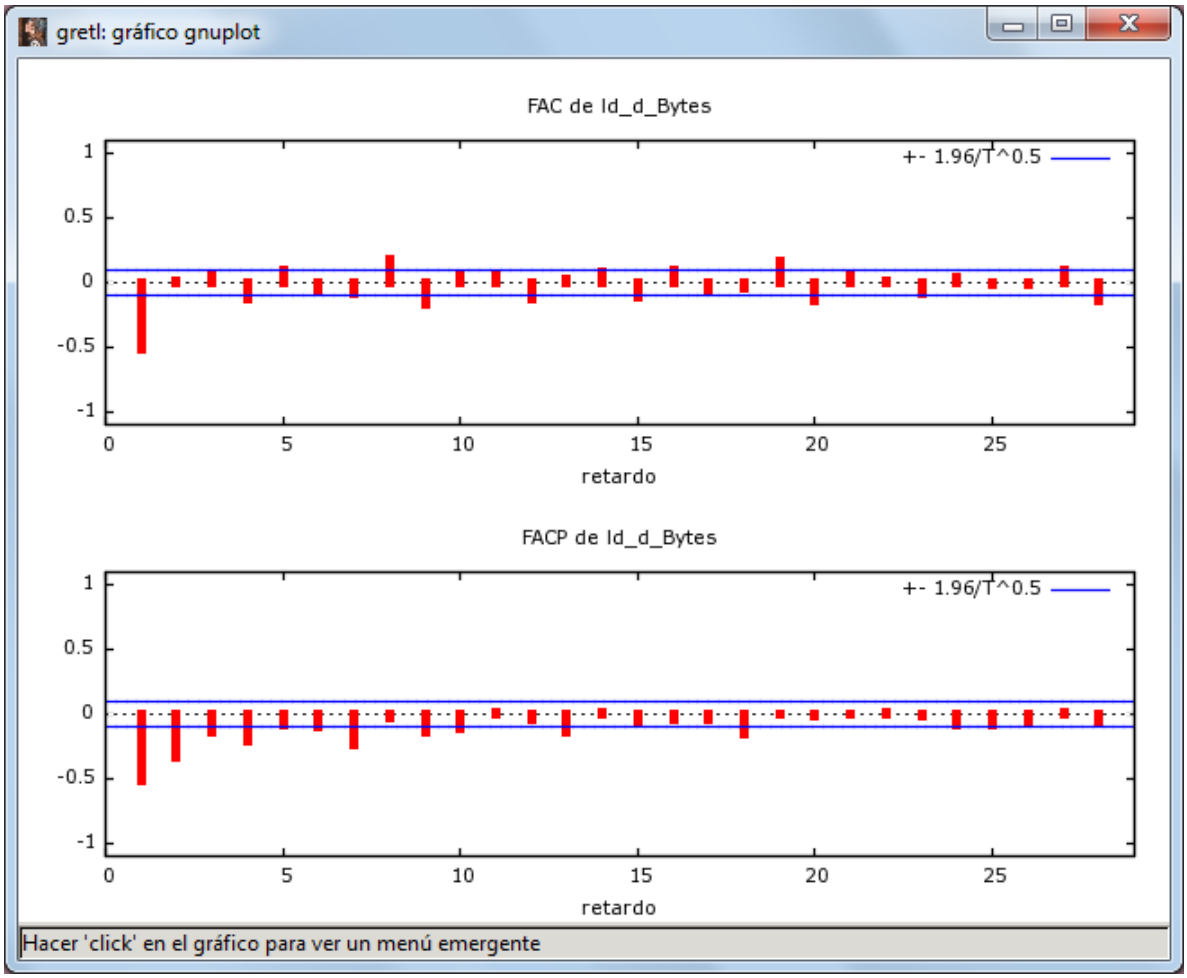

Figura 4.1.15

Se genera las series temporales fig. 4.1.16

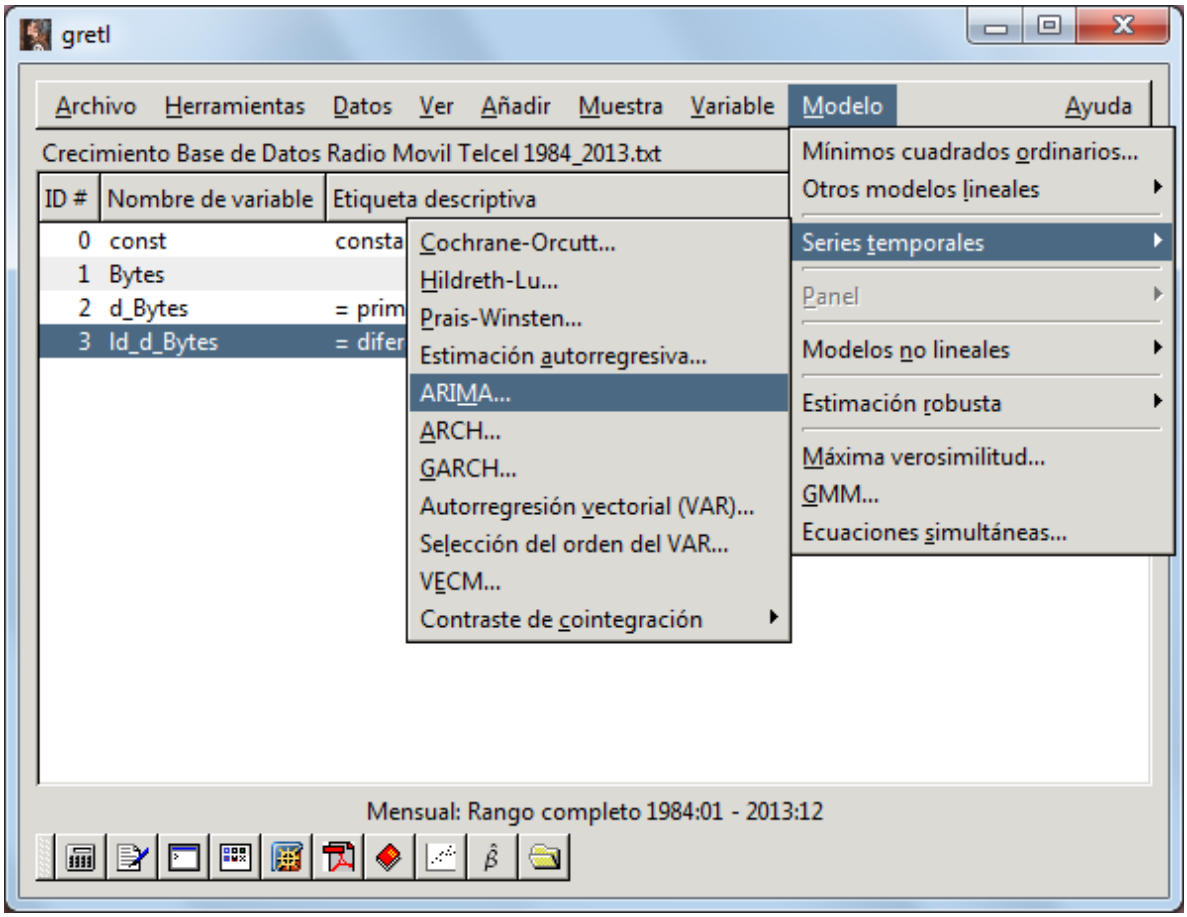

Figura 4.1.16

Se selecciona el modelo fig. 4.1.17

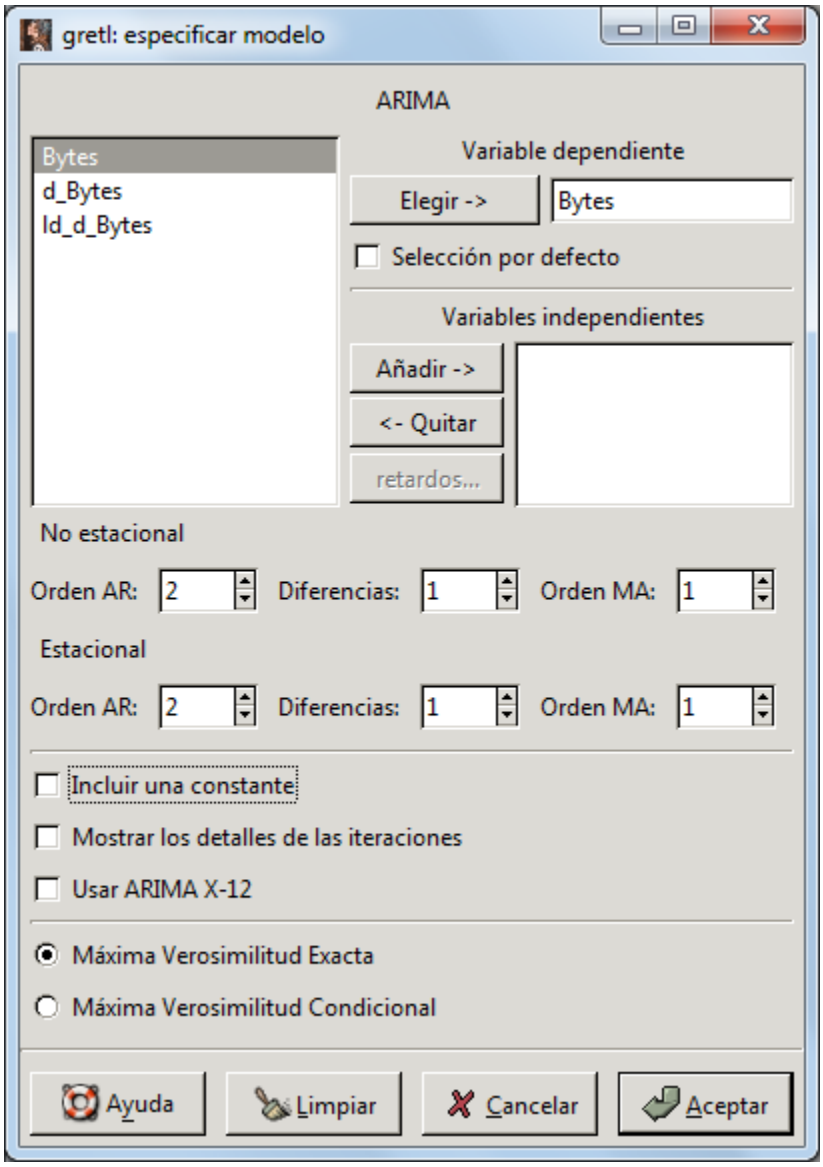

Figura 4.1.17

Se aprecia que tanto los valores de las variables como el criterio de Akaike (Brockwell, 2002 p. 170) son significativos y las raíces de MA también lo son, fig. 4.1.18

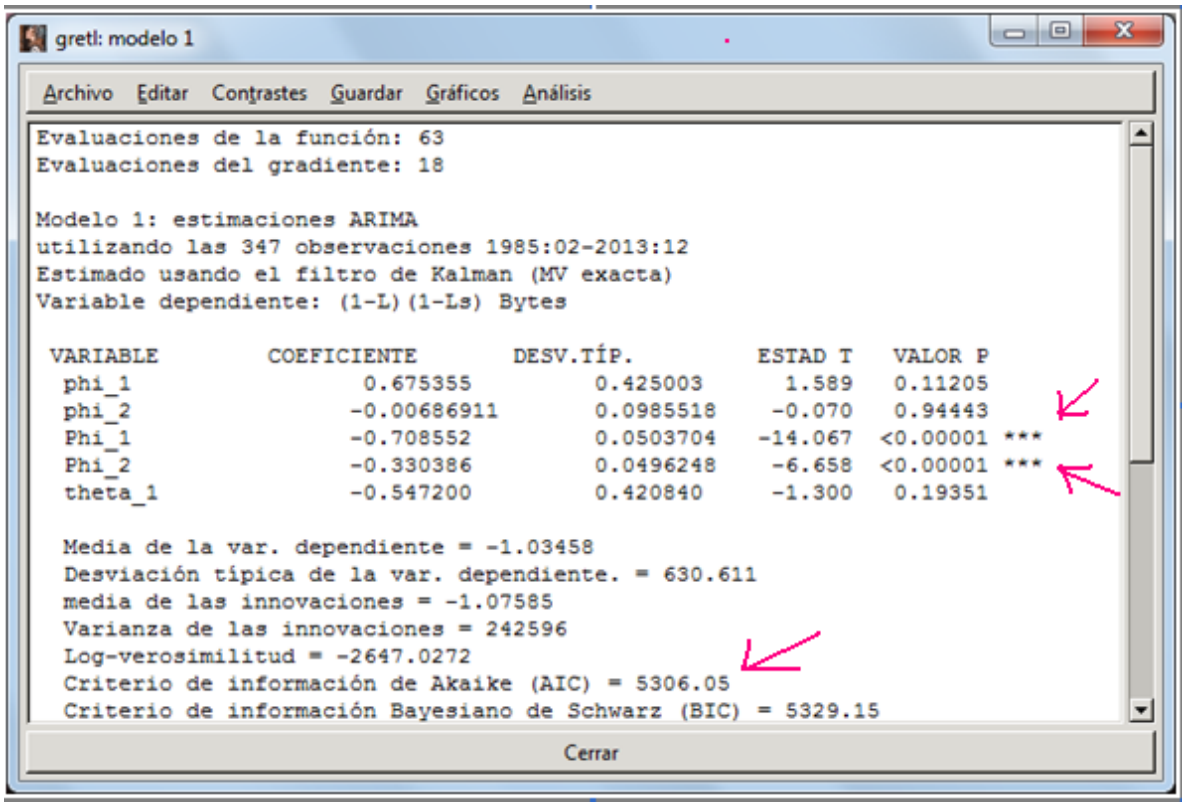

Figura 4.1.18

Se muestran las raíces de AR y MA, fig. 4.1.19

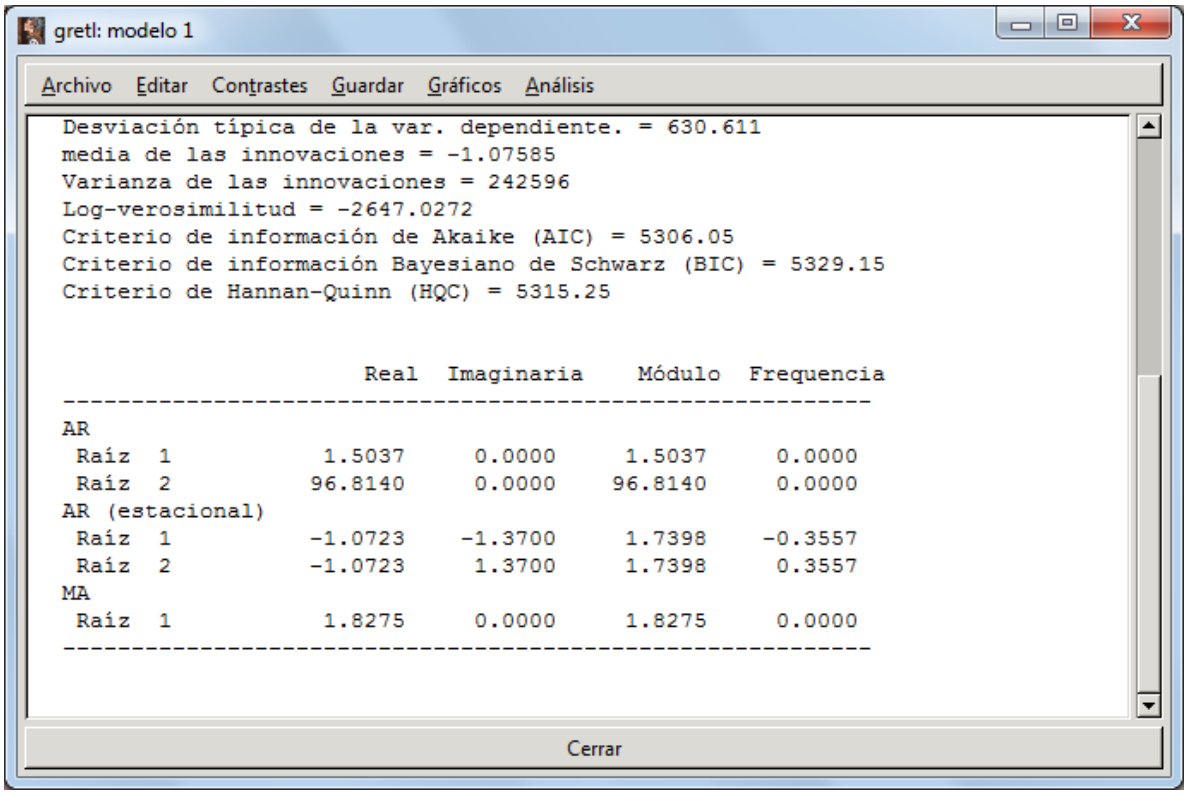

Figura 4.1.19

Se genera la predicción fig. 4.1.20

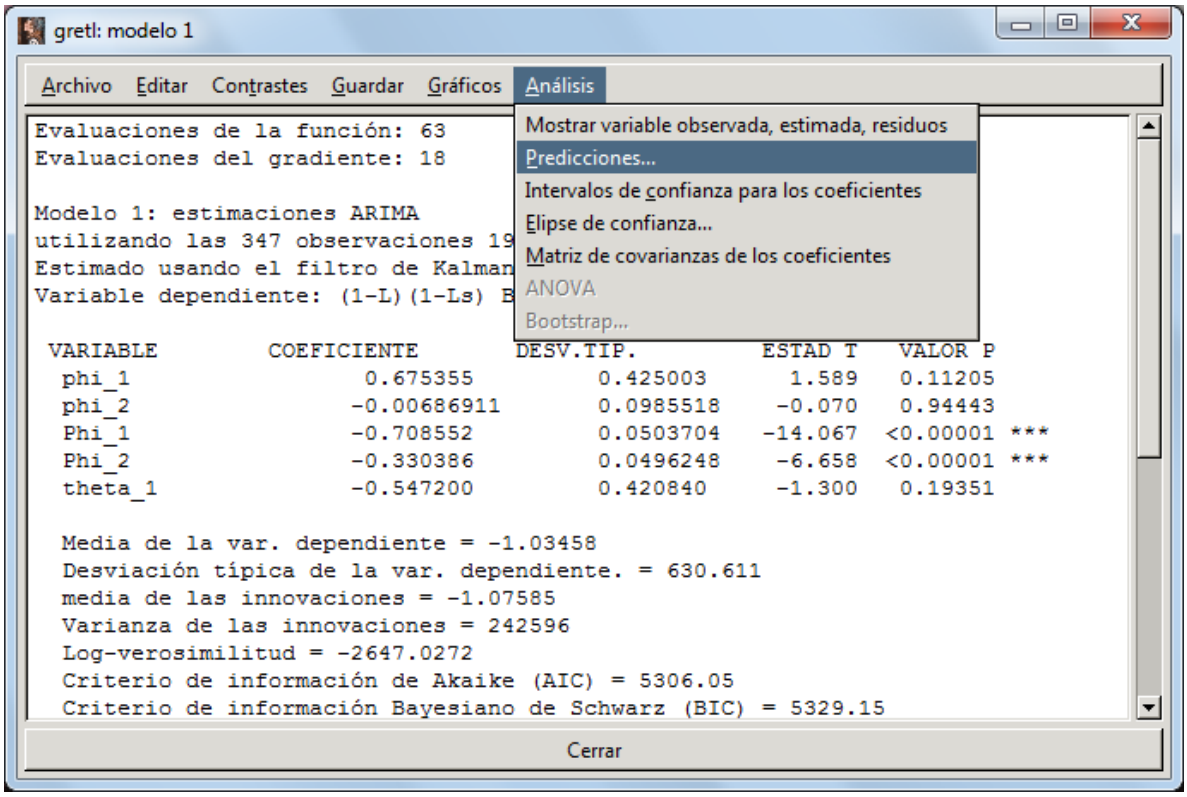

Figura 4.1.20

Se selecciona la predicción automática (dinámica fuera de la muestra) con un número de observaciones de 100, fig. 4.1.21

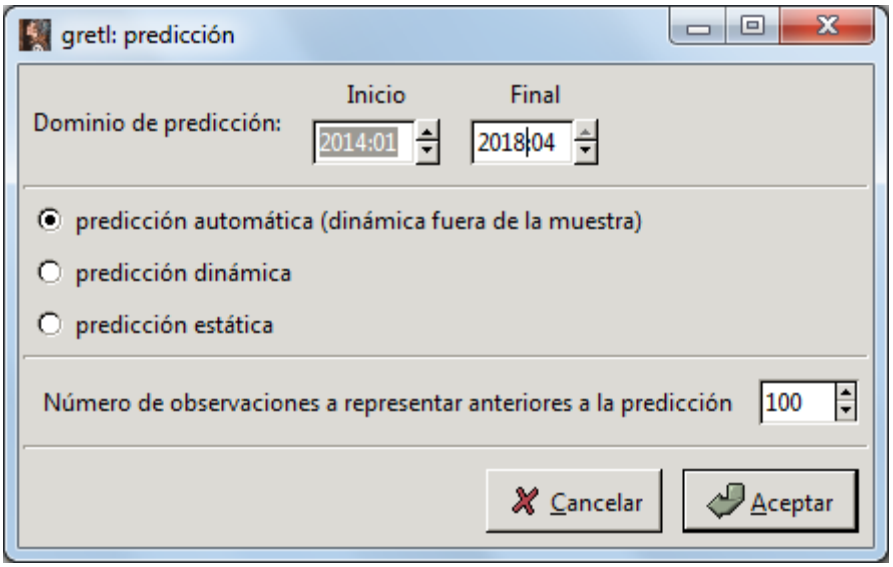

Figura 4.1.21

Aparece la serie esperada de la predicción fig. 4.1.22

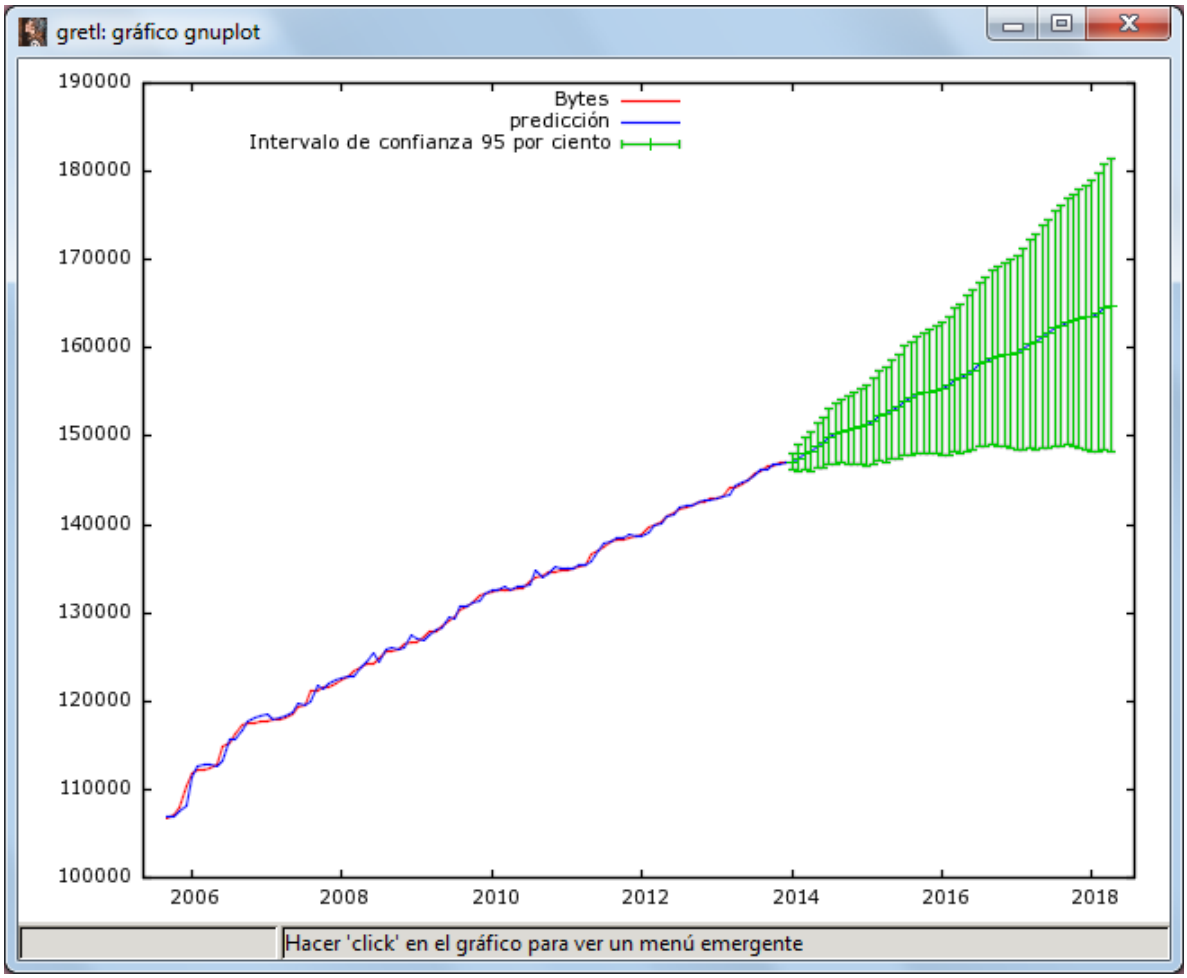

Figura 4.1.22

Los datos finales quedan de la siguiente forma en la que se presentan los valores pronosticados con la desviación estándar y un intervalo de confianza que muestra el rango de certeza con la que los datos se sometieron a análisis y poder saber si es confiable cada uno de los valores en cuanto al valor esperado y a la dispersión pronosticada del modelo fig. 4 .1.23:

Obs	MBytes	Predicción	DesviaciónTípica	Intervalo de confianza
2005:09	106806.00	106925.05		
2005:10	107153.00	106918.17		
2005:11	107968.00	107567.38		
2005:12	110392.00	108133.51		
2006:01	111724.00	111334.97		
2006:02	112147.00	112546.08		
2006:03	112176.00	112835.70		
2006:04	112343.00	112732.10		
2006:05	112809.00	112723.56		
2006:06	114922.00	113296.36		
2006:07	115301.00	115727.24		
2006:08	116273.00	115640.61		
2006:09	117355.00	116712.89		
2006:10	117559.00	117726.50		
2006:11	117566.00	118163.33		
2006:12	117615.00	118300.80		
2007:01	117671.00	118444.37		
2007:02	117818.00	117870.10		
2007:03	118013.00	118201.65		

2007:04	118219.00	118398.11		
2007:05	118621.00	118735.88		
2007:06	119280.00	119736.67		
2007:07	119454.00	119576.47		
2007:08	121092.00	119938.37		
2007:09	121166.00	121761.26		
2007:10	121547.00	121426.94		
2007:11	121583.00	122066.68		
2007:12	122084.00	122466.18		
2008:01	122459.00	122642.50		
2008:02	122720.00	122754.21		
2008:03	123503.00	122769.97		
2008:04	123871.00	123763.06		
2008:05	124132.00	124531.05		
2008:06	124188.00	125456.02		
2008:07	124867.00	124452.61		
2008:08	125562.00	125788.85		
2008:09	125602.00	125992.42		
2008:10	125779.00	125810.77		
2008:11	126520.00	125996.95		
2008:12	126587.00	127509.77		
2009:01	126764.00	127052.59		
2009:02	127187.00	126933.72		
2009:03	127866.00	127465.01		
2009:04	127874.00	128134.04		

2009:05	128592.00	128238.92		
2009:06	129045.00	129590.91		
2009:07	129550.00	129385.46		
2009:08	130319.00	130685.01		
2009:09	130810.00	130662.92		
2009:10	131209.00	131058.56		
2009:11	131925.00	131449.73		
2009:12	132249.00	132216.13		
2010:01	132367.00	132506.33		
2010:02	132491.00	132649.69		
2010:03	132538.00	133037.93		
2010:04	132668.00	132676.73		
2010:05	132712.00	133063.34		
2010:06	132861.00	133007.76		
2010:07	133694.00	133255.60		
2010:08	133919.00	134734.68		
2010:09	134234.00	134003.07		
2010:10	134678.00	134511.62		
2010:11	134715.00	135177.95		
2010:12	134758.00	134939.46		
2011:01	134807.00	134922.07		
2011:02	135007.00	135034.95		
2011:03	135266.00	135497.07		
2011:04	135496.00	135373.04		
2011:05	136642.00	135846.74		

2011:06	137117.00	136963.30		
2011:07	137543.00	137848.93		
2011:08	138150.00	138138.01		
2011:09	138213.00	138449.91		
2011:10	138309.00	138527.48		
2011:11	138580.00	138792.53		
2011:12	138695.00	138682.69		
2012:01	138952.00	138778.91		
2012:02	139624.00	139196.80		
2012:03	140002.00	139995.29		
2012:04	140350.00	140155.03		
2012:05	140828.00	140985.52		
2012:06	141262.00	141182.88		
2012:07	141711.00	141885.95		
2012:08	141962.00	142216.14		
2012:09	142240.00	142222.83		
2012:10	142479.00	142545.48		
2012:11	142603.00	142785.14		
2012:12	142884.00	142721.29		
2013:01	142914.00	143012.66		
2013:02	143217.00	143210.42		
2013:03	144153.00	143431.36		
2013:04	144192.00	144470.64		
2013:05	144641.00	144796.90		
2013:06	144972.00	144989.62		

2013:07	145768.00	145533.32		
2013:08	145999.00	146171.33		
2013:09	146629.00	146202.02		
2013:10	146729.00	146933.39		
2013:11	146943.00	146885.70		
2013:12	147016.00	147094.94		
2014:01		147135.91	492.540	146170.53 - 148101.29
2014:02		147543.39	742.534	146088.02 - 148998.75
2014:03		148044.01	951.458	146179.16 - 149908.87
2014:04		148262.52	1136.173	146035.62 - 150489.41
2014:05		148952.46	1303.203	146398.18 - 151506.74
2014:06		149369.78	1456.242	146515.55 - 152224.02
2014:07		149912.18	1597.801	146780.49 - 153043.87
2014:08		150274.88	1729.748	146884.58 - 153665.19
2014:09		150584.38	1853.535	146951.45 - 154217.31
2014:10		150735.58	1970.322	146873.75 - 154597.41
2014:11		150934.36	2081.045	146855.51 - 155013.20
2014:12		151099.87	2186.473	146814.39 - 155385.36
2015:01		151231.06	2333.419	146657.56 - 155804.56
2015:02		151686.42	2477.919	146829.70 - 156543.14
2015:03		152311.17	2618.210	147179.48 - 157442.86
2015:04		152504.57	2753.762	147107.20 - 157901.94
2015:05		153033.37	2884.486	147379.78 - 158686.96
2015:06		153423.56	3010.522	147522.94 - 159324.18
2015:07		154031.00	3132.121	147892.05 - 160169.96

2015:08		154306.99	3249.578	147937.82 - 160676.17
2015:09		154727.29	3363.196	148135.42 - 161319.15
2015:10		154888.13	3473.268	148080.53 - 161695.74
2015:11		155067.96	3580.066	148051.03 - 162084.89
2015:12		155236.64	3683.839	148016.32 - 162456.97
2016:01		155330.14	3843.568	147796.74 - 162863.53
2016:02		155717.05	4005.073	147867.11 - 163566.99
2016:03		156397.69	4165.293	148233.72 - 164561.67
2016:04		156549.57	4322.804	148076.88 - 165022.27
2016:05		157112.95	4476.865	148338.29 - 165887.60
2016:06		157493.84	4627.152	148424.63 - 166563.06
2016:07		158138.99	4773.584	148782.76 - 167495.21
2016:08		158432.90	4916.224	148797.11 - 168068.70
2016:09		158880.58	5055.207	148972.38 - 168788.79
2016:10		159017.68	5190.708	148843.89 - 169191.47
2016:11		159215.96	5322.915	148783.05 - 169648.87
2016:12		159351.83	5452.017	148665.88 - 170037.79
2017:01		159468.31	5644.959	148404.19 - 170532.43
2017:02		159887.90	5840.795	148439.94 - 171335.86
2017:03		160487.93	6035.967	148657.44 - 172318.43
2017:04		160677.52	6228.732	148469.21 - 172885.84
2017:05		161269.64	6418.121	148690.12 - 173849.16
2017:06		161666.09	6603.649	148722.94 - 174609.24
2017:07		162263.03	6785.123	148964.19 - 175561.87
2017:08		162572.89	6962.529	148926.33 - 176219.45

2017:09		162964.56	7135.957	148978.08 - 176951.04
2017:10		163115.30	7305.548	148796.42 - 177434.17
2017:11		163306.76	7471.473	148662.68 - 177950.85
2017:12		163464.84	7633.910	148502.37 - 178427.30
2018:01		163577.48	7839.393	148212.27 - 178942.70
2018:02		163996.53	8045.878	148226.61 - 179766.45
2018:03		164635.22	8251.033	148463.19 - 180807.24
2018:04		164811.81	8453.720	148242.52 - 181381.10

Figura 4.1.23

4.2 Resultado

Enfoques para pronosticar:

Existen dos enfoque generales para pronosticar, así como existen dos maneras de abordar todos los modelos de decisión:

Los pronósticos cualitativos o subjetivos involucran algunos factores importantes tales como la intuición, emociones, experiencias personales del que toma la decisión, y sistemas de valores para alcanzar un pronóstico.

Los pronósticos cuantitativos por su parte incorporan una gran variedad de modelos matemáticos que utilizan datos históricos y/o variables causales para pronosticar la demanda o ventas futuras. (Caba, 2011 p. 88)

Por otro lado existen técnicas univariantes y multivariantes, se usará una técnica univariante, las hay sencillas como las autoregresivas de primer orden o modelos de tendencia lineal o exponencial para este caso se usará una un poco más compleja denominada Box Jenkins.

Investigación

Para desarrollar este trabajo se tomaron datos de la empresa de telecomunicaciones radio móvil dipsa s.a de c.v., que es subsidiaria de la empresa América Móvil, está a su vez filial de grupo carso. Aunque el origen de la compañía como una empresa de servicios de telefonía se remonta al año 1981, los datos para este análisis se toman de la serie de tiempo del año 1984 al año 2013 para la serie original.

Se usó el software GRETL de distribución libre GNU versión 1.9.92 CVS que implícitamente hace uso del método de máxima verosimilitud, a diferencia de otros que usan el método de mínimos cuadrados.

Se usó el modelo autorregresivo integrado y de medias móviles: ARIMA. Entre los criterios de Akaike (AIC) y Schwartz (SC) o criterio de información bayesiana (BIC), se tomo al de Akaike por ser el que da la más baja suma de errores al cuadrado. (Montenegro, 2011 págs. 94-95)

En la grafica 4.1.2 se aprecia una tendencia ascendente a partir del año 2005 al año 2013 esto quiere decir que la serie de tiempo tiene heteroscedasticidad o la varianza no es continua y que por tal motivo tendrán que usarse métodos para hacerla estabilizar la tendencia. Aunque en el año 2009 se esperaría que cayera el crecimiento de datos en los esquemas y objetos que componen en su conjunto van ocupando espacio, esto por la baja generalizada del consumo en México por la alerta de pandemia del H1N1 así como una desaceleración en estados unidos por la crisis hipotecaria en la bolsa de valores, que como consecuencia impactaron al crecimiento de la mayoría de las empresas mexicanas, esto no impacto de manera considerable a América Móvil debido a que es una empresa preponderante en México en el sector de las telecomunicaciones y a esa fecha de la serie no existe un competidor considerable para el crecimiento exponencial de venta de servicios en este sector.

Por esta razón las ventas al menudeo de recarga amigo móvil (retail o venta al menudeo) no impactaron de manera considerable y esto como consecuencia no se ve reflejado en la grafica de la series temporales.

En la grafica solo se ven pequeñas variaciones que vistas desde la tendencia general no son significativas para tener algún cambio en el resultado final del pronóstico.

No obstante que del año 1998 al año 2001 si se nota claramente una tendencia en la que se aprecia un decaimiento ligeramente en la tendencia que anteriormente se presentaba de crecimiento sostenido, esto se estima que se presento por una desaceleración de las ventas de recarga amigo móvil por el cambio de periodo presidencial 2000 – 2006, aunque este ligero cambio en la tendencia tampoco fue significativo.

Cabe aclarar que la serie de este trabajo es muy parecida a la mayoría de series usadas en economía, esto facilita su predicción ya que los grados AR y MA a lo más usan el nivel 2 para su modelado.

En la figura 4.1.2 se abre la serie para posteriormente graficar en fig. 4.1.3, se generan las primeras diferencias Fig. 4.1.8, después de hacer una serie de iteraciones llegamos a la figura 4.1.15 en la que se aprecia que para la FAC hay un valor significativo y para la FACP son dos valores significativos lo cual permite suponer que hay un posible modelo de la forma ARIMA(2,0,1) de esta forma se nota que para el valor p es 2, máximo nivelo grado usado común mente en estos modelos ARIMA(*p,d,q*).

El modelo que se usará es ARIMA (2,1,1)x(2,1,1) esto justificado por el hecho de que en la figura 4.1.18 el valor p es significativo y así también el valor de Akaike.

Se ejecuta la predicción fig. 4.1.21 en un dominio de 2014 a 2018.se genera finalmente la grafica original denominada Mega Bytes para medir los megabytes y una grafica de predicción en color azul, así también el pronóstico en color verde con un intervalo de confianza del 95 %, notemos que tanto la grafica original como la de predicción son muy similares.

En la fig. 4.1.22 se observa la grafica original en color rojo la grafica predicción en color azul y en verde la grafica de predicción así como el umbral de confianza que es de un 95%.

Se nota que el crecimiento a partir de la grafica verde es uniforme y se estima el espacio necesario para poder hacer una solicitud de discos al área de infraestructura que administra el storage (arreglo de discos).

En enero de 1984 existen 978 MB de ocupación y para diciembre del 2013 se tienen 147016 MB de ocupación es decir 143 GB de espacio utilizado. Esto para la serie de tiempo real.

Para la serie estimada en el periodo de septiembre 2005 a diciembre de 2013 se tiene: inicio 104 GB, final 143 GB.

Finalmente para el pronóstico del periodo enero 2014 a abril del 2018 inicio 143 GB, final 164811.81 MB dividiendo entre 1024 son160 GB, que es el aproximado al que se debe considerar el crecimiento de espacio para la base de datos.

Cada disco SSD es de 60 GB así que se necesitaran 3 discos con un total de 180 GB, sin considerar la redundancia de Oracle que puede ser normal, alta o externa (tipo de espejeo).

Externa solo 3 discos de 60 GB, para la normal se reduce el tamaño necesario a la mitad es decir que se duplica a 320 el espacio requerido o el equivalente en discos 6 discos de 60 GB.

Para la configuración alta se reduce a una tercera parte es decir se requiere el triple de espacio físico, es decir: 480 GB o en discos son necesarios 8 discos de 60GB. Por ser una base de datos crítica de alta disponibilidad se usará la tercera opción o redundancia alta es decir 8 discos pero para dejar de reserva se usarán 9 discos de 60 GB, para dar en total de 540 GB que se reducirán a 180 GB una vez que se generen los grupos de discos en ASM.

Además de los discos que se usarán para los datos e índices, es necesario también configurar la FRA (Fast Recovery Area) y el área de archivamiento en automático archive log. Así que con la información que se obtuvo anteriormente usando el pronóstico Box-Jenkins se necesita el doble de espacio que tiene la base de datos para la FRA y así también el doble para el área de archive logs es decir 160 GB x 2 = 320, 320 GB para la FRA y 320 GB para el área archive log, con una redundancia normal es decir el doble para cada área. Es decir 9 discos de 60 GB para poder asignar, abajo se muestran la distribución de espacios, tamaño requerido, tipo de redundancia ASM, tamaño real y número de discos SSD de 60 GB.

AREA	ESPACIO REQUERIDO	ESPACIO ESTIMADO DISCOS 60 GB	REDUNDANCIA	ESPACIO TOTAL	TOTAL DISCOS
Datos	160 GB	180 GB	Alta 1/3	540 GB	9
Archivos ARC	320 GB	360 GB	Norma 1/2	720 GB	12
FRA	320 GB	360 GB	Externa 1/1	360 GB	6

Figura 4.2.1

Como conclusión de acuerdo al estudio realizado se requieren 27 discos de estado sólido para instalar la base de datos, el área de recuperación y el área de archivamiento.

De esta forma se estimó el espacio más óptimo con la finalidad de no generar pérdidas económicas al momento de implementar por primera vez la base de datos en cuestión.

Por otro lado se hizo la descripción a lo largo de este trabajo de las herramientas de estimación del modelo matemático con la finalidad de dar a conocer la tecnología que está disponible en este momento sin la necesidad de pagar alguna licencia costosa para fines académicos; pero también en ámbitos empresariales y de gobierno.

Finalmente se estableció un modelo general para poder ser usado en otro tipo de estimaciones de crecimiento en servidores a los que no se tiene acceso debido al costo y a la complejidad de instalación para el propósito de hacer un pronóstico a futuro.

En última instancia se podría hacer una modificación en los valores para generar resultados que excedan los umbrales máximos y mínimos de usos de espacios con la finalidad de modelar escenarios posibles para diagnosticar pérdidas en caso de excederse en algún posible desbordamiento de datos y como consecuencia compra y adición de discos no previstos o por el contrario subutilizar el espacio y tener discos que se compraron pero que no se utiliza al final del pronostico en su totalidad.

4.3 Trabajos futuros

En trabajos posteriores se puede usar esta metodología para hacer pronósticos de crecimiento de otras bases de datos principalmente cuando se tienen datos productivos en los que se requiera hacer una estimación de la tasa de crecimiento del espacio para recursos de hardware (discos).

Aunque hoy en día se está comenzando a usar la tecnología de almacenamiento en la nube, se puede también aplicar en estas empresas especializadas para almacenamiento en grandes cantidades con la cual pueden pronosticar un crecimiento masivo en función del crecimiento y demanda que presentan cada una de las empresas que compran el servicio de alojamiento en relación al tiempo.

Revisando un poco la historia en la década de los 80's aparecieron diversas técnicas para el análisis de series temporales cuyo procedimiento estaba basado en las relaciones lineales entre las variables; pero en el mundo real los datos no se relacionan de manera fácil, una de estas razones es debido a que este tipo de modelos presentan cambios abruptos. Las relaciones lineales normalmente son inadecuadas para hacer previsiones (Granger y Newbold, 1986), por tal motivo sería ideal que fuesen modelos no lineales. Aparecieron otros métodos como: modelos de umbral y modelos bilineales.

Otra metodología que se puede usar para series del tipo anteriormente tratado con la metodología Box-Jenkins en trabajos futuros, puede ser el área de las redes neuronales que también hacen estimaciones del comportamiento de varios resultados en el tiempo en función de algoritmos inteligentes de entrenamiento, utiliza también una secuencia de valores medidos en unidades de tiempo discretas o continuas y pueden ser también univariantes y multivariantes.

Las redes de neuronas artificiales son modelos matemáticos que emulan a las redes neuronales del cerebro humano las neuronas de este modelo se estructuran en capas las cuales están relacionadas entre sí, denominadas: capa de entrada, capa oculta y capa de salida.

En términos generales la capa de entrada recibe los datos, la capa oculta procesa estos para generar información relacionada y la capa de salida genera los resultados, internamente hay una función de transferencia que puede ser de tres tipos:

Lineal La actividad de salida es proporcional a la entrada ponderada total de inicio

De umbral — La salida queda fija a uno de dos niveles, dependiendo de un valor crítico (umbral)

Sigmoide — La salida varía continuamente dependiendo de la entrada; aunque esta no es lineal

El proceso de las redes neuronales integra tres etapas fundamentales:

Diseño de la Red — Número de neuronas, capa de entrada, capa oculta y salida, la arquitectura depende del problema en particular.

Computación cuántica y probabilidad:

> Existen investigaciones recientes que proponen una nueva manera de organizar los datos que tradicionalmente se almacenan como 1 verdadero y 0 falso en ciencias de la computación. Lo interesante de este nuevo paradigma es que también involucra valores aleatorios o estocásticos y no únicamente deterministas.

Es interesante entender que actualmente se está investigando la manera de almacenar datos desde el modelo de computación tradicional usando el comportamiento físico de los cuantums en la materia, o en átomos de cierta materia específicamente usando para ello campos magnéticos que obligan al comportamiento de esta a obtener el estado deseado para fines de almacenamiento.

La probabilidad y estadística intervienen en este campo de la computación en tanto que un qubit o unidad de almacenamiento en computación cuántica, está representado en un vector formado por números imaginarios en el que actualmente se trabaja para definir sus leyes de comportamiento.

Lo anterior no solo se puede usar para el almacenamiento de información en computadoras cuánticas, sino que se tiene previsto que también se puedan hacer simulaciones haciendo uso del comportamiento de la materia con campos magnéticos controlados para obtener un resultado optimo.

Una parte interesante que se pudo revisar aunque es para gran cantidad de datos “big data” fue la minería de datos con el software RapidMiner también de distribución GNU, se pueden hacer pronósticos con algoritmos bayesianos incluidos en sus librerías y además pueden tomarse decisiones en función de los resultados obtenidos.

Por la parte de bases de datos existen nuevas tecnologías, una de mucho interés es la de la base de datos HANA que se está implementando en las aplicaciones del ERP SAP R/3 o versiones más nuevas que funciona casi en su totalidad con memoria y discos de estado sólido, aunque en este momento estoy comenzando a configurar ambientes virtuales para probar esta tecnología que es más rápida incluso que Oracle Exadata pero eso no lo puedo asegurar solo es una suposición hasta este momento.

Conclusiones

Para finalizar, se tiene la serie estimada del periodo de septiembre 2005 a diciembre de 2013 con un tamaño de inicio de 104 GB y final de 143 GB solo considerando el área de datos, ya que las áreas de archivamiento y FRA se estiman en función del doble para cada una de ellas.

Para el pronóstico que se realizó del periodo de enero 2014 a abril de 2018, inicia con 143 GB y finaliza en 160 GB≅164811.81 MB, es decir, ésta es la tasa de crecimiento que se experimentará del periodo de 2014 a 2018. Por tal motivo de acuerdo con la tabla 4.2.1 de resultados presentada anteriormente se requieren1,620 GB de espacio o el equivalente a 27 discos.

La tasa de crecimiento permite hacer la estimación que se buscó durante la investigación, que en términos generales es el espacio necesario para hacer un dimensionamiento de discos de acuerdo a la redundancia que usa ASM:

- normal (se utiliza todo el espacio)
- externa(se usa la mitad del total)
- alta (se usa una tercera parte del total de espacios en los discos)

El crecimiento se derivó de la función del pronóstico del periodo del año 2014 al 2018 junto con su margen de certeza que es un porcentaje considerable para poder aceptarlo de acuerdo a cada uno de los puntos y la dispersión correspondiente.

Con lo anterior se conoció la tasa de crecimiento planteada originalmente en la hipótesis determinando el tipo de redundancia para cada área lógica de almacenamiento y el número de dispositivos que en este caso es de 9 discos SSD.

Así también se pudo pronosticar el crecimiento de la base de datos Oracle Exadata 11g R2 con la metodología Box-Jenkings y llegar a un resultado en la hipótesis en el que se conoció el incremento para obtener el dimensionamiento adecuado en el uso de espacio y en la optimización de costos al hacer la instalación.

Se logro encontrar una tasa aceptable del crecimiento de los espacios que permiten cumplir tanto con la hipótesis como con el objetivo principal de obtener el dimensionamiento adecuado y con ello hacer una adquisición adecuada de discos para el almacén de datos.

Ligado a lo anterior se pudo implementar por primera vez el espacio y evitar con ello pérdida que en el futuro podrían derivarse en gastos excesivos para la organización.

También se hizo una descripción de las herramientas matemáticas y computacionales como la metodología Box-Jenkins así como los filtros utilizados al usar las series temporales que ayudan a detectar el modelo más óptimo cuando se realiza el proceso de generar la serie estimada y el pronóstico a partir de que terminan los datos reales obtenidos de un proceso estocástico.

Se hizo también una descripción general del software GRETL que sirve para facilitar los cálculos numéricos que permiten ganar tiempo a la hora de obtener resultados y ser más precisos en los cálculos matemáticos.

También se explica de manera general el paradigma de almacenamiento de la empresa Oracle con tecnología Exadata que es un sistema de almacenamiento relacionado con una instancia ASM que hace el trabajo de una controladora de discos con diferencias particulares descritas en el trabajo. Además de incluir nuevas tecnologías como discos de estado sólido y conceptos como el de redundancia en ASM.

Finalmente se presentó el modelo general para hacer estimaciones de crecimiento en servidores a los que no se tiene acceso por su alto costo en el mercado y por que la tecnología no siempre es facilitada por las empresas debido a que su información es confidencial y a que no se tiene como tal un ambiente de pruebas para probar con el modelo mencionado.

Bibliografía

1. Abraham J, Bagal P. and other authors (2012) Automatic Storage Management Administrator's Guide 11g Release 2 (11.2), Parkway, Redwood City, CA 94065. Oracle America, Inc.
2. Alapati. S. (2009) Expert Oracle Database 11g Administration. 233 Spring Street, 6th Foor, New York, NY 10013. Springer-Verlag New York.
3. Alapati. S. (2009) Oracle Database 11g New Features for DBAs and Developers. 233 Spring Street, 6th Foor, New York, NY 10013. Springer-Verlag New York.
4. Artigas M. (1999), Filosofía de la ciencia , Ed. EUNSA, Plaza de los sauces 1 y 2 31010 BarañainNavarra España.
5. Ashdown L. and Kyte T. (2009) Oracle Database Concepts 11g Release 2, Parkway, Redwood City, CA 94065. Oracle America, Inc.
6. Bauwens. C. (2009) Oracle 11g: New Features for Administrators Volume I. 500 Oracle Parkway, Redwood Shores, California 94065 USA. Oracle Press.
7. Box, G.E.P. y Jenkins G.M. (1976) Time Series Analysis: Forecasting and Control. 2nd Ed. San Francisco: Holden Day.
8. Brockwell P. (2002) Introduction to Time Series and Forecasting, Printed and boundby R.R. Donnelley and Sons, Harrisonburg, VA.Springer-Verlag New York, Inc.
9. Bryla. B. (2008) DBA Oracle Database 11g DBA HandBook. New York, USA. Oracle Press.
10. Caba N. (2011) Gestión de la Producción y Operaciones, Colombia, Ed. Corporación para la Gestión del Conocimiento Asesores del 2000
11. Chan. I. (2011) Performance Tuning Guide 11g Release 2, Oracle Parkway, Redwood City, CA 94065, Oracle America, Inc.
12. Clarke J. (2013) Oracle Exadata Recipies, 233 Spring Street, 6th Floor,New York, NY 10013., Apress Media.
13. Cottrell A.(2014),Gretl User's Guide: Gnu Regression, Econometrics and Time-series Library, GNU Free Documentation License, Version 1.1 or any later version published by the Free SoftwareFoundation (see http://www.gnu.org/licenses/fdl.html).
14. Date. C. (2003) An Introduction to Database System (8th Edition). Boston, MA, USA. Addison-Wesley Longman Publishing Co., Inc.
15. Delgado C. (2000) Aplicaciones de Modelos Matemáticos en el parque de diversiones "La Feria de Chapulepec"., Av. Alcanfores y San Juan Totoltepec s/n, Santa Cruz Acatlán Naucalpan Edo de Mex, Fes Acatlán UNAM.
16. Devore L. (2001) Probabilidad y Estadística para Ingeniería Y Ciencias, Séneca Núm 53 Col. Polanco México, D.F., 11560, International Thomson Editores, S.A. de C.V.

17. Granger, C.W.J. y Newbold, P. (1986) Forecasting economic time series. 2nd Ed. Orlando, FL, Academic Press.
18. González P. (2009). Análisis de series temporales:Modelos ARIMA. Ed universidad del país vasco.
19. Green R. (1999) ,A Handbook of Time-Series Analysis, Signal Processing and Dynamics, 525 B Street, Suite 1900, San Diego, California 92101-4495, USA
20. Kuhn. D. (2007) RMAN Recipes for Oracle database 11g. 233 Spring Street, 6th Foor, New York, NY 10013. Springer-Verlag New York.
21. Loney. K. (2009) Oracle Database 11g The Complete Reference. San Francisco CA. McGraw-Hill Press.
22. Mendenhall W. (2010), Introducción a la probabilidad y estadística. Av. Santa FéNúm. 505, Piso 12, Col Cruz Manca, Santa Fé, CP 05349, México D.F. Cengage Learning Editores S.A. de C.V. Décima tercera edición.
23. Montenegro A.(2011), Análisis de series de Tiempo, Editorial pontificia Universidad Javeriana,Carretera 7 núm. 37 – 25 Piso 13 Bogotá DC
24. Orosio. L. (2008) Bases de Datos Relacionales Teoría y Práctica. Calle 73, No. 76ª 354 Medellín, Colombia. Fondo Editorial ITM, 1ra Edición.
25. Osborne. K. (2010) Expert Oracle Exadata. 233 Spring Street, New York NY 10013, U.S.A. Apress.
26. Pollock, D.S.G. (1999) A Handbook of Time-Series Analysis, Signal Processing and Dynamics. 24-28 Oval Road London NW1 7DX, UK, Academic Press.
27. Popper K. (1967), La lógica de la investigación científica, Ed. Tecnos, O'Donell 27 – 28009 Madrid España.
28. Price. J. (2008) Oracle Database 11g SQL, Master sql and pl/sql in the oracle database. San Francisco CA. McGraw-Hill Press.
29. Rockwell Software. (2012) Arena Contact Center User's Guide, 1201 South Second Street Milwaukee, WI 53204-2496 USA. Allen-Bradley Press.
30. Videgaray M. (2011) Pronósticos: Metodología de Box-Jenkins, Unidad de Servicios Editoriales de la FES Acatlán, FGB 47-B Col Ampliación Morelos México DF.
31. Wreenwald. R. (2009) Oracle Essentials, Oracle Database 11g. 1005 Gravenstein Highway North, Sebastopol, CA 95472. O'Reilly Media, Inc.
32. Xirau Ramón. (2011) Introducción a la histioria de la filosofía. Universidad Nacional Autónoma de México, Col Copilco Universidad, Del. Coyoacan 04360, México, D.F.

Glosario

Android: Sistema operativo basado en el kernel (núcleo) Linux diseñado principalmente para dispositivos móviles, teléfonos o tablets

ASM: Automatic Storage Management es una característica de Oracle para el manejo de espacios en las bases de datos que permiten administrar volúmenes y arreglos de discos, esta tecnología está basada en una instancia independiente ala que administra datos y sustituye a los programas que desarrolla cada compañía que administra los arreglos de discos

Autocorrelación: es la herramienta básica utilizada en la mayoría de los métodos avanzados, tales como ARMA de Word, ARIMA de Box-Jenkins, ARARMA de Parzen y filtrado de Kalman. Sirve para identificar el patrón básico y determinar el modelo apropiado que corresponde a la serie de tiempo (datos)

Box-Jenkins: es una metodología creada por George Box y Gwilym Jenkins en la década de los 70s y es usada para los modelos autorregresivos de media móvil ARMA o a los modelos autorregresivos integrados de media móvil ARIMA encuentra el mejor ajuste de una serie temporal de valores mas óptimos.

CRM : Sistema de administración basada en los clientes, (customer relationship management)

DBWR Es un proceso de segundo plano que se activa al iniciar la instancia del servidor Oracle con la finalidad de escribir a los data files o archivos de datos bajo ciertos cambios que se presentan durante la actividad del servidor

DSS: Son bases de datos que están diseñadas para consulta de gran cantidad de datos y toma de decisiones esta palabra viene como opción en el asistente de instalación de Oracle

Ext4 fourth extended filesystem, es un sistema de archivos transaccional aparece en octubre de 2006 como respuesta a una mejora de ext3 algunas mejoras son soporte de mas volumen y añadido de extent, menor uso de CPU mejoras en velocidad de lectura escritura.

Entropía: Es un concepto que se generó a partir del estudio de la energía por el físico Ludwing Boltzmann y es el concepto de el estado en el que se encuentra un sistema proporcionalmente a la libertad en la que sus componentes se encuentran, el orden sería lo contrario a un sistema en estado de libertad.

ERP: Sistema de planificación de recursos empresariales *enterprise resource planning* se encuentran en: logística, distribución, inventario, envíos, facturas y contabilidad de la compañía de forma modular. Ejemplos son SAP, Oracle Applications, People Soft, etc.

Exadata: Es una aplicación de base de datos que soporta OLTP (arquitectura de base de datos transaccional)y OLAP (Arquitectura analítica), Exadata originalmente fue diseñado por Oracle corporation y Hewlett Packard. Oracle diseño la base de datos con un sistema operativo basado en Linux y el almacenamiento en disco lo diseñó HP, cuando Oracle compra Sun Microsystems la versión 2 de Exadata se baso en Sun storage systems.

Extent o extensión en español significa el espacio o espacios adicionales que se van agregando al segmento a partir de ciertas reglas definidas desde la configuración.

FAC Función de auto correlación parcial es la contribución marginal o el peso de cada nueva variable autorregresiva Y_{t-k}, también es una herramienta que sirve para definir el modelo.

FACP Función de autocorrelación parcial es la grafica generada por los valores de la FAC contra los valores de los intervalos k de números naturales

FAT32 File Allocation Table es un formato que se origina del FAT creado por Bill Gates y que es una mejora del FAT16 y se mantiene la compatibilidad de MS-DOS utilizando direcciones de cluster de 32 bits aunque solo se utilizaron 28

GNU: es un acrónimo recursivo de *"GNU's Not Unix!"* (*en español:* GNU no es Unix),47 elegido porque el diseño de GNU es Unix-like, pero se diferencia de Unix en que es software libre y que no contiene código de Unix

GRETL Es un programa econométrico y estadístico de distribución libre, interactúa con R-project y LaTeX que se especializa en los algoritmos que resuelven modelos estadísticos entre ellos se encuentra la metodología Box-Jenkins, es de distribución libre GNU, esto quiere decir que no es necesario pagar una licencia.

Hard Parsing: Parseo duro significa que una consulta se creará por primera vez ya que los datos son traídos de disco duro

Homoscedasticidad: es cuando una función toma varios valores a lo largo del tiempo pero estos se mantienen más o menos constantes. Esto se logra al estabilizar la varianza.

Heteroscedasticidad: es cuando una función toma varios valores a lo largo del tiempo y estos varían a través del tiempo esto normalmente se presenta de forma ascendente.

LGWR Es un proceso de segundo plano que se activa al iniciar la instancia del servidor Oracle con la finalidad de escribir a los archivos log que guardan los cambios temporalmente hechos por la actividad de la base de datos.

Linux: Sistema operativo de licencia libre basado en un kernel parecido al de UNIX, asociado a la licencia GNU.

Infinidad: es una arquitectura de comunicación usada para conectar dos servidores (nodos) en la arquitectura Oracle Exadata.

LUN (Logical Unit Number) es una unidad de medida que se usa en almacenamiento computacional esta unidad es direccionada por el protocolo SCSI o la Storage Area Network el cual

usa SCSI ya sea con fibra óptica o SCSI. La LUN es soportada por cualquier dispositivo de almacenamiento que soporta lectura / escritura.

LVM (Logical Volumen Management) es un método que permite reservar espacio en dispositivos de almacenamiento masivo de manera más fácil, además los administradores pueden redimensionar o mover las particiones sin que haya interrupción en la operación del sistema.

MAC OS es un sistema operativo basado en un ambiente grafico para el usuario desarrollado por Apple. Inc. Apareció por primera vez en 1984 como system, posteriormente se le dio el nombre de MAC OS.

Modelo AR autorregresivo: Es aquel modelo en el que bajo ciertas condiciones, toda y_t puede expresarse como una combinación lineal de sus valores pasados (parte simétrica) mas un termino de error (innovación). La variable endógena de un periodo t es explicada por las observaciones de ella misma correspondientes a periodos anteriores y agregándose al final un término de error (ruido blanco), deben cumplirse tres condiciones para lo anterior: media nula, varianza constante y covarianza nula entre errores correspondientes a errores diferentes.

Modelo MA de medias móviles: Es aquel que explica el valor de una determinada variable en un periodo t en función de un término independiente y una sucesión de errores que corresponden a periodos precedentes, ponderados convenientemente, un modelo de medias móviles pude obtenerse a partir de un modelo autorregresivo sin más que realizar sustituciones sucesivas utilizando el teorema general de la descomposición de Wold.

MTBF (Mean Time Between Failure) Es una predicción de duración de un sistema computacional en este caso un disco duro, también se puede denominar como un promedio aritmético.

NTFS network technology file system, es un system de archives de windows NT está basado en el sistema de archivos HPFS IBM/Microsoft usado también en OS/2, es un sistema de archivos diseñado para particiones de gran tamaño.

OLAP: Procesamiento analítico en línea, On-Line Analytical Processing, solución utilizada en la inteligencia empresarial. Agiliza la consulta de grandes cantidades de datos

OLTP: Procesamiento de transacciones en línea, online transaction processing es un sistema de base de datos que esta balanceado en datos de consulta e inserción de nuevos datos.

Parsimonia el acto de elegir siempre el modelo más sencillo, agregando siempre al modelo elementos que no están plenamente justificados. Basado en el principio de William de Occam: “De entre cosas iguales, la solución más sencilla tiende a ser la mejor”

Proceso ergódico: La correlación entre una variable y su pasado va reduciéndose a medida que nos alejamos más en el tiempo del momento para el que estamos considerando dicha correlación, es por esto que se dice también que es un proceso invertible.

RAID (Redundant Array of Inexpensive Disks) es una tecnología para virtualidad discos de datos, combina múltiples controladores de discos en una unidad lógica para obtener una redundancia de mejora en desempeño.

RDBMS: Sistema manejador de bases de datos *Relational en ingles: Database Management System*

SAN Storage Área Network es una arquitectura basada en red que sirve para administrar dispositivos de almacenamiento arreglos de discos, librerías de cintas y discos ópticos

SAGE (Storage Appliance Grid Environment) Es una aplicación que maneja el almacenamiento distribuido en una base de datos Exadata

Segment Es una característica de almacenamiento de Oracle denominada en español segment que ocupa un espacio de algún objeto de la bases de datos de manera lógica ejemplos: taba, índice, procedimiento etc. Todo lo que ocupa un espacio en la estructura lógica de Oracle es un segment conformado por uno o más bloques Oracle.

Sinergia: Fenómeno en el que actúan varios factores en conjunto o influencias, observándose un resultado diferente de que hubiera resultado si cada uno de ellos fuera un proceso en ejecución por separado, el concepto se puso de moda en los sistemas en la década de los 90.

Soft Parsing Parseo suave significa que la consulta que ya está en la estructura de memoria de Oracle se vuelve a usar antes de ejecutar una consulta dura.

UNIX BSD Fue un sistema operativo UNIX con el nombre original de Berkeley Software Distribution, desarrollado por el Computer System Research Group de la universidad de Berkeley California y que sería la base de los sistemas operativos UNIX de la actualidad.